Efficient Robust Proper Learning of Log-concave Distributions

Ilias Diakonikolas[*]
University of Southern California
diakonik@usc.edu

Daniel M. Kane[†]
University of California, San Diego
dakane@cs.ucsd.edu

Alistair Stewart[‡]
University of Southern California
alistais@usc.edu

June 10, 2016

Abstract

We study the *robust proper learning* of univariate log-concave distributions (over continuous and discrete domains). Given a set of samples drawn from an unknown target distribution, we want to compute a log-concave hypothesis distribution that is as close as possible to the target, in total variation distance. In this work, we give the first computationally efficient algorithm for this learning problem. Our algorithm achieves the information-theoretically optimal sample size (up to a constant factor), runs in polynomial time, and is robust to model misspecification with nearly-optimal error guarantees.

Specifically, we give an algorithm that, on input $n = O(1/\epsilon^{5/2})$ samples from an unknown distribution f, runs in time $\widetilde{O}(n^{8/5})$, and outputs a log-concave hypothesis h that (with high probability) satisfies $d_{\mathrm{TV}}(h, f) = O(\mathrm{OPT}) + \epsilon$, where OPT is the minimum total variation distance between f and the class of log-concave distributions. Our approach to the robust proper learning problem is quite flexible and may be applicable to many other univariate distribution families.

[*] Part of this work was performed while the author was at the University of Edinburgh. Supported in part by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration grant.

[†] Supported in part by NSF Award CCF-1553288 (CAREER). Some of this work was performed while visiting the University of Edinburgh.

[‡] Part of this work was performed while the author was at the University of Edinburgh. Supported by EPSRC grant EP/L021749/1.

1 Introduction

1.1 Background and Motivation

Suppose that we are given a number of samples drawn from an unknown target distribution that belongs to (or is well-approximated by) a given family of distributions $\mathcal{D}$. Our goal is to approximately estimate (learn) the target distribution in a precise way. Estimating a distribution from samples is a fundamental unsupervised learning problem that has been studied in statistics since the late nineteenth century [Pea95]. During the past couple of decades, there has been a large body of work in computer science on this topic with a focus on computational efficiency [KMR+94].

The performance of a distribution learning (density estimation) algorithm is typically evaluated by the following criteria:

- *Sample Complexity:* For a given error tolerance, the algorithm should require a small number of samples, ideally matching the information-theoretic minimum.

- *Computational Complexity:* The algorithm should run in time polynomial in the number of samples provided as input.

- *Robustness:* The algorithm should provide error guarantees under model misspecification, i.e., even if the target distribution does not belong in the target family $\mathcal{D}$. The goal here is to be competitive with the best approximation of the unknown distribution by *any* distribution in $\mathcal{D}$.

In *non-proper* learning, the goal of the learning algorithm is to output an approximation to the target distribution without any constraints on its representation. In *proper* learning, we require in addition that the hypothesis is a member of the family $\mathcal{D}$. Note that these two notions of learning are essentially equivalent in terms of sample complexity (given any accurate hypothesis, we can do a brute-force search to find its closest distribution in $\mathcal{D}$), but not necessarily equivalent in terms of computational complexity.

In many learning situations it is desirable to compute a proper hypothesis, i.e., one that belongs to the underlying family $\mathcal{D}$. A proper hypothesis is usually preferable due to its interpretability. In particular, a practitioner may not want to use a density estimate, unless it is proper. For example, one may want the estimate to have the properties of the underlying family, either because this reflects some physical understanding of the inference problem, or because one might only be using the density estimate as the first stage of a more involved procedure.

The aforementioned discussion raises the following algorithmic question: *Can one obtain a* proper *learning algorithm for a given distribution family $\mathcal{D}$ whose running time matches that of the* best non-proper *algorithm for $\mathcal{D}$?* Perhaps surprisingly, our understanding of this natural question remains quite poor. In particular, little is known about the complexity of proper learning in the unsupervised setting of learning probability distributions. In contrast, the computational complexity of proper learning has been extensively investigated in the supervised setting of PAC learning Boolean functions [KV94, Fel15], with several algorithmic and computational intractability results obtained in the past decades.

In this work, we study the problem of *robust proper learning* for the family of univariate log-concave distributions (over $\mathbb{R}$ or $\mathbb{Z}$) (see Section 1.2 for a precise definition). Log-concave distributions constitute a broad non-parametric family that is very useful for modeling and inference [Wal09]. In the discrete setting, log-concave distributions encompass a range of fundamental types of discrete distributions, including binomial, negative binomial, geometric, hypergeometric, Poisson, Poisson Binomial, hyper-Poisson, Pólya-Eggenberger, and Skellam distributions (see Section 1 of [FBR11]). In the continuous setting, they include uniform, normal, exponential, logistic, extreme value, Laplace, Weibull, Gamma, Chi and Chi-Squared and Beta distributions (see [BB05]). Log-concave distributions have been studied in a wide range of different contexts including economics [An95], statistics and probability theory (see [SW14] for a recent survey), theoretical computer science [LV07], and algebra, combinatorics and geometry [Sta89].

1.2 Our Results and Comparison to Prior Work

The problem of density estimation for log-concave distributions is of central importance in the area of non-parametric shape constrained inference. As such, this problem has received significant attention in the statistics literature, see [CS10, DR09, DW16, CS13, KS14, BD14, HW16] and references therein, and, more recently, in theoretical computer science [CDSS13, CDSS14a, ADLS15, ADK15, CDGR16, DKS16]. In this section, we state our results and provide a brief comparison to the most relevant prior work. See Section 1.3 for a a more detailed summary of related work.

We study univariate log-concave distributions over both continuous and discrete domains.

Definition 1. A function $f : \mathbb{R} \to \mathbb{R}_+$ with respect to Lebesgue measure is log-concave if $f = \exp(\phi)$ where $\phi : \mathbb{R} \to [-\infty, \infty)$ is a concave function. A function $f : \mathbb{Z} \to [0, 1]$ is log-concave if $f^2(x) \geq f(x-1) \cdot f(x+1)$ for all $x \in \mathbb{Z}$ and f has no internal zeroes. We will denote by $\mathcal{LC}(D)$ the family of log-concave densities over D.

We use the following notion of agnostic learning under the total variation distance, denoted by d_{TV}:

Definition 2 (Agnostic Proper Learning)**.** Let $\mathcal{D}$ be a family of probability density functions on domain D. A randomized algorithm $A^{\mathcal{D}}$ is an *agnostic distribution learning algorithm for* $\mathcal{D}$, if for any $\epsilon > 0$, and any probability density function $f : D \to \mathbb{R}_+$, on input ϵ and sample access to f, with probability $9/10$, algorithm $A^{\mathcal{D}}$ outputs a hypothesis density $h \in \mathcal{D}$ such that $d_{\mathrm{TV}}(h, f) \leq O(\mathrm{OPT}) + \epsilon$, where $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{g \in \mathcal{D}} d_{\mathrm{TV}}(f, g)$.

Given the above terminology, we can state our main algorithmic result:

Theorem 3 (Main Result)**.** *There exists an algorithm that, given $n = O(\epsilon^{-5/2})$ samples from an arbitrary density $f : D \to \mathbb{R}_+$, where $D = \mathbb{R}$ or $D = \mathbb{Z}$, runs in time $\widetilde{O}(n^{8/5})$ and outputs a hypothesis $h \in \mathcal{LC}(D)$ such that with probability at least* $9/10$ *it holds $d_{\mathrm{TV}}(h, f) \leq O(\mathrm{OPT}) + \epsilon$, where* $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{g \in \mathcal{LC}(D)} d_{\mathrm{TV}}(f, g)$.

We note that the sample complexity of our algorithm is optimal (up to constant factors), as follows from previous work [DL01, CDSS13]. Our algorithm of Theorem 3 is the first polynomial time *agnostic proper* learning algorithm for the family of log-concave distributions.

In particular, previous polynomial time learning algorithms for log-concave distributions were either *non-proper* [CDSS13, CDSS14a, ADLS15] or *non-agnostic* [ADK15, CDGR16]. Specifically, the sequence of works [CDSS13, CDSS14a, ADLS15] give computationally efficient agnostic learning algorithms that are inherently non-proper. Two recent works [ADK15, CDGR16] give proper learning algorithms for discrete log-concave distributions that are provably non-agnostic. It should be noted that the sample complexity and running time of the non-robust proper algorithms in [ADK15, CDGR16] are significantly worse than ours. We elaborate on this point in the following subsection.

1.3 Related Work

Distribution Learning. Distribution learning is a paradigmatic inference problem with an extensive literature in statistics (see, e.g., the books [BBBB72, DG85, Sil86, Sco92, DL01]). A number of works in the statistics community have proposed proper estimators (relying on a maximum likelihood approach) for various distribution families. Alas, typically, these estimators are either intractable or their computational complexity is not analyzed.

A body of work in theoretical computer science has focused on distribution learning from a computational complexity perspective; see, e.g., [KMR$^+$94, FM99, AK01, CGG02, VW02, FOS05, BS10, KMV10, DDS12a, DDS12b, DDO$^+$13, CDSS13, CDSS14a, CDSS14b, ADLS15, DKS15b, DDKT15, DKS15a]. We note that, while the majority of the literature studies either non-proper learning or parameter estimation, proper learning algorithms have been obtained for a number of families, including mixtures of simple parametric models [FOS05, DK14, SOAJ14, LS15], and, Poisson binomial distributions [DKS15c].

Prior Work on Learning Log-concave Distributions. Density estimation of log-concave distributions has been extensively investigated in the statistics literature [DR09, GW09, Wal09, DW16, BJRP13, CS13, KS14, BD14] with a focus on analyzing the maximum likelihood estimator (MLE). For the *continuous* case, the sample complexity of the problem has been characterized [DL01], and it is known [KS14, HW16] that the MLE is sample efficient. It has been shown [DR11] that the MLE for continuous log-concave densities c an be formulated as a convex program, but no explicit upper bound on its running time is known. We remark here that the MLE is known to be non-agnostic with respect to the total variation distance, even for very simple settings (e.g., for Gaussian distributions).

Recent work in theoretical computer science [CDSS13, CDSS14a, ADLS15] gives sample-optimal, agnostic, and computationally efficient algorithms for learning log-concave distributions (both continuous and discrete). Alas, all of these algorithms are *non-proper*, i.e., they output a hypothesis that is not log-concave. For the case of *discrete* log-concave distributions supported on $[n]$, two recent papers [ADK15, CDGR16] obtain proper algorithms that use $\mathrm{poly}(1/\epsilon)$ samples and run in time $\mathrm{poly}(n/\epsilon)$. Roughly speaking, [ADK15, CDGR16] proceed by formulating the proper learning problem as a convex program.

Here we would like to emphasize three important differences between [ADK15, CDGR16] and the guarantees of Theorem 3. First, the algorithms of [ADK15, CDGR16] are *inherently non-agnostic*. Second, their sample complexity is sub-optimal, namely $\Omega(1/\epsilon^5)$, while our algorithm is sample-optimal. Third, the linear programming formulation that they employ has size (i.e., number of variables and constraints) $\Omega(n)$, i.e., its size depends on the support of the underlying distribution. As a consequence, the runtime of this approach is prohibitively slow, for large n. In sharp contrast, our algorithm's running time is independent of the support size, and scales sub-quadratically with the number of samples.

1.4 Overview of our Techniques

In this section, we provide a high-level overview of our techniques. Our approach to the proper learning problem is as follows: Starting with an accurate non-proper hypothesis, we fit a log-concave density to this hypothesis. This fitting problem can be formulated as a (non-convex) discrete optimization problem that we can solve efficiently by a combination of structural approximation results and dynamic programming. Specifically, we are able to phrase this optimization problem as a shortest path computation in an appropriately defined edge-weighted directed acyclic graph.

In more detail, our agnostic proper learning algorithm works in two steps: First, we compute an accurate *non-proper* hypothesis, g, by applying any efficient non-proper agnostic learning algorithm as a black-box (e.g., [CDSS14a, ADLS15]). In particular, we will use the non-proper learning algorithm of [ADLS15] that outputs a piecewise linear hypothesis distribution g. To establish the sample-optimality of the [ADLS15] algorithm, one requires the following structural result that we establish (Theorem 12): Any log-concave distribution (continuous or discrete) can be ϵ-approximated, in total variation distance, by a piecewise linear distribution with $O(\epsilon^{-1/2})$ interval pieces. Since $\Omega(\epsilon^{-1/2})$ interval pieces are required for such an approximation, our bound on the number of intervals is tight. It should be noted that a quantitatively similar structural result was shown in [CDSS14a] for *continuous* log-concave distributions, with a bound on the number of pieces that is sub-optimal up to logarithmic factors. For the discrete case, no such structural result was previously known.

Since g is not guaranteed to be log-concave, our main algorithmic step efficiently post-processes g to compute a log-concave distribution that is (essentially) as close to g as possible, in total variation distance. To achieve this, we prove a new structural result (Lemma 7) showing that the closest log-concave distribution can be well-approximated by a log-concave *piecewise exponential* distribution whose pieces are determined only by the mean and standard deviation of g. Furthermore, we show (Proposition 9) we can assume that the values of this approximation at the breakpoints can be appropriately discretized. These structural results are crucial for our algorithmic step outlined below.

From this point on, our algorithm proceeds via dynamic programming. Roughly speaking, we record the best possible error in approximating g by a function of the aforementioned form on the interval $(-\infty, x]$ for various values of x and for given values of $h(x), h'(x)$. Since knowing $h(x)$ and $h'(x)$ is all that we need in order to ensure that the rest of the function is log-concave, this is sufficient for our purposes. It turns out that this dynamic program can be expressed as a shortest path computation in a graph that we construct. The time needed to compute the edge weights of this graph depends on the description of the non-proper hypothesis g. In our case, g is a piecewise linear distribution and all these computations are manageable.

1.5 Organization

In Section 2 we record the basic probabilistic ingredients we will require. In Section 3 we prove our main result. Finally, we conclude with a few open problems in Section 4.

2 Preliminaries

For $n \in \mathbb{Z}_+$, we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. For $u \in \mathbb{R}$, we will denote $\exp(u) \stackrel{\text{def}}{=} e^u$. Let $f : \mathbb{R} \to \mathbb{R}$ be a Lebesgue measurable function. We will use $f(A)$ to denote $\int_{x \in A} f(x)dx$.

A Lebesgue measurable function $f : \mathbb{R} \to \mathbb{R}$ is a probability density function (pdf) if $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{\mathbb{R}} f(x)dx = 1$. We say that $f : \mathbb{R} \to \mathbb{R}$ is a pseudo-distribution if $f(x) \geq 0$ for all $x \in \mathbb{R}$. A function $f : \mathbb{Z} \to \mathbb{R}$ is a probability mass function (pmf) if if $f(x) \geq 0$ for all $x \in \mathbb{Z}$ and $\sum_{x \in \mathbb{Z}} f(x) = 1$. We will similarly use $f(A)$ to denote $\sum_{x \in A} f(x)$. We say that $f : \mathbb{Z} \to \mathbb{R}$ is a pseudo-distribution if $f(x) \geq 0$ for all $x \in \mathbb{Z}$.

For uniformity of the exposition, we will typically use D to denote the domain of our functions, where D is $\mathbb{R}$ for the continuous case and $\mathbb{Z}$ for the discrete case. We will use the term *density* to refer to either a pdf or pmf.

The L_1-distance between $f, g : D \to \mathbb{R}$ over $I \subseteq D$, denoted $\|f - g\|_1^I$, is $\int_I |f(x) - g(x)|dx$ for $D = \mathbb{R}$ and $\sum_{x \in I} |f(x) - g(x)|$ for $D = \mathbb{Z}$; when $I = D$ we suppress the superscript I. The *total variation distance* between densities $f, g : D \to \mathbb{R}_+$ is defined as $d_{\mathrm{TV}}(f, g) \stackrel{\text{def}}{=} (1/2) \cdot \|f - g\|_1$.

Our algorithmic and structural results make essential use of continuous piecewise exponential functions, that we now define:

Definition 4. Let $I = [\alpha, \beta] \subseteq D$, where $\alpha, \beta \in D$. A function $g : I \to \mathbb{R}_+$ is *continuous k-piecewise exponential* if there exist $\alpha \equiv x_1 < x_2 < \ldots < x_k < x_{k+1} \equiv \beta$, $x_i \in D$, such that for all $i \in [k]$ and $x \in I_i \stackrel{\text{def}}{=} [x_i, x_{i+1}]$ we have that $g(x) = g_i(x)$, where $g_i(x) \stackrel{\text{def}}{=} \exp(c_i x + d_i)$, $c_i, d_i \in \mathbb{R}$.

Note that the above definition implies that $g_i(x_{i+1}) = g_{i+1}(x_{i+1})$, for all $i \in [k]$.

We will also require a number of useful properties of log-concave densities, summarized in the following lemma:

Lemma 5. *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ and standard deviation σ. Then: (i) If $D = \mathbb{R}$ or $D = \mathbb{Z}$ and σ is at least a sufficiently large constant, $1/(8\sigma) \leq M_f \stackrel{\text{def}}{=} \max_{x \in D} f(x) \leq 1/\sigma$, and (ii) $f(x) \leq \exp\left(1 - |x - \mu| M_f / e\right) M_f$ for all $x \in D$.*

For the case of continuous log-concave densities, (i) appears as Lemma 5.5 in [LV07], and the discrete case follows similarly. To show (ii) we note that, since f is unimodal, $f(\mu + e/M_f)$ and $f(\mu - e/M_f)$ are each at most M_f/e. The claim then follows from log-concavity.

Lemma 6. *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ and standard deviation σ. Assume that either $D = \mathbb{R}$ or that σ is sufficiently large. Let $g : D \to \mathbb{R}_+$ be a density with $d_{\mathrm{TV}}(f, g) \leq 1/10$. Given an explicit description of g, we can efficiently compute values $\tilde{\mu}$ and $\tilde{\sigma}$ so that $|\mu - \tilde{\mu}| \leq 2\sigma$ and $3\sigma/10 \leq \tilde{\sigma} \leq 6\sigma$.*

The proof of the above lemma uses the log-concavity of f and is deferred to Appendix A.1.

3 Proof of Theorem 3: Our Algorithm and its Analysis

3.1 Approximating Log-concave Densities by Piecewise Exponentials

Our algorithmic approach relies on approximating log-concave densities by continuous piecewise exponential functions. Our first structural lemma states that we can approximate a log-concave density by a continuous piecewise exponential pseudo-distribution with an appropriately small set of interval pieces.

Lemma 7. *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ, standard deviation σ at least a sufficiently large constant, and $\epsilon > 0$. Let $I = [\alpha, \beta] \subseteq D$ be such that $\alpha < \mu < \beta$ and $|\alpha - \mu|, |\beta - \mu| = \Theta(\log(1/\epsilon)\sigma)$, with the implied constant sufficiently large. Let k be an integer so that either $k = \Theta(\log(1/\epsilon)/\epsilon)$, or $k = \beta - \alpha = O(\log(1/\epsilon)/\epsilon)$ and $D = \mathbb{Z}$. Consider the set of equally spaced endpoints $\alpha \equiv x_1 < x_2 < \ldots < x_k < x_{k+1} \equiv \beta$. There exist indices $1 \leq l < r \leq k+1$ and a log-concave continuous piecewise exponential pseudo-distribution $g : I \to \mathbb{R}_+$ with $\|f - g\|_1 \leq O(\epsilon)$ such that the following are satisfied:*

(i) $g(x) = 0$, for all $x \notin J \stackrel{\text{def}}{=} [x_l, x_r]$.

(ii) For all $l \leq i \leq r$ it holds $g(x_i) = f(x_i)$.

(iii) For $l \leq i < r$, g is exponential on $[x_i, x_{i+1}]$.

Proof. If $D = \mathbb{Z}$ and σ is less than a sufficiently small constant, then Lemma 5 implies that all but an ϵ-fraction of the mass of f is supported on an interval of length $O(\log(1/\epsilon))$. If we let I be this interval and take $k = |I|$, we can ensure that $g = f$ on I and our result follows trivially. Henceforth, we will assume that either $D = \mathbb{R}$ or that σ is sufficiently large.

The following tail bound is a consequence of Lemma 5 and is proved in Appendix A.2:

Claim 8. *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ and standard deviation σ. Let $\alpha \leq \mu - \Omega(\sigma(1 + \log(1/\epsilon)))$ and $\beta \geq \mu + \Omega(\sigma(1 + \log(1/\epsilon))))$. Then, $\|f\|_1^{(-\infty,\alpha)} \leq \epsilon$ and $\|f\|_1^{(\beta,\infty)} \leq \epsilon$.*

By Claim 8, it suffices to exhibit the existence of the function $g : I \to \mathbb{R}_+$ and show that $\|f - g\|_1^I \leq O(\epsilon)$. We note that if $D = \mathbb{Z}$ and $k = \beta - \alpha$, then we may take $g = f$ on I, and this follows immediately. Hence, it suffices to assume that $k > C(\log(1/\epsilon)/\epsilon)$, for some appropriately large constant $C > 0$.

Next, we determine appropriate values of l and r. In particular, we let l be the minimum value of i so that $f(x_i) > 0$, and let r be the maximum. We note that the probability measure of $\mathrm{supp}(f) \setminus J$ is at most $O(|\beta - \alpha|/k) = O(\epsilon\sigma)$. Since $M_f = O(1/\sigma)$, by Lemma 5(i), we have that $f(I \setminus J) = O(\epsilon)$. Therefore, it suffices to show that $\|f - g\|_1^J = O(\epsilon)$.

We take $k = \Omega(\log(1/\epsilon)/\epsilon)$. Since the endpoints $x_1, \ldots, x_{k+1}$ are equally spaced, it follows that $L = |x_{i+1} - x_i| = O(\epsilon\sigma)$.

For $l \leq i < r$, for $x \in [x_i, x_{i+1}]$ we let $g(x)$ be given by the unique exponential function that interpolates $f(x_i)$ and $f(x_{i+1})$. We note that this g clearly satisfies properties (i), (ii), and (iii). It remains to show that $\|f - g\|_1^J = O(\epsilon)$.

Let $I_j = [x_j, x_{j+1}]$ be an interval containing a mode of f. We claim that $\|f-g\|_1^{I_j} = O(\epsilon)$. This is deduced from the fact that $\max_x g(x) \leq \max_x f(x) \leq 1/\sigma$, where the first inequality is by the definition of g and the second follows by Lemma 5 (i). This in turn implies that the probability mass of both $f(I_j)$ and $g(I_j)$ is at most $L \cdot (1/\sigma) = O(\epsilon)$.

We now bound from above the contribution to the error coming from the intervals $I_1, \ldots, I_{j-1}$, i.e., the quantity $\sum_{i=1}^{j-1} \|f-g\|_1^{I_i}$. Since all I_i's have length L, and f, g are monotone non-decreasing agreeing on the endpoints, we have that the aforementioned error term is at most

$$L \cdot \sum_{i=1}^{j-1} (f(x_{i+1}) - f(x_i)) = O(\epsilon\sigma) \cdot M_f \leq O(\epsilon\sigma) \cdot (1/\sigma) = O(\epsilon) \ .$$

A symmetric argument shows the error coming from the intervals $I_{j+1}, \ldots, I_k$ is also $O(\epsilon)$. An application of the triangle inequality completes the proof. □

The following proposition establishes the fact that the log-concave piecewise exponential approximation can be assumed to be appropriately discretized:

Proposition 9. *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ, standard deviation σ at least a sufficiently large constant, and $\epsilon > 0$. Let $\tilde{\sigma} = \Theta(\sigma)$. Let $I = [\alpha, \beta] \subseteq D$ containing μ be such that $|\alpha - \mu|, |\beta - \mu| = \Theta(\log(1/\epsilon)\tilde{\sigma})$, where the implied constant is sufficiently large. Consider a set of equally spaced endpoints $\alpha \equiv x_1 < x_2 < \ldots < x_k < x_{k+1} \equiv \beta$, $x_i \in D$, where either $k = \Theta(\log(1/\epsilon)/\epsilon)$, or $D = \mathbb{Z}$ and $k = \beta - \alpha = O(\log(1/\epsilon)/\epsilon)$. There exist indices $1 \leq l < r \leq k+1$ and a log-concave continuous piecewise exponential pseudo-distribution $h : I \to \mathbb{R}_+$ with $\|f - h\|_1 \leq O(\epsilon)$ such that the following are satisfied:*

(i) $h(x) > 0$ if and only if $x \in J \stackrel{\text{def}}{=} [x_l, x_r]$.

(ii) For each endpoint x_i, $i \in [l, r]$, we have that (a) $\log(h(x_i)\tilde{\sigma})$ is an integer multiple of ϵ/k, and (b) $-O(\log(1/\epsilon)) \leq \log(h(x_i)\tilde{\sigma}) \leq O(1)$.

(iii) For any $i \in [l, r-1]$ we have $|\log(h(x_i)\tilde{\sigma}) - \log(h(x_{i+1})\tilde{\sigma})|$ is of the form $b \cdot \epsilon \cdot 2^c/k$, for integers $|b| \leq (1/\epsilon)\log(1/\epsilon)$ and $0 \leq c \leq O(\log(1/\epsilon))$.

Proof. Let $g : I \to \mathbb{R}_+$ be the pseudo-distribution given by the Lemma 7. We will construct our function $h : I \to \mathbb{R}_+$ such that $\|h - g\|_1 = O(\epsilon)$. For notational convenience, for the rest of this proof we will denote $a_i \stackrel{\text{def}}{=} \log(h(x_i)\tilde{\sigma})$ and $a'_i \stackrel{\text{def}}{=} \log(g(x_i)\tilde{\sigma})$ for $i \in [k+1]$.

We define the function h to be supported on the interval $J = [x_l, x_r]$ specified as follows: The point x_l is the leftmost endpoint such that $g(x_l) \geq \epsilon^2/\tilde{\sigma}$ or equivalently $a'_l \geq -2\log(1/\epsilon)$. Similarly, the point x_r is the rightmost endpoint such that $g(x_r) \geq \epsilon^2/\tilde{\sigma}$ or equivalently $a'_r \geq -2\log(1/\epsilon)$.

We start by showing that the probability mass of g outside the interval J is $O(\epsilon)$. This is because $g(x) \leq \epsilon^2/\tilde{\sigma}$ off of J, and so has total mass at most $\epsilon^2/\tilde{\sigma}(\beta - \alpha) = O(\epsilon)$.

To complete the proof, we need to appropriately define h so that it satisfies conditions (ii) and (iii) of the proposition statement, and in addition that $\|h - g\|_1^J \leq O(\epsilon)$.

We note that $-O(\log(1/\epsilon)) \leq a'_i \leq O(1)$ for all $i \in [l, r]$. Indeed, since $a'_l, a'_r \geq -2\log(1/\epsilon)$ and g is log-concave, we have that $a'_i \geq -\log(1/\epsilon)$ for all $i \in [l, r]$. Also, since $a'_i =$

$\log(g(x_i)\tilde{\sigma}) = \log(f(x_i)\tilde{\sigma})$, we obtain $a'_i \leq \log(M_f \tilde{\sigma}) \leq \log 6 \leq 1.8$. We will construct h so that $|a'_i - a_i| = O(\epsilon)$ for all $l \leq i \leq r$. We claim that this is sufficient since it would imply that $\log(g(x)/h(x)) = O(\epsilon)$ for all $x \in J$. This in turn implies that $h(x) = g(x) + O(\epsilon g(x))$, and thus that $\|h - g\|_1^J = \int_J O(\epsilon g(x))dx = O(\epsilon)$.

We are now ready to define a_i, $i \in [l, r]$. Let j be such that x_j is a mode of g. Let d_i be obtained by rounding $d'_i \stackrel{\text{def}}{=} a'_i - a'_{i-1}$ as follows: Let c_i be the least non-negative integer such that $|d'_i| \leq 2^{c_i} \log(1/\epsilon)/k$. Then, we define d_i to be d'_i rounded to the nearest integer multiple of $2^{c_i}\epsilon/k$ (rounding towards 0 in the case of ties). Let a_j be the nearest multiple of (ϵ/k) to a'_j. Let $a_i = a_j + \sum_{k=j+1}^{i} d_k$ for $j > i$, and $a_i = a_j - \sum_{k=i+1}^{j} d_k$ for $i < j$. We define h to be the continuous piecewise exponential function with $h(x_i) = \exp(a_i)/\tilde{\sigma}$, $i \in [l, r]$, that is exponential on each of the intervals $[x_i, x_i + 1]$, for $i \in [l, r - 1]$.

By construction, for all $i \in [l, r]$, a_i is an integer multiple of ϵ/k and $|a_i - a_{i+1}|$ is of the form $b \cdot \epsilon \cdot 2^c/k$ for integers $0 \leq b \leq (1/\epsilon)\log(1/\epsilon)$, and $0 \leq c \leq O(\log 1/\epsilon)$. Since g is log-concave, we have $a'_i - a'_{i-1} \geq a'_{i+1} - a'_i$. Note that the rounding of the d_i's is given by a monotone function and thus we also have $a_i - a_{i-1} \geq a_{i+1} - a_i$. Hence, h is also log-concave. Since $|a'_i| \leq \log(1/\epsilon)$, $i \in [l, r]$, the definition of the c_i's yields $\sum_i 2^{c_i} \log(1/\epsilon)/k = O(\log(1/\epsilon))$ or $\sum_i 2^{c_i} = O(k)$. Since $|d'_i - d_i| \leq 2^{c_i}\epsilon/k$, we have that $|a_i - a'_i| \leq (\epsilon/k) + \sum_i 2^{c_i}\epsilon/k \leq O(\epsilon)$. This completes the proof. □

3.2 Main Algorithm

Theorem 10. *Let $g : D \to \mathbb{R}_+$ be a density and let $\mathrm{OPT} = \inf_{f \in \mathcal{LC}(D)} d_{\mathrm{TV}}(g, f)$. There exists an algorithm that, given g and $\epsilon > 0$, outputs an explicit log-concave density h such that $d_{\mathrm{TV}}(g, h) = O(\mathrm{OPT} + \epsilon)$. The algorithm has running time $O((t + 1)\mathrm{polylog}(1/\epsilon)/\epsilon^4)$, where t is the average across the intervals of an upper bound on the time needed to approximate $\|g - h\|_1^{[x_i, x_{i+1}]}$ to within $O(\epsilon^2/\log(1/\epsilon))$.*

Proof. Let f be a log-concave density such that $d_{\mathrm{TV}}(g, f) = \mathrm{OPT}$. First, we compute the median $\tilde{\mu}$ and interquartile range $\tilde{\sigma}$. If $\mathrm{OPT} \leq 1/10$, Lemma 6 applies to these, and otherwise Theorem 10 is trivial. Using these approximations, we construct an interval $I \subseteq D$ containing at least $C\log(1/\epsilon)$ standard deviations about the mean of f, and of total length $O(\log(1/\epsilon)\sigma)$, where C is a sufficiently large constant. If $D = \mathbb{Z}$ and $1/\epsilon = O(\sigma)$, we let k be the length of I, otherwise, we let $k = \Theta(\log(1/\epsilon)/\epsilon)$ and ensure that the length of I divided by k is in D.

We will attempt to find a log-concave pseudo-distribution h satisfying the properties of Proposition 9 so that $d_{\mathrm{TV}}(h, g)$ is (approximately) minimized. Note that the proposition implies there exists a log-concave pseudo-distribution h with $d_{\mathrm{TV}}(f, h) = O(\epsilon)$, and thus $d_{\mathrm{TV}}(g, h) = O(\mathrm{OPT} + \epsilon)$. Given any such h with $d_{\mathrm{TV}}(h, g) = O(\mathrm{OPT} + \epsilon)$, re-normalizing gives an explicit log-concave density h' with $d_{\mathrm{TV}}(h', g) = O(\mathrm{OPT} + \epsilon)$.

We find the best such h via dynamic programming. In particular, if $x_1, \ldots, x_{k+1}$ are the interval endpoints, then h is determined by the quantities $a_i = \log(h(x_i)\tilde{\sigma})$, which are either $-m \cdot \epsilon/k$, where $m \in \mathbb{Z}_+$ $|m| \leq O((1/\epsilon)\log(1/\epsilon)k)$, or $-\infty$. The condition that h is log-concave is equivalent to the sequence a_i being concave.

Let S be the set of possible a_i 's, i.e., the multiples of ϵ/k in the range $[-\log(1/\epsilon), O(1)] \cup \{-\infty\}$. Let T be the set of possible $a_{i+1} - a_i$'s, i.e., numbers of the form $b \cdot \epsilon \cdot 2^c/k$, for integers $|b| \le (1/\epsilon)\log(1/\epsilon)$ and $0 \le c \le O(\log(1/\epsilon))$. Let H be the set of h which satisfy the properties of Proposition 9 except the bound on $\|h-f\|_1$. We use dynamic programming to determine for each $i \in [k], a \in S, d \in T$ the concave sequence $a_1, \ldots, a_i$ so that $a_i = a$, $a_i - a_{i-1} = d$ and $\|h-g\|_1^{[x_1,x_i]}$ is as small as possible, where $h(x)$ is the density obtained by interpolating the a_i's by a piecewise exponential function.

We write $e_i(a_i, a_{i+1})$ for the error in the i-th interval $[x_i, x_{i+1}]$. When a_i and a_{i+1} are both finite, we take $e_i(a_i, a_{i+1}) \overset{\text{def}}{=} \|g-h_i\|_1^{[x_i,x_{i+1}]}$, where $h_i(x) = \exp\left(\frac{x-x_i}{x_{i+1}-x_i}a_i + \frac{x_{i+1}-x_i}{x_{i+1}-x_i}a_{i+1}\right)/\sigma$. We define $e_i(a,-\infty) = e_i(-\infty,a) \overset{\text{def}}{=} \|g\|_1^{[x_i,x_{i+1}]}$. Thus, we have $\|g,h\|_1^{[x_1,x_i]} \le \sum_{i=1}^{k} e_i(a_i,a_{i+1})$. When $D = \mathbb{R}$, this an equality. However, when $D = \mathbb{Z}$, we double count the error in the endpoints in the interior of the support, and so have $\sum_{i=1}^{k} e_i(a_i,a_{i+1}) \le 2\|g-h\|_1^{[x_1,x_i]}$.

The algorithm computes $\tilde{e}_i(a_{i-1},a_i)$ with $|\tilde{e}_i(a_{i-1},a_i) - e_i(a_{i-1},a_i)| \le \epsilon/k$ for all $a_{i-1}, a_i \in S$ with $a_i - a_{i-1} \in T$.

```
Algorithm Compute $h$
Input: an oracle for computing $\tilde{e}_i(a,a')$
Output: a sequence $a_1,\ldots,a_n$ that minimizes $\sum_{i=1}^{k} e_{i+1}(a_i,a_{i+1})$

  1. Let $G$ be the directed graph with vertices of the form:
       (a) $(0,-\infty,-)$,$(k+1,\infty,+)$
       (b) $(i,a,d)$ for $i \in [k], a \in S\backslash\{-\infty\}, d \in T \cup \{\infty\}$
       (c) $(i,-\infty,s)$ for $i \in [k], s \in \{\pm\}$
     and weighted edges of the form
       (a) $(i,-\infty,-)$ to $(i+1,a,\infty)$ of weight $\tilde{e}_i(-\infty,a)$
       (b) $(i,a,d)$ to $(i+1,a+d,d)$ of weight $\tilde{e}_i(a,a+d)$
       (c) $(i,a,d)$ to $(i,a,d')$ with $d'$ the predecessor of $d$ in $T \cup \{\infty\}$ or weight 0.
       (d) $(i,a,d)$ to $(i+1,-\infty,+)$ of weight $e_i(a,-\infty)$
       (e) $(i,-\infty,s)$ to $(i+1,-\infty,s)$ of weight $e_i(-\infty,-\infty)$
  2. Using the fact that $G$ is a DAG compute the path $P$ from $(0,-\infty,-)$ to $(k+1,-\infty,+)$ of smallest weight.
  3. For each $i \in [k]$, let $a_i$ be the value such that $P$ passes through a vertex of the form $(i,a_i,d^*)$.
```

```
Algorithm Full-Algorithm
Input: A concise description of a distribution $g$ such that $d_{\mathrm{TV}}(f,g) \leq \mathrm{OPT}$ for some log-concave distribution $f$ and $\epsilon > 0$.
Output: A log-concave continuous piecewise exponential $h$ with $d_{\mathrm{TV}}(g,h) \leq O(\mathrm{OPT}+\epsilon)$

1. Compute the median $\tilde{\mu}$ and interquartile range $\tilde{\sigma}$ of $g(x)$
2. If $D = \mathbb{Z}$ and $\tilde{\sigma} = O(1/\epsilon)$,
3.     let $\alpha = \tilde{\mu} - \Theta(\log(1/\epsilon)/\epsilon)$ and $\beta = \tilde{\mu} + \Theta(\log(1/\epsilon)/\epsilon)$ be integers, $k = \beta - \alpha$ and $L = 1$,
4. else let $L = \Theta(\epsilon\tilde{\sigma})$ with $L \in \mathbb{Z}$ or $\mathbb{R}$ in the discrete and continuous cases respectively, $k = \Theta(\log(1/\epsilon)/\epsilon)$ be an even integer, $\alpha = \tilde{\mu} - (k/2)$ and $\beta = \tilde{\mu} - (k/2)$.
5. Let $x_i = \alpha + (i+1)L$ for $1 \leq i \leq k+1$.
6. Let $S$ be the set of the multiples of $\epsilon/k$ in the range $[-\ln(1/\epsilon), O(1)] \cup \{-\infty\}$. Let $T$ be the set of numbers of the form $b \cdot \epsilon \cdot 2^c/k$, for integers $|b| \leq (1/\epsilon)\ln(1/\epsilon)$ and $0 \leq c \leq O(\log(1/\epsilon))$.
7. Sort $T$ into ascending order.
8. Let $a_1, \ldots, a_{k+1}$ be the output of algorithm Compute $h$.
9. Return the continuous piecewise exponential $h(x)$ that has $h(x_i) = \exp(a_i)/\tilde{\sigma}$ for all $1 \leq i \leq k+1$ and has endpoints $x_l, x_{l+1}, \ldots, x_r$, where $l$ and $r$ and the least and greatest $i$ such that $a_i$ is finite.
```

Now we show correctness. For every $h' \in H$, with log probabilities at the endpoints a'_i, there is a path of weight $w_{h'} := \sum_{i=1}^{k} \tilde{e}_i(a'_i, a'_{i+1})$ which satisfies

$$\|g - h'\|_1^I - \epsilon \leq w_{h'} \leq 2\|g - h'\|_1^I + \epsilon \ .$$

Thus, the output $h(x)$ has $\|g - h\|_1^I \leq 2\epsilon + 2\min_{h' \in H} \|g - h'\|_1^I$. By Proposition 9, there is an $h^* \in H$ with $\|f - h^*\|_1^I \leq O(\epsilon)$, where $d_{\mathrm{TV}}(f,g) = \mathrm{OPT}$. Thus, $\|g - h^*\|_1^I \leq \mathrm{OPT} + O(\epsilon)$. Therefore, we have that

$$\|g - h\|_1^I \leq 2\|g - h^*\|_1^I + 2\epsilon \leq 2\mathrm{OPT} + O(\epsilon) \ .$$

Since the mass of f outside of I is $O(\epsilon)$, we have that the mass of g outside of I is at most $\mathrm{OPT} + O(\epsilon)$. Thus, $d_{\mathrm{TV}}(g,h) \leq \mathrm{OPT} + O(\epsilon) + \|g - h\|_1^I = O(\mathrm{OPT} + \epsilon)$, as required.

Finally we analyze the time complexity. The graph G has $k|S||T| + 2$ vertices. Each vertex has at most one in-edge of each type. Thus, we can find the shortest path in time $O(k|S||T|)$ plus the time it takes to compute every $\tilde{e}_i(a, a+d)$. There are $O(k|S||T|)$ such computations and they take average time at most t. Thus, the time complexity is

$$O(k|S||T|(t+1)) = O(\log(1/\epsilon)/\epsilon \cdot \log(1/\epsilon)^2/\epsilon^2 \cdot \log(1/\epsilon)^2/\epsilon \cdot (t+1)) = O((t+1)\log^5(1/\epsilon)/\epsilon^4).$$

□

3.3 Putting Everything Together

We are now ready to combine the various pieces that yield our main result. Our starting point is the following non-proper learning algorithm:

Theorem 11 ([ADLS15])**.** *There is an agnostic learning algorithm for t-piecewise linear distributions with sample complexity $O(t/\epsilon^2)$ and running time $O((t/\epsilon^2)\log(1/\epsilon))$. Moreover, the algorithm outputs an $O(t)$-piecewise linear hypothesis distribution.*

To establish that our overall learning algorithm will have the optimal sample complexity of $O(\epsilon^{-5/2})$, we make use of the following approximation theorem:

Theorem 12. *For any log-concave density f on either $\mathbb{R}$ or $\mathbb{Z}$, and $\epsilon > 0$, there exists a piecewise linear distribution g with $O(\epsilon^{-1/2})$ interval pieces so that $d_{\mathrm{TV}}(f, g) \leq \epsilon$.*

The proof of Theorem 12 is deferred to Appendix A.3. We now have all the ingredients to prove our main result.

Proof of Theorem 3. Let f' be a log-concave density with $d_{\mathrm{TV}}(f, f') = \mathrm{OPT}$. By Theorem 12, there is a piecewise linear density g' with $O(\epsilon^{-1/2})$ pieces that has $d_{\mathrm{TV}}(f', g') \leq \mathrm{OPT} + \epsilon$. By Theorem 11, there is an algorithm with sample complexity $O(1/\epsilon^{5/2})$ and running time $O((1/\epsilon^{5/2})\log(1/\epsilon))$ that computes a piecewise linear density g' with $O(\epsilon^{-1/2})$ pieces such that $d_{\mathrm{TV}}(f', g') \leq O(\mathrm{OPT} + \epsilon)$. We apply the algorithm of Theorem 10 to this g', which produces a piecewise exponential approximation $h(x)$ that satisfies $d_{\mathrm{TV}}(g', h) \leq O(\mathrm{OPT} + \epsilon)$, and therefore $d_{\mathrm{TV}}(h, f) \leq O(\mathrm{OPT} + \epsilon)$.

It remains to prove that the time complexity is $\tilde{O}(n^{8/5}) = \tilde{O}(1/\epsilon^4)$. To obtain this, we must show that $t = \mathrm{polylog}(1/\epsilon)$ in the statement of Theorem 10. When $D = \mathbb{Z}$ and the length of each interval is 1, we have $t = O(1)$. Otherwise, we divide into $k = \Theta((1/\epsilon)\log(1/\epsilon))$ pieces. Since $k \geq O(\epsilon^{-1/2})$, the average number of endpoints of $g'(x)$ in a piece of $h(x)$ is smaller than 1. Thus, to get the amortized time complexity to be $\mathrm{polylog}(1/\epsilon)$, it suffices to show this bound for an exponential and linear function on a single interval. The following claim is proved in Appendix A.4:

Claim 13. *Let $g(x) = ax + b$ and $h(x) = c\exp(dx)$. Let $I = [x', x' + L]$ be an interval with $g(x) \geq 0$ and $0 \leq h(x) \leq O(\epsilon/L)$ for all $x \in I$. There is an algorithm which approximates $\|g - h\|_1^I$ to within an additive $O(\epsilon^2/\log(1/\epsilon))$ in time* $\mathrm{polylog}(1/\epsilon)$.

This completes the proof of Theorem 3. □

4 Discussion and Future Directions

In this paper, we gave the first agnostic learning algorithm for log-concave distributions that runs in polynomial time. Our algorithm is sample-optimal and runs in time that is subquadratic in the size of its input sample. The obvious open problem is to obtain an agnostic proper learning algorithm that runs in near-linear time. More broadly, an interesting and challenging question is to generalize our techniques to the problem of learning log-concave distributions in higher dimensions.

We believe that our algorithmic approach should naturally extend to other structured distribution families, e.g., to monotone hazard rate (MHR) distributions, but we have not pursued this direction. Finally, as we point out in the following paragraph, our dynamic programming approach can be extended to properly learning mixtures of log-concave densities, alas with running time exponential in the number k of components, i.e., $(1/\epsilon)^{O(k)}$.

Indeed, the non-proper learning algorithm from [ADLS15] also applies to mixtures, so it suffices to efficiently compute a nearly optimal approximation of a given distribution by a mixture of k log-concave distributions. It is easy to see that we can assume each of the mixing weights is $\Omega(\epsilon)$. For our approach to work, we will need to approximate the mean and standard deviation of each distribution in the mixture. This can be done if we have $O(1)$ samples from each component, which can be accomplished by taking $O(1)$ samples from our original distribution and noting that with probability $\Omega(\epsilon)^{O(1)}$ it has chosen only samples from the desired component. After doing this, we will need to build a larger dynamic program. Our new dynamic program will attempt to approximate f as a mixture of functions h of the form given in Proposition 9. Specifically, it will need to have steps corresponding to each of the x_i's for each of the functions h, and will need to keep track of both the current value of each h and its current logarithmic derivative.

The aforementioned discussion naturally leads to our final open problem: Is there a proper learning algorithm for mixtures of k log-concave distributions with running time poly(k/ϵ)?

References

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *NIPS*, 2015.

[ADLS15] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[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.

[BB05] M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):pp. 445–469, 2005.

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

[BD14] F. Balabdaoui and C. R. Doss. Inference for a Mixture of Symmetric Distributions under Log-Concavity. Available at http://arxiv.org/abs/1411.4708, 2014.

[BJRP13] F. Balabdaoui, H. Jankowski, K. Rufibach, and M. Pavlides. Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):769–790, 2013.

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

[CDGR16] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *STACS*, pages 25:1–25:14, 2016.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDSS14b] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

[CGG02] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[CS10] M. Cule and R. Samworth. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B*, 72:545–607, 2010.

[CS13] Y. Chen and R. J. Samworth. Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, 23:1373–1398, 2013.

[DDKT15] C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. *CoRR*, abs/1511.03641, 2015. To appear in STOC 2016.

[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*, pages 1371–1385, 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.

[DK14] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT '14, pages 1183–1213, 2014.

[DKS15a] I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. *CoRR*, abs/1511.03592, 2015. To appear in STOC 2016.

[DKS15b] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015. To appear in COLT 2016.

[DKS15c] I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. *CoRR*, 2015. To appear in COLT 2016.

[DKS16] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning multivariate log-concave distributions. *CoRR*, abs/1605.08188, 2016.

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

[DR09] L. Dumbgen 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.

[DR11] L. Dümbgen and K. Rufibach. logcondens: Computations related to univariate log-concave density estimation. *J. Statist. Software*, 39(6), 2011.

[DW16] C. R. Doss and J. A. Wellner. Global rates of convergence of the mles of log-concave and s-concave densities. *Ann. Statist.*, 44(3):954–981, 06 2016.

[FBR11] H. Jankowski F. Balabdaoui and K. Rufibach. Maximum likelihood estimation and confidence bands for a discrete log-concave distribution, 2011.

[Fel15] V. Feldman. Hardness of proper learning (1988; pitt, valiant). In *Encyclopedia of Algorithms*. 2015.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual 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.

[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.

[HW16] Q. Han and J. A. Wellner. Approximation and estimation of s-concave densities via renyi divergences. *Ann. Statist.*, 44(3):1332–1359, 06 2016.

[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.

[KS14] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. Available at http://arxiv.org/abs/1404.2298, 2014.

[KV94] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, 1994.

[LS15] J. Li and L. Schmidt. A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k. *CoRR*, abs/1506.01367, 2015.

[LV07] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

[Pea95] K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Trans. of the Royal Society of London*, 186:343–414, 1895.

[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.

[SOAJ14] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *NIPS*, pages 1395–1403, 2014.

[Sta89] R. P. Stanley. Log-concave and unimodal sequences in algebra, combinatorics, and geometry. *Annals of the New York Academy of Sciences*, 576(1):500–535, 1989.

[SW14] A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: A review. *Statist. Surv.*, 8:45–114, 2014.

[VW02] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, pages 113–122, 2002.

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

A Omitted Proofs

A.1 Proof of Lemma 6

Lemma 6 *Let $f : D \to \mathbb{R}_+$ be a log-concave density with mean μ and standard deviation σ. Assume that either $D = \mathbb{R}$ or that σ is sufficiently large. Let $g : D \to \mathbb{R}_+$ be a density with $d_{\mathrm{TV}}(f, g) \le 1/10$. Given an explicit description of g, we can efficiently compute values $\tilde{\mu}$ and $\tilde{\sigma}$ so that $|\mu - \tilde{\mu}| \le 2\sigma$ and $3\sigma/10 \le \tilde{\sigma} \le 6\sigma$.*

Proof. We define $\tilde{\mu}$ to be the median of g and $\tilde{\sigma}$ to be the difference between the 25^{th} and 75^{th} percentiles of g. Since f and g are within total variation distance $1/10$, it follows that their Kolmogorov distance (i.e., the maximum distance between their cumulative distribution functions) is at most $1/10$. This implies that $\tilde{\mu}$ lies between the 40^{th} and 60^{th} percentiles of f. By Cantelli's inequality, we have that $\Pr_{X \sim f}[X - \mu \ge 2\sigma] \le 1/5$ and $\Pr_{X \sim f}[X - \mu \le -2\sigma] \le 1/5$. Thus, $|\mu - \tilde{\mu}| \le 2\sigma$.

Similarly, $\tilde{\sigma}$ lies between (a) the difference between the 65^{th} and 35^{th} percentile of f and (b) the difference between the 85^{th} and 15^{th} percentile of f. By Cantelli's inequality, we have that $\Pr_{X \sim f}[X - \mu \ge 3\sigma] \le 1/10$ and $\Pr_{X \sim f}[X - \mu \le -3\sigma] \le 1/10$. Thus, $\tilde{\sigma} \le 6\sigma$. For the other direction, note that $3/10$ of the probability mass of f lies between the 35^{th} and 65^{th} percentile. Since the maximum value of f is at most $1/\sigma$, by Lemma 5 (i), we conclude that the difference between the 65^{th} and 35^{th} percentiles is at least $3\sigma/10$. □

A.2 Proof of Claim 8

Claim 8 *Let f be a log-concave density with mean μ and standard deviation σ. Let $\alpha \le \mu - \Omega(\sigma(1 + \log(1/\epsilon)))$ and $\beta \ge \mu + \Omega(\sigma(1 + \log(1/\epsilon))))$. Then, $\|f\|_1^{(-\infty,\alpha)} \le \epsilon$ and $\|f\|_1^{(\beta,\infty)} \le \epsilon$.*

Proof. By Lemma 5 (ii), we have $f(x) \le \exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma$. In the case $D = \mathbb{R}$, we have $\int_{-\infty}^{\alpha} f(x)dx \le \int_{-\infty}^{\alpha} \exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma dx = 8e\sigma \exp\left(1 - \frac{|\alpha-\mu|}{8e\sigma}\right)$. This is at most ϵ when $|\alpha - \mu| \ge 8e\sigma(1 + \ln(8e/\epsilon))$, which holds by our bounds on $|\alpha - \mu|$.

In the case $D = \mathbb{Z}$, we have $\sum_{-\infty}^{\alpha-1} f(x) \le \sum_{-\infty}^{\alpha-1} \exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma$. Since $\exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma$ is monotonically increasing on $(\infty, \alpha]$, we have that $\exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma \le \int_x^{x+1} \exp\left(1 - \frac{|y-\mu|}{8e\sigma}\right)/\sigma dy$ for $x \le \alpha - 1$. Thus, we can bound this sum by the same integral to the one for the continuous case, i.e., $\sum_{-\infty}^{\alpha-1} \exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma \le \int_{-\infty}^{\alpha} \exp\left(1 - \frac{|x-\mu|}{8e\sigma}\right)/\sigma dx \le \epsilon$. A symmetric argument yields that the probability mass of f on (β, ∞) is $O(\epsilon)$. □

A.3 Proof of Theorem 12

Theorem 12 *If f is a log-concave density on either $\mathbb{R}$ or $\mathbb{Z}$, and $\epsilon > 0$, there exists a piecewise linear distribution g with $O(\epsilon^{-1/2})$ interval pieces so that $d_{\mathrm{TV}}(f, g) \le \epsilon$.*

We begin by proving this in the case where the range of f and the logarithmic derivative of f are both relatively small.

Lemma 14. *Let f be a log-concave function defined on an interval I in either $\mathbb{R}$ or $\mathbb{Z}$. Suppose furthermore, that the range of f is contained in an interval of the form $[a, 2a]$ for some a, and that the logarithmic derivative of f (or the log-finite difference of f in the discrete case) varies by at most $1/|I|$ on I. Then there exists a piecewise linear function g on I with $O(\epsilon^{-1/2})$ pieces so that $\|f-g\|_1 \leq O(\epsilon\|f\|_1)$.*

Proof. By scaling f, we may assume that $a = 1$. Note that the log-derivative or log-finite difference of f must be $O(1/|I|)$ everywhere. We now partition I into subintervals $J_1, J_2, \ldots, J_n$ so that on each J_i has length at most $\epsilon^{1/2}|I|$ and the logarithmic derivative (or finite difference) varies by at most $\epsilon^{1/2}/|I|$. Note that this can be achieved with $n = O(\epsilon^{-1/2})$ by placing an interval boundary every $O(\epsilon^{1/2}|I|)$ distance as well as every time the logarithmic derivative passes a multiple of $\epsilon^{1/2}/|I|$.

We now claim that on each interval J_i there exists a linear function g_i so that $\|g_i - f\|_\infty = O(\epsilon)$. Letting g be g_i on J_i will complete the proof.

Let $J_i = [y, z]$. We note that for $x \in J_i$ that

$$f(x) = f(y)\exp((x-y)\alpha)$$

for some α in the range spanned by the logarithmic derivative (or log finite difference) of f on J_i. Letting α_0 be some number in this range, we have that

$$\begin{aligned} f(x) &= f(y)\exp((x-y)\alpha_0 + (x-y)(\alpha-\alpha_0)) \\ &= f(y)\exp((x-y)\alpha_0)\exp(O(\epsilon^{1/2}|I|)O(\epsilon^{1/2}/|I|)) \\ &= (1+O(\epsilon))f(y)\exp((x-y)\alpha_0)\ . \end{aligned}$$

Noting that $(x-y)\alpha_0 = O(\epsilon^{1/2}|I|)O(1/|I|) = O(\epsilon^{1/2})$, this is

$$(1+O(\epsilon))f(y)(1+(x-y)\alpha_0+O((x-y)\alpha_0)^2) = (1+O(\epsilon))(f(y)+(x-y)\alpha_0+O(\epsilon)) = f(y)+(x-y)\alpha_0+O(\epsilon).$$

Therefore, taking $g_i(x) = f(y) + (x-y)\alpha_0$ suffices. This completes the proof. □

Next, we need to show that we can partition the domain of f into intervals I satisfying the above properties.

Proposition 15. *Let f be a log-concave distribution on either $\mathbb{R}$ or $\mathbb{Z}$. Then there exists a partition of $\mathbb{R}$ or $\mathbb{Z}$ into disjoint intervals $I_1, I_2, \ldots$ so that*

- *f satisfies the hypotheses of Lemma 14 on each I_i.*
- *For each m, there are only $O(m)$ values of i so that $f(I_i) > 2^{-m}$.*

Proof. Firstly, by splitting the domain of f into two pieces separated by the modal value, we may assume that f is monotonic. Henceforth, we assume that f is defined on $\mathbb{R}^+$ or $\mathbb{Z}^+$ and that f is both log-concave and monotonically decreasing.

We define the intervals $I_i = [a_i, b_i]$ inductively. We let $a_1 = 0$. Given a_i, we let b_i be the largest possible value so that f restricted to $[a_i, b_i]$ satisfies the hypotheses of Lemma 14. Given b_i we let a_{i+1} be either b_i (in the continuous case) or $b_i + 1$ (in the discrete case). Note that this causes the first condition to hold automatically.

We note that for each i, either $f(a_{i+1}) \leq f(a_i)/2$ or the logarithmic derivative of f at a_{i+1} is less than the logarithmic derivative at a_i by at least $1/(a_{i+1} - a_i)$. Note that in the latter case, since $f(a_{i+1}) > f(a_i)/2$, we have that the absolute value of the logarithmic derivative at $f(a_i)$ is at most $O(1/(a_{i+1} - a_i))$. Therefore, in this latter case, the absolute value of the logarithmic derivative of f at a_{i+1} is larger than the absolute value of the logarithmic derivative at a_i by at least a constant multiple.

Note that at the end of the first interval, we have that $f(a_2) = O(1/|I_1|)$ and that the absolute logarithmic derivative of f at a_2 is at least $\Omega(1/|I_1|)$. Note that each interval at least one of these increases by a constant multiple, therefore, there are only $O(m)$ many i so that both $f(a_i) > 2^{-m}/|I_1|$ and the absolute logarithmic derivative of f at a_i is less than $2^m/|I_1|$. We claim that if either of these fail to hold that the integral of f over I_i is $O(2^{-m})$.

If $f(a_i) < 2^{-m}/|I_1|$, then since the absolute logarithmic derivative of f on I_i is at least $\Omega(1/|I_1|)$, we have that the length of I_i is $O(|I_1|)$. Therefore, the mass of f on I_i is $O(2^{-m})$.

If on the other hand the absolute logarithmic derivative of f at a_i is at least $2^m/|I_1|$, since the value if f on I_i varies by at most a multiple of 2, we have that $|I_i| = O(|I_1|/2^m)$. Since f is decreasing, is has size $O(1/|I_1|)$ on I_i, and therefore, the integral of f of I_i is $O(2^{-m})$. This completes the proof of the second condition. □

We are now prepared to prove our Theorem 12:

Proof. We divide $\mathbb{R}$ or $\mathbb{Z}$ into intervals as described in Proposition 15. Call these intervals $I_1, I_2, \ldots$ sorted so that $f(I_i)$ is decreasing in i. Therefore, we have that $f(I_m) = O(2^{-\Omega(m)})$. In particular, there is a constant $c > 0$ so that $f(I_m) = O(2^{-cm})$.

For $m = 1, \ldots, 2\log(1/\epsilon)/c$, we use Lemma 14 to approximate f in I_m by a piecewise linear function g_m so that g_m has at most $O(\epsilon^{-1/2}2^{-cm/4})$ pieces and so that the L_1 distance between f and g_m on I_m is at most $f(I_m)O(\epsilon 2^{cm/2}) = O(\epsilon 2^{-cm/2})$. Let g be the piecewise linear function that is g_m on I_m for $m \leq c\log(1/\epsilon)/2$, and 0 elsewhere. g is piecewise linear on

$$\sum_{m=1}^{2\log(1/\epsilon)/c} O(\epsilon^{-1/2}2^{-cm/4}) = O(\epsilon^{-1/2})$$

intervals.

Furthermore the L_1 error between f and g on the I_m with $m \leq 2\log(1/\epsilon)/c$ is at most

$$\sum_{m=1}^{2\log(1/\epsilon)/c} O(\epsilon 2^{-cm/2}) = O(\epsilon).$$

The L_1 error from other intervals is at most

$$\sum_{m=2\log(1/\epsilon)/c}^{\infty} O(2^{-cm}) = O(\epsilon).$$

Therefore, $\|f - g\|_1 = O(\epsilon)$.

By replacing g by $\max(g, 0)$, we may ensure that it is positive (and at most double the number of pieces and decrease the distance from f). By scaling g, we may then ensure that it is a distribution. Finally by decreasing ϵ by an appropriate constant, we may ensure that $d_{\mathrm{TV}}(f, g) \leq \epsilon$. This completes the proof. □

A.4 Proof of Claim 13

Claim 13 *Let $g(x) = ax + b$ and $h(x) = c\exp(dx)$. Let $I = [x', x' + L]$ be an interval with $g(x) \geq 0$ and $0 \leq h(x) \leq O(\epsilon/L)$ for all $x \in I$. There is an algorithm which approximates $\|g - h\|_1^I$ to within an additive $O(\epsilon^2/\log(1/\epsilon))$ in time* $\mathrm{polylog}(1/\epsilon)$.

Proof. First, we claim that for any subinterval $I' \subseteq I$, we can compute $\|h\|_1^{I'}$ and $\|g\|_1^{I'}$ to within $O(\epsilon^2/\log(1/\epsilon))$ in time $\mathrm{polylog}(1/\epsilon)$. There are simple closed formulas for the integrals of these and for the sum of arithmetic and geometric series. The formula for the sum of a geometric series may have a cancellation issue when the denominator $1 - \exp(d)$ is small but note that when $|d| = O(\epsilon^2/\log(1/\epsilon))$ we can approximate the sum of $h(x)$ over I by its integral. These can all be computed in $\mathrm{polylog}(1/\epsilon)$ time.

Now it remains to approximate any crossing points, i.e., points x where $g(x) = h(x)$ for $x' \leq x \leq x' + L$ (which need not satisfy $x \in D$). If we find these with sufficient precision, then we can divide I and calculate $\|h\|_1^{I'}$ and $\|g\|_1^{I'}$ for each sub-interval I' to get the result. Note that $g(x)$ and $h(x)$ can have at most two crossing points, since there is at most one $x \in \mathbb{R}$ where the derivative of $h(x) - g(x)$ is 0. This can be calculated as $x^* = \ln(a/(cd))/d$, when $a/(cd) > 0$. If x^* lies in I we can subdivide and reduce to the case when there is a crossing point only if $g(x) - h(x)$ has different signs at the endpoints. In this case, if $g(x) = \Omega(\epsilon/L)$ at one endpoint, we can find a point at which $g(x) = \Theta(\epsilon/L)$, which is higher than our bound on $h(x)$, and we can divide there.

Thus, we can reduce to the case where there is exactly one crossing point in I and $h(x), g(x) \leq O(\epsilon/L)$. By performing $O(\log(1/\epsilon))$ bisections we can approximate this crossing point to within $O(\epsilon L/\log(1/\epsilon)$ (or $\max\{1, O(\epsilon^2 L/\log(1/\epsilon)\}$ when $D = \mathbb{Z}$). Then, we have that if J is the interval between the true crossing point and our estimate, then $\|g - h\|_1^J \leq \|g\|_1^J + \|h\|_1^J = O(L\epsilon/\log(1/\epsilon) \cdot \epsilon/L) = O(\epsilon^2/\log(1/\epsilon)$. Hence, if we divide here, each of the sub-intervals I' has $\|h - g\|_1^{I'} = |\|h\|_1^{I'} - \|g\|_1^{I'}| + O(\epsilon^2/\log(1/\epsilon))$.

Note that in all of the above cases, we only sub-divide I into at most $O(1)$ sub-intervals, and so it takes $\mathrm{polylog}(1/\epsilon)$ time to compute $\|g - h\|_1^I$ for the whole interval. □