Monotone Probability Distributions over the Boolean Cube Can Be Learned with Sublinear Samples

Ronitt Rubinfeld
CSAIL at MIT, Cambridge, MA, USA
Blavatnik School of Computer Science at Tel Aviv University, Israel
https://people.csail.mit.edu/ronitt/
ronitt@csail.mit.edu

Arsen Vasilyan
CSAIL at MIT, Cambridge, MA, USA
vasilyan@mit.edu

Abstract

A probability distribution over the Boolean cube is **monotone** if flipping the value of a coordinate from zero to one can only increase the probability of an element. Given samples of an unknown monotone distribution over the Boolean cube, we give (to our knowledge) the first algorithm that learns an approximation of the distribution in statistical distance using a number of samples that is sublinear in the domain.

To do this, we develop a structural lemma describing monotone probability distributions. The structural lemma has further implications to the sample complexity of basic testing tasks for analyzing monotone probability distributions over the Boolean cube: We use it to give nontrivial upper bounds on the tasks of estimating the distance of a monotone distribution to uniform and of estimating the support size of a monotone distribution. In the setting of monotone probability distributions over the Boolean cube, our algorithms are the first to have sample complexity lower than known lower bounds for the same testing tasks on arbitrary (not necessarily monotone) probability distributions.

One further consequence of our learning algorithm is an improved sample complexity for the task of testing whether a distribution on the Boolean cube is monotone.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms; Theory of computation

Keywords and phrases Learning distributions, monotone probability distributions, estimating support size

Digital Object Identifier 10.4230/LIPIcs.ITCS.2020.28

Funding *Ronitt Rubinfeld*: FinTech@CSAIL, MIT-IBM Watson AI Lab and Research Collaboration Agreement No. W1771646, and NSF Award Numbers: CCF-1650733, CCF-1740751, CCF-1733808, and IIS-1741137
Arsen Vasilyan: NSF grant IIS-1741137, EECS SuperUROP program

1 Introduction

1.1 Learning Monotone Distributions

Data generated from probability distributions is ubiquitous, and algorithms for understanding such data are of fundamental importance. In particular, a fundamental task is to *learn* an approximation to the probability distribution underlying the data. For probability distributions over huge discrete domains, the sample complexity and run-time bounds for the learning task can be prohibitive. In particular, learning an arbitrary probability distribution on a universe of N_{universe} elements up to sufficiently small constant total variation distance


© Ronitt Rubinfeld and Arsen Vasilyan;
licensed under Creative Commons License CC-BY

requires $\Omega(N_{\text{universe}})$ samples. However, when the probability distribution is known to belong to a more structured class of distributions, much better results are possible (cf. [16, 19, 22, 21, 20, 24, 13, 12, 6, 10, 25, 23, 15, 28]) – for example, learning an unknown Poisson binomial distribution up to variation distance ϵ can be achieved with only $\tilde{O}(1/\epsilon^3)$ samples and $\tilde{O}(\log(N_{\text{universe}})/\epsilon^3)$ run-time [13].

A fundamental class of probability distributions is the class of multidimensional monotone probability distributions, which broadly satisfy the following properties:

- The elements of the probability distribution have n different features.
- For every element, an increase in the value of one of the features can only increase its probability.

This basic class of distributions is of great interest because many commonly studied distributions are either monotone or can be approximated by a combination of monotone distributions. Furthermore, often the tools developed for monotone distributions are useful for other classes of distributions: for example, in the one dimensional setting, [12] use tools developed for testing monotone distributions in order to learn k-modal distributions. In [8], tools developed for testing properties of monotone distributions by [4] are used to develop testers for many other classes of distributions.

For the case of only one feature, or equivalently for monotone probability distributions over the totally ordered set $[k]$, a sample-efficient algorithm is known for learning the unknown distribution up total variation distance ϵ with $O(\log(k)/\epsilon^3)$ samples [6, 12]. In [1] it was also shown that an unknown probability distribution over $[k]^n$ can be learned up to χ^2 distance ϵ^2 with $O((n \log k/\epsilon^2)^n/\epsilon^2)$ samples (note that for constant ϵ, this sample complexity is non-trivial only when k is sufficiently large). Overall, the cases considered in the literature specialize on the regime when all the dimensions have a wide range that grows with n. Here we focus on a contrasting case, where each feature has only two possible values, 0 and 1, thus specializing on the Boolean cube:

▶ **Definition 1.** *A probability distribution ρ over $\{0,1\}^n$ is **monotone** if whenever for $x, y \in \{0,1\}^n$ we have that $x \preceq y$ (which means that for all i $x_i \leq y_i$), then we have that $\rho(x) \leq \rho(y)$.*

When studying multi-dimensional objects, focusing on the specific case of the Boolean cube is a common research theme, because the ideas and techniques developed for the Boolean cube are often applicable in the general case. A lower bound of $\Omega(2^{0.15n})$ for learning monotone probability distributions over the Boolean cube (up to sufficiently small constant variation distance) can be inferred from an entropy testing lower bound in [29, page 39] and an argument in [32] (see Claim 17 in Preliminaries). Though the dramatic exponential improvement as in [6, 12] for the totally ordered set is thereby impossible, this still leaves open the possibility of a sublinear sample algorithm for the Boolean cube.

We give, to the best of our knowledge, the first sublinear sample algorithm for learning a monotone probability distribution over the Boolean cube:

▶ **Theorem 2.** *For every positive ϵ, such that $0 < \epsilon \leq 1$ and for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(n^{1/5})}}$ samples from an unknown monotone probability distribution ρ over $\{0,1\}^n$, can reliably return a description of an estimate probability distribution $\hat{\rho}$, such that $d_{TV}(\rho, \hat{\rho}) \leq \epsilon$. The algorithm runs in time $O\left(2^{n+O_\epsilon(n^{1/5}\log n)}\right)$.*

Our algorithm relies on a new structural lemma describing monotone probability distributions on the Boolean cube, as described in Section 1.3. These structural insights also allow us to get improved sample complexity for certain testing tasks on monotone distributions – namely, estimating the closeness of a distribution to uniformity and the support size of the distribution, as presented in Section 1.2.

Theorem 2, together with the L_1 distance tester in [31], can be applied to give the best known sample complexity for testing whether a distribution is monotone. Specifically, one can test whether an unknown distribution ρ over the Boolean cube is monotone or ϵ-far from monotone with $O(\frac{2^n}{n\epsilon^2})$ samples as shown in Claim 18 in Preliminaries. Note that this does not follow from [32] directly, because monotonicity is not a symmetric property. The best previously known algorithm for testing monotonicity over the Boolean cube was presented in [5], requiring $\tilde{O}\left(\frac{2^n}{(n/\log n)^{1/4}}\text{poly}(1/\epsilon)\right)$ samples. The best sample complexity lower bound for testing monotonicity over $\{0,1\}^n$ is $\Omega(2^{(1-\Theta(\sqrt{\epsilon})+o(1))\cdot n})$, as presented in [3]. For the domain $[k]^n$, a monotonicity testing algorithm that requires $O\left(k^{n/2}/\epsilon^2 + \left(\frac{n\log k}{\epsilon^2}\right)^n \cdot \frac{1}{\epsilon^2}\right)$ samples is given and shown to be optimal in [1] (note that this is inapplicable to the Boolean setting, because this sample bound is non-trivial only for sufficiently large k).

1.2 Testing properties of monotone distributions

In addition to learning a distribution, several other basic tasks aimed at understanding distributions have received attention. These include estimating the entropy of a distribution, the size of the support and whether the distribution has certain "shape" properties (monotonicity, convexity, monotone hazard rate, etc.). For arbitrary probability distributions over huge domains, the sample complexity and run-time bounds for the above tasks can be prohibitive, provably requiring $\Omega\left(\frac{N_{\text{universe}}}{\log(N_{\text{universe}})}\right)$ samples. This is true in particular for the properties of support size, entropy and the distance to the uniform distribution [27, 32, 30, 33, 34].

This state of affairs motivates going beyond worst-case analysis and considering common classes of structured probability distributions, a direction that has been considered by many and with a large variety of results (cf. [4, 9, 24, 29, 14, 18]). Some specific examples include: In [4] it is shown that testing whether a monotone distribution is uniform requires only $\Theta(\log^3(N_{\text{universe}})/\epsilon^3)$ samples, in contrast to the $\Theta(\sqrt{N_{\text{universe}}}/\epsilon^2)$ samples required for testing arbitrary distributions for uniformity [26, 11, 17]. The situation is analogous for the tasks of testing whether two distributions given by samples are either the same or far, and testing whether a constant dimensional distribution is independent, which require only polylogarithmic samples if the unknown distributions are promised to be monotone on a total order [4].

Algorithms for testing properties of monotone probability distributions over the Boolean cube were studied in [29, 2]. It was shown that, given samples from a probability distribution over $\{0,1\}^n$ that is promised to be monotone, distinguishing the uniform distribution over $\{0,1\}^n$ from one that is ϵ-far from uniform can be done using only $O\left(\frac{n}{\epsilon^2}\right)$ samples, which is nearly optimal. In contrast, a number of other testing problems cannot have such dramatic improvements when the distribution is known to be monotone: for example in [29] it was shown that for sufficiently small constant ϵ the estimation of entropy up to an additive error of ϵn requires $2^{\Omega(n)}$ samples. However, no nontrivial[1] upper bounds on the sample complexity of any other computational tasks for monotone probability distributions over the Boolean cube are known.

[1] i.e. using monotonicity in an essential way and going beyond the bounds known for arbitrary probability distributions.

1.2.1 Estimating support size

We consider the task of additively estimating the support size of an unknown monotone probability distribution over the Boolean cube. The following assumption is standard in support size estimation:

▶ **Definition 3.** *A probability distribution over a universe of size $N_{universe}$ is called **well-behaved** (in context of support size estimation) if for every x in the set, the probability of x is either zero or at least $1/N_{universe}$.*

The purpose of this definition is to rule out pathological cases in which there are items that are in the support, yet have probability very close to zero. We henceforth adapt this definition to probability distributions over $\{0,1\}^n$, where we have $N_{\text{universe}} = 2^n$. We prove the following theorem:

▶ **Theorem 4.** *For every positive ϵ, the following is true: for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$ samples from an unknown well-behaved monotone probability distribution ρ over $\{0,1\}^n$, can reliably[2] approximate the support size of ρ with an additive error of up to ϵ. The algorithm runs in time $O_\epsilon\left(\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}\right)$*

We contrast this result to the results of [27, 30, 31, 32, 34] that show that one needs $\Omega(N_{\text{universe}}/\log(N_{\text{universe}}))$ samples to estimate the support size of an arbitrary distribution up to a sufficiently small constant, which equals to $\Omega(2^n/n)$ for a universe of size 2^n, such as the Boolean cube.

1.2.2 Estimating distance to uniformity

We now consider the task of additively estimating the distance from an unknown monotone probability distribution over the Boolean cube to the uniform distribution. We prove the following theorem:

▶ **Theorem 5.** *For every positive ϵ, the following is true: for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$ samples from an unknown monotone probability distribution ρ over $\{0,1\}^n$, can reliably approximate the distance between ρ and the uniform distribution over $\{0,1\}^n$ with an additive error of up to ϵ. The algorithm runs in time $O\left(2^{n+O_\epsilon(\sqrt{n}\log n)}\right)$.*

We, again, contrast this result to the results of [30, 31, 32] that show that one needs $\Omega(N_{\text{universe}}/\log(N_{\text{universe}}))$ samples to estimate the distance of an arbitrary distribution to the uniform distribution, which equals to $\Omega(2^n/n)$ for a universe of size 2^n, such as the Boolean cube.

We also have the following sample complexity lower bound on this task, which we prove using the sub-cube decomposition technique of [29]:

▶ **Theorem 6.** *For infinitely many positive integers n, there exist two probability distributions Δ_{Close} and Δ_{Far} over monotone distributions over $\{0,1\}^n$, satisfying:*

1. *Every distribution in Δ_{Far} is $1/2$-far from the uniform distribution.*
2. *Any algorithm that takes only $o\left(2^{\frac{n^{0.5-0.01}}{2}}\right)$ samples from a probability distribution, fails to reliably distinguish between Δ_{Close} and Δ_{Far}.*
3. *Every distribution in Δ_{Close} is $o(1)$-close to the uniform distribution.*

[2] By **reliably** we henceforth mean that the probability of success is at least $2/3$.

▶ Remark 7. In our construction, the distribution Δ_{Close} consists of only one probability distribution. Additionally, the constant 0.01 can be made arbitrarily small.

Recall that in [29, 2] it was shown that, given samples from a probability distribution over the Boolean cube that is promised to be monotone, distinguishing the uniform distribution from one that is ϵ-far from uniform can be done using only $O\left(\frac{n}{\epsilon^2}\right)$ samples. Yet, as the theorem above shows, the **tolerant** version of this problem, which requires one to distinguish a distribution that is $o(1)$-close to the uniform from a distribution that is $1/2$-far from uniform, requires $\Omega\left(2^{\frac{n^{0.5-0.01}}{2}}\right)$ samples, which is dramatically greater.

1.3 Technical overview

1.3.1 Structural results

Our analysis applies and builds upon the main structural lemma in [7]. To state it, recall that a **DNF** is a Boolean function that is formed as an OR of ANDs, and it is monotone if there are no negations. Each AND is referred to as a **clause**, with the number of variables in the AND is referred to as the **width** of the clause. Their structural lemma shows that each monotone function can be approximated by a DNF with only a constant number of distinct clause widths. Specifically:

▶ **Lemma 8** (Main Lemma in [7], abridged and restated)**.** *For every positive ϵ, for all sufficiently large n, let f be a monotone Boolean function over the domain $\{0,1\}^n$. There is a function $g = g_1 \vee ... \vee g_t$ with the following properties: (i) $t \leq 2/\epsilon$ (ii) each g_i is a monotone DNF with terms of width exactly k_i (iii) g disagrees with f at no more than $\epsilon \cdot 2^n$ elements of $\{0,1\}^n$ (iv) $g(x) \leq f(x)$ for all x in $\{0,1\}^n$.*

For Theorem 4, we use the lemma above on the indicator function of the support of the probability distribution, which allows us to prove the correctness of our algorithm. For the problems of learning and estimating the distance to uniform, we go a step further and prove an analogous structural lemma for *monotone probability distributions*.

There are some crucial differences between monotone Boolean functions in the setting of Boolean function approximation and monotone probability distributions in our setting. First of all, the basic properties of the two objects are different: a Boolean function always has one of the two values (zero or one), which is usually not the case for a probability distribution, but a probability distribution, summed over $\{0,1\}^n$, has to equal one. Secondly, the relevant notions of a function f_2 being well-approximated by a function f_1 are different: for Boolean functions we bound the fraction of points on which f_1 and f_2 disagree, whereas for monotone probability distributions we would like to bound the L_1 distance between f_1 and f_2.

To overcome these differences, we generalize to the setting of non-Boolean functions the main concept used in the proof of Lemma 8: the concept of a **minterm** of a monotone Boolean function. In [7] the minterm of a monotone Boolean function f is defined as follows:

$$\text{minterm}_f(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } f(x) = 1 \text{ and for all } y \prec x,\ f(y) = 0 \\ 0 & \text{otherwise} \end{cases}$$

Using this language, the function g in Lemma 8 can be characterized as a function, for which:

$$\sum_{h=0}^{n} \mathbf{1}_{\exists\, x \in \{0,1\}^n:\, ||x||=h \,\wedge\, \text{minterm}_g(x) \neq 0} \leq \frac{2}{\epsilon} \tag{1}$$

We introduce the notion of monotone **slack** that generalizes the notion of a minterm to non-Boolean functions:

$$\text{slack}_f(x) \stackrel{\text{def}}{=} f(x) - \max_{y \prec x} f(y) = f(x) - \max_{y \preceq x \text{ and } ||y||=||x||-1} f(y)$$

With such a definition at hand, one could hope to prove that every monotone probability distribution ρ is well-approximated in the L_1 norm by a monotone function f, for which $\sum_{h=0}^{n} \mathbf{1}_{\exists\, x \in \{0,1\}^n:\, ||x||=h \,\wedge\, \text{slack}_f(x) \neq 0}$ is bounded by a constant independent of n. We were not able to prove such a theorem, and instead we bound a related quantity that can be thought of as the weighted analogue of the expression in Equation 1: $\sum_{h=0}^{n} R_h \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n:\, ||x||=h \,\wedge\, \text{slack}_f(x) \neq 0}$, where the R_h are positive weights that can be chosen arbitrarily, as long as they satisfy a certain technical condition that ensures that not too many of these weights are too large. Precisely, our main lemma is:

▶ **Lemma 9** (Main Structural Lemma). *For all positive ζ, for all sufficiently large n, the following is true: Let ρ be a monotone probability distribution over $\{0,1\}^n$. Suppose, for each h between 0 and n we are given a positive value R_h, and it is the case that: $\sum_{h=0}^{n} R_h \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n} \binom{n}{j}} \leq \zeta$*

Then, there exists a positive monotone function f, mapping $\{0,1\}^n$ to positive real numbers, satisfying:

1. *For all x, it is the case that $\rho(x) \geq f(x)$.*
2. *It is the case that: $\sum_{x \in \{0,1\}^n} \rho(x) - f(x) \leq \zeta$*
3. *It is the case that: $\sum_{h=0}^{n} R_h \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n:\, ||x||=h \,\wedge\, slack_f(x) \neq 0} \leq 1$*

Now, as a corollary, we present a simple special case (proven to be so in Subsection 3.1) that not only illustrates the power of Lemma 9, but also is sufficient for our proof of Theorem 5:

▶ **Corollary 10.** *Let ρ be a monotone probability distribution over $\{0,1\}^n$ and let h_0 be an integer for which:*

$$\frac{\epsilon}{4} \leq \Pr_{x \sim \{0,1\}^n}[||x|| \geq h_0] \leq \frac{\epsilon}{2}$$

Then, there exists a positive monotone function $f : \{0,1\}^n \to R$ satisfying:

1. *For all x, it is the case that $\rho(x) \geq f(x)$.*
2. *It is the case that: $\sum_{x \in \{0,1\}^n} \rho(x) - f(x) \leq \frac{\epsilon}{4}$.*
3. *There exists a set of values $\{k_1, ..., k_t\}$ (ordered in an increasing order) with $t \leq \frac{16}{\epsilon^2}$, satisfying that if for some x in $\{0,1\}^n$ we have $||x|| < h_0$ and $slack_f(x) \neq 0$, then $||x|| = k_i$ for some i.*

For Theorem 2, however, we use the full power of Lemma 9.

1.3.2 Algorithmic ideas

Here we present an informal overview of the ideas involved in the design and analysis of our algorithms. Throughout we omit details and technicalities. As already mentioned, our algorithms for Theorems 4, 5 and 2 use respectively Lemma 8, Corollary 10 and Lemma 9 as their structural core. Here we present the algorithmic ideas in the order of increasing technical sophistication.

1.3.2.1 Support size estimation (Theorem 4)

The idea behind our support size estimation algorithm is as follows: if we received x as a sample, then not only x has to be in the support of ρ, but every y, satisfying $x \preceq y$ is in the support of ρ. For all such y, we say that y is **covered** by x. Our algorithms estimates the support size of ρ through estimating the number of all such y that are covered by at least one of the samples.

This algorithm can be made computationally efficient by standard methods in randomized algorithms, and the only non-trivial step is to show that $\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$ samples suffice. To show this, we first apply Lemma 8 to the indicator function of the support of ρ (which we from now on call the support function of ρ). This gives us a Boolean function g that approximates well the support function of ρ and has zero slack everywhere, except for a small number of levels[3] of $\{0,1\}^n$. For simplicity, assume that the support function of ρ itself has this property, and there are only a small number of levels of the Boolean cube on which the support function of ρ can have non-zero slack, which we call the **slacky** levels.

Now, we divide the elements of $\{0,1\}^n$ (which we also call **points**) into **good**[4] points and **bad** points, with the former defined as all the points sufficiently close to a slacky level, and the latter defined as all the other points. Clearly, a given level of $\{0,1\}^n$ consists either fully from good points or fully from bad points, so we also refer to levels as good or bad.

We argue that if a point y in the support of ρ is a good point, then it is likely to be covered by one of the samples, because there is a large number of values x in the support of ρ, for which $y \preceq x$.

We conclude by bounding the number of elements in the support of ρ that are bad, by using the fact that there cannot be too many slacky levels.

1.3.2.2 Estimation of distance to uniform (Theorem 5)

To estimate the distance of a monotone distribution to uniform, we pick a value h_0 as in Corollary 10 and break down the value of the total variation distance from ρ to uniform into contributions from two disjoint components: (i) $\{x \in \{0,1\}^n \text{ s.t. } ||x|| \geq h_0\}$ and (ii) all the other points of $\{0,1\}^n$. In other words, we use h_0 as the cutoff value for the Hamming weight, to separate $\{0,1\}^n$ into components (i) and (ii). The first contribution is straightforward to estimate simply through estimating how likely a random sample x from ρ is to have $||x|| \geq h_0$, because it is straightforward to prove that if one redistributes the probability mass of ρ in $\{x \in \{0,1\}^n \text{ s.t. } ||x|| \geq h_0\}$, while keeping the total amount of probability mass in this set fixed, the total variation distance between ρ and the uniform distribution cannot change by more than $O_\epsilon(1)$.

For any element x of the component (ii), we prepare an estimate of $\rho(x)$, which we call $\hat{\phi}(x)$. Our approach here is somewhat similar to the one for our support size estimation algorithm. In the case of support size estimation, we only registered whether x was covered by a sample from ρ or not. In this case, we actually need an estimate on $\rho(x)$ (as opposed to $\mathbf{1}_{\rho(x)\neq 0}$) which we obtain by studying the pattern of all the samples covering x. More precisely, suppose we draw N_2 samples from the distribution, which form a multiset S_2. We extract the estimate $\hat{\phi}(x)$ from the pattern of samples as follows:

$$\hat{\phi}(x) := \frac{1}{2^L} \cdot \frac{\max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y||=L} \left| \left\{ z \in S_2 : y \preceq z \preceq x \right\} \right|}{N_2}$$

[3] i.e. subsets of $\{0,1\}^n$ that have the same Hamming weight.

[4] We later re-define these notions in order to adapt them for the technical details we ignore in the introduction.

Here L is a parameter equal to $\Theta_\epsilon(\sqrt{n})$. We then estimate the contribution of set (ii) as $\sum_{x\in\{0,1\}^n:\ ||x||\geq h_0} \left|\hat{\phi}(x) - 1/2^n\right|$.

We show the correctness of our algorithm as follows. We use a tail bound to show that $\hat{\phi}(x)$ concentrates sufficiently closely to the value: .

$$\phi(x) \stackrel{\text{def}}{=} \frac{1}{2^L} \cdot \max_{y \text{ s.t. } y\preceq x \text{ and } ||x||-||y||=L} \Pr_{z\sim\rho}[y \preceq z \preceq x] =$$

$$\frac{1}{2^L} \cdot \max_{y \text{ s.t. } y\preceq x \text{ and } ||x||-||y||=L} \sum_{z \text{ s.t. } y\preceq z\preceq x} \rho(z)$$

Then, we apply Corollary 10, which implies that ρ is approximated well by a function f and a certain set of constraints on the slack of f holds.

Now for the sake of simplicity (analogously to the case of support size estimation), assume that ρ itself satisfies the condition that below the threshold h_0 there are at most $O_\epsilon(1)$ levels of $\{0,1\}^n$ on which there are points x with non-zero $\text{slack}_\rho(x)$ (in reality it is merely well-approximated by such a function). We now can (analogously to the case of support size estimation) introduce the concepts of **slacky** levels as levels on which ρ has non-zero slack, and **good** levels, which are below h_0 and farther than L from all **slacky** levels of ρ. Now, one can prove that for x on a good level the value of $\phi(x)$ equals precisely to $\rho(x)$, for the following reasons: First of all the inequality:

$$\max_{y \text{ s.t. } y\preceq x \text{ and } ||x||-||y||=L} \rho(y) \leq \phi(x) \leq \rho(x)$$

follows immediately from the monotonicity of ρ and the definition of ϕ. Secondly, if ρ has no slack on the levels between $||x||$ and $||x|| - L$ (inclusive), then from the definition of slack it follows immediately using induction on L that:

$$\max_{y \text{ s.t. } y\preceq x \text{ and } ||x||-||y||=L} \rho(y) = \rho(x)$$

Therefore, it has to be the case that $\rho(x) = \phi(x)$.

Finally, we bound the contribution to the L_1 distance between $\hat{\phi}$ and ρ of all the levels below h_0 that are not good (which we again call the **bad** levels). We do this by upper-bounding the number of bad levels, and then upper bounding the total probability mass on a single level below h_0.

1.3.2.3 Learning a monotone probability distribution (Theorem 2)

As we saw, our algorithm for the estimation of the distance to the uniform distribution contained a component that learned in L_1 distance the restriction of ρ on the levels below the cutoff h_0. The main challenge here is to extend these ideas to levels above h_0. To this, we make the following changes to our setup:

- Instead of having one fixed constant L defining whether a point is close to a slacky level, we make this value level-dependent. In other words, for every h we define L_h, and then after drawing N samples, which form a multiset S, we compute:

$$\hat{\phi}(x) := \frac{1}{2^{\lfloor L_{||x||}\rfloor}} \cdot \frac{\max_{y \text{ s.t. } y\preceq x \text{ and } ||y||-||x||=\lfloor L_{||x||}\rfloor} \left|\left\{z \in S : y \preceq z \preceq x\right\}\right|}{N}$$

- Instead of using Corollary 10, we use the the full power of Lemma 9. This, again gives us a function f that approximates ρ closely and has a restriction on its slacky levels.

Finally, we pick values of L_h in the algorithm and R_h in the analysis so we balance (i) The random error from the deviation of $\hat{\phi}(x)$ from its expectation and (ii) The systematic error introduced by the slacky levels of f and the levels close to them. As a result, we find that $\frac{2^n}{2^{\Theta(n^{1/5})}}$ samples suffice.

2 Preliminaries

We use the following basic definitions and notation:

▶ **Definition 11.** *For $x \in \{0,1\}^n$, its **Hamming weight** is denoted as $||x||$ and is equal to $\sum_i x_i$.*

▶ **Definition 12.** *For a function $f : \{0,1\}^n \to R$, we define the **average value on level** k (with $0 \le k \le n$) as: $\mu_f(k) = \frac{1}{\binom{n}{k}} \sum_{x \in \{0,1\}^n : ||x|| = k} f(x)$. We also refer to average value on level k for a probability distribution ρ, which we denote $\mu_\rho(k)$. By this we mean the average value on level k of the density function of ρ.*

▶ **Definition 13.** *For a monotone function $f : \{0,1\}^n \to R$, we define the **monotone slack** $slack_f(x)$ at point $x \in \{0,1\}^n$ as follows: $slack_f(x) \stackrel{\text{def}}{=} f(x) - \max_{y \prec x} f(y) = f(x) - \max_{y \preceq x \text{ and } ||y|| = ||x|| - 1} f(y)$. We also stipulate that $slack_f(0^n) = f(0^n)$.*

▶ **Definition 14.** *The **total variation distance** between two probability distributions ρ_1 and ρ_2 is defined as:*
$d_{TV}(\rho_1, \rho_2) \stackrel{\text{def}}{=} \frac{1}{2} \sum_x |\rho_1(x) - \rho_2(x)|$.

The following are well-known facts, which were also used in [7]:

▶ **Fact 15.** *For a monotone function $f : \{0,1\}^n \to R$, for all k_1, k_2 satisfying $0 \le k_1 \le k_2 \le n$, it is the case that $\mu_f(k_1) \le \mu_f(k_2)$.*

▶ **Fact 16.** *For all k, it is the case that $\binom{n}{k} \le \frac{2}{\sqrt{n}} \cdot 2^n$.*

Now, we justify two claims we made in the introduction:

▷ Claim 17. For sufficiently small ϵ_0, for all sufficiently large n, any algorithm that learns an unknown monotone probability distribution over $\{0,1\}^n$ requires at least $\Omega(2^{0.15n})$ samples from the distribution.

Proof. From the argument in [32, pages 1937-1938] it follows that if two probability distributions are ϵ-close in total variation distance, then their entropy values are within $2\log(N_{\text{universe}})\epsilon = 2\epsilon n$. Therefore, the task of estimating the entropy of an unknown monotone probability up to an additive error $2\epsilon n$ is not harder than learning it to withing total variation distance ϵ. But in [29, page 39] it is shown that at least $\sqrt{T}/10$ samples are required for the task of distinguishing whether the unknown monotone probability distribution has entropy at least $0.81n$ or at most $n/2 + \log T$. Picking $T = 2^{0.3n}$ gives us the desired learning lower bound. ◁

▷ Claim 18. Given Theorem 2, one can test whether an unknown distribution ρ over the Boolean cube is monotone or ϵ-far from monotone with $O(\frac{2^n}{n\epsilon^2})$ samples.

Proof. This can be done in the following way: (1) Use our learning algorithm with an error parameter $\epsilon/4$. This gives us a description of a distribution $\hat{\rho}$, which is $\epsilon/4$-close to ρ if ρ is monotone. (2) Estimate, using the estimator of [31], the total variation distance between

ρ and $\hat{\rho}$ up to $\epsilon/4$. If the result is closer to ϵ than to zero, output NO. (3) Compute the total variation distance between $\hat{\rho}$ and the closest monotone probability distribution. If this distance estimate is closer to ϵ than to zero, output NO, otherwise output YES. For constant ϵ, the sample complexity is dominated by step (3), which is $O(\frac{2^n}{\epsilon^2 n})$. It is easy to see that a monotone probability distribution will pass this test, whereas a distribution that is ϵ-far from monotone will fail either step (2) or step (3). ◁

3 Learning monotone probability distributions

Algorithm 1 Algorithm for learning a monotone probability distribution over the Boolean cube (given sample access from a distribution ρ, which is monotone over $\{0,1\}^n$).

1. Set $A := \frac{1}{2n} \cdot e^{\frac{1}{2000} \cdot n^{1/5}}$. For all $h \geq n/2$, set $L_h := \max\left(\log\left(2nA \cdot \frac{\binom{n}{h}}{2^n}\right), 0\right)$
   Similarly, for all h, satisfying $n/2 > h \geq 0$, set: $L_h := L_{n/2} = \log\left(2nA \cdot \frac{\binom{n}{n/2}}{2^n}\right)$.
2. Set $N := \frac{2^n}{A} \cdot \frac{192}{\epsilon^2} \cdot (n + 9\sqrt{n} + 4)$
   Draw N samples from the probability distribution ρ and denote the multiset of these samples as S.
3. For all x in $\{0,1\}^n$, if $||x|| < 9\sqrt{n}$, then set $\hat{\phi}(x) = 0$, otherwise compute:
$$\hat{\phi}(x) := \frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \frac{\max_{y \text{ s.t. } y \preceq x \text{ and } ||y|| - ||x|| = \lfloor L_{||x||} \rfloor} \left| \left\{ z \in S : y \preceq z \preceq x \right\} \right|}{N}$$
   Do this by first making a look-up table, which given arbitrary $z \in \{0,1\}^n$ returns the number of times z was encountered in S. Then, use this look-up table to compute the necessary values of $|\{z \in S : y \preceq z \preceq x\}|$ by querying all these values of z in the lookup table and summing the results up.
4. For all x in $\{0,1\}^n$, compute the following: $\hat{\rho}(x) = \hat{\phi}(x) + \frac{1}{2^n}\left(1 - \sum_{y \in \{0,1\}^n} \hat{\phi}(y)\right)$
5. Output the value table of $\hat{\rho}$.

In this section we prove our upper-bound on the sample complexity of learning an unknown monotone probability distribution over the Boolean cube. We restate the theorem:

▶ **Theorem 2.** *For every positive ϵ, such that $0 < \epsilon \leq 1$ and for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(n^{1/5})}}$ samples from an unknown monotone probability distribution ρ over $\{0,1\}^n$, can reliably return a description of an estimate probability distribution $\hat{\rho}$, such that $d_{TV}(\rho, \hat{\rho}) \leq \epsilon$. The algorithm runs in time $O\left(2^{n + O_\epsilon(n^{1/5} \log n)}\right)$.*

Proof. We present the algorithm as Algorithm 1. The number of samples drawn from ρ is $N = \frac{2^n}{2^{\Theta_\epsilon(n^{1/5})}}$. The run-time, in turn, is dominated by computing the values of $\hat{\phi}$ in step (3), in which the construction of the lookup table takes $O(n \cdot 2^n)$ time, and the time spent computing each $\hat{\phi}(x)$ can be upper bounded by the product of: (i) the number of pairs (y, z) that simultaneously satisfy $y \preceq z \preceq x$ and $||y|| - ||x|| = L_{||x||}$, which can be upper-bounded by $O(n^{L_{||x||}} \cdot 2^{L_{||x||}})$ and (ii) the time it takes to look up a given z in the lookup table, which can be upper-bounded by $O(n)$. Overall, this gives us a run-time upper bound of $O(2^{n + O_\epsilon(n^{1/5} \log n)})$.

Now, the only thing to prove is correctness. Here is our main claim:

▷ Claim 19. If the following conditions are the case:

a) As a function of h, L_h is non-increasing.

b) For all h, we have that $L_h \leq 9\sqrt{n}$.

c)

$$\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} \leq \frac{1}{2}$$

d)

$$\sum_{h=9\sqrt{n}}^{n} L_h \cdot \left(\begin{cases} \frac{400}{n^{2.5}} & \text{if } h \leq n/2 - \sqrt{n\ln(n)} \\ \frac{40000}{n} & \text{if } n/2 - \sqrt{n\ln(n)} < h < n/2 + \sqrt{n} \\ 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) \leq \frac{\epsilon^2}{20000}$$

Then, with probability at least $2/3$, it is the case that $\sum_{x\in\{0,1\}^n \text{ s.t. } 9\sqrt{n}\leq||x||} \left|\hat{\phi}(x) - \rho(x)\right| \leq \frac{\epsilon}{2}$.

We verify in Appendix A, subsection 7.1, that L_h indeed satisfy the conditions above. In fact, the values of L_h and A were chosen specifically to satisfy the constraints above. We prove Claim 19 in Section 3.2, after we develop our main structural lemma in Section 3.1.

We now bound the contribution to the L_1 distance between $\hat{\phi}$ to ρ that comes from points of Hamming weight less than $9\sqrt{n}$. Since $\sum_{x\in\{0,1\}^n} \rho(x) = 1$ and ρ is monotone, then whenever $||x|| \leq n/2$ we have $\rho(x) \leq 1/2^{n/2}$. Therefore, for sufficiently large n we have:

$$\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||<9\sqrt{n}} \left|\hat{\phi}(x) - \rho(x)\right| = \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||<9\sqrt{n}} \rho(x) \leq \frac{n^{9\sqrt{n}}}{2^{n/2}} \leq \frac{\epsilon}{2}$$

Combining this with the bound in Claim 19 we get:

$$\sum_{x\in\{0,1\}^n} \left|\hat{\phi}(x) - \rho(x)\right| \leq \epsilon$$

Overall, we have:

$$2 \cdot d_{\mathrm{TV}}(\rho, \hat{\rho}) = \sum_{x\in\{0,1\}^n} \left|\hat{\rho}(x) - \rho(x)\right| =$$

$$\sum_{x\in\{0,1\}^n} \left|\hat{\phi}(x) - \rho(x) + \frac{1}{2^n}\left(1 - \sum_{y\in\{0,1\}^n} \hat{\phi}(y)\right)\right| \leq$$

$$\sum_{x\in\{0,1\}^n} \left|\hat{\phi}(x) - \rho(x)\right| + \left|1 - \sum_{y\in\{0,1\}^n} \hat{\phi}(y)\right| =$$

$$\sum_{x\in\{0,1\}^n} \left|\hat{\phi}(x) - \rho(x)\right| + \left|\sum_{x\in\{0,1\}^n} \rho(x) - \hat{\phi}(x)\right| \leq 2 \cdot \sum_{x\in\{0,1\}^n} \left|\hat{\phi}(x) - \rho(x)\right| \leq 2 \cdot \epsilon$$

Thus, with probability at least $2/3$, we have $d_{\mathrm{TV}}(\rho, \hat{\rho}) \leq \epsilon$. ◀

3.1 Main lemma

Here we prove the following structural lemma. The lemma, as well as its proof are inspired by the main structural lemma of [7] (i.e. Lemma 8). Recall that the slack of a monotone function was given in Definition 13.

▶ **Lemma 9** (Main Structural Lemma). *For all positive ζ, for all sufficiently large n, the following is true: Let ρ be a monotone probability distribution over $\{0,1\}^n$. Suppose, for each h between 0 and n we are given a positive value R_h, and it is the case that:* $\sum_{h=0}^{n} R_h \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n}\binom{n}{j}} \leq \zeta$

Then, there exists a positive monotone function f, mapping $\{0,1\}^n$ to positive real numbers, satisfying:

1. *For all x, it is the case that $\rho(x) \geq f(x)$.*
2. *It is the case that: $\sum_{x\in\{0,1\}^n} \rho(x) - f(x) \leq \zeta$*
3. *It is the case that: $\sum_{h=0}^{n} R_h \cdot \mathbf{1}_{\exists\, x\in\{0,1\}^n:\, ||x||=h \,\wedge\, slack_f(x)\neq 0} \leq 1$*

Proof. We use the following process to obtain f:

a) Set $f^* = \rho$.

b) For $h = 0$ to n:

- If it is the case that:

$$\frac{1}{\binom{n}{h}} \cdot \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||=h} \text{slack}_{f^*}(x) < R_h \cdot \frac{1}{\sum_{j=h}^{n}\binom{n}{h}} \tag{2}$$

  Then, for all x in $\{0,1\}^n$, satisfying $||x|| = h$ set: $f^*(x) := f^*(x) - \text{slack}_{f^*}(x)$.

c) Set $f = f^*$ and output f.

By inspection, f^* remains monotone and positive at every iteration of the process. Therefore, f is also monotone and positive.

Property (1) in the Lemma is true, because at every step of the process, values of f^* only decrease.

To see why Property (2) is the case, note that the value $\sum_{x\in\{0,1\}^n} \rho(x) - f(x)$ is zero in the beginning of the process, and at a step h it either stays the same or decreases by at most $R_h \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n}\binom{n}{j}}$. Therefore we can upper-bound:

$$\sum_{x\in\{0,1\}^n} \rho(x) - f(x) \leq \sum_{h=0}^{n} R_h \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n}\binom{n}{j}} \leq \zeta$$

Now, the only thing left to prove is that property (3) holds.

From the definition of monotone slack, it follows that modifying the value of a function on points of Hamming weight j does not affect the slack on any point with Hamming weight lower than j. Therefore, the value $\frac{1}{\binom{n}{j}} \cdot \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||=j} \text{slack}_{f^*}(x)$ will not change as f^* changes after the jth iteration. Therefore, this value will be equal to $\frac{1}{\binom{n}{j}} \cdot \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||=j} \text{slack}_{f}(x)$. Thus, the value of $\frac{1}{\binom{n}{j}} \cdot \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||=j} \text{slack}_{f}(x)$ is either zero or at least $R_h \cdot \frac{1}{\sum_{j=h}^{n}\binom{n}{h}}$.

Now, we need the following generalization of Fact 15:

▶ **Observation 20.** *Let f be an arbitrary monotone function $\{0,1\}^n \to R$. Then, for any k in $[0, n-1]$ it is the case that:*

$$\mu_f(k+1) \geq \mu_f(k) + \frac{1}{\binom{n}{k+1}} \cdot \sum_{x\in\{0,1\}^n \text{ s.t. } ||x||=k+1} slack_f(x)$$

Proof. For all x with $||x|| = k+1$ we have that:

$$f(x) = \text{slack}_f(x) + \max_{y \in \{0,1\}^n \text{ s.t. } ||y||=k \text{ and } y \preceq x} f(y)$$

We have that:

$$\max_{y \in \{0,1\}^n \text{ s.t. } ||y||=k \text{ and } y \preceq x} f(y) \geq \mathbb{E}_{y \sim \{0,1\}^n \text{ conditioned on } ||y||=k \text{ and } y \preceq x}[f(y)]$$

Therefore:

$$f(x) \geq \text{slack}_f(x) + \mathbb{E}_{y \sim \{0,1\}^n \text{ conditioned on } ||y||=k \text{ and } y \preceq x}[f(y)]$$

Averaging the both sides, we get:

$$\mu_f(k+1) \geq \frac{1}{\binom{n}{k+1}} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } ||x||=k+1} \text{slack}_f(x) +$$

$$\mathbb{E}_{x \sim \{0,1\}^n \text{ conditioned on } ||x||=k+1} \mathbb{E}_{y \sim \{0,1\}^n \text{ conditioned on } ||y||=k \text{ and } y \preceq x}[f(y)] =$$

$$\frac{1}{\binom{n}{k+1}} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } ||x||=k+1} \text{slack}_f(x) + \mathbb{E}_{y \sim \{0,1\}^n \text{ conditioned on } ||y||=k}[f(y)] =$$

$$\frac{1}{\binom{n}{k+1}} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } ||x||=k+1} \text{slack}_f(x) + \mu_f(k) \quad (3)$$

Above, the penultimate equality followed from a simple probabilistic fact: if one picks a random n-bit string of Hamming weight $k+1$ and then sets to zero a random bit that equals to one, this is equivalent to picking a random n-bit string of weight k. ◀

Using the Observation 20 repeatedly and recalling that in Definition 13 we defined $\text{slack}_f(0^n) = f(0^n)$, we get that for all h:

$$\mu_f(h) \geq \mu_f(0) + \sum_{k=1}^{h} \frac{1}{\binom{n}{k}} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } ||x||=k} \text{slack}_f(x) =$$

$$\sum_{k=0}^{h} \frac{1}{\binom{n}{k}} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } ||x||=k} \text{slack}_f(x) \geq \sum_{k=0}^{h} R_k \cdot \frac{1}{\sum_{j=k}^{n} \binom{n}{j}} \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n :\, ||x||=k \,\wedge\, \text{slack}_f(x) \neq 0}$$

Summing this up over all h and changing the order of summations, we get:

$$1 = \sum_{x \in \{0,1\}^n} \rho(x) \geq \sum_{x \in \{0,1\}^n} f(x) = \sum_{h=0}^{n} \binom{n}{h} \mu_f(h) \geq$$

$$\sum_{h=0}^{n} \binom{n}{h} \sum_{k=0}^{h} R_k \cdot \frac{1}{\sum_{j=k}^{n} \binom{n}{j}} \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n :\, ||x||=k \,\wedge\, \text{slack}_f(x) \neq 0} =$$

$$\sum_{k=0}^{n} \sum_{h=k}^{n} \binom{n}{h} R_k \cdot \frac{1}{\sum_{j=k}^{n} \binom{n}{j}} \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n :\, ||x||=k \,\wedge\, \text{slack}_f(x) \neq 0} =$$

$$\sum_{k=0}^{n} R_k \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n :\, ||x||=k \,\wedge\, \text{slack}_f(x) \neq 0}$$

This finishes the proof of the lemma. ◀

Now, we prove the following corollary:

▶ **Corollary 10.** *Let ρ be a monotone probability distribution over $\{0,1\}^n$ and let h_0 be an integer for which:*

$$\frac{\epsilon}{4} \leq \Pr_{x \sim \{0,1\}^n}[||x|| \geq h_0] \leq \frac{\epsilon}{2}$$

Then, there exists a positive monotone function $f : \{0,1\}^n \to R$ satisfying:

1. *For all x, it is the case that $\rho(x) \geq f(x)$.*
2. *It is the case that: $\sum_{x \in \{0,1\}^n} \rho(x) - f(x) \leq \frac{\epsilon}{4}$.*
3. *There exists a set of values $\{k_1, ..., k_t\}$ (ordered in an increasing order) with $t \leq \frac{16}{\epsilon^2}$, satisfying that if for some x in $\{0,1\}^n$ we have $||x|| < h_0$ and $slack_f(x) \neq 0$, then $||x|| = k_i$ for some i.*

Proof. We use Lemma 9, setting $\zeta = \epsilon/4$ and

$$R_h = \begin{cases} \frac{\epsilon^2}{16} & \text{if } h \leq h_0 \\ 0 & \text{otherwise} \end{cases}$$

We verify the precondition to Lemma 9, by using that $\sum_{x \in \{0,1\}^n} \rho(x) - f(x) \leq \frac{\epsilon}{4}$:

$$\sum_{h=0}^{n} R_h \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n} \binom{n}{j}} = \sum_{h=0}^{h_0} \frac{\epsilon^2}{16} \cdot \frac{\binom{n}{h}}{\sum_{j=h}^{n} \binom{n}{j}} \leq \sum_{h=0}^{h_0} \frac{\epsilon^2}{16} \cdot \frac{\binom{n}{h}}{\sum_{j=h_0}^{n} \binom{n}{j}} \leq$$

$$\sum_{h=0}^{h_0} \frac{\epsilon^2}{16} \cdot \frac{\binom{n}{h}}{2^n \cdot \epsilon/4} = \frac{\epsilon}{4} \cdot \sum_{h=0}^{h_0} \frac{\binom{n}{h}}{2^n} \leq \frac{\epsilon}{4}$$

Now, we simply check that properties (1), (2) and (3) of the Lemma directly imply the properties (1), (2) and (3) of the Corollary respectively. This completes the proof. ◀

To use Lemma 9, we need an upper bound on the value of $\frac{\binom{n}{h}}{\sum_{j \geq h}^{n} \binom{n}{j}}$. The following claim provides such an upper bound:

▷ Claim 21. For all sufficiently large n, for all h, satisfying $0 \leq h \leq n$, it is the case that:

$$\frac{\binom{n}{h}}{\sum_{j \geq h}^{n} \binom{n}{j}} \leq \left(\begin{cases} \frac{2}{n^2} & \text{if } h \leq n/2 - \sqrt{n \ln(n)} \\ \frac{200}{\sqrt{n}} & \text{if } n/2 - \sqrt{n \ln(n)} < h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h - n/2}{n} & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right)$$

Proof. See Appendix A, Subsection 7.2 ◁

3.2 Proof of Claim 19

For all x in $\{0,1\}^n$, satisfying $9\sqrt{n} \leq ||x||$, we define the following quantity:

$$\phi(x) \stackrel{\text{def}}{=} \frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x|| - ||y|| = \lfloor L_{||x||} \rfloor} \Pr_{z \sim \rho}[y \preceq z \preceq x] =$$

$$\frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x|| - ||y|| = \lfloor L_{||x||} \rfloor} \sum_{z \text{ s.t. } y \preceq z \preceq x} \rho(z) \quad (4)$$

Observe that since for every such x and y there are $2^{\lfloor L_{||x||} \rfloor}$ values of z satisfying $y \preceq z \preceq x$, and ρ is a monotone probability distribution, it has to be the case that $\phi(x) \leq \rho(x)$ for all x on which $\phi(x)$ is defined.

More interestingly, we will be claiming that ϕ is (in terms of L_1 distance) a good approximation to ρ, but first we will show that $\hat{\phi}$ is a good approximation to ϕ, assuming that the values L_h are not too small:

▷ Claim 22. If it is the case that $\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} \leq \frac{1}{2}$, then, with probability at least $7/8$, it is the case that:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \hat{\phi}(x) - \phi(x) \right| \leq \frac{\epsilon}{4} \tag{5}$$

Proof. See Appendix A, Subsection 7.3, for the proof, which follows using tail bounds. ◁

Now, we apply Lemma 9 to ρ, with value $\zeta := \epsilon/100$. For now, we postpone setting the values of R_h, which we will do later in our derivation (of course, we will then check that the required constraint is indeed satisfied by these values).

This gives a positive monotone function f that satisfies the three conditions of Lemma 9. We separate all the values of x in $\{0,1\}^n$ for which $9\sqrt{n} \leq ||x||$ into two kinds: **good** and **bad**. We say that x is **bad** if there is some y for which $0 \leq ||x|| - ||y|| < \lfloor L_{||x||} \rfloor$ and $\text{slack}_f(y)$ is non-zero. Otherwise, x if **good**. Clearly, for a given Hamming weight value, wither every point with this Hamming weight is good, or every such point is bad.

We can write:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} |\phi(x) - f(x)| = \sum_{\text{good } x} |\phi(x) - f(x)| + \sum_{\text{bad } x} |\phi(x) - f(x)| \tag{6}$$

Now, we bound the two terms above separately. If x is good, then it is the case that for all y satisfying $||x|| - \lfloor L_{||x||} \rfloor < ||y|| \leq ||x||$ we have $\text{slack}_f(x) = 0$, and therefore $f(y) = \max_{y' \in \{0,1\}^n \text{ s. t. } y' \preceq y \text{ and } ||y|| - ||y'|| = 1} f(y')$. Using this relation recursively, we obtain that:

$$f(x) = \max_{y \in \{0,1\}^n \text{ s. t. } y \preceq x \text{ and } ||x|| - ||y|| = \lfloor L_{||x||} \rfloor} f(y)$$

Therefore, since f is monotone, we obtain that:

$$f(x) = \frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x|| - ||y|| = \lfloor L_{||x||} \rfloor} \sum_{z \text{ s.t. } y \preceq z \preceq x} f(z)$$

By Lemma 9, it is the case $\rho(x) \geq f(x)$. This, together with the equation above and Equation 4 implies:

$$\phi(x) \geq f(x)$$

But we also know that $\rho(x) \geq \phi(x)$. Therefore:

$$\sum_{\text{good } x} |\phi(x) - f(x)| \leq \sum_{\text{good } x} |\rho(x) - f(x)| \leq \frac{\epsilon}{4} \tag{7}$$

Where the last inequality follows from Lemma 9.

Now, we bound the contribution of bad points. Since $\phi(x) \leq \rho(x)$, $f(x) \leq \rho(x)$ and recalling the definition of a bad point, we get:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \sum_{\text{bad } x} \max(\phi(x), f(x)) \leq \sum_{\text{bad } x} \rho(x) \leq \sum_{h_2=9\sqrt{n}}^{n} \mu_\rho(h_2) \cdot \binom{n}{h_2} \cdot \mathbf{1}_{\exists\, x\in\{0,1\}^n:\, \left(h_2 - \lfloor L_{h_2} \rfloor < ||x|| \leq h_2\right) \wedge \text{slack}_f(x) \neq 0} \quad (8)$$

Since Lemma 9 gives us a bound on a weighed sum of indicator variables of the form $\mathbf{1}_{\exists\, x\in\{0,1\}^n:\, ||x||=h \wedge \text{slack}_f(x)\neq 0}$, we would like to upper-bound the expression above by such a weighted sum. To do this, to every Hamming weight value h that has a point x with non-zero $\text{slack}_f(x)$ (we call such Hamming weight value h **slacky**) we "charge" every value h_2, for which points of Hamming weight h_2 are rendered bad because h is slacky. This will happen only if $h_2 \geq h$ and $h_2 - \lfloor L_{h_2} \rfloor < h$. But since $\lfloor L_{h_2} \rfloor$ can only decrease as h_2 increases, the latter can happen only if $h_2 - \lfloor L_h \rfloor < h$. Therefore:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \sum_{h=0}^{n} \left(\sum_{h_2=h}^{h+\lfloor L_h \rfloor - 1} \mu_\rho(h_2) \cdot \binom{n}{h_2} \right) \cdot \mathbf{1}_{\exists\, x\in\{0,1\}^n:\, ||x||=h \wedge \text{slack}_f(x)\neq 0} \quad (9)$$

Now, to upper-bound $\mu_\rho(h_2)$, we need the following claim:

▷ Claim 23. For any monotone probability distribution ρ it is the case that for all h:

$$\mu_\rho(h) \leq \frac{1}{\sum_{j=h}^{n} \binom{n}{j}}$$

Proof. This follows immediately from Fact 15 and that $\sum_{x\in\{0,1\}^n} \rho(x) = 1$. ◁

Claim 23, Equation 8 and Claim 21 together imply:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \sum_{h=0}^{n} \left(\sum_{h_2=h}^{h+\lfloor L_h \rfloor - 1} \frac{\binom{n}{h_2}}{\sum_{j=h_2}^{n} \binom{n}{j}} \right) \mathbf{1}_{\left(\substack{\exists\, x\in\{0,1\}^n: \\ ||x||=h \wedge \text{slack}_f(x)\neq 0}\right)} \leq$$

$$\sum_{h=0}^{n} \left(\sum_{h_2=h}^{h+\lfloor L_h \rfloor - 1} \left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h_2 < n/2 + \sqrt{n} \\ 200 \cdot \frac{h_2 - n/2}{n} & \text{if } h_2 \geq n/2 + \sqrt{n} \end{cases} \right) \right) \mathbf{1}_{\left(\substack{\exists\, x\in\{0,1\}^n: \\ ||x||=h \wedge \text{slack}_f(x)\neq 0}\right)} \leq$$

$$\sum_{h=0}^{n} L_h \cdot \left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h + L_h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h + L_h - n/2}{n} & \text{if } h + L_h \geq n/2 + \sqrt{n} \end{cases} \right) \mathbf{1}_{\left(\substack{\exists\, x\in\{0,1\}^n: \\ ||x||=h \wedge \text{slack}_f(x)\neq 0}\right)} \quad (10)$$

Now, we claim that:

$$\left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h + L_h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h + L_h - n/2}{n} & \text{if } h + L_h \geq n/2 + \sqrt{n} \end{cases} \right) \leq 10 \cdot \left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h - n/2}{n} & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) \quad (11)$$

This follows by considering three cases (i) $h + L_h < n/2 + \sqrt{n}$, in which case this is equivalent to $\frac{200}{\sqrt{n}} \leq \frac{2000}{\sqrt{n}}$, which is trivially true. (ii) $h \geq n/2 + \sqrt{n}$, in which case since $L_h \leq 9\sqrt{n}$, we have that $\frac{h + L_h - n/2}{n} \leq 10 \cdot \frac{h - n/2}{n}$ (iii) $h + L_h \geq n/2 + \sqrt{n}$, but $h < n/2 + \sqrt{n}$, in which case since $L_h \leq 9\sqrt{n}$, we have that $\frac{h + L_h - n/2}{n} \leq \frac{\sqrt{n} + L_h}{n} \leq 10\sqrt{n}$.

Combining Equations 10 and 11, we get:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq$$
$$\sum_{h=0}^{n} L_h \cdot 10 \cdot \left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h-n/2}{n} & \text{otherwise} \end{cases} \right) \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n \colon ||x|| = h \,\wedge\, \text{slack}_f(x) \neq 0} \quad (12)$$

Recall that we postponed setting the values of R_h. The equation above motivates us to set:

$$R_h := \frac{200}{\epsilon} \cdot L_h \cdot \left(\begin{cases} \frac{200}{\sqrt{n}} & \text{if } h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h-n/2}{n} & \text{otherwise} \end{cases} \right)$$

Now, we check the constraint on R_h in Lemma 9. Using Claim 21 and the premise of Claim 19:

$$\sum_{h=0}^{n} R_h \cdot \frac{\binom{n}{h}}{\sum_{j \geq h}^{n} \binom{n}{j}} \leq \sum_{h=0}^{n} R_h \cdot \left(\begin{cases} \frac{2}{n^2} & \text{if } h \leq n/2 - \sqrt{n \ln(n)} \\ \frac{200}{\sqrt{n}} & \text{if } n/2 - \sqrt{n \ln(n)} < h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h-n/2}{n} & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) =$$
$$\frac{200}{\epsilon} \cdot \sum_{h=0}^{n} L_h \cdot \left(\begin{cases} \frac{400}{n^{2.5}} & \text{if } h \leq n/2 - \sqrt{n \ln(n)} \\ \frac{40000}{n} & \text{if } n/2 - \sqrt{n \ln(n)} < h < n/2 + \sqrt{n} \\ 40000 \cdot \left(\frac{h-n/2}{n} \right)^2 & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) \leq \frac{\epsilon}{100} = \zeta$$

Therefore, Lemma 9, together with Equation 12 implies that:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \sum_{h=0}^{n} \frac{\epsilon}{20} \cdot R_h \cdot \mathbf{1}_{\exists\, x \in \{0,1\}^n \colon ||x|| = h \,\wedge\, \text{slack}_f(x) \neq 0} \leq \frac{\epsilon}{20}$$

Now, using triangle inequality and then combining the inequality above with Equations 28, 6 and 7 we get:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \hat{\phi}(x) - \rho(x) \right| \leq$$
$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \hat{\phi}(x) - \phi(x) \right| + \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \phi(x) - \rho(x) \right| \leq$$
$$\frac{\epsilon}{4} + \frac{\epsilon}{100} + \frac{\epsilon}{20} \leq \frac{\epsilon}{2} \quad (13)$$

4 Estimating the distance to uniform

In this section we prove our upper-bound on the sample complexity of estimating the distance from uniform of an unknown monotone probability distribution over the Boolean cube. We restate the theorem:

▶ **Theorem 5.** *For every positive ϵ, the following is true: for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$ samples from an unknown monotone probability distribution ρ over $\{0,1\}^n$, can reliably approximate the distance between ρ and the uniform distribution over $\{0,1\}^n$ with an additive error of up to ϵ. The algorithm runs in time $O\left(2^{n+O_\epsilon(\sqrt{n}\log n)}\right)$.*

Algorithm 2 Algorithm for the estimation of distance to uniform efficiently (given sample access from a distribution ρ, which is monotone over $\{0,1\}^n$).

1. Pick set h_0 to be an integer for which it is the case that:

$$\frac{\epsilon}{4} \leq \Pr_{x \sim \{0,1\}^n}[||x|| \geq h_0] \leq \frac{\epsilon}{2} \tag{14}$$

   Do this by going through every integer candidate $h_{\text{candidate}}$ in the interval and computing the fraction of points x in $\{0,1\}$ for which $||x|| \geq h_{\text{candidate}}$. Finally, pick h_0 to be one of $h_{\text{candidate}}$ for which the relation above holds.
2. Set $N_1 := \frac{32 \ln 2}{\epsilon^2}$. Draw N_1 samples from the probability distribution ρ and denote the multiset of these samples as S_1.
3. Set:

$$\hat{d}_1 := \frac{1}{2} \cdot \frac{\left|\left\{z \in S_1 : ||z|| \geq h_0\right\}\right|}{N_1}$$

4. Set $L := \left\lfloor \frac{\sqrt{n}\epsilon^4}{512} \right\rfloor$.
5. Set

$$N_2 := \frac{2^n}{2^L} \cdot \frac{192}{\epsilon^2} \cdot \left(n \ln 2 + L \ln n + 4 \ln 2\right)$$

   Draw N_2 samples from the probability distribution ρ and denote the multiset of these samples as S_2.
6. For all x, satisfying $L \leq ||x|| < h_0$, compute:

$$\hat{\phi}(x) := \frac{1}{2^L} \cdot \frac{\max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y||=L} \left|\left\{z \in S_2 : y \preceq z \preceq x\right\}\right|}{N_2}$$

   Do this by first making a look-up table, which given arbitrary $z \in \{0,1\}^n$ returns the number of times z was encountered in S_2. Then, use this look-up table to compute the necessary values of $|\{z \in S_2 : y \preceq z \preceq x\}|$ by querying all these values of z in the lookup table and summing the results up.
7. Compute the following:

$$\hat{d}_2 := \frac{1}{2} \cdot \sum_{x \text{ s.t. } L \leq ||x|| < h_0} \left|\hat{\phi}(x) - \frac{1}{2^n}\right|$$

8. Output $\hat{d}_1 + \hat{d}_2$.

Proof. We present the algorithm as Algorithm 2. The number of samples drawn from ρ is $N_1 + N_2 = \frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$. The run-time, in turn, is dominated[5] by computing the values of $\hat{\phi}$ in step (6), in which the construction of the lookup table takes $O(n \cdot 2^n)$ time and the time spent computing each $\hat{\phi}(x)$ can be upper bounded by the product of: (i) the number of pairs (y, z) that simultaneously satisfy $y \preceq z \preceq x$ and $||x|| - ||y|| = L$, which can be upper-bounded by $O(n^L \cdot 2^L)$ and (ii) the time it takes to look up a given z in the lookup table, which can be upper-bounded by $O(n)$. Overall, this gives us a run-time upper bound of $O(2^{n+O_\epsilon(\sqrt{n}\log n)})$.

Now, the only thing left to prove is correctness. First of all, it is not a priori clear that there exists a value of h_0 satisfying Equation 14 (in Algorithm 2). This is true for the following reason: imagine changing $h_{\text{candidate}}$ from n to 0 by decrementing it in steps of one. Then $\Pr_{x\in\{0,1\}^n}[||x|| \geq h_{\text{candidate}}]$ will increase from $\frac{1}{2^n}$ to 1 and by Fact 16 it will not increase by more than $\frac{2}{\sqrt{n}}$ at any given step. For sufficiently large n we have $\frac{2}{\sqrt{n}} < \frac{\epsilon}{4}$. Then it is impossible to skip over the interval between $\frac{\epsilon}{4}$ and $\frac{\epsilon}{2}$ in just one step of length at most $\frac{2}{\sqrt{n}}$, and therefore Equation 14 (in Algorithm 2) will be the case for some value of $h_{\text{candidate}}$.

We decompose the total variation distance between ρ and the uniform distribution into three terms:

$$\frac{1}{2}\cdot\sum_{x\in\{0,1\}^n}\left|\rho(x)-\frac{1}{2^n}\right| = \frac{1}{2}\cdot\sum_{\substack{x\in\{0,1\}^n \text{ s.t.}\\ ||x||\geq h_0}}\left|\rho(x)-\frac{1}{2^n}\right| + \frac{1}{2}\cdot\sum_{\substack{x\in\{0,1\}^n \text{ s.t.}\\ L\leq||x||<h_0}}\left|\rho(x)-\frac{1}{2^n}\right| + \frac{1}{2}\cdot\sum_{\substack{x\in\{0,1\}^n \text{ s.t.}\\ ||x||<L}}\left|\rho(x)-\frac{1}{2^n}\right| \tag{15}$$

We argue that the first term is well approximated by $\hat{d}_1$, the second term is well approximated by $\hat{d}_2$, and the third term is negligible. As the reader will see, out of these three terms, the middle term is the least trivial to prove guarantees for.

We will first handle the first term: From the triangle inequality, Hoeffding's bound and Equation 14 (in Algorithm 2) it follows immediately that with probability at least 7/8 it is the case that:

$$\left|\hat{d}_1 - \frac{1}{2}\cdot\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||\geq h_0}\left|\rho(x)-\frac{1}{2^n}\right|\right| \leq \left|\hat{d}_1 - \frac{1}{2}\cdot\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||\geq h_0}\rho(x)\right| + \frac{1}{2}\cdot\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||\geq h_0}\frac{1}{2^n} \leq \frac{\epsilon}{8}+\frac{\epsilon}{4} = \frac{3\epsilon}{8} \tag{16}$$

Now, we use the two following facts: (i) Since $\sum_x \rho(x) = 1$ and ρ is monotone, for every x with $||x|| \leq L$ it should be the case that $\rho(x) \leq \frac{1}{2^{n-L}}$. (ii) The number of different values of x in $\{0,1\}^n$ for which $||x|| \leq L$ can be upper bounded by n^L. We get for sufficiently large n:

$$\frac{1}{2}\cdot\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||<L}\left|\rho(x)-\frac{1}{2^n}\right| \leq \frac{1}{2}\cdot\sum_{x\in\{0,1\}^n \text{ s.t. } ||x||<L}\left(\frac{1}{2^n}+\rho(x)\right) \leq \frac{1}{2}\cdot n^L\cdot\left(\frac{1}{2^n}+\frac{1}{2^{n-L}}\right) = o(1) \leq \frac{\epsilon}{8} \tag{17}$$

[5] Step 1 requires only $2^n\text{poly}(n)$ time, which is less than what step (6) requires. By inspection, other steps require even less run-time. Incidentally, the task in step 1 can be done much faster by randomized sampling, but since this is not the run-time bottleneck, we use this direct approach for the sake of simplicity.

The rest of this section will be dedicated to proving the following claim:

▷ Claim 24. With probability at least 7/8 it is the case that:

$$\left|\hat{d}_2 - \frac{1}{2} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \left|\rho(x) - \frac{1}{2^n}\right|\right| \leq \frac{\epsilon}{2} \tag{18}$$

Once this is proven, it follows by a union bound that with probability at least 3/4 both Equations 16 and 18 will be the case. This, together with Equation 17, when substituted into Equation 15 will imply that:

$$\left| \sum_{x \in \{0,1\}^n} \left|\rho(x) - \frac{1}{2^n}\right| - (\hat{d}_1 + \hat{d}_2)\right| \leq \epsilon$$

This will imply the correctness of our algorithm. ◀

4.1 Proof of Claim 24

For all x in $\{0,1\}^n$, satisfying $L \leq ||x|| < h_0$, we define the following quantity:

$$\phi(x) \stackrel{\text{def}}{=} \frac{1}{2^L} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y||=L} \Pr_{z \sim \rho}[y \preceq z \preceq x] = \frac{1}{2^L} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y||=L} \sum_{z \text{ s.t. } y \preceq z \preceq x} \rho(z) \tag{19}$$

Observe that since for every such x and y there are 2^L values of z satisfying $y \preceq z \preceq x$, and ρ is a monotone probability distribution, it has to be the case that $\phi(x) \leq \rho(x)$ for all x on which $\phi(x)$ is defined.

We will be claiming that $\phi(x)$ is (in terms of L_1 distance) a good approximation to $\rho(x)$, but first we will show that $\hat{\phi}(x)$ is a good approximation to $\phi(x)$:

▷ Claim 25. With probability at least 7/8, it is the case that:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \left|\hat{\phi}(x) - \phi(x)\right| \leq \frac{\epsilon}{4} \tag{20}$$

Proof. We claim that for any pair (x, y), such that ϕ is defined on x and $||x|| - ||y|| = L$, with probability at least $1 - \frac{1}{8 \cdot 2^n \cdot n^L}$ the following holds:

$$\frac{1}{2^L}\left|\Pr_{z \sim \rho}[y \preceq z \preceq x] - \frac{|\{z \in S : y \preceq z \preceq x\}|}{N_2}\right| \leq \frac{\epsilon}{8} \cdot \max\left(\frac{1}{2^n}, \frac{1}{2^L}\Pr_{z \sim \rho}[y \preceq z \preceq x]\right) \tag{21}$$

We use Chernoff's bound to prove this as follows. Denote by q the value $\Pr_{z \sim \rho}[y \preceq z \preceq x]$. If $q \geq \frac{2^L}{2^n}$ then by Chernoff's bound we have:

$$\Pr\left[\left||\{z \in S : y \preceq z \preceq x\}| - qN_2\right| \geq \frac{\epsilon}{8}qN_2\right] \leq 2\exp\left(-\frac{1}{3}\left(\frac{\epsilon}{8}\right)^2 qN_2\right) \leq 2\exp\left(-\frac{1}{3}\left(\frac{\epsilon}{8}\right)^2 \frac{2^L}{2^n} \cdot N_2\right) = \frac{1}{8 \cdot 2^n \cdot n^L}$$

Otherwise, if we have $q < \frac{2^L}{2^n}$, then by Chernoff's bound:

$$\Pr\left[\left||\{z \in S : y \preceq z \preceq x\}| - qN_2\right| \geq \frac{\epsilon}{8} \cdot \frac{2^L}{2^n} \cdot N_2\right] \leq 2\exp\left(-\frac{1}{3}\left(\frac{\epsilon}{8} \cdot \frac{2^L}{2^n} \cdot \frac{1}{q}\right)^2 qN_2\right) \leq$$
$$2\exp\left(-\frac{1}{3}\left(\frac{\epsilon}{8}\right)^2 \frac{2^L}{2^n} \cdot N_2\right) = \frac{1}{8 \cdot 2^n \cdot n^L}$$

Now, by taking a union bound, it follows that with probability 7/8 for all such pairs (x, y) Equation 21 will be the case. For all x on which ϕ is defined it then will be the case that:

$$\left|\hat{\phi}(x) - \phi(x)\right| \leq \frac{\epsilon}{8} \cdot \max\left(\frac{1}{2^n}, \phi(x)\right)$$

Summing this for all x in the domain of ϕ we get:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \left|\hat{\phi}(x) - \phi(x)\right| \leq \frac{\epsilon}{8} \cdot \left(2^n \cdot \frac{1}{2^n} + \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \phi(x)\right) \leq$$
$$\frac{\epsilon}{8} \cdot \left(1 + \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \rho(x)\right) \leq \frac{\epsilon}{4}$$

◁

Now, we apply Corollary 10 to ρ. This gives a positive monotone function f that satisfies the three conditions of Corollary 10. We separate all the values of x in $\{0,1\}^n$ for which $L \leq ||x|| < h_0$ into two kinds: **good** and **bad**. Recall that by Corollary 10 an element x of $\{0,1\}^n$ for which $L \leq ||x|| < h_0$ can have $\text{slack}_f(x) \neq 0$ only if $||x|| = k_i$ for some i between 1 and t. We say that x is **bad** if there is some k_i for which $0 \leq ||x|| - k_i \leq L$. Otherwise, x if **good**.

We can write:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\phi(x) - f(x)| = \sum_{\text{good } x} |\phi(x) - f(x)| + \sum_{\text{bad } x} |\phi(x) - f(x)| \tag{22}$$

Now, we bound the two terms above separately. If x is good, then it is the case that for all z satisfying $||x|| - L \leq ||z|| \leq ||x||$ we have $\text{slack}_f(x) = 0$, and therefore $f(z) = \max_{z' \in \{0,1\}^n \text{ s. t. } z' \preceq z \text{ and } ||z|| - ||z'|| = 1} f(z')$. Using this relation recursively, we obtain that:

$$f(x) = \max_{z \in \{0,1\}^n \text{ s. t. } z \preceq x \text{ and } ||x|| - ||z|| = L} f(z)$$

Therefore, since f is monotone, we obtain that:

$$f(x) = \frac{1}{2^L} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x|| - ||y|| = L} \sum_{z \text{ s.t. } y \preceq z \preceq x} f(z)$$

By Corollary 10, it is the case $\rho(x) \geq f(x)$. This, together with the equation above and Equation 19 implies:

$$\phi(x) \geq f(x)$$

But we also know that $\rho(x) \geq \phi(x)$. Therefore:

$$\sum_{\text{good } x} |\phi(x) - f(x)| \leq \sum_{\text{good } x} |\rho(x) - f(x)| \leq \frac{\epsilon}{4} \tag{23}$$

Where the last inequality follows from Corollary 10.

Now, we bound the contribution of bad points.

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \sum_{\text{bad } x} \max(\phi(x), f(x)) \leq \sum_{\text{bad } x} \rho(x) = \sum_{k \in [L, h_0] \text{ s.t. for some } k_i\text{: } |k - k_i| \leq L} \mu_\rho(k) \cdot \binom{n}{k}$$

Now, by Claim 23 we have $\mu_\rho(k) \leq \frac{1}{2^n} \cdot \frac{4}{\epsilon}$ and by Fact 16 we have that $\binom{n}{k} \leq \frac{2}{\sqrt{n}} \cdot 2^n$. Combining these two facts with the inequality above we get:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \left(\frac{1}{2^n} \cdot \frac{4}{\epsilon}\right) \cdot \left(\frac{2}{\sqrt{n}} \cdot 2^n\right) \cdot (L \cdot t) = \frac{4}{\epsilon} \cdot \frac{2}{\sqrt{n}} \cdot L \cdot t$$

Substituting the value of L and the upper bound on t from Corollary 10 we get:

$$\sum_{\text{bad } x} |\phi(x) - f(x)| \leq \frac{4}{\epsilon} \cdot \frac{2}{\sqrt{n}} \cdot \frac{\epsilon^4 \sqrt{n}}{512} \cdot \frac{16}{\epsilon^2} = \frac{\epsilon}{4}$$

Combining this with Equations 22 and 23 we get:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\phi(x) - f(x)| \leq \frac{\epsilon}{2} \tag{24}$$

Overall, we have:

$$\left| \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\rho(x) - 1/2^n| - \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\hat{\phi}(x) - 1/2^n| \right| \leq$$
$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\hat{\phi}(x) - \rho(x)| \leq \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\hat{\phi}(x) - \phi(x)| +$$
$$\sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\phi(x) - f(x)| + \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |f(x) - \rho(x)|$$

This three terms can be bound using respectively Equation 20, Corollary 10 and Equation 24. This gives us:

$$\left| 2 \cdot \hat{d}_2 - \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} \left| \rho(x) - \frac{1}{2^n} \right| \right| =$$
$$\left| \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\rho(x) - 1/2^n| - \sum_{x \in \{0,1\}^n \text{ s.t. } L \leq ||x|| < h_0} |\hat{\phi}(x) - 1/2^n| \right| \leq \epsilon$$

Therefore, with probability at least 7/8 Equation 18 holds, which proves Claim 24 and completes the proof of correctness.

5 Estimating the support size

In this section we prove our upper-bound on the sample complexity of estimating the support size of an unknown monotone probability distribution over the Boolean cube. Recall that a probability distribution ρ is **well-behaved** if for every x either $\rho(x) = 0$ or $\rho(x) \geq 1/2^n$. We restate the theorem:

▶ **Theorem 4.** *For every positive ϵ, the following is true: for all sufficiently large n, there exists an algorithm, which given $\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$ samples from an unknown well-behaved monotone probability distribution ρ over $\{0,1\}^n$, can reliably[6] approximate the support size of ρ with an additive error of up to ϵ. The algorithm runs in time $O_\epsilon\left(\frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}\right)$*

Proof. The algorithm we use is listed as Algorithm 3.

Algorithm 3 Algorithm for the estimation of support size (given sample access from the distribution).

1. Set
$$M_1 = \frac{2^n}{2^{\frac{\epsilon^2}{64}\sqrt{n}}}\left(\ln\frac{32}{\epsilon}+1\right)$$
2. Take M_1 samples from the probability distributions. Call the set of these samples S_1.
3. Set
$$M_2 = \frac{32\ln 2}{\epsilon^2}$$
4. Pick M_2 elements of $\{0,1\}^n$ uniformly at random. Call these samples S_2.
5. We say that a point y is **covered** if in S_1 there exists at least one z, so that $z \preceq y$. One can check if a point y is covered by going through all the M_1 elements in S_1. Using this checking procedure, compute the fraction $\hat{\eta}$ of the elements in S_2 that are covered.
6. Output $\hat{\eta}$.

Clearly, the sample complexity is:

$$O\left(\frac{2^n}{2^{\frac{\epsilon^2}{64}\sqrt{n}}}\left(\ln\frac{32}{\epsilon}+1\right)\right) = \frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$$

In turn, the run-time is:

$$O\left(\frac{2^n}{2^{\frac{\epsilon^2}{64}\sqrt{n}}}\left(\ln\frac{32}{\epsilon}+1\right)\cdot\frac{32\ln 2}{\epsilon^2}\right) = \frac{2^n}{2^{\Theta_\epsilon(\sqrt{n})}}$$

Now, all is left to prove is correctness.

Let η denote the fraction of elements in $\{0,1\}^n$ that are covered by our samples in S_1. Then, a random element of $\{0,1\}^n$ is covered with probability η. Therefore, by the Hoeffding bound it follows that:

$$\Pr_{S_2}\left[|\hat{\eta}-\eta| > \frac{\epsilon}{4}\right] \leq 2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 M_2\right) = \frac{1}{8} \tag{25}$$

The last equality follows by substituting the value of M_2.

[6] By **reliably** we henceforth mean that the probability of success is at least 2/3.

Since ρ is monotone, it has to be the case that every point that is covered is in the support of ρ. Hence, the support size of ρ is at least $\eta \cdot 2^n$.

Now, all we need to show is that $\eta \cdot 2^n$ is not likely to be much smaller than the support size of ρ. We call a point x in the support of ρ **good** if there are at least $2^{\frac{\epsilon^2}{64}\sqrt{n}}$ points y each of which satisfying: (i) y belongs to the support of ρ. (ii) $x \preceq y$. If a point in the support of ρ is not good, then it is **bad**. We will show that the bad points are few, while a lot of the good points are likely to be covered.

Let f_{support} be defined as follows:

$$f_{\text{support}} \stackrel{\text{def}}{=} \begin{cases} 1 \text{ if } \rho(x) \neq 0 \\ 0 \text{ otherwise} \end{cases}$$

In other words, f_{support} is the indicator function of the support of ρ. Since ρ is a monotone probability distribution, f_{support} is a monotone function. Therefore, applying Lemma 8, there exists a function $g = g_1 \vee ... \vee g_t$ that $\epsilon/4$-approximates f_{support}, where $t \leq 8/\epsilon$ and each g_i is a monotone DNF with terms of width exactly k_i. Additionally, $g(x) \leq f(x)$ for all x in $\{0,1\}^n$

▷ Claim 26. For all i, g_i contains at most $\frac{\epsilon^2}{32} \cdot 2^n$ bad points.

Proof. Recall that g_i is k_i-regular, and therefore every point x on which $g_i(x) = 1$ needs to have Hamming weight $||x|| \geq k_i$.

▷ Claim 27. If x satisfies $g_i(x) = 1$ and $||x|| \geq k_i + \frac{\epsilon^2}{64}\sqrt{n}$, then x has to be good

Proof. Since g_i is a DNF and $g_i(x) = 1$ then x satisfies at least one of the terms of g_i. If there are more than one, arbitrarily pick one of them. Let this term be

$$t(y) = \bigwedge_{j \in H_1} y_j$$

The width of this AND has to be k_i, therefore $|H| = k_i$. Since $||x|| \geq k_i + \frac{\epsilon^2}{64}\sqrt{n}$, there must be $\frac{\epsilon^2}{64}\sqrt{n}$ values of j for which $x_j = 1$ but j is not in H_1. Denote the set of these values of j as H_2.

Now, consider an element $y \in \{0,1\}^n$ satisfying the criteria:

- For all j in H_1, $y_j = 1$.
- For all j neither in H_1 nor in H_2, $y_j = 0$.

Clearly, $t(y) = 1$, which implies $g_i(y) = 1$, $g(y) = 1$, and $f_{\text{support}}(y) = 1$. Also, for all j, we have $y_j \leq x_j$, and therefore $y \preceq x$. Finally, for all j in H_2 the value of y_j can be set arbitrarily to zero or one, and therefore there are $2^{|H_2|}$ such points, which is at least $2^{\frac{\epsilon^2}{64}\sqrt{n}}$. Therefore, x is a good point. ◁

Thus, we can upper-bound the number of bad points x on which $g_i(x) = 1$ by:

$$\left|\left\{x \in \{0,1\}^n : g(x) = 1 \text{ and } k_i + \frac{\epsilon^2}{64}\sqrt{n} > \sum_j x_j \geq k_i\right\}\right| \leq$$

$$\left|\left\{x \in \{0,1\}^n : k_i + \frac{\epsilon^2}{64}\sqrt{n} > \sum_j x_j \geq k_i\right\}\right| = \sum_{j=k_i}^{k_i + \frac{\epsilon^2}{64}\sqrt{n} - 1} \binom{n}{j} \leq \frac{\epsilon^2}{64}\sqrt{n}\binom{n}{n/2}$$

By Fact 16, for sufficiently large n, it is the case that $\binom{n}{n/2} \leq 2 \cdot \frac{2^n}{\sqrt{n}}$. This implies that the expression above is upper-bounded by $\frac{\epsilon^2}{32} \cdot 2^n$, which completes the proof of this claim. ◁

Since $g = g_1 \vee ... \vee g_t(x)$, our claim implies the following:

$$\left|\left\{x : g(x) = 1 \text{ and } x \text{ is bad}\right\}\right| \leq \sum_{i=1}^{t} \left|\left\{x : g_i(x) = 1 \text{ and } x \text{ is bad}\right\}\right| \leq$$
$$t \cdot \frac{\epsilon^2}{32} \cdot 2^n \leq \frac{8}{\epsilon} \cdot \frac{\epsilon^2}{32} \cdot 2^n = \frac{\epsilon}{4} \cdot 2^n$$

In addition, there could be at most $\frac{\epsilon}{4} \cdot 2^n$ bad points among the points on which f_{support} and g disagree. Thus, in total, there are at most $\frac{\epsilon}{2} \cdot 2^n$ bad points.

Finally, we need to argue that it is likely that many of the good points get covered:

▷ Claim 28. Suppose there are G good points. Then, with probability at least $7/8$ it will be the case that at least $1 - \epsilon/4$ fraction of these good points are covered.

Proof. For every good point x there exist least $2^{\frac{\epsilon^2}{64}\sqrt{n}}$ values of y for which i) $x \preceq y$ and ii) y is in the support of ρ. Since $x \preceq y$, if y is ever picked from the distribution, then x will be covered. Since y is in the support of ρ, and ρ is well-behaved, we have $\rho(y) \geq \frac{1}{2^n}$. Together, these imply that the probability that a random sample from ρ covers x is at least $\frac{2^{\frac{\epsilon^2}{64}\sqrt{n}}}{2^n}$. Hence, the probability that any of the M_1 i.i.d. samples taken from ρ does not cover x is at most:

$$\left(1 - \frac{2^{\frac{\epsilon^2}{64}\sqrt{n}}}{2^n}\right)^{M_1} = \left(1 - \frac{2^{\frac{\epsilon^2}{64}\sqrt{n}}}{2^n}\right)^{\frac{2^n}{2^{\frac{\epsilon^2}{64}\sqrt{n}}}\left(\ln\frac{32}{\epsilon}+1\right)} \leq \frac{1}{e^{\ln\frac{32}{\epsilon}}} = \frac{\epsilon}{32}$$

Let C denote a random variable, whose value equals to the number of the good points (out of total G) covered after taking M_1 i.i.d. samples from ρ.

The value of C has to satisfy these two constraints: (i) It has to be between 0 and G (ii) By linearity of expectation, $E[C] \geq (1 - \frac{\epsilon}{32})G$. Thus, to finish the proof of the Lemma, it is sufficient to show the following claim:

▷ Claim 29. If, for some fixed G, a random variable C is supported on $[0, G]$ and $E[C] \geq (1 - \frac{\epsilon}{32})G$, then $\Pr[C \geq (1 - \epsilon/4)G] \geq 7/8$.

Proof. This is immediate from Markov's inequality for the random variable $G - C$. ◁

◁

Now, we put it all together. Suppose that the bad events we previously identified do not happen. In particular, we know that with probability at least $7/8$ we have:

$$\left|\hat{\eta} - \eta\right| \leq \frac{\epsilon}{4}$$

Additionally, we also know that with probability at least $7/8$ it is the case that:

$$\left|\frac{\left|\left\{x : \substack{f_{\text{support}}(x)=1 \text{ and} \\ x \text{ is good and covered}}\right\}\right|}{2^n} - \frac{\left|\left\{x : \substack{f_{\text{support}}(x)=1 \text{ and} \\ x \text{ is good}}\right\}\right|}{2^n}\right| \leq$$
$$\frac{\epsilon}{4} \cdot \frac{|\{x : f_{\text{support}}(x) = 1 \text{ and } x \text{ is good}\}|}{2^n}$$

By union bound, the probability that none of this bad events happens is at least 3/4, which we will henceforth assume. Using the inequalities above together with the fact that the fraction of bad points is at most $\epsilon/2$ we get:

$$\left|\hat{\eta} - \frac{|\{x : f_{\text{support}}(x) = 1\}|}{2^n}\right| \leq \left|\hat{\eta} - \eta\right| + \left|\eta - \frac{|\{x : f_{\text{support}}(x) = 1\}|}{2^n}\right| =$$
$$\left|\hat{\eta} - \eta\right| + \left|\frac{|\{x : f_{\text{support}}(x) = 1 \text{ and } x \text{ is covered}\}|}{2^n} - \frac{|\{x : f_{\text{support}}(x) = 1\}|}{2^n}\right| \leq \left|\hat{\eta} - \eta\right| +$$
$$\left|\frac{\left|\left\{x : \begin{smallmatrix} f_{\text{support}}(x)=1 \text{ and} \\ x \text{ is good and covered} \end{smallmatrix}\right\}\right|}{2^n} - \frac{\left|\left\{x : \begin{smallmatrix} f_{\text{support}}(x)=1 \text{ and} \\ x \text{ is good} \end{smallmatrix}\right\}\right|}{2^n}\right| +$$
$$\frac{|\{x : f_{\text{support}}(x) = 1 \text{ and } x \text{ is bad}\}|}{2^n} \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} \cdot \frac{|\{x : f_{\text{support}}(x) = 1 \text{ and } x \text{ is good}\}|}{2^n} + \frac{\epsilon}{2} \leq$$
$$\frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{2} = \epsilon$$

This completes the proof of correctness. ◀

6 A lower bound on tolerant testing of uniformity

In this section we prove a sample complexity lower bound on the problem of tolerantly testing the uniformity of an unknown monotone probability distribution over $\{0,1\}^n$: the task of distinguishing a distribution that is $o(1)$-close to uniform from a distribution that is sufficiently far from uniform. Recall the theorem:

▶ **Theorem 6.** *For infinitely many positive integers n, there exist two probability distributions Δ_{Close} and Δ_{Far} over monotone distributions over $\{0,1\}^n$, satisfying:*

1. *Every distribution in Δ_{Far} is $1/2$-far from the uniform distribution.*
2. *Any algorithm that takes only $o\left(2^{\frac{n^{0.5-0.01}}{2}}\right)$ samples from a probability distribution, fails to reliably distinguish between Δ_{Close} and Δ_{Far}.*
3. *Every distribution in Δ_{Close} is $o(1)$-close to the uniform distribution.*

Proof. A basic building block of our construction is the following:

▶ **Definition 30.** *For a member of the Boolean cube x, the **subcube distribution** S_x is the probability distribution that picks y uniformly, subject to $y \succeq x$.*

All our distributions will be mixtures of such subcube distributions. For all the mixtures we will use, each subcube in the mixture is given the same weight. This method involving subcube distributions was used in [29] to prove property testing lower bounds for monotone probability distributions.

We construct Δ_{Close} to have only one member, which is equal to the uniform mixture of S_x for all $\binom{n}{n^{0.5-0.01}}$ values of x with Hamming weight $n^{0.5-0.01}$.

We define a random member of Δ_{Far} to be the uniform mixture of $\frac{1}{2}2^{n^{0.5-0.01}}$ subcube distributions S_{x_j}, where each of the x_j is picked randomly among all the members of the Boolean cube with Hamming weight $n^{0.5-0.01}$.

We show that any member of Δ_{Far} is sufficiently far from uniform by upper-bounding the size of its support (i.e. the number of elements that have non-zero probability). Each of the subcube distributions has a support size of $2^{n-n^{0.5-0.01}}$. The support size of a mixture of distributions is at most the sum of the supports sizes of the respective distributions.

Therefore, the support size of a member of Δ_{Far} is at most:

$$2^{n-n^{0.5-0.01}} \cdot \frac{1}{2} 2^{n^{0.5-0.01}} = \frac{1}{2} 2^n$$

This is sufficient to conclude that any member of Δ_{Far} is 1/2-far from uniform.

A random member D_1 of Δ_{Far} and the sole member D_2 of Δ_{Close} cannot be reliably distinguished using only $o\left(2^{\frac{n^{0.5-0.01}}{2}}\right)$ samples. This follows by the argument used in [29]: Because of the number of samples, with probability at least 0.99, the samples drawn from a random distribution from D_1 will all be from different subcube distributions. Also with probability at least 0.99, this will also be true for the sole distribution of D_2. If both of these things happen (which is the case with probability at least 0.98), the samples will be statistically indistinguishable. Thus, no tester can distinguish between D_1 and D_2 with an advantage greater than 0.02.

Finally, we need to prove that D_2 is $o(1)$-close to the uniform distribution. Here, the proof goes as follows. Both D_2 and the uniform distribution are symmetric with respect to a change of indices. This implies that the distance between these probability distributions equals to the distance between random variables R_2 and R_1, where R_1 is distributed as the Hamming weight of a random sample from D_2, whereas R_2 is distributed as the Hamming weight of uniformly random element of the Boolean cube. It is not hard to see that R_1 and R_2 are distributed according to binomial distributions with slightly different parameters. Now, the problem is equivalent to proving that the two following probability distributions are $o(1)$-close in total variation distance:

- A sum of n i.i.d. uniform random variables from $\{0,1\}$.
- A sum of $n - n^{0.5-0.01}$ i.i.d. uniform random variables from $\{0,1\}$.

It is convenient to first bound the variation distance between 1) the sum of k i.i.d. uniform random variables from $\{0,1\}$ and 2) $k+1$ i.i.d. uniform random variables from $\{0,1\}$, where k. We write the total variation distance as:

$$\frac{1}{2^k} + \sum_{i=1}^{k-1} \left| \frac{1}{2^k}\binom{k}{i} - \frac{1}{2^{k-1}}\binom{k-1}{i} \right| = \frac{1}{2^k}\left(1 + \sum_{i=1}^{k-1} \left| \binom{k-1}{i} - \binom{k-1}{i-1} \right| = \right) =$$

$$\frac{1}{2^k}\left(1 + \sum_{i=1}^{(k-1)/2} \left(\binom{k-1}{i} - \binom{k-1}{i-1} \right) + \sum_{i=(k-1)/2}^{k-1} \left(\binom{k-1}{i-1} - \binom{k-1}{i} \right)\right) =$$

$$\frac{1}{2^k}\left(1 + \binom{k-1}{(k-1)/2} - 1 + \binom{k-1}{(k-1)/2} - 1\right) = O\left(\frac{1}{\sqrt{k}}\right)$$

We telescoped the sums, and used the inequality that for all k, we have that $\binom{k}{k/2} \leq O\left(\frac{2^k}{\sqrt{k}}\right)$. For simplicity, we assumed above that $k-1$ is even, the odd case can be handled analogously. Thus, we have an upper bound of $O(1/\sqrt{k})$ on the total variation distance.

Using this, together with the triangle inequality for total variation distance, we bound the variation distance between 1) the sum of n i.i.d. uniform random variables from $\{0,1\}$ and 2) the sum of $n - n^{0.5-0.01}$ i.i.d. uniform random variables from $\{0,1\}$ by

$$O\left(\frac{n^{0.5-0.01}}{n^{0.5}}\right) = o(1).$$

This finishes the proof. ◀

References

1 Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal Testing for Properties of Distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3591–3599, 2015. URL: http://papers.nips.cc/paper/5839-optimal-testing-for-properties-of-distributions.

2 Michal Adamaszek, Artur Czumaj, and Christian Sohler. Testing Monotone Continuous Distributions on High-dimensional Real Cubes. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 56–65, 2010. doi:10.1137/1.9781611973075.6.

3 Maryam Aliakbarpour, Themis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Towards Testing Monotonicity of Distributions Over General Posets. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 34–82, 2019. URL: http://proceedings.mlr.press/v99/aliakbarpour19a.html.

4 Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2004.

5 Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *ICS*, pages 239–252, 2011.

6 Lucien Birgé et al. Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, 15(3):995–1012, 1987.

7 Eric Blais, Johan Håstad, Rocco A Servedio, and Li-Yang Tan. On DNF approximators for monotone boolean functions. In *International Colloquium on Automata, Languages, and Programming*, pages 235–246. Springer, 2014.

8 Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.

9 Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian Networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 370–448, 2017. URL: http://proceedings.mlr.press/v65/canonne17a.html.

10 Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. Society for Industrial and Applied Mathematics, 2013.

11 Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1193–1203, 2014. doi:10.1137/1.9781611973402.88.

12 Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-Modal Distributions via Testing. *Theory of Computing*, 10:535–570, 2014. doi:10.4086/toc.2014.v010a020.

13 Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning Poisson Binomial Distributions. *Algorithmica*, 72(1):316–357, 2015. doi:10.1007/s00453-015-9971-3.

14 Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-Modal Distributions: Optimal Algorithms via Reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1833–1852, 2013. doi:10.1137/1.9781611973105.131.

15 Constantinos Daskalakis and Gautam Kamath. Faster and Sample Near-Optimal Algorithms for Proper Learning Mixtures of Gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1183–1213, 2014.

16 Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the Structure, Covering, and Learning of Poisson Multinomial Distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1203–1217, 2015. doi:10.1109/FOCS.2015.77.

17 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-Based Testers are Optimal for Uniformity and Closeness. *Chicago J. Theor. Comput. Sci.*, 2019, 2019. URL: http://cjtcs.cs.uchicago.edu/articles/2019/1/contents.html.

18 Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. Society for Industrial and Applied Mathematics, 2015.

19 Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint*, 2016. arXiv:1606.03077.

20 Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. In *Conference on Learning Theory*, pages 831–849, 2016.

21 Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Conference on Learning Theory*, pages 850–878, 2016.

22 Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. *arXiv preprint*, 2018. arXiv:1802.08513.

23 Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning Mixtures of Gaussians in High Dimensions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 761–770, 2015. doi:10.1145/2746539.2746616.

24 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-SIGAI symposium on Principles of Database Systems*, pages 15–22. ACM, 2012.

25 Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 553–562, 2010. doi:10.1145/1806689.1806765.

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

27 Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

28 Oded Regev and Aravindan Vijayaraghavan. On Learning Mixtures of Well-Separated Gaussians. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 85–96, 2017. doi:10.1109/FOCS.2017.17.

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

30 Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694. ACM, 2011.

31 Gregory Valiant and Paul Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412. IEEE, 2011.

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

33 Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

34 Yihong Wu, Pengkun Yang, et al. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

7 Appendix A

7.1 Verifying the conditions on L_h

Recall that we defined A and L_h as follows:

- $A := \frac{1}{2n} \cdot e^{\frac{1}{2000} \cdot n^{1/5}}$
- For all $h \geq n/2$, we set $L_h := \max\left(\log\left(2nA \cdot \frac{\binom{n}{h}}{2^n}\right), 0\right)$
- For all h, satisfying $n/2 > h \geq 9\sqrt{n}$, we set: $L_h := L_{n/2} = \log\left(2nA \cdot \frac{\binom{n}{n/2}}{2^n}\right)$.

Here we prove that these values of A and L_h satisfy the following four conditions:

a) As a function of h, L_h is non-increasing.

b) For all h, we have that $L_h \leq 9\sqrt{n}$.

c)

$$\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} \leq \frac{1}{2}$$

d)

$$\sum_{h=9\sqrt{n}}^{n} L_h \cdot \left(\begin{cases} \frac{400}{n^{2.5}} & \text{if } h \leq n/2 - \sqrt{n\ln(n)} \\ \frac{40000}{n} & \text{if } n/2 - \sqrt{n\ln(n)} < h < n/2 + \sqrt{n} \\ 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) \leq \frac{\epsilon^2}{20000}$$

We will need the following standard fact can be proven, for example, by comparing $\sum_{i=0}^{N} i^k$ and $\int_{i=0}^{N} i^k \, di$:

▶ **Fact 31.** *For any positive constant k and for sufficiently large n, it is the case that:* $\sum_{i=0}^{n} i^k = (1 + o(1)) \, \frac{n^{k+1}}{k+1}$.

The truth of conditions (a) and (b) follows immediately by inspection. In fact a statement stronger than (b) is the case: for sufficiently large n we have $L_h \leq \log(n \cdot A) \leq 2 \cdot n^{1/5}$. Regarding condition (c), we have:

$$\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} =$$

$$\frac{A}{2^n} \left(\sum_{h=9\sqrt{n}}^{n/2} \binom{n}{h} \frac{1}{2nA} \cdot \frac{2^n}{\binom{n}{n/2}} + \sum_{h=n/2}^{n} \binom{n}{h} \cdot \min\left(\frac{1}{2nA} \cdot \frac{2^n}{\binom{n}{h}}, 1\right) \right) \leq \sum_{h=9\sqrt{n}}^{n} \frac{1}{2n} \leq \frac{1}{2}$$

Finally, recall that for all h, we have $L_h \leq 2 \cdot n^{1/5}$. For sufficiently large n, we have:

$$\sum_{h=9\sqrt{n}}^{n} L_h \cdot \left(\begin{cases} \frac{400}{n^{2.5}} & \text{if } h \leq n/2 - \sqrt{n\ln(n)} \\ \frac{40000}{n} & \text{if } n/2 - \sqrt{n\ln(n)} < h < n/2 + \sqrt{n} \\ 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 & \text{if } h \geq n/2 + \sqrt{n} \end{cases} \right) \leq$$

$$\sum_{h=9\sqrt{n}}^{n/2-\sqrt{n\ln n}} 2 \cdot n^{1/5} \cdot \frac{400}{n^{2.5}} + \sum_{h=n/2-\sqrt{n\ln n}}^{n+\sqrt{n}} 2 \cdot n^{1/5} \cdot \frac{40000}{n} + \sum_{h=n/2+\sqrt{n}}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot L_h \leq$$

$$\frac{\epsilon^2}{40000} + \sum_{h=n/2+\sqrt{n}}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot L_h \quad (26)$$

We now bound the last term using Hoeffding's inequality to bound the value of $\binom{n}{h}$, making a change of variables with $i := \frac{n-h}{2}$ and then using Fact 31 to bound the resulting summation. Precisely, we have the following chain of inequalities (some of which are only true for sufficiently large n):

$$\sum_{h=n/2+\sqrt{n}}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot L_h \leq$$

$$\sum_{h=n/2}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot \max\left(\log\left(2nA \cdot \frac{\binom{n}{h}}{2^n}\right), 0\right) \leq$$

$$\sum_{h=n/2}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot \max\left(\log\left(2nA \cdot \exp\left(-2\frac{(h-n/2)^2}{n}\right)\right), 0\right) =$$

$$\sum_{i=0}^{\sqrt{\frac{n}{2}\ln(2nA)}} \frac{40000}{\ln 2} \cdot \left(\frac{i}{n}\right)^2 \left(\ln(2nA) - 2 \cdot \frac{i^2}{n}\right) =$$

$$\frac{40000}{\ln 2} \cdot \left((1+o(1)) \frac{\left(\sqrt{\frac{n}{2}\ln(2nA)}\right)^3}{3n^2} \ln(2nA) - (1+o(1))\, 2 \cdot \frac{\left(\sqrt{\frac{n}{2}\ln(2nA)}\right)^5}{5n^3}\right)$$

Finally, simplifying and substituting the value of A we get:

$$\sum_{h=n/2+\sqrt{n}}^{n} 40000 \cdot \left(\frac{h-n/2}{n}\right)^2 \cdot L_h \leq (1+o(1)) \cdot \frac{40000}{\ln 2} \cdot \frac{\sqrt{2}}{30\sqrt{n}} \left(\ln(2nA)\right)^{5/2} \leq \frac{\epsilon^2}{40000}$$

Condition (d) is verified by combining Equation 26 with the equation above.

7.2 Proof of Claim 21

Here we prove that for all sufficiently large n, for all h, satisfying $0 \leq h \leq n$, it is the case that:

$$\frac{\binom{n}{h}}{\sum_{j\geq h}^{n}\binom{n}{j}} \leq \left(\begin{cases} \frac{2}{n^2} & \text{if } h \leq n/2 - \sqrt{n\ln(n)} \\ \frac{200}{\sqrt{n}} & \text{if } n/2 - \sqrt{n\ln(n)} < h < n/2 + \sqrt{n} \\ 200 \cdot \frac{h-n/2}{n} & \text{if } h \geq n/2 + \sqrt{n} \end{cases}\right)$$

We first handle the case when $h \geq n/2 + \sqrt{n}$. If, furthermore, $h > 11n/20$, then it is sufficient to prove that $\frac{\binom{n}{h}}{\sum_{j\geq h}^{n}\binom{n}{j}} \leq 10$, which is trivially true. Thus, we now assume that $h \leq 11n/20$

It is the case that:

$$\frac{\binom{n}{k}}{\binom{n}{k-1}} = \frac{1 - \frac{k-n/2-1}{n/2}}{1 + \frac{k-n/2}{n/2}} \tag{27}$$

Therefore, we can write:

$$\frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}} = \sum_{j=h}^{n} \prod_{k=h+1}^{j} \frac{1 - \frac{k-n/2-1}{n/2}}{1 + \frac{k-n/2}{n/2}}$$

Since $n/2+\sqrt{n} \leq h \leq 11n/20$, for sufficiently large n we have that $h+\frac{1}{4}\cdot\frac{n}{h-n/2} \leq n$. Using this, we can truncate the sum above, and then lower-bound the result by the product of the smallest summand with the total number of summands, getting:

$$\frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}} \geq \sum_{j=h}^{h+\frac{1}{4}\cdot\frac{n}{h-n/2}} \prod_{k=h+1}^{j} \frac{1-\frac{k-n/2-1}{n/2}}{1+\frac{k-n/2}{n/2}} \geq \frac{1}{4}\cdot\frac{n}{h-n/2}\cdot \prod_{k=h+1}^{h+\frac{1}{4}\cdot\frac{n}{h-n/2}} \frac{1-\frac{k-n/2-1}{n/2}}{1+\frac{k-n/2}{n/2}}$$

Now, we analogously lower-bound the product by lower-bounding each of the factors, and then use the fact that since $h \geq n/2+\sqrt{n}$, it is the case that $\frac{n}{h-n/2} \leq h-n/2$. We get:

$$\frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}} \geq \frac{1}{4}\cdot\frac{n}{h-n/2}\cdot\left(\frac{1-\frac{h+\frac{1}{4}\cdot\frac{n}{h-n/2}-n/2-1}{n/2}}{1+\frac{h+\frac{1}{4}\cdot\frac{n}{h-n/2}-n/2}{n/2}}\right)^{\frac{1}{4}\cdot\frac{n}{h-n/2}} \geq$$

$$\frac{1}{4}\cdot\frac{n}{h-n/2}\cdot\left(\frac{1-1.25\frac{h-n/2}{n/2}}{1+1.25\frac{h-n/2}{n/2}}\right)^{\frac{1}{4}\cdot\frac{n}{h-n/2}}$$

Finally, we use the fact that for all w between zero and one we have that $\frac{1}{1+w} = 1-w+w^2-... \geq 1-w$. We get:

$$\frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}} \geq \frac{1}{4}\cdot\frac{n}{h-n/2}\cdot\left(1-1.25\frac{h-n/2}{n/2}\right)^{\frac{1}{2}\cdot\frac{2n}{h-n/2}}$$

Now, recall that for any value w between zero and one, we have that $\ln(1-w) = -\sum_{i=1}^{\infty}\frac{w^i}{i} \geq -\sum_{i=1}^{\infty} w^i = -\frac{w}{1-w}$. Using this, and recalling that $h \leq 11n/20$, we get that:

$$\ln\left(1-1.25\frac{h-n/2}{n/2}\right) \geq -\frac{1.25\frac{h-n/2}{n/2}}{1-1.25\frac{h-n/2}{n/2}} \geq -\frac{1.25\frac{h-n/2}{n/2}}{1-1.25\frac{11n/20-n/2}{n/2}} = -\frac{20}{7}\frac{h-n/2}{n}$$

Combining the two previous equations together we get:

$$\frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}} \geq \frac{1}{4}\cdot\frac{n}{h-n/2}\cdot\exp\left(-\frac{20}{7}\frac{h-n/2}{n}\cdot\frac{1}{2}\cdot\frac{n}{h-n/2}\right) \geq \frac{1}{200}\cdot\frac{n}{h-n/2}$$

This completes the proof in the case $h \geq n/2+\sqrt{n}$.

Given our bound in the range $h \geq n/2+\sqrt{n}$, to show the desired bound in the range $n/2-\sqrt{n\ln(n)} < h < n/2+\sqrt{n}$ it is sufficient to show that $\frac{\binom{n}{h}}{\sum_{j\geq h}^{n}\binom{n}{j}}$ is non-decreasing, as a function of h. If $h < n/2$, this follows immediately, because, as a function of h, the numerator is non-decreasing, whereas the denominator is decreasing. If $h \geq n/2$, then using Equation 27, we get:

$$\frac{\sum_{j\geq h+1}^{n}\binom{n}{j}}{\binom{n}{h+1}} = \frac{1+\frac{h+1-n/2}{n/2}}{1-\frac{h-n/2}{n/2}}\cdot\frac{1}{\binom{n}{h}}\cdot\sum_{j\geq h+1}^{n}\frac{1-\frac{j-n/2-1}{n/2}}{1+\frac{j-n/2}{n/2}}\binom{n}{j-1} \leq$$

$$\frac{1}{\binom{n}{h}}\cdot\sum_{j\geq h+1}^{n}\binom{n}{j-1} \leq \frac{1}{\binom{n}{h}}\cdot\sum_{j\geq h+1}^{n+1}\binom{n}{j-1} = \frac{\sum_{j\geq h}^{n}\binom{n}{j}}{\binom{n}{h}}$$

Which implies that $\frac{\binom{n}{h+1}}{\sum_{j\geq h+1}^{n}\binom{n}{j}} \geq \frac{\binom{n}{h}}{\sum_{j\geq h}^{n}\binom{n}{j}}$

Finally, for the range $h \leq n/2-\sqrt{n\ln(n)}$ we can use Hoeffding's bound:

$$\frac{\binom{n}{h}}{\sum_{j\geq h}^{n}\binom{n}{j}} \leq \frac{\binom{n}{h}}{\frac{1}{2}\cdot 2^n} \leq 2\cdot\Pr_{x\sim\{0,1\}^n}\left[x \leq n/2-\sqrt{n\ln(n)}\right] \leq 2\cdot\exp\left(-2\ln n\right) = \frac{2}{n^2}$$

7.3 Proof fo Claim 22

Recall that:

$$N \stackrel{\text{def}}{=} \frac{2^n}{A} \cdot \frac{192}{\epsilon^2} \cdot (n + 9\sqrt{n} + 4)$$

Our algorithm for learning a monotone probability distribution drew N samples from the probability distribution ρ and the resulting multiset of samples was denoted as S. For all x in $\{0,1\}^n$, if $||x|| < 9\sqrt{n}$, we set $\hat{\phi}(x) = 0$, otherwise we set:

$$\hat{\phi}(x) := \frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \frac{\max_{y \text{ s.t. } y \preceq x \text{ and } ||y||-||x|| = \lfloor L_{||x||} \rfloor} \left| \left\{ z \in S : y \preceq z \preceq x \right\} \right|}{N}$$

Where L_h is a specific value associated to each value of h. We also defined for all x with $x \geq 9\sqrt{n}$ the value:

$$\phi(x) \stackrel{\text{def}}{=} \frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y|| = \lfloor L_{||x||} \rfloor} \Pr_{z \sim \rho}[y \preceq z \preceq x] =$$
$$\frac{1}{2^{\lfloor L_{||x||} \rfloor}} \cdot \max_{y \text{ s.t. } y \preceq x \text{ and } ||x||-||y|| = \lfloor L_{||x||} \rfloor} \sum_{z \text{ s.t. } y \preceq z \preceq x} \rho(z)$$

Here we prove that if it is the case that $\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} \leq \frac{1}{2}$, then, with probability at least $7/8$, it is the case that:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \hat{\phi}(x) - \phi(x) \right| \leq \frac{\epsilon}{4}$$

We claim that for any pair (x, y), such that ϕ is defined on x and $||x|| - ||y|| = \lfloor L_{||x||} \rfloor$, with probability at least $1 - \frac{1}{8 \cdot 2^n \cdot n^{9\sqrt{n}}}$ the following holds:

$$\left| \Pr_{z \sim \rho}[y \preceq z \preceq x] - \frac{|\{z \in S : y \preceq z \preceq x\}|}{N} \right| \leq \frac{\epsilon}{8} \cdot \max\left(\frac{A}{2^n}, \Pr_{z \sim \rho}[y \preceq z \preceq x] \right) \tag{28}$$

We use Chernoff's bound to prove this as follows. Denote by q the value $\Pr_{z \sim \rho}[y \preceq z \preceq x]$. If $q \geq \frac{A}{2^n}$ then by Chernoff's bound we have:

$$\Pr\left[\left| |\{z \in S : y \preceq z \preceq x\}| - qN \right| \geq \frac{\epsilon}{8} qN \right] \leq$$
$$2 \exp\left(-\frac{1}{3} \left(\frac{\epsilon}{8} \right)^2 qN \right) \leq 2 \exp\left(-\frac{1}{3} \left(\frac{\epsilon}{8} \right)^2 \frac{A}{2^n} \cdot N \right) = \frac{1}{8 \cdot 2^n \cdot n^{9\sqrt{n}}}$$

Otherwise, if we have $q < \frac{A}{2^n}$, then by Chernoff's bound:

$$\Pr\left[\left| |\{z \in S : y \preceq z \preceq x\}| - qN \right| \geq \frac{\epsilon}{8} \cdot \frac{A}{2^n} \cdot N \right] \leq 2 \exp\left(-\frac{1}{3} \left(\frac{\epsilon}{8} \cdot \frac{A}{2^n} \cdot \frac{1}{q} \right)^2 qN \right) \leq$$
$$2 \exp\left(-\frac{1}{3} \left(\frac{\epsilon}{8} \right)^2 \frac{A}{2^n} \cdot N \right) = \frac{1}{8 \cdot 2^n \cdot n^{9\sqrt{n}}}$$

Now, by taking a union bound, it follows that with probability $7/8$ for all such pairs (x, y) Equation 28 will be the case. Recalling the definition of ϕ, for all x on which ϕ is defined it then will be the case that:

$$\left| \hat{\phi}(x) - \phi(x) \right| \leq \frac{\epsilon}{8} \cdot \max\left(\frac{1}{2^{\lfloor L_{||x||} \rfloor}} \frac{A}{2^n}, \phi(x) \right)$$

Now, we sum this for all x in the domain of ϕ and use the fact that $\lfloor L_h \rfloor \geq L_h + 1$, and then use that $\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} \leq \frac{1}{2}$. We get:

$$\sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \left| \hat{\phi}(x) - \phi(x) \right| \leq$$

$$\frac{\epsilon}{8} \cdot \left(\frac{1}{2^n} \cdot \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \frac{A}{2^{\lfloor L_{||x||} \rfloor}} + \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \phi(x) \right) =$$

$$\frac{\epsilon}{8} \cdot \left(\frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{\lfloor L_h \rfloor}} + \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \phi(x) \right) \leq$$

$$\frac{\epsilon}{8} \cdot \left(2 \cdot \frac{1}{2^n} \cdot \sum_{h=9\sqrt{n}}^{n} \binom{n}{h} \cdot \frac{A}{2^{L_h}} + \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \phi(x) \right) \leq$$

$$\frac{\epsilon}{8} \cdot \left(1 + \sum_{x \in \{0,1\}^n \text{ s.t. } 9\sqrt{n} \leq ||x||} \rho(x) \right) \leq \frac{\epsilon}{4}$$