JMLR: Workshop and Conference Proceedings vol 49:1–28, 2016

Learning and Testing Junta Distributions

Maryam Aliakbarpour * MARYAMA@MIT.EDU
CSAIL, MIT, Cambridge MA 02139

Eric Blais† ERIC.BLAIS@UWATERLOO.CA
David Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada

Ronitt Rubinfeld ‡ RONITT@CSAIL.MIT.EDU
CSAIL, MIT, Cambridge MA 02139 and the Blavatnik School of Computer Science, Tel Aviv University

Abstract

We consider the problem of learning distributions in the presence of irrelevant features. This problem is formalized by introducing a new notion of *k-junta distributions*. Informally, a distribution $\mathcal{D}$ over the domain $\mathcal{X}^n$ is a *k-junta distribution* with respect to another distribution $\mathcal{U}$ over the same domain if there is a set $J \subseteq [n]$ of size $|J| \leq k$ that captures the difference between $\mathcal{D}$ and $\mathcal{U}$.

We show that it is possible to learn k-junta distributions with respect to the uniform distribution over the Boolean hypercube $\{0,1\}^n$ in time $\text{poly}(n^k, 1/\epsilon)$. This result is obtained via a new Fourier-based learning algorithm inspired by the Low-Degree Algorithm of Linial, Mansour, and Nisan (1993).

We also consider the problem of testing whether an unknown distribution is a k-junta distribution with respect to the uniform distribution. We give a nearly-optimal algorithm for this task. Both the analysis of the algorithm and the lower bound showing its optimality are obtained by establishing connections between the problem of testing junta distributions and testing uniformity of weighted collections of distributions.

Keywords: Learning distributions, Property testing, Juntas, Fourier analysis

1. Introduction

A central challenge in machine learning is learning target concepts in the presence of irrelevant features. Due to its importance and ubiquity, this challenge has inspired much research in the statistics and machine learning communities over the last decades (see, e.g., Guyon and Elisseeff (2003), Liu and Motoda (2012), and Chandrashekar and Sahin (2014)). This challenge also lends itself to an elegant formalization using the notion of juntas. The function $f : \mathcal{X}^n \to \mathcal{Y}$ is a *k-junta* for $k \leq n$ if there is a set $J = \{j_1, \ldots, j_k\} \subseteq [n]$ such that for every $x \in \mathcal{X}^n$, the value $f(x)$ is determined by the k values $x_{j_1}, \ldots, x_{j_k}$. When this is the case, the *relevant* features of f are contained in J and the variables outside of J are *irrelevant*. As Blum (1994) and Blum and Langley (1997) originally proposed, learning with irrelevant features in the PAC model corresponds to the problem of designing a

* Research supported by NSF grants CCF-1420692 and CCF-1065125
† Research supported by an NSERC Discovery Grant
‡ Research supported by NSF grants CCF-1420692, CCF-1065125, and ISF grant 1536/14

polynomial-time algorithm for learning k-junta functions from samples $\big(x, f(x)\big)$ where x is drawn from some fixed distribution over $\mathcal{X}^n$.

In this work, we consider the problem of learning with irrelevant features in a different setting: where the target concept is a *distribution* over a high-dimensional domain $\mathcal{X}^n$ and the learning algorithm observes samples $x \in \mathcal{X}^n$ drawn from the target distribution. The problem of learning distributions has a long and rich history—see Diakonikolas (2016) for a great introduction to the topic—yet we are not aware of any result that directly addresses the problem of learning distributions in the presence of irrelevant features in its full generality.

To formalize this problem, we must first specify the notion of distributions that only have a few relevant features. We do so with a definition that is analogous to that of junta functions. For this definition, we need to first introduce some notation: given a distribution $\mathcal{D}$ over a high-dimensional domain $\mathcal{X}^n$, a set $J \subseteq [n]$, and a vector x in the support of $\mathcal{D}$, let $\mathcal{D}_{J \leftarrow x}$ be the conditional distribution of a random variable drawn from $\mathcal{D}$ conditioned on the event that $X_j = x_j$ for every $j \in J$.

Definition 1 (Junta distributions) *Fix any distribution $\mathcal{U}$ over $\mathcal{X}^n$.*[1] *A distribution $\mathcal{D}$ over $\mathcal{X}^n$ with support $\mathrm{supp}(\mathcal{D}) \subseteq \mathrm{supp}(\mathcal{U})$ is a* k-junta distribution with respect to $\mathcal{U}$ *if there is a set $J \subseteq [n]$ of $|J| \leq k$ coordinates such that for every x in the support of $\mathcal{D}$, the distributions $\mathcal{D}_{J \leftarrow x}$ and $\mathcal{U}_{J \leftarrow x}$ are identical.*

The definition is perhaps best illustrated with an example. Let $\mathcal{U}$ be the uniform distribution on $\{0,1\}^n$ and let $\mathcal{D}$ be the uniform distribution on the set of strings $x \in \{0,1\}^n$ that satisfy $x_1 \oplus \cdots \oplus x_k = 1$. The distribution $\mathcal{D}$ is a k-junta with respect to $\mathcal{U}$, but it is far from being a $(k-1)$-junta with respect to $\mathcal{U}$.

We consider the problem of learning the class of distributions that are k-juntas with respect to the uniform distribution over the Boolean hypercube $\{0,1\}^n$. As we will see shortly, even this idealized setting captures much of the rich structure of the problem of learning junta distributions. Furthermore, this setting also reveals some important connections to the analysis of Boolean functions. We also consider the complementary problem of *testing* if a distribution is a k-junta distribution with respect to the uniform distribution over $\{0,1\}^n$. We describe our results in more details below.

1.1. Learning junta distributions

We consider the problem of learning distributions in the model introduced by Kearns et al. (1994). A *class* of distributions is a set of distributions, and the total variation distance between two distributions $\mathcal{D}$ and $\mathcal{D}'$ with probability density functions $p, p' : \mathcal{X}^n \to [0,1]$ is $\mathrm{d_{TV}}(\mathcal{D}, \mathcal{D}') = \sum_{x \in \mathcal{X}^n} |p(x) - p'(x)|/2$. An *$\epsilon$-learner* for a class $\mathcal{C}$ of distributions is an algorithm that draws i.i.d. samples from an unknown distribution $\mathcal{D}$ in the class $\mathcal{C}$ and outputs a hypothesis distribution $\widetilde{\mathcal{D}}$ that satisfies $\mathrm{d_{TV}}(\mathcal{D}, \widetilde{\mathcal{D}}) \leq \epsilon$ with probability at least $\frac{2}{3}$.

The easiest way to establish an upper bound on the time complexity of the problem of learning k-junta distributions is via the cover method (Devroye and Lugosi (2001);

1. The notation $\mathcal{U}$ is chosen to reflect that $\mathcal{U}$ is the *universal*, or *underlying* distribution.

Daskalakis et al. (2014); Diakonikolas (2016)), which can be used to show that there is an ϵ-learner for k-juntas with respect to the uniform distribution over $\{0,1\}^n$ with time complexity $\tilde{O}(\binom{n}{k}2^{k2^k/\epsilon})$.[2] For completeness, we include the details of this construction in Appendix A. Note, however, that the running time of this algorithm is *doubly-exponential* in k and also *exponential* in $1/\epsilon$.

Our first main result shows that we can learn k-junta distributions with respect to the uniform distribution much more efficiently: it is possible to learn this class of distributions with a time complexity that is only *singly-exponential* in k and *polynomial* in $1/\epsilon$.

Theorem 2 *Fix $\epsilon > 0$ and $1 \leq k \leq n$. There is an ϵ-learner for k-junta distributions with respect to the uniform distribution over $\{0,1\}^n$ with sample complexity $O(2^{2k}k\log n/\epsilon^4)$ and running time $\tilde{O}(\min(n^k, 2^n)\cdot 2^{2k}\cdot k/\epsilon^4)$.*

The starting point for the proof of Theorem 2 is the observation that a distribution with probability mass function (pmf) $p : \{0,1\}^n \to [0,1]$ is a k-junta with respect to the uniform distribution if and only if p is a k-junta function. This characterization naturally suggests the use of the Low-Degree Algorithm (LDA) of Linial et al. (1993) as a promising approach for learning p. Indeed, LDA can be implemented even in the distribution learning setting: by drawing samples from the distribution, we can estimate the Fourier coefficients $\hat{p}(S)$ of the function p for every set of size $|S| \leq k$ (and, thus, of all the non-zero Fourier coefficients of p).

The Low-Degree Algorithm generates a hypothesis function p' whose Fourier coefficients are determined by the estimates of the corresponding coefficients of p. This approach, however, cannot be used in the context of learning distributions for two reasons. The first is that the accuracy guarantee of LDA is in terms of a bound on the L_2 norm $\sum(p'(x)-p(x))^2$. To obtain a valid distribution learning algorithm, however, we must bound the L_1 norm $\sum|p'(x)-p(x)|$ between the hypothesis and target distributions' pmfs, which is a stronger requirement (see Kalai et al. (2008)). The second is that in any case, for even the L_2 guarantee to be sufficiently strong, the estimates to the Fourier coefficients $\hat{p}(S)$ need a level of accuracy that can only be achieved with a prohibitively large number of samples.

To bypass both of these issues, we use the estimated Fourier coefficients in a completely different way. We compute a score for each candidate junta $J \subseteq [n]$ of size $|J| = k$ corresponding to the estimated total Fourier mass of the subsets of J. If our estimates were exact, the candidate with the largest score would be the correct junta; the bulk of our analysis lies in showing that this characterization is robust, in that for every junta candidate J' whose score is close to that of the actual junta, p is close to a junta on J'.

Our upper bounds on the sample complexity of the problem of learning k-juntas have a logarithmic dependence on n and an exponential complexity in k. We show that both of these dependencies are necessary. We also show that any significant improvement on the running time of our algorithm would also yield a corresponding improvement in the time complexity of the problem of learning k-junta Boolean functions.

Theorem 3 *Fix $\epsilon > 0$ and $n > k \geq 1$. Any algorithm that learns k-junta distributions with respect to the uniform distribution over $\{0,1\}^n$ has sample complexity $\Omega(2^k/\epsilon^2 + k\log(n)/\epsilon)$.*

2. Here and throughout the paper, the $\tilde{O}(\cdot)$ notation is used to hide polylogarithmic factors in the argument.

Furthermore, if there is an algorithm for this task with running time t, the time complexity of the problem of learning k-junta Boolean functions in the PAC model is bounded above by $O(2^k t)$.

In particular, a learning algorithm for k-junta distributions with respect to the uniform distribution with running time $\mathrm{poly}(n, 2^k, 1/\epsilon)$ would provide a solution to the celebrated *learning juntas* problem (see Blum (1994), Blum and Langley (1997), and Mossel et al. (2004)). In fact, even the more modest improvement of our algorithm's running time to $O(n^{ck})$ for any constant $c < \omega/4 < 0.6$, where ω is the the matrix multiplication exponent, would yield an improvement on the best current upper bound of Valiant (2012) on the time complexity for learning k-junta functions.

1.2. Testing junta distributions

We next turn our attention to the problem of testing whether an unknown distribution is a k-junta distribution with respect to the uniform distribution in the property testing framework introduced by Batu et al. (2000). A *property* of distributions is simply a class (or a set) of distributions. A distribution $\mathcal{D}$ is *ϵ-far* from having property P when $\mathrm{d_{TV}}(\mathcal{D}, P) := \min_{\mathcal{D}' \in P} \mathrm{d_{TV}}(\mathcal{D}, \mathcal{D}') \geq \epsilon$; otherwise, $\mathcal{D}$ is *ϵ-close* to P. An *ϵ-tester* for property P is a randomized algorithm with bounded error that distinguishes between distributions with property P from those that are ϵ-far from having the same property. We show that it is possible to test k-juntas with respect to the uniform distribution over $\{0,1\}^n$ with a number of samples that is sublinear in the size $N = 2^n$ of the domain of the distribution.

Theorem 4 *Fix $\epsilon > 0$ and $n > k \geq 1$. There is an ϵ-tester for k-juntas with respect to the uniform distribution over $\{0,1\}^n$ with sample complexity $O(2^{n/2} k \log n/\epsilon^2) = \tilde{O}(\sqrt{N}/\epsilon^2)$.*

The proof of Theorem 4 is obtained by reducing the problem of testing juntas to the problem of testing uniformity of weighted collections of distributions, a framework originally introduced by Levi et al. (2013). A *(weighted) collection of distributions* is defined to be a set of m distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ on a common domain $\mathcal{X}$ of size N and a set of m weights $w_1, \ldots, w_m \in [0,1]$ such that $\sum_{i=1}^m w_i = 1$. We denote such a collection by $\{\mathcal{D}_i | w_i\}_{i=1}^m$. When we draw a sample from $\{\mathcal{D}_i | w_i\}_{i=1}^m$, we obtain a pair $(\mathcal{D}_i, j)$ such that $\mathcal{D}_i$ is picked with probability w_i and then j is a sample drawn from $\mathcal{D}_i$.

An *ϵ-tester* of collections of distributions for a property P is a randomized algorithm with bounded error that distinguishes collections $\{\mathcal{D}_i | w_i\}_{i=1}^m$ where each $\mathcal{D}_i$ has property P from those that satisfy $\sum_{i=1}^m w_i \cdot \mathrm{d_{TV}}(\mathcal{D}_i, P) \geq \epsilon$. The key technical result in the proof of Theorem 4 is that we can test the uniformity of collections of m distributions on a domain of size N with roughly $\sqrt{mN}$ samples. In our reduction, the original junta-testing problem corresponds to the problem of testing the uniformity of a collection of $m = 2^k$ distributions on a domain of size $N = 2^{n-k}$, yielding the sample complexity bound in Theorem 4.

When $k \ll n$, the sample complexity of the junta testing algorithm is much larger than (i.e., doubly-exponential in) the sample complexity of the learning algorithm. We show that this gap is unavoidable and, in fact, that the bound in Theorem 4 is nearly optimal.

Theorem 5 *Fix $0 \leq k < n$ and $0 < \epsilon < 1$. Every ϵ-tester for k-juntas with respect to the uniform distribution on $\{0,1\}^n$ has sample complexity $\Omega(2^{n/2}/\epsilon^2) = \Omega(\sqrt{N}/\epsilon^2)$.*

The lower bound again uses the connection to the problem of testing collections of distributions. In this case, this is done by constructing a distribution over k-junta distributions and distributions that are far from k-juntas such that any algorithm that distinguishes between the two with sample complexity $o(2^{n/2}/\epsilon^2)$ would also be able to test collections of distributions for uniformity with a number of samples that violates a lower bound of Levi et al. (2013).

1.3. Future directions

The current work initiates the study of learning and testing distributions in the presence of irrelevant features. There are a several directions in which further study of this topic will likely yield valuable results. The most immediate next step is to extend our results on learning and testing k-junta distributions with respect to a wider class of underlying distributions $\mathcal{U}$ and over non-Boolean domains. Similar extensions have been achieved in the functional setting with appropriate generalizations of the Fourier analysis tools used in the Boolean domain, and it remains an open problem to determine whether similar generalizations can be used to extend the analysis of our current algorithms (or natural extensions of them) to other settings or whether entirely new testing and learning approaches are required.

Another particularly intriguing question is whether our time and sample complexity lower bounds can be bypassed by considering stronger sampling models or additional structural restrictions on the distributions themselves. For example, recent results have shown that distribution property testing problems can have dramatically smaller sample complexity in the *conditional sampling* model introduced by Canonne et al. (2015) and Chakraborty et al. (2013). Is this also the case for the problem of testing junta distributions?

1.4. Organization

We present the algorithm for learning junta distributions and complete the proof of Theorem 2 in Section 2. The algorithm for testing junta distributions and the proof of Theorem 4 are presented in Section 3. Finally, we complete the proofs of Theorems 3 and 5 establishing the sample and time complexity lower bounds for learning and testing junta distributions in Section 4.

2. Learning junta distributions

In this section, we complete the proof of Theorem 2 by introducing and analyzing an algorithm for ϵ-learning k-junta distributions with respect to the uniform distribution over $\{0,1\}^n$.

For an input $x \in \{0,1\}^n$ and a set $J \subseteq [n]$ of coordinates, we write $x^{(J)} \in \{0,1\}^{|J|}$ to denote the restriction of x to the coordinates in J. (For example, $101000^{(\{1,3\})} = 11$.) A distribution $\mathcal{P}$ with pmf $p : \{0,1\}^n \to [0,1]$ is a k-junta distribution with respect to the uniform distribution over $\{0,1\}^n$ iff there is a set $J^* \subseteq [n]$ of size $|J^*| = k$ and a set of 2^k parameters $\{a_y\}_{y \in \{0,1\}^k}$ such that for every $x \in \{0,1\}^n$, $p(x) = a_{x^{(J^*)}}/2^{n-k}$. Our learning algorithm describes its hypothesis by specifying the corresponding set J^* and parameters

Algorithm 1 An ϵ-learner for junta distributions

$x_1, x_2, \ldots, x_s \leftarrow$ draw $s = O(\min(k \log n, n) \cdot 2^{2k}/\epsilon^4)$ samples from $\mathcal{P}$.
for all subsets $J \subseteq [n]$ of size $|J| = k$ **do**
  Initialize $\tilde{f}(J) \leftarrow 0$
  for all $S \subseteq J$ where $S \neq \emptyset$ **do**
    $t \leftarrow |\{i \in [s] : \chi_S(x_i) = 1\}|$.
    $\tilde{f}(J) \leftarrow \tilde{f}(J) + \left(\frac{2t}{s} - 1\right)^2$
  end for
end for
Output $J^\dagger = \text{argmax}_J \tilde{f}(J)$ (breaking ties arbitrarily).
for all $y \in \{0,1\}^k$ **do**
  $\tilde{a}_y \leftarrow \frac{1}{s} \cdot |\{i \in [s] : x_i^{(J)} = y\}|$
  Output $\tilde{a}_y$.
end for

a_y of its hypothesis distribution. (The algorithm is a *proper learning* algorithm, in that it always outputs a distribution that is a k-junta with respect to the uniform distribution.)

We turn to Fourier analysis over the Boolean hypercube to solve the learning problem. The *Fourier basis function* corresponding to a set $S \subseteq [n]$ is the function $\chi_S : \{0,1\}^n \to \{-1,1\}$ defined by $\chi_S(x) = (-1)^{\sum_{i \in S} x_i} = (-1)^{\oplus_{i \in S} x_i}$ when S is not empty and $\chi_\emptyset(x) = 1$. The pmf $p : \{0,1\}^n \to [0,1]$ of a distribution $\mathcal{P}$ can be expressed in terms of the Fourier basis functions as $p(x) = \sum_{S \subseteq [n]} \widehat{p}(S) \cdot \chi_S(x)$ where the *Fourier coefficient* of p corresponding to S is $\widehat{p}(S) = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} p(x) \cdot \chi_S(x) = \mathrm{E}[p(x) \cdot \chi_S(x)]$. Here and throughout this section, we write $\mathrm{E}[f(x)]$ to refer to the expected value of $f(x)$ when x is picked uniformly at random and $\mathrm{E}_{x \sim \mathcal{P}}[f(x)]$ to refer to the expected value of $f(x)$ when x is drawn from $\mathcal{P}$. In the following, we will also abuse notation slightly and use the same symbol $\mathcal{P}$ to denote a distribution and its pmf $\mathcal{P} : \{0,1\}^n \to [0,1]$. For a complete introduction to Fourier analysis over the Boolean hypercube, we highly recommend the book of O'Donnell (2014).

The proof of Theorem 2 is obtained via the following result.

Theorem 6 *Algorithm 1 is an ϵ-learner for k-junta distributions with respect to the uniform distribution over $\{0,1\}^n$ with sample complexity $s = O(\min(k \log n, n) \cdot 2^{2k}/\epsilon^4)$ and running time $O(n^k \cdot 2^{3k} \cdot k \cdot \min(k \log n, n)/\epsilon^4)$. With minor changes in implementation, the running time can be improved to $O(\min(2^n, n^k) \cdot \min(k \log n, n) \cdot 2^{2k}/\epsilon^4)$.*

Proof To prove the theorem, we show that the algorithm outputs a distribution $\widetilde{\mathcal{P}}_{J^\dagger}$ that is ϵ-close to the target distribution $\mathcal{P}$ with large constant probability. To do this, we define an intermediate distribution $\mathcal{P}_{J^\dagger}$ and show that $\mathcal{P}$ and $\widetilde{\mathcal{P}}_J$ are both close to $\mathcal{P}_{J^\dagger}$. More precisely, for any set $J \subseteq [n]$ of size k, we let $\mathcal{P}_J$ be the distribution with the pmf defined by

$$\mathcal{P}_J(x) := \Pr_{y \sim \mathcal{P}} \left[y^{(J)} = x^{(J)} \right] / 2^{n-k}. \tag{1}$$

This construction guarantees that $\mathcal{P}_J$ is a junta distribution on the set J. In particular, letting J^* be a set containing the relevant features of the target distribution $\mathcal{P}$, we observe

that $\mathcal{P}_{J^*}$ is identical to $\mathcal{P}$. We complete the analysis of correctness of Algorithm 1 by showing that with large constant probability, $\mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_{J^\dagger}) \leq \frac{\epsilon}{2}$ and $\mathrm{d_{TV}}(\mathcal{P}_{J^\dagger}, \widetilde{\mathcal{P}}_{J^\dagger}) \leq \frac{\epsilon}{2}$ and applying the triangle inequality.

The first task, showing that $\mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_{J^\dagger}) \leq \frac{\epsilon}{2}$, is the more demanding one. We say that a set J is *invalid* if $\mathcal{P}_J$ is $\frac{\epsilon}{2}$-far from $\mathcal{P}$. We want to show that the probability that Algorithm 1 outputs an invalid set $J^\dagger$ is small. In order to do so, we define two functions f and h on the subsets of $[n]$ of size k by setting

$$h(J) = 2^{2n} \cdot \mathrm{E}[(\mathcal{P}_J(x) - 1/2^n)^2] \tag{2}$$

and

$$f(J) = 2^{2n} \cdot \sum_{S \subseteq J, S \neq \emptyset} \widehat{\mathcal{P}}(S)^2. \tag{3}$$

We bound the total variation distance between $\mathcal{P}$ and $\mathcal{P}_{J^\dagger}$ in three steps.

- **Step 1.** We show that for every set J, $h(J^*) - h(J)$ is at least $4 \cdot \mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_J)^2$.
- **Step 2.** We show that $f(J)$ is always equal to $h(J)$. Therefore, $f(J^*) - f(J) \geq \epsilon^2$ for any invalid set J.
- **Step 3.** We show that with large constant probability, for every set J of size k, $|\tilde{f}(J) - f(J)| < \epsilon^2/2$.

The three steps are completed by establishing the following three lemmas.

Lemma 7 (Step 1) *Let $\mathcal{P}$ be a k-junta distribution on the set J^* and $\mathcal{P}_J$ be a k-junta distributions defined in Equation 1. Then,*

$$\mathrm{E}[(\mathcal{P}(x) - 1/2^n)^2] - \mathrm{E}[(\mathcal{P}_J(x) - 1/2^n)^2] \geq 4 \cdot \mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_J)^2 / 2^{2n}. \tag{4}$$

The starting point for the proof of this lemma is the observation that for every set J, the distribution $\mathcal{P}_J$ is a junta distribution over the set $J \cap J^*$. Two other critical ingredients in the proof of the lemma are the identity

$$\mathcal{P}_J(x) = \left(\Pr_{y \sim \mathcal{P}} \left[y^{(J \cap J^*)} = x^{(J \cap J^*)} \right] \right) / 2^{n-k+|J \setminus J^*|} \tag{5}$$

and the additional observation that $\mathrm{E}[\mathcal{P}_J(x)(\mathcal{P}(x) - \mathcal{P}_J(x))] = 0$.

Lemma 8 (Step 2) *With f and h as defined in Equation (3) and Equation (2), for any $J \subset [n]$ of size k we have $f(J) = h(J)$.*

The proof of this lemma is established with Parseval's theorem and a characterization of the structure of the nonzero Fourier coefficients of the pmfs of $\mathcal{P}_J$ and $\mathcal{P}$.

Lemma 9 (Step 3) *Let $\mathcal{P}$ be a junta distribution on the set J^* of size k. Suppose we draw $s = 72 \cdot 2^{2k} \cdot \ln(12 \min(n^k, 2^n))/\epsilon^4$ samples from $\mathcal{P}$. For any set J of size k, we estimate $f(J)$, as defined in (3), by*

$$\tilde{f}(J) = \sum_{S \subseteq J, S \neq \emptyset} \left(\frac{2 \cdot [\#\ \text{samples } x \text{ with } \chi_S(x) = 1]}{s} - 1 \right)^2.$$

With probability $5/6$ all of the J's we have $|f(J) - \tilde{f}(J)| < \epsilon^2$.

Algorithm 2 An ϵ-tester for junta distributions

$x_1, x_2, \ldots, x_s \leftarrow$ draw $s = \mathcal{S}(2^k, 2^{n-k}, \epsilon, \left(3\binom{n}{k}\right)^{-1})$ samples from $\mathcal{P}$.
for all subsets $J \subseteq [n]$ of size $|J| = k$ **do**
  Convert each sample x_j to a pair $p^J(x_j) = (i, z)$ such that $z = x_j^{([n]\setminus J)}$ and the binary encoding of i is $x_j^{(J)}$.
  Run the uniformity test of the collection of distributions using p_j^J's
  if the test accepts **then**
    return Accept
  end if
end for
return Reject.

The proofs of Equation (5) and Lemmas 7–9 are presented in Appendices B–E. Combining the three lemmas, we obtain that with large constant probability, every invalid set J has estimated value $\tilde{f}(J)$ less than $\tilde{f}(J^*)$ and, therefore, that the set $J^\dagger$ output by Algorithm 1 is not invalid. To complete the proof of correctness of the algorithm, it remains to show that with large constant probability, $\mathrm{d_{TV}}(\mathcal{P}_{J^\dagger}, \widetilde{\mathcal{P}}_{J^\dagger}) \leq \frac{\epsilon}{2}$. This is done via a standard use of Hoeffding's inequality and the union bound.

Finally, to analyze the running time of the algorithm, observe that we consider $\binom{n}{k}$ subsets of size k. Each of them has $2^k - 1$ non-empty subsets (the S's). We compute $\chi_S(x)$ for each sample x in time $O(k)$. Thus, the time complexity of our algorithm is $O(n^k \cdot 2^{3k} \cdot k \cdot \min(k \log n, n)/\epsilon^4)$. We can obtain a faster algorithm using dynamic programming and the fact that $\chi_S(x) = (-1)^{x_i} \cdot \chi_{S\setminus\{i\}}(x)$. We compute the number of samples with $\chi_S(x_i) = 1$ for each S of size at most k in time $O(s \cdot \sum_{i=1}^{k} \binom{n}{i})$, or $O(s \cdot \min(2^n, n^k))$. Then, the running time of the algorithm can be improved to $O(\min(2^n, n^k) \cdot \min(k \log n, n) \cdot 2^{2k}/\epsilon^4)$. ■

3. Testing junta distributions

In this section we consider the problem of testing junta distributions: how do we determine whether there exists a subset of coordinates of size k, namely J, such that conditioning on any setting of $x^{(J)}$, we get the uniform distribution? One way to cast the problem of testing that a distribution is a k-junta distribution is as the problem of testing whether a collection of 2^k distributions are all uniform. In Section 3.1, we provide a uniformity test for a collection of distributions, which is a natural problem in its own right. In Algorithm 2, we describe the reduction and prove its correctness in the following theorem.

Theorem 10 *Assume there exists an ϵ-tester for uniformity of a collection of m distributions over $[n]$ that uses $\mathcal{S}(m, n, \epsilon, \delta)$ samples. Then, Algorithm 2 is an ϵ-tester for k-junta distributions using $\mathcal{S}(2^k, 2^{n-k}, \epsilon, \left(3\binom{n}{k}\right)^{-1})$ samples.*

Proof We use $\mathcal{P}$ to denote the underlying distribution that we want to test. Fix a set J of size k. Given J, we view $\mathcal{P}$ as a collection of 2^k distributions over the domain $\{0,1\}^{n-k}$.

We denote the binary encoding of i over k bits by C_i. Let $\mathcal{P}_i^J$ be the marginal distribution over the domain $\{0,1\}^{n-k}$ such that

$$\mathcal{P}_i^J(z) := \Pr_{x\sim\mathcal{P}}\left[x^{([n]\setminus J)} = z | x^{(J)} = C_i\right].$$

Assume x is a sample drawn from $\mathcal{P}$. We convert x to a pair $p^J(x) = (i,z)$ such that $z = x^{([n]\setminus J)}$ and the binary encoding of i is $x^{(J)}$. Drawing x from $\mathcal{P}$ can be viewed as drawing $p^J(x)$ from a collection of distributions according to the following process. First, i is drawn from $\{0,1,\ldots,2^k-1\}$ with probability $\Pr_{x\sim\mathcal{P}}[x^{(J)} = C_i]$. Then, z is drawn from $\mathcal{P}_i^J$. Using this conversion, we can view $\mathcal{P}$ as a distribution over a collection of 2^k distributions with the domain of size 2^{n-k} elements.

Now, we prove the correctness of Algorithm 2. We first show that if $\mathcal{P}$ is a junta distribution, then it passes with probability at least $2/3$: Assume $\mathcal{P}$ is a junta distribution on the set J^*. By definition, the $\mathcal{P}_i^{J^*}$'s are uniform, i.e., $\mathcal{P}_i^{J^*}(z) = 1/2^{n-k}$ for $z \in \{0,1\}^{n-k}$. This means that in the iteration where $J = J^*$ in the algorithm, the collection of distributions ($\mathcal{P}_i^J$'s) should be accepted by the *uniformity test of a collection*. Thus, the algorithm outputs the correct answer with probability at least $2/3$.

Second, we show that the algorithm rejects every distribution $\mathcal{P}$ which is ϵ-far from being a junta distribution with probability at least $2/3$. Recall $\mathcal{P}_J$ as defined in (1). By definition it is a $|J|$-junta distribution and so, by assumption, $\mathcal{P}$ is ϵ-far from $\mathcal{P}_J$. We compute the distance of $\mathcal{P}_J$ and $\mathcal{P}$. Let X_i be the set of all x's such that $x^{(J)} = C_i$. Then,

$$\begin{aligned}
2\,\mathrm{d_{TV}}(\mathcal{P},\mathcal{P}_J) &= \sum_x |\mathcal{P}(x) - \mathcal{P}_J(x)| = \sum_{i=1}^{2^k-1}\sum_{x\in X_i} |\mathcal{P}(x) - \mathcal{P}_J(x)| \\
&= \sum_{i=0}^{2^k-1}\sum_{x\in X_i} \Pr_{y\sim\mathcal{P}}[y\in X_i]\cdot\left|\frac{\mathcal{P}(x)}{\Pr_{y\sim\mathcal{P}}[y\in X_i]} - \frac{\mathcal{P}_J(x)}{\Pr_{y\sim\mathcal{P}}[y\in X_i]}\right| \\
&= \sum_{i=0}^{2^k-1}\sum_{x\in X_i} \Pr_{y\sim\mathcal{P}}[y\in X_i]\cdot\left|\frac{\mathcal{P}(x)}{\Pr_{y\sim\mathcal{P}}[y\in X_i]} - \frac{\mathcal{P}_J(x)}{\Pr_{y\sim\mathcal{P}_J}[y\in X_i]}\right| \\
&= \sum_{i=0}^{2^k-1}\sum_{x\in X_i} \Pr_{y\sim\mathcal{P}}[y\in X_i]\cdot\left|\mathcal{P}_i^J(x^{([n]\setminus J)}) - \frac{1}{2^{n-k}}\right| = 2\sum_{i=0}^{2^k-1}\Pr_{y\sim\mathcal{P}}[y\in X_i]\cdot\mathrm{d_{TV}}(\mathcal{P}_i^J,\mathcal{U}).
\end{aligned}$$

The third line is from $\Pr_{y\sim\mathcal{P}}[y\in X_i] = \Pr_{y\sim\mathcal{P}}[y^{(J)} = C_i] = \Pr_{y\sim\mathcal{P}_J}[y^{(J)} = C_i]$ by the definition of $\mathcal{P}_J$. Note that if we view the distribution as a collection of $\mathcal{P}_i^J$'s then $\Pr_{y\sim\mathcal{P}}[y\in X_i]$ is the weight of $\mathcal{P}_i^J$ in the collection, namely w_i. Thus, the value of $\mathrm{d_{TV}}(\mathcal{P},\mathcal{P}_J)$ or (equivalently $\mathrm{d_{TV}}(\mathcal{P},\mathcal{P}_J)$) is equivalent to the weighted distance of the collection. Since $\mathcal{P}$ is ϵ-far from any junta distribution, $\mathrm{d_{TV}}(\mathcal{P},\mathcal{P}_J)$ is at least ϵ. Thus, the collection is ϵ-far from being a collection of uniform distributions and will be rejected by the uniformity test for all J's with high probability. ■

Note that by applying Theorem 11 with standard amplification and union bound arguments, we can assume that all tests return the correct answer with probability at least $2/3$.

In addition, by setting $m = 2^k$ and $n = 2^{n-k}$ in Theorem 11, it is not hard to see that the total number of samples is $\tilde{O}(k \cdot \log n \cdot \mathcal{S}(m, n, \epsilon, \delta)) = \tilde{O}(2^{n/2}k^4/\epsilon^3 + 2^k/\epsilon)$.

3.1. Uniformity Test of a Collection of Distributions

In this section we consider the problem of testing uniformity of a collection of m distributions $\{\mathcal{P}_i|w_i\}_{i=1}^m$ on the domain $[n]$. To draw a sample (i, j) from the collection, first we pick the distribution $\mathcal{P}_i$ with probability w_i and then we draw a sample j from $\mathcal{P}_i$. The naïve approach to solving this problem is to test uniformity of each distribution separately using $\Omega(m\sqrt{n})$ samples. However, this is not optimal. For example if we know all the w_i's are equal to $1/m$ then our problem can be converted to the uniformity test over the domain $[m] \times [n]$, which requires $O(\sqrt{m\,n})$ samples (Paninski (2008)). This observation suggests that when the weights w_i are similar to each other, it may be preferable not to test the distributions separately. Implementing this idea, we introduce an ϵ-tester for the special case where all the weights w_i's are within a constant factor of each other (See Appendix F). To generalize this idea to general weights, we use a bucketing argument such that the $\mathcal{P}_i$'s in each bucket have roughly the same w_i weights. Then, we perform the uniformity test in each bucket separately and integrate the result to conclude whether the collection is uniform or not. The sample complexity of the resulting algorithm is $\tilde{O}(\sqrt{m\,n} + m)$.

Another problem that is closely related to that of testing uniformity is the more general problem of testing equivalence of distributions. This more general problem has been considered before by Levi et al. (2013) and Diakonikolas and Kane (2016). Levi et al. (2013) consider several different testing models, including the model of testing collections of distributions that we consider here. The sample complexity of their algorithm for testing equivalence of distributions in this model is $\tilde{O}(n^{2/3}\,m^{1/3} + m)$. In the regime of parameters where $m \leq n$, our algorithm requires fewer samples.

Another testing model introduced in Levi et al. (2013) considers the setting where the testing algorithm knows the weights in advance. Levi et al. (2013) present an algorithm for testing equivalence of distributions in this model with $\tilde{O}\left(\max(n^{2/3}\,m^{1/3}, \sqrt{m\,n})\right)$ samples. Very recently, Diakonikolas and Kane (2016) provide another algorithm for the same task with sample complexity $O\left(\max(n^{2/3}\,m^{1/3}, \sqrt{m\,n})\right)$. This result is optimal in terms of m, n, and ϵ. We did not explicitly provide a result in this model. However, it is not hard to see that if our algorithm knows the weights w_i's, it will only uses $\tilde{O}\sqrt{m\,n})$ samples, which outperforms the other two algorithms for the special case of testing uniformity in the setting where $m \leq n$.

Theorem 11 *Algorithm 3 is an ϵ-tester for uniformity of a collection of distributions $\{\mathcal{P}_i|w_i\}$ using $s = \tilde{O}(\sqrt{mn}/\epsilon^3 + m/\epsilon)$ samples.*

Proof In the algorithm, instead of drawing a fixed number of samples, we use the "Poissonization method"[3] and draw s samples where s is a random variable drawn from a Poisson

3. Observe that when we draw a fixed number of samples, the number of appearances of each element depends on others. This would usually convolute the analysis of the algorithm. However, the Poisson distribution has the convenient property that the number of appearances of each symbol is independent from the others. It is known that if a single distribution $\mathcal{P}$ is sampled Poi(n) times, then the number of samples equal to a symbol x is a random variable from a Poisson distribution with mean $n\mathcal{P}(x)$ (see

Algorithm 3 An ϵ-tester for testing uniformity of a collection of distributions

$B \leftarrow \lceil \log(4m/\epsilon) \rceil$
$\mathcal{S} \leftarrow 40\, m \log(12(m+1))/\epsilon$
Draw a sample, namely s, from Poi($\mathcal{S}$).
if $s > 2\mathcal{S}$ **then**
  Reject and halt.
end if
$x_1, x_2, \ldots, x_s \leftarrow$ draw s samples from the collection $\{\mathcal{P}_i|w_i\}_{i=1}^m$.
$s_i \leftarrow$ number of samples from $\mathcal{P}_i$ for $i \in [m]$.
$\hat{w}_i \leftarrow s_i/\mathcal{S}$ number of samples from $\mathcal{P}_i$ for $i \in [m]$.
for $\ell = 1, \ldots, B$ **do**
  $B_\ell \leftarrow \{i \mid 2^{\ell-1}\epsilon/(4m)\} \leq \hat{w}_i < 2^\ell \epsilon/(4m)$
  $\widehat{W_\ell} \leftarrow \sum_{i \in B_\ell} \hat{w}_i$
  $S_l \leftarrow \sum_{i \in B_\ell} s_i$
  if $\widehat{W_\ell} \geq \epsilon/4B$ **then**
    Run *bucket uniformity test* with distance parameter $\epsilon/2B$ and maximum error probability $1/6B$
    if the test rejects **then**
      return Reject.
    end if
  end if
end for
return Accept.

distribution with mean $\mathcal{S}$. Thus, we can assume the number of samples from each distribution $\mathcal{P}_i$, namely s_i, is distributed as $Poi(w_i \cdot \mathcal{S})$ and is independent from the rest of s_j's. Now, we show in the following concentration lemma that the s_i's are not far from their mean. Equivalently, we prove that $\hat{w}_i = s_i/\mathcal{S}$ is close to w_i.

Lemma 12 *Suppose we draw $s \sim Poi(\mathcal{S})$ samples from a collection of distributions $\{\mathcal{P}_i|w_i\}_{i=1}^m$ such that $\mathcal{S} \geq 40\, m \log 12(m+1)/\epsilon$. Let $\hat{w}_i = s_i/\mathcal{S}$ where s_i is the number of samples from $\mathcal{P}_i$. With probability of 5/6 all of the following events happen.*

- *s is in the range $[\mathcal{S}/2, 2\mathcal{S}]$.*
- *For any i if $w_i \geq \epsilon/8m$, then $\hat{w}_i$ is in the range $[\frac{1}{2}w_i, 2w_i]$.*
- *For any i if $w_i < \epsilon/8m$, then $\hat{w}_i \leq \epsilon/4m$.*

Proof Here we need to use concentration inequalities for Poisson distribution. (See Theorem 5.4 in Mitzenmacher and Upfal (2005)) For a Poisson random variable X with mean μ

$$\Pr(X \geq (1+\beta)\mu) \leq \left(\frac{e^\beta}{(1+\beta)^{(1+\beta)}}\right)^\mu;$$

for example Mitzenmacher and Upfal (2005)). This also implies that the number of appearances of each symbol is independent of the others.

and

$$\Pr(X \leq (1-\beta)\mu) \leq \left(\frac{e^{-\beta}}{(1-\beta)^{(1-\beta)}}\right)^{\mu}.$$

It is not hard to see

$$1 - \Pr[\mu/2 \leq X \leq 2\mu] < 0.68^{\mu} + 0.86^{\mu} < 2 \cdot 2^{-\mu/5} \leq \frac{1}{6(m+1)}.$$

where the last inequality holds for $\mu \geq 5\,\log(12\,(m+1))$. Thus, s is in the range $[\mathcal{S}/2, 2\,\mathcal{S}]$ with probability $1 - 1/6(m+1)$.

Note that by properties of the Poissonization method (Mitzenmacher and Upfal (2005)), the s_i's are distributed as independent draws from $\mathrm{Poi}(w_i \cdot \mathcal{S})$. For a fixed $w_i \geq \epsilon/8m$, since $w_i \cdot \mathcal{S}$ is at least $5\,\log(12\,(m+1))$, we can conclude that s_i is in the range $[w_i \cdot \mathcal{S}/2, 2w_i \cdot \mathcal{S}]$ or equivalently $\hat{w_i}$ is in the range $[w_i/2, 2w_i]$ with probability at least $1 - 1/6(m+1)$. Now assume w_i is less than $\epsilon/8m$. Clearly, the expected value of s_i is less than $\mathcal{S} \cdot \epsilon/8m$. Consider another random variable X which is drawn from $\mathrm{Poi}(\mathcal{S} \cdot \epsilon/8m)$. Thus,

$$\Pr[s_i > \mathcal{S} \cdot \epsilon/4m] \leq \Pr[X > \mathcal{S} \cdot \epsilon/4m] \leq \frac{1}{6(m+1)}$$

Thus, by the union bound over the s_i's and the s with probability at least $5/6$ the conclusions of the lemma hold. ■

Partitioning into buckets: Based on the idea that uniformity test of a collection of distributions is easier when w_i's are uniform, we partition the distributions into buckets such that w_i in the same buckets are within a constant factor of each other. Assume we have $B = \lceil \log(4m/\epsilon) \rceil$ buckets where the ℓ-th buckets contains all the distributions $\mathcal{P}_i$'s such that $2^{\ell-1}\epsilon/4m < \hat{w_i} \leq 2^{\ell}\epsilon/4m$. By Lemma 12, the w_i's are in the range $[\epsilon/8m2^{\ell-1}, \epsilon/2m2^{\ell}]$. Observe that each bucket ℓ can be viewed as a (sub-)collection of $m_\ell = |B_\ell|$ distributions with the new weights w_i/W_ℓ where W_ℓ is the total weight of the ℓ-th bucket.

Reduction to the *bucket uniformity test*: Here, we want to show that there is a reduction between uniformity test of a collection of distributions and uniformity test of each bucket as a sub-collection of distributions. For uniformity test of a collection, we partition the collection into buckets as explained before. Then for each bucket, we invoke the *bucket uniformity test* with distance parameter $\epsilon/2B$ and with error probability of at most $1/6B$. To prove the correctness of the reduction, we consider the two following cases:

- $\{\mathcal{P}_i|w_i\}_{i=1}^m$ **is a collection of uniform distributions.** Since all of the distributions are uniform, all buckets contain only uniform distributions. Then, all the B invocations of *bucket uniformity test* should accept with probability at least $1 - 1/6B$. Thus, none of them rejects with probability $1 - 1/6$ by the union bound.

- $\{\mathcal{P}_i|w_i\}_{i=1}^m$ **is ϵ-far from being a collection of uniform distributions.** We prove that at least one bucket should be rejected with high probability. Note that in our bucketing method we ignore the distributions with $\hat{w_i} \leq \epsilon/4m$: by Lemma 12, the total weight of these distributions is at most $\epsilon/2$ and since the total variation distance

is at most one, they can not contribute to the weighted distance by more than $\epsilon/2$. Thus,

$$\sum_{l=1}^{B} \sum_{i \in B_\ell} w_i \cdot \mathrm{d_{TV}}(\mathcal{P}_i, \mathcal{U}) \geq \epsilon/2.$$

By averaging there is at least one bucket, namely ℓ, such that $\sum_{i \in B_\ell} w_i \cdot \mathrm{d_{TV}}(\mathcal{P}_i, \mathcal{U}) \geq \epsilon/2B$. Let W_ℓ be the total weight of the ℓ-th bucket. Since the total variation distance is at most one, $W_\ell = \sum_{i \in B_\ell} w_i \geq \epsilon/2B$. Therefore, $\widehat{W_\ell} = \sum_{i \in B_\ell} \hat{w}_i \geq \epsilon/4B$. In addition, we consider this bucket as a separate collection. Since $W_\ell \leq 1$, if we renormalize the weights, we also see that

$$\sum_{i \in B_\ell} \frac{w_i}{W_\ell} \cdot \mathrm{d_{TV}}(\mathcal{P}_i, \mathcal{U}) \geq \frac{\epsilon}{2BW_\ell} \geq \frac{\epsilon}{2B}.$$

Now if we show the assumptions of Corollary 27 are satisfied, then the *bucket uniformity test* rejects the bucket ℓ with probability at least $1 - 1/6B$. It is not hard to see that our estimation of the new weight of the i-th distribution in bucket ℓ is $\hat{w}_i/\widehat{W_\ell}$, which is in the range $[w_i/4W_\ell, 4w_i/W_\ell]$ by Lemma 12. Moreover, since the $\hat{w}_i$'s are in the range $[2^{\ell-1}\epsilon/8m, 2^\ell\epsilon/2m]$, every w_i/W_ℓ is at most $8/m_i$ where $m_i = |B_\ell|$.

Thus the sample complexity for testing the buckets is

$$\sum_{i=1}^{B} s_i = O\left(\sum_{i=1}^{B} B^2\sqrt{m_i\, n}\log(6B)/\epsilon^2 + \log(6B) m_i \log m_i\right).$$

Using the Cauchy-Schwarz inequality and since $B = O(\log(m/\epsilon))$, it is not hard to see that the total sample complexity is $\tilde{O}(\sqrt{m\, n}/\epsilon^2 + m)$.

Using the union bound, the test does not fail with probability more than a $1/3$. The total sample complexity is $\tilde{O}(\sqrt{m\, n}/\epsilon^2 + m/\epsilon)$. Hence, the proof is complete. ■

4. Lower Bounds

4.1. Sample complexity lower bound for learning juntas

We now complete the proof of the first part of Theorem 3.

Theorem 13 (Restatement of the first part of Theorem 3) *Fix $0 < \epsilon < \frac{1}{2}$ and $1 \leq k < n$. Any ϵ-learner for k-junta distributions with respect to the uniform distribution over $\{0,1\}^n$ must have sample complexity $s = \Omega(\max\{2^k/\epsilon^2, \log\binom{n}{\leq k}/\epsilon\})$.*

Proof The first part of the lower bound, $s = \Omega(2^k/\epsilon^2)$ follows from the (folklore) lower bound on the number of samples required to learn a general discrete distribution over a domain of size N. $\Omega(N/\epsilon^2)$ samples are required for this task. Observing that the set of juntas on the set $J = \{1, 2, \ldots, k\}$ is a set of general discrete functions on a domain of size $N = 2^k$, we conclude that any k-junta learning algorithm must draw $\Omega(2^k/\epsilon^2)$ samples—even if it is given the identity of the junta coordinates.

We now want to show that $s = \Omega(\log \binom{n}{\leq k}/\epsilon)$. By Yao's minimax principle, it suffices to show that there is a distribution $\mathcal{P}$ on k-junta distributions such that any deterministic ϵ-learner for $D \sim \mathcal{P}$ must draw at least $s = \Omega(\log \binom{n}{\leq k}/\epsilon)$ samples from D. For non-empty sets $S \subseteq [n]$, let D_S be the distribution with the probability mass function

$$p_S(x) = \begin{cases} (\frac{1}{2} + \epsilon)/2^{n-1} & \text{if } \bigoplus_{i \in S} x_i = 1 \\ (\frac{1}{2} - \epsilon)/2^{n-1} & \text{if } \bigoplus_{i \in S} x_i = 0. \end{cases}$$

Let $D_\emptyset$ be the uniform distribution on $\{0,1\}^n$. We let $\mathcal{P}$ be the distribution defined by $\mathcal{P}(D_\emptyset) = \frac{1}{2}$ and $\mathcal{P}(D_S) = \frac{1}{2(\binom{n}{\leq k}-1)}$ for every set of size $1 \leq |S| \leq k$. Every function in the support of $\mathcal{P}$ is a k-junta distribution, and they are all ϵ-far from each other.

Fix any deterministic learning algorithm $\mathcal{A}$ that is ϵ-learner for the k-junta distributions drawn from $\mathcal{P}$. Let X be a sequence of s samples drawn from D. The success probability of $\mathcal{A}$ guarantees that

$$\begin{aligned} \frac{2}{3} &\leq \Pr[\mathcal{A} \text{ identifies the correct distribution}] \\ &= \sum_{S \in \binom{n}{\leq k}} \mathcal{P}(D_S) \sum_{X \in \{0,1\}^{n \times s}} p_S(X) \cdot \mathbf{1}[\mathcal{A} \text{ outputs } D_S \text{ on } X] \\ &\leq \sum_{X \in \{0,1\}^{n \times s}} \max_{S \in \binom{n}{\leq k}} \mathcal{P}(D_S) \cdot p_S(X). \end{aligned}$$

We can partition the set of s-tuples of samples, $\{0,1\}^{n \times s}$ into $\binom{n}{\leq k}$ parts χ_S, $S \in \binom{n}{\leq k}$ such that $X \in \chi_S$ iff $\mathcal{P}(D_S) \cdot p_S(X) = \max_{T \in \binom{n}{\leq k}} \mathcal{P}(D_T) \cdot p_T(X)$ (breaking ties arbitrarily).

If every X belongs to $\chi_\emptyset$, then it means that our algorithm always outputs the uniform distribution, which clearly does not satisfy the learning requirement. (With probability 1/2, the hypothesis from this algorithm will be epsilon-far from the target function.) So some X must belong in χ_S for a non-empty set S. For any set of samples X, we have that $\mathcal{P}(D_\emptyset) \cdot p_S(X) = \frac{1}{2} \cdot 2^{-ns}$ since $D_\emptyset$ is the uniform distribution. This means that if $X \in \chi_S$ for some $S \neq \emptyset$, then $\mathcal{P}(D_S) \cdot p_S(X) \geq 2^{-ns-1}$ and hence $p_S(X) \geq (\binom{n}{\leq k} - 1)2^{-ns}$. Let $\kappa_S(X)$ denote the number of samples $x \in X$ such that $\bigoplus_{i \in S} x_i = 1$. Then from the above inequality we have

$$\begin{aligned} \left(\binom{n}{\leq k} - 1\right) \cdot 2^{-ns} \leq p_S(X) &= (\frac{1}{2} + \epsilon)^{\kappa_S(X)} (\frac{1}{2} - \epsilon)^{s - \kappa_S(X)} 2^{-s(n-1)} \\ &\leq (1 + 2\epsilon)^{\kappa_S(X)} \cdot 2^{-ns} \leq e^{2\epsilon \kappa_S(X)} \cdot 2^{-ns}. \end{aligned}$$

Therefore, $s \geq \kappa_S(X) = \Omega(\log \binom{n}{\leq k}/\epsilon)$, as we wanted to show.

■

4.2. Time complexity lower bound for learning juntas

The time complexity component of Theorem 3 is established via the intermediate problem of learning k-junta pmfs in the standard functional setting.

Definition 14 *The function $f : \{0,1\}^n \to [0,1]$ is a k-*junta pmf *if $\sum_{x\in\{0,1\}^n} f(x) = 1$ and f is a k-junta.*

An *ϵ-learner* for a class $\mathcal{C}$ of pmfs $f : \{0,1\}^n \to [0,1]$(under the uniform distribution) is a learning algorithm that draws s samples $x_1, \ldots, x_s$ independently and uniformly at random from $\{0,1\}^n$, observes the s pairs $(x_1, f(x_1)), \ldots, (x_s, f(x_s))$, and outputs a hypothesis function $h : \{0,1\}^n \to \mathbb{R}$ such that with probability at least $\frac{2}{3}$, $\sum_{x\in\{0,1\}^n} |f(x) - h(x)| \leq \epsilon$. An *$\epsilon$-learner* for Boolean functions $f : \{0,1\}^n \to \{0,1\}$ is defined analogously, except that the guarantee on the hypothesis $h : \{0,1\}^n \to \{0,1\}$ is that it satisfies $\Pr_x[f(x) \neq h(x)] \leq \epsilon$ with probability at least $\frac{2}{3}$.

Lemma 15 *If there is an ϵ-learner for k-junta distributions with time complexity t and sample complexity s, then there is an ϵ-learner for the class of k-junta pmfs which runs in time $O(t + 2^k s)$.*

Proof Let $f : \{0,1\}^n \to [0,1]$ be the input to the k-junta pmf learning problem, and let $\mathcal{A}$ be an algorithm for learning k-junta distributions with sample complexity s and time complexity t.

Consider now the algorithm $\mathcal{A}'$ that draws $x \in \{0,1\}^n$ uniformly at random, observes $(x, f(x))$, and then with probability $2^{n-k} f(x)$, passes x along as a sample to $\mathcal{A}$. After passing s samples to $\mathcal{A}$ in this way, $\mathcal{A}'$ returns the function $h : \{0,1\}^n \to [0,1]$ that corresponds to the pmf of the distribution $\mathcal{D}'$ learned by $\mathcal{A}$.

Note that since f is a k-junta, for every $x \in \{0,1\}^n$ there are at least 2^{n-k} inputs $y \in \{0,1\}^n$ that have the same value as x on the (at most) k relevant coordinates and so satisfy $f(y) = f(x)$. Taking the sum of $f(y)$ over all such inputs, we get $\sum f(y) = 2^{n-k} f(x) \leq 1$, so the expression $2^{n-k} f(x)$ defined above is indeed a valid probability.

For any given $x \in \{0,1\}^n$, the probability that $\mathcal{A}'$ draws x and passes it along to $\mathcal{A}$ is $\Pr[x \text{ is drawn}] \cdot \Pr[x \text{ is accepted}] = 2^{-n} \cdot 2^{n-k} f(x) = 2^{-k} f(x)$. The probability that the current sample is accepted is therefore $\sum_z 2^{-k} f(z) = 2^{-k}$. So the probability that the next sample to be accepted is x is $\frac{2^{-k} f(x)}{2^{-k}} = f(x)$, guaranteeing that the distribution on samples passed to $\mathcal{A}$ indeed has pmf f, and $O(2^k s)$ initial samples are sufficient to generate the s samples required by $\mathcal{A}$ with large constant probability. When this condition is satisfied, then $\mathcal{A}$ returns a distribution with pmf $h : \{0,1\}^n \to [0,1]$ that with large constant probability satisfies $\sum_x |f(x) - h(x)| \leq \epsilon$, as required. ■

Lemma 16 *If there is an $\frac{\epsilon}{2}$-learner for k-junta pmfs with time complexity t, then there is an ϵ-learner for the class of k-junta Boolean functions which runs in time $O(t + 2^{2k})$.*

Proof Let $\mathcal{A}$ be an algorithm for $\frac{\epsilon}{2}$-learning k-junta pmfs with sample complexity s and time complexity t. We can design an algorithm $\mathcal{A}'$ for learning k-junta Boolean functions as follows.

Let $f : \{0,1\}^n \to \{0,1\}$ be the input to the Boolean function k-junta learning problem. Define $\Sigma(f) := \sum_{x\in\{0,1\}^n} f(x)$ and let $g : \{0,1\}^n \to [0,1]$ be the pmf defined by $g(x) = f(x)/\Sigma(f)$.

Since f is a k-junta, $\Sigma(f) \in \{0, 2^{n-k}, 2 \cdot 2^{n-k}, \ldots, 2^k \cdot 2^{n-k}\}$ and so the algorithm $\mathcal{A}'$ can learn $\Sigma(f)$ exactly with large constant probability with $O(2^{2k})$ samples. It can then use $\mathcal{A}$ to learn a hypothesis $\tilde{g}$ of g by passing along the sample $(x, f(x)/\Sigma(f))$ when it observes the sample $(x, f(x))$. The total running time of the estimation and simulation is $O(2^{2k} + t)$, as required. Furthermore, $\mathcal{A}$ guarantees that with large constant probability, the hypothesis $\tilde{g}$ satisfies $\sum |\tilde{g}(x) - g(x)| \leq \frac{\epsilon}{2}$. Let $\mathcal{A}'$ output the function $\tilde{f}$ defined by $\tilde{f}(x) = \mathbf{1}\left[\tilde{g}(x) \geq \frac{1}{2\Sigma(f)}\right]$. By this construction, for every input $x \in \{0,1\}^n$, if $f(x) \neq \tilde{f}(x)$ then $|\tilde{g}(x) - g(x)| \geq \frac{1}{2\Sigma(f)}$. So by Markov's inequality,

$$\Pr_x[\tilde{f}(x) \neq f(x)] = \Pr_x\left[|\tilde{g}(x) - g(x)| \geq \tfrac{1}{2\Sigma(f)}\right] \leq 2\Sigma(f) \cdot \frac{1}{2^n} \sum_{x \in \{0,1\}^n} |\tilde{g}(x) - g(x)| \leq 2\mathrm{E}[f] \cdot \frac{\epsilon}{2} \leq \epsilon$$

and $\mathcal{A}'$ is indeed an ϵ-learner for k-junta Boolean functions, as we wanted to show. ■

4.3. Lower bound for testing juntas

In this section, we prove a lower bound for testing junta distributions.

Theorem 17 *There is no ϵ-tester for k-junta distributions using $o(2^{n/2}/\epsilon^2)$ samples for sufficiently small ϵ.*

Proof We show a reduction from testing uniformity of a collection of distributions to testing k-junta distributions. Then, we use the lower bounds for testing uniformity of a collection of distributions to prove the theorem.

Assume we have a collection of distributions, $\mathcal{C} = \{\mathcal{P}_i | w_i\}_{i=1}^{M}$, over the domain $[N]$. Without loss of generality assume $M = 2^{k_1}$ and $N = 2^{k_2}$. If they are not powers of two, round each of them to the next power of two and this does not affect M, N and ϵ in our proof by more than a constant factor.

Given $\mathcal{C}$, we construct a distribution over $\{0,1\}^{k_1+k_2+1}$.

$$\mathcal{D}(x) = \begin{cases} 0 & \text{if } x^{([k_1+1])} \text{ has odd parity.} \\ \mathcal{P}_{i_x+1}(j_x + 1) & \text{if } x^{([k_1+1])} \text{ has even parity.} \end{cases}$$

such that $C_{i_x} = x^{([k_1])}$ and $C_{j_x} = x^{([k_1+k_2+1]\setminus[k_1+1])}$. Note that we use C_i to indicate the binary encoding of i. Let $k = k_1 + 1$. If $\mathcal{C}$ is a collection of uniform distributions, it is not hard to see that $\mathcal{D}$ is k-junta on the set $[k]$. Moreover, if $\mathcal{D}$ is a k-junta distribution, then $\mathcal{C}$ has to be a collection of uniform distributions. Below, we show this fact, by proving that $\mathcal{D}$ is $1/4$-far from being k-junta on every set $J \neq [k]$. This implies that $\mathcal{D}$ has to be a k-junta distribution on the set $[k]$ which means that all $\mathcal{P}_i$'s are uniform.

Fix an arbitrary set $J \neq [k]$ of size k. Let $\mathcal{D}'$ be any arbitrary k-junta distribution on the set J. Let $a_i = \Pr_{x \sim \mathcal{D}}\left[x^{(J)} = C_i\right]$ for $i = 0, 1, \ldots, 2^k - 1$. We define X_i to be the set of all x's such that $x^{(J)} = C_i$. Now, we show that at least half of the elements in X_i have probability zero. Since $J \neq [k]$ and $|J| = k$, there exists a coordinate $\ell \in [k]$ such that ℓ is not in the set J. Consider an element $x \in X_i$. Let y be x with the ℓ-th bit flipped. Since

ℓ is not in J, $x^{(J)} = y^{(J)}$. Thus, y is also in X_i. On the other hand, the parity of the first k bits of x and y is not the same, because $\ell \in [k]$. Thus, one of them has probability zero. Note that we can pair up all the elements in X_i as we did for to x and y. Therefore, at least half of the elements in X_i have probability zero. By definition, we know $a_i = \sum_{x \in X_i} \mathcal{D}'(x)$ for $i = 0, 1, \ldots, 2^k - 1$ and $\mathcal{D}'(x) = a_i/2^{k_2}$ for $x \in X_i$. We can conclude that $\mathcal{D}'$ is 1/4-far from $\mathcal{D}$, since

$$\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') &= \frac{1}{2}\sum_{x} |\mathcal{D}(x) - \mathcal{D}'(x)| = \frac{1}{2}\sum_{i=0}^{2^k-1}\sum_{x \in X_i} |\mathcal{D}(x) - \mathcal{D}'(x)| \\
&= \frac{1}{2}\sum_{i=0}^{2^k-1}\sum_{x \in X_i} |\mathcal{D}(x) - a_i/2^{k_2}| \geq \frac{1}{2}\sum_{i=0}^{2^k-1}\sum_{x \in X_i : \mathcal{D}(x)=0} a_i/2^{k_2} \geq \sum_{i=0}^{2^k-1}\frac{a_i}{4} \geq \frac{1}{4}
\end{aligned}$$

where the first inequality follows from the fact that at least 2^{k_2-1} elements in X_i have probability zero. Since $\mathcal{D}'$ is an arbitrary k-junta distribution on an arbitrary set $J \neq [k]$, it follows that $\mathcal{D}$ is 1/4-far from being a junta distribution on any $J \neq [k]$.

As we mentioned earlier, this implies that $\mathcal{D}$ is a k-junta distribution iff $\mathcal{C}$ is a collection of uniform distributions. It is not hard to see that we can convert a sample drawn from $\mathcal{C}$ to get a sample from $\mathcal{D}$. Thus, any ϵ-tester for a k-junta distribution can be used as an ϵ-tester for a collection of uniform distributions. Diakonikolas and Kane (2016) (in section 3.1.1) have shown that $\Omega(\sqrt{MN}/\epsilon^2) = \Omega(2^{(k_1+k_2)/2})$ samples are required to distinguish a collection of M uniform distributions from a collection which is ϵ-far from being uniform with probability 2/3. This implies that we need $\Omega(2^{n/2}/\epsilon^2)$ samples to test k-junta distributions. ■

5. Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. CCF-1420692 and CCF-1065125. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. We thank anonymous reviewers for their insightful comments on the preliminary version of this paper.

References

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269, 2000. doi: 10.1109/SFCS.2000.892113. URL http://dx.doi.org/10.1109/SFCS.2000.892113.

Avrim Blum. Relevant examples and relevant features: Thoughts from computational learning theory. In *AAAI Fall Symposium on 'Relevance'*, volume 5, 1994.

Avrim Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, December 1997.

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

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1859-4. doi: 10.1145/2422436.2422497. URL http://doi.acm.org/10.1145/2422436.2422497.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28, 2014.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-modal distributions via testing. *Theory of Computing*, 10(20):535–570, 2014. doi: 10.4086/toc.2014.v010a020. URL http://www.theoryofcomputing.org/articles/v010a020.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

Ilias Diakonikolas. Learning structured distributions. In *CRC Handbook of Big Data*. 2016.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. *CoRR*, abs/1601.05557, 2016. URL http://arxiv.org/abs/1601.05557.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008.

Michael J. Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 273–282, 1994. doi: 10.1145/195058.195155. URL http://doi.acm.org/10.1145/195058.195155.

Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *Journal of the ACM (JACM*, 40(3):607–620, July 1993.

Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media, 2012.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005. ISBN 0521835402.

Elchanan Mossel, Ryan O'Donnell, and Rocco A Servedio. Learning functions of k relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434, November 2004.

Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, October 2014.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theor.*, 54(10):4750–4755, October 2008. ISSN 0018-9448. doi: 10.1109/TIT.2008.928987. URL http://dx.doi.org/10.1109/TIT.2008.928987.

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. *FOCS*, pages 11–20, 2012.

Appendix A. Learning juntas with the cover method

Fix any class $\mathcal{C}$ of distributions over $\{0,1\}^n$. An *ϵ-cover* of $\mathcal{C}$ is a collection $\mathcal{C}_\epsilon$ of distributions on $\{0,1\}^n$ such that for every distribution $D \in \mathcal{C}$, there is a distribution $D' \in \mathcal{C}_\epsilon$ such that $\mathrm{d_{TV}}(D, D') \leq \epsilon$. We can obtain a good learning algorithm for $\mathcal{C}$ by designing a small ϵ-cover for it and using the following lemma.

Lemma 18 *Let $\mathcal{C}$ be an arbitrary family of distributions and $\epsilon > 0$. Let $\mathcal{C}_\epsilon \subseteq \mathcal{C}$ be an ϵ-cover of $\mathcal{C}$ of cardinality N. Then there is an algorithm that draws $O(\epsilon^{-2} \log N)$ samples from an unknown distribution $D \in \mathcal{C}$ and, with probability $9/10$, outputs a distribution $D' \in \mathcal{C}_\epsilon$ that satisfies $\mathrm{d_{TV}}(D, D') \leq 6\epsilon$. The running time of this algorithm is $O(N \log N/\epsilon^2)$.*

See Diakonikolas (2016); Devroye and Lugosi (2001); Daskalakis et al. (2014) for good introductions to the lemma itself and its application to distribution learning problems. We are now ready to use it to complete the proof of Theorem 19.

Theorem 19 *Fix $\epsilon > 0$ and $1 \leq k \leq n$. Define $t = \binom{n}{k}\binom{2^k - 1 + 2^k/\epsilon}{2^k - 1}$. There is an algorithm A with sample complexity $O(\log t/\epsilon^2) = O(k \log n/\epsilon^2 + (2^k \log(1/\epsilon))/\epsilon^2)$ and running time $O(t \log t/\epsilon^2) = \binom{n}{k}(c/\epsilon)^{2^k+2}\left(k \log n + 2^k \log(1/\epsilon)\right)$ for some constant c that, given samples from a k-junta distribution $\mathcal{D}$, with probability at least $\frac{2}{3}$ outputs a distribution $\mathcal{D}'$ such that $\mathrm{d_{TV}}(\mathcal{D}, \mathcal{D}') := \sum_{x \in \{0,1\}^n} |\mathcal{D}(x) - \mathcal{D}'(x)|/2 \leq \epsilon$.*

Proof By Lemma 18, it suffices to show that the class of all k-junta distributions has a cover of size $N = \binom{n}{k}\binom{2^k - 1 + 2^k/\epsilon}{2^k - 1}$. This, in turn, follows directly from the fact that we can simply let $\mathcal{C}_\epsilon$ be the set of all k-juntas with probability mass function p where $p(x)$ is a multiple of $\epsilon/2^n$ for each element $x \in \{0,1\}^n$. There are $\binom{n}{k}$ ways to choose the set $J \subseteq [n]$ of junta coordinates and at most $(2^k)^{2^k/\epsilon}$ ways to allocate the probability mass in $\epsilon/2^k$ increments among the 2^k different restrictions of x on J. ■

Appendix B. Proof of Equation (5)

We establish some basic properties of $\mathcal{P}_J$ as described in Section 2. First, we introduce the notation we use below. For a fixed set J, define the biases b_i's, $i \in \{0, 1, \ldots, 2^k - 1\}$, to be the probability of $x^{(J)} = C_i$ where x is drawn from $\mathcal{P}$ and C_i is the binary encoding of i with k bits. For a subset I of size k, we define a function $r_I : I \to [k]$ such that $r_I(c)$ indicates the rank of the coordinate $c \in I$ (rank the smallest first). Basically, when $x^{(I)} = C_i$, the c-th bit of x is equal to the $r_I(c)$-th bit of C_i for all $c \in I$. In addition, we define $r_I(S)$ to be the image of subset S under the function r_I. In particular, if $x^{(I)} = C_i$, we have $x^{(S)} = C_i^{(r_I(S))}$ for any $S \subseteq I$. For a junta distribution on the set J^*, let $L(i, T) = \{j \mid 0 \leq j < 2^k \text{ and } C_j^{r_{J^*}(T)} = C_i^{r_{J^*}(T)}\}$ for any $i \in \{0, 1, \ldots, 2^k - 1\}$ and $T \subseteq J^*$. In other words, $L(i, T)$ is a set of j's such that all the C_j's agree on the setting of the coordinates in T.

Lemma 20 *Assume $\mathcal{P}$ is a k-junta distribution on the set J^*. Let J be a subset of $[n]$ such that $|J| = k$. We define $\mathcal{P}_J$ similar to Equation 1. Then, for any $i \in \{0, 1, \ldots, 2^k - 1\}$, we*

have

$$\Pr_{y\sim\mathcal{P}}\left[y^{(J)}=C_i\right]=2^{-|J\setminus J^*|}\sum_{j\in L(i,J\cap J^*)}\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right] \tag{6}$$

Proof If $J=J^*$, then $L(i,J\cap J^*)$ has just one member, i, and the proof is clear. Thus, assume $J\neq J^*$. Let $t=|J\setminus J^*|$ and it is at least one. Then, we have

$$\begin{aligned}b_i&=\Pr_{y\sim\mathcal{P}}\left[y^{(J)}=C_i\right]=\Pr_{y\sim\mathcal{P}}\left[y^{(J\setminus J^*)}=C_i^{r_J(J\setminus J^*)}\wedge y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}\right]\\&=\sum_{j=0}^{2^k-1}\Pr_{y\sim\mathcal{P}}\left[y^{(J\setminus J^*)}=C_i^{r_J(J\setminus J^*)}\wedge y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}|y^{(J^*)}=C_j\right]\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right]\end{aligned}$$

Observe that the assumption on P implies that the non-junta coordinates (including $J\setminus J^*$) are distributed uniformly. Thus, the probability of each setting for the coordinates in $J\setminus J^*$ is $2^{-|J\setminus J^*|}=2^{-t}$. Consequently, it is independent of the junta coordinates. Therefore, we conclude

$$\begin{aligned}b_i&=\sum_{j=0}^{2^k-1}\Pr_{y\sim\mathcal{P}}\left[y^{(J\setminus J^*)}=C_i^{r_J(J\setminus J^*)}\right]\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}|y^{(J^*)}=C_j\right]\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right]\\&=\sum_{l=0}^{2^k-1}2^{-t}\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}|y^{(J^*)}=C_j\right]\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right].\end{aligned}$$

Note that if $y^{(J^*)}=C_j$, then the values of x_i's on all the coordinates in $J\cap J^*$ are determined. Therefore, both binary encodings C_i and C_j should appoint the same value to these coordinates. In other words, if $C_i^{r_J(J\cap J^*)}\neq C_j^{r_{J^*}(J\cap J^*)}$, then the probability of $y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}$ is zero given the fact $y^{(J\cap J^*)}=C_j^{r_{J^*}(J\cap J^*)}$. Otherwise, it is one. Therefore,

$$\begin{aligned}b_i&=\sum_{j\in L(i,J\cap J^*)}2^{-t}\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J\cap J^*)}=C_i^{r_J(J\cap J^*)}|y^{(J^*)}=C_j\right]\cdot\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right]\\&=2^{-t}\cdot\sum_{j\in L(i,J\cap J^*)}\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right].\end{aligned}$$

Since $t=|J\setminus J^*|$, the proof is complete. ■

Moreover, by definition of $\mathcal{P}_J$ and Equation 6, and setting $C_i=x^{(J)}$

$$\begin{aligned}\mathcal{P}_J(x)&=\Pr_{y\sim\mathcal{P}}\left[y^{(J)}=x^{(J)}\right]/2^{n-k}\\&=\sum_{j\in L(i,J\cap J^*)}\Pr_{y\sim\mathcal{P}}\left[y^{(J^*)}=C_j\right]/2^{n-k+|J\setminus J^*|}\\&=\left(\Pr_{y\sim\mathcal{P}}\left[y^{(J\cap J^*)}=x^{(J\cap J^*)}\right]\right)/2^{n-k+|J\setminus J^*|}\end{aligned}$$

where the last equality comes from the fact that $L(i,J\cap J^*)$ contains all C_j's such that $C_j^{(J\cap J^*)}=x^{(J\cap J^*)}$. Thus, the proof of Equation (5) is complete.

Appendix C. Proof of Lemma 7

Lemma 21 (Step 1) *Let $\mathcal{P}$ be a k-junta distribution on the set J^* and $\mathcal{P}_J$ be a k-junta distributions defined in Equation 1. Then,*

$$\mathrm{E}[(\mathcal{P}(x) - 1/2^n)^2] - \mathrm{E}[(\mathcal{P}_J(x) - 1/2^n)^2] \geq 4 \cdot \mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_J)^2/2^{2n}. \tag{4}$$

Proof Before we go to prove the inequality, we prove the following equality

$$\mathrm{E}[\mathcal{P}_J(x)(\mathcal{P}(x) - \mathcal{P}_J(x))] = 0. \tag{7}$$

Assume we partition all x's into X_i's such that for any two vectors x_1 and x_2 in X_i then $x_1^{(J\cap J^*)} = x_2^{(J\cap J^*)}$. We prove that for each X_i, $\sum_{x\in X_i} \mathcal{P}_J(x)(\mathcal{P}(x) - \mathcal{P}_J(x))$ is zero which yields the Equation 7. By Equation 5, $\mathcal{P}_J$ is a junta distribution on the set $J \cap J^*$. Therefore, for any two vectors x_1 and x_2 in X_i for a fixed i, we have $\mathcal{P}_J(x_1) = \mathcal{P}_J(x_2)$. Thus, we just need to prove $\sum_{x\in X_i} \mathcal{P}_J(x) = \sum_{x\in X_i} \mathcal{P}(x)$ which follows from Equation 5 directly. Moreover, observe that $\mathrm{E}[\mathcal{D}(x)] = 1/2^n$ for any distribution $\mathcal{D}$. Therefore, we have

$$\begin{aligned}\mathrm{E}[(\mathcal{D}(x) - 1/2^n)^2] &= \mathrm{E}[\mathcal{D}(x)^2] + \mathrm{E}[1/2^{2n}] - 2\mathrm{E}[\mathcal{D}(x)/2^n] \\ &= \mathrm{E}[\mathcal{D}(x)^2] + 1/2^{2n} - 2/2^{2n} = \mathrm{E}[\mathcal{D}(x)^2] - 1/2^{2n} = \mathrm{E}[\mathcal{D}(x)^2] - \mathrm{E}[\mathcal{D}(x)]^2.\end{aligned} \tag{8}$$

Now, we prove (4). Since $\mathcal{P}$ and $\mathcal{P}_J$ are distributions, their pmfs satisfy $\mathrm{E}[\mathcal{P}(x)] = \mathrm{E}[\mathcal{P}_J(x)] = 1/2^n$. By this fact and linearity of expectation,

$$\begin{aligned}\mathrm{E}[(\mathcal{P}(x) - 1/2^n)^2] - \mathrm{E}[(\mathcal{P}_J(x) - 1/2^n)^2] &= \mathrm{E}[\mathcal{P}(x)^2] - \mathrm{E}[\mathcal{P}_J(x)^2] \\ &= \mathrm{E}[(\mathcal{P}(x) - \mathcal{P}_J(x) + \mathcal{P}_J(x))^2] - \mathrm{E}[\mathcal{P}_J(x)^2] \\ &= \mathrm{E}[(\mathcal{P}(x) - \mathcal{P}_J(x))^2] + 2\mathrm{E}[((\mathcal{P}_J(x)(\mathcal{P}(x) - \mathcal{P}_J(x))] + \mathrm{E}[\mathcal{P}_J(x)^2] - \mathrm{E}[\mathcal{P}_J(x)^2] \\ &= \mathrm{E}[(\mathcal{P}(x) - \mathcal{P}_J(x))^2] \geq (\mathrm{E}[|\mathcal{P}(x) - \mathcal{P}_J(x)|])^2 = 4 \cdot \mathrm{d_{TV}}(\mathcal{P}, \mathcal{P}_J)^2/2^{2n}\end{aligned}$$

where the second to last equality comes from Equation 7. ■

Appendix D. Proof of Lemma 8

Below, we first show that the Fourier coefficient $\widehat{\mathcal{P}}(S)$ of a set $S \not\subseteq J^*$ is zero. This lemma allows us to infer that it is enough to compute the low degree Fourier coefficients, because the other ones are zero. Intuitively, such a high degree S contains a coordinate that will be either zero or one each with probability a half. Therefore, the Fourier coefficient of S is zero. We prove this formally in Lemma 22. Leveraging this lemma, we prove that the values of $h(J)$ and $f(J)$ are equal in Lemma 8.

Lemma 22 *For any $J \subset [n]$, let $\mathcal{D}$ be a junta distribution with J being the set of junta coordinates. For any $S \not\subseteq J$, $\widehat{\mathcal{D}}(S)$ is zero.*

Proof Observe that J might be the empty set, in which case $\mathcal{D}$ is a uniform distribution. Since S is not a subset of J, there is a coordinate i such that i is in S but not J. Thus, the i-th coordinate in each sample x is one or zero, each with probability a half. We simply pair up all x's based on their agreement on $x^{([n]\setminus\{i\})}$ and denote a pair by (x_0, x_1). Since i is not in the junta, $\mathcal{D}(x_0) = \mathcal{D}(x_1)$. And since i is in S, $\chi_S(x_0) = -\chi_S(x_1)$. Therefore,

$$\begin{aligned}\widehat{\mathcal{D}}(S) = \frac{1}{2^n}\sum_{x\in\{0,1\}^n}\mathcal{D}(x)\cdot\chi_S(x) &= \frac{1}{2^n}\sum_{(x_0,x_1)}\left(\mathcal{D}(x_0)\cdot\chi_S(x_0)+\mathcal{D}(x_1)\cdot\chi_S(x_1)\right)\\ &= \frac{1}{2^n}\sum_{(x_0,x_1)}\left(\mathcal{D}(x_0)\cdot\chi_S(x_0)-\mathcal{D}(x_0)\cdot\chi_S(x_0)\right) = 0\end{aligned}$$

as we wanted to show. ■

Now we are ready to prove that $f(J)$ is equal to $h(J)$ for any $J \subseteq [n]$ of size k.

Lemma 23 (Step 2) *With f and h as defined in Equation (3) and Equation (2), for any $J \subset [n]$ of size k we have $f(J) = h(J)$.*

Proof By (8),

$$h(J) = 2^{2n}\cdot\mathrm{E}[(\mathcal{P}_J(x)-1/2^n)^2] = 2^{2n}\cdot\left(\mathrm{E}[\mathcal{P}_J(x)^2]-\mathrm{E}[\mathcal{P}_J(x)]^2\right) = 2^{2n}\cdot\left(\sum_S\widehat{\mathcal{P}}_J(S)^2-\widehat{\mathcal{P}}_J(\emptyset)^2\right)$$

where the last equality follows by Parseval's Theorem and the fact that

$$\widehat{\mathcal{P}_J}(\emptyset) = \frac{1}{2^n}\sum_{x\in\{0,1\}^n}\mathcal{P}_J(x)\cdot\chi_\emptyset(x) = \mathrm{E}[\mathcal{P}_J(x)].$$

In addition, note that by (5), any $\mathcal{P}_J$ is a junta distribution over the set $J \cap J^*$. By Lemma 22, for any $S \not\subseteq (J \cap J^*)$, $\widehat{\mathcal{P}}_J(S)$ is zero. Thus,

$$h(J) = 2^{2n}\cdot\left(\sum_S\widehat{\mathcal{P}}_J(S)^2-\widehat{\mathcal{P}}_J(\emptyset)^2\right) = 2^{2n}\cdot\sum_{S\subseteq(J\cap J^*),S\neq\emptyset}\widehat{\mathcal{P}}_J(S)^2.$$

Now, it is clear that $h(J^*) = f(J^*)$. Assume $J \neq J^*$. Let S be a non-empty set of $J \cap J^*$ and c be a fixed binary vector of size $|S|$. By definition of $\mathcal{P}_J$, it is not hard to see $\Pr_{x\sim\mathcal{P}}[x^{(S)} = c] = \Pr_{x\sim\mathcal{P}_J}[x^{(S)} = c]$. Thus, by conditioning over all possible c, we can prove $\Pr_{x\sim\mathcal{P}}[\chi_S(x) = b] = \Pr_{x\sim\mathcal{P}_J}[\chi_S(x) = b]$ when b is $+1$ or -1. Therefore,

$$\begin{aligned}\widehat{\mathcal{P}}_J(S) = 2^{-n}\cdot\sum_x\mathcal{P}_J(x)\chi_S(x) &= 2^{-n}\cdot\left(\Pr_{x\sim\mathcal{P}_J}[\chi_S(x)=1]-\Pr_{x\sim\mathcal{P}_J}[\chi_S(x)=-1]\right)\\ &= 2^{-n}\cdot\left(\Pr_{x\sim\mathcal{P}}[\chi_S(x)=1]-\Pr_{x\sim\mathcal{P}}[\chi_S(x)=-1]\right) = 2^{-n}\cdot\sum_x\mathcal{P}_J(x)\chi_S(x) = \widehat{\mathcal{P}}(S).\end{aligned}$$

In this way, for any non-empty subset S of $J \cap J^*$, $\widehat{\mathcal{P}}_J(S)$ is equal to $\widehat{\mathcal{P}}(S)$. By Lemma 22 for any $S \subseteq J$ which is not subset of J^* then $\widehat{\mathcal{P}}(S)$ is zero. Thus,

$$h(J) = 2^{2n} \cdot \sum_{S \subseteq (J \cap J^*), S \neq \emptyset} \widehat{\mathcal{P}}_J(S)^2 = 2^{2n} \cdot \sum_{S \subseteq (J \cap J^*), S \neq \emptyset} \widehat{\mathcal{P}}(S)^2$$
$$= 2^{2n} \cdot \left(\sum_{S \subseteq (J \cap J^*), S \neq \emptyset} \widehat{\mathcal{P}}(S)^2 + \sum_{S \subseteq J, S \not\subseteq J^*} \widehat{\mathcal{P}}(S)^2 \right) = 2^{2n} \cdot \sum_{S \subseteq (J), S \neq \emptyset} \widehat{\mathcal{P}}(S)^2 = f(J)$$

and the proof is complete. ■

Appendix E. Proof of Lemma 9

Lemma 24 (Step 3) *Let $\mathcal{P}$ be a junta distribution on the set J^* of size k. Suppose we draw $s = 72 \cdot 2^{2k} \cdot \ln(12 \min(n^k, 2^n))/\epsilon^4$ samples from $\mathcal{P}$. For any set J of size k, we estimate $f(J)$, as defined in (3), by*

$$\tilde{f}(J) = \sum_{S \subseteq J, S \neq \emptyset} \left(\frac{2 \cdot [\#\ \textit{samples}\ x\ \textit{with}\ \chi_S(x) = 1]}{s} - 1 \right)^2.$$

With probability $5/6$ all of the J's we have $|f(J) - \tilde{f}(J)| < \epsilon^2$.

Proof By the definition of Fourier coefficients, we have

$$f(J) = 2^{2n} \cdot \sum_{S \subseteq J, S \neq \emptyset} \widehat{\mathcal{P}}(S)^2 = \sum_{S \subseteq J, S \neq \emptyset} \left(\sum_x \mathcal{P}(x) \chi_S(x)) \right)^2$$
$$= \sum_{S \subseteq J, S \neq \emptyset} \left(\Pr_{x \sim \mathcal{P}}[\chi_S(x) = 1] - \Pr_{x \sim \mathcal{P}}[\chi_S(x) = -1] \right)^2 = \sum_{S \subseteq J, S \neq \emptyset} \left(2 \Pr_{x \sim \mathcal{P}}[\chi_S(x) = 1] - 1 \right)^2.$$

Now for abbreviation, let $P_S = 2 \cdot \Pr[\chi_S(x) = 1] - 1$ and let $\tilde{P}_S$ be $2 \cdot [\#$ samples x with $\chi_S(x) = 1]/s - 1$. First, notice that $\tilde{P}_S$ is an estimator for P_S such that their difference is not likely to be more than $\epsilon' = \epsilon^2/(6 \cdot 2^k)$, because by the Hoeffding bound we have

$$\Pr\left[\left| P_S - \tilde{P}_S \right| > \epsilon' \right] = \Pr\left[\left| \frac{[\#\ \text{samples}\ x\ \text{with}\ \chi_S(x) = 1]}{s} - \Pr_{x \sim \mathcal{P}}[\chi_S(x) = 1] \right| > \frac{\epsilon'}{2} \right] \leq 2 \cdot e^{-\frac{s\epsilon'^2}{2}}.$$

Note that in the Algorithm 1 we estimate this value for all subsets of size at most k. It is well known that $\sum_{i=1}^{k} \binom{n}{i} \leq \min(n^k, 2^n)$. Thus, there are at most $\min(n^k, 2^n)$ many sets. Therefore, for $s \geq 2 \ln(12 \min(n^k, 2^n))/\epsilon'^2$, it is not hard to see that the probability of estimating at least one P_S inaccurately is at most $1/6$ by the union bound. We can assume $\alpha_S = P_S - \tilde{P}_S$ is in the range $[-\epsilon', \epsilon']$ with probability $5/6$. The maximum error of $\tilde{f}(J)$ is

Algorithm 4 An ϵ-test for testing uniformity of a bucket

Input: s_i samples from each $\mathcal{P}_i$ in the collection.
$Y \leftarrow$ the number of unique elements in the samples.
For each sample (i, x), replace x with x' uniformly chosen from $[n]$.
$\lambda \leftarrow \frac{\epsilon^2 \cdot s^2}{c^2 \cdot m \cdot n}$
$Y' \leftarrow$ the number of unique elements in these s samples.
if $|Y - Y'| > \lambda/2$, **then**
  return Reject
else
  return Accept
end if

bounded by

$$|f(J) - \tilde{f}(J)| = \left| \sum_{S \subseteq J, S \neq \emptyset} P_S^2 - \sum_{S \subseteq J, S \neq \emptyset} \tilde{P}_S^2 \right| = \left| \sum_{S \subseteq J, S \neq \emptyset} (2P_S \alpha_S - \alpha_S^2) \right|$$
$$\leq \sum_{S \subseteq J, S \neq \emptyset} (2P_S |\alpha_S| + \alpha_S^2) < 2^k (2\epsilon' + \epsilon'^2) < \epsilon^2/2$$

where the last inequality follows by $\epsilon' = \epsilon^2/(6 \cdot 2^k) < 1$. ■

Appendix F. Uniformity Test within a Bucket

In this section, we provide a uniformity test for a collection of distributions when the weights are bounded. In other words, the algorithm distinguishes whether the weighted distance is zero or at least ϵ. Our algorithm is based on counting the number of unique elements which is also negatively related to the number of the coincidences. This idea was proposed before in Paninski (2008); Batu et al. (2000) for uniformity test of a single distribution. The high level idea is to estimate the expected value of the number of unique elements when the underlying collection is an unknown collection and compare that value to the case when it is a collection of uniform distributions. If these values are close enough to each other we can infer that the unknown collection is actually a collection of uniform distributions. Otherwise it is not. More formally, we represent Algorithm 4 and prove its correctness in the following theorem.

Theorem 25 *Assume we have a collection of distributions $\mathcal{C} = \{\mathcal{P}_i | w_i\}_{i=1}^m$. We draw* $\mathrm{Poi}(s)$ *samples from $\mathcal{C}$. Assume the following conditions hold.*

- *For all $i \in [m]$, w_i is at most T.*
- *The number of samples from $\mathcal{P}_i$, namely s_i, is in $[w_i \cdot s/c, cw_i \cdot s]$ for a constant c.*
- *s is at least $c^3 m \sqrt{30Tn}/\epsilon^2$.*

Then, Algorithm 4 distinguishes whether $\mathcal{C}$ is a collection of uniform distributions or it is ϵ-far from it with probability $2/3$.

Proof Let Y be a random variable that indicates the number of unique elements in a set of samples drawn from $\mathcal{P}$ which is the underlying distribution over $[m] \times [n]$. Notice that we consider each sample as an ordered pair (i, x) which means x is drawn based on $\mathcal{P}_i$. Thus, (i, x) is not equal to (j, x). Similarly, let Y_i denote the number of unique elements from distribution $\mathcal{P}_i$. It is not hard to see $\sum_{i=1}^{m} Y_i = Y$. We use $\mathrm{E}_{\{\mathcal{P}_i\}}[Y]$ to denote the expected value of Y when samples are drawn from the underlying collection $\{\mathcal{P}_i|w_i\}_{i=1}^{m}$. In addition, we denote the expected value of Y by $\mathrm{E}_{\{\mathcal{U}\}}[Y]$ when the underlying collection is a set of uniform distributions with the same weights as $\mathcal{P}$ (i.e. $\{\mathcal{U}|w_i\}_{i=1}^{m}$).

Now we need to answer this question: Does the number of unique elements indicate that the collection is a set of uniform distribution or not? The answer is Yes. We show that $\mathrm{E}_{\{\mathcal{P}_i\}}[Y]$ is smaller than $\mathrm{E}_{\{\mathcal{U}\}}[Y]$ if the collection is far from being a collection of uniform distributions. Therefore, if we see a meaningful difference between $\mathrm{E}_{\{\mathcal{P}_i\}}[Y]$ and $\mathrm{E}_{\{\mathcal{U}\}}[Y]$, we can conclude $\{\mathcal{P}_i|w_i\}_{i=1}^{m}$ is not a collection of uniform distributions. For a single distribution $\mathcal{P}$, in Paninski (2008), Paninski showed that the difference $\mathrm{E}_{\mathcal{P}}[Y]$ and $\mathrm{E}_{\mathcal{U}}[Y]$ is related to the distance between $\mathcal{P}$ and the uniform distribution.

$$\mathrm{E}_{\mathcal{U}}[Y] - \mathrm{E}_{\mathcal{P}}[Y] \geq \frac{s^2 \cdot (d_{L_1}(\mathcal{P},\ \mathcal{U}))^2}{n}.$$

Since we are looking for $\mathrm{E}_{\{\mathcal{U}\}}[Y]$ (not $\mathrm{E}_{\mathcal{U}_{[m]\times[n]}}[Y]$), we can not use this inequality directly over the domain $[m] \times [n]$. However, we use this inequality for each $\mathcal{P}_i$ separately. Observe that the way that we convert the samples allows us to get the same number of samples, s_i from $\mathcal{P}_i$ and $\mathcal{U}$ over the domain of size n. Thus, we can use the above inequality separately. Hence, by linearity of expectation and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y] &= \sum_{i=1}^{m} (\mathrm{E}_{\mathcal{U}}[Y_i] - \mathrm{E}_{\mathcal{P}_i}[Y_i]) \geq \frac{1}{n} \sum_{i=1}^{m} (s_i \cdot d_{L_1}(\mathcal{P}_i,\ \mathcal{U}))^2 \\ &\geq \frac{s^2}{c^2 \cdot n} \sum_{i=1}^{m} (w_i \cdot d_{L_1}(\mathcal{P}_i,\ \mathcal{U}))^2 \geq \frac{s^2}{c^2 \cdot n \cdot m} \left(\sum_{i=1}^{m} w_i \cdot d_{L_1}(\mathcal{P}_i,\ \mathcal{U}) \right)^2\end{aligned}$$

where the first inequality follows from Paninski (2008), the second follows from that $s_i \in [w_i \cdot s/c, cw_i \cdot s]$. Set λ to $(4s^2\epsilon^2)/(c^2 \cdot m \cdot n)$. Therefore, if $\mathcal{C}$ is ϵ-far from being a collection of uniform distributions, then

$$\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y] \geq \frac{4s^2 \cdot \epsilon^2}{c^2 n \cdot m} = \lambda, \tag{9}$$

because the weighted L_1 distance is at least 2ϵ. However, these two expected values cannot be calculated directly since the w_i's and $\mathcal{P}_i$'s are unknown. Thus, we need to estimate them. By definition, the number of unique elements in s samples, Y, is an unbiased estimator for $\mathrm{E}_{\{\mathcal{P}_i\}}[Y]$. To estimate $\mathrm{E}_{\{\mathcal{U}\}}[Y]$, we reuse the samples we get from the collection and change each sample (i, x) to (i, x') where x' is chosen uniformly at random from $[n]$. Since i and x' are picked with probability w_i and $1/n$ respectively, we can assume the sample (i, x') is

drawn from the collection $\{\mathcal{U}|w_i\}_{i=1}^m$. Therefore, the number of unique elements in the new sample set, namely Y', is an unbiased estimator for $\mathrm{E}_{\{\mathcal{U}\}}[Y]$. Below, we formally prove that the number of unique elements, Y, (and Y') cannot be far from their own expected value using the Chebyshev's inequality. To do so, we need to bound the variance.

Lemma 26 *If the constraints of Theorem 25 hold, then*

$$\mathrm{Var}[Y] \leq \mathrm{E}_{\mathcal{U}}[Y] - \mathrm{E}_{\mathcal{P}_i}[Y] + \frac{c^2 s^2 T}{n}.$$

Proof Bounding the variance of the number of unique elements has been studied in Paninski (2008). Paninski showed the following inequality

$$\mathrm{Var}_{\mathcal{P}}[Y] \leq \mathrm{E}_{\mathcal{U}}[Y] - \mathrm{E}_{\mathcal{P}}[Y] + \frac{s_i^2}{n}$$

Here, since we know the s_i's are independent (by using the standard Poissonization method), we have

$$\mathrm{Var}[Y] = \sum_{i=1}^{m} \mathrm{Var}[Y_i] \leq \sum_{i=1}^{m} \left(\mathrm{E}_{\mathcal{U}}[Y_i] - \mathrm{E}_{\mathcal{P}_i}[Y_i] + \frac{s_i^2}{n} \right) \leq \mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y] + \frac{c^2 s^2}{n} \left(\sum_{i=1}^{m} w_i^2 \right).$$

On the other hand, it is not hard to see that since w_i's are less than T, we have

$$\sum_i w_i^2 \leq \left(\sum_i w_i \right) \cdot T \leq T.$$

Combining the two above inequalities we get $\mathrm{Var}[Y] \leq \mathrm{E}_{\mathcal{U}}[Y] - \mathrm{E}\mathcal{P}_i[Y] + \frac{c^2 s^2 T}{n}$. ■

Now, we are ready to use the Chebyshev's inequality to prove that we are able to estimate Y accurately. Below we consider two cases based on the underlying collection.

- **Case 1: $\mathcal{C}$ is a collection of uniform distribution:** In this case $\mathrm{E}_{\{\mathcal{P}_i\}}[Y]$ is equal to $\mathrm{E}_{\{\mathcal{U}\}}[Y]$, so by Lemma 26 the variance of Y is at most $c^2 s^2 T/n$. Thus by the Chebyshev's inequality we have

  $$\Pr[|Y - E_{\{\mathcal{U}\}}[Y]| \geq \lambda/4] \leq 16\mathrm{Var}[Y]/\lambda^2 \leq \frac{c^6\, T\, n\, m^2}{s^2 \epsilon^4}.$$

  It is not hard to see that for $s \geq c^3 m\sqrt{6Tn}/\epsilon^2$ the above probability is at most $1/6$. Similar to Y, we can prove that the probability that Y' is $\lambda/4$ far away from its mean is less than $1/6$. Therefore, $Y' - Y$ is at most $\lambda/2$ with probability at least $1 - 1/3$.

- **Case 2: $\mathcal{C}$ is ϵ-far from being a collection of uniform distributions:**. Therefore by Equation 9, $\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y]$ is at least λ. Similar to the above, we use the

Chebyshev's inequality. By Lemma 26 and (9) we have

$$\begin{aligned}\Pr\left[|Y - \mathrm{E}_{\{\mathcal{P}_i\}}[Y]| \geq \tfrac{1}{4}(\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y])\right] &\leq \mathrm{Var}[Y]/(\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y]/4)^2 \\ &\leq \frac{\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y] + c^3 s^2 T/n}{(\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y]/4)^2} \\ &\leq \frac{16}{\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y]} + \frac{16c^2 s^2 T}{n \cdot (\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y])^2} \\ &\leq \frac{16}{\lambda} + \frac{16c^2 s^2 T}{n \cdot \lambda^2} \leq \frac{4\,c^2\,n\,m}{s^2\epsilon^2} + \frac{c^6\,T\,n\,m^2}{s^2\epsilon^4} < \frac{5\,c^6\,T\,n\,m^2}{s^2\epsilon^4}.\end{aligned}$$

Note that T by definition can not be less than $1/m$ that's why the last inequality is true. It is straightforward that for $s \geq c^3 m\sqrt{30Tn}/\epsilon^2$ the above probability is at most $1/6$. On the other hand, similar to what we had in case one, Y' cannot go far from its mean too. Thus,

$$\Pr[|Y' - E_{\{\mathcal{U}\}}[Y]| \geq \frac{1}{4}(\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y])] \leq \Pr[|Y' - E_{\{\mathcal{U}\}}[Y]| \geq \lambda/4] \leq 1/6.$$

Therefore, $Y' - Y$ is at least $(\mathrm{E}_{\{\mathcal{U}\}}[Y] - \mathrm{E}_{\{\mathcal{P}_i\}}[Y])/2 \geq \lambda/2$ with probability at least $1 - 1/3$.

In both cases, the uniformity test outputs the correct answer with probability at least $2/3$. ■

Corollary 27 *Assume we have a collection of distributions $\mathcal{C} = \{\mathcal{P}_i | w_i\}_{i=1}^m$. Assume for all $i \in [m]$, w_i is in $[1/8m, 8/m]$. Then, there exists an algorithm that distinguishes whether $\mathcal{C}$ is a collection of uniform distributions or it is ϵ-far from being a collection of uniform distributions with probability $1 - \delta$ using $O(\sqrt{m\,n}\log(\delta^{-1})/\epsilon^2 + \log(\delta^{-1})m\log m)$ samples.*

Proof Let $\mathcal{S} = \max(c^3 m\sqrt{30Tn}/\epsilon^2, 40m\log(12(m+1))$. Let s be a sample drawn from $\mathrm{Poi}(\mathcal{S})$. Using Lemma 12 and since all the $w_i \geq 1/8m$, $\hat{w_i}$ is in the range $[w_i/2, 2w_i]$ for all i's; and $s \geq 2\mathcal{S}$ with probability $5/6$. Now, using Theorem 25 and setting $T = 8/m$ and $c = 2$, with probability $\frac{5}{6} \cdot \frac{2}{3} = 5/9$, we can distinguish whether $\mathcal{C}$ is a collection of uniform distributions or not. Using standard amplification, by repeating this argument $O(\log \delta^{-1})$ times, the algorithm answers correctly with probability at least $1 - \delta$. Moreover, the sample complexity is $O(\sqrt{m\,n}\log(\delta^{-1})/\epsilon^2 + \log(\delta^{-1})m\log m)$. ■