Efficient Density Estimation via Piecewise Polynomial Approximation

Citation for published version:
Chan, S-O, Diakonikolas, I, Servedio, RA & Sun, X 2014, Efficient Density Estimation via Piecewise Polynomial Approximation. in Proceedings of the 46th Annual ACM Symposium on Theory of Computing. STOC '14, ACM, New York, NY, USA, pp. 604-613. DOI: 10.1145/2591796.2591848

Digital Object Identifier (DOI):
10.1145/2591796.2591848

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 46th Annual ACM Symposium on Theory of Computing

Efficient Density Estimation via Piecewise Polynomial Approximation

Siu-On Chan[*]
UC Berkeley
siuon@cs.berkeley.edu.

Ilias Diakonikolas[†]
University of Edinburgh
ilias.d@ed.ac.uk.

Rocco A. Servedio[‡]
Columbia University
rocco@cs.columbia.edu.

Xiaorui Sun[§]
Columbia University
xiaoruisun@cs.columbia.edu.

May 14, 2013

Abstract

We give a highly efficient "semi-agnostic" algorithm for learning univariate probability distributions that are well approximated by piecewise polynomial density functions. Let p be an arbitrary distribution over an interval I which is τ-close (in total variation distance) to an unknown probability distribution q that is defined by an unknown partition of I into t intervals and t unknown degree-d polynomials specifying q over each of the intervals. We give an algorithm that draws $\tilde{O}(t(d+1)/\varepsilon^2)$ samples from p, runs in time $\mathrm{poly}(t, d, 1/\varepsilon)$, and with high probability outputs a piecewise polynomial hypothesis distribution h that is $(O(\tau)+\varepsilon)$-close (in total variation distance) to p. This sample complexity is essentially optimal; we show that even for $\tau = 0$, any algorithm that learns an unknown t-piecewise degree-d probability distribution over I to accuracy ε must use $\Omega(\frac{t(d+1)}{\mathrm{poly}(1+\log(d+1))} \cdot \frac{1}{\varepsilon^2})$ samples from the distribution, regardless of its running time. Our algorithm combines tools from approximation theory, uniform convergence, linear programming, and dynamic programming.

We apply this general algorithm to obtain a wide range of results for many natural problems in density estimation over both continuous and discrete domains. These include state-of-the-art results for learning mixtures of log-concave distributions; mixtures of t-modal distributions; mixtures of Monotone Hazard Rate distributions; mixtures of Poisson Binomial Distributions; mixtures of Gaussians; and mixtures of k-monotone densities. Our general technique yields computationally efficient algorithms for all these problems, in many cases with provably optimal sample complexities (up to logarithmic factors) in all parameters.

[*] Supported by NSF award DMS-1106999, DOD ONR grant N000141110140 and NSF award CCF-1118083.

[†] Part of this work was done while the author was at UC Berkeley supported by a Simons Postdoctoral Fellowship.

[‡] Supported by NSF grants CCF-0915929 and CCF-1115703.

[§] Supported by NSF grant CCF-1149257.

1 Introduction

Over the past several decades, many works in computational learning theory have addressed the general problem of learning an unknown Boolean function from labeled examples. A recurring theme that has emerged from this line of work is that state-of-the-art learning results can often be achieved by analyzing *polynomials* that compute or approximate the function to be learned, see e.g. [LMN93, KM93, Jac97, KS04, MOS04, KOS04].

In the current paper we show that this theme extends to the well-studied unsupervised learning problem of *density estimation*; namely, learning an unknown *probability distribution* given i.i.d. samples drawn from the distribution. We propose a new approach to density estimation based on establishing the existence of *piecewise polynomial density functions* that approximate the distributions to be learned. The key tool that enables this approach is a new and highly efficient general algorithm that we provide for learning univariate probability distributions that are well approximated by piecewise polynomial density functions. Combining our general algorithm with structural results showing that probability distributions of interest can be well approximated using piecewise polynomial density functions, we obtain learning algorithms for those distributions.

We demonstrate the efficacy of this approach by showing that for many natural and well-studied types of distributions, there do indeed exist piecewise polynomial densities that approximate the distributions to high accuracy. For all of these types of distributions our general approach gives a state-of-the-art computationally efficient learning algorithm with the best known sample complexity (number of samples that are required from the distribution) to date; in many cases the sample complexity of our approach is provably optimal, up to logarithmic factors in the optimal sample complexity.

1.1 Related work.

Density estimation is a well-studied topic in probability theory and statistics (see [DG85, Sil86, Sco92, DL01] for book-length introductions). There is a number of generic techniques for density estimation in the mathematical statistics literature, including histograms, kernels (and variants thereof), nearest neighbor estimators, orthogonal series estimators, maximum likelihood (and variants thereof) and others (see Chapter 2 of [Sil86] for a survey of existing methods). In recent years, theoretical computer science researchers have also studied density estimation problems, with an explicit focus on obtaining *computationally efficient* algorithms (see e.g. [KMR$^+$94, FM99, FOS05, BS10, KMV10, MV10, DDS12a, DDS12b].

We work in a PAC-type model similar to that of [KMR$^+$94] and to well-studied statistical frameworks for density estimation. The learning algorithm has access to i.i.d. draws from an unknown probability distribution p. It must output a hypothesis distribution h such that with high probability the total variation distance $d_{\mathrm{TV}}(p, h)$ between p and h is at most ε. (Recall that the total variation distance between two distributions p and h is $\frac{1}{2}\int |p(x) - h(x)|dx$ for continuous distributions, and is $\frac{1}{2}\sum |p(x) - h(x)|$ for discrete distributions.) We shall be centrally concerned with obtaining learning algorithms that both use few samples and are computationally efficient.

The previous work that is most closely related to our current paper is the recent work [CDSS13]. (That paper dealt with distributions over the discrete domain $[n] = \{1, \ldots, n\}$, but since the current work focuses mostly on the continuous domain, in our description of the [CDSS13] results below we translate them to the continuous domain. This translation is straightforward.) To describe the main result of [CDSS13] we need to introduce the notions of *mixture distributions* and *piecewise constant* distributions. Given distributions $p_1, \ldots, p_k$ and non-negative values $\mu_1, \ldots, \mu_k$ that sum to 1, we say that $p = \sum_{i=1}^{k} \mu_i p_i$ is a *k-mixture* of *components* $p_1, \ldots, p_k$ with *mixing weights* $\mu_1, \ldots, \mu_k$. A draw from p is obtained by choosing $i \in [k]$ with probability μ_i and then making a draw from p_i. A distribution q over an interval I is (ε, t)*-piecewise constant* if there is a partition of

I into t disjoint intervals $I_1, \ldots, I_t$ such that p is ε-close (in total variation distance) to a distribution q such that $q(x) = c_j$ for all $x \in I_j$ for some $c_j \geq 0$.

The main result of [CDSS13] is an efficient algorithm for learning any k-mixture of (ε, t)-piecewise constant distributions:

Theorem 1. *There is an algorithm that learns any k-mixture of (ε, t)-piecewise constant distributions over an interval I to accuracy $O(\varepsilon)$, using $O(kt/\varepsilon^3)$ samples and running in $\tilde{O}(kt/\varepsilon^3)$ time.*[1]

1.2 Our main result. As our main algorithmic contribution, we give a significant strengthening and generalization of Theorem 1 above. First, we improve the ε-dependence in the sample complexity of Theorem 1 from $1/\varepsilon^3$ to a near-optimal $\tilde{O}(1/\varepsilon^2)$. [2] Second, we extend Theorem 1 from piecewise constant distributions to *piecewise polynomial* distributions. More precisely, we say that a distribution over an interval I is *(ε, t)-piecewise degree-d* if there is a partition of I into t disjoint intervals $I_1, \ldots, I_t$ such that p is ε-close (in total variation distance) to a distribution q such that $q(x) = q_j(x)$ for all $x \in I_j$, where each of $q_1, \ldots, q_t$ is a univariate degree-d polynomial.[3] (Note that being (ε, t)-piecewise constant is the same as being (ε, t)-piecewise degree-0.) We say that such a distribution q is a *t-piecewise degree-d distribution.*

Our main algorithmic result is the following (see Theorem 23 for a fully detailed statement of the result):

Theorem 2. *[Informal statement] There is an algorithm that learns any k-mixture of (ε, t)-piecewise degree-d distributions over an interval I to accuracy $O(\varepsilon)$, using $\tilde{O}((d+1)kt/\varepsilon^2)$ samples and running in* $\mathrm{poly}((d+1), k, t, 1/\varepsilon)$ *time.*

As we describe below, the applications that we give for Theorem 2 crucially use both aspects in which it strengthens Theorem 1 (degree d rather than degree 0, and $\tilde{O}(1/\varepsilon^2)$ samples rather than $O(1/\varepsilon^3)$) to obtain near-optimal sample complexities.

A different view on our main result, which may also be illuminating, is that it gives a "semi-agnostic" algorithm for learning piecewise polynomial densities. (Since any k-mixture of t-piecewise degree-d distributions is easily seen to be a kt-piecewise degree-d distribution, we phrase the discussion below only in terms of t-piecewise degree-d distributions rather than mixtures.) Let $\mathcal{P}_{t,d}(I)$ denote the class of all t-piecewise degree-d distributions over interval I. Let p be any distribution over I. Our algorithm, given parameters t, d, ε and $\tilde{O}(t(d+1)/\varepsilon^2)$ samples from p, outputs an $O(t)$-piecewise degree-d hypothesis distribution h such that $d_{\mathrm{TV}}(p, h) \leq 4\mathrm{opt}_{t,d}(1+\varepsilon) + \varepsilon$, where

$$\mathrm{opt}_{t,d} := \inf_{r \in \mathcal{P}_{t,d}(I)} d_{\mathrm{TV}}(p, r).$$

(See Theorem 25.)

We prove the following lower bound (see Theorem 8 for a precise statement), which shows that the number of samples that our algorithm uses is optimal up to logarithmic factors:

[1] Here and throughout the paper we work in a standard unit-cost model of computation, in which a sample from distribution p is obtained in one time step (and is assumed to fit into one register) and basic arithmetic operations are assumed to take unit time. Our algorithms, like the [CDSS13] algorithm, only performs basic arithmetic operations on "reasonable" inputs.

[2] Recall the well-known fact that $\Omega(1/\varepsilon^2)$ samples are required for essentially every nontrivial distribution learning problem. In particular, any algorithm that distinguishes the uniform distribution over $[-1, 1]$ from the piecewise constant distribution with pdf $p(x) = \frac{1}{2}(1-\varepsilon)$ for $-1 \leq x \leq 0$, $p(x) = \frac{1}{2}(1+\varepsilon)$ for $0 < x \leq 1$, must use $\Omega(1/\varepsilon^2)$ samples.

[3] Here and throughout the paper, whenever we refer to a "degree-d polynomial," we mean a polynomial of degree at most d.

Theorem 3. *[Informal statement] Any algorithm that learns an unknown t-piecewise degree-d distribution q over an interval I to accuracy ε must use $\Omega(\frac{t(d+1)}{\mathrm{poly}(1+\log(d+1))} \cdot \frac{1}{\varepsilon^2})$ samples.*

Note that the lower bound holds even when the unknown distribution is exactly a t-piecewise degree-d distribution, i.e. $\mathrm{opt}_{t,d} = 0$ (in fact, the lower bound still applies even if the $t-1$ "breakpoints" defining the t interval boundaries within I are fixed to be evenly spaced across I).

1.3 Applications of Theorem 2. Using Theorem 2 we obtain highly efficient algorithms for a wide range of specific distribution learning problems over both continuous and discrete domains. These include learning mixtures of log-concave distributions; mixtures of t-modal distributions; mixtures of Monotone Hazard Rate distributions; mixtures of Poisson Binomial Distributions; mixtures of Gaussians; and mixtures of k-monotone densities. (See Table 1 for a concise summary of these results and a comparison with previous results.) All of our algorithms run in polynomial time in all of the relevant parameters, and for all of the mixture learning problems listed in Table 1, our results improve on previous state-of-the-art results by a polynomial factor. (In some cases, such as t-piecewise degree-d polynomial distributions and mixtures of t bounded k-monotone distributions, we believe that we give the first nontrivial learning results for the distribution classes in question.) In many cases the sample complexities of our algorithms are provably optimal, up to logarithmic factors in the optimal sample complexity. Detailed descriptions of all of the classes of distributions in the table, and of our results for learning those distributions, are given in Section 4.

We note that all the learning results indicated with theorem numbers in Table 1 (i.e. results proved in this paper) are in fact *semi-agnostic* learning results for the given classes as described in the previous subsection; hence all of these results are highly robust even if the target distribution does not exactly belong to the specified class of distributions. More precisely, if the target distribution is τ-close to some member of the specified class of distributions, then the algorithm uses the stated number of samples and outputs a hypothesis that is $(O(\tau) + \varepsilon)$ close to the target distribution.

1.4 Our Approach and Techniques. As stated in [Sil86], "the oldest and most widely used density estimator is the histogram": Given samples from a density f, the method partitions the domain into a number of intervals (bins) $I_1, \ldots, I_k$, and outputs the empirical density which is constant within each bin. Note that the number k of bins and the width of each bin are parameters and may depend on the particular class of distributions being learned. Our proposed technique may naturally be viewed as a very broad generalization of the histogram method, where instead of approximating the distribution by a *constant* within each bin, we approximate it by a *low-degree polynomial.* We believe that such a generalization is very natural; the recent paper [PA13] also proposes using splines for density estimation. (However, this is not the main focus of the paper and indeed [PA13] does not provide or analyze algorithms for density estimation.) Our generalization of the histogram method seems likely to be of wide applicability. Indeed, as we show in this paper, it can be used to obtain many computationally efficient learners for a wide class of concrete learning problems, yielding several new and nearly optimal results.

The general algorithm. At a high level, our algorithm uses a rather subtle dynamic program (roughly, to discover the "correct" intervals in each of which the underlying distribution is close to a degree-d polynomial) and linear programming (roughly, to learn a single degree-d sub-distribution on a given interval). We note, however, that many challenges arise in going from this high-level intuition to a working algorithm.

Consider first the special case in which there is only a single known interval (see Section 3.3). In this special case our problem is somewhat reminiscent of the problem of learning a "noisy

Class of Distributions	**Number of samples**	**Reference**
Continuous distributions over an interval I		
t-piecewise constant	$O(t/\epsilon^3)$	[CDSS13]
t-piecewise constant	$\tilde{O}(t/\epsilon^2)$ (†)	Theorem 23
t-piecewise degree-d polynomial	$\tilde{O}(td/\epsilon^2)$ (†)	Theorem 23, Theorem 8
log-concave	$O(1/\varepsilon^{5/2})$ (†)	folklore [DL01]
mixture of k log-concave distributions	$\tilde{O}(k/\varepsilon^{5/2})$ (†)	Theorem 26
mixture of t bounded 1-monotone distributions	$\tilde{O}(t/\epsilon^3)$ (†)	Theorem 33
mixture of t bounded 2-monotone distributions	$\tilde{O}(t/\epsilon^{5/2})$ (†)	Theorem 33
mixture of t bounded k-monotone distributions	$\tilde{O}(tk/\epsilon^{2+1/k})$	Theorem 33
mixture of k Gaussians	$\tilde{O}(k/\epsilon^2)$ (†)	Corollary 37
Discrete distributions over $\{1, 2, \ldots, N\}$		
t-modal	$\tilde{O}(t \log(N)/\epsilon^3) + \tilde{O}(t^3/\varepsilon^3)$	[DDS12a]
mixture of k t-modal distributions	$O(kt \log(N)/\epsilon^4)$	[CDSS13]
mixture of k t-modal distributions	$\tilde{O}(kt \log(N)/\epsilon^3)$ (†)	Theorem 39
mixture of k monotone hazard rate distributions	$\tilde{O}(k \log(N)/\epsilon^4)$	[CDSS13]
mixture of k monotone hazard rate distributions	$\tilde{O}(k \log(N)/\epsilon^3)$ (†)	Theorem 40
mixture of k log-concave distributions	$\tilde{O}(k/\epsilon^4)$	[CDSS13]
mixture of k log-concave distributions	$\tilde{O}(k/\epsilon^3)$	Theorem 41
Poisson Binomial Distribution	$\tilde{O}(1/\epsilon^3)$	[DDS12b, CDSS13]
mixture of k Poisson Binomial Distributions	$\tilde{O}(k/\epsilon^4)$	[CDSS13]
mixture of k Poisson Binomial Distributions	$\tilde{O}(k/\epsilon^3)$	Theorem 41

Table 1: Known algorithmic results for learning various classes of probability distributions. "Number of samples" indicates the number of samples that the algorithm uses to learn to total variation distance ε. Results given in this paper are indicated with a reference to the corresponding theorem. A (†) indicates that the given upper bound on sample complexity is known to be optimal up to at most logarithmic factors (i.e. "$\tilde{O}(m)$ (†)" means that there is a known lower bound of $\Omega(m)$).

polynomial" that was studied by Arora and Khot [AK03]. We stress, though, that our setting is considerably more challenging in the following sense: in the [AK03] framework, each data point is a pair (x, y) where y is assumed to be close to the value $p(x)$ of the target polynomial at x. In our setting the input data is *unlabeled* – we only get points x drawn from a *distribution* that is τ-close to some polynomial pdf. However, we are able to leverage some ingredients from [AK03] in our context. We carry out a careful error analysis using probabilistic inequalities (the VC inequality and tail bounds) and ingredients from basic approximation theory to show that $\tilde{O}(d/\varepsilon^2)$ samples suffice for our linear program to achieve an $O(\mathrm{opt}_{1,d} + \varepsilon)$-accurate hypothesis with high probability.

Additional challenges arise when we go from a single interval to the general case of t-piecewise polynomial densities (see Section 3.4). The "correct" intervals can of course only be approximated rather than exactly identified, introducing an additional source of error that needs to be carefully managed. We formulate a dynamic program that uses the algorithm from Section 3.3 as a "black box" to achieve our most general learning result.

The applications. Given our general algorithm, in order to obtain efficient learning algorithms for specific classes of distributions, it is sufficient to establish the existence of piecewise polynomial (or piecewise constant) approximations to the distributions that are to be learned. In some cases such existence results were already known; for example, Birgé [Bir87b] provides the necessary existence result that we require for discrete t-modal distributions, and classical results in approximation theory [Dud74, Nov88] give the necessary existence results for concave distributions over continuous domains. For log-concave densities over continuous domains, we prove a new structural result on approximation by piecewise linear densities (Lemma 27) which, combined with our general algorithm, leads to an optimal learning algorithm for (mixtures of) such densities. Finally, for *k-monotone* distributions we are able to leverage a recent (and quite sophisticated) result from the approximation theory literature [KL04, KL07] to obtain the required approximation result.

Structure of this paper: In Section 2 we include some basic preliminaries. In Section 3 we present our main learning result and in Section 4 we describe our applications.

2 Preliminaries

Throughout the paper for simplicity we consider distributions over the interval $[-1,1)$. It is easy to see that the general results given in Section 3 go through for distributions over an arbitrary interval I. (In the applications given in Section 4 we explicitly discuss the different domains over which our distributions are defined.)

Given a value $\kappa > 0$, we say that a distribution p over $[-1,1)$ is *κ-well-behaved* if $\sup_{x\in[-1,1)} \Pr_{x\sim p}[x] \leq \kappa$, i.e. no individual real value is assigned more than κ probability under p. Any probability distribution with no atoms (and hence any piecewise polynomial distribution) is κ-well-behaved for all $\kappa > 0$, but for example the distribution which outputs the value 0.3 with probability $1/100$ and otherwise outputs a uniform value in $[-1,1)$ is only κ-well-behaved for $\kappa \geq 1/100$. Our results apply for general distributions over $[-1,1)$ which may have an atomic part as well as a non-atomic part.

Throughout the paper we assume that the density p is measurable. Note that throughout the paper we only ever work with the probabilities $\Pr_{x\sim p}[x = z]$ of single points and probabilities $\Pr_{x\sim p}[x \in S]$ of sets S that are finite unions of intervals and single points.

Given a function $p : I \to \mathbb{R}$ on an interval $I \subseteq [-1,1)$ and a subinterval $J \subseteq I$, we write $p(J)$ to denote $\int_J p(x)dx$. Thus if p is the pdf of a probability distribution over $[-1,1)$, the value $p(J)$ is the probability that distribution p assigns to the subinterval J. We sometimes refer to a function p over an interval (which need not necessarily integrate to 1 over the interval) as a "subdistribution."

Given m independent samples $s_1, \ldots, s_m$, drawn from a distribution p over $[-1,1)$, the *empirical distribution* $\widehat{p}_m$ over $[-1,1)$ is the discrete distribution supported on $\{s_1, \ldots, s_m\}$ defined as follows: for all $z \in [-1,1)$, $\Pr_{x\sim\widehat{p}_m}[x = z] = |\{j \in [m] \mid s_j = x\}|/m$.

Optimal piecewise polynomial approximators. Fix a distribution p over $[-1,1)$. We write $\mathrm{opt}_{t,d}$ to denote the value

$$\mathrm{opt}_{t,d} := \inf_{r\in\mathcal{P}_{t,d}([-1,1))} d_{\mathrm{TV}}(p, r).$$

Standard closure arguments can be used to show that the above infimum is attained by some $r \in \mathcal{P}_{t,d}([-1,1))$; however this is not actually required for our purposes. It is straightforward to verify that any distribution $\tilde{r} \in \mathcal{P}_{t,d}([-1,1))$ such that $d_{\mathrm{TV}}(p,\tilde{r})$ is at most (say) $\mathrm{opt}_{t,d} + \varepsilon/100$ is sufficient for all our arguments.

Refinements. Let $\mathcal{I} = \{I_1, \ldots, I_s\}$ be a partition of $[-1,1)$ into s disjoint intervals, and $\mathcal{J} = \{J_1, \ldots, J_t\}$ be a partition of $[-1,1)$ into t disjoint intervals. We say that $\mathcal{J}$ is a *refinement* of $\mathcal{I}$ if each interval in $\mathcal{I}$ is a union of intervals in $\mathcal{J}$, i.e. for every $a \in [s]$ there is a subset $S_a \subseteq [t]$ such that $I_a = \cup_{b\in S_a} J_b$.

For $\mathcal{I} = \{I_i\}_{i=1}^r$ and $\mathcal{I}' = \{I'_i\}_{i=1}^s$ two partitions of $[-1,1)$ into r and s intervals respectively, we say that the *common refinement* of $\mathcal{I}$ and $\mathcal{I}'$ is the partition $\mathcal{J}$ of $[-1,1)$ into intervals obtained from $\mathcal{I}$ and $\mathcal{I}'$ in the obvious way, by taking all possible nonempty intervals of the form $I_i \cap I'_j$. It is clear that $\mathcal{J}$ is both a refinement of $\mathcal{I}$ and of $\mathcal{I}'$ and that $\mathcal{J}$ contains at most $r+s$ intervals.

Approximation theory. We will need some basic notation and results from approximation theory. We write $\|p\|_\infty$ to denote $\sup_{x \in [-1,1)} |p(x)|$. We recall the famous inequalities of Bernstein and Markov bounding the derivative of univariate polynomials:

Theorem 4. *For any real-valued degree-d polynomial p over $[-1,1)$, we have*

- *(Bernstein's Inequality) $\|p'\|_\infty \leq \|p\|_\infty \cdot d^2$; and*
- *(Markov's Inequality) $\|p'\|_\infty \leq \frac{d}{\sqrt{1-x^2}} \cdot \|p\|_\infty$ for all $-1 \leq x \leq 1$.*

The VC inequality. Given a family of subsets $\mathcal{A}$ over $[-1,1)$, define $\|p\|_{\mathcal{A}} = \sup_{A \in \mathcal{A}} |p(A)|$. The *VC dimension* of $\mathcal{A}$ is the maximum size of a subset $X \subset [-1,1)$ that is shattered by $\mathcal{A}$ (a set X is shattered by $\mathcal{A}$ if for every $Y \subseteq X$, some $A \in \mathcal{A}$ satisfies $A \cap X = Y$). If there is a shattered subset of size s for all s then we say that the VC dimension of $\mathcal{A}$ is ∞. The well-known *Vapnik-Chervonenkis (VC) inequality* says the following:

Theorem 5 (VC inequality, [DL01, p.31])**.** *Let $\widehat{p}_m$ be an empirical distribution of m samples from p. Let $\mathcal{A}$ be a family of subsets of VC dimension d. Then $\mathbb{E}[\|p - \widehat{p}_m\|_{\mathcal{A}}] \leq O(\sqrt{d/m})$.*

2.1 Partitioning into intervals of approximately equal mass.

As a basic primitive, we will often need to decompose a κ-well-behaved distribution p into $\Theta(1/\kappa)$ intervals each of which has probability $\Theta(\kappa)$ under p. The following lemma lets us achieve this using $\tilde{O}(1/\kappa)$ samples; the simple proof is given in Appendix A.

Lemma 6. *Given $0 < \kappa < 1$ and access to samples from an $\kappa/64$-well-behaved distribution p over $[-1,1)$, the procedure* `Approximately-Equal-Partition` *uses $\tilde{O}(1/\kappa)$ samples from p, runs in time $\tilde{O}(1/\kappa)$, and with probability at least $99/100$ outputs a partition of $[-1,1)$ into $\ell = \Theta(1/\kappa)$ intervals such that $p(I_j) \in [\frac{1}{2\kappa}, \frac{3}{\kappa}]$ for all $1 \leq j \leq \ell$.*

3 Main result: Learning mixtures of piecewise polynomial distributions with near-optimal sample complexity

In this section we present and analyze our main algorithm for learning mixtures of (τ, t)-piecewise degree-d distributions over $[-1,1)$.

We start by giving a simple information-theoretic argument (Proposition 7, Section 3.1) showing that there is a (computationally inefficient) algorithm to learn any distribution p to accuracy $3\mathrm{opt}_{t,d}+\varepsilon$ using $O(t(d+1)/\varepsilon^2)$ samples, where $\mathrm{opt}_{t,d}$ is the smallest variation distance between p and any t-piecewise degree-d distribution. Next, we contrast this information-theoretic positive result with an information-theoretic lower bound (Theorem 8, Section 3.2) showing that any algorithm, regardless of its running time, for learning a t-piecewise degree-d distribution to accuracy ε must use $\Omega(\frac{t(d+1)}{\mathrm{poly}(1+\log(d+1))} \cdot \frac{1}{\varepsilon^2})$ samples. We then build up to our main result in stages by giving efficient algorithms for successively more challenging learning problems.

In Section 3.3 we give an efficient "semi-agnostic" algorithm for learning a single degree-d pdf. More precisely, the algorithm draws $\tilde{O}((d+1)/\varepsilon^2)$ samples from any well-behaved distribution p, and with high probability outputs a degree-d pdf h such that $d_{\mathrm{TV}}(p,h) \leq 3\mathrm{opt}_{1,d}(1+\varepsilon)+\varepsilon$. This algorithm uses ingredients from approximation theory and linear programming. In Section 3.4 we extend the approach using dynamic programming to obtain an efficient "semi-agnostic" algorithm for t-piecewise degree-d pdfs. The extended algorithm draws $\tilde{O}(t(d+1)/\varepsilon^2)$ samples from any well-behaved distribution p, and with high probability outputs a $(2t-1)$-piecewise degree-d pdf h such that $d_{\mathrm{TV}}(p,h) \le 3\mathrm{opt}_{t,d}(1+\varepsilon)+\varepsilon$. In Section 3.5 we extend the result to k-mixtures of well-behaved distributions. Finally, in Section 3.6 we show how we may get rid of the "well-behaved" requirement, and thereby prove Theorem 2.

3.1 An information-theoretic sample complexity upper bound.

Proposition 7. *There is a (computationally inefficient) algorithm that draws $O(t(d+1)/\varepsilon^2)$ samples from any distribution p over $[-1,1)$, and with probability $9/10$ outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(p,h) \le 3\mathrm{opt}_{t,d}+\varepsilon$.*

Proof. The main idea is to use Theorem 5, the VC inequality. Let p be the target distribution and let q be a t-piecewise degree-d distribution such that $d_{\mathrm{TV}}(p,q)=\mathrm{opt}_{t,d}$. The algorithm draws $m=O(t(d+1)/\varepsilon^2)$ samples from p; let $\widehat{p}_m$ be the resulting empirical distribution of these m samples.

We define the family $\mathcal{A}$ of subsets of $[-1,1)$ to consist of all unions of up to $2t(d+1)$ intervals. Since $d_{\mathrm{TV}}(p,q) \le \mathrm{opt}_{t,d}$ we have that $\|p-q\|_{\mathcal{A}} \le \mathrm{opt}_{t,d}$. Since the VC dimension of $\mathcal{A}$ is $4t(d+1)$, Theorem 5 implies that $\mathbb{E}[\|p-\widehat{p}_m\|_{\mathcal{A}}] \le \varepsilon/40$, and hence by Markov's inequality, with probability at least $19/20$ we have that $\|p-\widehat{p}_m\|_{\mathcal{A}} \le \varepsilon/2$. By the triangle inequality for $\|\cdot\|_{\mathcal{A}}$-distance, this means that $\|q-\widehat{p}_m\|_{\mathcal{A}} \le \mathrm{opt}_{t,d}+\varepsilon/2$.

The algorithm outputs a t-piecewise degree-d distribution h that minimizes $\|h-\widehat{p}_m\|_{\mathcal{A}}$. Since q is a t-piecewise degree-d distribution that satisfies $\|q-\widehat{p}_m\|_{\mathcal{A}} \le \mathrm{opt}_{t,d}+\varepsilon/2$, the distribution h satisfies $\|h-\widehat{p}_m\|_{\mathcal{A}} \le \mathrm{opt}_{t,d}+\varepsilon/2$. Hence the triangle inequality gives $\|h-q\|_{\mathcal{A}} \le 2\mathrm{opt}_{t,d}+\varepsilon$.

Now since h and q are both t-piecewise degree-d distributions, they must have at most $2t(d+1)$ crossings. (Taking the common refinement of the intervals for p and the intervals for q, we get at most $2t$ intervals. Within each such interval both h and q are degree-d polynomials, so there are at most $2t(d+1)$ crossings in total (where the extra $+1$ comes from the endpoints of each of the $2t$ intervals).) Consequently we have that $d_{\mathrm{TV}}(h,q)=\|h-q\|_{\mathcal{A}} \le 2\mathrm{opt}_{t,d}+\varepsilon$. The triangle inequality for variation distance gives that $d_{\mathrm{TV}}(h,p) \le 3\mathrm{opt}_{t,d}+\varepsilon$, and the proof is complete. □

It is not hard to see that the dependence on each of the parameters $t, d, 1/\varepsilon$ in the above upper bound is information-theoretically optimal.

Note that the algorithm described above is not efficient because it is by no means clear how to construct a t-piecewise degree-d distribution h that minimizes $\|h-\widehat{p}_m\|_{\mathcal{A}}$ in a computationally efficient way. Indeed, several approaches to solve this problem yield running times that grow exponentially in t, d. Starting in Section 3.3, we give an algorithm that achieves almost the same sample complexity but runs in time $\mathrm{poly}(t,d,1/\varepsilon)$. The main idea is that minimizing $\|\cdot\|_{\mathcal{A}}$ (which involves infinitely many inequalities) can be approximately achieved by minimizing a small number of inequalities (Theorems 11 and 14), and this can be achieved with a linear program.

3.2 An information-theoretic sample complexity lower bound. To complement the information-theoretic upper bound from the previous subsection, in this subsection we prove an information-theoretic lower bound showing that even if $\mathrm{opt}_{t,d}=0$ (i.e. the target distribution p is exactly a t-piecewise degree-d distribution), $\tilde{\Omega}(t(d+1)/\varepsilon^2)$ samples are required for any algorithm to learn to accuracy ε:

Theorem 8. *Let p be an unknown t-piecewise degree-d distribution over $[-1,1)$ where $t \ge 1$, $d \ge 0$ satisfy $t+d>1$.* [4] *Let L be any algorithm which, given as input t, d, ε and access to independent*

[4] Note that $t=1$ and $d=0$ is a degenerate case where the only possible distribution p is the uniform distribution over $[-1,1)$.

samples from p, outputs a hypothesis distribution h such that $\mathbb{E}[d_{\mathrm{TV}}(p,h)] \le \varepsilon$, where the expectation is over the random samples drawn from p and any internal randomness of L. Then L must use at least $\Omega(\frac{t(d+1)}{(1+\log(d+1))^2} \cdot \frac{1}{\varepsilon^2})$ samples.

Theorem 8 is proved using a well known lemma of Assouad [Ass83], together with carefully tailored constructions of polynomial probability density functions to meet the conditions of Assouad's lemma. The proof of Theorem 8 is deferred to Appendix A.2.

3.3 Semi-agnostically learning a degree-d polynomial density with near-optimal sample complexity. In this section we prove the following:

Theorem 9. *Let p be an $\frac{\varepsilon}{64(d+1)}$-well-behaved pdf over $[-1,1)$. There is an algorithm* `Learn-WB-Single-Poly`(d,ε) *which runs in poly$(d+1, 1/\varepsilon)$ time, uses $\tilde{O}((d+1)/\varepsilon^2)$ samples from p, and with probability at least $9/10$ outputs a degree-d polynomial q which defines a pdf over $[-1,1)$ such that $d_{\mathrm{TV}}(p,q) \le 3\mathrm{opt}_{1,d}(1+\varepsilon) + O(\varepsilon)$.*

Some preliminary definitions will be helpful:

Definition 10 (Uniform partition)**.** Let p be a subdistribution on an interval $I \subseteq [-1,1)$. A partition $\mathcal{P} = \{I_1, \ldots, I_\ell\}$ of I is (p,η)*-uniform* if $p(I_j) \le \eta$ for all $1 \le j \le \ell$.

Definition 11. Let $\mathcal{P} = \{[i_0, i_1), \ldots, [i_{r-1}, i_r)\}$ be a partition of an interval $I \subseteq [-1,1)$. Let $p, q : I \to \mathbb{R}$ be two functions on I. We say that p and q satisfy the $(\mathcal{P}, \eta, \varepsilon)$*-inequalities over* I if

$$|p([i_j, i_\ell)) - q([i_j, i_\ell))| \le \sqrt{\varepsilon(\ell - j)} \cdot \eta$$

for all $0 \le j < \ell \le r$.

We will also use the following notation: For this subsection, let $I = [-1,1)$ (I will denote a subinterval of $[-1,1)$ when the results are applied in the next subsection). We write $\|f\|_1^{(I)}$ to denote $\int_I |f(x)|dx$, and we write $d_{\mathrm{TV}}^{(I)}(p,q)$ to denote $\|p-q\|_1^{(I)}/2$. We write $\mathrm{opt}_{1,d}^{(I)}$ to denote the infimum of the statistical distance $d_{\mathrm{TV}}^{(I)}(p,g)$ between p and any degree-d subdistribution g on I that satisfies $g(I) = p(I)$.

The key step of `Learn-WB-Single-Poly` is Step 3 where it calls the `Find-Single-Polynomial` procedure. In this procedure $T_i(x)$ denotes the degree-i Chebychev polynomial of the first kind. The function `Find-Single-Polynomial` should be thought of as the CDF of a "quasi-distribution" f; we say that $f = F'$ is a "quasi-distribution" and not a bona fide probability distribution because it is not guaranteed to be non-negative everywhere on $[-1,1)$. Step 2 of `Find-Single-Polynomial` processes f slightly to obtain a polynomial q which is an actual distribution over $[-1,1)$.

We note that while the `Find-Single-Polynomial` procedure may appear to be more general than is needed for this section, we will exploit its full generality in the next subsection where it is used as a key subroutine for semi-agnostically learning t-piecewise polynomial distributions.

Algorithm `Learn-WB-Single-Poly`**:**

Input: parameters d, ε
Output: with probability at least $9/10$, a degree-d distribution q such that $d_{\mathrm{TV}}(p,q) \le 3 \cdot \mathrm{opt}_{1,d} + O(\varepsilon)$

1. Run Algorithm `Approximately-Equal-Partition` on input parameter $\varepsilon/(d+1)$ to partition $[-1,1)$ into $z = \Theta((d+1)/\varepsilon)$ intervals $I_0 = [i_0, i_1)$, $\ldots$, $I_{z-1} = [i_{z-1}, i_z)$, where $i_0 = -1$ and $i_z = 1$, such that for each $j \in \{1, \ldots, z\}$ we have $p([i_{j-1}, i_j)) = \Theta(\varepsilon/(d+1))$.
2. Draw $m = \tilde{O}((d+1)/\varepsilon^2)$ samples and let $\widehat{p}_m$ be the empirical distribution defined by these samples.
3. Call `Find-Single-Polynomial`(d, ε, $\eta := \Theta(\varepsilon/(d+1))$, $\{I_0, \ldots, I_{z-1}\}$, $\widehat{p}_m$) and output the hypothesis q that it returns.

Subroutine `Find-Single-Polynomial`:

Input: degree parameter d; error parameter ε; parameter η; (p,η)-uniform partition $\mathcal{P}_I = \{I_1, \ldots, I_z\}$ of interval $I = \cup_{i=1}^{z} I_i$ into z intervals such that $\sqrt{\varepsilon z} \cdot \eta \le \varepsilon/2$; a subdistribution $\widehat{p}_m$ on I such that $\widehat{p}_m$ and p satisfy the $(\mathcal{P}, \eta, \varepsilon)$-inequalities over I
Output: a number τ and a degree-d subdistribution q on I such that $q(I) = \widehat{p}_m(I)$,

$$d_{\mathrm{TV}}^{(I)}(p,q) \le 3\mathrm{opt}_{1,d}^{(I)}(1+\varepsilon) + \sqrt{\varepsilon r(d+1)} \cdot \eta + \mathrm{error},$$

$0 \le \tau \le \mathrm{opt}_{1,d}^{(I)}(1+\varepsilon)$ and $\mathrm{error} = O((d+1)\eta)$.

1. Let τ be the solution to the following LP:

   minimize τ subject to the following constraints:

   (Below $F(x) = \sum_{i=0}^{d+1} c_i T_i(x)$ where $T_i(x)$ is the degree-i Chebychev polynomial of the first kind, and $f(x) = F'(x) = \sum_{i=0}^{d+1} c_i T_i'(x)$.)

   (a) $F(-1) = 0$ and $F(1) = \widehat{p}_m(I)$;

   (b) For each $0 \le j < k \le z$,

   $$\left| \left(\widehat{p}_m([i_j, i_k)) + \sum_{j \le \ell < k} w_\ell \right) - (F(i_k) - F(i_j)) \right| \le \sqrt{\varepsilon \cdot (k-j)} \cdot \eta; \qquad (1)$$

   (c)

   $$\sum_{0 \le \ell < z} w_\ell = 0, \qquad (2)$$

   $$-y_\ell \le w_\ell \le y_\ell \qquad \text{for all } 0 \le \ell < z, \qquad (3)$$

   $$\sum_{0 \le \ell < z} y_\ell \le 2\tau(1+\varepsilon); \qquad (4)$$

   (d) The constraints $|c_i| \le \sqrt{2}$ for $i = 0, \ldots, d+1$;

   (e) The constraints

   $$0 \le F(z) \le 1 \quad \text{for all } z \in J,$$

   where J is a set of $O(d+1)^6$ equally spaced points across $[-1,1]$;

(f) The constraints

$$\sum_{i=0}^{d} c_i T_i'(x) \geq 0 \quad \text{for all } x \in K,$$

where K is a set of $O((d+1)^2/\varepsilon)$ equally spaced points across $[-1,1)$.

2. Define $q(x) = \varepsilon f(I)/|I| + (1-\varepsilon)f(x)$. Output q as the hypothesis pdf.

The rest of this subsection gives the proof of Theorem 9. The claimed sample complexity bound is obvious (observe that Steps 1 and 2 of `Learn-WB-Single-Poly` are the only steps that draw samples), as is the claimed running time bound (the computation is dominated by solving the poly$(d, 1/\varepsilon)$-size LP in `Find-Single-Poly`), so it suffices to prove correctness.

Before launching into the proof we give some intuition for the linear program. Intuitively $F(x)$ represents the cdf of a degree-d polynomial distribution f where $f = F'$. Constraint 1(a) captures the endpoint constraints that any cdf must obey if it has the same total mass as $\widehat{p}_m$. Intuitively, constraint 1(b)(1) ensures that for each interval $[i_j, i_k)$, the value $F(i_k) - F(i_j)$ (which we may alternately write as $f([i_j, i_k)))$ is close to the mass $\widehat{p}_m([i_j, i_k))$ that the empirical distribution puts on the interval. Recall that by assumption p is $\mathrm{opt}_{1,d}$-close to some degree-d polynomial r. Intuitively the variable w_ℓ represents $\int_{[i_\ell, i_{\ell+1})}(r-p)$ (note that these values sum to zero by constraint 1(c)(2)), and y_ℓ represents the absolute value of w_ℓ (see constraint 1(c)(3)). The value τ, which by constraint 1(c)(4) is at least the sum of the y_ℓ's, represents a lower bound on $\mathrm{opt}_{1,d}$. (The factor 2 on the RHS of constraint 1(c)(4) is present because $\|p - r\|_1 = 2d_{\mathrm{TV}}(p,r)$.) The constraints in 1(d) and 1(e) reflect the fact that as a cdf, F should be bounded between 0 and 1 (more on this below), and the 1(f) constraints reflect the fact that the pdf $f = F'$ should be everywhere nonnegative (again more on this below).

We begin by showing that with high probability `Learn-WB-Single-Poly` calls `Find-Single-Polynomial` with input parameters that satisfy `Find-Single-Polynomial`'s input requirements:

(I) the intervals $I_0, \ldots, I_{z-1}$ are (p, η)-uniform; and

(II) $\widehat{p}_m$ and p satisfy the $(\mathcal{P}, \eta, \varepsilon)$-inequalities over $[-1,1)$.

We further show that given that this happens, `Find-Single-Polynomial`'s LP is feasible and has a high-quality optimal solution.

Lemma 12. *Suppose p is an $\frac{\varepsilon}{64(d+1)}$-well-behaved pdf over $[-1,1)$. Then with overall probability at least $37/40$ over the random draws performed in steps 1 and 2 of* `Learn-WB-Single-Poly`*, conditions (I) and (II) above hold; the LP defined in step 1 of* `Find-Single-Polynomial` *is feasible; and the optimal solution τ is at most* $\mathrm{opt}_{1,d} \cdot (1+\varepsilon)$.

Proof. By Lemma 6, we have that with probability at least 99/100, every pair $j < k$ is such that the true probability mass $p([i_j, i_k))$ is $\Theta((k-j)\varepsilon/(d+1))$. (Note that the assumption that p is $\frac{\varepsilon}{64(d+1)}$-well-behaved was required to apply Lemma 6.) This gives (I). The multiplicative Chernoff bound (and a union bound) tells us that for every pair (j,k) with $1 \leq j < k \leq z$, with probability at least 39/40 we have

$$\widehat{p}_m([i_j, i_k)) \in (1 \pm \tau)p([i_j, i_k)) \qquad \text{for } \tau = \sqrt{\frac{\varepsilon}{k-j}}, \tag{5}$$

and hence

$$|\widehat{p}_m([i_j, i_k)) - p([i_j, i_k))| \leq \frac{1}{2} \cdot \sqrt{\varepsilon(k-j)} \cdot \frac{\varepsilon}{(d+1)}, \tag{6}$$

which implies (II). We assume that all these events hold going forth, and show that then the LP is feasible.

As above, let r be a degree-d polynomial pdf such that $\mathrm{opt}_{1,d} = d_{\mathrm{TV}}(p, r)$ and $r(I) = p(I)$. Let $\overline{r}$ be r renormalized by the empirical mass $\widehat{p}_m$, so $\overline{r} = r \cdot \widehat{p}_m(I)/p(I)$. Similarly let $\overline{p} = p \cdot \widehat{p}_m(I)/p(I)$ be the renormalization of p. We exhibit a feasible solution as follows: take F to be the cdf of $\overline{r}$ (a degree d polynomial). Take w to be $\int_{[i_\ell, i_{\ell+1})}(\overline{r} - \overline{p})$, and take y_ℓ to be $|w_\ell|$. Finally, take τ to be $\frac{1}{2}\sum_{0 \leq \ell < z} y_\ell$.

We first argue feasibility of the above solution. We first take care of the easy constraints: since F is the cdf of a subdistribution over I it is clear that constraints 1(a) and 1(e) are satisfied, and since both r and p are pdfs with the same total mass it is clear that constraints 1(c)(2) and 1(f) are both satisfied. Constraints 1(c)(3) and 1(c)(4) also hold, because $\frac{1}{2}\sum y_\ell = \frac{1}{2}\|r - p\|_1 \cdot \widehat{p}_m(I)/p(I) \leq d_{\mathrm{TV}}(p, r) \cdot (1 + \varepsilon)$, where we have used $(\mathcal{P}, \eta, \varepsilon)$-inequalities and the assumption $\sqrt{\varepsilon z} \cdot \eta \leq \varepsilon/2$ to show $\widehat{p}_m(I)/p(I) \in [1 - \varepsilon/2, 1 + \varepsilon/2]$. So it remains to argue constraints 1(b) and 1(d).

Claim 13. *If $\widehat{p}_m$ and p satisfy $(\mathcal{P}, \eta, \varepsilon/4)$-inequalities on $I \subseteq [-1, 1)$, then $\widehat{p}_m + \overline{r} - p$ and $\overline{r}$ satisfy $(\mathcal{P}, \eta, \varepsilon)$-inequalities on I.*

Proof. For an interval $J = [i_j, i_k) \in \mathcal{P}$, the LHS of $(\mathcal{P}, \eta, \varepsilon)$-inequalities between $\widehat{p} + (\overline{r} - p)$ and $\overline{r}$ is

$$|\widehat{p}_m(J) + (\overline{r} - p)(J) - \overline{r}(J)| = |\widehat{p}_m(J) - p(J)|.$$

Therefore it suffices to bound $|\widehat{p}_m(J) - p(J)|$ and $|\overline{r}(J) - p(J)|$. We can bound $|\widehat{p}_m(J) - p(J)|$ by $(\mathcal{P}, \eta, \varepsilon/4)$-inequalities between $\widehat{p}_m$ and p in our assumption. We also have

$$|\overline{r}(J) - p(J)| \leq \frac{\varepsilon}{2} p(J)$$

because $\widehat{p}_m(J)/p(J) \in [1 - \varepsilon/2, 1 + \varepsilon/2]$. □

Note that constraint 1(b) is equivalent to $\widehat{p}_m + (\overline{r} - p)$ and $\overline{r}$ satisfying $(\mathcal{P}, \varepsilon/(d+1), \varepsilon)$-inequalities, therefore this constraint is satisfied by Eq. (6) and Theorem 13.

To see that constraint 1(d) is satisfied we recall some of the analysis of Arora and Khot [AK03, Section 3]. This analysis shows that since r is a cdf (a function bounded between 0 and 1 on I) each of its Chebychev coefficients is at most $\sqrt{2}$ in magnitude. Therefore F is bounded between 0 and $1 + \varepsilon$, and likewise its coefficients are bounded by $\sqrt{2}(1 + \varepsilon)$.

To conclude the proof of the lemma we need to argue that $\tau \leq \mathrm{opt}_{1,d} \cdot (1 + \varepsilon)$. Since $w_\ell = \int_{[i_\ell, i_{\ell+1})}(\overline{r} - \overline{p})$ it is easy to see that $2\tau = \sum_{0 \leq \ell < z} y_\ell = \sum_{0 \leq \ell < z} |w_\ell| \leq \|\overline{p} - \overline{r}\|_1$, and hence indeed $\tau \leq d_{\mathrm{TV}}(p, r) \cdot \widehat{p}_m(I)/p(I) \leq \mathrm{opt}_{1,d} \cdot (1 + \varepsilon)$ as required. □

Having established that with high probability the LP is indeed feasible, henceforth we let τ denote the optimal solution to the LP and F, f, w_ℓ, c_i, y_ℓ denote the values in the optimal solution. A simple argument (see e.g. the proof of [AK03, Theorem 8]) gives that $\|F\|_\infty \leq 2(1 + \varepsilon)$. Given this bound on $\|F\|_\infty$, the Bernstein–Markov inequality implies that $\|f\|_\infty = \|F'\|_\infty \leq O((d+1)^2)$. Together with (1f) this implies that $f(z) \geq -\varepsilon/2$ for all $z \in [-1, 1)$. Consequently $q(z) \geq 0$ for all $z \in [-1, 1)$, and

$$\int_{-1}^{1} q(x)dx = \varepsilon + (1 - \varepsilon)\int_{-1}^{1} f(x)dx = \varepsilon + (1 - \varepsilon)(F(1) - F(-1)) = 1.$$

So $q(x)$ is indeed a degree-d pdf. To prove Theorem 9 it remains to show that $d_{\mathrm{TV}}(q,p) \le 3\mathrm{opt}_{1,d} + O(\varepsilon)$.

We sketch the argument that we shall use to bound $d_{\mathrm{TV}}(p,q)$. A key step in achieving this bound is to bound the $\|\cdot\|_{\mathcal{A}}$ distance between f and $\widehat{p}_m + w$ where $\mathcal{A} = \mathcal{A}_{d+1}$ is the class of all unions of $d+1$ intervals and w is a function based on the w_ℓ values (see Eq. (9) below). Similar to Section 3.1 the VC theorem gives us that $\|p - \widehat{p}_m\|_{\mathcal{A}} \le \varepsilon$ with probability at least 39/40, so if we can bound $\|(\widehat{p}_m + w) - f\|_{\mathcal{A}} \le O(\varepsilon)$ then it will not be difficult to show that $\|r - f\|_{\mathcal{A}} \le 2\mathrm{opt}_{1,d} + O(\varepsilon)$. Since r and f are both degree-d polynomials we have $d_{\mathrm{TV}}(r,f) = \|r - f\|_{\mathcal{A}} \le 2\mathrm{opt}_{1,d} + O(\varepsilon)$, so the triangle inequality (recalling that $d_{\mathrm{TV}}(p,r) = \mathrm{opt}_{1,d}$) gives $d_{\mathrm{TV}}(p,f) \le 3\mathrm{opt}_{1,d} + O(\varepsilon)$. From this point a simple argument (Proposition 15) gives that $d_{\mathrm{TV}}(p,q) \le d_{\mathrm{TV}}(p,f) + O(\varepsilon)$, which gives the theorem.

We will use the following lemma that translates $(\mathcal{P},\eta,\varepsilon)$-inequalities into a bound on $\mathcal{A}_{d+1}$ distance.

Lemma 14. *Let $\mathcal{P} = \{I_0 = [i_0, i_1), \ldots, I_{z-1} = [i_{z-1}, i_z)\}$ be a (p,η)-uniform partition of I. Let $\widehat{p}_m$ be a subdistribution on I such that $\widehat{p}_m$ and p satisfy $(\mathcal{P},\eta,\varepsilon)$-inequalities on I. If $h : I \to \mathbb{R}$ and $\widehat{p}_m$ also satisfy the $(\mathcal{P},\eta,\varepsilon)$-inequalities, then*

$$\|\widehat{p}_m - h\|_{\mathcal{A}_{d+1}}^{(I)} \le \sqrt{\varepsilon z(d+1)} \cdot \eta + \mathrm{error},$$

where error $= O((d+1)\eta)$.

Proof. To analyze $\|\widehat{p}_m - h\|_{\mathcal{A}_{d+1}}$, consider any union of $d+1$ disjoint non-overlapping intervals $S = J_1 \cup \cdots \cup J_{d+1}$. We will bound $\|\widehat{p}_m - h\|_{\mathcal{A}_{d+1}}$ by bounding $|\widehat{p}_m(S) - h(S)|$.

We lengthen intervals in S slightly to obtain $T = J_1' \cup \cdots \cup J_{d+1}'$ so that each J_j' is a union of intervals of the form $[i_\ell, i_{\ell+1})$. Formally, if $J_j = [a,b)$, then $J_j' = [a', b')$, where $a' = \max_\ell\{i_\ell \mid i_\ell \le a\}$ and $b' = \min_\ell\{i_\ell \mid i_\ell \ge b\}$. We claim that

$$|\widehat{p}_m(S) - h(S)| \le O((d+1)\eta) + |\widehat{p}_m(T) - f(T)|. \tag{7}$$

Indeed, consider any interval of the form $J = [i_\ell, i_{\ell+1})$ such that $J \cap S \ne J \cap T$. We have

$$|\widehat{p}_m(J \cap S) - \widehat{p}_m(J \cap T)| \le \widehat{p}_m(J) \le O(\eta), \tag{8}$$

where the first inequality uses nonegativity of $\widehat{p}_m$ and the second inequality follows from $(\mathcal{P},\eta,\varepsilon)$-inequalities (between $\widehat{p}_m$ and p) and the bound $p([i_\ell, i_{\ell+1})) \le \eta$. The $(\mathcal{P},\eta,\varepsilon)$-inequalities (between h and $\widehat{p}_m$) implies that the inequalities in Eq. (8) also hold with h in place of $\widehat{p}_m$. Now Eq. (7) follows by adding Eq. (8) across all $J = [i_\ell, i_{\ell+1})$ such that $J \cap S \ne J \cap T$ (there are at most $2(d+1)$ such intervals J), since each interval J_j in S can change at most two such J's when lengthened.

Now rewrite T as a disjoint union of $s \le d+1$ intervals $[i_{L_1}, i_{R_1}) \cup \cdots \cup [i_{L_s}, i_{R_s})$. We have

$$|\widehat{p}_m(T) - h(T)| \le \sum_{j=1}^{s} \sqrt{R_j - L_j} \cdot \sqrt{\varepsilon}\eta$$

by $(\mathcal{P},\eta,\varepsilon)$-inequalities between $\widehat{p}_m$ and h. Now observing that that $0 \le L_1 \le R_1 \cdots \le L_s \le R_s \le t = O((d+1)/\varepsilon)$, we get that the largest possible value of $\sum_{j=1}^{s} \sqrt{R_j - L_j}$ is $\sqrt{sz} \le \sqrt{(d+1)z}$, so the RHS of (7) is at most $O((d+1)\eta) + \sqrt{(d+1)z\varepsilon}\eta$, as desired. □

Recall from above that F, f, w_ℓ, c_i, y_ℓ, τ denote the values in the optimal solution. We claim that

$$\|(\widehat{p}_m + w) - f\|_{\mathcal{A}} = O(\varepsilon), \tag{9}$$

where w is the sub-distribution which is constant on each $[i_\ell, i_{\ell+1})$ and has mass w_ℓ there, so in particular $\|w\|_1 \le 2\tau \le 2\mathrm{opt}_{1,d}(1+\varepsilon)$. Indeed, this equality follows by applying Theorem 14 with $h = f - w$. The lemma requires h and $\widehat{p}_m$ to satisfy $(\mathcal{P}, \eta, \varepsilon)$-inequalities, which follows from constraint 1(b) ($(\mathcal{P}, \eta, \varepsilon)$-inqualities between $\widehat{p}_m + w$ and f) and observing that $(\widehat{p}_m + w) - f = \widehat{p}_m - (f - w)$. We have also used $\eta = \Theta(\varepsilon/(d+1))$ to bound the error term of the lemma by $O(\varepsilon)$.

Next, by the triangle inequality we have (writing $\mathcal{A}$ for $\mathcal{A}_{d+1}$)

$$\|r - f\|_{\mathcal{A}} \le \|r - (p+w)\|_{\mathcal{A}} + \|(p+w) - (\widehat{p}_m + w)\|_{\mathcal{A}} + \|(\widehat{p}_m + w) - f\|_{\mathcal{A}}.$$

The last term on the RHS has just been shown to be $O(\varepsilon)$. The second term equals $\|p - \widehat{p}_m\|_{\mathcal{A}}$ and is $O(\varepsilon)$ with probability at least 39/40 by the VC inequality. The first term is bounded by

$$\|r - (p+w)\|_{\mathcal{A}} \le d_{\mathrm{TV}}(r, p+w) = \|r - (p+w)\|_1/2 \le (\|r-p\|_1 + \|w\|_1)/2 \le 2\mathrm{opt}_{1,d}(1+\varepsilon).$$

Altogether, we get that $\|r - f\|_{\mathcal{A}} \le 2\mathrm{opt}_{1,d}(1+\varepsilon) + O(\varepsilon)$.

Since r and f are degree d polynomials, $d_{\mathrm{TV}}(r, f) = \|r - f\|_{\mathcal{A}} \le 2\mathrm{opt}_{1,d}(1+\varepsilon) + O(\varepsilon)$. This implies $d_{\mathrm{TV}}(p, f) \le d_{\mathrm{TV}}(p, r) + d_{\mathrm{TV}}(r, f) \le 3\mathrm{opt}_{1,d}(1+\varepsilon) + O(\varepsilon)$. Finally, we turn our quasidistribution f which has value $\ge -\varepsilon/2$ everywhere into a distribution q (which is nonnegative), by redistributing the mass. The following simple proposition bounds the error incurred.

Proposition 15. *Let f and p be any sub-quasidistribution on I. If $q = \varepsilon f(I)/|I| + (1-\varepsilon)f$, then $\|q - p\|_1 \le \|f - p\|_1 + \varepsilon(f(I) + p(I))$.*

Proof. We have

$$q - p = \varepsilon(f(I)/|I| - p) + (1-\varepsilon)(f - p).$$

Therefore

$$\|q - p\|_1 \le \varepsilon \, \|f(I)/|I| - p\|_1 + (1-\varepsilon)\, \|f - p\|_1 \le \varepsilon(f(I) + p(I)) + \|f - p\|_1. \qquad \square$$

We have $d_{\mathrm{TV}}(p, q) \le d_{\mathrm{TV}}(p, f) + O(\varepsilon)$ by Proposition 15, and we are done with the proof of Theorem 9. □

3.4 Efficiently learning (ε, t)-piecewise degree-d distributions.

In this section we extend the previous result to semi-agnostically learn t-piecewise degree-d distributions. We prove the following:

Theorem 16. *Let p be an $\frac{\varepsilon}{64t(d+1)}$-well-behaved pdf over $[-1, 1)$. There is an algorithm* `Learn-WB-Piecewise-Poly`(t, d, ε) *which runs in poly$(t, d+1, 1/\varepsilon)$ time, uses $\tilde{O}(t(d+1)/\varepsilon^2)$ samples from p, and with probability at least $9/10$ outputs a $(2t-1)$-piecewise degree-d distribution q such that $d_{\mathrm{TV}}(p, q) \le 3\mathrm{opt}_{t,d}(1+\varepsilon) + O(\varepsilon)$.*

At a high level, `Learn-WB-Piecewise-Poly`(t, d, ε) breaks down $[-1, 1)$ into t/ε subintervals (denoted as the partition $\mathcal{P}' = \{I'_0, \ldots, I'_{t/\varepsilon - 1}\}$ in subsequent discussion; this partition is constructed in step (2)) and calls the subroutine `Find-Single-Polynomial`$(d, \varepsilon, \eta, \{I'_\ell, \ldots, I'_{j-1}\}, \widehat{p}_m)$ on blocks of consecutive intervals from $\mathcal{P}'$ (see Theorem 17). As shown in the previous subsection, the subroutine `Find-Single-Polynomial` returns a degree-d polynomial h that is close to the optimal degree-d polynomial over $I'_\ell \cup \cdots \cup I'_{j-1}$. An exhaustive search over all ways of breaking $[-1, 1)$ up into t intervals would require running time exponential in t; to improve efficiency, dynamic programming is used to combine the different h's obtained as described above and efficiently construct an overall high-accuracy piecewise degree-d hypothesis.

Remark 17. *The subroutine* Find-Single-Polynomial *from the previous section assumes the domain I is $[-1,1)$. The following modification extends the subroutine to arbitrary domain I.*

Map the interval $I=[a,b)$ to $[-1,1)$ via

$$\phi_I(a+\lambda(b-a)) = -1+2\lambda \quad \forall \lambda \in [0,1).$$

We write $\phi=\phi_I$ when I is clear from the context. Then the transformation $f \mapsto f_\phi$, where

$$f_\phi(x) = \frac{b-a}{2} \cdot f(\phi^{-1}(x)),$$

is a linear map taking distributions over I to distributions over $[-1,1)$ (and in fact, a linear isomorphism from $L_1(I)$ to $L_1[-1,1)$.) This transformation is also a bijection between degree-d polynomials over I and those over $[-1,1)$. As a result, if we represent f_ϕ by

$$f_\phi(x) = \sum_{i=0}^{d} c_i T_i(x) \quad \forall x \in [-1,1),$$

where $T_i : [-1,1) \to \mathbb{R}$ are Chebyshev polynomials of degree i, we get a representation of $f : I \to \mathbb{R}$ via

$$f(y) = \frac{2}{b-a} \sum_{i=0}^{d} c_i T_i(\phi(y)). \tag{10}$$

Note that if f is bounded on I and $b-a \le 2$, then the same is true for f_ϕ on $[-1,1)$, and

$$\|f_\phi\|_\infty^{([-1,1))} \le \|f\|_\infty^{(I)}.$$

(The same inequality is also true with the RHS multiplied by $(b-a)/2 \le 1$, but we only need the weaker inequality above.)

Further, since $f \mapsto f_\phi$ preserves distances between subdistributions, the assumptions and conclusions in the subroutine remain unchanged.

Algorithm `Learn-WB-Piecewise-Poly`:

Input: parameters t, d, ε

Output: with probability at least $9/10$, a t-piecewise degree-d distribution q such that $d_{\mathrm{TV}}(p,q) \le 3 \cdot \mathrm{opt}_{t,d}(1+\varepsilon) + O(\varepsilon)$

```
1. Run Algorithm Approximately-Equal-Partition on input parameter $\varepsilon/(t(d+1))$ to partition $[-1,1)$ into $z = \Theta(t(d+1)/\varepsilon)$ intervals $I_0 = [i_0, i_1)$, ..., $I_z = [i_{z-1}, i_z)$, where $i_0 = 0$ and $i_z = 1$, such that for each $j \in \{1,\ldots,t\}$ we have $p([i_{j-1}, i_j)) = \Theta(\varepsilon/(t(d+1)))$.
2. Let $s = z/(d+1) = \Theta(t/\varepsilon)$. Set $i'_j = i_{(d+1)j}$ and define interval $I'_j = [i'_j, i'_{j+1})$ for $0 \le j < s$.
3. Draw $m = \tilde{O}(t(d+1)/\varepsilon^2)$ samples to define an empirical distribution $\widehat{p}_m$ over $[-1,1)$.
4. Initialize $T(i,j) = \infty$ for $i \in \{0,\ldots,2t-1\}$, $j \in \{0,\ldots,s\}$, except that $T(0,0) = 0$.
5. For $i \in \{1,\ldots,2t-1\}$, $j \in \{1,\ldots,s\}$, $\ell \in \{0,\ldots,j-1\}$:
    (a) Call subroutine Find-Single-Polynomial ($d$, $\varepsilon$, $\eta = \Theta(\varepsilon/(t(d+1)))$, $\{I'_\ell, \ldots, I'_{j-1}\}$, $\widehat{p}_m$)
    (b) Let $\tau$ be the solution to the LP found by Find-Single-Polynomial and $h$ be the degree-$d$ hypothesis sub-distribution that it returns.
    (c) If $T(i,j) > T(i-1,\ell) + \tau$, then
        i. Update $T(i,j)$ to $T(i-1,\ell) + \tau$
        ii. Store the polynomial $h$ in a table $H(i,j)$.
6. Recover a piecewise degree-$d$ distribution $h$ from the table $H(\cdot,\cdot)$.
```

Let $\checkmark_1$ be the event that step (1) of Subroutine `Find-Piecewise-Polynomial` succeeds (i.e. the intervals $[i_j, i_{j+1})$ all have mass within a constant factor of $\varepsilon/t(d+1)$). In step (2) of `Learn-WB-Piecewise-Poly`, the algorithm effectively constructs a coarsening $\mathcal{P}'$ of $\mathcal{P}$ by merging every $d+1$ consecutive intervals from $\mathcal{P}$. These super-intervals are used in the dynamic programming in step (5). The table entry $T(i,j)$ stores the minimum sum of errors τ (returned by the subroutine Find-Single-Polynomial) when the interval $[i'_0, i'_j)$ is partitioned into i pieces. The dynamic program above only computes an estimate of $\mathrm{opt}_{t,d}$; one can use standard techniques to also recover a t-piecewise degree-d polynomial q close to p.

For step (3), let $\checkmark_2$ be the event that p and $\widehat{p}_m$ satisfy $(\mathcal{P}, \varepsilon/(t(d+1)), \varepsilon/4)$-inequalities. In particular, when $\checkmark_2$ holds $\widehat{p}_m(I)/p(I) \leq \varepsilon/2$ for all $I \in \mathcal{P}$. By multiplicative Chernoff and union bound (over the m samples in step (3)), event $\checkmark_2$ holds with probability at least 19/20.

Proposition 18. *If $\checkmark_1$ and $\checkmark_2$ hold and p is τ-close to some t-piecewise degree-d distribution, then there is a coarsening $\mathcal{P}^*$ of $\mathcal{P}'$ and degree-d polynomials $g_i : I^*_i \to \mathbb{R}$ such that $\sum_i d_{\mathrm{TV}}(p, g_i) \leq \tau + O(\varepsilon)$. Further, the g_i functions can be chosen to satisfy constraints 1a, 1d–1f in the subroutine* Find-Piecewise-Polynomial.

Proof. Suppose p is τ-close to a t-piecewise degree-d distribution. In other words, there exists a partition $\{J_1, \ldots, J_t\}$ of $[-1,1)$ and degree-d polynomials $h_i : J_i \to \mathbb{R}$ such that $\sum_{1 \leq i \leq t} d_{\mathrm{TV}}(p, h_i) \leq \tau$.

Let $\{[i'_0, i'_1), \ldots, [i'_{s-1}, i'_s)\}$ be $\mathcal{P}'$. Except in degenerate cases, the coarsening $\mathcal{P}^*$ contains $2t-1$ intervals, corresponding to the t intervals on which p is a polynomial and $t-1$ small intervals containing "breakpoints" between the polynomials. More precisely, if we denote by $\{\alpha_0, \ldots, \alpha_j\}$ the breakpoints of $J_1, \ldots, J_t$ (so that $J_j = [\alpha_{j-1}, \alpha_j)$), and define

$$J'_j := \cup\{[\alpha_a, \alpha_b) \mid [\alpha_a, \alpha_b) \subset J_j\}$$

as the maximal subinterval of J_j with endpoints from $\{\alpha_j\}$, then $\mathcal{P}^*$ is the partition containing all the J'_j's together with the intervals between consecutive J'_j's. As a result, $\mathcal{P}^*$ is a partition of $[-1,1)$ into at most $2t-1$ non-empty intervals.

For an interval I^*_i not containing any breakpoint, the corresponding polynomial $g_i : I^*_i \to \mathbb{R}$ is simply h_i rescaled by the empirical mass on I^*_i, so

$$g_i(x) = h_i(x) \cdot \frac{\widehat{p}_m(I^*_i)}{h_i(I^*_i)} \quad \text{for } x \in I^*_i \neq \emptyset.$$

Then g_i clearly satisfies constraints 1a and 1f. Constraints 1d and 1e are also satisfied: $(h_i)_{\phi_i}$ is a degree-d polynomial on $[-1,1)$ bounded by 1 in absolute value (here $\phi_i = \phi_{I_i^*}$), and $\widehat{p}_m(I_i^*)/h_i(I_i^*) \leq \varepsilon/2$ when $\checkmark_2$ holds.

For an interval I_i^* containing a breakpoint, we simply set g_i to be the constant function with total mass $\widehat{p}_m(I_i^*)$ on I_i^*. As before, g_i satisfies 1d–1f. The contribution of such g_i's (there are at most $t-1$ of them) to $\sum_i d_{\mathrm{TV}}(p, g_i)$ is at most $(t-1)\cdot 2\varepsilon/t = O(\varepsilon)$, using the fact that $\mathcal{P}'$ is $(\widehat{p}_m, 4\varepsilon/t)$-uniform when $\checkmark_1$ and $\checkmark_2$ hold. □

When event $\checkmark_2$ holds, p and $\widehat{p}_m$ satisfy the $(\mathcal{P}_{I_i^*}, \varepsilon/(t(d+1)), \varepsilon/4)$-inequalities. But this is the same as g_i and $\widehat{p}_m + (g_i - p)$ satisfying the $(\mathcal{P}_{I_i^*}, \varepsilon/(t(d+1)), \varepsilon)/4$-inequalities, because $p - \widehat{p}_m = g_i - (\widehat{p}_m + g_i - p)$. Therefore Theorem 13 tells us that constraint 1b is satisfied. Constraints 1c are satisfied for similar reasons as in Section 3.3. Together with Theorem 18, the LP in the subroutine `Find-Single-Polynomial` will be feasible, provided the partition $\mathcal{P}^*$ is chosen correctly in the dynamic program.

We have the following restatement of Theorem 14, and a robust version as a corollary (which follows by combining Theorem 14 and the proof of Theorem 7).

Lemma 19 (Theorem 14 restated)**.** *Let $\mathcal{P}$ be a (p,η)-partition of $I \subseteq [-1,1)$ into r intervals. Let $\widehat{p}_m$ be a subdistribution on I such that $\widehat{p}_m$ and p satisfy the $(\mathcal{P},\eta,\varepsilon)$-inequalities. If $f : I \to \mathbb{R}$ and $\widehat{p}_m$ also satisfy the $(\mathcal{P},\eta,\varepsilon)$-inequalities, then*

$$\|\widehat{p}_m - f\|_{\mathcal{A}_d}^{(I)} \leq \sqrt{\varepsilon r(d+1)} \cdot \eta + \mathrm{error},$$

where the error *is $O((d+1)\eta)$.*

Corollary 20. *Let p be a degree-d subdistribution on I. Let $\mathcal{P}$ be a (p,η)-partition of $I \subseteq [-1,1)$ into r intervals. Let $\widehat{p}_m$ be a subdistribution on I such that $\widehat{p}_m$ and p satisfy $(\mathcal{P},\eta,\varepsilon)$-inequalities. If $h : I \to \mathbb{R}$ and $\widehat{p}_m + w$ also satisfy $(\mathcal{P},\eta,\varepsilon)$-inequalities, then*

$$d_{\mathrm{TV}}^{(I)}(p,h) \leq 3\tau(1+\varepsilon) + \sqrt{\varepsilon r(d+1)} \cdot \eta + \mathrm{error},$$

where $2\tau = \|w\|_1$ and error $= O((d+1)\eta)$*.*

Proof of Theorem 16. Since p is τ-close to a t-piecewise degree-d distribution, there are a partition $\{J_1,\ldots,J_t\}$ of $[-1,1)$ and degree-d polynomials $g_i : J_i \to \mathbb{R}$ such that $\sum_{1\leq i\leq t} \tau_i \leq \tau$, where $\tau_i = d_{\mathrm{TV}}(p, g_i)$. Let $\mathcal{P}^* = \{I_1^*,\ldots,I_{2t-1}^*\}$ be the coarsening of $\mathcal{P}'$ as in the proof of Theorem 18.

When $\checkmark_1$ and $\checkmark_2$ hold, it follows by a simple induction on $i \in \{0,\ldots,2t-1\}$ that the algorithm will output a $(2t-1)$-piecewise degree-d distribution h satisfying

$$d_{\mathrm{TV}}(p,h) \leq \sum_{1\leq i\leq t} \left(3\tau_i(1+\varepsilon) + \sqrt{\varepsilon r_i(d+1)} \cdot \frac{\varepsilon}{t(d+1)} + O\left((d+1)\cdot\frac{\varepsilon}{t(d+1)}\right)\right) + O(\varepsilon). \qquad (11)$$

The first term comes from Theorem 20 (with $\eta = O(\varepsilon/(t(d+1))))$, and the second term comes from the $t-1$ intervals containing the breakpoints (see the proof of Theorem 18). Here r_i denotes the number of intervals from $\mathcal{P}$ contained in I_i^*. Therefore the RHS of Eq. (11) is at most

$$3\tau(1+\varepsilon) + \sum_{1\leq i\leq t} \sqrt{\varepsilon r_i(d+1)} \cdot \frac{\varepsilon}{t(d+1)} + O(\varepsilon).$$

The second term of this expression is bounded by ε using Cauchy–Schwarz and the fact that $\mathcal{P}$ contains $t(d+1)/\varepsilon$ intervals. □

3.5 Learning k-mixtures of well-behaved (τ, t)-piecewise degree-d distributions. In this subsection we prove Theorem 2 under the additional restriction that the target polynomial p is well-behaved:

Theorem 21. *Let p be an $\frac{\varepsilon}{64kt(d+1)}$-well-behaved k-mixture of (τ, t)-piecewise degree-d distributions over $[-1, 1)$. There is an algorithm that runs in $\mathrm{poly}(k, t, d + 1, 1/\varepsilon)$ time, uses $\tilde{O}((d + 1)kt/\varepsilon^2)$ samples from p, and with probability at least $9/10$ outputs a $(2kt - 1)$-piecewise degree-d hypothesis h such that $d_{\mathrm{TV}}(p, h) \le 3\mathrm{opt}_{t,d}(1 + \varepsilon) + O(\varepsilon)$.*

As we shall see, the algorithm of the previous subsection in fact suffices for this result. The key to extending Theorem 16 to yield Theorem 21 is the following structural result, which says that any k-mixture of (τ, t)-piecewise degree-d distributions must itself be an (τ, kt)-piecewise degree-d distribution.

Lemma 22. *Let $p_1, \ldots, p_k$ each be an (τ, t)-piecewise degree-d distribution over $[-1, 1)$ and let $p = \sum_{j=1}^{k} \mu_j p_j$ be a k-mixture of components $p_1, \ldots, p_k$. Then p is a (τ, kt)-piecewise degree-d distribution.*

The simple proof is essentially the same as the proof of Lemma 3.2 of [CDSS13] and is given in Appendix A.

We may rephrase Theorem 16 as follows:

Alternate Phrasing of Theorem 16. *Let p be an $\frac{\varepsilon}{64t(d+1)}$-well-behaved (τ, t)-piecewise degree-d pdf over $[-1, 1)$. Algorithm* `Learn-WB-Piecewise-Poly`(t, d, ε) *runs in $\mathrm{poly}(t, d + 1, 1/\varepsilon)$ time, uses $\tilde{O}(t(d + 1)/\varepsilon^2)$ samples from p, and with probability at least $9/10$ outputs a $(2t - 1)$-piecewise degree-d distribution q such that $d_{\mathrm{TV}}(p, q) \le 3\tau(1 + \varepsilon) + O(\varepsilon)$.*

Theorem 21 follows immediately from Theorem 16 and Lemma 22.

3.6 Proof of Theorem 2. In this subsection we show how to remove the well-behavedness assumption from Theorem 21 and thus prove Theorem 2. More precisely we prove the following theorem which is a more detailed version of Theorem 2:

Theorem 23. *Let p be any k-mixture of (τ, t)-piecewise degree-d distributions over $[-1, 1)$. There is an algorithm that runs in $\mathrm{poly}(k, t, d+1, 1/\varepsilon)$ time, uses $\tilde{O}((d+1)kt/\varepsilon^2)$ samples from p, and with probability at least $9/10$ outputs a $(2kt - 1)$-piecewise degree-d hypothesis h such that $d_{\mathrm{TV}}(p, h) \le 4\mathrm{opt}_{t,d}(1 + \varepsilon) + O(\varepsilon)$.*

To prove Theorem 23 we will need the following simple procedure, which (approximately) outputs all the points in $[-1, 1)$ that are γ-heavy under a distribution p:

Algorithm `Find-Heavy`**:**

Input: parameter $\gamma > 0$, sample access to distribution p over $[-1, 1)$
Output: With probability at least 99/100, a set $S \subset [-1, 1)$ such that for all $x \in [-1, 1)$,

1. if $\Pr_{x\sim p}[x] \ge 2\gamma$ then $x \in S$;
2. if $\Pr_{x\sim p}[x] < \gamma/2$ then $x \notin S$.

Draw $m = \tilde{O}(1/\gamma)$ samples from p. For each $x \in [-1, 1)$ let $\widehat{p}(x)$ equal $1/m$ times the number of occurrences of x in these m draws. Return the set S which contains all x such that $\widehat{p}(x) \ge \gamma$.

It is clear that the set S returned by Find-Heavy(γ) has $|S| \leq 1/\gamma$. We now prove that Find-Heavy performs as claimed:

Lemma 24. *With probability at least* $99/100$, Find-Heavy(γ) *returns a set S satisfying conditions (1) and (2) in the "Output" description.*

We give the straightforward proof in Appendix A.

To prove Theorem 23 it suffices to prove the following result (which is an extension of Theorem 16 that does not require the well-behavedness condition on p):

Theorem 25. *Let p be a pdf over $[-1,1)$. There is an algorithm* Learn-Piecewise-Poly(t,d,ε) *which runs in poly$(t,d{+}1,1/\varepsilon)$ time, uses $\tilde{O}(t(d{+}1)/\varepsilon^2)$ samples from p, and with probability at least* $9/10$ *outputs a $(2t-1)$-piecewise degree-d distribution q such that* $d_{\mathrm{TV}}(p,q) \leq 4\mathrm{opt}_{t,d}(1+\varepsilon)+O(\varepsilon)$. *where* $\mathrm{opt}_{t,d}$ *is the smallest variation distance between p and any t-piecewise degree-d distribution.*

Using the arguments of Section 3.5, Theorem 23 follows from Theorem 25 exactly as Theorem 21 follows from Theorem 16.

Proof of Theorem 25. The algorithm Learn-Piecewise-Poly$(t,d,1/\varepsilon)$ works as follows: it first runs Find-Heavy(γ) where $\gamma = O(\frac{\varepsilon}{t(d+1)})$ to obtain a set $S \subset [-1,1)$. It then runs Learn-WB-Piecewise-Poly-$(t,d,1/\varepsilon)$ but using the distribution $p_{[-1,1)\setminus S}$ (i.e. p conditioned on $[-1,1)\setminus S$) in place of p throughout the algorithm. Each time a draw from $p_{[-1,1)\setminus S}$ is required, it simply draws repeatedly from p until a point outside of S is obtained.

Let p be any distribution over $[-1,1)$. Since the conclusion of the theorem is trivial if $\mathrm{opt}_{t,d} \geq 1/4$, we may assume that $\mathrm{opt}_{t,d} < 1/4$.

Consider an execution of Learn-Piecewise-Poly$(t,d,1/\varepsilon)$. We assume that conditions (1) and (2) of Find-Heavy indeed hold for the set S that it constructs. Let $S' \supseteq S$ be defined as $S' = \{x \in [-1,1) : \Pr_{x\sim p}[x] \geq \gamma/2\}$. Since every t-piecewise degree-d distribution q has $d_{TV}(p,q) \geq \Pr_{x\sim p}[x \in S']$ (because p assigns probability $\Pr_{x\sim p}[x \in S']$ to S' whereas q assigns probability 0 to this finite set of points), it must be the case that $\Pr_{x\sim p}[x \in S] \leq \Pr_{x\sim p}[x \in S'] \leq \mathrm{opt}_{t,d}$. Hence a draw from $p_{[-1,1)\setminus S}$ is indeed a valid draw from $p_{[-1,1)\setminus S}$ except with failure probability at most $\mathrm{opt}_{t,d} < 1/4$. It follows easily from this and the sample complexity bound of Theorem 16 that the sample complexity of algorithm Learn-Piecewise-Poly$(t,d,1/\varepsilon)$ is as claimed.

Verifying correctness is also straightforward. Recall that $\mathrm{opt}_{t,d}$ denotes the infimum of $d_{TV}(p,q)$ where q is any t-piecewise degree-d distribution. Fix a q which achieves $d_{TV}(p,q) = \mathrm{opt}_{t,d}$; we claim that this q also satisfies $d_{TV}(p_{[-1,1)\setminus S}, q) \leq \mathrm{opt}_{t,d}$. (To see this, note that we may write $d_{TV}(p,q)$ as $A+B$ where A is the contribution from points in $[-1,1)\setminus S$ and B is the contribution from S. Since $\Pr_{x\sim q}[x \in B]$ is zero it must be the case that $B = \frac{1}{2}\Pr_{x\sim p}[S]$, where the "$\frac{1}{2}$" is the factor relating L_1 norm and total variation distance. Now write $d_{TV}(p_{[-1,1)\setminus S}, q)$ as $A'+B'$ where A' is the contribution from points in $[-1,1)\setminus S$ and B is the contribution from S. Clearly B' is now 0, and A' can be at most $B = \frac{1}{2}\Pr_{x\sim p}[S]$ larger than A.) By Lemma 24 we have that $p_{[-1,1)\setminus S}$ is $O(\frac{\varepsilon}{t(d+1)})$-well-behaved. Hence by Theorem 16, when Learn-WB-Piecewise-Poly$(t,d,1/\varepsilon)$ is run on $p_{[-1,1)\setminus S}$ it succeeds with high probability to give a hypothesis h such that $d_{TV}(h, p_{[-1,1)\setminus S}) \leq 3\mathrm{opt}_{t,d}(1+\varepsilon)+O(\varepsilon)$. Since $d_{TV}(p, p_{[-1,1)\setminus S}) \leq \mathrm{opt}_{t,d}$ using the triangle inequality we get that $d_{TV}(h,p) \leq 4\mathrm{opt}_{t,d}(1+\varepsilon)+O(\varepsilon)$, and Theorem 25 is proved. □

4 Applications

In this section we use Theorem 23 to obtain a wide range of concrete learning results for natural and well-studied classes of distributions over both continuous and discrete domains. Throughout this section we do not aim to exhaustively cover all possible applications of Theorem 23, but rather to give some selected applications that are indicative of the generality and power of our methods.

We first (Section 4.1) give a range of applications of Theorem 23 to semi-agnostically learn various natural classes of continuous distributions. These include non-parametric classes such as concave, log-concave, and k-monotone densities, mixtures of these densities, and parametric classes such as mixtures of univariate Gaussians.

Next, turning to discrete distributions we first show (Section 4.2) how the $d = 0$ case of Theorem 23 can be easily adapted to learn *discrete* distributions that are well-approximated by piecewise flat distributions. Using this general result, we improve prior results on learning mixtures of discrete t-modal distributions, mixtures of discrete monotone hazard rate (MHR) distributions, and mixtures of discrete log-concave distributions (including mixtures of Poisson Binomial Distributions), in most cases giving essentially optimal results in terms of sample complexity. While we have not pursued this direction in the current paper, which focuses chiefly on continuous distributions, we suspect that with additional work Theorem 23 can be adapted to discrete domains in its full generality (of polynomials of degree d for arbitrary d). We conjecture that such an adaptation may give essentially optimal sample complexity bounds for all of the classes of discrete distributions that we discuss in this paper.

4.1 Applications to Distributions over Continuous Domains.

In this section we apply our general approach to obtain efficient learning algorithms for mixtures of many different types of continuous probability distributions. We focus chiefly on distributions that are defined by various kinds of "shape restrictions" on the pdf. Nonparametric density estimation for shape restricted classes has been a subject of study in statistics since the 1950s (see [BBBB72] for an early book on the topic), and has applications to a range of areas including reliability theory (see [Reb05] and references therein). The shape restrictions that have been studied in this area include monotonicity and concavity of pdfs [Gre56, Bru58, Rao69, Weg70, HP76, Gro85, Bir87a, Bir87b]. More recently, motivated by statistical applications (see e.g. Walther's recent survey [Wal09]), researchers in this area have considered other types of shape restrictions including log-concavity and k-monotonicity [BW07, DR09, BRW09, GW09, BW10, KM10].

As we will see, our general method provides a single unified approach that gives a highly-efficient algorithm (both in terms of sample complexity and computational complexity) for all the aforementioned shape restricted densities (and mixtures thereof). In most cases the sample complexities of our efficient algorithms are optimal up to log factors.

4.1.1 Concave and Log-concave Densities.

Let $I \subseteq \mathbb{R}$ be a (not necessarily finite) interval. Recall that a function $g : I \to \mathbb{R}$ is called *concave* if for any $x, y \in I$ and $\lambda \in [0, 1]$ it holds $g(\lambda x + (1 - \lambda)y) \geq \lambda g(x) + (1 - \lambda)g(y)$. A function $h : I \to \mathbb{R}_+$ is called *log-concave* if $h(x) = \exp(g(x))$, where $g : I \to \mathbb{R}$ is concave.

In this section we show that our general technique yields nearly-optimal efficient algorithms to learn (mixtures of) concave and (more generally) log-concave densities. (Because of the concavity of the log function it is easy to see that every positive and concave function is log-concave.) In particular, we show the following:

Theorem 26. *Let $f : I \to \mathbb{R}_+$ be any k-mixture of log-concave densities, where $I = [a, b]$ is an arbitrary (not necessarily finite) interval. There is an algorithm that runs in poly(k/ε) time, draws $\tilde{O}(k/\varepsilon^{5/2})$ samples from f, and with probability at least $9/10$ outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(f, h) \leq \varepsilon$.*

We note that the above sample complexity is information-theoretically optimal (up to logarithmic factors). In particular, it is known (see e.g. Chapter 15 of [DL01]) that learning a single concave density (recall that a concave density is necessarily log-concave) over $[0,1]$ requires $\Omega(\varepsilon^{-5/2})$ samples. This lower bound can be easily generalized to show that learning a k-mixture of log-concave distributions over $[0,1]$ requires $\Omega(k/\varepsilon^{5/2})$ samples. As far as we know, ours is the first computationally efficient algorithm with essentially optimal sample complexity for this problem.

To prove our result we proceed as follows: We show that any log-concave density $f : I \to \mathbb{R}_+$ has an (ε,t)-piecewise linear (degree-1) decomposition for $t = \tilde{O}(1/\sqrt{\varepsilon})$. A continuous version of the argument in Theorem 4.1 of [CDSS13] can be used to show the existence of an (ε,t)-piecewise *constant* (degree-0) decomposition with $t = \tilde{O}(1/\varepsilon)$. Unfortunately, the latter bound is essentially tight, hence cannot lead to an algorithm with sample complexity better than $\Omega(\varepsilon^{-3})$.

Classical approximation results (see e.g. [Dud74, Nov88]) provide optimal piecewise linear decompositions of concave functions. While these results have a dependence on the domain size of the function, they can rather easily be adapted to establish the existence of (ε,t)-piecewise linear decompositions for concave densities with $t = O(1/\sqrt{\varepsilon})$. However, we are not aware of prior work establishing the existence of piecewise linear decompositions for *log-concave* densities. We give such a result by proving the following structural lemma:

Lemma 27. *Let $f : I \to \mathbb{R}_+$ be any log-concave density, where $I = [a,b]$ is an arbitrary (not necessarily finite) interval. There exists an (ε,t)-piecewise linear decomposition of f for $t = \tilde{O}(1/\sqrt{\varepsilon})$.*

We note that our proof of Lemma 27 is significantly different from the aforementioned known arguments establishing the existence of piecewise linear approximations for concave functions. In particular, these proofs critically exploit concavity, namely the fact that for a concave function f, the line segment $(x, f(x))$, $(y, f(y))$ lies below the graph of the function. Before giving the proof of our lemma, we note that the $\tilde{O}(1/\sqrt{\varepsilon})$ bound is best possible (up to log factors) even for concave densities. This can be verified by considering the concave density over $[0,1]$ whose graph is given by the upper half of a circle. We further note that the [DL01] $\Omega(1/\varepsilon^{5/2})$ lower bound implies that no significant strengthening can be achieved by using our general results for learning piecewise degree-d polynomials for $d > 1$.

Theorem 26 follows as a direct corollary of Lemma 27 and Theorem 2.

Proof of Lemma 27: We begin by recalling the following fact which is a basic property (in fact an alternate characterization) of log-concave densities:

Fact 28. *([An95], Lemma 1) Let $f : \mathbb{R} \to \mathbb{R}_+$ be log-concave. Suppose that $\{x \mid f(x) > 0\} = (a,b)$. Then, for all $x_1, x_2 \in (a,b)$ with $x_1 < x_2$ and all $\delta \geq 0$ such that $x_1 + \delta, x_2 + \delta \in (a,b)$ we have*

$$\frac{f(x_1+\delta)}{f(x_1)} \geq \frac{f(x_2+\delta)}{f(x_2)}.$$

Let f be an arbitrary log-concave density over $\mathbb{R}$. Well known concentration bounds for log-concave densities (see [An95]) imply that $1-\varepsilon$ fraction of the total probability mass lies in a *finite* interval $[a,b]$. Let $m \in [a,b]$ be a mode of f so that f is non-decreasing in $[a,m]$ and non-increasing in $[m,b]$. (Recall the well-known fact [An95] that every log-concave density is unimodal, so such a mode must exist.) It suffices to analyze the second portion of the density, i.e., a non-increasing log-concave (sub)-distribution over $[m,b]$. We may further assume without loss of generality that $[m,b] = [0,1]$. (It will be clear that in what follows nothing changes in the calculations as a result of this assumption – the length of the interval is irrelevant.)

So let $f : [0,1] \to \mathbb{R}_+$ be a non-increasing log-concave density and let $c = f(0) = \max_{x\in[0,1]} f(x)$. It follows from elementary calculus that f is continuous in its support. We assume without loss of generality that f is strictly decreasing in this domain. (It follows from Fact 28 that for any non-increasing log-concave density over $[0,1]$ there exists $x_0 \in [0,1]$ such that f is constant in $[0,x_0]$ and strictly decreasing in $[x_0,1]$.)

We proceed to construct the desired piecewise-linear approximation in two stages:

(a) Let $r, s \in \mathbb{Z}_+$ with $r = \Theta((1/\varepsilon)\log(1/\varepsilon))$ and $s = \lceil \log_{1/(1-\varepsilon)} \frac{f(0)}{f(1)} \rceil = \lceil \log_{1-\varepsilon} \frac{f(1)}{f(0)} \rceil$.

We divide the domain $[0,1]$ into $t' \stackrel{\text{def}}{=} \min\{r,s\} = O((1/\varepsilon)\log(1/\varepsilon))$ intervals (disjoint except at the endpoints) $\mathcal{I} = \{I_i\}_{i=1}^{t'}$, where $I_i = [x_{i-1}, x_i]$, $i \in [t']$. The point $x_i \in [0,1]$ is the point that satisfies

$$f(x_i) = \max\{f(x_0)(1-\varepsilon)^i, f(1)\}. \tag{12}$$

Since the function is strictly decreasing and continuous, such a point exists and is unique. Note that the definition with the "max" above addresses the case that $s \leq r$. In this case, we will have that $x_{t'} = x_s = 1$. If $s > r$, then we will have that $f(x_i) = f(x_0)(1-\varepsilon)^i$ for $i \in [t']$ and $x_{t'} < 1$.

We now proceed to establish a couple of useful properties of this decomposition. The first property is that the length of the intervals I_i is non-increasing as a function of i for $i \in [t']$.

Claim 29. *For all $i \in [t'-1]$ we have that $|I_i| \geq |I_{i+1}|$.*

Proof. Consider two consecutive intervals $I_i = [x_{i-1}, x_i]$ and $I_{i+1} = [x_i, x_{i+1}]$, $i \in [t'-1]$. It is easy to see that by the definition of the intervals we have that

$$\frac{f(x_{i+1})}{f(x_i)} \geq \frac{f(x_i)}{f(x_{i-1})}$$

or equivalently

$$\frac{f(x_i + |I_{i+1}|)}{f(x_i)} \geq \frac{f(x_{i-1} + |I_i|)}{f(x_{i-1})}.$$

Since $x_{i-1} < x_i$, by Fact 28 we have

$$\frac{f(x_{i-1} + |I_i|)}{f(x_{i-1})} \geq \frac{f(x_i + |I_i|)}{f(x_i)}.$$

Combining the above two inequalities yields that $f(x_i + |I_{i+1}|) \geq f(x_i + |I_i|)$. Since f is non-increasing we conclude that $x_i + |I_{i+1}| \leq x_i + |I_i|$ and the proof is complete. □

The second property is that the probability mass that f puts in the interval $[x_{t'}, 1]$ is bounded by ε.

Claim 30. *We have that $f([x_{t'}, 1]) \leq \varepsilon$.*

Proof. We consider two cases. If $t' = s$, then $x_{t'} = 1$ and the desired probability is zero.

It thus suffices to analyze the case $t' = r$. In this case $x_{t'} < 1$ and for all $i \in [t']$ it holds $f(x_i) = f(x_0)(1-\varepsilon)^i$. Note that $f(x_{t'}) = f(0)(1-\varepsilon)^{t'} \leq f(0)\varepsilon/2 = c\varepsilon/2$. For the purposes of the analysis, suppose we decompose $[x_{t'}, 1]$ into a sequence of intervals $\{I_i\}_{i>t'}$, where $I_i = [x_{i-1}, x_i]$ and point x_i is defined by (12). That is, we have a total of s intervals $I_1, \ldots, I_s$

partitioning $[0,1]$ where by Claim 29 $|I_1| \geq |I_2| \geq \ldots \geq |I_s|$. Clearly, $\sum_{i=1}^{s} f(I_i) = 1$ and since f is non-increasing

$$c(1-\varepsilon)^i |I_i| \leq f(x_i)|I_i| \leq f(I_i) \leq f(x_{i-1})|I_i| = c(1-\varepsilon)^{i-1}|I_i|. \tag{13}$$

Combining the above yields

$$c \cdot \sum_{i=1}^{s} (1-\varepsilon)^i |I_i| \leq 1. \tag{14}$$

We want to show that $f([x_{t'}, 1]) = \sum_{i=t'+1}^{s} f(I_i) \leq \varepsilon$. Indeed, we have

$$\sum_{i=t'+1}^{s} f(I_i) \leq \sum_{i=t'+1}^{s} c(1-\varepsilon)^{i-1}|I_i| \leq \frac{c\varepsilon}{2(1-\varepsilon)} \cdot \sum_{i=1}^{s-t'} (1-\varepsilon)^i |I_{i+t'}| \tag{15}$$

where the first inequality uses (13) and the second uses the fact that $(1-\varepsilon)^{t'} \leq \varepsilon/2$. By Claim 29 it follows that $|I_{i+t'}| \leq |I_i|$ which yields

$$\sum_{i=t'+1}^{s} f(I_i) \leq \frac{c\varepsilon}{2(1-\varepsilon)} \cdot \sum_{i=1}^{s-t'} (1-\varepsilon)^i |I_i| \leq \frac{c\varepsilon}{2(1-\varepsilon)} \cdot \sum_{i=1}^{s} (1-\varepsilon)^i |I_i| \leq \varepsilon$$

where the last inequality follows from (14) for $\varepsilon \leq 1/2$. □

In fact, it is now easy to show that $\mathcal{I}$ is an $(O(\varepsilon), t')$-flat decomposition of f, but we will not make direct use of this in the subsequent analysis.

(b) In the second step, we group consecutive intervals of $\mathcal{I}$ (in increasing order of i) to obtain an $(O(\varepsilon), t)$ piecewise linear decomposition $\mathcal{J} = \{J_\ell\}_{\ell=1}^{t}$ for f, where $t = \tilde{O}(\varepsilon^{-1/2})$. Suppose that we have constructed the super-intervals $J_1, \ldots, J_{\ell-1}$ and that $\cup_{s=1}^{\ell-1} J_s = \cup_{k=1}^{i} I_k = [x_0, x_i]$. If $i = t'$ then t is set to $\ell - 1$, and if $i \leq t'$ then the super-interval J_ℓ contains the intervals $I_{i+1}, \ldots, I_j$, where $j \in \mathbb{Z}_+$ is the maximum value which is $\leq t'$ and satisfies:

(1) $f(x_j) \geq f(x_i)(1-\varepsilon)^{1/\sqrt{\varepsilon}}$, and

(2) $|I_j| \geq (1-\sqrt{\varepsilon})|I_{i+1}|$.

Within each super-interval $J_\ell = \cup_{k=i+1}^{j} I_k = [x_i, x_j]$ we approximate f by the linear function $\tilde{f}$ satisfying $\tilde{f}(x_i) = f(x_i)$ and $\tilde{f}(x_j) = f(x_j)$. This completes the description of the construction.

We proceed to show correctness. Our first claim is that it is sufficient, in the construction described in (b) above, to take only $t = \tilde{O}(\varepsilon^{-1/2})$ super-intervals, because the probability mass under f that lies to the right of the rightmost of these super-intervals is at most ε:

Claim 31. *Suppose that $t = \Omega(\varepsilon^{-1/2}\log(1/\varepsilon))$ and $J_t = [x_u, x_v]$ is the rightmost super-interval. Then, $f([x_v, 1]) \leq \varepsilon$.*

Proof. Consider a generic super-interval $J_\ell = \cup_{k=i+1}^{j} I_k$. Since j is the maximum value that satisfies both (1) and (2) we conclude that either

$$j + 1 - i > 1/\sqrt{\varepsilon} \tag{16}$$

(this inequality follows from the negation of (1) and the definition of $f(x_i)$, $f(x_j)$) or

$$|I_{j+1}| < (1-\sqrt{\varepsilon})|I_{i+1}|. \tag{17}$$

Suppose we have $t = \Omega(\varepsilon^{-1/2}\log(1/\varepsilon))$ super-intervals. Then, either (16) is satisfied for at least $t/2$ super-intervals or (17) is satisfied for at least $t/2$ super-intervals. Denote the rightmost super-interval by $J_t = [x_u, x_v]$. In the first case, for an appropriate constant in the big-Omega, we have $v = t'$ and the desired result follows from Claim 30.

In the second case, for an appropriate constant in the big-Omega we will have $|I_v| \le \varepsilon^3|I_1|$. To show that $f([x_v,1]) \le \varepsilon$ in this case, we consider further partitioning the interval $[x_v, 1]$ into a sequence of intervals $\{I_i\}_{i>v}$, where $I_i = [x_{i-1}, x_i]$ and point x_i is defined by (12). By Claim 29 we will have that $|I_i| \le |I_v|$, $i > v$. We can therefore bound the desired quantity by

$$\sum_{i=v+1}^{s} f(I_i) \le \sum_{i=v+1}^{s} c(1-\varepsilon)^{i-1}|I_i| \le \sum_{i=v+1}^{s} c(1-\varepsilon)^{i-1}\varepsilon^3|I_1| \le \varepsilon^3 c|I_1| \sum_{i=1}^{\infty}(1-\varepsilon)^{i-1} \le \frac{\varepsilon^3}{(1-\varepsilon)^2}\cdot\frac{1-\varepsilon}{\varepsilon} \le \varepsilon,$$

where the first inequality used the first inequality of (15) and the penultimate inequality uses the fact that $c(1-\varepsilon)|I_1| \le p(I_1) \le 1$. This completes the proof of the claim.

□

The main claim we are going to establish for the piecewise-linear approximation $\mathcal{J}$ is the following:

Claim 32. *For any super-interval $J_\ell = \cup_{k=i+1}^{j} I_k$ and any $i \le m \le j$ we have that*

$$|\tilde{f}(x_m) - f(x_m)| = O(\varepsilon)f(x_m).$$

Assuming the above claim it is easy to argue that $\mathcal{J}$ is indeed an $(O(\varepsilon), t)$ piecewise linear approximation to f. Let $\tilde{f}$ be the piecewise linear function over $[0,1]$ which is linear over J_ℓ (as described above) and identically zero in the interval $[x_v, 1]$.

Indeed, we have that

$$\begin{aligned}
\|\tilde{f} - f\|_1 &\le \sum_{\ell=1}^{t} \int_{J_\ell} |\tilde{f}(y) - f(y)|dy + f([x_v,1]) \\
&\le \sum_{i=1}^{v} \int_{y=x_{i-1}}^{x_i} |\tilde{f}(y) - f(y)|dy + \varepsilon \\
&\le \sum_{i=1}^{v} O(\varepsilon) f(x_m)|I_i| + \varepsilon \\
&\le \sum_{i=1}^{v} O(\varepsilon) f(x_{i-1})|I_i| + \varepsilon \\
&= O(\varepsilon)
\end{aligned}$$

where the second inequality used Claim 31, the third inequality used Claim 32, the fourth inequality used the fact that f is non-increasing, and the final inequality used the fact that

$$\sum_{i=1}^{v} f(x_{i-1})|I_i| \le \frac{1}{1-\varepsilon}\sum_{i=1}^{v} f(x_i)|I_i| \le \frac{1}{1-\varepsilon}\sum_{i=1}^{v} f(I_i) \le 1/(1-\varepsilon),$$

which follows by the definition of the $f(x_i)$'s.

We are now ready to give the proof of the claim.

Proof of Claim 32. If $\tilde{f}$ is the approximating line between x_i and x_j we can write

$$\tilde{f}(x_m) = f(x_i) + (f(x_j) - f(x_i)) \cdot \frac{\sum_{k=i+1}^{m} |I_k|}{\sum_{k=i+1}^{j} |I_k|}.$$

Note that $f(x_j) - f(x_i) = f(x_i)\big((1-\varepsilon)^{j-i} - 1\big)$. We also recall that

$$(1-\varepsilon)^{j-i} = 1 - \varepsilon(j-i) + \varepsilon^2(j-i)^2/2 + O(\varepsilon^3(j-i)^3).$$

Since i, j are in the same super-interval, we have that $j - i \leq 1/\sqrt{\varepsilon}$, which implies that the above error term is $O(\varepsilon^{3/2})$. We will use this approximation henceforth, which is also valid for any $m \in [i, j]$.

Also by condition (2) defining the lengths of the intervals in the same super-interval and the monotonicity of the lengths themselves, we obtain

$$\frac{m-i}{j-i} \cdot (1-\sqrt{\varepsilon}) \leq \frac{\sum_{k=i+1}^{m} |I_k|}{\sum_{k=i+1}^{j} |I_k|} \leq \frac{m-i}{j-i} \cdot \frac{1}{1-\sqrt{\varepsilon}}.$$

By carefully combining the above inequalities we obtain the desired result. In particular, we have that

$$\tilde{f}(x_m) \leq f(x_i)\left[1 - \varepsilon\big(1 + O(\sqrt{\varepsilon})\big)(m-i) + (\varepsilon^2/2)(j-i)(m-i)\big(1 + O(\sqrt{\varepsilon})\big) + O(\varepsilon^{3/2})\right].$$

Also

$$f(x_m) = f(x_i)\left[1 - \varepsilon(m-i) + (\varepsilon^2/2)(m-i)^2 + O(\varepsilon^{3/2})\right].$$

Therefore, using the fact that $j - i, m - i \leq 1/\sqrt{\varepsilon}$, we get that

$$\tilde{f}(x_m) - f(x_m) \leq O(\varepsilon) f(x_i).$$

In an analogous manner we obtain that

$$f(x_m) - \tilde{f}(x_m) \leq O(\varepsilon) f(x_i).$$

By the definition of a super-interval, the maximum and minimum values of f within the super-interval are within a $1 + o(1)$ factor of each other. This completes the proof of Claim 32. □

This completes the proof of Lemma 27. □

4.1.2 k-monotone Densities. Let $I = [a, b] \subseteq \mathbb{R}$ be a (not necessarily finite) interval. A function $f : I \to \mathbb{R}_+$ is said to be 1-*monotone* if it is non-increasing. It is 2-*monotone* if it is non-increasing and convex, and k-*monotone* for $k \geq 3$ if $(-1)^j f^{(j)}$ is non-negative, non-increasing and convex for $j = 0, \ldots, k-2$. The problem of density estimation for k-monotone densities has been extensively investigated in the mathematical statistics community during the past few years (see [BW07, GW09, BW10, Ser10] and references therein) due to its significance in both theory and applications [BW07]. For example, as pointed out in [BW07], the problem of learning an unknown k-monotone density arises in a generalization of Hampel's bird-watching problem [Ham87].

The aforementioned papers from the statistics community focus on analyzing the rate of convergence of the Maximum Likelihood Estimator (MLE) under various metrics. In this section we show that our approach yields an efficient algorithm to learn bounded k-monotone densities over $[0,1]$ (i.e., k-monotone densities p such that $\sup_{x\in[0,1]} p(x) = O(1)$), and mixtures thereof, with sample complexity $\tilde{O}(k/\varepsilon^{2+1/k})$. This bound is provably optimal (up to log factors) for $k = 1$ by [Bir87a] and for $k = 2$ (see e.g. Chapter 15 of [DL01]) and we conjecture that it is similarly tight for all values of k.

Our main algorithmic result for k-monotone densities is the following:

Theorem 33. *Let $k \in \mathbb{Z}_+$ and $f : [0,1] \to \mathbb{R}_+$ be a t-mixture of bounded k-monotone densities. There is an algorithm that runs in $\mathrm{poly}(k, t, 1/\varepsilon)$ time, uses $\tilde{O}(tk/\varepsilon^{2+1/k})$ samples, and outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(h, f) \le \varepsilon$.*

The above theorem follows as a corollary of Theorem 2 and the following structural result:

Lemma 34 (Implicit in [KL04, KL07])**.** *Let $f : [0,1] \to \mathbb{R}_+$ be a k-monotone density such that $\sup_x |f(x)| = O(1)$. There exists an (ε, t)-piecewise degree-$(k-1)$ approximation of f with $t = O(\varepsilon^{1/k})$.*

As we now explain the above lemma can be deduced from recent work in approximation theory [KL04, KL07]. To state the relevant theorem we need some terminology: Let $s \in \mathbb{Z}_+$, and for a real function f over interval I, let $\Delta_\tau^s f(t) = \sum_{i=0}^{s} (-1)^{s-i} \binom{s}{i} f(t + i\tau)$ be the sth difference of the function x with step $\tau > 0$, where $[t, t+s\tau] \subseteq I$. For $r \in \mathbb{Z}_+^*$, let $W_1^r(I)$ be the set of real functions f over I that are absolutely continuous in every compact subinterval of I and satisfy $\|f^{(r)}\|_1 = O(1)$. We denote by $\Delta_+^s W_1^r(I)$ the subset of functions f in $W_1^r(I)$ that satisfy $\Delta_\tau^s f(t) \ge 0$ for all $\tau > 0$ such that $[t, t+s\tau] \subseteq I$. (Note that if f is s-times differentiable the latter condition is tantamount to saying that $f^{(s)} \ge 0$.) We have the following:

Theorem 35 (Theorem 1 in [KL07])**.** *Let $s \in \mathbb{Z}_+$, $r, \nu, n \in \mathbb{Z}_+^*$ such that $\nu \ge \max\{r, s\}$. For any $f \in \Delta_+^s W_1^r(I)$ there exists a piecewise degree-$(\nu - 1)$ polynomial approximation h to f with n pieces such that $\|h - f\|_1 = O(n^{-\max\{r,s\}})$.*

(In fact, it is shown in [KL07] that the above bound is quantitatively optimal up to constant factors.) Let $f : [0,1] \to \mathbb{R}_+$ be a k-monotone density such that $\sup |f| = O(1)$. It is easy to see that Lemma 34 follows from Theorem 35 for the following setting of parameters: $s = k$, $r = 1$ and $\nu = \max\{r, s\} = k$. Indeed, since $(-1)^{k-2} f^{(k-2)}$ is convex, it follows that $\Delta_\tau^k f(t)$ is nonnegative for even k and nonpositive for odd k.

Since f is a non-increasing bounded density, it is clear that $\|f'\|_1 = |\int_0^1 f'(t)dt| = f(0) - f(1) = O(1)$. Hence, for even k Theorem 35 is applicable to f and yields Lemma 34. For odd k, Lemma 34 follows by applying Theorem 35 to the function $-f$.

4.1.3 Mixtures of Univariate Gaussians. As a final example illustrating the power and generality of Theorem 2, we now show how it very easily yields a computationally efficient and essentially optimal (up to logarithmic factors) sample complexity algorithm for learning mixtures of k univariate Gaussians. As will be evident from the proof, similar results could be obtained via our techniques for a wide range of mixture distribution learning problems for different types of parametric univariate distributions beyond Gaussians.

Lemma 36. *Let $p = N(\mu, \sigma^2)$ be a univariate Gaussian. Then p is an $(\varepsilon, 3)$-piecewise degree-d distribution for $d = O(\log(1/\varepsilon))$.*

Since Theorem 23 is easily seen to extend to semi-agnostic learning of k-mixtures of t-piecewise degree-d distributions, Lemma 36 immediately gives the following semi-agnostic learning result for mixtures of k one-dimensional Gaussians:

Theorem 37. *Let p be any distribution that has $d_{\mathrm{TV}}(p,q) \leq \varepsilon$ where q is any one-dimensional mixture of k Gaussians. There is a* $\mathrm{poly}(k, 1/\varepsilon)$*-time algorithm that uses $\tilde{O}(k/\varepsilon^2)$ samples and with high probability outputs a hypothesis h such that $d_{\mathrm{TV}}(h,p) \leq O(\varepsilon)$.*

It is straightforward to show that $\Omega(k/\varepsilon^2)$ samples are information-theoretically necessary for learning a mixture of k Gaussians, and thus our sample complexity is optimal up to logarithmic factors.

Discussion. Moitra and Valiant [MV10] recently gave an algorithm for *parameter estimation* (a stronger requirement than the density estimation guarantees that we provide) of any mixture of k *n-dimensional* Gaussians. Their algorithm has sample complexity that is exponential in k, and indeed they prove that any algorithm that does parameter estimation even for a mixture of k one-dimensional Gaussians must use $2^{\Omega(k)}$ samples. In contrast, our result shows that it is possible to perform *density estimation* for any mixture of k one-dimensional Gaussians with a computationally efficient algorithm that uses *exponentially fewer* (linear in k) samples than are required for parameter estimation. Moreover, unlike the parameter estimation results of [MV10], our density estimation algorithm is semi-agnostic: it succeeds even if the target distribution is ε-far from a mixture of Gaussians.

Proof of Lemma 36: Without loss of generality we may take p to be the standard Gaussian $N(0,1)$, which has pdf $p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Let $I_1 = (-\infty, -C\sqrt{\log(1/\varepsilon)})$, $I_2 = [-C\sqrt{\log(1/\varepsilon)}, C\sqrt{\log(1/\varepsilon)})$ and $I_3 = [C\sqrt{\log(1/\varepsilon)}, \infty)$ where $C > 0$ is an absolute constant. We define the distribution q as follows: $q(x) = 0$ for all $x \in I_1 \cup I_3$, and $q(x)$ is given by the degree-d Taylor expansion of $p(x)$ about 0 for $x \in I_2$, where $d = O(\log(1/\varepsilon))$. Clearly q is a 3-piecewise degree-d polynomial. To see that $d_{\mathrm{TV}}(p,q) \leq \varepsilon$, we first observe that by a standard Gaussian tail bound the regions I_1 and I_3 contribute at most $\varepsilon/2$ to $d_{\mathrm{TV}}(p,q)$ so it suffices to argue that

$$\int_{I_2} |p(x) - q(x)| dx \leq \varepsilon/2. \tag{18}$$

Fix any $x \in I_2$. Taylor's theorem gives that $|p(x) - q(x)| \leq p^{(d+1)}(x')x^{d+1}/(d+1)!$ for some $x' \in [0,x]$. Recalling that the $(d+1)$-st derivative $p^{(d+1)}(x')$ of the pdf of the standard Gaussian equals $H_{d+1}(x')p(x')$, where H_{d+1} is the Hermite polynomial of order $d+1$, standard bounds on the Hermite polynomials together with the fact that $|x| \leq C\sqrt{\log(1/\varepsilon)}$ give that for $d = O(\log \frac{1}{\varepsilon})$ we have $|p(x) - q(x)| \leq \varepsilon^2$ for all $x \in I_2$. This gives the lemma. □

4.2 Learning discrete distributions. For convenience in this subsection we consider discrete distributions over the $2N$-point finite domain

$$D := \left\{ -\frac{N}{N}, -\frac{N-1}{N}, \ldots, -\frac{1}{N}, 0, \frac{1}{N}, \ldots, \frac{N-1}{N} \right\}.$$

We say that a discrete distribution q over domain D is *t-flat* if there exists a partition of D into t intervals $I_1, \ldots, I_t$ such that $q(i) = q(j)$ for all $i, j \in I_\ell$ for all $\ell = 1, \ldots, t$. We say that a distribution p over D is *(ε, t)-flat* if $d_{\mathrm{TV}}(p,q) \leq \varepsilon$ for some distribution q over D that is t-flat.

We begin by giving a simple reduction from learning (ε, t)-flat distributions over D to learning (ε, t)-piecewise degree-0 distributions over $[-1, 1]$. Together with Theorem 23 this reduction gives us an essentially optimal algorithm for learning discrete (ε, t)-flat distributions (see Theorem 38).

We then apply Theorem 38 to obtain highly efficient algorithms (in most cases with provably near-optimal sample complexity) for various specific classes of discrete distributions essentially resolving a number of open problems from previous works.

4.2.1 A reduction from discrete to continuous. Given a discrete distribution p over D, we define $\tilde{p}$ to be the distribution over $[-1,1)$ defined as follows: a draw from $\tilde{p}$ is obtained by drawing a value i/N from p, and then outputting $i + x/N$ where x is distributed uniformly over $[0,1)$. It is easy to see that if distribution p (over domain D) is t-flat, then the distribution $\tilde{p}$ (over domain $[-1,1)$) is t-piecewise degree-0. Moreover, if p is τ-close to some t-flat distribution q over D, then $\tilde{p}$ is τ-close to $\tilde{q}$.

In the opposite direction, for p a distribution over $[-1,1)$ we define p^* to be the following distribution supported on D: a draw from p^* is obtained by sampling x from p and then outputting the value obtained by rounding x down to the next integer multiple of $1/N$. It is easy to see that if p, q are distributions over $[-1,1)$ then $d_{\mathrm{TV}}(p,q) = d_{\mathrm{TV}}(p^*,q^*)$. It is also clear that for p a distribution over D we have $(\tilde{p})^* = p$.

With these relationships in hand, we may learn a (τ,t)-flat distribution p over D as follows: run Algorithm `Learn-Piecewise-Poly`$(t, d=0, \varepsilon)$ on the distribution $\tilde{p}$. Since p is (τ,t)-flat, $\tilde{p}$ is τ-close to some t-piecewise degree-0 distribution q over $[-1,1)$, so the algorithm with high probability constructs a hypothesis h over $[-1,1)$ such that $d_{\mathrm{TV}}(h,\tilde{p}) \le O(\tau+\varepsilon)$. The final hypothesis is h^*; for this hypothesis we have

$$d_{\mathrm{TV}}(h^*, p) = d_{\mathrm{TV}}(h^*, (\tilde{p})^*) = d_{\mathrm{TV}}(h, \tilde{p}) \le O(\tau+\varepsilon)$$

as desired.

The above discussion and Theorem 23 together give the following:

Theorem 38. *Let p be a mixture of k (τ,t)-flat discrete distributions over D. There is an algorithm which uses $\tilde{O}(kt/\varepsilon^2)$ samples from p, runs in time* $\mathrm{poly}(k,t,1/\varepsilon)$*, and with probability at least* $9/10$ *outputs a hypothesis distribution h over D such that $d_{\mathrm{TV}}(p,h) \le O(\varepsilon+\tau)$.*

We note that this is essentially a stronger version of Corollary 3.1 (the main technical result) of [CDSS13], which gave a similar guarantee but with an algorithm that required $O(kt/\varepsilon^3)$ samples. We also remark that $\Omega(kt/\varepsilon^2)$ samples are information-theoretically required to learn an arbitrary k-mixture of t-flat distributions. Hence, our sample complexity is optimal up to logarithmic factors (even for the case $\tau = 0$).

We would also like to mention the relation of the above theorem to a recent work by Indyk, Levi and Rubinfeld [ILR12]. Motivated by a database application, [ILR12] consider the problem of learning a k-flat distribution over $[n]$ *under the L_2 norm* and give an efficient algorithm that uses $O(k^2 \log(n)/\varepsilon^4)$ samples. Since the total variation distance is a stronger metric, Theorem 38 immediately implies an improved sample bound of $\tilde{O}(k/\varepsilon^2)$ for their problem.

4.2.2 Learning specific classes of discrete distributions.

Mixtures of t-modal discrete distributions. Recall that a distribution over an interval $I = [a,b] \cap D$ is said to be *unimodal* if there is a value $y \in I$ such that its pdf is monotone non-decreasing on $I \cap [-1,y]$ and monotone non-increasing on $I \cap (y,1)$. For $t > 1$, a distribution p over D is t-modal if there is a partition of D into t intervals $I_1,\dots,I_t$ such that the conditional distributions $p_{I_1},\dots,p_{I_t}$ are each unimodal.

In [CDSS13, DDS$^+$13] (building on [Bir87b]) it is shown that every t-modal distribution over D is $(\varepsilon, t\log(N)/\varepsilon)$-flat. By using this fact together with Theorem 38 in place of Corollary 3.1 of

[CDSS13], we improve the sample complexity of the [CDSS13] algorithm for learning mixtures of t-modal distributions and obtain the following:

Theorem 39. *For any $t \geq 1$, let p be any k-mixture of t-modal distributions over D. There is an algorithm that runs in time* $\mathrm{poly}(k, t, \log N, 1/\varepsilon)$*, draws $\tilde{O}(kt \log(N)/\varepsilon^3)$ samples from p, and with probability at least* $9/10$ *outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(p, h) \leq \varepsilon$.*

We note that an easy adaptation of Birgé's lower bound [Bir87a] for learning monotone distributions (see the discussion at the end of Section 5 of [CDSS13]) gives that any algorithm for learning a k-mixture of t-modal distributions over D must use $\Omega(kt \log(N/(kt))/\varepsilon^3)$ samples, and hence the sample complexity bound of Theorem 39 is optimal up to logarithmic factors. We further note that even the $t = 1$ case of this result compares favorably with the main result of [DDS12a], which gave an algorithm for learning t-modal distributions over D that uses $O(t \log(N)/\varepsilon^3) + \tilde{O}(t^3/\varepsilon^3)$ samples. The [DDS12a] result gave an optimal bound only for small settings of t, specifically $t = \tilde{O}((\log N)^{1/3})$, and gave a quite poor bound as t grows large; for example, at $t = (\log N)^2$ the optimal bound would be $O((\log N)^3/\varepsilon^3)$ but the [DDS12a] result only gives $\tilde{O}((\log N)^9/\varepsilon^3)$. In contrast, our new result gives an essentially optimal bound (up to log factors in the optimal sample complexity) for *all* settings of t.

Mixtures of monotone hazard rate distributions. Let p be a distribution supported on D. The *hazard rate* of p is the function $H(i) \stackrel{\mathrm{def}}{=} \frac{p(i)}{\sum_{j \geq i} p(j)}$; if $\sum_{j \geq i} p(j) = 0$ then we say $H(i) = +\infty$. We say that p has *monotone hazard rate* (MHR) if $H(i)$ is a non-decreasing function over D.

[CDSS13] showed that every MHR distribution over D is $(\varepsilon, O(\log(N/\varepsilon)/\varepsilon))$-flat. Theorem 38 thus gives us the following:

Theorem 40. *Let p be any k-mixture of MHR distributions over D. There is an algorithm that runs in time* $\mathrm{poly}(k, \log N, 1/\varepsilon)$*, draws $\tilde{O}(k \log(N)/\varepsilon^3)$ samples from p, and with probability at least* $9/10$ *outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(p, h) \leq \varepsilon$.*

In [CDSS13] it is shown that any algorithm to learn k-mixtures of MHR distributions over D must use $\Omega(k \log(N/k)/\varepsilon^3)$ samples, so Theorem 40 is essentially optimal in its sample complexity.

Mixtures of discrete log-concave distributions. A probability distribution p over D is said to be *log-concave* if it satisfies the following conditions: (i) if $i < j < k \in D$ are such that $p(i)p(k) > 0$ then $p(j) > 0$; and (ii) $p(k/N)^2 \geq p((k-1)/N)p((k+1)/N)$ for all $k \in \{-N+1, \ldots, -1, 0, 1, \ldots, N-2\}$.

In [CDSS13] it is shown that every log-concave distribution over D is $(\varepsilon, O(\log(1/\varepsilon))/\varepsilon)$-flat. Hence Theorem 38 gives:

Theorem 41. *Let p be any k-mixture of log-concave distributions over D. There is an algorithm that runs in time* $\mathrm{poly}(k, 1/\varepsilon)$*, draws $\tilde{O}(k/\varepsilon^3)$ samples from p, and with probability at least* $9/10$ *outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(p, h) \leq \varepsilon$.*

As in the previous examples, this improves the [CDSS13] sample complexity by essentially a factor of $1/\varepsilon$. We note that as a special case of Theorem 41 we get an efficient $O(k/\varepsilon^3)$-sample algorithm for learning any mixture of k *Poisson Binomial Distributions.* (A Poisson Binomial Distribution, or PBD, is a random variable of the form $X_1 + \cdots + X_N$ where the X_i's are independent 0/1 random variables that may have arbitrary and non-identical means.) The main result of [DDS12b] gave an efficient $\tilde{O}(1/\varepsilon^3)$-sample algorithm for learning a single PBD; here we achieve the same sample complexity, with an efficient algorithm, for learning any mixture of any constant number of PBDs.

Acknowledgements. We would like to thank Dany Leviatan for useful correspondence regarding his recent works [KL04, KL07].

References

[AK03] Sanjeev Arora and Subhash Khot. Fitting algebraic curves to noisy data. *J. Comput. Syst. Sci.*, 67(2):325–340, 2003.

[An95] M. Y. An. Log-concave probability distributions: Theory and statistical testing. Technical Report Economics Working Paper Archive at WUSTL, Washington University at St. Louis, 1995.

[Ass83] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I*, 296:1021–1024, 1983.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions.* Wiley, New York, 1972.

[Bir87a] L. Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

[Bir87b] L. Birgé. On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, 15(3):1013–1022, 1987.

[Bru58] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958.

[BRW09] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009.

[BS10] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[BW07] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007.

[BW10] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, 2013.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In SODA, 2012.

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS$^+$13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA, to appear*, 2013.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View.* John Wiley & Sons, 1985.

[DGJ$^+$10] I. Diakoniokolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics, Springer, 2001.

[DR09] L. D umbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[Dud74] R.M Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227 – 236, 1974.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 183–192, 1999.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005.

[Gre56] U. Grenander. On the theory of mortality measurement. *Skand. Aktuarietidskr.*, 39:125–153, 1956.

[Gro85] P. Groeneboom. Estimating a monotone density. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 539–555, 1985.

[GW09] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52:1525–1538, 2009.

[Ham87] F. R. Hampel. Design, data & analysis. chapter Design, modelling, and analysis of some biological data sets, pages 93–128. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[HP76] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6):pp. 1038–1050, 1976.

[ILR12] P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *PODS*, pages 15–22, 2012.

[Jac97] J. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55:414–440, 1997.

[KL04] V. N. Konovalov and D. Leviatan. Free-knot splines approximation of s-monotone functions. *Adv. Comput. Math.*, 20(4):347–366, 2004.

[KL07] V. N. Konovalov and D. Leviatan. Freeknot splines approximation of sobolev-type classes of s -monotone functions. *Adv. Comput. Math.*, 27(2):211–236, 2007.

[KM93] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. on Computing*, 22(6):1331–1348, 1993.

[KM10] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Statist.*, 38(5):2998–3027, 2010.

[KMR$^+$94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Symposium on Theory of Computing*, pages 273–282, 1994.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[KOS04] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004.

[KS04] A. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *Journal of Computer & System Sciences*, 68(2):303–318, 2004.

[LMN93] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[MOS04] E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *Journal of Computer & System Sciences*, 69(3):421–434, 2004. Preliminary version in *Proc. STOC'03*.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, 1995.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[Nov88] E. Novak. *Deterministic and Stochastic Error Bounds In Numerical Analysis*. Springer-Verlag, 1988.

[PA13] D. Papp and F. Alizadeh. Shape constrained estimation using nonnegative splines. *Journal of Computational and Graphical Statistics*, 0(ja):null, 2013.

[Rao69] B.L.S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Ser. A*, 31:23–36, 1969.

[Reb05] L. Reboul. Estimation of a function under shape restrictions. Applications to reliability. *Ann. Statist.*, 33(3):1330–1356, 2005.

[Sco92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

[Ser10] A. Seregin. Uniqueness of the maximum likelihood estimator for k-monotone densities. *Proceedings of The American Mathematical Society*, 138:4511–4511, 2010.

[Sil86] B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[Wal09] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

[Weg70] E.J. Wegman. Maximum likelihood estimation of a unimodal density. I. and II. *Ann. Math. Statist.*, 41:457–471, 2169–2174, 1970.

A Omitted proofs

A.1 Proof of Lemma 6. Recall Lemma 6:

Lemma 6. *Given $0 < \kappa < 1$ and access to samples from an $\kappa/64$-well-behaved distribution p over $[-1, 1)$, the procedure* `Approximately-Equal-Partition` *uses $\tilde{O}(1/\kappa)$ samples from p, runs in time $\tilde{O}(1/\kappa)$, and with probability at least $99/100$ outputs a partition of $[-1, 1)$ into $\ell = \Theta(1/\kappa)$ intervals such that $p(I_j) \in [\frac{1}{2\kappa}, \frac{3}{\kappa}]$ for all $1 \leq j \leq \ell$.*

Proof of Lemma 6: Let n denote $1/\kappa$ (we assume wlog that n is an integer). Let S be a sample of $m = \Theta(n \log n)$ i.i.d. draws from p, where m is an integer multiple of n. For $1 \leq i \leq m$ let $U_{(i)}$ denote the i-th order statistic of S, i.e. the i-th largest element of S. Let $U_{(0)} := -1$.

Our goal is to show that with high probability, for each $j \in \{1, \ldots, n\}$ we have $p([U_{(\frac{j-1}{n} \cdot m)}, U_{(\frac{j}{n} \cdot m)})) \in [\frac{1}{2n}, \frac{2}{n}]$. This means that simply greedily taking the intervals I_1, $I_2, \ldots$ from left to right, where the left endpoint of I_0 is -1, the left (closed) endpoint of the j-th interval is the same as the right (open) endpoint of the $(j-1)$st interval, and the j-th interval ends at $U_{(\frac{j}{n} \cdot m)}$, the resulting intervals have probability masses as desired. (These intervals cover $[-1, U_{(m)}]$; an easy argument shows that with probability at least $1 - 1/n$, the uncovered region $(U_{(m)}, 1)$ has mass at most $1/n$ under p , so we may add it to the final interval.)

Let P denote the cumulative density functions associated with p. For $0 \leq \alpha < \beta \leq 1$ let $\#_S[\alpha, \beta)$ denote the number of elements $x \in S$ that have $P(x) \in [\alpha, \beta)$. A multiplicative Chernoff bound and a union bound together straightforwardly give that with probability at least $99/100$, for each $i \in \{1, \ldots, 8n\}$ we have $\#_S[\frac{i-1}{8n}, \frac{i}{8n}) \in [\frac{1}{16} \cdot \frac{m}{n}, \frac{3}{16} \cdot \frac{m}{n}]$. (Note that since p is $\frac{1}{64n}$-well-behaved, the amount of mass that p puts on $P^{-1}([\frac{i-1}{8n}, \frac{i}{8n}))$ lies in $[\frac{3}{32n}, \frac{5}{32n}]$.) As an immediate consequence of this we get that $p([U_{(\frac{j-1}{n} \cdot m)}, U_{(\frac{j}{n} \cdot m)}]) \in [\frac{1}{2n}, \frac{2}{n}]$ for each $j \in \{1, \ldots, n\}$, which establishes the lemma. □

A.2 Proof of Theorem 8. Recall Theorem 8:

Theorem 8. *Let p be an unknown t-piecewise degree-d distribution over $[-1, 1)$ where $t \geq 1$, $d \geq 0$ satisfy $t + d > 1$. Let L be any algorithm which, given as input t, d, ε and access to independent samples from p, outputs a hypothesis distribution h such that $\mathbb{E}[d_{\mathrm{TV}}(p, h)] \leq \varepsilon$, where the expectation is over the random samples drawn from p and any internal randomness of L. Then L must use at least $\Omega(\frac{t(d+1)}{(1+\log(d+1))^2} \cdot \frac{1}{\varepsilon^2})$ samples.*

We first observe that if $d = 0$ then the claimed $\Omega(t/\varepsilon^2)$ lower bound follows easily from the standard fact that this many samples are required to learn an unknown distribution over the t-element set $\{1, \ldots, t\}$. (This fact follows easily from Assouad's lemma; we will essentially prove it using Assouad's lemma in Section A.2.1 below.) Thus we may assume below that $d > 0$; in fact, we can (and do) assume that $d \geq C$ where C may be taken to be any fixed absolute constant.

In what follows we shall use Assouad's lemma to establish an $\Omega(\frac{d}{(\log d)^2} \cdot \frac{1}{\varepsilon^2})$ lower bound for learning a single degree-d distribution over $[-1, 1)$ to accuracy ε. The same argument applied to a concatenation of t equally weighted copies of this lower bound construction over t disjoint intervals $[-1, -1 + \frac{2}{t}), \ldots, [1 - \frac{2}{t}, 1)$ (again using Assouad's lemma) yields Theorem 8. Thus to prove Theorem 8 for general t it is enough to prove the following lower bound, corresponding to $t = 1$. (For ease of exposition in our later arguments, we take the domain of p below to be the interval $[0, 2k)$ rather than $[-1, 1)$.)

Theorem 42. *Fix an integer $d \geq C$. Let p be an unknown degree-d distribution over $[0, 2k)$. Let L be any algorithm which, given as input d, ε and access to independent samples from p, outputs* *a hypothesis distribution h such that $\mathbb{E}[d_{\mathrm{TV}}(p,h)] \le \varepsilon$. Then L must use at least $\Omega(\frac{d}{(\log d)^2} \cdot \frac{1}{\varepsilon^2})$ samples.*

Our main tool for proving Theorem 42 is Assouad's Lemma [Ass83]. We recall the statement of Assouad's Lemma from [DG85] below. (The statement below is slightly tailored to our context, in that we have taken the underlying domain to be $[0, 2k)$ and the partition of the domain to be $[0, 2), [2, 4), \ldots, [2k-2, 2k)$.)

Theorem 43. *[Theorem 5, Chapter 4, [DG85]] Let $k \ge 1$ be an integer. For each $b = (b_1, \ldots, b_k) \in \{-1, 1\}^k$, let p_b be a probability distribution over $[0, 2k)$.*

Suppose that the distributions p_b satisfy the following properties: Fix any $\ell \in [k]$ and any $b \in \{-1, 1\}^k$ with $b_\ell = 1$. Let $b' \in \{-1, 1\}^k$ be the same as b but with $b'_\ell = -1$. The properties are that

1. *$\int_{2\ell-2}^{2\ell} |p_b(x) - p_{b'}(x)| dx \ge \alpha$, and*

2. *$\int_0^{2k} \sqrt{p_b(x) p_{b'}(x)} dx \ge 1 - \gamma > 0$.*

Then for any any algorithm L that draws n samples from an unknown $p \in \{p_b\}_{b \in \{-1,1\}^k}$ and outputs a hypothesis distribution h, there is some $b \in \{-1, 1\}^k$ such that if the target distribution p is p_b, then

$$\mathbb{E}[d_{\mathrm{TV}}(p_b, h)] \ge (k\alpha/4)(1 - \sqrt{2n\gamma}). \tag{19}$$

We will use this lemma in the following way: Fix any $d \ge C$ and any $0 < \varepsilon < 1/2$. We will exhibit a family of 2^k distributions p_b, where each p_b is a degree-d polynomial distribution and $k = \Theta(d/(\log d)^2)$. We will show that all pairs $b, b' \in \{-1, 1\}^k$ as specified in Theorem 43 satisfy condition (1) with $\alpha = \Omega(\varepsilon/k)$, and satisfy condition (2) with $\gamma = O(\varepsilon^2/k)$. With these conditions, consider an algorithm L that draws $n = 1/(8\gamma)$ samples from the unknown target distribution p. The right-hand side of (19) simplifies to $k\alpha/8 = \Omega(\varepsilon)$, and hence by Theorem 43, the expected variation distance error of algorithm L's hypothesis h is $\Omega(\varepsilon)$. This yields Theorem 42.

Thus, in the rest of this subsection, to prove Theorem 42 and thus establish Theorem 8, it suffices for us to describe the 2^k distributions p_b and establish conditions (1) and (2) with the claimed bounds $\alpha = \Omega(\varepsilon/k)$ and $\gamma = O(\varepsilon^2/k)$. We do this below.

A.2.1 The idea behind the construction.

We provide some intuition before entering into the details of our construction. Intuitively, each polynomial p_b (for a given $b \in \{-1, 1\}^k$) is an approximation, over the interval $[0, 2k)$ of interest, of a $2k$-piecewise constant distribution S_b that we describe below. To do this, first let us define the $2k$-piecewise constant distribution

$$R_b(x) = R_{b,1}(x) + \ldots + R_{b,k}(x)$$

over $[0, 2k)$, where $R_{b,i}(x)$ is a function which is 0 outside of the interval $[2i-2, 2i)$. For $x \in [2i-2, 2i-1)$ we have $R_{b,i}(x) = (1 + b_i \cdot \varepsilon)/(2k)$, and for $x \in [2i-1, 2i)$ we have $R_{b,i}(x) = (1 - b_i \cdot \varepsilon)/(2k)$. So note that regardless of whether b_i is 1 or -1, we have $\int_{2i-2}^{2i} R_{b,i}(x) dx = 1/k$ and hence $\int_0^{2k} R_b(x) dx = 1$, so R_b is indeed a probability distribution over the domain $[0, 2k)$.

The distribution S_b over $[0, 2k)$ is defined as

$$S_b(x) = \frac{1}{10} \cdot \frac{1}{2k} + \frac{9}{10} \cdot R_b(x). \tag{20}$$

(The reason for "mixing" R_b with the uniform distribution will become clear later; roughly, it is to control the adverse effect on condition (2) of having only a polynomial approximation p_b instead of the actual piecewise constant distribution.)

To motivate the goal of constructing polynomials p_b that approximate the piecewise constant distributions S_b, let us verify that the distributions $\{S_b\}_{b\in\{-1,1\}^k}$ satisfy conditions (1) and (2) of Theorem 43 with the desired parameters. So fix any $b \in \{-1,1\}^k$ with $b_\ell = 1$ and let $b' \in \{-1,1\}^k$ differ from b precisely in the ℓ-th coordinate. For (1), we immediately have that

$$\int_{2\ell-2}^{2\ell} |S_b(x) - S_{b'}(x)|dx = \frac{9}{10}\int_{2\ell-2}^{2\ell} |R_{b,\ell}(x) - R_{b',\ell}(x)|dx = \frac{9}{5}\cdot\frac{\varepsilon}{k}.$$

For (2), we have that for any two distributions f, g,

$$\int_0^{2k} \sqrt{f(x)g(x)}dx = 1 - h(f,g)^2$$

where $h(f,g)^2$ is the squared Hellinger distance between f and g,

$$h(f,g)^2 = \frac{1}{2}\int_0^{2k} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx.$$

Applying this to S_b and $S_{b'}$, we get

$$\begin{aligned} h(S_b, S_{b'})^2 &= \frac{1}{2}\int_0^{2k} \left(\sqrt{S_b(x)} - \sqrt{S_{b'}(x)}\right)^2 dx \\ &= \frac{1}{2}\int_{2\ell-2}^{2\ell} \left(\sqrt{\frac{1}{20k} + \frac{9}{10}\cdot\frac{1+\varepsilon}{2k}} - \sqrt{\frac{1}{20k} + \frac{9}{10}\cdot\frac{1-\varepsilon}{2k}}\right)^2 dx \\ &= \Theta(\varepsilon^2/k), \end{aligned}$$

as desired. We now turn to the actual construction.

A.2.2 The construction.

Fix any $b \in \{-1,1\}^k$. Our goal is to give a degree-d polynomial p_b that is a high-quality approximator of $S_b(x)$ over $[0, 2k)$. We shall do this by approximating each $R_{b,i}(x)$ and combining the approximators in the obvious way.

We can write each $R_{b,i}(x)$ as $R_{b,i,1}(x) + R_{b,i,2}(x)$ where $R_{b,i,1}(x)$ is 0 outside of $[2i-2, 2i-1)$ and $R_{b,i,2}(x)$ is 0 outside of $[2i-1, 2i)$. So $R_b(x)$ is the sum of $2k$ many functions each of which is of the form $\omega_{b,j} \cdot I_j(x)$, i.e.

$$R_b(x) = \sum_{j=1}^{2k} \omega_{b,j} \cdot I_j(x) \tag{21}$$

where each $\omega_{b,j}$ is either $(1+\varepsilon)/2k$ or is $(1-\varepsilon)/2k$ and I_j is the indicator function of the interval $[j-1, j)$: i.e. $I_j(x) = 1$ if $x \in [j-1, j)$ and is 0 elsewhere.

We shall approximate each indicator function $I_j(x)$ over $[0, 2k)$ by a low-degree univariate polynomial which we shall denote $\tilde{I}_j(x)$; then we will multiply each $\tilde{I}_j(x)$ by $\omega_{b,j}$ and sum the results to obtain our polynomial approximator $\tilde{R}_b(x)$ to $R_b(x)$, i.e.

$$\tilde{R}_b(x) := \sum_{j=1}^{2k} \omega_{b,j}\tilde{I}_j(x). \tag{22}$$

The starting point of our construction is the polynomial whose existence is asserted in Lemma 3.7 of [DGJ$^+$10]; this is essentially a low-degree univariate polynomial which is a high-accuracy approximator to the function $\mathrm{sign}(x)$ over $[-1, 1)$ except for values of x that have small absolute value. Taking $k = M \log(1/\varepsilon)$ in Claim 3.8 of [DGJ$^+$10] for M a sufficiently large constant (rather than $M = 15$ as is done in [DGJ$^+$10]), the construction employed in the proof of Lemma 3.7 gives the following:

Fact 44. *For $0 \le \tau \le c$, where $c < 1$ is an absolute constant, there is a polynomial $A(x)$ of degree $O((\log(1/\tau))^2/\tau)$ such that*

1. *For all $x \in [-1, -\tau)$ we have $A(x) \in [-1, -1 + \tau^{10}]$;*
2. *For all $x \in (\tau, 1]$ we have $A(x) \in [1 - \tau^{10}, 1]$;*
3. *For all $x \in [-\tau, \tau]$ we have $A(x) \in [-1, 1]$.*

For $-1/4 \le \theta \le 1/4$ let $B_\theta(x)$ denote the polynomial $B_\theta(x) = (A(x) - A(x - \theta))/2$. Given Fact 44, it is easy to see that $B_\theta(x)$ has degree $O((\log(1/\tau))^2/\tau)$ and, over the interval $[-1/2, 1/2]$, is a high-accuracy approximation to the indicator function of the interval $[0, \theta]$ except on "error regions" of width at most τ at each of the endpoints $0, \theta$.

Next, recall that $k = \Theta(d/(\log d)^2)$ where d is at least some universal constant C. Choosing $\tau = \delta/k$ for a suitably small positive absolute constant δ, and performing a suitable linear scaling and shifting of the polynomial $B_\theta(x)$, we get the following:

Fact 45. *Fix any integer $1 \le j \le 2k$. There is a polynomial $C_j(x)$ of degree at most d which is such that*

1. *For $x \in [j - 0.999, j - 0.001)$ we have $C_j(x) \in [1 - 1/k^5, 1)]$;*
2. *For $x \in [0, j - 1) \cup [j, 2k)$ we have $C_j(x) \in [0, 1/k^5]$;*
3. *For $x \in [j - 1, j - 0.999) \cup [j - 0.001, j)$ we have $0 \le C_j(x) \le 1$.*

The desired polynomial $\tilde{I}_j(x)$ which is an approximator of the indicator function $I_j(x)$ is obtained by renormalizing C_j so that it integrates to 1 over the domain $[0, 2k)$; i.e. we define

$$\tilde{I}_j(x) = C_j(x) / \int_0^{2k} C_j(x) dx. \tag{23}$$

By Fact 45 we have that $\int_0^{2k} C_j(x)dx \in [0.997, 1.003]$, and thus we obtain the following:

Fact 46. *Fix any integer $1 \le j \le 2k$. The polynomial $\tilde{I}_j(x)$ has degree at most d and is such that*

1. *For $x \in [j - 0.999, j - 0.001)$ we have $\tilde{I}_j(x) \in [0.996, 1.004)]$;*
2. *For $x \in [0, j - 1) \cup [j, 2k)$ we have $\tilde{I}_j(x) \in [0, 1/k^4]$;*
3. *For $x \in [j - 1, j - 0.999) \cup [j - 0.001, j)$ we have $0 \le \tilde{I}_j \le 1.004$; and*
4. *$\int_0^{2k} \tilde{I}_j(x)dx = 1$.*

Recall that from (22) the polynomial approximator $\tilde{R}_b(x)$ for $R_b(x)$ is defined as $\tilde{R}_b(x) = \sum_{j=1}^{2k} \omega_{b,j}\tilde{I}_j(x)$. We define the final polynomial $p_b(x)$ as

$$p_b(x) = \frac{1}{10} \cdot \frac{1}{2k} + \frac{9}{10} \cdot \tilde{R}_b(x). \tag{24}$$

Since $\sum_{j=1}^{2k} \omega_{b,j} = 1$ for every $b \in \{-1,1\}^k$, the polynomial p_b does indeed define a legitimate probability distribution over $[0, 2k)$.

It will be useful for us to take the following alternate view on $p_b(x)$. Define

$$\tilde{J}_j(x) = \frac{1}{10} \cdot \frac{1}{2k} + \frac{9}{10} \cdot \tilde{I}_j(x). \tag{25}$$

Recalling that $\sum_{j=1}^{2k} \omega_{b,j} = 1$, we may alternately define p_b as

$$p_b(x) = \sum_{j=1}^{2k} \omega_{b,j}\tilde{J}_j(x). \tag{26}$$

The following is an easy consequence of Fact 46:

Fact 47. *Fix any $1 \le j \le 2k$. The polynomial $\tilde{J}_j(x)$ has degree at most d and is such that*

1. *For $x \in [j - 0.999, j - 0.001)$ we have $\tilde{J}_j(x) \in [0.896 + 0.1/(2k), 0.9004 + 0.1/(2k)]$;*
2. *For $x \in [0, j-1) \cup [j, 2k)$ we have $\tilde{I}_j(x) \in [0.1/(2k), 0.1/(2k) + 1/k^4]$;*
3. *For $x \in [j-1, j-0.999) \cup [j-0.001, j)$ we have $\tilde{J}_j(x) \in [0.1/(2k), 0.9004 + 0.1/(2k))$; and*
4. *$\int_0^{2k} \tilde{J}_j(x)dx = 1$.*

We are now ready to prove that the distributions $\{p_b\}_{b\in\{-1,1\}^k}$ satisfy properties (1) and (2) of Assouad's lemma with $\alpha = \Omega(\varepsilon/k)$ and $\gamma = O(\varepsilon^2/k)$ as described in the discussion following Theorem 43. Fix $b \in \{-1,1\}^k$ with $b_\ell = 1$ and $b' \in \{-1,1\}^k$ which agrees with b except in the ℓ-th coordinate. We establish properties (1) and (2) in the following two claims:

Claim 48. *We have $\int_{2\ell-2}^{2\ell} |p_b(x) - p_{b'}(x)|dx \ge \Omega(\varepsilon/k)$.*

Proof. Recall from (26) that

$$p_b(x) = \sum_{j=1}^{2k} \omega_{b,j} \cdot \tilde{J}_j(x) \quad \text{and} \quad p_{b'}(x) = \sum_{j=1}^{2k} \omega_{b',j} \cdot \tilde{J}_j(x).$$

We have that $\omega_{b,j} = \omega_{b',j}$ for all but exactly two (adjacent) values of j, which are $j = 2\ell - 1$ and $j = 2\ell$. For those values we have

$$\omega_{b,2\ell-1} = (1+\varepsilon)/(2k), \quad \omega_{b',2\ell-1} = (1-\varepsilon)/(2k)$$

while

$$\omega_{b,2\ell} = (1-\varepsilon)/(2k), \quad \omega_{b',2\ell} = (1+\varepsilon)/(2k).$$

So we have

$$\begin{aligned}\int_{2\ell-2}^{2\ell} |p_b(x) - p_{b'}(x)|dx &= \int_{2\ell-2}^{2\ell} |(\omega_{b,2\ell-1}\tilde{J}_{2\ell-1}(x) + \omega_{b,2\ell}\tilde{J}_{2\ell}(x)) - (\omega_{b',2\ell-1}\tilde{J}_{2\ell-1}(x) + \omega_{b',2\ell}\tilde{J}_{2\ell}(x))|dx \\ &= (\varepsilon/k) \cdot \int_{2\ell-2}^{2\ell} |\tilde{J}_{2\ell-1}(x) - \tilde{J}_{2\ell}(x)|dx.\end{aligned}$$

Claim 48 now follows immediately from

$$\int_{2\ell-2}^{2\ell} |\tilde{J}_{2\ell-1}(x) - \tilde{J}_{2\ell}(x)|dx = \Omega(1),$$

which is an easy consequence of Fact 47. □

Claim 49. *We have* $\int_0^{2k} \sqrt{p_b(x)p_{b'}(x)}dx \geq 1 - O(\varepsilon^2/k)$*, i.e.* $h(p_b, p_{b'})^2 \leq O(\varepsilon^2/k)$*.*

Proof. As above $\omega_{b,j} = \omega_{b',j}$ for all but exactly two (adjacent) values of j which are $j = 2\ell - 1$ and $j = 2\ell$. For those values we have

$$\omega_{b,2\ell-1} = (1+\varepsilon)/(2k), \quad \omega_{b',2\ell-1} = (1-\varepsilon)/(2k), \quad \omega_{b,2\ell} = (1-\varepsilon)/(2k), \quad \omega_{b',2\ell} = (1+\varepsilon)/(2k).$$

We have

$$h(p_b, p_{b'})^2 = \frac{1}{2}\int_0^{2k} \left(\sqrt{p_b}(x) - \sqrt{p_{b'}(x)}\right)^2 dx = A/2 + B/2,$$

where

$$A = \int_{[2k]\setminus[2\ell-2,2\ell)} \left(\sqrt{p_b(x)} - \sqrt{p_{b'}(x)}\right)^2 dx$$

and

$$B = \int_{[2\ell-2,2\ell]} \left(\sqrt{p_b(x)} - \sqrt{p_{b'}(x)}\right)^2 dx.$$

We first bound B, by upper bounding the value of the integrand $\left(\sqrt{p_b(x)} - \sqrt{p_{b'}(x)}\right)^2$ on any fixed $x \in [2k] \setminus [2\ell-2, 2\ell]$. Recall that $p_b(x)$ is a sum of the $2k$ values $\omega_{b,j} \cdot \tilde{J}_j(x)$. The $0.1/(2k)$ contribution to each $\tilde{J}_j(x)$ ensures that $p_b(x) \geq 0.1/(2k)$ for all $x \in [0, 2k]$, and it is easy to see from the construction that $p_b(x) \leq 2/(2k)$ for all $x \in [0, 2k]$. The difference between the values $p_b(x)$ and $p_{b'}(x)$ comes entirely from $(\varepsilon/k)(\tilde{J}_{2\ell-1}(x) - \tilde{J}_{2\ell}(x))$, which has magnitude at most $(\varepsilon/k) \cdot (1/k^4) = \varepsilon/k^5$. So we have that $\left(\sqrt{p_b}(x) - \sqrt{p_{b'}(x)}\right)^2$ is at most the following (where $c_x \in [0.1, 2]$ for each $x \in [2k] \setminus [2\ell-2, 2\ell]$):

$$\left[\sqrt{\frac{c_x}{k} + \frac{\varepsilon}{k^5}} - \sqrt{\frac{c_x}{k}}\right]^2 = (c_x/k) \cdot \left[\sqrt{1 + \frac{\varepsilon}{c_x k^4}} - 1\right]^2 = (c_x/k) \cdot [\Theta(\varepsilon/k^4)]^2 = \Theta(\varepsilon^2/k^9).$$

Integrating over the region of width $2k - 2$, we get that $B = O(\varepsilon^2/k^8)$.

It remains to bound A. Fix any $x \in [2\ell-2, 2\ell]$. As above we have that $p_b(x)$ equals c_x/k for some $c_x \in [0.1, 2]$, and (26) implies that $p_b(x)$ and $p_{b'}(x)$ differ by at most $\Theta(\varepsilon/k)$. So we have

$$\left(\sqrt{p_b}(x) - \sqrt{p_{b'}(x)}\right)^2 \leq \left[\sqrt{\frac{c_x}{k}} - \sqrt{\frac{c_x}{k} - \frac{\Theta(\varepsilon)}{k}}\right]^2 = \frac{c_x}{k}\left[1 - \sqrt{1 - \Theta(\varepsilon)}\right]^2 = \frac{c_x}{k}\Theta(\varepsilon^2) = \Theta(\varepsilon^2/k).$$

Integrating over the region of width 2, we get that $A = O(\varepsilon^2/k)$. □

This concludes the proof of Theorem 42 and with it the proof of Theorem 8.

A.3 Proof of Lemma 22.

Recall Lemma 22:

Lemma 22. *Let $p_1, \dots, p_k$ each be an (τ, t)-piecewise degree-d distribution over $[-1, 1)$ and let $p = \sum_{j=1}^{k} \mu_j p_j$ be a k-mixture of components $p_1, \dots, p_k$. Then p is a (τ, kt)-piecewise degree-d distribution.*

Proof of Lemma 22: For $1 \le j \le k$, let $\mathcal{P}_j$ denote the intervals $I_{j,1}, \dots, I_{j,t}$ such that p_j is τ-close to a distribution g_j whose pdf is given by polynomials $g_{j_1}, \dots, g_{j,t}$ over intervals $I_{j,1}, \dots, I_{j,t}$ respectively. Let $\mathcal{P}$ be the common refinement of $\mathcal{P}_1, \dots, \mathcal{P}_k$. It is clear that $\mathcal{P}$ is a partition of $[-1, 1)$ into at most kt intervals.

For each I in $\mathcal{P}$ and for each $1 \le j \le k$, let $g_{j,I} \in \{g_{j,1}, \dots, g_{j,t}\}$ be the polynomial corresponding to I. We claim that $p = \sum_{j=1}^{k} \mu_j p_j$ is τ-close to the kt-piecewise degree-d distribution g which has the polynomial $\sum_{j=1}^{k} \mu_j g_{j,I}$ as its pdf over interval I, for each $I \in \mathcal{P}$. To see this, for each interval $I \in \mathcal{P}$ let us write $\tilde{p}_{j,I}$ to denote the function which equals p_j on I and equals 0 elsewhere, and likewise for $\tilde{g}_{j,I}$. With this notation we may write the condition that p_j is τ-close to g_j in total variation distance as

$$\left\| \sum_{I \in \mathcal{P}} \tilde{p}_{j,I} - \tilde{g}_{j,I} \right\|_1 \le 2\tau. \tag{27}$$

We then have

$$\|p - g\|_1 = \left\| \sum_{I \in \mathcal{P}} \left(\sum_{j=1}^{k} \mu_j \tilde{p}_{j,I} - \mu_j \tilde{g}_{j,I} \right) \right\|_1 \le \sum_{j=1}^{k} \mu_j \left\| \sum_{I \in \mathcal{P}} (\tilde{p}_{j,I} - \tilde{g}_{j,I}) \right\|_1 \le 2\tau,$$

and the proof is complete. □

A.4 Proof of Lemma 24.

Recall Lemma 24:

Lemma 24. *With probability at least* 99/100, `Find-Heavy`(γ) *returns a set S satisfying conditions (1) and (2) in the "Output" description.*

Proof of Lemma 24: Fix any $x \in [-1, 1)$ such that $\Pr_{x \sim p}[x] \ge 2\gamma$. A standard multiplicative Chernoff bound implies that x is placed in S except with failure probability at most $\frac{1}{200} \cdot \frac{1}{2\gamma}$. Since there are at most $\frac{1}{2\gamma}$ values $x \in [-1, 1)$ such that $\Pr_{x \sim p}[x] \ge 2\gamma$, we get that condition (1) holds except with failure probability at most $\frac{1}{200}$.

For the second bullet, first consider any x such that $\Pr_{x \sim p}[x] \in [\frac{\gamma}{2^c}, \frac{\gamma}{2}]$ (here $c > 0$ is a universal constant). A standard multiplicative Chernoff bound gives that each such x satisfies $\widehat{p}(x) \ge 2 \Pr_{x \sim p}[x]$ with probability at most $\frac{1}{400} \cdot \frac{2^c}{\gamma}$, and hence each such x satisfies $\widehat{p}(x) \ge \gamma$ with probability at most $\frac{1}{400} \cdot \frac{2^c}{\gamma}$. Since there are at most $2^c/\gamma$ such x's, we get that with probability at least $1 - \frac{1}{400}$ no such x belongs to S.

To finish the analysis we recall the following version of the multiplicative Chernoff bound:

Fact 50. *[[MR95], Theorem 4.1] Let $Y_1, \dots, Y_m$ be i.i.d. 0/1 random variables with* $\Pr[Y_i = 1] = q$ *and let $Q = mq = \mathbb{E}[\sum_{i=1}^{m} Y_i]$. Then for all $\tau > 0$ we have*

$$\Pr\left[\sum_{i=1}^{m} Y_i \ge (1 + \tau)Q \right] \le \left(\frac{e^\tau}{(1 + \tau)^{1+\tau}} \right)^Q \le \left(\frac{e}{(1 + \tau)} \right)^{(1+\tau)Q}.$$

Fix any integer $r \geq c$ and fix any x such that $\Pr_{x \sim p}[x] \in [\frac{\gamma}{2^{r+1}}, \frac{\gamma}{2^r}]$. Taking $1+\tau$ in Fact 50 to equal 2^r, we get that

$$\Pr[x \in S] \leq \left(\frac{e}{2^r}\right)^{\Theta(m\gamma)} = \left(\frac{e}{2^r}\right)^{\Theta(\log(1/\gamma))}.$$

Summing over all (at most $2^{r+1}/\gamma$ many) x such that $\Pr_{x \sim p}[x] \in [\frac{\gamma}{2^{r+1}}, \frac{\gamma}{2^r}]$, we get that the probability that any such x is placed in S is at most $\frac{2^{r+1}}{\gamma} \cdot \left(\frac{e}{2^r}\right)^{\Theta(\log(1/\gamma))} \leq \frac{1}{400} \cdot \frac{1}{2^r}$. Summing over all $r \geq c$, the total failure probability incurred by such x is at most $1/400$. This proves the lemma. □