Efficient Density Estimation via Piecewise Polynomial Approximation

[Extended Abstract]

Siu-On Chan[*]
Microsoft Research
Cambridge, MA
siuon@cs.berkeley.edu

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

Rocco A. Servedio[‡]
Columbia University
New York, NY
rocco@cs.columbia.edu

Xiaorui Sun[§]
Columbia University
New York, NY
xiaoruisun@cs.columbia.edu

ABSTRACT

We give a computationally 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, and suppose that p 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 $\tilde{O}(t(d+1)/\epsilon^2)$ samples from p, runs in time $\mathrm{poly}(t, d+1, 1/\epsilon)$, and with high probability outputs a piecewise polynomial hypothesis distribution h that is $(14\tau+\epsilon)$-close to p in total variation distance. Our algorithm combines tools from real approximation theory, uniform convergence, linear programming, and dynamic programming. Its sample complexity is simultaneously near-optimal in all three parameters t, d and ϵ; 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 $\Omega(\frac{t(d+1)}{\mathrm{poly}(\log(d+2))} \cdot \frac{1}{\epsilon^2})$ samples from the distribution, regardless of its running time.

We apply this general algorithm to obtain a wide range of results for many natural density estimation problems 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 gives improved results, with provably optimal sample complexities (up to logarithmic factors) in all parameters in most cases, for all these problems via a single unified algorithm.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Distribution functions*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Nonparametric statistics*; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Theory

Keywords

Computational Learning Theory; Unsupervised Learning; Learning Distributions

1. INTRODUCTION

Density estimation — the problem of constructing a highly accurate hypothesis distribution given i.i.d. samples drawn from an unknown probability distribution — is a well-studied topic in probability theory and statistics; book-length introductions to the field can be found in [DG85, Sil86, Sco92, DL01]. A number of generic techniques for density estimation are known 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 these problems, with an explicit focus on obtaining *computationally efficient* algorithms; see e.g., [KMR+94, FM99,

[*]Supported by NSF award DMS-1106999, DOD ONR grant N000141110140 and NSF award CCF-1118083. Some of this work was done while the author was at UC Berkeley.

[†]Supported in part by a SICSA PECE grant. Part of this work was done while the author was at UC Berkeley supported by a Simons Postdoctoral Fellowship.

[‡]Supported by NSF grants CCF-1115703 and CCF-1319788.

[§]Supported by NSF grant CCF-1149257.

STOC '14, May 31 - June 03 2014, New York, NY, USA

http://dx.doi.org/10.1145/2591796.2591800 .

FOS05, BS10, KMV10, MV10, DDS12a, DDS12b, DDO$^+$13] for some representative recent works.

In this paper we propose a new approach to density estimation. Our approach is based on establishing the existence of *piecewise polynomial density functions* that approximate the distribution to be learned. The key tool that enables this new approach is a computationally 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 effectiveness 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 most cases the sample complexity of our approach is provably optimal up to logarithmic factors. We remark that these questions correspond to fundamental statistical estimation tasks; hence, we believe it is of substantial interest from a theoretical standpoint to obtain computationally efficient and sample-*optimal* algorithms for these problems. From a practical perspective, it is critical to use our data efficiently. In particular, in many natural settings, even modest asymptotic differences in the sample size can play a big role.

1.1 Our main general result.

We work in a PAC-type model similar to that of [KMR$^+$94] and to well-studied statistical frameworks for density estimation. In our model, the learning algorithm has access to i.i.d. draws from an unknown probability distribution p supported on a finite interval $I \subset \mathbb{R}$.It draws i.i.d. samples from p, and 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 are concerned with learning algorithms that both (i) use few samples — ideally, close to the information-theoretic minimum possible number — and (ii) are computationally efficient.

We say that a distribution q over I is a *t-piecewise degree-d distribution* if there is a partition of I into t disjoint intervals $I_1,\ldots,I_t$ such that $q(x)=q_j(x)$ for all $x \in I_j$, where each of $q_1,\ldots,q_t$ is a degree-d polynomial.[1] Our main result is a semi-agnostic algorithm for learning piecewise polynomial distributions. Let $\mathcal{P}_{t,d}(I)$ denote the class of all t-piecewise degree-d distributions over interval I. Our main result, stated informally, is the following:

THEOREM 1. *[Informal statement] Let p be any unknown target distribution over I. Our algorithm, given parameters t, d, ϵ and $\tilde{O}(t(d+1)/\epsilon^2)$ samples from p, runs in time*[2] $\mathrm{poly}(t, d+1, 1/\epsilon)$ *and with high probability outputs an $O(t)$-piecewise degree-d hypothesis distribution h such that $d_{\mathrm{TV}}(p,h) \le 14\mathrm{opt}_{t,d}+\epsilon$, where $\mathrm{opt}_{t,d} := \inf_{r\in\mathcal{P}_{t,d}(I)} d_{\mathrm{TV}}(p,r)$ is the error of the best t-piecewise degree-d distribution for p.*

(See Theorem 5 for a precise statement.) We view this result as providing quite a strong guarantee, since (as we will see through our applications) t-piecewise degree-d distributions constitute a broad and flexible class that is capable of accurately approximating many distributions of interest.

One strength of our result is that it is extremely efficient — in fact, essentially optimal — in terms of the number of samples it uses. We prove the following lower bound (see Theorem 7 for a precise statement), which shows that the number of samples that our algorithm uses is simultaneously optimal in all three parameters up to logarithmic factors:

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

Note that this 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 are guaranteed to be evenly spaced across I.

1.2 Applications of Theorem 1.

Using Theorem 1 we obtain highly efficient estimators for a wide range of specific distribution learning problems over both continuous and discrete domains. Note that we do not aim to exhaustively cover all possible applications of Theorem 1, but rather to give some selected applications that are indicative of the generality and power of our methods.

In the continuous setting 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 1950's (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 various 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]. Our general method provides a single unified approach that yields a 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 logarithmic factors. Our approach also yields efficient and sample optimal estimators for various parametric classes, such as mixtures of univariate Gaussians. Next, turning to discrete distributions, using our methods, we improve prior results to obtain sample-optimal estimators for various nonparametric

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

[2] 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 only performs basic arithmetic operations on "reasonable" inputs.

classes, including mixtures of discrete t-modal distributions and mixtures of discrete monotone hazard rate (MHR) distributions.

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 most 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 them, are given in the full paper.

We note that all the learning results indicated with theorem numbers in Table 1, i.e., all the 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 $(14\tau + \epsilon)$-close to the target distribution.

1.3 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) $B_1, \ldots, B_k$, 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 technique may 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, and note that the recent paper [PA13] also proposes using splines for density estimation. (However, this is not the main focus of that paper, and indeed [PA13] does not provide or analyze algorithms for density estimation.)

Our 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 subdistribution on a given interval). However, 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 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 define a linear program and carry out a careful error analysis using probabilistic inequalities (the VC inequality and tail bounds) and ingredients from basic approximation theory to show that $\tilde{O}(d/\epsilon^2)$ samples suffice for our LP to achieve an $O(\mathrm{opt}_{1,d} + \epsilon)$-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 3.5). 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 approach, in order to obtain efficient estimators for specific classes of distributions, it is sufficient to establish the existence of piecewise polynomial approximations to the distributions that are to be learned. In some cases, results establishing the existence of piecewise constant or piecewise linear approximations were already known; for example, [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 17) which, combined with our general algorithm, leads to an optimal learning algorithm for (mixtures of) such densities. In particular, we show that any log-concave density can be ϵ-approximated by a piecewise linear distribution with $\tilde{O}(\epsilon^{-1/2})$ pieces. The proof of Lemma 17 is quite elaborate and significantly different from the aforementioned known arguments establishing the existence of piecewise linear approximations for concave functions. 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.

1.4 Related work.

As noted above, our work may be viewed as a broad generalization of the "histogram method." To the best of our knowledge, the state of the art for histogram-type algorithms was given in [CDSS13]. (That paper dealt with distributions over the discrete domain $[N] = \{1, \ldots, N\}$, but for simplicity we translate the [CDSS13] results to the continuous domain. This translation is straightforward.) Recall that given distributions $p_1, \ldots, p_k$ and non-negative values $\mu_1, \ldots, \mu_k$ that sum to 1, the distribution $p = \sum_{i=1}^{k} \mu_i p_i$ is a *k-mixture* of *components* $p_1, \ldots, p_k$ with *mixing weights* $\mu_1, \ldots, \mu_k$ if 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 said to be *(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 $x \in I_j$.

[CDSS13] gave an algorithm that learns any k-mixture of (ϵ, t)-piecewise constant distributions over an interval I to accuracy $O(\epsilon)$, using $O(kt/\epsilon^3)$ samples and running in $\tilde{O}(kt/\epsilon^3)$ time. Our main result gives both a quantitative and qualitative improvement over this [CDSS13] result. Quantitatively, we improve the ϵ-dependence from $1/\epsilon^3$ to a near-

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 5
t-piecewise degree-d polynomial	$\tilde{O}(td/\epsilon^2)$ (†)	Theorem 5, Theorem 7
log-concave	$O(1/\epsilon^{5/2})$ (†)	folklore [DL01]
mixture of k log-concave distributions	$\tilde{O}(k/\epsilon^{5/2})$ (†)	Theorem 16
mixture of t bounded 1-monotone distributions	$\tilde{O}(t/\epsilon^3)$ (†)	full version
mixture of t bounded 2-monotone distributions	$\tilde{O}(t/\epsilon^{5/2})$ (†)	full version
mixture of t bounded k-monotone distributions	$\tilde{O}(tk/\epsilon^{2+1/k})$	full version
mixture of k Gaussians	$\tilde{O}(k/\epsilon^2)$ (†)	full version
Discrete distributions over $\{1, 2, \ldots, N\}$		
t-modal	$\tilde{O}(t\log(N)/\epsilon^3) + \tilde{O}(t^3/\epsilon^3)$	[DDS12a]
mixture of k t-modal distributions	$O(kt\log(N)/\epsilon^4)$	[CDSS13]
mixture of k t-modal distributions	$\tilde{O}(kt\log(N)/\epsilon^3)$ (†)	full version
mixture of k monotone hazard rate distributions	$\tilde{O}(k\log(N)/\epsilon^4)$	[CDSS13]
mixture of k monotone hazard rate distributions	$\tilde{O}(k\log(N)/\epsilon^3)$ (†)	full version

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 optimal up to at most logarithmic factors (i.e., "$\tilde{O}(m)$ (†)" means that there is a known lower bound of $\Omega(m)$).

optimal $\tilde{O}(1/\epsilon^2)$. Qualitatively, we extend the [CDSS13] result from piecewise *constant* distributions to piecewise *polynomial* distributions. The applications of our main result crucially use both the quantitative and qualitative aspects in which it strengthens the [CDSS13] 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. Because of space constraints many proofs are omitted from this extended abstract, and can be found in the full version.

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.

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 nonnegative function p over an interval (which need not necessarily integrate to one 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 $z \in [-1, 1)$, $\Pr_{x\sim\widehat{p}_m}[x = z] = |\{j \in [m] \mid s_j = z\}|/m$.

Given a value $\kappa > 0$, a distribution or subdistribution 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 Lebesgue measurable. Note that 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.

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) $\mathrm{opt}_{t,d} + \epsilon/100$ is sufficient for all our arguments. We will sometimes say that a distribution p is *(τ, t)-piecewise degree-d* to indicate that p is τ-close in variation distance to a t-piecewise degree-d distribution.

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 $\mathcal{I}'$ 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.

The VC inequality. Let $p : [-1, 1) \to \mathbb{R}$ be a Lebesgue measurable function. 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 $X \subseteq [-1, 1)$ 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 $s \in \mathbb{Z}_+$, then we say that the VC dimension of $\mathcal{A}$ is ∞. The well-known *Vapnik-Chervonenkis (VC) inequality* states the following:

THEOREM 3 (VC INEQUALITY, [DL01, P.31]). *Let $p : I \to \mathbb{R}_+$ be a probability density function over $I \subseteq \mathbb{R}$ and $\widehat{p}_m$ be the empirical distribution obtained after drawing m samples from p. Let $\mathcal{A} \subseteq 2^I$ be a family of subsets with VC dimension d. Then $\mathbb{E}[\|p - \widehat{p}_m\|_{\mathcal{A}}] \leq O(\sqrt{d/m})$.*

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.

LEMMA 4. *Given $\kappa \in (0,1)$ and access to samples from a $\kappa/64$-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 $p(I_j) \in [\kappa/2, 3\kappa]$ for all $1 \leq j \leq \ell$.*

3. MAIN RESULT: LEARNING PIECEWISE POLYNOMIAL DISTRIBUTIONS WITH NEAR-OPTIMAL SAMPLE COMPLEXITY

In this section we present and analyze our main algorithm for learning (τ, t)-piecewise degree-d distributions. The structure of this section is as follows: We start by giving a simple information-theoretic argument (Proposition 6, Section 3.1) showing that there is a (computationally inefficient) algorithm to learn any distribution p to accuracy $3\mathrm{opt}_{t,d}+\epsilon$ using $O(t(d+1)/\epsilon^2)$ samples, where $\mathrm{opt}_{t,d}$ is the smallest variation distance between p and any t-piecewise degree-d distribution. Next, we prove an information-theoretic lower bound (Theorem 7, Section 3.2) showing that the aforementioned sample complexity is essentially optimal: Any algorithm, regardless of its running time, for learning a t-piecewise degree-d distribution to accuracy ϵ must use $t \cdot \tilde{\Omega}(d+1)/\epsilon^2$ samples.

We then proceed to 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 $\tilde{O}((d+1)/\epsilon^2)$ samples from any well-behaved distribution p, and with high probability outputs a degree-d pdf h such that $d_{\mathrm{TV}}(p,h) \leq 7\mathrm{opt}_{1,d} + \epsilon$. This algorithm is based on linear programming and its analysis uses ingredients from real approximation theory and uniform convergence results in probability theory. In Section 3.4 we generalize this algorithm so that it can be used to learn a *sub-distribution* over an arbitrary interval. In Section 3.5 we use the result of Section 3.4 to obtain an efficient "semi-agnostic" algorithm for t-piecewise degree-d pdfs. The approach of this section uses a subtle dynamic program on top of the linear programming approach mentioned above. The extended algorithm draws $\tilde{O}(t(d+1)/\epsilon^2)$ samples from any well-behaved distribution p, and with high probability outputs a $(2t-1)$-piecewise degree-d pdf h such that $d_{\mathrm{TV}}(p,h) \leq 13\mathrm{opt}_{t,d} + \epsilon$. In the full version we point out that this result can be straightforwardly extended to k-mixtures of well-behaved distributions, and we show how we may get rid of the "well-behaved" requirement (at the cost of another additive $\mathrm{opt}_{t,d}$ factor), and thereby prove the following generalization of Theorem 1:

THEOREM 5. *Let p be any k-mixture of (τ, t)-piecewise degree-d distributions over $[-1,1)$. There is an algorithm that runs in $\mathrm{poly}(k, t, d+1, 1/\epsilon)$ time, uses $\tilde{O}((d+1)kt/\epsilon^2)$ samples from p, and with probability at least $9/10$ outputs a $(2kt-1)$-piecewise degree-d hypothesis h such that $d_{\mathrm{TV}}(p,h) \leq 14\mathrm{opt}_{t,d} + O(\epsilon)$.*

3.1 An information-theoretic sample complexity upper bound.

PROPOSITION 6. *There is a (computationally inefficient) algorithm that draws $O(t(d+1)/\epsilon^2)$ samples from any distribution p over $[-1,1)$, and with probability $9/10$ outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(p,h) \leq 3\mathrm{opt}_{t,d} + \epsilon$.*

PROOF. The main idea is to use Theorem 3, 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 $m = O(t(d+1)/\epsilon^2)$ 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 $d_{\mathrm{TV}}(p,q) \leq \mathrm{opt}_{t,d}$ we have that $\|p-q\|_{\mathcal{A}} \leq \mathrm{opt}_{t,d}$. Since the VC dimension of $\mathcal{A}$ is $4t(d+1)$, Theorem 3 implies that $\mathbb{E}[\|p-\widehat{p}_m\|_{\mathcal{A}}] \leq \epsilon/40$, and hence by Markov's inequality, with probability at least $19/20$ we have that $\|p - \widehat{p}_m\|_{\mathcal{A}} \leq \epsilon/2$. By the triangle inequality for $\|\cdot\|_{\mathcal{A}}$-distance, this means that $\|q-\widehat{p}_m\|_{\mathcal{A}} \leq \mathrm{opt}_{t,d}+\epsilon/2$.

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 $\|q - \widehat{p}_m\|_{\mathcal{A}} \leq \mathrm{opt}_{t,d} + \epsilon/2$, the distribution h satisfies $\|h-\widehat{p}_m\|_{\mathcal{A}} \leq \mathrm{opt}_{t,d}+\epsilon/2$. Hence, the triangle inequality gives $\|h - q\|_{\mathcal{A}} \leq 2\mathrm{opt}_{t,d} + \epsilon$.

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 $d_{\mathrm{TV}}(h,q) = \|h-q\|_{\mathcal{A}} \leq 2\mathrm{opt}_{t,d} + \epsilon$. The triangle inequality for variation distance gives that $d_{\mathrm{TV}}(h,p) \leq 3\mathrm{opt}_{t,d} + \epsilon$, as desired. □

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 natural 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 nearly the same sample complexity and runs in time $\mathrm{poly}(t, d+1, 1/\epsilon)$. 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 10 and 13), and that the corresponding optimization problem can be expressed as an appropriate 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 record 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), $\tilde{\Omega}(t(d+1)/\epsilon^2)$ samples are required for any algorithm to learn to accuracy ϵ:

THEOREM 7. *Let p be an unknown t-piecewise degree-d distribution over $[-1,1)$ where $t \geq 1$, $d \geq 0$ satisfy $t+d > 1$.* [3] *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 $\mathbb{E}[d_{\mathrm{TV}}(p,h)] \leq \epsilon$, where the expectation is over the random samples drawn from p and any internal randomness of L. Then L must use at least $\Omega(\frac{t(d+1)}{(\log(d+2))^2} \cdot \frac{1}{\epsilon^2})$ samples.*

To prove Theorem 7 we use a well-known information-theoretic tool, known as Assouad's lemma [Ass83]. To apply this lemma in our context we make essential use of recent nontrivial constructions of certain low-degree polynomial approximations to the sign function [DGJ+10]. At a high-level, we leverage the existence of such polynomials to construct a set of polynomial probability density functions that meet the conditions of Assouad's lemma.

3.3 Semi-agnostically learning a degree-d polynomial density with near-optimal sample complexity.

Throughout this subsection p will be a probability distribution with support contained in $[-1,1)$. Hence for this subsection we define $d_{\mathrm{TV}}(p,q)$, where p is a Lebesgue measurable density and q is a polynomial, to be $\frac{1}{2}\int_{-1}^{1} |p(x) - q(x)|dx$.

In this subsection we prove the following:

THEOREM 8. *Let p be an $\frac{\epsilon}{64(d+1)}$-well-behaved pdf over $[-1,1)$. There is an algorithm* `Learn-WB-Single-Poly`(d,ϵ) *which runs in* $\mathrm{poly}(d+1, 1/\epsilon)$ *time, uses $\tilde{O}((d+1)/\epsilon^2)$ 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 $d_{\mathrm{TV}}(p,q) \leq 7 \cdot \mathrm{opt}_{1,d} + \epsilon$.*

We start by providing a sketch of our computationally efficient algorithm. For the sake of this intuitive explanation, let us assume that the unknown distribution p is *exactly* a degree-d pdf – as opposed to close to one. Given $\epsilon > 0$ and sample access to $p : [-1,1) \to \mathbb{R}_+$ we proceed as follows: We start by partitioning the domain $[-1,1)$ into intervals $I_1, I_2, \ldots, I_z$ each of probability mass $\eta = \Theta(\epsilon/(d+1))$, where $z = \Theta(1/\eta)$. Note that this step can be achieved with $\tilde{O}((d+1)/\epsilon)$ samples from p (Lemma 4).

To compute a hypothesis degree-d polynomial f, we draw $m = \tilde{O}((d+1)/\epsilon^2)$ samples from p and consider the corresponding empirical distribution $\widehat{p}_m$. We then formulate and solve an appropriate linear program, of size polynomial in d and $1/\epsilon$, based on these samples. Ideally, one would like to impose the constraint that f agrees with p everywhere on $[-1,1)$. However, this is not possible, since (i) we only have approximate information about p, via the empirical distribution $\widehat{p}_m$, and (ii) we are allowed to use at most a finite number of constraints. As our main result of this section, we show that a polynomial number of linear constraints suffices to approximate the unknown pdf p up to an additive ϵ in total variation distance, as explained below.

Let F be the hypothesis "cumulative distribution function" (cdf) corresponding to f, i.e., $F : [-1,1) \to \mathbb{R}$ with $F(x) = \int_{-1}^{x} f(u)du$. (We used quotation marks above because F will end up satisfying the properties of a bona-fide cdf only approximately.) First, we require that $F(-1) = 0$ and $F(1) = 1$. Moreover, we impose the following crucial set of constraints: the probability mass that the hypothesis f assigns to unions of consecutive intervals from our decomposition, i.e., $f(\cup_{\ell=j}^{k} I_\ell)$ is "approximately" the same as $\widehat{p}_m(\cup_{\ell=j}^{k} I_\ell)$, i.e., the mass that the empirical distribution assigns to the same set. The "closeness" of this approximation is selected carefully so as to guarantee that (i) these constraints are satisfied (w.h.p.), and (ii) the quantitative bounds are sufficient for our error analysis. In addition, we would ideally like to guarantee that $F([-1,1)) \subseteq [0,1]$ and $f([-1,1)) \subseteq [0,\infty]$. Similarly, this is not possible to achieve with a finite number of inequalities. Exploiting the fact that f is a degree-d polynomial, we show that it is essentially sufficient for our purposes to impose these constraints on a sufficiently dense discrete "grid" of the domain. For technical reasons related to these aforementioned "grid" constraints, we express our hypothesis distribution in the basis of the Chebychev polynomials.

We are now ready to formally describe the algorithm and prove its correctness. Some preliminary definitions will help:

DEFINITION 9 (UNIFORM PARTITION). *Let p be a sub-distribution on an interval $I \subseteq [-1,1)$. A partition $\mathcal{I} = \{I_1, \ldots, I_z\}$ of I is (p,η)-*uniform *if $p(I_\ell) \leq \eta$ for all $1 \leq \ell \leq z$.*

DEFINITION 10. *Let $\mathcal{I} = \{I_1 = [x_0, x_1), \ldots, I_z = [x_{z-1}, x_z)\}$ 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 $(\mathcal{I}, \eta, \epsilon)$-*inequalities over I *if*

$$|p(\cup_{\ell=j}^{k-1} I_\ell) - q(\cup_{\ell=j}^{k-1} I_\ell)| \leq \sqrt{\epsilon(k-j)} \cdot \eta \qquad (1)$$

for all $1 \leq j < k \leq z+1$.

Our learning algorithm `Learn-WB-Single-Poly` is described in detailed pseudocode below. The key step of `Learn-WB-Single-Poly` is Step 3 where the `Find-Single-Polynomial` procedure is called. In this procedure $T_i(x)$ denotes the degree-i Chebychev polynomial of the first kind. The function F computed by the procedure `Find-Single-Polynomial` may be thought of as the CDF of a degree-d "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 4 of `Learn-WB-Single-Poly` processes the degree-d polynomial f slightly to obtain a polynomial q which is an actual distribution over $[-1,1)$.

Algorithm `Learn-WB-Single-Poly`
Input: parameters d, ϵ
Output: with probability at least $9/10$, a degree-d distribution q over $[-1,1)$ such that $d_{\mathrm{TV}}(p,q) \leq 7 \cdot \mathrm{opt}_{1,d} + O(\epsilon)$

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

1. Run Algorithm `Approximately-Equal-Partition` on input parameter $\epsilon/(d+1)$ to partition $[-1,1)$ into $z = \Theta((d+1)/\epsilon)$ intervals $I_1 = [x_0, x_1)$, $\ldots$, $I_z = [x_{z-1}, x_z)$, where $x_0 = -1$ and $x_z = 1$, such that for each $\ell \in [z]$ we have $p(I_\ell) = \Theta(\epsilon/(d+1))$. Let $\mathcal{I}$ denote this partition of $[-1,1)$ into $I_1, \ldots, I_z$, i.e., $\mathcal{I} = \{I_\ell\}_{\ell=1}^z$.
2. Draw $m = \tilde{O}((d+1)/\epsilon^2)$ samples and let $\widehat{p}_m$ be the empirical distribution defined by these samples. Set $\eta := \Theta(\epsilon/(d+1))$.
3. Call `Find-Single-Polynomial`$(d, \epsilon, \eta, \mathcal{I}, \widehat{p}_m)$ and let f^* be the degree-d polynomial that it returns.
4. Define $q(x) = \epsilon'/2 + (1-\epsilon')f^*(x)$, where ϵ' is defined to be $\epsilon' = \eta/10$ as in Step 1(f) of `Find-Single-Polynomial`, and output q as the hypothesis pdf.

Subroutine `Find-Single-Polynomial`

Input: degree parameter d; error parameter ϵ; parameter η; (p,η)-uniform partition $\mathcal{I} = \{I_1, \ldots, I_z\}$ of interval $[-1,1)$ into z intervals $I_\ell = [x_{\ell-1}, x_\ell)$; a distribution $\widehat{p}_m$ on $[-1,1)$ such that $\widehat{p}_m$ and p satisfy the $(\mathcal{I}, \eta, \epsilon)$-inequalities over $[-1,1)$

Output: a number τ^* such that $0 \le \tau^* \le \mathrm{opt}_{1,d}$ and a degree-d polynomial f^* on $[-1,1)$ such that $f^*([-1,1)) = 1$ and

$$d_{\mathrm{TV}}(p, f^*) \le 7 \cdot \mathrm{opt}_{1,d} + \sqrt{\epsilon z (d+1)} \cdot \eta + \text{error},$$

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

1. Let τ^* be the optimal value of 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, the c_i's are variables of the LP, and $f(x) = F'(x) = \sum_{i=0}^{d+1} c_i T_i'(x)$.)

   (a) $F(-1) = 0$ and $F(1) = 1$;

   (b) For each $1 \le j < k \le z+1$,

   $$\left|\widehat{p}_m(\cup_{\ell=j}^{k-1} I_\ell) + \sum_{j \le \ell < k} w_\ell - f(\cup_{\ell=j}^{k-1} I_\ell)\right| \le \sqrt{\epsilon \cdot (k-j)} \cdot \eta;$$

   (c)

   $$-y_\ell \le w_\ell \le y_\ell \qquad \text{for all } 1 \le \ell \le z, \quad (2)$$
   $$\sum_{1 \le \ell \le z} y_\ell \le 2\tau; \quad (3)$$

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

   (e) The constraints $0 \le F(x) \le 1$ for all $x \in J$, where J is a set of $O((d+1)^3)$ equally spaced points across $[-1,1)$;

   (f) The constraints $\sum_{i=0}^{d+1} c_i T_i'(x) \ge 0$ for all $x \in K$, where K is a set of $O(d^2(d+1)^2/\epsilon')$ equally spaced points across $[-1,1)$ and $\epsilon' \stackrel{\text{def}}{=} \eta/10$.

2. Define the degree-d polynomial $f^*(x) = \sum_{i=0}^{d+1} c_i^* T_i'(x)$ where the c_i^*'s correspond to an optimal solution to the LP, and output τ^* and f^*.

In the rest of this subsection we sketch the proof of Theorem 8. 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 $\mathrm{poly}(d, 1/\epsilon)$-size LP in `Find-Single-Polynomial`), so it suffices to prove correctness.

Before launching into the proof we give some additional 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. Constraint 1(b)(1) ensures that for each interval $\cup_{\ell=j}^{k-1} I_\ell = [x_{j-1}, x_{k-1})$, the value $F(x_{k-1}) - F(x_{j-1}) = f([x_{j-1}, x_{k-1}))$ is close to the mass $\widehat{p}_m([x_{j-1}, x_{k-1}))$ that the empirical distribution puts on the same interval. Recall that by assumption p is $\mathrm{opt}_{1,d}$-close in total variation distance to some degree-d polynomial r. Intuitively the variable w_ℓ represents $\int_{I_\ell}(r-p)$. The variable 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, and the 1(f) constraints reflect the fact that the pdf $f = F'$ should be everywhere nonnegative. Since we are imposing these constraints on discrete sets of points, we will not be able to guarantee that the desired inequalities hold for all $x \in [-1,1)$. In particular, it will not necessarily be the case that a feasible solution f to the LP satisfies $f(x) \ge 0$ everywhere. However, we will be able to show that f is "essentially" nonnegative everywhere, and this turns out to be good enough for our purposes.

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 $I_1, \ldots, I_z$ are (p,η)-uniform; and

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

LEMMA 11. *Suppose p is an $\frac{\epsilon}{64(d+1)}$-well-behaved pdf over $[-1,1)$. Then with probability at least $19/20$ over the random draws performed in Steps 1 and 2 of* `Learn-WB-Single-Poly`*, conditions (I) and (II) above hold.*

As our first main lemma for this section, we show that given that conditions (I) and (II) are satisfied, `Find-Single-Polynomial`'s LP is feasible and has a high-quality optimal solution.

LEMMA 12. *Suppose that conditions (I) and (II) above hold. Then the LP defined in Step 1 of the routine* `Find-Single-Polynomial` *is feasible, and its optimal value τ^* is at most $\mathrm{opt}_{1,d}$.*

Having established that with high probability the LP is indeed feasible, henceforth we let τ^* denote the optimal value of the LP and F^*, f^*, w^*_ℓ, c^*_i, y^*_ℓ denote the values of the variables in an optimal solution. From this point, using basic approximation theory it can be shown that for all $x \in [-1,1)$ we have

$$f^*(x) \geq 0 - \frac{\|(f^*)'\|_\infty}{|K|} \geq -\epsilon'/2 \tag{4}$$

(recall that $F^*(x) = \sum_{i=0}^{d+1} c^*_i T_i(x)$ and f^*, the derivative of F^*, is $f^*(x) = \sum_{i=0}^{d+1} c^*_i T'_i(x)$).

Error Analysis. To prove Theorem 8 it remains to show that

$$d_{\mathrm{TV}}(q,p) \leq 7 \cdot \mathrm{opt}_{1,d} + O(\epsilon). \tag{5}$$

We sketch the argument that we use to prove (5). A key step in achieving this is to bound from above the $\|\cdot\|_{\mathcal{A}_{d+1}}$ distance between f^* and $\widehat{p}_m + w^*$, where $w^* : [-1,1) \to \mathbb{R}$ is a piecewise constant function defined based on the w^*_ℓ values. Similar to Section 3.1, the VC theorem gives us that $\|p - \widehat{p}_m\|_{\mathcal{A}_{d+1}} \leq \epsilon$ with probability at least 39/40. Conditioning on this event, our main claim is that

$$\|(\widehat{p}_m + w^*) - f^*\|_{\mathcal{A}_{d+1}} \leq 4 \cdot \mathrm{opt}_{1,d} + O(\epsilon). \tag{6}$$

Recall that r denotes a degree-d polynomial pdf such that $\mathrm{opt}_{1,d} = d_{\mathrm{TV}}(p,r)$. Given (6), it will not be difficult to show that $\|r - f^*\|_{\mathcal{A}_{d+1}} \leq 6 \cdot \mathrm{opt}_{1,d} + O(\epsilon)$. Since r and f^* are both degree-d polynomials we have $d_{\mathrm{TV}}(r,f^*) = \|r - f^*\|_{\mathcal{A}_{d+1}} \leq 6 \cdot \mathrm{opt}_{1,d} + O(\epsilon)$, so recalling that $d_{\mathrm{TV}}(p,r) = \mathrm{opt}_{1,d}$, the triangle inequality gives $d_{\mathrm{TV}}(p,f^*) \leq 7 \cdot \mathrm{opt}_{1,d} + O(\epsilon)$. From this point a straightforward argument using Equation (4) gives that $d_{\mathrm{TV}}(p,q) \leq d_{\mathrm{TV}}(p,f^*) + O(\epsilon)$, which concludes the proof.

To prove (6), we make essential use of the following lemma that translates $(\mathcal{I},\eta,\epsilon)$-inequalities into a bound on $\mathcal{A}_k$ distance. (While we will only use the lemma with $k = d+1$ in this subsection, the fact that the lemma holds for all k will prove crucial in the analysis of our final dynamic programming based algorithm, which uses a generalization of this lemma that we establish in the next subsection.)

LEMMA 13. *Let $\mathcal{I} = \{I_1 = [x_0,x_1), \ldots, I_z = [x_{z-1},x_z)\}$ be a (p,η)-uniform partition of $[-1,1)$. Let s be a distribution on $[-1,1)$ such that s and p satisfy $(\mathcal{I},\eta,\epsilon)$-inequalities on $[-1,1)$. Let $h : [-1,1) \to \mathbb{R}$ be such that $h(x) \geq -\epsilon'$ for all $x \in [-1,1)$ and $g : [-1,1) \to \mathbb{R}$. If $h+g$ and s satisfy the $(\mathcal{I},\eta,\epsilon)$-inequalities, then for all $k \geq 0$ we have that*

$$\|s - (h+g)\|_{\mathcal{A}_k} \leq 2\|g\|_1 + \sqrt{\epsilon z k} \cdot \eta + O(k\eta).$$

3.4 Generalized `Find-Single-Polynomial` Routine.

In this subsection we generalize the routine from the previous section to work for an arbitrary *subdistribution* with unknown total mass over a given input interval $[a,b)$. Note that the results we give in this subsection provide guarantees of a somewhat different form (in terms of $\mathcal{A}_k$-distance) from Theorem 8; this form is convenient for the dynamic programming algorithm that we give in Section 3.5.

We define $\widetilde{\mathrm{opt}}_{1,d;[a,b)}$ to be the infimum of $d_{\mathrm{TV}}(p,g)$ (where both p,g are viewed as having domain $[a,b)$) between p and any degree-d subdistribution g over $[a,b)$.

THEOREM 14. *Let p be an $\frac{\epsilon}{64(d+1)}$-well-behaved subdistribution over $[a,b)$, where $[a,b)$ is a subinterval of $[-1,1)$. Then there is an algorithm* `Gen-Find-Single-Polynomial-Arbitrary-Interval` *with the following properties. It takes as input a degree parameter d; error parameter ϵ; parameter η; a (p,η)-uniform partition $\mathcal{I}$ of $[a,b)$ into z intervals $\{I_\ell\}_{1 \leq \ell \leq z}$; and a non-negative function $\widehat{p}_m$ on $[a,b)$ such that $\widehat{p}_m$ and p satisfy the $(\mathcal{I},\eta,\epsilon)$-inequalities. It outputs a number τ^* with $0 \leq \tau^* \leq \widetilde{\mathrm{opt}}_{1,d;[a,b)}$ and a degree-d polynomial f^* on $[a,b)$ satisfying $f^*([a,b)) = \widehat{p}_m([a,b))$ and $f^*(x) \geq -\eta/(10(b-a))$ for $x \in [a,b)$ such that for all $k \geq 1$ we have*

$$\|(\widehat{p}_m + w^*) - f^*\|_{\mathcal{A}_k} \leq 4\tau^* + \sqrt{\epsilon z k} \cdot \eta + O(k\eta),$$

where w^* is a piecewise constant function on $[a,b)$ which is constant on each I_ℓ and satisfies $\|w^*\|_1 \leq 2\tau^*$.

3.5 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 15. *Let p be an $\frac{\epsilon}{64t(d+1)}$-well-behaved pdf over $[-1,1)$. The algorithm* `Learn-WB-Piecewise-Poly`(t,d,ϵ) *runs in poly$(t,d+1,1/\epsilon)$ time, uses $\widetilde{O}(t(d+1)/\epsilon^2)$ samples from p, and with probability at least 9/10 outputs a $(2t-1)$-piecewise degree-d distribution q such that $d_{\mathrm{TV}}(p,q) \leq 13\mathrm{opt}_{t,d} + O(\epsilon)$.*

At a high level, `Learn-WB-Piecewise-Poly`(t,d,ϵ) breaks down $[-1,1)$ into $t(d+1)/\epsilon$ subintervals (denoted as the partition $\mathcal{I} = \{I_0,\ldots,I_z\}$ in subsequent discussion; this partition is constructed in Step 1) and calls the subroutine `Gen-Find-Single-Polynomial-Arbitrary-Interval`$(d,\epsilon,\eta,\{I_a,\ldots,I_b\},\widehat{p}_m)$ on blocks of consecutive intervals from $\mathcal{I}$. As shown in the previous subsection, the subroutine `Gen-Find-Single-Polynomial-Arbitrary-Interval` returns a degree-d polynomial h that satisfies the guarantee of Theorem 14. 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.

Algorithm `Learn-WB-Piecewise-Poly`:

Input: parameters t,d,ϵ
Output: with probability at least 9/10, a $(2t-1)$-piecewise degree-d distribution q such that $d_{\mathrm{TV}}(p,q) \leq 3 \cdot \mathrm{opt}_{t,d} + O(\epsilon)$

1. Run Algorithm `Approximately-Equal-Partition` on input parameter $\epsilon/(t(d+1))$ to partition $[-1,1)$ into $z = \Theta(t(d+1)/\epsilon)$ intervals $I_0 = [x_0,x_1), \ldots, I_z = [x_{z-1},x_z)$, where $x_0 = -1$ and $x_z = 1$, such that for each $j \in \{1,\ldots,t\}$ we have $p([x_{j-1},x_j)) = \Theta(\epsilon/(t(d+1)))$. Let $\mathcal{I}$ denote this partition of $[-1,1)$ into $I_0,\ldots,I_z$.
2. Let $s = z/(d+1) = \Theta(t/\epsilon)$. Set $x'_j = x_{(d+1)j}$ and define interval $I'_j = [x'_j, x'_{j+1})$ for $0 \leq j < s$. Let $\mathcal{I}'$ denote this partition of $[-1,1)$ into $I'_0,\ldots,I'_{s-1}$.

3. Draw $m = \tilde{O}(t(d+1)/\epsilon^2)$ 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 `Gen-Find-Single-Polynomial-Arbitrary-Interval` $(d,\ \epsilon,\ \eta = \Theta(\epsilon/(t(d+1))),\ \{I_a,\ldots,I_b\},\ \widehat{p}_m)$ where $I_a,\ldots,I_b$ are consecutive intervals from $\mathcal{I}$ whose union is $I'_\ell \cup \cdots \cup I'_{j-1}$.

   (b) Let τ be the solution to the LP found by `Gen-Find-Single-Polynomial-Arbitrary-Interval` and h be the degree-d hypothesis subdistribution 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 $H(\cdot,\cdot)$.

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 GEN-FIND-SINGLE-POLYNOMIAL-ARBITRARY-INTERVAL) when the interval $[x'_0, x'_j)$ is partitioned into i pieces. The dynamic program above 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. The analysis of the dynamic program, given in the full version, crucially uses the fact that Theorem 14 provides a guarantee for all $k \geq 1$.

4. APPLICATIONS

We use Theorem 5 to obtain a wide range of concrete learning results for natural and well-studied classes of distributions over both continuous and discrete domains. Because of space constraints in this version we only present an application of Theorem 5 to learn mixtures of log-concave distributions over $[-1,1)$.

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 16. *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 $\mathrm{poly}(k/\epsilon)$ time, draws $\tilde{O}(k/\epsilon^{5/2})$ samples from f, and with probability at least $9/10$ outputs a hypothesis distribution h such that $d_{\mathrm{TV}}(f,h) \leq \epsilon$.*

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 $\Omega(\epsilon^{-5/2})$ samples. This lower bound can be easily generalized to show that learning a k-mixture of log-concave distributions over $[0,1]$ requires $\Omega(k/\epsilon^{5/2})$ 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 $t = \tilde{O}(1/\sqrt{\epsilon})$. 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 $t = \tilde{O}(1/\epsilon)$. Unfortunately, the latter bound is essentially tight, hence cannot lead to an algorithm with sample complexity better than $\Omega(\epsilon^{-3})$.

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 $t = O(1/\sqrt{\epsilon})$. 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 17. *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 $t = \tilde{O}(1/\sqrt{\epsilon})$.*

We note that our proof of Lemma 17 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. We note that the $\tilde{O}(1/\sqrt{\epsilon})$ 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] $\Omega(1/\epsilon^{5/2})$ 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 16 follows as a direct corollary of Lemma 17 and Theorem 5.

5. REFERENCES

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

[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. *Ann. Math. Statist.*, 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. *Ann. Statist.*, 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. *Ann. Statist.*, 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.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 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.

[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 *Proc. 12th COLT*, pages 183–192, 1999.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th IEEE 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. Berkeley Conf. in Honor of J. Neyman and J. 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: Math.*, 52:1525–1538, 2009.

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

[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.

[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 *Proc. 26th STOC*, 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.

[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. *J. Comput. & Graph. Statist.*, 0, 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.

[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.