A Polynomial Time Algorithm for Log-Concave Maximum Likelihood via Locally Exponential Families

Brian Axelrod
Department of Computer Science
Stanford University
baxelrod@cs.stanford.edu

Ilias Diakonikolas
Department of Computer Science
University of Southern California
ilias.diakonikolas@gmail.com

Anastasios Sidiropoulos
Department of Computer Science
University of Illinois at Chicago
sidiropo@gmail.com

Alistair Stewart
Web3 Foundation
stewart.al@gmail.com

Gregory Valiant
Department of Computer Science
Stanford University
gvaliant@stanford.edu

Abstract

We consider the problem of computing the maximum likelihood multivariate log-concave distribution for a set of points. Specifically, we present an algorithm which, given n points in $\mathbb{R}^d$ and an accuracy parameter $\epsilon > 0$, runs in time $\mathrm{poly}(n, d, 1/\epsilon)$, and returns a log-concave distribution which, with high probability, has the property that the likelihood of the n points under the returned distribution is at most an additive ϵ less than the maximum likelihood that could be achieved via any log-concave distribution. This is the first computationally efficient (polynomial time) algorithm for this fundamental and practically important task. Our algorithm rests on a novel connection with exponential families: the maximum likelihood log-concave distribution belongs to a class of structured distributions which, while not an exponential family, "locally" possesses key properties of exponential families. This connection then allows the problem of computing the log-concave maximum likelihood distribution to be formulated as a convex optimization problem, and solved via an approximate first-order method. Efficiently approximating the (sub) gradients of the objective function of this optimization problem is quite delicate, and is the main technical challenge in this work.

Authors are in alphabetical order.
Preprint. Under review.

1 Introduction

A distribution on $\mathbb{R}^d$ is log-concave if the logarithm of its probability density function is concave:

Definition 1 (Log-concave Density)**.** *A probability density function $f : \mathbb{R}^d \to \mathbb{R}_+$, $d \in \mathbb{Z}_+$, is called* log-concave *if there exists an upper semi-continuous concave function $\phi : \mathbb{R}^d \to [-\infty, \infty)$ such that $f(x) = e^{\phi(x)}$ for all $x \in \mathbb{R}^d$. We will denote by $\mathcal{F}_d$ the set of upper semi-continuous, log-concave densities with respect to the Lebesgue measure on $\mathbb{R}^d$.*

Log-concave densities form a broad nonparametric family encompassing a wide range of fundamental distributions, including the uniform, normal, exponential, logistic, extreme value, Laplace, Weibull, Gamma, Chi and Chi-Squared, and Beta distributions (see, e.g., [5]). Log-concave probability measures have been extensively investigated in several scientific disciplines, including economics, probability theory and statistics, computer science, and geometry (see, e.g., [60, 3, 54, 62, 59]). The problem of *density estimation* for log-concave distributions is of central importance in the area of non-parametric estimation (see, e.g., [62, 59, 58]) and has received significant attention during the past decade in statistics [22, 38, 36, 21, 49, 6, 45] and computer science [18, 19, 2, 15, 32, 33, 16].

One reason the class of log-concave distributions has attracted this attention, both from the theoretical and practical communities, is that log-concavity is a very natural "shape constraint," which places significantly fewer assumptions on the distribution in question than most parameterized classes of distributions. In extremely high-dimensional settings when the amount of available data is not too much larger than the dimensionality, fitting a multivariate Gaussian (or some other parametric distribution) to the data might be all one can hope to do. For many practical settings, however, the dimensionality is modest (e.g., 5-20) and the amount of data is significantly larger (e.g., hundreds of thousands or millions). In such settings, making a strong assumption on the parametric form of the underlying distribution is unnecessary—there is sufficient data to fit a significantly broader class of distributions, and log-concave distributions are one of the most natural such classes. From a practical perspective, even in the univariate setting, computing the log-concave density that maximizes the likelihood of the available data is a useful primitive, with the R implementation of Rufibach and Duembgen having over 39,000 downloads [39]. As we discuss below, the amount of data required to *learn* a log-concave distribution scales exponentially in the dimension, in contrast to most parametric classes of distributions. Nevertheless, for the many practical settings with modest dimensionality and large amounts of data, there *is* sufficient data to learn. The question now is computational: how does one compute the best-fit log-concave distribution? We focus on this algorithmic question:

Is there an efficient algorithm to compute the log-concave MLE for datapoints in $\mathbb{R}^d$?

Obtaining an understanding of the above algorithmic question is of interest for a number of reasons. First, the log-concave MLE is *the* prototypical statistical estimator for the class, is fully automatic (in contrast to kernel-based estimators, for example), and was very recently shown to achieve the minimax optimal sample complexity for the task of learning a log-concave distribution (up to logarithmic factors) [16, 23]. The log-concave MLE also has an intriguing geometry that is of interest from a purely theoretical standpoint [22, 57]. Developing an efficient algorithm for computing the log-concave MLE is of significant theoretical interest, and would also allow this general non-parametric class of distributions to be leveraged in the many practical settings where the dimensionality is moderate and the amount of data is large. We refer the reader to the recent survey [58] for a more thorough justification for why the log-concave MLE is a desirable distribution to compute.

1.1 Our Results and Techniques

The main result of this paper is the first efficient algorithm to compute the multivariate log-concave MLE. For concreteness, we formally define the log-concave MLE:

Definition 2 (Log-concave MLE)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$. The log-concave MLE, $\widehat{f}_n = \widehat{f}_n(X_1, \ldots, X_n)$, is the density $\widehat{f}_n \in \mathcal{F}_d$ which maximizes the log-likelihood $\ell(f) \stackrel{\text{def}}{=} \sum_{i=1}^n \ln(f(X_i))$ over $f \in \mathcal{F}_d$.*

As shown in [22], the log-concave MLE $\widehat{f}_n$ exists and is unique. Our main result is the first efficient algorithm to compute it up to any desired accuracy.

Theorem 1 (Main Result). *Fix $d \in \mathbb{Z}_+$ and $0 < \epsilon, \tau < 1$. There is an algorithm that, on input any set of points $X_1, \ldots, X_n$ in $\mathbb{R}^d$, and $0 < \epsilon$, $\tau < 1$, runs in $\mathrm{poly}(n, d, 1/\epsilon, \log(1/\tau))$ time and with probability at least $1 - \tau$ outputs a succinct description of a log-concave density $h^* \in \mathcal{F}_d$ such that $\ell(h^*) \geq \ell(\widehat{f}_n) - \epsilon$.*

Our algorithm does *not* require that the input points $X_1, \ldots, X_n$ in $\mathbb{R}^d$ are i.i.d. samples from a log-concave density, i.e., it efficiently solves the MLE optimization problem for any input set of points. We also note that the succinct output description of h^* allows for both efficient evaluation and efficient sampling. That is, we can efficiently approximate the density at a given point (within multiplicative accuracy), and efficient sample from a distribution that is close in total variation distance.

Recent work [16, 23] has shown that the log-concave MLE is minimax optimal, within a logarithmic factor, with respect to squared Hellinger distance. In particular, the minimax rate of convergence with n samples is $\tilde{\Theta}_d\big(n^{-2/(d+1)}\big)$. Combining this sample complexity bound with our Theorem 1, we obtain the first sample near-optimal and computationally efficient *proper* learning algorithm for multivariate log-concave densities. See Theorem 4 in Appendix B.

Technical Overview Here we provide an overview of our algorithmic approach. Notably, our algorithm does *not* require the assumption that the input points are samples from a log-concave distribution. It runs in $\mathrm{poly}(n, d, 1/\epsilon)$ on *any* set of input points and outputs an ϵ-accurate solution to the log-concve MLE. Our algorithm proceeds by convex optimization: We formulate the problem of computing the log-concave MLE of a set of n points in $\mathbb{R}^d$ as a convex optimization problem that we solve via an appropriate first-order method. It should be emphasized that one needs to overcome several non-trivial technical challenges to implement this plan.

The first difficulty lies in choosing the right (convex) formulation. Previous work [22] has considered a convex formulation of the problem that inherently fails, i.e., it cannot lead to a polynomial time algorithm. Given our convex formulation, a second difficulty arises: we do not have direct access to the (sub-)gradients of the objective function and the naive algorithm to compute a subgradient at a point takes exponential time. Hence, a second challenge is how to obtain an efficient algorithm for this task. One of our main contributions is a randomized polynomial time algorithm to approximately compute a subgradient of the objective function. Our algorithm for this task leverages structural results on log-concave densities established in [16] combined with classical algorithmic results on approximating the volume of convex bodies and uniformly sampling from convex sets [48, 53, 52].

We now proceed to explain our convex optimization formulation. Our starting point is a key structural property of the log-concave MLE, shown in [22]: The logarithm of the log-concave MLE $\ln \widehat{f}_n$, is a "tent" function, whose parameters are the values $y_1, \ldots, y_n$ of the log density at the n input points $x^{(1)}, \ldots, x^{(n)}$, and whose log-likelihoods correspond to polyhedra. Our conceptual contribution lies in observing that while tent distributions are not an exponential family, they "locally" retain many properties of exponential families (Definition 4). This high-level similarity can be leveraged to obtain a convex formulation of the log-concave MLE that is similar in spirit to the standard convex formulation of the exponential family MLE [61]. Specifically, we seek to maximize the log-likelihood of the probability density function obtained by normalizing the log-concave function whose logarithm is the convex hull of the log densities at the samples. This objective function is a concave function of the parameters, so we end up with a (non-differentiable) convex optimization problem. The crucial observation is that the subgradient of this objective at a given point y is given by an expectation under the current hypothesis density at y.

Given our convex formulation, we would like to use a first-order method to efficiently find an ϵ-approximate optimum. We note that the objective function is not differentiable everywhere, hence we need to work with subgradients. We show that the subgradient of the objective function is bounded in ℓ_2-norm at each point, i.e., the objective function is Lipschitz. Another important structural result (Lemma 2) allows us to essentially restrict the domain of our optimization problem to a compact convex set of appropriately bounded diameter $D = \mathrm{poly}(n, d)$. This is crucial for us, as the diameter bound implies an upper bound on the number of iterations of a first-order method. Given the above, we can in principle use a projected subgradient method to find an approximate optimum to our optimization problem, i.e., find a log-concave density whose log-likelihood is ϵ-optimal.

It remains to describe how we can efficiently compute a subgradient of our objective function. Note that the log density of our hypothesis can be considered as an unbounded convex polytope. The previous approach to calculate the subgradient in [22] relied on decomposing this polytope into faces and obtaining a closed form for the underlying integral over these faces (that gives their contribution to the subgradient). However, this convex polytope is given by n vertices in d dimensions, and therefore the number of its faces can be $n^{\Omega(d)}$. So, such an algorithm cannot run in polynomial time.

Instead, we note that we can use a linear program (see proof of Lemma 1) to evaluate a function proportional to the hypothesis density at a point in time polynomial in n and d. To use this oracle for the density in order to produce samples from the hypothesis density, we use Markov Chain Monte Carlo (MCMC) methods. In particular, we use MCMC to draw samples from the uniform distribution on super-level sets and estimate their volumes. With appropriate rejection sampling, we can use these samples to obtain samples from a distribution that is close to the hypothesis density. See Lemma 3. (We note that it does not suffice to simply run a standard log-concave density sampling technique such as hit-and-run [51]. These random walks require a hot start which is no easier than the sampling technique we propose.)

Since the subgradient of the objective can be expressed as an expectation over this density, we can use these samples to sample from a distribution whose expectation is close to a subgradient. We then use stochastic subgradient descent to find an approximately optimal solution to the convex optimization problem. The hypothesis density this method outputs has log-likelihood close to the maximum.

1.2 Related Work

There are two main strands of research in density estimation. The first one concerns the learnability of high-dimensional parametric distributions, e.g., mixtures of Gaussians. The sample complexity of learning parametric families is typically polynomial in the dimension and the challenge is to design computationally efficient algorithms. The second research strand — which is the focus of this paper — considers the problem of learning a probability distribution under various non-parametric assumptions on the shape of the underlying density, typically focusing on the univariate or small constant dimensional regime. There has been a long line of work in this vein within statistics since the 1950s, dating back to the pioneering work of [42] who analyzed the MLE of a univariate monotone density. Since then, shape constrained density estimation has been an active research area with a rich literature in mathematical statistics and, more recently, in computer science. The reader is referred to [10] for a summary of the early work and to [44] for a recent book on the subject.

The standard method used in statistics for density estimation problems of this form is the MLE. See [14, 55, 63, 46, 43, 11, 12, 40, 17, 7, 47, 38, 9, 41, 8, 50, 62, 21, 49, 6, 45, 16] for a partial list of works analyzing the MLE for various distribution families. During the past decade, there has been a body of algorithmic work on shape constrained density estimation in computer science with a focus on both sample and computational efficiency [24–26, 18–20, 1, 2, 29, 30, 27, 31, 33, 34]. The majority of this literature has studied the univariate (one-dimensional) setting which is by now fairly well-understood for a wide range of distributions. On the other hand, the *multivariate* setting is significantly more challenging and wide gaps in our understanding remain even for $d = 2$.

For the specific problem of learning a log-concave distribution, a line of work in statistics [22, 38, 36, 21, 6] has characterized the global consistency properties of the log-concave multivariate MLE. Regarding finite sample bounds, [49, 23] gave a sample complexity *lower bound* of $\Omega_d\left((1/\epsilon)^{(d+1)/2}\right)$ for $d \in \mathbb{Z}_+$ that holds for *any* estimator, and [49] gave a near-optimal sample complexity *upper bound* for the log-concave MLE for $d \leq 3$. [33] established the first finite sample complexity upper bound for learning multivariate log-concave densities under global loss functions. Their estimator (which is different than the MLE and seems hard to compute in multiple dimensions) learns log-concave densities on $\mathbb{R}^d$ within squared Hellinger loss ϵ with $\tilde{O}_d\left((1/\epsilon)^{(d+5)/2}\right)$ samples. [16] showed a sample complexity upper bound of $\tilde{O}_d\left((1/\epsilon)^{(d+3)/2}\right)$ for the multivariate log-concave MLE with respect to squared Hellinger loss, thus obtaining the first finite sample complexity upper bound for this estimator in dimension $d \geq 4$. Building on their techniques, this bound was subsequently improved in [23] to a near-minimax optimal bound of $\tilde{O}_d\left((1/\epsilon)^{(d+1)/2}\right)$. Alas, the computational complexity of the log-concave MLE has remained open in the multivariate case. Finally, we note that a recent work [28] obtained a non-proper estimator for multivariate log-concave densities with sample complexity $\tilde{O}_d((1/\epsilon)^{d+2})$ (i.e., at least quadratic in that of the MLE) and runtime $\tilde{O}_d((1/\epsilon)^{2d+2})$.

On the empirical side, recent work [56] proposed a non-convex optimization approach to the problem of computing the log-concave MLE, which seems to exhibit superior performance in practice in comparison to previous implementations (scaling to 6 or higher dimensions). Unfortunately, their method is of a heuristic nature, in the sense that there is no guarantee that their solution will converge to the log-concave MLE.

The present paper is a merger of two independent works [4, 35], proposing essentially the same algorithm to compute the log-concave MLE. Here we provide a unified presentation of these works with an arguably conceptually cleaner analysis.

2 Preliminaries

Notation. We denote by $X_1, \ldots, X_n \in \mathbb{R}^d$ the sequence of samples. We denote by $S_n = \mathrm{Conv}(\{X_i\}_{i=1}^n)$ the convex hull of $X_1, \ldots, X_n$, and by X the $d \times n$ matrix with columns vectors $X_1, \ldots, X_n$. We write $\mathbb{1}$ for the all-ones vector of the appropriate length. For a set $Y \subset Z$, $\mathbb{1}_Y$ denotes the indicator function for Y.

Tent Densities. We start by defining tent functions and tent densities:

Definition 3 (Tent Function)**.** *For $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and a set of points $X_1, \ldots, X_n$ in $\mathbb{R}^d$, we define the tent function $h_{X,y} : \mathbb{R}^d \to \mathbb{R}$ as follows:*

$$h_{X,y}(x) = \begin{cases} \max\{z \in \mathbb{R} \textit{ such that } (x, z) \in \mathrm{Conv}(\{(X_i, y_i)\}_{i=1}^n)\} & \textit{if } x \in S_n \\ -\infty & \textit{if } x \notin S_n \end{cases}$$

The points (X_i, y_i) are referred to as *tent poles*. (See Figure 1 for the graph of an example tent function.)

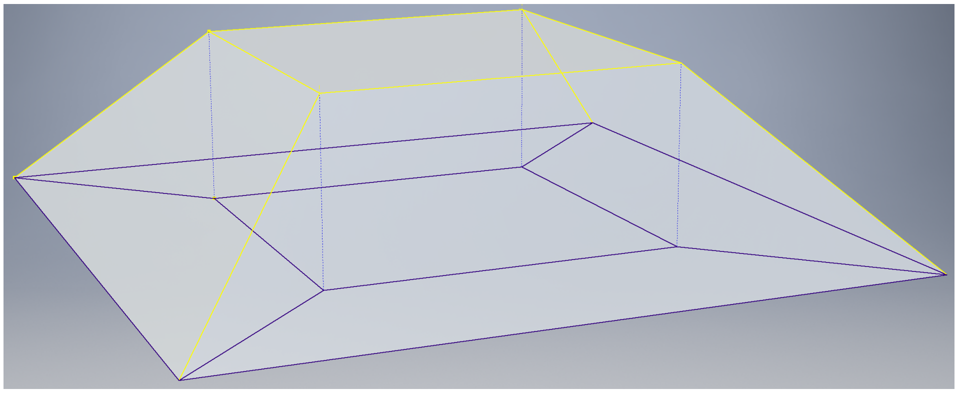

Figure 1: An example of a tent function and its corresponding regular subdivision. Notice that the regular subdivision is *not* a regular triangulation.

Let $p_{X,y}(x) = c \exp(h_{X,y}(x))$ with c chosen such that $p_{X,y}(x)$ integrates to one. We refer to $p_{X,y}$ as a *tent density* and the corresponding distribution as a *tent distribution*. Note that the support of a *tent distribution* must be within the convex hull of $X_1, \ldots, X_n$. For the remainder of the paper, we choose a scaling such that $\mathbb{1}^T y = 0$. This scaling is arbitrary, and has no significant effect on either the algorithm or its analysis.

Tent densities are notable because they contain solutions to the log-concave MLE [22]. The solution to the log-concave MLE over $X_1, \ldots, X_n$ is always a tent density, because tent densities with tent poles $X_1, \ldots, X_n$ are the minimal log-concave functions with log densities $y_1, \ldots, y_n$ at points $X_1, \ldots, X_n$.

The algorithm which we present can be thought of as an optimization over tent functions. In Section 3.1, we will show that tent distributions retain important properties of exponential families which will be useful to establish the correctness of our algorithm.

Regular Subdivisions. Given a tent function $h_{X,y}$ with $h_{X,y}(X_i) = y_i$, its associated *regular subdivision* $\Delta_{X,y}$ of X is a collection of subsets of $X_1, \ldots, X_n \in \mathbb{R}^d$ whose convex hulls are the regions of linearity of $h_{X,y}$. See Figure 1 for an illustration of a tent function and its regular subdivision. We refer to these polytopes of linearity as *cells*. We say that $\Delta_{X,y}$ is a *regular triangulation* of X if every cell is a $d-$dimensional simplex.

It is helpful to think of regular subdivisions in the following way: Consider the hyperplane H in $\mathbb{R}^{d+1}$ obtained by fixing the last coordinate. Consider the function $h_{X,y}$ as a polytope and project each face onto H. Each cell is a projection of a face, and together the cells partition the convex hull of $X_1, \ldots, X_n$. Observe that regular subdivisions may vary with y. Figure 2 provides one example of how changing the y vector changes the regular subdivision.

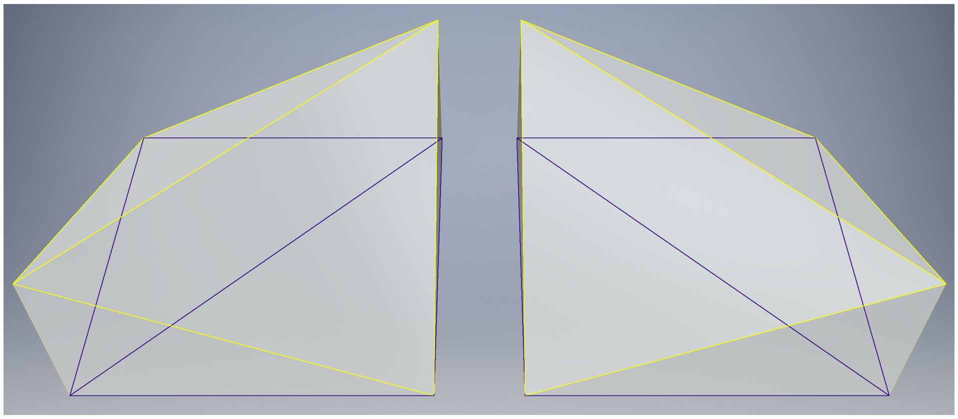

Figure 2: Changing the height of the tent poles can change the induced regular subdivision (shown in purple).

For a given regular triangulation Δ, the associated *consistent neighborhood* N_Δ is the set of all $y \in \mathbb{R}^n$, such that $\Delta_{X,y} = \Delta$. That is, consistent neighborhoods are the sets of parameters where the *regular triangulation* remains fixed. Note that these neighborhoods are open and their closures cover the whole space. See Figure 2 for an example of how crossing between consistent neighborhoods results in different subdivisions. We note that for fixed X, when y is chosen in general position, $\Delta_{X,y}$ is always a regular triangulation.

3 Locally Exponential Convex Programs

In this section, we lay the foundations for the algorithm presented in the next section. We present the "locally" exponential form of tent distributions and show it has the necessary properties to enable efficient computation of the log-concave MLE. Though they form a broader class of distributions, "locally" exponential distributions share some important properties of exponential families. Namely, the log-likelihood optimization is convex, and the expectation of the sufficient statistic is a subgradient. This will allows us to formulate a convex program which we will be able to solve in polynomial time.

Definition 4. *Let T be some function (possibly parametrized by y) and let $q_y = \exp(\langle T(x), y\rangle - A(y))$ be a family of probability densities parametrized by y with $A(y)$ acting to normalize the density so it integrates to 1. We say that the family $\{q_y\}$ is* locally-exponential *if the following hold: (1) $A(y)$ is convex in y, and (2) $\mathbb{E}_{x\sim q_y}[T(x)] \in \partial_y A(y)$.*

Note that the above definition differs from an exponential family in that for exponential families T may not depend on y.

In this section, we derive a sufficient statistic, the *polyhedral statistic*, that shows that tent distributions are in fact locally exponential. More formally, we show:

Lemma 1. *For tent poles $X_1, \ldots, X_n$, there exists a function $T_{X,y} : \mathbb{R}^d \to \mathbb{R}^n$ (the polyhedral statistic) such that $p_{X,y}(x) = \exp(\langle T_{X,y}(x), y\rangle - A(y))$ corresponds to the family of tent-distributions such that $\{p_{X,y}\}$ is locally exponential. Furthermore, $T_{X,y}$ is computable in time $\mathrm{poly}(n, d)$.*

Since we know that the log-concave MLE is a tent distribution, and all tent-distributions are log-concave, we know that the optimum of the maximum likelihood convex program in Equation (3.1) corresponds to the log-concave MLE.

$$\text{MLE of tents } = \max_y \sum_i h_{X,y}(X_i) - \log \int \exp h_{X,y}(x)dx = \max_y \quad \sum_i y_i - A(y) \tag{3.1}$$

Combining the above with the fact that the sufficient statistic allows us to compute the stochastic subgradient suggests that Algorithm 1 can compute the log-concave MLE in polynomial time.

Algorithm 1 ComputeLogConcaveMLE($X_1, \ldots, X_n, \epsilon$)

```
$y \leftarrow 0$; $c \leftarrow 8n^2 d \log(2nd)$; $m \leftarrow \frac{2c^2}{\epsilon^2}$

for $i \leftarrow 1, m$ do
    $\eta \leftarrow c/\sqrt{i}$
    $s \sim p_{X,y}$                                              ▷ Using Lemma 3
    $y \leftarrow y + \eta\left(\frac{1}{n}\mathbb{1} - T_{X,y}(s)\right)$    ▷ $T$ computed via Lemma 1. $\frac{1}{n}\mathbb{1}$ follows from Equation (3.1)
return $y$
```

3.1 The Polyhedral Sufficient Statistic

Consider a regular triangulation Δ corresponding to tent distribution parametrized by X and y. The *polyhedral statistic* is the function

$$T_{X,y}(x) : S_n \to [0,1]^n,$$

that expresses x as a convex combination of corners of the cell containing x in Δ_y. That is $x = XT_{X,y}(x)$ where $||T_y(x)||_1 = 1$ and $T_y(x)_i = 0$ if X_i is not a corner of the cell containing x. The polyhedral statistic gives an alternative way of writing tent functions and tent densities:

$$h_{X,y}(x) = \langle T_y(x), y\rangle \qquad p_{X,y}(x) = \exp(\langle T_y(x), y\rangle) \ .$$

If we restrict y such that $\sum_i y_i = 0$ and define $A(y) = \log \int_x p_{X,y}(x)dx$, then we can see that for every consistent neighborhood N_Δ we have an exponential family of the form

$$\exp\left(\langle T_y(x), \theta\rangle - A(y)\right) \quad \text{for } \theta \in N_\Delta \ . \tag{3.2}$$

While Equation (3.2) shows how subsets of tent distributions are exponential families, it also helps highlight why tent distributions are *not* an exponential family. The sufficient statistic depends on y through the regular subdivision. This means that tent distributions do not admit the same factorized form as exponential families since the sufficient statistic depends on y.

Note that we can use any ordering of $X_1, \ldots, X_n$ to define the polyhedral sufficient statistic everywhere including on regular subdivisions that are *not* regular triangulations. Also note that, assuming that no $X_i = X_j, i \neq j$, eliminating the last coordinate using the constraint $\mathbb{1}_n^T\theta = 0$ makes each exponential family minimal. In other words, over regions where the regular subdivision does not change (for example the consistent neighborhoods), tent distributions are minimal exponential families. This means the set of tent distribution can be seen as the finite union of a set of minimal exponential families. We refer to Equation (3.3) as the exponential form for tent densities:

$$p_{X,y}(x) = \exp\left(\langle T_{X,y}(x), y\rangle - A(y)\right) \mathbb{1}_{S_n}(x). \tag{3.3}$$

Both the polyhedral statistic and tent density queries can be computed in polynomial time with the packing linear program presented in Equation (3.4). For a point x, the value of y yields the log-density and the vector α corresponds to polyhedral statistic.

$$\max y \text{ s.t. } (x, y) = \sum_i \alpha_i(X_i, y_i), \sum_i \alpha_i = 1, \alpha_i \geq 0 \tag{3.4}$$

Note that the above combined with tent distributions being exponential families on consistent neighborhoods gives us that the properties from Lemma 1 hold true on consistent neighborhoods. We extend the proof to the full result below.

Proof. Convexity follows by iteratively applying known operations that preserve convexity of a function. Since a sum of convex functions is convex (see, e.g., page 79 of [13]), it suffices to show that the function $G(y) = \ln(\int \exp(h_{X,y}(x))\mathrm{d}x)$ is convex. Since $h_{X,y}(x)$ is a convex function of y, by definition, $\exp(h_{X,y}(x))$ is log-convex as a function of y. Since an integral of log-convex functions is log-convex (see, e.g., page 106 of [13]), it follows that $\int \exp(h_y(x))\mathrm{d}x$ is log-convex. Therefore, G is convex. We have therefore established that Equation (3.1) is convex, as desired.

$\mathbb{E}_{x\sim p_{X,y}}[T_{X,y}(x)] \in \partial_y A(y)$: Note that when y is in the interior of a consistent neighborhood, the polyhedral statistic LP has a unique solution and $\mathbb{E}_{x\sim p_{X,y}}[T(x)] \in \partial_y A(y)$ (by Fact 3). When y is on the boundary the solution set to the LP corresponds to the convex hull of solutions corresponding to each adjacent consistent neighborhood. This corresponds to the convex hull of limiting gradients from each neighboring consistent neighborhood and is the set of subgradients. □

4 Algorithm and Detailed Analysis

Recall that we compute the log-concave MLE via a first-order method on the optimization formulation presented in Equation (3.1). The complete method is presented in Algorithm 1. The algorithm is based on the stochastic gradient computation presented in the previous section, a standard application of the stochastic gradient method, and a sampler to be described later in this section.

4.1 Analysis

We now provide the main technical ingredients used to prove Theorem 1. Specifically, we bound the rate of convergence of the stochastic subgradient method, and we provide an efficient procedure for sampling from a log-concave distribution.

4.1.1 Stochastic Subgradient Method

Recall that algorithm 1 is simply applying the stochastic subgradient method to the following convex program with $\mathbb{1}^T y = 0$: $h(y) = \left\langle \frac{1}{n}\mathbb{1}_n, y \right\rangle - A(y)$.

We will require a slight strengthening of the following standard result, see, e.g., Theorem 3.4.11 in [37]:

Fact 1. *Let $\mathcal{C}$ be a compact convex set of diameter $\mathrm{diam}(\mathcal{C}) < \infty$. Suppose that the projections $\pi_{\mathcal{C}}$ are efficiently computable, and there exists $M < \infty$ such that for all $y \in \mathcal{C}$ we have that $\|g\|_2 \leq M$ for all stochastic subgradients. Then, after $K = \Omega\left(M \cdot \mathrm{diam}(\mathcal{C}) \log(1/\tau)/\epsilon^2\right)$ iterations of the projected stochastic subgradient method (for appropriate step sizes), with probability at least $1-\tau$, we have that $F\left(\bar{y}^{(K)}\right) - \min_{y\in\mathcal{C}} F(y) \leq \epsilon$, where $\bar{y}^{(K)} = (1/K)\sum_{i=1}^{K} y^{(i)}$.*

We note that Fact 1 assumes that, in each iteration, we can efficiently calculate an *unbiased* stochastic subgradient, i.e., a vector $g^{(k)}$ such that $\mathbb{E}[g^{(k)}] \in \partial_y F(y^{(k)})$. Unfortunately, this is not the case in our setting, because we can only *approximately* sample from log-concave densities. However, it is straightforward to verify that the conclusion of Fact 1 continues to hold if in each iteration we can compute a random vector $\widetilde{g}^{(k)}$ such that $\|\mathbb{E}[\widetilde{g}^{(k)}] - g^{(k)}\|_2 < \delta \overset{\mathrm{def}}{=} \epsilon/(2\mathrm{diam}(\mathcal{C}))$, for some $g^{(k)} \in \partial_y F(y^{(k)})$. This slight generalization is the basic algorithm we use in our setting.

We now return to the problem at hand. We note that since T represents the coefficients of a convex combination $||T(x)|| < 1$ for all x, bounding M by 1.

Lemma 2 will show that $\mathrm{diam}(\mathcal{C}) = O(2n^2 d\log(2nd))$. This implies that if we let $c = 8n^2 d\log(2nd)$ and run SGD for $\frac{2c^2}{\epsilon^2}$ iterations, the resulting point will have objective value within ϵ of the log-concave MLE.

Lemma 2. *Let $X_1, \ldots, X_n$ be a set of points in $\mathbb{R}^d$ and $\hat{f}$ be the corresponding log-concave MLE. Then, we have that $R_\infty \overset{\mathrm{def}}{=} \frac{\max_{i\in[n]} \hat{f}(X_i)}{\min_{i\in[n]} \hat{f}(X_i)} \leq (2nd)^{2nd}$. Converting to an ℓ_2 norm yields a bound on the diameter of $\mathcal{C}$: $\mathrm{diam}(\mathcal{C}) \leq 2n^2 d\log(2nd)$.*

Let us briefly sketch the proof of Lemma 2. The main idea is to show that if R_∞ were too high, then $\widehat{f}_n$ would have a lower likelihood than the uniform distribution on the convex hull of the samples S_n. More specifically, if the maximum value M of the density $\widehat{f}_n$ is large, then the volume of the set

$\{x \in \mathbb{R}^d : \widehat{f}_n(x) \geq M/R\}$ is small. For a fixed R, this set contains S_n and thus R_∞ must be large compared to $M\mathrm{vol}(S_n)$. Since $\widehat{f}_n$ has likelihood at least as high as the uniform distribution over S_n, R must be small compared to $M\mathrm{vol}(S_n)$. Combining these two observations yields a bound on R.

We now proceed with the complete proof.

Proof of Lemma 2. Let $V = \mathrm{vol}(S_n)$ be the volume of the convex hull of the sample points and $M = \max_x \widehat{f}_n(x)$ be the maximum pdf value of the MLE. By basic properties of the log-concave MLE (see, e.g., Theorem 2 of [22]), we have that $\widehat{f}_n(x) > 0$ for all $x \in S_n$ and $\widehat{f}_n(x) = 0$ for all $x \notin S_n$. Moreover, by the definition of a tent function, it follows that $\widehat{f}_n$ attains its global maximum value and its global non-zero positive value in one of the points X_i.

We can assume without loss of generality that $\widehat{f}_n$ is not the uniform distribution on S_n, since otherwise $R_\infty = 1$ and the lemma follows. Under this assumption, we have that $R_\infty > 1$ or $\ln R_\infty > 0$, which implies that $M > 1/V$. The following fact bounds the volume of upper level sets of any log-concave density:

Fact 2 (see, e.g., Lemma 8 in [16])**.** *Let $f \in \mathcal{F}_d$ with maximum value M_f. Then for all $w > 0$, we have* $\mathrm{vol}(L_f(M_f e^{-w})) \leq w^d/M_f$.

By Fact 2 applied to the MLE $\widehat{f}_n$, for $w = \ln R_\infty$, we get that $\mathrm{vol}(L_{\widehat{f}_n}(M/R_\infty)) \leq (\ln R_\infty)^d/M$. Since the pdf value of $\widehat{f}_n$ at any point in the convex hull S_n is at least that of the smallest sample point X_i, i.e., M/R_∞, it follows that S_n is contained in $L_{\widehat{f}_n}(M/R_\infty)$. Therefore,

$$V \leq (\ln R_\infty)^d/M \; . \tag{4.1}$$

On the other hand, the log-likelihood of $\widehat{f}_n$ is at least the log-likelihood of the uniform distribution U_{S_n} on S_n. Since at least one sample point X_i has pdf value $\widehat{f}_n(X_i) = M/R_\infty$ and the other $n-1$ sample points have pdf value $\widehat{f}_n(X_i) \leq M$, we have that

$$\ln(M/R_\infty) + (n-1)\ln M \geq \ell(\widehat{f}_n) \geq \ell(U_{S_n}) = n\ln(1/V) \; ,$$

or $n \ln M - \ln R_\infty \geq -n \ln V$, and therefore $\ln(MV) \geq (\ln R_\infty)/n$. This gives that

$$R_\infty^{1/n} \leq MV \; . \tag{4.2}$$

Combining (4.1) and (4.2) gives

$$R_\infty \leq (\ln R_\infty)^{nd} \; . \tag{4.3}$$

Since $\ln x < x$, $x \in \mathbb{R}$, setting $x = R_\infty^{\frac{1}{2nd}}$ gives that $\ln R_\infty < 2nd \cdot R_\infty^{\frac{1}{2nd}}$ or

$$(\ln R_\infty)^{nd} < (2nd)^{nd} \cdot R_\infty^{1/2} \; . \tag{4.4}$$

By (4.3) and (4.4) we deduce that $R_\infty \leq (2nd)^{nd} \cdot R_\infty^{1/2}$ or

$$R_\infty \leq (2nd)^{2nd} \; .$$

This completes the proof of Lemma 2. □

4.1.2 Efficient Sampling and Log-Partition Function Evaluation

In this section, we establish the following result, which gives an efficient algorithm for sampling from the log-concave distribution computed by our algorithm.

Lemma 3 (Efficient Sampling)**.** *There exist algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ satisfying the following: Let $\delta, \tau > 0$, let $X = X_1, \ldots, X_n \in \mathbb{R}^d$, let $y \in \mathbb{R}^n$ be a parameter of a tent-density in exponential form. Then the following conditions hold:*

(1) On input X, y, δ, and τ, algorithm $\mathcal{A}_1$ outputs a random vector $Z \in \mathbb{R}^d$, distributed according to some probability distribution with density $\widetilde{\phi}$, such that $\|\widetilde{\phi} - p_{X,y}\|_1 = O(\delta)$, in time $\mathrm{poly}(n, d, \|y\|_\infty, 1/\delta, \log(1/\tau))$, *with probability at least* $1 - \tau$.

(2) On input X, y, δ, and τ, algorithm $\mathcal{A}_2$ outputs some $\gamma' > 0$, such that $\gamma'/(1 + O(\delta)) \leq \int \exp(h_{X,y}(x))\mathrm{d}x \leq \gamma' \cdot (1 + O(\delta))$, in time $\mathrm{poly}(n, d, \|y\|_\infty, 1/\delta, \log(1/\tau))$, *with probability at least* $1 - \tau$.

The algorithm used in the proof of Lemma 3 is concerned mainly with part (1) in its statement. The pseudocode of this sampling procedure is given in Algorithm 2.

Using the notation from Algorithm 2, part (2) is easier to describe and we thus omit the pseudocode. We note that the following exposition of Algorithm 2 assumes that the input vector y is bounded. In the execution of Algorithm 1, $\|y\|_\infty$ is bounded linearly by the number of SGD iterates. Thus, the dependence of the sampling runtime on $\|y\|_\infty$ increases the overall runtime by at most a polynomial.

Algorithm 2 Algorithm to sample from $p_{X,y}$

```
procedure SAMPLE($X_1, \dots, X_n, y$)
Input: Sequence of points $X = \{X_i\}_{i=1}^n$ in $\mathbb{R}^d$, vector $y \in \mathbb{R}^n$, parameter $0 < \delta < 1$.
Output: A random vector $Z \in \mathbb{R}^d$ sampled from a probability distribution with density function $\widetilde{\phi}$, such that $\|\widetilde{\phi} - p_{X,y}\|_1 \le \delta$.
Step 1. Let $m = \lceil 1 + 2\|y\|_\infty \rceil$. Let $M = \max_{x \in \mathbb{R}^d} \exp(h_{X,y}(x))$. For any $i \in [m]$, let $L_i = \{x \in \mathbb{R}^d : \exp(h_{X,y}(x)) \ge M \cdot 2^{-i}\}$. For each $i \in [m]$ compute an estimate $\widetilde{\mathrm{vol}}(L_i)$ of $\mathrm{vol}(L_i)$ such that
    $\mathrm{vol}(L_i)/(1+\delta) \le \widetilde{\mathrm{vol}}(L_i) \le \mathrm{vol}(L_i)(1+\delta).$
Step 2. For $i \in [m]$, let $u_i$ be the uniform probability distribution on $L_i$, and let $\widetilde{u}_i$ be an efficiently samplable probability distribution such that
    $\|\widetilde{u}_i - u_i\|_1 \le \delta.$
Step 3. Let $\widetilde{c} = \sum_{i=1}^m 2^{-i}\widetilde{\mathrm{vol}}(L_i) + 2^{-m}\widetilde{\mathrm{vol}}(L_m)$.
Step 4. Let $\widehat{D}$ be the probability distribution on $[m]$ with
    $\Pr_{I \sim \widetilde{D}}[I = i] = \begin{cases} \widetilde{\mathrm{vol}}(L_i) \cdot 2^{-i}/\widetilde{c} & \text{if } i \in \{1, \dots, m-1\} \\ 2 \cdot \widetilde{\mathrm{vol}}(L_m) \cdot 2^{-m}/\widetilde{c} & \text{if } i = m \end{cases}$
Step 5. Sample $I \sim \widetilde{D}$.
Step 6. Sample $Z \sim \widetilde{u}_I$.
Step 7. For any $x \in \mathbb{R}^d$ let
    $G_{X,y}(x) = M \cdot 2^{-\lfloor \log_2(M/\exp(h_{X,y}(x))) \rfloor}$
Step 8. With probability $1 - \exp(h_{X,y}(Z))/G_{X,y}(Z)$ go to Step 5.
return $Z$.
```

We now present the proof of Lemma 3. The pseudocode of the sampling procedure is given in Algorithm 2. As stated in Section 4.1.2, Algorithm 2 uses subroutines for approximating the volume of a convex body given by a membership oracle, and a procedure for sampling from the uniform distribution supported on such a body. For these procedures we use the algorithms by [48], which are summarized in Theorems 2 and 3 respectively.

Theorem 2 ([48])**.** *The volume of a convex body K in $\mathbb{R}^d$, given by a membership oracle, can be approximated to within a relative error of δ with probability $1 - \tau$ using*

$$d^5 \cdot \mathrm{poly}(\log d, 1/\delta, \log(1/\tau))$$

oracle calls.

Theorem 3 ([48])**.** *Given a convex body $K \subset \mathbb{R}^d$, with oracle access, and some $\delta > 0$, we can generate a random point $u \in K$ that is distributed according to a distribution that is at most δ away from uniform in total variation distance, using*

$$d^5 \cdot \mathrm{poly}(\log d, 1/\delta)$$

oracle calls.

For all $X = X_1, \dots, X_n \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, and $x \in \mathbb{R}^d$, we use the notation $H_{X,y}(x) = \exp(h_{X,y}(x))$.

In order to use the algorithms in Theorems 2 and 3 in our setting, we need a membership oracle for the superlevel sets of the function $H_{X,y}$. Such an oracle can clearly be implemented using the LP (3.4). We also need a separation oracle for these superlevel sets, which is given in the following lemma:

Lemma 4 (Efficient Separation)**.** *There exists a* $\mathrm{poly}(n, d)$ *time separation oracle for the superlevel sets of* $H_{X,y}(x) = \exp(h_{X,y}(x))$.

Proof. To construct our separation oracle, we will rely on the covering LP that is dual to the packing LP used to evaluate a tent function. The dual to the packing LP looks for the hyperplane that is above all the (X_i, y_i) that has minimal y at x. More specifically, it is the following LP:

$$\begin{array}{ll} \text{minimize} & \beta_0 + \sum_{j=1}^{d} \beta_j x_j \\ \text{subject to} & \beta \in \mathbb{R}^{d+1}, \beta_0 + \sum_{j=1}^{d} \beta_j X_{i,j} \geq y_i, i \in [n] \ , \end{array} \tag{4.5}$$

where $X_{i,j}$ is the j-th coordinate of the vector X_i. Now suppose that we are interested in a super level set $L_{H_{X,y}}(l)$. We can use the above LP to compute $h_{X,y}(x)$ (and thus $H_{X,y}(x)$) and check if it is in the superlevel set. Suppose that it is not, then there will be a solution $\beta \in \mathbb{R}^{d+1}$ whose value is below $\ln l$, say $\ln l - \delta$ for some $\delta > 0$. Consider an x' in the halfspace $\beta_0 + \sum_{j=1}^{d} \beta_j x'_j \leq \ln l - \delta/2$ which has x in the interior. Since x does not appear in the objective, β is a feasible solution for the dual LP (4.5) with y, x', and so $h_y(x') \leq \ln l - \delta/2$, which implies that x' is not in the superlevel set. Therefore, $\beta_0 + \sum_j \beta_j x'_j = \ln l - \delta/2$ is a separating hyperplane for x and the level set. This completes the proof. □

Given all of the above ingredients, we are now ready to prove the main result of this section.

Proof of Lemma 3. We first prove part (1) of the assertion. To that end we analyze the sampling procedure described in Algorithm 2. Recall that $m = 1 + \lceil \|y\|_\infty \rceil$, and for any $i \in [m]$, we define the superlevel set

$$L_i = \{x \in \mathbb{R}^d : H_{X,y}(x) \geq M_{H_{X,y}} \cdot 2^{-i}\} \ .$$

For any $x \in \mathbb{R}^d$ recall that

$$G_{X,y}(x) = M_{H_{X,y}} 2^{-\lfloor \log_2(M_{H_{X,y}}/H_{X,y}(x)) \rfloor} \ .$$

For any $A \subseteq \mathbb{R}^d$, let $\chi_A : \mathbb{R}^d \to \{0, 1\}$ be the indicator function for A. It is immediate that for all $x \in \mathbb{R}^d$,

$$\begin{aligned} G_{X,y}(x) &= M_{H_{X,y}} \sum_{i=1}^{\infty} 2^{-i} \chi_{L_i}(x) \\ &= M_{H_{X,y}} \sum_{i=1}^{m} 2^{-i} \chi_{L_i}(x) + 2^{-m} \chi_{L_i}(m) \qquad \text{(since } H_{X,y}(x) = 0 \text{ for all } x \notin L_m) \end{aligned}$$

Let

$$c = \sum_{i=1}^{m} 2^{-i} \mathrm{vol}(L_i) + 2^{-m} \mathrm{vol}(L_m).$$

We have

$$\int_{\mathbb{R}^d} G_{X,y}(x) \mathrm{d}x = M_{H_{X,y}} \left(\sum_{i=1}^{m} 2^{-i} \mathrm{vol}(L_i) + 2^{-m} \mathrm{vol}(L_m) \right) = M_{H_{X,y}} c. \tag{4.6}$$

Let

$$\widehat{G}_{X,y}(x) = G_{X,y}(x)/(M_{H_{X,y}} c).$$

It follows by (4.6) that $\widehat{G}_{X,y}$ is a probability density function.

Let D be the probability distribution on $\{1, \ldots, m\}$, where

$$\Pr_{I \sim D}[I = i] = \left\{ \begin{array}{ll} \mathrm{vol}(L_i) \cdot 2^{-i}/c & \text{if } i \in \{1, \ldots, m-1\} \\ 2 \cdot \mathrm{vol}(L_m) \cdot 2^{-m}/c & \text{if } i = m \end{array} \right.$$

For any $i \in [m]$, let u_i be the uniform probability density function on L_i. To sample from $\widehat{G}_{X,y}$, we can first sample $I \sim D$, and then sample $Z \sim u_I$.

Recall that $\widehat{p}_{X,y} : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is the probability density function obtained by normalizing $H_{X,y}$; that is, for all $x \in \mathbb{R}^d$ we have

$$p_{X,y}(x) = H_{X,y}(x)/c',$$

where

$$c' = \int_{\mathbb{R}^d} H_{X,y}(x) \mathrm{d}x.$$

Consider the following random experiment: first sample $Z \sim \widehat{G}_y$, and then accept with probability $H_{X,y}(Z)/G_{X,y}(Z)$; conditioning on accepting, the resulting random variable $Z \in \mathbb{R}^d$ is distributed according to $\widehat{H}_{X,y}$. Note that since for all $x \in \mathbb{R}^d$, $G_{X,y}(x)/2 \leq H_{X,y}(x) \leq G_{X,y}(x)$, it follows that we always accept with probability at least $1/2$. Let α be the probability of accepting. Then

$$\alpha = \int_{\mathbb{R}^d} \widehat{G}_{X,y}(x)(H_{X,y}(x)/G_{X,y}(x)) \mathrm{d}x,$$

and thus

$$\begin{aligned} \int_{\mathbb{R}^d} H_{X,y}(x) \mathrm{d}x &= \int_{\mathbb{R}^d} G_{X,y}(x)(H_{X,y}(x)/G_{X,y}(x)) \mathrm{d}x \\ &= M_{H_{X,y}} c \int_{\mathbb{R}^d} \widehat{G}_{X,y}(x)(H_{X,y}(x)/G_{X,y}(x)) \mathrm{d}x \\ &= M_{H_{X,y}} c\alpha \,. \end{aligned} \tag{4.7}$$

By Theorem 2, for each $i \in [m]$, we compute an estimate, $\widetilde{\mathrm{vol}}(L_i)$, to $\mathrm{vol}(L_i)$, to within relative error δ, using $\mathrm{poly}(d, 1/\delta, \log(1/\tau'))$ oracle calls, with probability at least τ', where $\tau' = \tau/n^b$, for some constant $b > 0$ to be determined; moreover, by Theorem 3, we can efficiently sample, using $\mathrm{poly}(d, 1/\delta)$ oracle calls, from a probability distribution $\widetilde{u}_i$ with $\|u_i - \widetilde{u}_i\| \leq \delta$. Each of these oracle calls is a membership query in some superlevel set of $H_{X,y}$. This membership query can clearly be implemented if we can compute that value H_y at the desired query point x, which can be done in time $\mathrm{poly}(n, d)$ using LP (3.4). Thus, each oracle call takes time $\mathrm{poly}(n, d)$. Let

$$\widetilde{c} = \sum_{i=1}^{m} 2^{-i} \widetilde{\mathrm{vol}}(L_i) + 2^{-m} \widetilde{\mathrm{vol}}(L_m). \tag{4.8}$$

Since for all $i \in [m]$, $\mathrm{vol}(L_i)/(1+\delta) \leq \widetilde{\mathrm{vol}}(L_i) \leq \mathrm{vol}(L_i)(1+\delta)$, it is immediate that

$$c/(1+\delta) \leq \widetilde{c} \leq c(1+\delta) \,.$$

Recall that Algorithm 2 uses the probability distribution $\widetilde{D}$ on $[m]$, where

$$\mathrm{Pr}_{I \sim \widetilde{D}}[I = i] = \begin{cases} \widetilde{\mathrm{vol}}(L_i) \cdot 2^{-i}/\widetilde{c} & \text{if } i \in \{1, \ldots, m-1\} \\ 2 \cdot \widetilde{\mathrm{vol}}(L_m) \cdot 2^{-m}/\widetilde{c} & \text{if } i = m \end{cases}$$

Consider the following random experiment, which corresponds to Steps 5–6 of Algorithm 2: We first sample $I \sim \widetilde{D}$, and then we sample $Z \sim \widetilde{u}_I$. The resulting random vector $Z \in \mathbb{R}^d$ is distributed according to

$$\widetilde{G}_{X,y}(x) = \frac{1}{\widetilde{c}} \left(\sum_{i=1}^{m} 2^{-i} \widetilde{\mathrm{vol}}(L_i) \widetilde{u}_i(x) + 2^{-m} \widetilde{\mathrm{vol}}(L_m) \widetilde{u}_m(x) \right).$$

Next, consider the following random experiment, which captures Steps 5–8 of Algorithm 2: We sample $Z \sim \widetilde{G}_{X,y}$, and we accept with probability $H_{X,y}(Z)/G_{X,y}(Z)$. Let $\widetilde{H}_{X,y}$ be the resulting probability density function supported on $\mathbb{R}^d$ obtained by conditioning the above random experiment on accepting. Let $\widetilde{\alpha}$ be the acceptance probability. We have

$$\widetilde{\alpha} = \int_{\mathbb{R}^d} (H_{X,y}(x)/G_{X,y}(x)) \widetilde{G}(x) \mathrm{d}x.$$

We have

$$\begin{aligned}
\|D_i - \widetilde{D}_i\|_1 &= \sum_{i=1}^{m-1} 2^{-i} \cdot \left| \frac{\mathrm{vol}(L_i)}{c} - \frac{\widetilde{\mathrm{vol}}(L_i)}{\widetilde{c}} \right| + 2 \cdot 2^{-m} \cdot \left| \frac{\mathrm{vol}(L_m)}{c} - \frac{\widetilde{\mathrm{vol}}(L_m)}{\widetilde{c}} \right| \\
&= \sum_{i=1}^{m-1} 2^{-i} \cdot \left| \frac{\mathrm{vol}(L_i)}{c} - \frac{\mathrm{vol}(L_i)(1+\delta)}{c/(1+\delta)} \right| + 2 \cdot 2^{-m} \cdot \left| \frac{\mathrm{vol}(L_m)}{c} - \frac{\mathrm{vol}(L_m)(1+\delta)}{c/(1+\delta)} \right| \\
&\leq \sum_{i=1}^{m-1} 2^{-i} \frac{\mathrm{vol}(L_i)}{c} 3\delta + 2 \cdot 2^{m} \frac{\mathrm{vol}(L_m)}{c} 3\delta \\
&= 3\delta.
\end{aligned}$$

It follows that

$$\|\widehat{G}_{X,y} - \widetilde{G}_{X,y}\|_1 \leq \|D_i - \widetilde{D}_i\| + \max_i \|u_i - \widetilde{u}_i\|_1 \leq 3\delta + \delta \leq 4\delta,$$

and so

$$|\alpha - \widetilde{\alpha}| \leq \int_{\mathbb{R}^d} \frac{H_{X,y}(x)}{G_{X,y}(x)} \left| \widehat{G}_{X,y}(x) - \widetilde{G}_{X,y}(x) \right| \mathrm{d}x \leq \int_{\mathbb{R}^d} \left| \widehat{G}_{X,y}(x) - \widetilde{G}_{X,y}(x) \right| \mathrm{d}x \leq \|\widehat{G}_{X,y} - \widetilde{G}_{X,y}\|_1 \leq 4\delta.$$

Note that $p_{X,y}(x)/\alpha = \widehat{G}_{X,y}(x) \frac{H_{X,y}(x)}{G_{X,y}(x)}$ and $\widetilde{H}_{X,y}(x)/\widetilde{\alpha} = \widetilde{G}_{X,y}(x) \frac{H_{X,y}(x)}{G_{X,y}(x)}$ and so

$$\begin{aligned}
\|\widetilde{H}_{X,y} - p_{X,y}\|_1 &\leq \alpha \left(\|\widetilde{H}_{X,y}/\alpha - p_{X,y}/\alpha\|_1 + \|p_{X,y}/\widetilde{\alpha} - p_{X,y}/\alpha\|_1 \right) && \text{(by the triangle inequality)} \\
&= \alpha \left(\|\widetilde{H}_{X,y}/\alpha - p_{X,y}/\alpha\|_1 + |1/\widetilde{\alpha} - 1/\alpha| \right) \\
&= \alpha \int_{\mathbb{R}^d} (H_{X,y}(x)/G_{X,y}(x)) |\widetilde{G}_{X,y}(x) - p_{X,y}(x)| + |\alpha - \widetilde{\alpha}|/\widetilde{\alpha} \\
&\leq \|p_{X,y} - \widetilde{G}_{X,y}\|_1 + 2|\alpha - \widetilde{\alpha}| \\
&\leq 12\delta,
\end{aligned}$$

which establishes that the random vector Z that Algorithm 2 outputs is distributed according to a probability distribution $\widetilde{\phi}$ such that $\|\widetilde{\phi} - p_{X,y}\|_1 \leq 10\delta$, as required.

In order to bound the running time, we observe that all the steps of the algorithm can be implemented in time $\mathrm{poly}(n, d, \|y\|_\infty, 1/\delta, \log(1/\tau))$. The most expensive operation is approximating the volume of a superlevel set L_i and sampling for L_i, using Theorems 2 and 3. By the above discussion, using LP (3.4) and Lemma 4 each of these operations can be implemented in time $\mathrm{poly}(n, d, 1/\delta, \log(1/\tau))$. The algorithm succeeds if all the invocations of the algorithm of Theorem 2 are successful; by the union bound, this happens with probability at least $1 - \tau' \mathrm{poly}(n) = 1 - \tau' n^b \mathrm{poly}(n) \geq 1 - \tau$, where the inequality follows by choosing some sufficiently large constant $b > 0$. This establishes part (1) of the Lemma.

It remains to prove part (2). By (4.7) we have that $\gamma = M_{H_{X,y}} c\alpha$. Algorithm $\mathcal{A}_2$ proceeds as follows. First, we compute $M_{H_{X,y}}$. By the convexity of $h_{X,y}$, it follows that the maximum value of $M_{H_{X,y}}$ is attained on some sample point x_i; that is, $M_{H_{X,y}} = \max_{i \in [n]} H_{X,y}(x_i)$. Since we can evaluate H_y in polynomial time using LP (3.4), it follows that we can also compute $M_{H_{X,y}}$ in polynomial time. Next, we compute $\widetilde{c}$ using formula 4.8. Arguing as in part (1), this can be done in time $\mathrm{poly}(n, 1/\delta, \log(1/\tau))$, and with probability at least $1 - \tau/2$. Finally, we estimate $\widetilde{\alpha}$. The value of $\widetilde{\alpha}$ is precisely the acceptance probability of the random experiment described in Steps 5–8 of Algorithm 2. Since $\alpha \geq 1/2$, and $|\alpha - \widetilde{\alpha}| \leq 4\delta$, it follows that for $\delta < 1/16$, we can compute an estimate $\bar{\alpha}$ of the value of $\widetilde{\alpha}$, to within error $1 + O(\delta)$, with probability at least $1 - \tau/2$, after $O(\log(1/\tau))$ repetitions of the random experiment. The output of algorithm $\mathcal{A}_2$ is $\gamma' = M_{H_{X,y}} \widetilde{c} \bar{\alpha}$. We obtain that, with probability at least $1 - \tau$, we have

$$\gamma' = M_{H_{X,y}} \widetilde{c} \bar{\alpha} \leq M_{H_{X,y}} c(1+\delta)\alpha(1 + O(\delta)) = \gamma(1 + O(\delta)) \ ,$$

and

$$\gamma' = M_{H_{X,y}} \widetilde{c} \bar{\alpha} \geq M_{H_{X,y}} (c/(1+\delta))(\alpha/(1 + O(\delta))) = \gamma/(1 + O(\delta)) \ ,$$

which concludes the proof. □

5 Conclusions

In this paper, we gave a $\mathrm{poly}(n, d, 1/\epsilon)$ time algorithm to compute an ϵ-approximation of the log-concave MLE based on n points in $\mathbb{R}^d$. Ours is the first algorithm for this problem with a sub-exponential dependence in the dimension d. We hope that our approach may lead to more practical methods for computing the log-concave MLE in higher dimensions than was previously possible.

One concrete open question is whether there exists an algorithm for computing the log-concave MLE that runs in time $\mathrm{poly}(n, d, \log(1/\epsilon))$, instead of the $\mathrm{poly}(n, d, 1/\epsilon)$ that we achieve. Such an algorithm would likely be technically interesting as it may require going beyond the first-order methods we employ. More broadly, it seems worth investigating whether the MLE can be efficiently computed for other natural classes of non-parametric distributions. Alternately, one could hope that there is a simple set of natural properties such that, if a class of distributions satisfies those properties, then the MLE can be efficiently computed.

Acknowledgments: Ilias Diakonikolas was supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. Alistair Stewart was supported by a USC startup grant. Anastasios Sidiropoulos was supported by NSF Award CCF-1453472 (CAREER) and NSF grants CCF-1423230 and CCF-1815145. Brian Axelrod was supported by NSF Fellowship grant DGE-1656518, NSF award CCF-1763299, and a Finch family fellowship. Brian Axelrod and Gregory Valiant were supported by NSF awards CCF-1704417, and an ONR Young Investigator Award (N00014-18-1-2295).

References

[1] J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. Fast and near-optimal algorithms for approximating distributions by histograms. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015*, pages 249–263, 2015.

[2] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1278–1289, 2017. Available at https://arxiv.org/abs/1506.00671.

[3] 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.

[4] B. Axelrod and G. Valiant. An efficient algorithm for high-dimensional log-concave maximum likelihood. *CoRR*, abs/1811.03204, 2018. URL `http://arxiv.org/abs/1811.03204`.

[5] M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):pp. 445–469, 2005. ISSN 09382259. URL `http://www.jstor.org/stable/25055959`.

[6] F. Balabdaoui and C. R. Doss. Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli*, 24(2):1053–1071, 05 2018. doi: 10.3150/16-BEJ864.

[7] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics*, 35(6):pp. 2536–2564, 2007. ISSN 00905364.

[8] 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.

[9] F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):pp. 1299–1331, 2009. ISSN 00905364.

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

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

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

[13] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. URL `http://www.stanford.edu/~boyd/cvxbook.html`.

[14] H. D. Brunk. On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):pp. 437–454, 1958. ISSN 00034851.

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

[16] T. Carpenter, I. Diakonikolas, A. Sidiropoulos, and A. Stewart. Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In *Conference On Learning Theory, COLT 2018*, pages 1234–1262, 2018. URL `http://proceedings.mlr.press/v75/carpenter18a.html`.

[17] K.S. Chan and H. Tong. Testing for multimodality with dependent data. *Biometrika*, 91(1): 113–123, 2004.

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

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

[20] 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.

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

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

[23] Y. Dagan and G. Kur. The log-concave maximum likelihood estimator is optimal in high dimensions. *CoRR*, abs/1903.05315, 2019. URL `http://arxiv.org/abs/1903.05315`.

[24] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

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

[26] 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.

[27] C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, 2016.

[28] A. De, P. M. Long, and R. A. Servedio. Density estimation for shift-invariant multidimensional distributions. *CoRR*, abs/1811.03744, 2018. URL `http://arxiv.org/abs/1811.03744`.

[29] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 831–849, 2016. Full version available at https://arxiv.org/abs/1505.00662.

[30] I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 850–878, 2016. Full version available at https://arxiv.org/abs/1511.04066.

[31] I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of STOC'16*, 2016.

[32] I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient Robust Proper Learning of Log-concave Distributions. Arxiv report, 2016.

[33] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning multivariate log-concave distributions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 711–727, 2017. URL `http://proceedings.mlr.press/v65/diakonikolas17a.html`.

[34] I. Diakonikolas, J. Li, and L. Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. In *Conference On Learning Theory, COLT 2018*, pages 819–842, 2018.

[35] I. Diakonikolas, A. Sidiropoulos, and A. Stewart. A polynomial time algorithm for maximum likelihood estimation of multivariate log-concave densities. *CoRR*, abs/1812.05524, 2018. URL `http://arxiv.org/abs/1812.05524`.

[36] 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.

[37] J. C. Duchi. Introductory lectures on stochastic convex optimization. *Park City Mathematics Institute, Graduate Summer School Lectures*, 2016.

[38] 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.

[39] Lutz Dümbgen and Kaspar Rufibach. logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 39(6):1–28, 2011. URL `http://www.jstatsoft.org/v39/i06/`.

[40] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997.

[41] 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.

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

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

[44] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

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

[46] D. L. Hanson and G. Pledger. Consistency in concave regression. *The Annals of Statistics*, 4(6): pp. 1038–1050, 1976. ISSN 00905364.

[47] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[48] R. Kannan, L. Lovász, and M. Simonovits. Random walks and an o*(n5) volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

[49] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. *Ann. Statist.*, 44(6):2756–2779, 12 2016. Available at http://arxiv.org/abs/1404.2298.

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

[51] L. Lovász and S. Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 57–68. IEEE, 2006.

[52] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4): 985–1005, 2006.

[53] L. Lovász and S. Vempala. Simulated annealing in convex bodies and an o*(n4) volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

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

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

[56] F. Rathke and C. Schnörr. Fast multivariate log-concave density estimation. *CoRR*, abs/1805.07272, 2018. URL `https://arxiv.org/abs/1805.07272`.

[57] E. Robeva, B. Sturmfels, and C. Uhler. Geometry of Log-Concave Density Estimation. *ArXiv e-prints*, 2017. Available at https://arxiv.org/abs/1704.01910.

[58] R. J. Samworth. Recent progress in log-concave density estimation. *ArXiv e-prints*, 2017.

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

[60] 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. ISSN 1749-6632. doi: 10.1111/j.1749-6632.1989.tb16434.x. URL `http://dx.doi.org/10.1111/j.1749-6632.1989.tb16434.x`.

[61] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

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

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

Appendix

A Introduction To Exponential Families

In this section, we give a brief overview of exponential families that covers just the material necessary to appreciate the connection between exponential families and the log-concave maximum likelihood problem. We refer to [61] for a more complete treatment of exponential families.

An *exponential family* parameterized by $\theta \in \mathbb{R}^n$ with *sufficient statistic* $T(x)$, with carrier density h measurable and non-negative is a family of probability distributions of the form

$$p_\theta(x) = \exp(\langle T(x), \theta \rangle - A(\theta))h(x).$$

The *log-partition* function $A(\theta)$ is defined to normalize the integral of the density

$$A(\theta) = \log \int \exp(\langle T(x), \theta \rangle)h(x)dx.$$

It makes sense to restrict our attention to values of θ that give a valid probability density. The set of *Canonical Parameters* Θ is defined such that $\Theta = \{\theta \mid A(\theta) < \infty\}$.

We say that an exponential family is *minimal* if $\theta_1 \neq \theta_2$ implies $p_{\theta_1} \neq p_{\theta_2}$. This is necessary and sufficient for statistical identifiability.

One reason exponential families are well studied is that we have an algorithm that computes the maximum likelihood estimate via a convex program.

The maximum likelihood parameters $\theta^\star$ for a set of iid samples $X_1, \dots, X_n$ are:

$$\begin{aligned}\theta^\star &= \arg\max_\theta \prod_i p_\theta(X_i) = \arg\max_\theta \log \prod_i p_\theta(X_i) \\ &= \arg\max_\theta \sum_i \langle T(X_i), \theta \rangle - nA(\theta) - \sum_i \log h(x_i) = \arg\max_\theta \left\langle \frac{1}{n} \sum_i T(X_i), \theta \right\rangle - A(\theta)\end{aligned} \tag{A.1}$$

We refer to the optimization in Equation (A.1) as the *exponential maximum likelihood optimization*. The last equation helps highlight why $T(x)$ is referred to as the sufficient statistic. No other information is needed about the data points to compute both the likelihood and the maximum likelihood estimator.

One reason why exponential families are important is that the geometry of the optimization in Equation (A.1) has several nice properties.

Fact 3. *$A(\theta)$ of exponential families satisfies the following properties: (a) $A(\theta) \in C^\infty$ on Θ. (b) $A(\theta)$ is convex. (c) $\Delta A(\theta) = \mathbb{E}_{x \sim p(\theta)}[T(x)]$. (d) If the exponential family is minimal, $A(\theta)$ is strictly convex.*

Note that properties $(b), (c)$ are very similar to the definition of locally exponential families. The fact that tent distributions maintain some of these properties is exactly what enables the efficient algorithm in this paper.

A.1 Analogy Between Log-Concave MLE and Exponential Family MLE

In the case of exponential families, at each time step, the algorithm maintains a distribution (from the hypothesis class) and generates a single sample from this distribution. The sufficient statistic of the exponential family can then be used to compute a subgradient. The computational efficiency follows from the convexity of the log-likelihood function, and existence of efficient samplers and procedures for computing the sufficient statistic. We portray this stochastic gradient method for exponential families, together with the analogous form of our algorithm for log-concave distributions.

Exponential Family MLE

Optimization Formulation:

$$\max_y \langle \mu, y \rangle - \log \int \exp\left(\langle T(x), y \rangle\right) dx$$

Log-Concave MLE

Optimization Formulation:

$$\max_y \langle \mathbb{1}, y \rangle - \log \int \exp\left(\langle T_{X,y}(x), y \rangle\right) dx$$

Algorithm 3 Stochastic First Order Algorithm

```
function ComputeExpFamMLE($X_1, ...X_n$)
    $y \leftarrow y_{init}$
    for $i \leftarrow 1, m$ do
        $s \sim p(y)$                                ▷ sample
        $y \leftarrow y + \eta_i \left(\mu - T(s)\right)$        ▷ subgradient
    return $y$
```

Algorithm 4 Stochastic First Order Algorithm

```
function ComputeLogConMLE($X_1, ...X_n$)
    $y \leftarrow 0$
    for $i \leftarrow 1, m$ do
        $s \sim p(X, y)$
        $y \leftarrow y + \eta_i \left(\frac{1}{n}\mathbb{1}_n - T_{X,y}(s)\right)$
    return $y$
```

B Learning Multivariate Log-Concave Densities

In this section, we combine our Theorem 1 with known sample complexity bounds to give the first computationally efficient and sample near-optimal proper learner for multivariate log-concave densities.

Recall that the squared Hellinger loss between two distributions with densities $f, g : \mathbb{R}^d \to \mathbb{R}_+$ is $h^2(f, g) = (1/2) \cdot \int_{\mathbb{R}^d} (\sqrt{f(x)} - \sqrt{g(x)})^2 dx$. Combined with the known rate of convergence of the log-concave MLE with respect to the squared Hellinger loss [16, 23], Theorem 1 implies the following:

Theorem 4. *Fix $d \in \mathbb{Z}_+$ and $0 < \epsilon, \tau < 1$. Let $n = \tilde{\Omega}\left((d^2/\epsilon)\ln(1/\tau)\right)^{(d+1)/2}$. There is an algorithm that, given n iid samples from an unknown log-concave density $f_0 \in \mathcal{F}_d$, runs in $\mathrm{poly}(n)$ time and outputs a log-concave density $h^* \in \mathcal{F}_d$ such that with probability at least $1 - \tau$, we have that $h^2(h^*, f_0) \le \epsilon$.*

We note that Theorem 4 yields the first efficient proper learning algorithm for multivariate log-concave densities under a global loss function. The proof follows by combining Theorem 1 with the following lemma:

Lemma 5. *Let $n = \Omega_d\left((1/\epsilon)\ln(1/(\epsilon\tau))\right)^{(d+1)/2}$. Let $\widehat{f_n}$ be the MLE of n samples drawn from $f_0 \in \mathcal{F}_d$. Let h^* be a log-concave density that is supported on the convex hull of the samples with $\ell(h^*) \ge \ell(\widehat{f_n}) - \epsilon/16$. Then with probability at least $1 - \tau$ over the samples, $h^2(h^*, f_0) \le \epsilon$.*

We write f_n for the empirical density over the samples $X_1, \ldots, X_n$. The proof is a minor modification of the arguments in Section 3 of [16], using the following lemma [23]:

Lemma 6 (Theorem 4 from [23])**.** *For any $t > 0$, we have except with probability $2\exp(-2t^2)$ that for any convex set C,*

$$|f_n(C) - f_0(C)| \le O_d(n^{-2/(d+1)}) + t/\sqrt{n}\ .$$

Proof. The proof follows Section 3 of [16], except that we need to replace Lemma 10 of that paper with Lemma 6 and that we use h^* in place of $\widehat{f_n}$. We will sketch the proof here and highlight the modified components of that proof.

Lemma 10 of [16] had that, except with probability $\tau/3$, for all convex sets C, $|f_n(C) - f_0(C)| \le \epsilon/32\ln(100n^4/\tau^2)$. We take $n = \Omega_d\left((1/\epsilon)\ln(1/(\epsilon\tau))\right)^{(d+1)/2}$ and $t = \sqrt{\ln(6/\tau)/2}$ in Lemma 6 and so $n^{-2/(d+1)} = O_d(\epsilon/\ln(1/\epsilon\tau)) = O_d(\epsilon/\ln(n/\tau)$ and $t/\sqrt{n} \le \sqrt{\ln(\tau)}(\epsilon/(\ln(\epsilon\tau)))^{-(d+1)/2} \le O(\epsilon/\ln(n/\tau))$ for $d \ge 2$. With a sufficiently large constant in the Ω_d, we obtain that $|f_n(C) - f_0(C)| \le \epsilon/K\ln(100n^4/\tau^2)$ except with probability $\tau/3$ where K is a constant large enough to make the subsequent proof work.

This gives the improved sample complexity. We now need to argue that replacing $\widehat{f_n}$ with h^* does not affect the proof.

Corollary 9 of [16] gave that except with probability $\tau/10$, all samples lie in a set S, which is the set where $f_0(x) \geq p_{\min}$ for $p_{\min} = M_{f_0}/(n^4 100/\tau^2)$, where we use the notation M_f for the maximum value of a density f.. When this holds both $\widehat{f}_n$ and h^* are supported on S. Examination of the proof of Lemma 18 from [16] shows that we can relax the inequality $\ell(f) \leq \ell(f_0)$ to $\ell(f) \leq \ell(f_0) - \epsilon/16$ for any f with maximum value M_f has $M_f = \Omega(\ln(100n^4/\tau^2))$. In partuclar, since $\ell(h^*) \geq \ell(\widehat{f}_n)) - \epsilon/16 \geq \ell(f_0) - \epsilon/16$, we have $M_{h^*} = O(\ln(100n^4/\tau^2))$.

Then we define $g_h(x)$ supported on S as the normalisation of $\max\{p_{\min}, h^*(x)\}$ for $x \in S$. The proof of Lemma 17 in [16] required only that $\widehat{f}_n$ is supported on S and so we can obtain the same result for g_h and $h^*(x)\}$ i.e. that $g_h(x) = \alpha \max\{p_{\min}, h^*(x)\}$ for $1 - \epsilon/32 \leq \alpha \leq 1$ and that the total variation distance is small,

$$d_{\mathrm{TV}}(g_h, h^*) \leq 3\epsilon/64 \; . \tag{B.1}$$

Note that since the superlevel sets of $\ln \max\{p_{\min}, h^*(x)\}$ are convex, we can use our application of Lemma 6 to bound the error in it's expectation as

$$\begin{aligned} |\mathbb{E}_{X\sim f_0}[1_S \ln(\max\{h^*(X), p_{\min}\})] - \mathbb{E}_{X\sim f_n}[1_S \ln(\max\{h^*(X), p_{\min}\})]| &\leq (M_{h^*} - p_{\min})\epsilon/K \ln(100n^4/\tau^2) \\ &\leq \epsilon/4 \end{aligned} \tag{B.2}$$

for large enough K.

We now follow the proof of Lemma 19 in [16]. We have that

$$\begin{aligned} \mathbb{E}_{X\sim f_0}[\ln g_h(X)] &= \mathbb{E}_{X\sim f_0}[1_S(x) \ln(\alpha \max\{p_{\min}, h^*(x)\})] \\ &\geq \mathbb{E}_{X\sim f_0}[1_S(x) \ln \max\{p_{\min}, h^*(x)\}] - \epsilon/16 && \text{(since } a > 1 - \epsilon/32\text{)} \\ &\geq \mathbb{E}_{X\sim f_0}[1_S \ln(\max\{h^*(X), p_{\min}\})] - \epsilon/16 \\ &\geq \mathbb{E}_{X\sim f_n}[1_S \ln(\max\{h^*(X), p_{\min}\})] - 3\epsilon/16 && \text{by (B.2)} \\ &\geq \frac{1}{n}\sum_i \ln h^*(X_i) - 3\epsilon/16 \\ &\geq \frac{1}{n}\sum_i \ln \widehat{f}_n(X_i) - \epsilon/4 \\ &\geq \frac{1}{n}\sum_i \ln f_0(X_i) - \epsilon/4 \\ &\geq \mathbb{E}_{X\sim f_0}[\ln f_0(X)] - 3\epsilon/8. && \text{(using Lemma 14 of [16])} \end{aligned} \tag{B.3}$$

Thus, we obtain that

$$\mathrm{KL}(f_0 || g) = \mathbb{E}_{X\sim f_0}[\ln f_0(X)] - \mathbb{E}_{X\sim f_0}[\ln g_h(X)] \leq 3\epsilon/8. \tag{B.4}$$

For the next derivation, we use that the Hellinger distance is related to the total variation distance and the Kullback-Leibler divergence in the following way: For probability functions $k_1, k_2 : \mathbb{R}^d \to \mathbb{R}$, we have that $h^2(k_1, k_2) \leq d_{\mathrm{TV}}(k_1, k_2)$ and $h^2(k_1, k_2) \leq \mathrm{KL}(k_1 || k_2)$. Therefore, we have that

$$\begin{aligned} h(f_0, h^*) &\leq h(f_0, g_h) + h(g_h, h^*) \\ &\leq \mathrm{KL}(f_0 || g_h)^{1/2} + d_{\mathrm{TV}}(g_h, h^*)^{1/2} \\ &= (3\epsilon/8)^{1/2} + (3\epsilon/64)^{1/2} && \text{(by (B.4) and (B.1))} \\ &\leq \epsilon^{1/2} \; , \end{aligned}$$

concluding the proof. □