Optimal Testing for Properties of Distributions

Jayadev Acharya, Constantinos Daskalakis, Gautam Kamath
EECS, MIT
{jayadev, costis, g}@mit.edu

Abstract

Given samples from an unknown discrete distribution p, is it possible to distinguish whether p belongs to some class of distributions $\mathcal{C}$ versus p being far from every distribution in $\mathcal{C}$? This fundamental question has received tremendous attention in statistics, focusing primarily on asymptotic analysis, and in information theory and theoretical computer science, where the emphasis has been on small sample size and computational complexity. Nevertheless, even for basic properties of discrete distributions such as monotonicity, independence, log-concavity, unimodality, and monotone-hazard rate, the optimal sample complexity is unknown.

We provide a general approach via which we obtain sample-optimal and computationally efficient testers for all these distribution families. At the core of our approach is an algorithm which solves the following problem: Given samples from an unknown distribution p, and a known distribution q, are p and q close in χ^2-distance, or far in total variation distance?

The optimality of our testers is established by providing matching lower bounds, up to constant factors. Finally, a necessary building block for our testers and an important byproduct of our work are the first known computationally efficient proper learners for discrete log-concave, monotone hazard rate distributions.

1 Introduction

The quintessential scientific question is whether an unknown object has some property, i.e. whether a model from a specific class fits the object's observed behavior. If the unknown object is a probability distribution, p, to which we have sample access, we are typically asked to distinguish whether p belongs to some class $\mathcal{C}$ or whether it is sufficiently far from it.

This question has received tremendous attention in the field of statistics (see, e.g.,[1, 2]), where test statistics for important properties such as the ones we consider here have been proposed. Nevertheless, the emphasis has been on asymptotic analysis, characterizing the rates of convergence of test statistics under null hypotheses, as the number of samples tends to infinity. In contrast, we wish to study the following problem in the small sample regime:

> $\Pi(\mathcal{C}, \varepsilon)$: Given a family of distributions $\mathcal{C}$, some $\varepsilon > 0$, and sample access to an unknown distribution p over a discrete support, how many samples are required to distinguish between $p \in \mathcal{C}$ versus $d_{\mathrm{TV}}(p, \mathcal{C}) > \varepsilon$?[1]

The problem has been studied intensely in the literature on property testing and sublinear algorithms [3, 4, 5], where the emphasis has been on characterizing the optimal tradeoff between p's support size and the accuracy ε in the number of samples. Several results have been obtained, roughly clustering into three groups, where (i) $\mathcal{C}$ is the class of monotone distributions over $[n]$, or more

[1] We want success probability at least $2/3$, which can be boosted to $1 - \delta$ by repeating the test $O(\log(1/\delta))$ times and taking the majority.

generally a poset [6, 7]; (ii) $\mathcal{C}$ is the class of independent, or k-wise independent distributions over a hypergrid [8, 9]; and (iii) $\mathcal{C}$ contains a single-distribution q, and the problem becomes that of testing whether p equals q or is far from it [8, 10, 11].

With respect to (iii), [11] exactly characterizes the number of samples required to test identity to each distribution q, providing a single tester matching this bound simultaneously for all q. Nevertheless, this tester and its precursors are not applicable to the composite identity testing problem that we consider. If our class $\mathcal{C}$ were finite, we could test against each element in the class, albeit this would not necessarily be sample optimal. If our class $\mathcal{C}$ were a continuum, we would need *tolerant* identity testers, which tend to be more expensive in terms of sample complexity [12], and result in substantially suboptimal testers for the classes we consider. Or we could use approaches related to generalized likelihood ratio test, but their behavior is not well-understood in our regime, and optimizing likelihood over our classes becomes computationally intense.

Our Contributions We obtain sample-optimal and computationally efficient testers for $\Pi(\mathcal{C}, \varepsilon)$ for the most fundamental shape restrictions to a distribution. Our contributions are the following:

1. For a known distribution q over $[n]$, and sample access to p, we show that distinguishing the cases: (a) whether the χ^2-distance between p and q is at most $\varepsilon^2/2$, versus (b) the ℓ_1 distance between p and q is at least 2ε, requires $\Theta(\sqrt{n}/\varepsilon^2)$ samples. As a corollary, we obtain an alternate argument that shows that identity testing requires $\Theta(\sqrt{n}/\varepsilon^2)$ samples (previously shown in [11]).

2. For the class $\mathcal{C} = \mathcal{M}_n^d$ of monotone distributions over $[n]^d$ we require an optimal $\Theta\left(n^{d/2}/\varepsilon^2\right)$ number of samples, where prior work requires $\Omega\left(\sqrt{n}\log n/\varepsilon^6\right)$ samples for $d = 1$ and $\tilde{\Omega}\left(n^{d-1/2}\mathrm{poly}\left(1/\varepsilon\right)\right)$ for $d > 1$ [6, 7]. Our results improve the exponent of n with respect to d, shave all logarithmic factors in n, and improve the exponent of ε by at least a factor of 2.

   (a) A useful building block and interesting byproduct of our analysis is extending Birgé's oblivious decomposition for single-dimensional monotone distributions [13] to monotone distributions in $d \geq 1$, and to the stronger notion of χ^2-distance. See Section C.1.

   (b) Moreover, we show that $O(\log^d n)$ samples suffice to learn a monotone distribution over $[n]^d$ in χ^2-distance. See Lemma 3 for the precise statement.

3. For the class $\mathcal{C} = \Pi_d$ of product distributions over $[n_1] \times \cdots \times [n_d]$, our algorithm requires $O\left(\left(\left(\prod_\ell n_\ell\right)^{1/2} + \sum_\ell n_\ell\right)/\varepsilon^2\right)$ samples. We note that a product distribution is one where all marginals are independent, so this is equivalent to testing if a collection of random variables are all independent. In the case where n_ℓ's are large, then the first term dominates, and the sample complexity is $O((\prod_\ell n_\ell)^{1/2}/\varepsilon^2)$. In particular, when d is a constant and all n_ℓ's are equal to n, we achieve the optimal sample complexity of $\Theta(n^{d/2}/\varepsilon^2)$. To the best of our knowledge, this is the first result for $d \geq 3$, and when $d = 2$, this improves the previously known complexity from $O\left(\frac{n}{\varepsilon^6}\mathrm{polylog}(n/\varepsilon)\right)$ [8, 14], significantly improving the dependence on ε and shaving all logarithmic factors.

4. For the classes $\mathcal{C} = \mathcal{LCD}_n$, $\mathcal{C} = \mathcal{MHR}_n$ and $\mathcal{C} = \mathcal{U}_n$ of log-concave, monotone-hazard-rate and unimodal distributions over $[n]$, we require an optimal $\Theta\left(\sqrt{n}/\varepsilon^2\right)$ number of samples. Our testers for $\mathcal{LCD}_n$ and $\mathcal{C} = \mathcal{MHR}_n$ are to our knowledge the first for these classes for the low sample regime we are studying—see [15] and its references for statistics literature on the asymptotic regime. Our tester for $\mathcal{U}_n$ improves the dependence of the sample complexity on ε by at least a factor of 2 in the exponent, and shaves all logarithmic factors in n, compared to testers based on testing monotonicity.

   (a) A useful building block and important byproduct of our analysis are the first computationally efficient algorithms for properly learning log-concave and monotone-hazard-rate distributions, to within ε in total variation distance, from $\mathrm{poly}(1/\varepsilon)$ samples, independent of the domain size n. See Corollaries 4 and 6. Again, these are the first computationally efficient algorithms to our knowledge in the low sample regime. [16] provide algorithms for density estimation, which are non-proper, i.e. will approximate an unknown distribution from these classes with a distribution that does not belong to these classes. On the other hand, the statistics literature focuses on maximum-likelihood estimation in the asymptotic regime—see e.g. [17] and its references.

5. For all the above classes we obtain matching lower bounds, showing that the sample complexity of our testers is optimal with respect to n, ε and when applicable d. See Section 8. Our lower bounds are based on extending Paninski's lower bound for testing uniformity [10].

Our Techniques At the heart of our tester lies a novel use of the χ^2 statistic. Naturally, the χ^2 and its related ℓ_2 statistic have been used in several of the afore-cited results. We propose a new use of the χ^2 statistic enabling our optimal sample complexity. The essence of our approach is to first draw a small number of samples (independent of n for log-concave and monotone-hazard-rate distributions and only logarithmic in n for monotone and unimodal distributions) to approximate the unknown distribution p in χ^2 distance. If $p \in \mathcal{C}$, our learner is required to output a distribution q that is $O(\varepsilon)$-close to $\mathcal{C}$ in total variation and $O(\varepsilon^2)$-close to p in χ^2 distance. Then some analysis reduces our testing problem to distinguishing the following cases:

- p and q are $O(\varepsilon^2)$-close in χ^2 distance; this case corresponds to $p \in \mathcal{C}$.
- p and q are $\Omega(\varepsilon)$-far in total variation distance; this case corresponds to $d_{\mathrm{TV}}(p, \mathcal{C}) > \varepsilon$.

We draw a comparison with *robust identity testing*, in which one must distinguish whether p and q are $c_1\varepsilon$-close or $c_2\varepsilon$-far in total variation distance, for constants $c_2 > c_1 > 0$. In [12], Valiant and Valiant show that $\Omega(n/\log n)$ samples are required for this problem – a nearly-linear sample complexity, which may be prohibitively large in many settings. In comparison, the problem we study tests for χ^2 closeness rather than total variation closeness: a relaxation of the previous problem. However, our tester demonstrates that this relaxation allows us to achieve a substantially sublinear complexity of $O(\sqrt{n}/\varepsilon^2)$. On the other hand, this relaxation is still tight enough to be useful, demonstrated by our application in obtaining sample-optimal testers.

We note that while the χ^2 statistic for testing hypothesis is prevalent in statistics providing optimal error exponents in the large-sample regime, to the best of our knowledge, in the small-sample regime, *modified-versions* of the χ^2 statistic have only been recently used for *closeness-testing* in [18, 19] and for testing uniformity of monotone distributions in [20]. In particular, [18] design an unbiased statistic for estimating the χ^2 distance between two *unknown* distributions.

Organization In Section 4, we show that a version of the χ^2 statistic, appropriately excluding certain elements of the support, is sufficiently well-concentrated to distinguish between the above cases. Moreover, the sample complexity of our algorithm is optimal for most classes. Our base tester is combined with the afore-mentioned extension of Birgé's decomposition theorem to test monotone distributions in Section 5 (see Theorem 2 and Corollary 1), and is also used to test independence of distributions in Section 6 (see Theorem 3).

In Section 7, we give our results on testing unimodal, log-concave and monotone hazard rate distributions. Naturally, there are several bells and whistles that we need to add to the above skeleton to accommodate all classes of distributions that we are considering. In Remark 1 we mention the additional modifications for these classes.

Related Work. For the problems that we study in this paper, we have provided the related works in the previous section along with our contributions. We cannot do justice to the role of shape restrictions of probability distributions in probabilistic modeling and testing. It suffices to say that the classes of distributions that we study are fundamental, motivating extensive literature on their learning and testing [21]. In the recent times, there has been work on shape restricted statistics, pioneered by Jon Wellner, and others. [22, 23] study estimation of monotone and k-monotone densities, and [24, 25] study estimation of log-concave distributions. Due to the sheer volume of literature in statistics in this field, we will restrict ourselves to those already referenced.

As we have mentioned, statistics has focused on the asymptotic regime as the number of samples tends to infinity. Instead we are considering the low sample regime and are more stringent about the behavior of our testers, requiring 2-sided guarantees. We want to accept if the unknown distribution is in our class of interest, and also reject if it is far from the class. For this problem, as discussed above, there are few results when $\mathcal{C}$ is a whole class of distributions. Closer related to our paper is the line of papers [6, 7, 26] for monotonicity testing, albeit these papers have sub-optimal sample complexity as discussed above. Testing independence of random variables has a long history in statisics [27, 28]. The theoretical computer science community has also considered the problem of

testing independence of two random variables [8, 14]. While our results sharpen the case where the variables are over domains of equal size, they demonstrate an interesting asymmetric upper bound when this is not the case. More recently, Acharya and Daskalakis provide optimal testers for the family of Poisson Binomial Distributions [29].

Finally, contemporaneous work of Canonne et al [30] provides a generic algorithm and lower bounds for the single-dimensional families of distributions considered here. We note that their algorithm has a sample complexity which is suboptimal in both n and ε, while our algorithms are optimal. Their algorithm also extends to mixtures of these classes, though some of these extensions are not computationally efficient. They also provide a framework for proving lower bounds, giving the optimal bounds for many classes when ε is sufficiently large with respect to $1/n$. In comparison, we provide these lower bounds unconditionally by modifying Paninski's construction [10] to suit the classes we consider.

2 Preliminaries

We use the following probability distances in our paper.

The *total variation distance* between distributions p and q is $d_{\mathrm{TV}}(p,q) \stackrel{\mathrm{def}}{=} \sup_A |p(A) - q(A)| = \frac{1}{2}\|p-q\|_1$. The χ^2*-distance* between p and q over $[n]$ is defined as $\chi^2(p,q) \stackrel{\mathrm{def}}{=} \sum_{i\in[n]} \frac{(p_i-q_i)^2}{q_i}$. The *Kolmogorov distance* between two probability measures p and q over an ordered set ($e.g.$, $\mathbf{R}$) with cumulative density functions F_p and F_q is $d_{\mathrm{K}}(p,q) \stackrel{\mathrm{def}}{=} \sup_{x\in\mathbb{R}} |F_p(x) - F_q(x)|$.

Our paper is primarily concerned with testing against classes of distributions, defined formally as:

Definition 1. *Given $\varepsilon \in (0,1]$ and sample access to a distribution p, an algorithm is said to* test *a class $\mathcal{C}$ if it has the following guarantees:*

- *If $p \in \mathcal{C}$, the algorithm outputs* ACCEPT *with probability at least $2/3$;*
- *If $d_{\mathrm{TV}}(p,\mathcal{C}) \geq \varepsilon$, the algorithm outputs* REJECT *with probability at least $2/3$.*

We note the following useful relationships between these distances [31]:

Proposition 1. $d_{\mathrm{K}}(p,q)^2 \leq d_{\mathrm{TV}}(p,q)^2 \leq \frac{1}{4}\chi^2(p,q)$.

Definition 2. *An* η-effective support *of a distribution p is any set S such that $p(S) \geq 1-\eta$.*

The *flattening* of a function f over a subset S is the function $\bar{f}$ such that $\bar{f}_i = p(S)/|S|$.

Definition 3. *Let p be a distribution, and support $I_1, \ldots$ is a partition of the domain. The flattening of p with respect to $I_1, \ldots$ is the distribution $\bar{p}$ which is the flattening of p over the intervals $I_1, \ldots$.*

Poisson Sampling Throughout this paper, we use the standard Poissonization approach. Instead of drawing exactly m samples from a distribution p, we first draw $m' \sim \mathrm{Poisson}(\mathrm{m})$, and then draw m' samples from p. As a result, the number of times different elements in the support of p occur in the sample become independent, giving much simpler analyses. In particular, the number of times we will observe domain element i will be distributed as $\mathrm{Poisson}(\mathrm{mp_i})$, independently for each i. Since $\mathrm{Poisson}(\mathrm{m})$ is tightly concentrated around m, this additional flexibility comes only at a sub-constant cost in the sample complexity with an inversely exponential in m, additive increase in the error probability.

3 The Testing Algorithm – An Overview

Our algorithm for testing a class $\mathcal{C}$ can be decomposed into three steps.

Near-proper learning in χ^2-distance. Our first step requires learning with very specific guarantees. Given sample access to $p \in \mathcal{C}$, we wish to output q such that (i) q is *close* to $\mathcal{C}$ in total variation distance, and (ii) p and q are $O(\varepsilon^2)$-close in χ^2-distance on an $\varepsilon-$effective support[2] of p. When

[2]We also require the algorithm to output a description of an effective support for which this property holds. This requirement can be slightly relaxed, as we show in our results for testing unimodality.

p is not in $\mathcal{C}$, we do not guarantee anything about q. From an information theoretic standpoint, this problem is harder than learning the distribution in total variation, since χ^2-distance is more restrictive than total variation distance. Nonetheless, for the structured classes we consider, we are able to learn in χ^2 by modifying the approaches to learn in total variation.

Computation of distance to class. The next step is to see if the hypothesis q is close to the class $\mathcal{C}$ or not. Since we have an explicit description of q, this step requires no further samples from p, i.e. it is purely computational. If we find that q is far from the class $\mathcal{C}$, then it must be that $p \notin \mathcal{C}$, as otherwise the guarantees from the previous step would imply that q is close to $\mathcal{C}$. Thus, if it is not, we can terminate the algorithm at this point.

χ^2-testing. At this point, the previous two steps guarantee that our distribution q is such that:

- If $p \in \mathcal{C}$, then p and q are close in χ^2 distance on a (known) effective support of p;

- If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, then p and q are far in total variation distance.

We can distinguish between these two cases using $O(\sqrt{n}/\varepsilon^2)$ samples with a simple statistical χ^2-test, that we describe in Section 4.

Using the above three-step approach, our tester, as described in the next section, can directly test monotonicity, log-concavity, and monotone hazard rate. With an extra trick, using Kolmogorov's max inequality, it can also test unimodality.

4 A Robust χ^2-ℓ_1 Identity Test

Our main result in this section is Theorem 1.

Theorem 1. *Given $\varepsilon \in (0, 1]$, a class of probability distributions $\mathcal{C}$, sample access to a distribution p, and an explicit description of a distribution q, both over $[n]$ with the following properties:*

> ***Property 1.*** $d_{\mathrm{TV}}(q, \mathcal{C}) \leq \frac{\varepsilon}{2}$.
>
> ***Property 2.*** *If $p \in \mathcal{C}$, then $\chi^2(p, q) \leq \frac{\varepsilon^2}{500}$.*

Then there exists an algorithm such that: If $p \in \mathcal{C}$, it outputs ACCEPT *with probability at least $2/3$; If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, it outputs* REJECT *with probability at least $2/3$. The time and sample complexity of this algorithm are $O\left(\sqrt{n}/\varepsilon^2\right)$.*

Proof. Algorithm 1 describes a χ^2 testing procedure that gives the guarantee of the theorem.

Algorithm 1 Chi-squared testing algorithm

1: **Input:** ε; an explicit distribution q; (Poisson) m samples from a distribution p, where N_i denotes the number of occurrences of the ith domain element.
2: $\mathcal{A} \leftarrow \{i : q_i \geq \varepsilon^2/50n\}$
3: $Z \leftarrow \sum_{i \in \mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
4: **if** $Z \leq m\varepsilon^2/10$ **return** close
5: **else return** far

In Section A we compute the mean and variance of the statistic Z (defined in Algorithm 1) as:

$$\mathbb{E}[Z] = m \cdot \sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} = m \cdot \chi^2(p_{\mathcal{A}}, q_{\mathcal{A}}), \quad \mathrm{Var}[Z] = \sum_{i \in \mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m \cdot \frac{p_i \cdot (p_i - q_i)^2}{q_i^2} \right] \quad (1)$$

where by $p_{\mathcal{A}}$ and $q_{\mathcal{A}}$ we denote respectively the vectors p and q restricted to the coordinates in $\mathcal{A}$, and we slightly abuse notation when we write $\chi^2(p_{\mathcal{A}}, q_{\mathcal{A}})$, as these do not then correspond to probability distributions.

Lemma 1 demonstrates the separation in the means of the statistic Z in the two cases of interest, *i.e.*, $p \in \mathcal{C}$ versus $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, and Lemma 2 shows the separation in the variances in the two cases. These two results are proved in Section B.

Lemma 1. *If $p \in \mathcal{C}$, then $\mathbb{E}[Z] \le m\varepsilon^2/500$. If $d_{\mathrm{TV}}(p,\mathcal{C}) \ge \varepsilon$, then $\mathbb{E}[Z] \ge m\varepsilon^2/5$.*

Lemma 2. *Let $m \ge 20000\sqrt{n}/\varepsilon^2$. If $p \in \mathcal{C}$ then $\mathrm{Var}[Z] \le \frac{1}{500000} m^2\varepsilon^4$. If $d_{\mathrm{TV}}(p,\mathcal{C}) \ge \varepsilon$, then $\mathrm{Var}[Z] \le \frac{1}{100} E[Z]^2$.*

Assuming Lemmas 1 and 2, Theorem 1 is now a simple application of Chebyshev's inequality.

When $p \in \mathcal{C}$, we have that $\mathbb{E}[Z] + \sqrt{3\,\mathrm{Var}[Z]} \le \left(1/500 + \sqrt{3/500000}\right) m\varepsilon^2 \le m\varepsilon^2/200$. Thus, Chebyshev's inequality gives

$$\Pr\left[Z \ge m\varepsilon^2/10\right] \le \Pr\left[Z \ge m\varepsilon^2/200\right] \le \Pr\left[Z - \mathbb{E}[Z] \ge \sqrt{3}\,\mathrm{Var}[Z]^{1/2}\right] \le 1/3.$$

When $d_{\mathrm{TV}}(p,\mathcal{C}) \ge \varepsilon$, $\mathbb{E}[Z] - \sqrt{3\,\mathrm{Var}[Z]} \ge \left(1 - \sqrt{3/100}\right) E[Z] \ge 3m\varepsilon^2/20$. Therefore,

$$\Pr\left[Z \le m\varepsilon^2/10\right] \le \Pr\left[Z \le 3m\varepsilon^2/20\right] \le \Pr\left[Z - \mathbb{E}[Z] \le -\sqrt{3}\,\mathrm{Var}[Z]^{1/2}\right] \le 1/3. \qquad \square$$

This proves the correctness of Algorithm 1. For the running time, we divide the summation in Z into the elements for which $N_i > 0$ and $N_i = 0$. When $N_i = 0$, the contribution of the term to the summation is mq_i, and we can sum them up by subtracting the total probability of all elements appearing at least once from 1.

Remark 1. *To apply Theorem 1, we need to learn distribution in $\mathcal{C}$ and find a q that is $O(\varepsilon^2)$-close in χ^2-distance to p. For the class of monotone distributions, we are able to efficiently obtain such a q, which immediately implies sample-optimal learning algorithms for this class. However, for some classes, we may not be able to learn a q with such strong guarantees, and we must consider modifications to our base testing algorithm.*

For example, for log-concave and monotone hazard rate distributions, we can obtain a distribution q and a set S with the following guarantees:

- If $p \in \mathcal{C}$, then $\chi^2(p_S, q_S) \le O(\varepsilon^2)$ and $p(S) \ge 1 - O(\varepsilon)$;

- If $d_{\mathrm{TV}}(p,\mathcal{C}) \ge \varepsilon$, then $d_{\mathrm{TV}}(p,q) \ge \varepsilon/2$. In this scenario, the tester will simply pretend that the support of p and q is S, ignoring any samples and support elements in $[n] \setminus S$. Analysis of this tester is extremely similar to Theorem 1. In particular, we can still show that the statistic Z will be separated in the two cases. When $p \in \mathcal{C}$, excluding $[n] \setminus S$ will only reduce Z. On the other hand, when $d_{\mathrm{TV}}(p,\mathcal{C}) \ge \varepsilon$, since $p(S) \ge 1 - O(\varepsilon)$, p and q must still be far on the remaining support, and we can show that Z is still sufficiently large. Therefore, a small modification allows us to handle this case with the same sample complexity of $O(\sqrt{n}/\varepsilon^2)$.

For unimodal distributions, we are even unable to identify a large enough subset of the support where the χ^2 approximation is guaranteed to be tight. But we can show that there exists a light enough piece of the support (in terms of probability mass under p) that we can exclude to make the χ^2 approximation tight. Given that we only use Chebyshev's inequality to prove the concentration of the test statistic, it would seem that our lack of knowledge of the piece to exclude would involve a union bound and a corresponding increase in the required number of samples. We avoid this through a careful application of Kolmogorov's max inequality in our setting. See Theorem 7 of Section 7.

5 Testing Monotonicity

As the first application of our testing framework, we will demonstrate how to test for monotonicity. Let $d \ge 1$, and $\mathbf{i} = (i_1, \ldots, i_d), \mathbf{j} = (j_1, \ldots, j_d) \in [n]^d$. We say $\mathbf{i} \succcurlyeq \mathbf{j}$ if $i_l \ge j_l$ for $l = 1, \ldots, d$. A distribution p over $[n]^d$ is monotone (decreasing) if for all $\mathbf{i} \succcurlyeq \mathbf{j}$, $p_{\mathbf{i}} \le p_{\mathbf{j}}$[3].

We follow the steps in the overview. The learning result we show is as follows (proved in Section C).

[3]This definition describes monotone non-increasing distributions. By symmetry, identical results hold for monotone non-decreasing distributions.

Lemma 3. *Let $d \geq 1$. There is an algorithm that takes $m = O((d\log(n)/\varepsilon^2)^d/\varepsilon^2)$ samples from a distribution p over $[n]^d$, and outputs a distribution q such that if p is monotone, then with probability at least $5/6$, $\chi^2(p,q) \leq \frac{\varepsilon^2}{500}$. Furthermore, the distance of q to monotone distributions can be computed in time $poly(m)$.*

This accomplishes the first two steps in the overview. In particular, if the distance of q from monotone distributions is more than $\varepsilon/2$, we declare that p is not monotone. Therefore, Property 1 in Theorem 1 is satisfied, and the lemma states that Property 2 holds with probability at least $5/6$. We then proceed to the $\chi^2 - \ell_1$ test. At this point, we have precisely the guarantees needed to apply Theorem 1 over $[n]^d$, directly implying our main result of this section:

Theorem 2. *For any $d \geq 1$, there exists an algorithm for testing monotonicity over $[n]^d$ with sample complexity*

$$O\left(\frac{n^{d/2}}{\varepsilon^2} + \left(\frac{d\log n}{\varepsilon^2}\right)^d \cdot \frac{1}{\varepsilon^2}\right)$$

and time complexity $O\left(n^{d/2}/\varepsilon^2 + \mathrm{poly}(\log n, 1/\varepsilon)^d\right)$.

In particular, this implies the following optimal algorithms for monotonicity testing for all $d \geq 1$:

Corollary 1. *Fix any $d \geq 1$, and suppose $\varepsilon > \sqrt{d\log n}/n^{1/4}$. Then there exists an algorithm for testing monotonicity over $[n]^d$ with sample complexity $O\left(n^{d/2}/\varepsilon^2\right)$.*

We note that the class of monotone distributions is the simplest of the classes we consider. We now consider testing for log-concavity, monotone hazard rate, and unimodality, all of which are much more challenging to test. In particular, these classes require a more sophisticated structural understanding, more complex proper χ^2-learning algorithms, and non-trivial modifications to our χ^2-tester. We have already given some details on the required adaptations to the tester in Remark 1.

Our algorithms for learning these classes use convex programming. One of the main challenges is to enforce log-concavity of the PDF when learning $\mathcal{LCD}_n$ (respectively, of the CDF when learning $\mathcal{MHR}_n$), while simultaneously enforcing closeness in total variation distance. This involves a careful choice of our variables, and we exploit structural properties of the classes to ensure the soundness of particular Taylor approximations. We encourage the reader to refer to the proofs of Theorems 7, 8, and 9 for more details.

6 Testing Independence of Random Variables

Let $\mathcal{X} \stackrel{\text{def}}{=} [n_1] \times \ldots \times [n_d]$, and let Π_d be the class of all product distributions over $\mathcal{X}$. Similar to learning monotone distributions in χ^2 distance we prove the following result in Section E.

Lemma 4. *There is an algorithm that takes $O\left((\sum_{\ell=1}^d n_\ell)/\varepsilon^2\right)$ samples from a distribution p and outputs a $q \in \Pi_d$ such that if $p \in \Pi_d$, then with probability at least $5/6$, $\chi^2(p,q) \leq O(\varepsilon^2)$.*

The distribution q always satisfies Property 1 since it is in Π_d, and by this lemma, with probability at least $5/6$ satisfies Property 2 in Theorem 1. Therefore, we obtain the following result.

Theorem 3. *For any $d \geq 1$, there exists an algorithm for testing independence of random variables over $[n_1] \times \ldots [n_d]$ with sample and time complexity $O\left(\left((\prod_{\ell=1}^d n_\ell)^{1/2} + \sum_{\ell=1}^d n_\ell\right)/\varepsilon^2\right)$.*

When $d = 2$ and $n_1 = n_2 = n$ this improves the result of [8] for testing independence of two random variables.

Corollary 2. *Testing if two distributions over $[n]$ are independent has sample complexity $\Theta(n/\varepsilon^2)$.*

7 Testing Unimodality, Log-Concavity and Monotone Hazard Rate

Unimodal distributions over $[n]$ (denoted by $\mathcal{U}_n$) are all distributions p for which there exists an i^* such that p_i is non-decreasing for $i \leq i^*$ and non-increasing for $i \geq i^*$. Log-concave distributions over $[n]$ (denoted by $\mathcal{LCD}_n$), is the sub-class of unimodal distributions for which $p_{i-1}p_{i+1} \leq p_i^2$.

Monotone hazard rate (MHR) distributions over $[n]$ (denoted by $\mathcal{MHR}_n$), are distributions p with CDF F for which $i < j$ implies $\frac{f_i}{1-F_i} \leq \frac{f_j}{1-F_j}$.

The following theorem bounds the complexity of testing these classes (for moderate ε).

Theorem 4. *Suppose $\varepsilon > n^{-1/5}$. For each of the classes, unimodal, log-concave, and MHR, there exists an algorithm for testing the class over $[n]$ with sample complexity $O(\sqrt{n}/\varepsilon^2)$.*

This result is a corollary of the specific results for each class, which is proved in the appendix. In particular, a more complete statement for unimodality, log-concavity and monotone-hazard rate, with precise dependence on both n and ε is given in Theorems 7, 8 and 9 respectively. We mention some key points about each class, and refer the reader to the respective appendix for further details.

Testing Unimodality Using a union bound argument, one can use the results on testing monotonicity to give an algorithm with $O\left(\sqrt{n}\log n/\varepsilon^2\right)$ samples. However, this is unsatisfactory, since our lower bound (and as we will demonstrate, the true complexity of this problem) is $\sqrt{n}/\varepsilon^2$. We overcome the logarithmic barrier introduced by the union bound, by employing a non-oblivious decomposition of the domain, and using Kolmogorov's max-inequality.

Testing Log-Concavity The key step is to design an algorithm to learn a log-concave distribution in χ^2 distance. We formulate the problem as a linear program in the logarithms of the distribution and show that using $O(1/\varepsilon^5)$ samples, it is possible to output a log-concave distribution that has a χ^2 distance at most $O(\varepsilon^2)$ from the underlying log-concave distribution.

Testing Monotone Hazard Rate For learning MHR distributions in χ^2 distance, we formulate a linear program in the logarithms of the CDF and show that using $O(\log(n/\varepsilon)/\varepsilon^5)$ samples, it is possible to output a MHR distribution that has a χ^2 distance at most $O(\varepsilon^2)$ from the underlying MHR distribution.

8 Lower Bounds

We now prove sharp lower bounds for the classes of distributions we consider. We show that the example studied by Paninski [10] to prove lower bounds on testing uniformity can be used to prove lower bounds for the classes we consider. They consider a class $\mathcal{Q}$ consisting of $2^{n/2}$ distributions defined as follows. Without loss of generality assume that n is even. For each of the $2^{n/2}$ vectors $z_0 z_1 \ldots z_{n/2-1} \in \{-1,1\}^{n/2}$, define a distribution $q \in \mathcal{Q}$ over $[n]$ as follows.

$$q_i = \begin{cases} \frac{(1+z_\ell c\varepsilon)}{n} & \text{for } i = 2\ell+1 \\ \frac{(1-z_\ell c\varepsilon)}{n} & \text{for } i = 2\ell. \end{cases} \tag{2}$$

Each distribution in $\mathcal{Q}$ has a total variation distance $c\varepsilon/2$ from U_n, the uniform distribution over $[n]$. By choosing c to be an appropriate constant, Paninski [10] showed that a distribution picked uniformly at random from $\mathcal{Q}$ cannot be distinguished from U_n with fewer than $\sqrt{n}/\varepsilon^2$ samples with probability at least $2/3$.

Suppose $\mathcal{C}$ is a class of distributions such that (i) The uniform distribution U_n is in $\mathcal{C}$, (ii) For appropriately chosen c, $d_{\mathrm{TV}}(\mathcal{C},\mathcal{Q}) \geq \varepsilon$, then testing $\mathcal{C}$ is not easier than distinguishing U_n from $\mathcal{Q}$. Invoking [10] immediately implies that testing the class $\mathcal{C}$ requires $\Omega(\sqrt{n}/\varepsilon^2)$ samples.

The lower bounds for all the one dimensional distributions will follow directly from this construction, and for testing monotonicity in higher dimensions, we extend this construction to $d \geq 1$, appropriately. These arguments are proved in Section H, leading to the following lower bounds for testing these classes:

Theorem 5.

- *For any $d \geq 1$, any algorithm for testing monotonicity over $[n]^d$ requires $\Omega(n^{d/2}/\varepsilon^2)$ samples.*
- *For $d \geq 1$, testing independence over $[n_1]\times\cdots\times[n_d]$ requires $\Omega\left((n_1 n_2 \ldots n_d)^{1/2}/\varepsilon^2\right)$ samples.*
- *Testing unimodality, log-concavity, or monotone hazard rate over $[n]$ needs $\Omega(\sqrt{n}/\varepsilon^2)$ samples.*

References

[1] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[2] E. Lehmann and J. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[3] E. Fischer, "The art of uninformed decisions: A primer to property testing," *Science*, 2001.

[4] R. Rubinfeld, "Sublinear-time algorithms," in *International Congress of Mathematicians*, 2006.

[5] C. L. Canonne, "A survey on distribution testing: your data is big, but is it blue," *ECCC*, 2015.

[6] T. Batu, R. Kumar, and R. Rubinfeld, "Sublinear algorithms for testing monotone and unimodal distributions," in *Proceedings of STOC*, 2004.

[7] A. Bhattacharyya, E. Fischer, R. Rubinfeld, and P. Valiant, "Testing monotonicity of distributions over general partial orders," in *ICS*, 2011, pp. 239–252.

[8] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proceedings of FOCS*, 2001.

[9] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie, "Testing k-wise and almost k-wise independence," in *Proceedings of STOC*, 2007.

[10] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data." *IEEE Transactions on Information Theory*, vol. 54, no. 10, 2008.

[11] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," in *FOCS*, 2014.

[12] ——, "Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of STOC*, 2011.

[13] L. Birgé, "Estimating a density under order restrictions: Nonasymptotic minimax risk," *The Annals of Statistics*, vol. 15, no. 3, pp. 995–1012, September 1987.

[14] R. Levi, D. Ron, and R. Rubinfeld, "Testing properties of collections of distributions," *Theory of Computing*, vol. 9, no. 8, pp. 295–347, 2013.

[15] P. Hall and I. Van Keilegom, "Testing for monotone increasing hazard rate," *Annals of Statistics*, pp. 1109–1137, 2005.

[16] S. O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun, "Learning mixtures of structured distributions over discrete domains," in *Proceedings of SODA*, 2013.

[17] M. Cule and R. Samworth, "Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density," *Electronic Journal of Statistics*, vol. 4, pp. 254–270, 2010.

[18] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh, "Competitive classification and closeness testing," in *COLT*, 2012, pp. 22.1–22.18.

[19] S. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proceedings of SODA*, 2014, pp. 1193–1203.

[20] J. Acharya, A. Jafarpour, A. Orlitsky, and A. Theertha Suresh, "A competitive test for uniformity of monotone distributions," in *Proceedings of AISTATS*, 2013, pp. 57–65.

[21] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical Inference under Order Restrictions*. New York: Wiley, 1972.

[22] H. K. Jankowski and J. A. Wellner, "Estimation of a discrete monotone density," *Electronic Journal of Statistics*, vol. 3, pp. 1567–1605, 2009.

[23] F. Balabdaoui and J. A. Wellner, "Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds," *Statistica Neerlandica*, vol. 64, no. 1, pp. 45–70, 2010.

[24] F. Balabdaoui, H. Jankowski, and K. Rufibach, "Maximum likelihood estimation and confidence bands for a discrete log-concave distribution," 2011. [Online]. Available: http://arxiv.org/abs/1107.3904v1

[25] A. Saumard and J. A. Wellner, "Log-concavity and strong log-concavity: a review," *Statistics Surveys*, vol. 8, pp. 45–114, 2014.

[26] M. Adamaszek, A. Czumaj, and C. Sohler, "Testing monotone continuous distributions on high-dimensional real cubes," in *SODA*, 2010, pp. 56–65.

[27] J. N. Rao and A. J. Scott, "The analysis of categorical data from complex sample surveys," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 221–230, 1981.

[28] A. Agresti and M. Kateri, *Categorical data analysis*. Springer, 2011.

[29] J. Acharya and C. Daskalakis, "Testing Poisson Binomial Distributions," in *Proceedings of SODA*, 2015, pp. 1829–1840.

[30] C. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, "Testing shape restrictions of discrete distributions," *arXiv preprint arXiv:1507.03558*, 2015.

[31] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, dec 2002.

[32] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Efficient compression of monotone and m-modal distributions," in *ISIT*, 2014.

[33] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *COLT*, 2015.

[34] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 07 1990.

A Moments of the Chi-Squared Statistic

We analyze the mean and variance of the statistic

$$Z = \sum_{i \in \mathcal{A}} \frac{(X_i - mq_i)^2 - X_i}{mq_i},$$

where each X_i is independently distributed according to $\mathrm{Poisson}(mp_i)$.

We start with the mean:

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{i \in \mathcal{A}} \mathbb{E}\left[\frac{(X_i - mq_i)^2 - X_i}{mq_i}\right] \\
&= \sum_{i \in \mathcal{A}} \frac{\mathbb{E}\left[X_i^2\right] - 2mq_i\mathbb{E}[X_i] + m^2q_i^2 - \mathbb{E}[X_i]}{mq_i} \\
&= \sum_{i \in \mathcal{A}} \frac{m^2p_i^2 + mp_i - 2m^2q_ip_i + m^2q_i^2 - mp_i}{mq_i} \\
&= m\sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} \\
&= m \cdot \chi^2(p_{\mathcal{A}}, q_{\mathcal{A}})
\end{aligned}$$

Next, we analyze the variance. Let $\lambda_i = \mathbb{E}[X_i] = mp_i$ and $\lambda_i' = mq_i$.

$$\begin{aligned}
\mathrm{Var}[Z] &= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} \mathrm{Var}\left[(X_i - \lambda_i)^2 + 2(X_i - \lambda_i)(\lambda_i - \lambda_i') - (X_i - \lambda_i)\right] \\
&= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} \mathrm{Var}\left[(X_i - \lambda_i)^2 + (X_i - \lambda_i)(2\lambda_i - 2\lambda_i' - 1)\right] \\
&= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} \mathbb{E}\left[(X_i - \lambda_i)^4 + 2(X_i - \lambda_i)^3(2\lambda_i - 2\lambda_i' - 1) + (X_i - \lambda_i)^2(2\lambda_i - 2\lambda_i' - 1)^2 - \lambda_i^2\right] \\
&= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} [3\lambda_i^2 + \lambda_i + 2\lambda_i(2\lambda_i - 2\lambda_i' - 1) + \lambda_i(2\lambda_i - 2\lambda_i' - 1)^2 - \lambda_i^2] \\
&= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} [2\lambda_i^2 + \lambda_i + 4\lambda_i(\lambda_i - \lambda_i') - 2\lambda_i + \lambda_i(4(\lambda_i - \lambda_i')^2 - 4(\lambda_i - \lambda_i') + 1)] \\
&= \sum_{i \in \mathcal{A}} \frac{1}{\lambda_i'^2} [2\lambda_i^2 + 4\lambda_i(\lambda_i - \lambda_i')^2] \\
&= \sum_{i \in \mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m \cdot \frac{p_i \cdot (p_i - q_i)^2}{q_i^2}\right] \qquad (3)
\end{aligned}$$

The third equality is by noting the random variable has expectation λ_i and the fourth equality substitutes the values of centralized moments of the Poisson distribution.

B Analysis of our χ^2-Test Statistic

We first prove the key lemmas in the analysis of our χ^2-test.

Proof of Lemma 1: The former case is straightforward from (1) and Property 2 of q.

We turn to the latter case. Recall that $\mathcal{A} = \{i : q_i \geq \varepsilon^2/50n\}$, and thus $q(\bar{\mathcal{A}}) \leq \varepsilon^2/50$. We first show that $d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geq \frac{6\varepsilon}{25}$, where $p_{\mathcal{A}}, q_{\mathcal{A}}$ are defined as above and in our slight abuse of notation we use $d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}})$ for non-probability vectors to denote $\frac{1}{2}\|p_{\mathcal{A}} - q_{\mathcal{A}}\|_1$.

Partitioning the support into $\mathcal{A}$ and $\bar{\mathcal{A}}$, we have

$$d_{\mathrm{TV}}(p, q) = d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) + d_{\mathrm{TV}}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}). \qquad (4)$$

We consider the following cases separately:

- $p(\bar{\mathcal{A}}) \le \varepsilon/2$**:** In this case,

$$d_{\mathrm{TV}}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) = \frac{1}{2}\sum_{i\in\bar{\mathcal{A}}} |p_i - q_i| \le \frac{1}{2}(p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}})) \le \frac{1}{2}\left(\frac{\varepsilon}{2} + \frac{\varepsilon^2}{50}\right) = \frac{13\varepsilon}{50}.$$

  Plugging this in (4), and using the fact that $d_{\mathrm{TV}}(p,q) \ge \varepsilon$ shows that $d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \ge \frac{6\varepsilon}{25}$.

- $p(\bar{\mathcal{A}}) > \varepsilon/2$**:** In this case, by the reverse triangle inequality,

$$d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \ge \frac{1}{2}(q(\mathcal{A}) - p(\mathcal{A})) \ge \frac{1}{2}((1 - \varepsilon^2/50) - (1 - \varepsilon/2)) = \frac{6\varepsilon}{25}.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\chi^2(p_{\mathcal{A}}, q_{\mathcal{A}}) &\ge 4\frac{d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}})^2}{q(\mathcal{A})} \\ &\ge \frac{\varepsilon^2}{5}.\end{aligned}$$

Plugging in (1) proves the result. □

Proof of Lemma 2: We bound the terms of (1) separately, starting with the first.

$$\begin{aligned}
2\sum_{i\in\mathcal{A}} \frac{p_i^2}{q_i^2} &= 2\sum_{i\in\mathcal{A}} \left(\frac{(p_i - q_i)^2}{q_i^2} + \frac{2p_iq_i - q_i^2}{q_i^2}\right) \\
&= 2\sum_{i\in\mathcal{A}} \left(\frac{(p_i - q_i)^2}{q_i^2} + \frac{2q_i(p_i - q_i) + q_i^2}{q_i^2}\right) \\
&\le 2n + 2\sum_{i\in\mathcal{A}} \left(\frac{(p_i - q_i)^2}{q_i^2} + 2\frac{(p_i - q_i)}{q_i}\right) && (5) \\
&\le 4n + 4\sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^2}{q_i^2} && (6) \\
&\le 4n + \frac{200n}{\varepsilon^2}\sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^2}{q_i} && (7) \\
&= 4n + \frac{200n}{\varepsilon^2}\frac{E[Z]}{m} && (8) \\
&\le 4n + \frac{1}{100}\sqrt{n}E[Z] && (9)
\end{aligned}$$

(5) uses $|\mathcal{A}| \le n$, (6) is the AM-GM inequality, the (7) uses that $q_i \ge \frac{\varepsilon}{50n}$ for all $i \in \mathcal{A}$, (8) uses (1), and (9)substitutes a value $m \ge 20000\frac{\sqrt{n}}{\varepsilon^2}$.

The second term can be similarly bounded:

$$\begin{aligned}
4m\sum_{i\in\mathcal{A}} \frac{p_i(p_i - q_i)^2}{q_i^2} &\le 4m\left(\sum_{i\in\mathcal{A}} \frac{p_i^2}{q_i^2}\right)^{1/2}\left(\sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^4}{q_i^2}\right)^{1/2} \\
&\le 4m\left(4n + \frac{1}{100}\sqrt{n}E[Z]\right)^{1/2}\left(\sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^4}{q_i^2}\right)^{1/2} \\
&\le 4m\left(2\sqrt{n} + \frac{1}{10}n^{1/4}E[Z]^{1/2}\right)\left(\sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^2}{q_i}\right) \\
&= \left(8\sqrt{n} + \frac{2}{5}n^{1/4}E[Z]^{1/2}\right)E[Z]
\end{aligned}$$

The first inequality is Cauchy-Schwarz, the second inequality uses (9), the third inequality uses the monotonicity of the ℓ_p norms, and the equality uses (1).

Combining the two terms, we get

$$\mathrm{Var}\left[Z\right] \leq 4n + 9\sqrt{n}\mathbb{E}\left[Z\right] + \frac{2}{5}n^{1/4}\mathbb{E}\left[Z\right]^{3/2}.$$

We now consider the two cases in the statement of our lemma.

- When $p \in \mathcal{C}$, we know from Lemma 1 that $\mathbb{E}\left[Z\right] \leq \frac{1}{500}m\varepsilon^2$. Combined with a choice of $m \geq 20000\frac{\sqrt{n}}{\varepsilon^2}$ and the above expression for the variance, this gives:

$$\mathrm{Var}\left[Z\right] \leq \frac{4}{20000^2}m^2\varepsilon^4 + \frac{9}{20000 \cdot 500}m^2\varepsilon^4 + \frac{\sqrt{10}}{12500000}m^2\varepsilon^4 \leq \frac{1}{500000}m^2\varepsilon^4.$$

- When $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, Lemma 1 and $m \geq 20000\frac{\sqrt{n}}{\varepsilon^2}$ give:

$$\mathbb{E}\left[Z\right] \geq \frac{1}{5}m\varepsilon^2 \geq 4000\sqrt{n}.$$

  Combining this with our expression for variance we get:

$$\mathrm{Var}\left[Z\right] \leq \frac{4}{4000^2}\mathbb{E}\left[Z\right]^2 + \frac{9}{4000}\mathbb{E}\left[Z\right]^2 + \frac{2}{5\sqrt{4000}}\mathbb{E}\left[Z\right]^2 \leq \frac{1}{100}\mathbb{E}\left[Z\right]^2.$$

□

C Details on Testing Monotonicity

In this section, we prove Lemma 3 necessary for our monotonicity testing result.

Our analysis starts with a structural lemma about monotone distributions. In [13], Birgé showed that any monotone distribution p over $[n]$ can be *obliviously* decomposed into $O(\log(n)/\varepsilon)$ intervals, such that the flattening $\bar{p}$ (recall Definition 3) of p over these intervals is ε-close to p in total variation distance. [32] extend this result, giving a bound between the χ^2-distance of p and $\bar{p}$. We strengthen these results by extending them to monotone distributions over $[n]^d$. In particular, we partition the domain $[n]^d$ of p into $O((d\log(n)/\varepsilon^2)^d)$ rectangles, and compare it with $\bar{p}$, the flattening over these rectangles. The following result is proved in Section C.1.

Lemma 5. *Let $d \geq 1$. There is an oblivious decomposition of $[n]^d$ into $O((d\log(n)/\varepsilon^2)^d)$ rectangles such that for any monotone distribution p over $[n]^d$, its flattening $\bar{p}$ over these rectangles satisfy $\chi^2(p, \bar{p}) \leq \varepsilon^2$.*

This effectively reduces the support size to logarithmic in n. At this point, we can apply the Laplace estimator (along the lines of [33]) and learn a q such that if p was monotone, then q will be $O(\varepsilon^2)$-close in χ^2-distance. The following result is proved in Section C.2.

Lemma 6. *Let $d \geq 1$, and p be a monotone distribution over $[n]^d$. There is an algorithm which outputs a distribution q such that $\mathbb{E}\left[\chi^2(p, q)\right] \leq \frac{\varepsilon^2}{500}$. The time and sample complexity are both $O((d\log(n)/\varepsilon^2)^d/\varepsilon^2)$.*

Applying Markov's inequality gives the χ^2 distance guarantee in Lemma 3.

The final step before we apply our χ^2-tester is to compute the distance between q and $\mathcal{M}_n^d$. This subroutine is similar to the one introduced by [6]. The key idea is to write a linear program, which searches for any distribution f which is close to q in total variation distance. We note that the desired properties of f (i.e., monotonicity, normalization, and ε-closeness to q) are easy to enforce as linear constraints. Note that the linear program operates over the oblivious decomposition used in our structural result, so the complexity is polynomial in $(d\log(n)/\varepsilon)^d$, rather than the naive n^d.

These results when combined, give precisely the guarantees of Lemma 3.

C.1 A Structural Result for Monotone Distributions on the Hypergrid

Birgé [13] showed that any monotone distribution is estimated to a total variation ε with a $O(\log(n)/\varepsilon)$-piecewise constant distribution. Moreover, the intervals over which the output is constant is independent of the distribution p. This result, was strengthened to the Kullback-Leibler divergence by [32] to study the compression of monotone distributions. They upper bound the KL divergence by χ^2 distance and then bound the χ^2 distance. We extend this result to $[n]^d$. We divide $[n]^d$ into b^d rectangles as follows. Let $\{I_1, \ldots, I_b\}$ be a partition of $[n]$ into consecutive intervals defined as:

$$|I_j| = \begin{cases} 1 & \text{for } 1 \leq j \leq \frac{b}{2}, \\ \lfloor 2(1+\gamma)^{j-b/2} \rfloor & \text{for } \frac{b}{2} < j \leq b. \end{cases}$$

For $\mathbf{j} = (j_1, \ldots, j_d) \in [b]^d$, let $I_{\mathbf{j}} \stackrel{\text{def}}{=} I_{j_1} \times I_{j_2} \times \ldots \times I_{j_d}$.

The χ^2 distance between p and $\bar{p}$ can be bounded as

$$\begin{aligned}\chi^2(p, \bar{p}) &= \left[\sum_{\mathbf{j} \in [b]^d} \sum_{\mathbf{i} \in I_{\mathbf{j}}} \frac{p_{\mathbf{i}}^2}{\bar{p}_{\mathbf{i}}}\right] - 1 \\ &\leq \left[\sum_{\mathbf{j} \in [b]^d} p_{\mathbf{j}}^+ |I_{\mathbf{j}}|\right] - 1\end{aligned}$$

For $\mathbf{j} = (j_1, \ldots, j_d) \in \mathcal{S}_{\text{large}}$, let $\mathbf{j}^* = (j_1^*, \ldots, j_b^*)$ be

$$j_i^* = \begin{cases} j_i & \text{if } j_i \leq b/2 + 1 \\ j_i - 1 & \text{otherwise.} \end{cases}$$

We bound the expression above as follows.

Let $T \subseteq [d]$ be any subset of d. Suppose the size of T is ℓ. Let $\bar{T}$ be the set of all $\mathbf{j}$ that satisfy $\mathbf{j}_i = b/2 + 1$ for $i \in T$. In other words, over the dimensions determined by T, the value of the index is equal to $d/2 + 1$. The map $\mathbf{j} \to \mathbf{j}^*$ restricted to T is one-to-one, and since at most $d - \ell$ of the coordinates drop,

$$|I_{\mathbf{j}}| \leq |I_{\mathbf{j}^*}| \cdot (1 + \gamma)^{d-\ell}.$$

Since there are ℓ coordinates that do not change, and each of them have $2(1 + \gamma)$ coordinates, we obtain

$$\begin{aligned}\sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}} &\leq \sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}^*}^- \cdot |I_{\mathbf{j}}| \cdot (2(1 + \gamma))^{\ell} \cdot (1 + \gamma)^{d-\ell} \\ &= \sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}^*}^- \cdot |I_{\mathbf{j}^*}| \cdot 2^{\ell}(1 + \gamma)^d.\end{aligned}$$

Since the mapping is one-to-one, the probability of observing as element in $\bar{T}$ is the probability of observing $b/2 + 1$ in ℓ coordinates, which is at most $(2/(b + 2))^{\ell}$ under any monotone distribution. Therefore,

$$\sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}} \leq \left(\frac{2}{b + 2}\right)^{\ell} \cdot 2^{\ell}(1 + \gamma)^d.$$

For any ℓ there are $\binom{d}{\ell}$ choices for T. Therefore,

$$\begin{aligned}\chi^2(p, \bar{p}) &\leq \sum_{\ell=0}^{d} \binom{d}{\ell} \left(\frac{4}{b + 2}\right)^{\ell} (1 + \gamma)^d - 1 \\ &= (1 + \gamma)^d \left(1 + \frac{4}{b + 2}\right)^d - 1 \\ &= \left(1 + \gamma + \frac{4}{b + 2} + \frac{4\gamma}{b + 2}\right)^d - 1\end{aligned}$$

Recall that $\gamma = 2\log(n)/b > 1/b$, implies that the expression above is at most $(1 + 2\gamma)^d - 1$. This implies Lemma 5.

C.2 Monotone Learning

Our algorithm requires a distribution q satisfying the properties discussed earlier. We learn a monotone distribution from samples as follows.

Before proving this result, we prove a general result for χ^2 learning of arbitrary discrete distributions, adapting the result from [33]. For a distribution p, and a partition of the domain into b intervals $I_1, \ldots, I_b$, let $\bar{p}_i = p(I_i)/|I_i|$ be the flattening of p over these intervals. We saw that for monotone distributions there exists a partition of the domain such that $\bar{p}$ is *close* to the underlying distribution in χ^2 distance.

Suppose we are given m samples from a distribution p and a partition $I_1, \ldots, I_b$. Let m_j be the number of samples that fall in I_j. For $i \in I_j$, let

$$q_i \stackrel{\text{def}}{=} \frac{1}{|I_j|} \frac{m_j + 1}{m + b}.$$

Let $S_j = \sum_{i \in I_j} p_i^2$. The expected χ^2 distance between p and q can be bounded as follows.

$$\begin{aligned}
\mathbb{E}\left[\chi^2(p,q)\right] &= \left[\sum_{j=1}^{b}\sum_{i\in I_j}\sum_{\ell=0}^{m}\binom{m}{\ell}(p(I_j))^{\ell}(1-p(I_j))^{m-\ell}\frac{p_i^2}{(\ell+1)/(|I_j|(m+b))}\right] - 1 \\
&= \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\left(\sum_{\ell=0}^{m}\binom{m+1}{\ell+1}(p(I_j))^{\ell+1}(1-p(I_j))^{m+1-\ell+1}\right)\right] - 1 \\
&= \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\left(1-(1-p(I_j)^{m+1}\right)\right] - 1 \\
&\leq \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\right] - 1 \\
&= \left[\frac{m+b}{m+1}\left(\chi^2(p,\bar{p})+1\right)\right] - 1 \\
&= \frac{m+b}{m+1}\cdot\chi^2(p,\bar{p}) + \frac{b}{m+1}. \qquad (10)
\end{aligned}$$

Suppose $\gamma = O(\log(n)/b)$, and $b = O(d\cdot\log(n)/\varepsilon^2)$. Then, by Lemma 5,

$$\chi^2(p,\bar{p}) \leq \varepsilon^2. \qquad (11)$$

Combining this with (10) gives Lemma 3.

D Details on testing Unimodality

One striking feature of Birgé's result is that the decomposition of the domain is oblivious to the samples, and therefore to the unknown distribution. However, such an oblivious decomposition will not work for the unimodal distribution, since the mode is unknown. Suppose we know where the mode of the unknown distribution might be, then the problem can be decomposed into monotone functions over two intervals. Therefore, in theory, one can modify the monotonicity testing algorithm by iterating over all the possible n modes. Indeed, by applying a union bound, it then follows that

Theorem 6. *(Follows from Monotone) For $\varepsilon > 1/n^{1/4}$, there exists an algorithm for testing unimodality over $[n]$ with sample complexity $O\left(\sqrt{n}\log n/\varepsilon^2\right)$.*

Our main result for testing unimodality is the following theorem.

Theorem 7. *Suppose $\varepsilon > n^{-1/4}$. Then there exists an algorithm for testing unimodality over $[n]$ with sample complexity $O(\sqrt{n}/\varepsilon^2)$.*

Recall that to circumvent Birgé's decomposition, we want to decompose the interval into disjoint intervals such that the probability of each interval is about $O(1/b)$, where b is a parameter, specified later. In particular we consider a decomposition of $[n]$ with the following properties:

1. For each element i with probability at least $1/b$, there is an $I_\ell = \{i\}$.
2. There are at most two intervals with $p(I) \leq 1/2b$.
3. Every other interval I satisfies $p(I) \in \left[\frac{1}{2b}, \frac{2}{b}\right]$.

Let $I_1, \ldots, I_L$ denote the partition of $[n]$ corresponding to these intervals. Note that $L = O(b)$.

Claim 1. *There is an algorithm that takes $O(b\log b)$ samples and outputs $I_1, \ldots, I_L$ satisfying the properties above.*

The first step in our algorithm is to estimate the *total probability* within each of these intervals. In particular,

Lemma 7. *There is an algorithm that takes $m' = O(b\log b/\varepsilon^2)$ samples from a distribution p, and with probability at least 9/10 outputs a distribution $\bar{q}$ that is constant on each I_L. Moreover, for any j such that $p(I_j) > 1/2b$, $\bar{q}(I_j) \in (1\pm\varepsilon)p(I_j)$.*

Proof. Consider any interval I_j with $p(I_j) \geq 1/2b$. The number of samples N_{I_j} that fall in that interval is distributed $Binomial(m', p(I_j)$. Then by Chernoff bounds for $m' > 12b \log b/\varepsilon^2$,

$$\Pr\left(|N_{I_j} - m'p(I_j)| > \varepsilon m'p(I_j)\right) \leq 2\exp\left(\varepsilon^2 m'p(I_j)/2\right) \tag{12}$$

$$\leq \frac{1}{b^2}, \tag{13}$$

where the last inequality uses the fact that $p(I_j) \geq 1/2b$. □

The next step is estimate the distance of q from $\mathcal{U}_n$. This is possible by a simple dynamic program, similar to the one used for monotonicity. If the estimated distance is more than $\varepsilon/2$, we output REJECT.

Our next step is to remove certain intervals. This will be to ensure that when the underlying distribution is unimodal, we are able to estimate the distribution *multiplicatively* over the remaining intervals. In particular, we do the following preprocessing step:

- $A = \emptyset$.
- For interval I_j,
  - If

$$q(I_j) \notin ((1-\varepsilon)\cdot q(I_{j+1}), (1+\varepsilon)\cdot q(I_{j+1})) \quad \text{OR} \tag{14}$$

$$q(I_j) \notin ((1-\varepsilon)\cdot q(I_{j-1}), (1+\varepsilon)\cdot q(I_{j-1}))\,, \tag{15}$$

    add I_j to A.
- Add the (at most 2) intervals with mass at most $1/2b$ to A.
- Add all intervals j with $q(I_j)/|I_j| < \varepsilon/50n$ to A

If the distribution is unimodal, we can prove the following about the set of intervals A^c.

Lemma 8. *If p is unimodal then,*

- $p(I_{A^c}) \geq 1 - \varepsilon/25 - 1/b - O\left(\log n/(\varepsilon b)\right)$.
- *Except* at most one *interval in A^c every other interval I_j satisfies,*

$$\frac{p_j^+}{p_j^-} \leq (1+\varepsilon).$$

If this holds, then the χ^2 distance between p and q constrained to A^c, is at most ε^2. This lemma follows from the following result.

Lemma 9. *Let $C > 2$. For a unimodal distribution over $[n]$, there are at most $\frac{4\log(50n/\varepsilon)}{C\varepsilon}$ intervals I_j that satisfy $\frac{p_j^+}{p_j^-} < (1+\varepsilon/C)$.*

Proof. To the contrary, if there are more than $\frac{4\log(50n/\varepsilon)}{C\varepsilon}$ intervals, then at least half of them are on one side of the mode, however this implies that the ratio of the largest probability and smallest probability is at least $(1+\varepsilon/C)^j$, and if $j > \frac{2\log(50n/\varepsilon)}{C\varepsilon}$, is at least $50n/\varepsilon$, contradicting that we have removed all such elements. □

We have one additional pre-processing step here. We compute $q(A^c)$ and if it is smaller than $1 - \varepsilon/25$, we output REJECT.

Suppose there are L' intervals in A^c. Then, except at most one interval in L' we know that the χ^2 distance between p and q is at most ε^2 when p is unimodal, and the TV distance between p and q is at least $\varepsilon/2$ over A^c. We propose the following simple modification to take into account, the one interval that might introduce a high χ^2 distance in spite of having a small total variation. If we knew the interval, we can simply remove it and proceed. Since we do not know where the interval lies, we do the following.

1. Let Z_j be the χ^2 statistic over the ith interval in A^c, computed with $O(\sqrt{n}/\varepsilon^2)$ samples.
2. Let Z_l be the largest among all Z_j's.
3. If $\sum_{j,j\neq l} Z_j > m\varepsilon^2/10$, output REJECT.
4. Output ACCEPT.

The objective of removing the largest χ^2 statistic is our substitute for not knowing the largest interval. We now prove the correctness of this algorithm.

Case 1 $p \in UM_n$**:** We only concentrate on the final step. The χ^2 statistic over all but one interval are at most $c \cdot m\varepsilon^2$, and the variance is bounded as before. Since we remove the largest statistic, the expected value of the new statistic is *strictly dominated* by that of these intervals. Therefore, the algorithm outputs ACCEPT with at least the same probability as if we removed the spurious interval.

Case 2 $p \notin UM_n$**:** This is the hard case to prove for unimodal distributions. We know that the χ^2 statistic is large in this case, and we therefore have to prove that it remains large even after removing the largest test statistic Z_l.

We invoke Kolmogorov's Maximal Inequality to this end.

Lemma 10 (Kolmogorov's Maximal Inequality)**.** *For independent zero mean random variables $X_1, \ldots, X_L$ with finite variance, let $S_\ell = X_1 + \ldots X_\ell$. Then for any $\lambda > 0$,*

$$\Pr\left(\max_{1 \le \ell \le L} |S_\ell| \ge \lambda\right) \le \frac{1}{\lambda^2} \cdot Var\left(S_L\right). \tag{16}$$

As a corollary, it follows that $\Pr\left(\max_\ell |X_\ell| > 2\lambda\right) \le \frac{1}{\lambda^2} \cdot Var\left(S_L\right)$.

In the case we are interested in, we let $X_i = Z_\ell - \mathbb{E}\left[Z_\ell\right]$. Then, similar to the computations before, and the fact that each interval has a small mass, it follows that that the variance of the summation is at most $\mathbb{E}\left[Z_\ell\right]^2/100$. Taking $\lambda = \mathbb{E}\left[S_L - m\varepsilon^2/3\right]^2/100$, it follows that the statistic does not fall below to $\sqrt{n}$. This completes the proof of Theorem 7.

E Learning product distributions in χ^2 distance

In this section we prove Lemma 4, thus proving Theorem 3. The proof is analogous to the proof for learning monotone distributions, and hinges on the following result of [33].

Given m samples from a distribution q over n elements, the add-1 estimator (Laplace estimator) q satisfies:

$$\mathbb{E}\left[\chi^2(p, q)\right] \le \frac{n}{m+1}.$$

To handle the χ^2 distribution of product distributions, we first bound the χ^2-distance between product distributions in terms of the individual coordinates.

Lemma 11. *Let $p = p^1 \times p^2 \ldots \times p^d$, and $q = q^1 \times q^2 \ldots \times q^d$ be two distributions in Π_d. Then*

$$\chi^2(p, q) = \prod_{\ell=1}^{d} (1 + \chi^2(p^\ell, q^\ell)) - 1.$$

Proof. By the definition of χ^2-distance and exchanging the product and summation,

$$\chi^2(p, q) = \sum_{\mathbf{i} \in \mathcal{X}} \frac{(p_\mathbf{i} - q_\mathbf{i})^2}{q_\mathbf{i}} = \sum_{\mathbf{i} \in \mathcal{X}} \frac{p_\mathbf{i}^2}{q_\mathbf{i}} - 1 = \prod_{\ell=1}^{d} \left[\sum_{i \in [n_\ell]} \frac{\left(p_i^\ell\right)^2}{q_i^\ell}\right] - 1 = \prod_{\ell=1}^{d} \left(1 + \chi^2(p^\ell, q^\ell)\right) - 1. \qquad \square$$

Now, suppose p is a product distribution over $\mathcal{X} = [n_1] \times \cdots \times [n_d]$. We simply perform the add-1 estimation over each coordinate independently, giving a distribution $q^1 \times \cdots \times q^d$. Since p is a product distribution the estimates in each coordinate is independent. Therefore, a simple application of the previous result and independence of the coordinates implies

$$\begin{aligned}
\mathbb{E}\left[\chi^2(p, q)\right] &= \prod_{l=1}^{d} \left(1 + \mathbb{E}\left[\chi^2(p^l, q^l)\right]\right) - 1 \\
&\le \prod_{l=1}^{d} \left(1 + \frac{n_l}{m+1}\right) - 1 \\
&\le \exp\left(\frac{\sum_l n_l}{m+1}\right) - 1, \qquad (17)
\end{aligned}$$

where (17) follows from $e^x \ge 1 + x$. Using $e^x \le 1 + 2x$ for $0 \le x \le 1$, we have

$$\mathbb{E}\left[\chi^2(p, q)\right] \le 2\frac{\sum_l n_l}{m+1}, \tag{18}$$

when $m \ge \sum_l n_l$. Therefore, following an application of Markov's inequality, when $m = \Omega((\sum_l n_l)/\varepsilon^2)$, Lemma 4 is proved.

F Details on Testing Log-Concavity

Our main result for testing log-concavity is as follows:

Theorem 8. *There exists an algorithm for testing log-concavity over $[n]$ with sample complexity $O\left(\sqrt{n}/\varepsilon^2 + 1/\varepsilon^5\right)$ and time complexity $\mathrm{poly}(n, 1/\varepsilon)$.*

In particular, this implies the following optimal tester for this class:

Corollary 3. *Suppose $\varepsilon > 1/n^{1/5}$. Then there exists an algorithm for testing log-concavity over $[n]$ with sample complexity $O\left(\sqrt{n}/\varepsilon^2\right)$.*

Our algorithm will fit into the structure of our general framework. We first perform a very particular type of learning algorithm, whose guarantees are summarized in the following lemma:

Lemma 12. *Given $\varepsilon > 0$ and sample access to p, there exists an algorithm such that:*

- If $p \in \mathcal{LCD}_n$, the algorithm outputs a distribution $q \in \mathcal{LCD}_n$ and an $O(\varepsilon)$-effective support S of p such that $\chi^2(p_S, q_S) \leq \varepsilon^2/500$ with probability at least $5/6$;

- If $d_{\mathrm{TV}}(p, \mathcal{LCD}_n) \geq \varepsilon$, the algorithm either outputs a distribution $q \in \mathcal{LCD}_n$ or REJECT. *The sample complexity is $O(1/\varepsilon^5)$ and the time complexity is $\mathrm{poly}(n, 1/\varepsilon)$.*

We note that as a corollary, one immediately obtains a $O(1/\varepsilon^5)$ proper learning algorithm for log-concave distributions. The result is immediate from the first item of Lemma 12 and Proposition 1. We can actually do a bit better – in the proof of Lemma 12, we partition $[n]$ into intervals of probability mass $\Theta(\varepsilon^{3/2})$. If one instead partitions into intervals of probability mass $\Theta(\varepsilon/\log(1/\varepsilon))$ and works directly with total variation distance instead of χ^2 distance, one can show that $\tilde{O}(1/\varepsilon^4)$ samples suffice.

Corollary 4. *Given $\varepsilon > 0$ and sample access to a distribution $p \in \mathcal{LCD}_n$, there exists an algorithm which outputs a distribution $q \in \mathcal{LCD}_n$ such that $d_{\mathrm{TV}}(p, q) \leq \varepsilon$. The sample complexity is $\tilde{O}(1/\varepsilon^4)$ and the time complexity is $\mathrm{poly}(n, 1/\varepsilon)$.*

Then, given the guarantees of Lemma 12, Theorem 8 follows from Theorem 1[4]. The details of these results are presented in Section F.

It will suffice to prove Lemma 12.

Proof of Lemma 12: We first draw samples from p and obtain a $O(1/\varepsilon^{3/2})$-piecewise constant distribution f by appropriately flattening the empirical distribution. The proof is now in two parts. In the first part, we show that if $p \in \mathcal{LCD}_n$ then f will be close to p in χ^2 distance over its effective support. The second part involves proper learning of p. We will use a linear program on f to find a distribution $q \in \mathcal{LCD}_n$. This distribution is such that if $p \in \mathcal{LCD}_n$, then $\chi^2(p, q)$ is small, and otherwise the algorithm will either output some $q \in \mathcal{LCD}_n$ (with no other relevant guarantees) or REJECT.

We first construct f. Let $\hat{p}$ be the empirical distribution obtained by sampling $O(1/\varepsilon^5)$ samples from p.

The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality gives a generic algorithm for learning any distribution with respect to the Kolmogorov distance.

Lemma 13. *(See [34]) Suppose we have n i.i.d. samples $X_1, \ldots X_n$ from a distribution with CDF F. Let $F_n(x) \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$ be the empirical CDF. Then $\Pr[d_{\mathrm{K}}(F, F_n) \geq \varepsilon] \leq 2e^{-2n\varepsilon^2}$. In particular, if $n = \Omega((1/\varepsilon^2) \cdot \log(1/\delta))$, then $\Pr[d_{\mathrm{K}}(F, F_n) \geq \varepsilon] \leq \delta$.*

This implies that with probability at least $5/6$, $d_{\mathrm{K}}(p, \hat{p}) \leq \varepsilon^{5/2}/10$. In particular, note that $|p_i - \hat{p}_i| \leq \varepsilon^{5/2}/10$. Condition on this event in the remainder of the proof.

Let a be the minimum i such that $p_i \geq \varepsilon^{3/2}/5$, and let b be the maximum i satisfying the same condition. Let $M = \{a, \ldots, b\}$ or $\emptyset$ if a and b are undefined. By the guarantee provided by the DKW inequality, $p_i \geq \varepsilon^{3/2}/10$ for all $i \in M$. Furthermore, $\hat{p}_i \in p_i \pm \varepsilon^{3/2}/10 \in (1 \pm \varepsilon) \cdot p_i$. For each $i \in M$, let $f_i = \hat{p}_i$. We note that $|M| = O(1/\varepsilon)$, so this contributes $O(1/\varepsilon)$ constant pieces to f.

We now divide the rest of the domain into t intervals, all but constantly many of measure $\Theta(\varepsilon^{3/2})$ (under p). This is done via the following iterative procedure. As a base case, set $r_0 = 0$. Define I_j as $[l_j, r_j]$, where $l_j = r_{j-1} + 1$ and r_j is the largest $j \in [n]$ such that $\hat{p}(I_j) \leq 9\varepsilon^{3/2}/10$. The exception is if I_j would intersect M – in this case, we "skip" M: set $r_j = a - 1$ and $l_{j+1} = b + 1$. If such a j exists, denote it by

[4]To be more precise, we require the modification of Theorem 8 which is described in Section 4, in order to handle the case where the χ^2-distance guarantees only hold for a known effective support.

j^*. We note that $p(I_j) \leq \hat{p}(I_j) + \varepsilon^{5/2}/10 \leq \varepsilon^{3/2}$. Furthermore, for all j except j^* and t, $r_j + 1 \notin M$, so $p(I_j) \geq 9\varepsilon^{3/2}/10 - \varepsilon^{3/2}/5 - \varepsilon^{5/2}/10 \geq 3\varepsilon^{3/2}/5$. Observe that this lower bound implies that $t \leq \frac{2}{\varepsilon^{3/2}}$ for ε sufficiently small.

Part 1. For this part of the algorithm, we only care about the guarantees when $p \in \mathcal{LCD}_n$, so we assume this is the case.

For the domain $[n] \setminus M$, we let f be the flattening of $\hat{p}$ over the intervals $I_1, \ldots I_t$. To analyze f, we need a structural property of log-concave distributions due to Chan, Diakonikolas, Servedio, and Sun [16]. This essentially states that a log-concave distribution cannot have a sudden increase in probability.

Lemma 14 (Lemma 4.1 in [16])**.** *Let p be a distribution over $[n]$ that is non-decreasing and log-concave on $[1, x] \subseteq [n]$. Let $I = [x, y]$ be an interval of mass $P(I) = \tau$, and suppose that the interval $J = [1, x-1]$ has mass $p(J) = \sigma > 0$. Then*

$$p(y)/p(x) \leq 1 + \tau/\sigma.$$

Recall that any log-concave distribution is unimodal, and suppose the mode of p is at i_0. We will first focus on the intervals $I_1, \ldots, I_{t_L}$ which lie entirely to the left of i_0 and M. We will refer to I_j as L_j for all $j \leq t_L$. Note that p is non-decreasing over these intervals.

The next steps to the analysis are as follows. First we show that the flattening of p over L_j is a multiplicative $(1 + O(1/j))$ estimate for each $p_i \in L_j$. Then, we show that flattening the empirical distribution $\hat{p}$ over L_j is a multiplicative $(1 + O(1/j))$ estimate of $p(i)$ for each $i \in L_j$. Finally, we exclude a small number of intervals (those corresponding to $O(\varepsilon)$ mass at the left and right side of the domain, as well as j^*) in order to get the χ^2 approximation we desire on an effective support.

- First, recall that $p(L_j) \leq \varepsilon^{3/2}$ for all j. Also, letting $J_j = [1, r_{j-1}]$, we have that $p(J_j) \geq (j-1) \cdot 3\varepsilon^{3/2}/5$. Thus by Lemma 14, $p(r_j) \leq p(l_j)(1 + 2/(j-1))$. Since the distribution is non-decreasing in L_j, the flattening $\bar{p}$ of p is such that $\bar{p}(i) \in p(i)(1 \pm \frac{2}{j-1})$ for all $i \in L_j$.
- We have that $p(L_j) \geq 3\varepsilon^{3/2}/5$, and $\hat{p}(L_j) \in p(L_j) \pm \varepsilon^{5/2}/10$, so $\hat{p}(L_j) \in p(L_j) \cdot (1 \pm \frac{\varepsilon}{6})$, and hence $\hat{p}(i) \in \bar{p}(i) \cdot (1 \pm \frac{\varepsilon}{6})$ for all $i \in L_j$. Combining with the previous point, we have that

$$\hat{p}(i) \in p(i) \cdot \left(1 \pm \left(\frac{2\varepsilon}{3(j-1)} + \frac{\varepsilon}{6} + \frac{2}{j-1}\right)\right) \in p(i) \cdot \left(1 \pm \frac{11}{3(j-1)}\right).$$

A symmetric statement holds for the intervals that lie entirely to the right of i_0 and M. We will refer to I_j as R_{t-j} for all $j > t_L$.

To summarize, we have the following guarantees for the distribution f:

- For all $i \in M$, $f(i) \in p(i) \cdot (1 \pm \varepsilon)$;
- For all $i \in L_j$ (except L_1 and L_{j^*}), $f(i) \in p(i) \cdot \left(1 \pm \frac{22}{3j}\right)$;
- For all $i \in R_j$ (except R_1), $f(i) \in p(i) \cdot \left(1 \pm \frac{22}{3j}\right)$;

Note that, in particular, we have multiplicative estimates for all intervals, except those in L_1, L_{j^*}, R_1 and the interval containing i_0. Let S be the set of all intervals except L_{j^*}, L_j and R_j for $j \leq 1/\sqrt{\varepsilon}$, and the one containing i_0 Then, since each interval has probability mass at most $O(\varepsilon^{3/2})$ and we are excluding $O(1/\sqrt{\varepsilon})$ intervals, $p(S) > 1 - O(\varepsilon)$.

We now compute the χ^2-distance induced by this approximation for elements in S. For an element $i \in L_j \cap S$, we have

$$\frac{(f(i) - p(i))^2}{p(i)} \leq \frac{60p(i)}{j^2}.$$

Summing over all $i \in L_j \cap S$ gives

$$\frac{60\varepsilon^{3/2}}{j^2}$$

since the probability mass of L_j is at most $\varepsilon^{3/2}$. Summing this over all L_j for $j \geq 1/\sqrt{\varepsilon}$ and $j \neq j^*$ gives

$$\begin{aligned} 60\varepsilon^{3/2} \sum_{j=1/\sqrt{\varepsilon}}^{2/\varepsilon^{3/2}} \frac{1}{j^2} &\leq 60\varepsilon^{3/2} \int_{1/\sqrt{\varepsilon}}^{\infty} \frac{1}{x^2} dx \\ &= 60\varepsilon^{3/2}(\sqrt{\varepsilon}) \\ &= O(\varepsilon^2) \end{aligned}$$

as desired.

Part 2. To obtain a distribution $q \in \mathcal{LCD}_n$, we write a linear program. We will work in the log domain, so our variables will be Q_i, representing $\log q(i)$ for $i \in [n]$. We will use $F_i = \log f(i)$ as parameters in our LP. There will be no objective function, we simply search for a feasible point. Our constraints will be

$$Q_{i-1} + Q_{i+1} \leq 2Q_i \ \ \forall i \in [n-1]$$
$$Q_i \leq 0 \ \ \forall i \in [n]$$
$$\log(1+\varepsilon) \leq |Q_i - F_i| \leq \log(1+\varepsilon) \text{ for } i \in M$$
$$\log\left(1 - \frac{22}{3j}\right) \leq |Q_i - F_i| \leq \log\left(1 + \frac{22}{3j}\right) \text{ for } i \in L_j, j \geq 1/\sqrt{\varepsilon} \text{ and } j \neq j^*$$
$$\log\left(1 - \frac{22}{3j}\right) \leq |Q_i - F_i| \leq \log\left(1 + \frac{22}{3j}\right) \text{ for } i \in R_j, j \geq 1/\sqrt{\varepsilon}$$

If we run the linear program, then after a rescaling and summing the error over all the intervals in the LP gives us that the distance between p and q to be $O(\varepsilon^2)$ χ^2-distance in a set S which has measure $p(S) \geq 1 - 4\varepsilon$, as desired.

If the linear program finds a feasible point, then we obtain a $q \in \mathcal{LCD}_n$. Furthermore, if $p \in \mathcal{LCD}_n$, this also tells us that (after a rescaling of ε), summing the error over all intervals implies that $\chi^2(p_S, q_S) \leq \frac{\varepsilon^2}{500}$ for a known set S with $p(S) \geq 1 - O(\varepsilon)$, as desired. If $M \neq \emptyset$, this algorithm works as described. The issue is if $M = \emptyset$, then we don't know when the L intervals end and the R intervals begin. In this case, we run $O(1/\varepsilon)$ LPs, using each interval as the one containing i_0, and thus acting as the barrier between the L intervals (to its left) and the R intervals (to its right). If p truly was log-concave, then one of these guesses will be correct and the corresponding LP will find a feasible point. □

G Details on MHR testing

In this section, we give our main result for testing for monotone hazard rate:

Theorem 9. *There exists an algorithm for testing monotone hazard rate over $[n]$ with sample complexity $O\left(\sqrt{n}/\varepsilon^2 + \log(n/\varepsilon)/\varepsilon^4\right)$ and time complexity $\mathrm{poly}(n, 1/\varepsilon)$.*

This implies the following optimal tester for the class:

Corollary 5. *Suppose $\varepsilon > \sqrt{\log(n/\varepsilon)}/n^{1/4}$. Then there exists an algorithm for testing monotone hazard rate over $[n]$ with sample complexity $O\left(\sqrt{n}/\varepsilon^2\right)$.*

We obey the same framework as before, first applying a χ^2-learner with the following guarantees:

Lemma 15. *Given $\varepsilon > 0$ and sample access to p, there exists an algorithm such that:*

- If $p \in \mathcal{MHR}_n$, the algorithm outputs a distribution $q \in \mathcal{MHR}_n$ and an $O(\varepsilon)$-effective support S of p such that $\chi^2(p_S, q_S) \leq \varepsilon^2/500$ with probability at least $5/6$;

- If $d_{\mathrm{TV}}(p, \mathcal{MHR}_n) \geq \varepsilon$, the algorithm either outputs a distribution $q \in \mathcal{MHR}_n$ and a set $S \subseteq [n]$ or REJECT. *The sample complexity is $O(\log(n/\varepsilon)/\varepsilon^4)$ and the time complexity is $\mathrm{poly}(n, 1/\varepsilon)$.*

As with log-concave distributions, this implies the following proper learning result:

Corollary 6. *Given $\varepsilon > 0$ and sample access to a distribution $p \in \mathcal{MHR}_n$, there exists an algorithm which outputs a distribution $q \in \mathcal{MHR}_n$ such that $d_{\mathrm{TV}}(p, q) \leq \varepsilon$. The sample complexity is $O(\log(n/\varepsilon)/\varepsilon^4)$ and the time complexity is $\mathrm{poly}(n, 1/\varepsilon)$.*

Proof of Lemma 15: As with log-concave distributions, our method for MHR distributions can be split into two parts. In the first step, if $p \in \mathcal{MHR}_n$, we obtain a distribution q which is $O(\varepsilon^2)$-close to p in χ^2 distance on a set $\mathcal{A}$ of intervals such that $p(\mathcal{A}) \geq 1 - O(\varepsilon)$. q will achieve this by being a multiplicative $(1 + O(\varepsilon))$ approximation for each element within these intervals. This step is very similar to the decomposition used for unimodal distributions (described in Section D), so we sketch the argument and highlight the key differences.

The second step will be to find a feasible point in a linear program. If $p \in \mathcal{MHR}_n$, there should always be a feasible point, indicating that q is close to a distribution in $\mathcal{MHR}_n$ (leveraging the particular guarantees for our algorithm for generating q). If $d_{\mathrm{TV}}(p, \mathcal{MHR}_n) \geq \varepsilon$, there may or may not be a feasible point, but when there is, it should imply the existence of a distribution $p^* \in \mathcal{MHR}_n$ such that $d_{\mathrm{TV}}(q, p^*) \leq \varepsilon/2$.

The analysis will rely on the following lemma from [16], which roughly states that an MHR distribution is "almost" non-decreasing.

Lemma 16 (Lemma 5.1 in [16])**.** *Let p be an MHR distribution over $[n]$. Let $I = [a, b] \subset [n]$ be an interval, and $R = [b+1, n]$ be the elements to the right of I. Let $\eta = p(I)/p(R)$. Then $p(b+1) \geq \frac{1}{1+\eta} p(a)$.*

Part 1. As before, with unimodal distributions, we start by taking $O(\frac{b \log b}{\varepsilon^2})$ samples, with the goal of partitioning the domain into intervals of mass approximately $\Theta(1/b)$. First, we will ignore the left and rightmost intervals of mass $\Theta(\varepsilon)$. For all "heavy" elements with mass $\geq \Theta(1/b)$, we consider them as singletons. We note that Lemma 16 implies that there will be at most $O(1/\varepsilon)$ contiguous intervals of such elements. The rest of the domain is greedily divided (from left to right) into intervals of mass $\Theta(1/b)$, cutting an interval short if we reach one of the heavy elements. This will result in the guarantee that all but potentially $O(1/\varepsilon)$ intervals have $\Theta(1/b)$ mass.

Next, similar to unimodal distributions, considering the flattened distribution, we discard all intervals for which the per-element probability is not within a $(1 \pm O(\varepsilon))$ multiplicative factor of the same value for both neighboring intervals. The claim is that all remaining intervals will have the property that the per-element probability is within a $(1 \pm O(\varepsilon))$ multiplicative factor of the true probability. This is implied by Lemma 16. If there were a point in an interval which was above this range, the distribution must decrease slowly, and the next interval would have a much larger per-element weight, thus leading to the removal of this interval. A similar argument forbids us from missing an interval which contains a point that lies outside this range. Relying on the fact that truncating the left and rightmost intervals eliminates elements with low probability mass, similar to the unimodal case, one can show that we will remove at most $\log(n/\varepsilon)/\varepsilon$ intervals, and thus a $\log(n/\varepsilon)/b\varepsilon$ probability mass. Choosing $b = \Omega(\varepsilon^2/\log(n/\varepsilon))$ limits this to be $O(\varepsilon)$, as desired. At this point, if p is indeed MHR, the multiplicative estimates guarantee that the result is $O(\varepsilon^2)$-close in χ^2-distance among the remaining intervals.

Part 2. We note that an equivalent condition for distribution f being MHR is log-concavity of $\log(1-F)$, where F is the CDF of f. Therefore, our approach for this part will be similar to the approach used for log-concave distributions.

Given the output distribution q from the previous part of this algorithm, our goal will be check if there exists an MHR distribution f which is $O(\varepsilon)$-close to q. We will run a linear program with variables $\mathfrak{f}_i = \log(1-F_i)$. First, we ensure that f is a distribution. This can be done with the following constraints:

$$\begin{aligned} \mathfrak{f}_i &\leq 0 \qquad \forall i \in [n] \\ \mathfrak{f}_i &\geq \mathfrak{f}_{i+1} \qquad \forall i \in [n-1] \\ \mathfrak{f}_n &= -\infty \end{aligned}$$

To ensure that f is MHR, we use the following constraint:

$$\mathfrak{f}_{i-1} + \mathfrak{f}_{i-1} \leq 2\mathfrak{f}_i \qquad \forall i \in [2, n-1]$$

Now, ideally, we would like to ensure f and q are ε-close in total variation distance by ensuring they are pointwise within a multiplicative $(1 \pm \varepsilon)$ factor of each other:

$$(1-\varepsilon) \leq f_i/q_i \leq (1+\varepsilon)$$

We note that this is a stronger condition than f and q being ε-close, but if $p \in \mathcal{MHR}_n$, the guarantees of the previous step would imply the existence of such an f.

We have a separate treatment for the identified singletons (i.e., those with probability $\geq 1/b$) and the remainder of the support. For each element q_i identified to have $\geq 1/b$ mass, we add two constraints:

$$\log((1-b\varepsilon/2)(1-Q_i)) \leq \mathfrak{f}_i \leq \log((1+b\varepsilon/2)(1-Q_i))$$

$$\log((1-b\varepsilon/2)(1-Q_{i-1})) \leq \mathfrak{f}_{i-1} \leq \log((1+b\varepsilon/2)(1-Q_{i-1}))$$

If we satisfy these constraints, it implies that

$$q_i - b\varepsilon \leq f_i \leq q_i + b\varepsilon.$$

Since $q_i \geq 1/b$, this implies

$$(1-\varepsilon)q_i \leq f_i \leq (1+\varepsilon)q_i$$

as desired.

Now, the remaining elements each have $\leq 1/b$ mass. For each such element q_i, we create a constraint

$$(1-O(\varepsilon))\frac{q_i}{1-Q_{i-1}} \leq \mathfrak{f}_{i-1} - \mathfrak{f}_i \leq (1+O(\varepsilon))\frac{q_i}{1-Q_{i-1}}$$

Note that the middle term is

$$-\log\left(\frac{1-F_i}{1-F_{i-1}}\right) = -\log\left(1 - \frac{f_i}{1-F_{i-1}}\right) \in \frac{f_i}{1-F_{i-1}}\left(1 \pm 2\varepsilon\right),$$

where the second equality uses the Taylor expansion and the facts that $f_i \leq 1/b$ and $1 - F_{i-1} \geq \varepsilon$ (since during the previous part, we ignored the rightmost $O(\varepsilon)$ probability mass). If we satisfy the desired constraints, it implies that

$$f_i \in \frac{1}{(1 \pm 2\varepsilon)} \frac{1 - F_{i-1}}{1 - Q_{i-1}} (q + O(\varepsilon)) q_i.$$

Since we are taking $\Omega(1/\varepsilon^4)$ samples and $1 - F_{i-1} \geq \Omega(\varepsilon)$, Lemma 13 implies that f_i is indeed a multiplicative $(1 \pm \varepsilon)$ approximation for these points as well.

We note that all points which do not fall into these two cases make up a total of $O(\varepsilon)$ probability mass. Therefore, f may be arbitrary at these points and only incur $O(\varepsilon)$ cost in total variation distance.

If we find a feasible point for this linear program, it implies the existence of an MHR distribution within $O(\varepsilon)$ total variation distance. In this case, we continue to the testing portion of the algorithm. Furthermore, if $p \in \mathcal{MHR}_n$, our method for generating q certifies that such a distribution exists, and we continue on to the testing portion of the algorithm. □

H Details of the Lower Bounds

In this section, for the class of distributions $\mathcal{Q}$ described in discussion on lower bounds and a class of interest $\mathcal{C}$, we show that $d_{\mathrm{TV}}(\mathcal{C}, \mathcal{Q}) \geq \varepsilon$, thus implying a lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for testing $\mathcal{C}$.

H.1 Monotone distributions

We first consider $d = 1$ and prove that for appropriately chosen c, any monotone distribution over $[n]$ is ε-far from all distributions in $\mathcal{Q}$. Consider any $q \in \mathcal{Q}$. For this distribution, we say that $i \in [n]$ is a *raise-point* if $q_i < q_{i+1}$. Let R_q be the set of raise points of q. For $q \in \mathcal{Q}$, (2) implies at least one in every four consecutive integers in $[n]$ is a raise point, and therefore, $|R_q| \geq n/4$. Moreover, note that if i is a raise-point, then $i + 1$ is not a raise point. For any monotone (decreasing) distribution p, $p_i \geq p_{i+1}$. For any raise-point $i \in R_q$, by the triangle inequality,

$$|p_i - q_i| + |p_{i+1} - q_{i+1}| \geq |p_i - p_{i+1} + q_{i+1} - q_i| \geq q_{i+1} - q_i = \frac{2c\varepsilon}{n}. \tag{19}$$

Summing over the set R_q, we obtain $d_{\mathrm{TV}}(p, q) \geq \frac{1}{2}|R_q| \cdot \frac{2c\varepsilon}{n} \geq c\varepsilon/4$. Therefore, if $c \geq 4$, then $d_{\mathrm{TV}}(\mathcal{M}_n, q) \geq \varepsilon$. This proves the lower bound for $d = 1$.

This argument can be extended to $[n]^d$. Consider the following class of distributions on $[n]^d$. For each point $\mathbf{i} = (i_1, \ldots, i_d) \in [n]^d$, where i_1 is even, generate a random $z \in \{-1, 1\}$, and assign to $\mathbf{i}$ a probability of $(1 + zc\varepsilon)/n^d$. Let $\mathbf{e}_1 \stackrel{\text{def}}{=} (1, 0, \ldots, 0)$. Similar to $d = 1$, assign a probability $(1 - zc\varepsilon)/n^d$ to the point $\mathbf{i} + \mathbf{e}_1 = (i_1 + 1, i_2, \ldots, i_d)$. This class consists of $2^{\frac{n^d/2}{2}}$ distributions, and Paninski's arguments extend to give a lower bound of $\Omega(n^{d/2}/\varepsilon^2)$ samples to distinguish this class from the uniform distribution over $[n]^d$. It remains to show that all these distributions are ε far from $\mathcal{M}_n^d$. Call a point $\mathbf{i}$ as a raise point if $p_{\mathbf{i}} < p_{\mathbf{i}+\mathbf{e}_1}$. For any $\mathbf{i}$, one of the points $\mathbf{i}, \mathbf{i} + \mathbf{e}_1, \mathbf{i} + 2\mathbf{e}_1, \mathbf{i} + 3\mathbf{e}_1$ is a raise point, and the number of raise points is at least $n^d/4$. Invoking the triangle inequality (identical to (19)) over the raise-points, in the first dimension shows that any monotone distribution over $[n]^d$ is at a distance $\frac{c\varepsilon}{4}$ from any distribution in this class. Choosing $c = 4$ yields a bound of ε.

H.2 Testing Product Distributions

Our idea for testing independence is similar to the previous section. We sketch the construction of a class of distributions on $\mathcal{X} = [n_1] \times \cdots \times [n_d]$. Then $|\mathcal{X}| = n_1 \cdot n_2 \ldots \cdot n_d$. For each element in $\mathcal{X}$ assign a value $(1 \pm c\varepsilon)$ and then for each such assignment, normalize the values so that they add to 1, giving rise to a distribution. This gives us a class of $2^{|\mathcal{X}|}$ distributions. The key argument is to show that a *large* fraction of these distributions are far from being a product distribution. This follows since the degrees of freedom of a product distribution is exponentially smaller than the number of possible distributions. The second step is to simply apply Paninski's argument, now over the larger set of distributions, where we show that distinguishing the collection of distributions we constructed from the uniform distribution over $\mathcal{X}$ (which is a product distribution) requires $\sqrt{|\mathcal{X}|}/\varepsilon^2$ samples.

H.3 Log-concave and Unimodal distributions

We will show that any log-concave or unimodal distribution is ε-far from all distributions in $\mathcal{Q}$. Since $\mathcal{LCD}_n \subset \mathcal{U}_n$, it will suffice to show this for every unimodal distribution. Consider any unimodal distribution p, with

mode ℓ. Then, p is monotone non-decreasing over the interval $[\ell]$ and non-increasing over $\{\ell+1,\ldots,n\}$. By the argument for monotone distributions, the total variation distance between p and any distribution q over elements greater than ℓ is at least $\frac{n-\ell-1}{n}\frac{c\varepsilon}{4}$, and over elements less than ℓ is at least $\frac{\ell-1}{n}\frac{c\varepsilon}{4}$. Summing these two gives the desired bound.

H.4 Monotone Hazard distributions

We will show that any monotone hazard rate distribution is ε-far from all distributions in $\mathcal{Q}$.

Let p be any monotone-hazard distribution. Any distribution $q \in \mathcal{Q}$ has mass at least $1/2$ over the interval $I = [n/4, 3n/4]$. Therefore, by Lemma 16, for any $i \in I$, $p_{i+1}\left(1+\frac{p_i}{1/4}\right) \geq p_i$. As noted before, at least $n/8$ of the raise-points are in I.

For any $i \in I \cap R_q$, $q_i = (1+c\varepsilon)/n$, $q_{i+1} = (1-c\varepsilon)/n$

$$d_i = |p_i - q_i| + |p_{i+1} - q_{i+1}|. \tag{20}$$

If $p_i \geq (1+2c\varepsilon)/n$ or $p_i \leq 1/n$, then the first term, and therefore d_i is at least $c\varepsilon/n$. If $p_i \in (1/n, (1+2c\varepsilon)/n)$, then for $n > 5/(c\varepsilon)$

$$p_{i+1} \geq \frac{1}{n}\cdot\frac{1}{1+\frac{4}{n}} \geq \frac{1-c\varepsilon/2}{n}.$$

Therefore the second term of d_i is at $c\varepsilon/2n$. Since there are at least $n/8$ raise points in I,

$$d_{\mathrm{TV}}(p,q) \geq \frac{1}{2}\frac{n}{8}\cdot\frac{c\varepsilon}{2n} \geq \frac{c\varepsilon}{16}. \tag{21}$$

Thus any MHR distribution is ε-far from $\mathcal{Q}$ for $c \geq 16$.