Proceedings of Machine Learning Research vol 75:1–29, 2018 31st Annual Conference on Learning Theory

Near-Optimal Sample Complexity Bounds for Maximum Likelihood Estimation of Multivariate Log-concave Densities

Timothy Carpenter CARPENTER.454@OSU.EDU
The Ohio State University

Ilias Diakonikolas DIAKONIK@USC.EDU
University of Southern California

Anastasios Sidiropoulos SIDIROPO@GMAIL.COM
University of Illinois at Chicago

Alistair Stewart ALISTAIS@USC.EDU
University of Southern California

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the problem of learning multivariate log-concave densities with respect to a global loss function. We obtain the first upper bound on the sample complexity of the maximum likelihood estimator (MLE) for a log-concave density on $\mathbb{R}^d$, for all $d \geq 4$. Prior to this work, no finite sample upper bound was known for this estimator in more than 3 dimensions.

In more detail, we prove that for any $d \geq 4$ and $\epsilon > 0$, given $\tilde{O}_d((1/\epsilon)^{(d+3)/2})$ samples drawn from an unknown log-concave density f_0 on $\mathbb{R}^d$, the MLE outputs a hypothesis h that with high probability is ϵ-close to f_0, in squared Hellinger loss. For any $d \geq 2$, a sample complexity lower bound of $\Omega_d((1/\epsilon)^{(d+1)/2})$ was previously known for any learning algorithm that achieves this guarantee. We thus establish that the sample complexity of the log-concave MLE is near-optimal for $d \geq 4$, up to an $\tilde{O}(1/\epsilon)$ factor.

1. Introduction

1.1. Background

The general task of estimating a probability distribution under certain qualitative assumptions about the *shape* of its probability density function has a long history in statistics, dating back to the pioneering work of Grenander (1956) who analyzed the maximum likelihood estimator of a univariate monotone density. Since then, shape constrained density estimation has been a very active research area with a rich literature in mathematical statistics and, more recently, in computer science. A wide range of shape constraints have been studied, including unimodality, convexity and concavity, k-modality, log-concavity, and k-monotonicity. The reader is referred to Barlow et al. (1972) for a summary of the early work and to Groeneboom and Jongbloed (2014) for a recent book on the subject. (See Section 1.3 for a succinct summary of prior work.) The majority of the literature has studied the univariate (one-dimensional) setting, which is by now fairly well-understood for a range of distributions. On the other hand, the multivariate setting and specifically the regime of *fixed* dimension is significantly more challenging and poorly understood for many natural distribution families.

In this work, we focus on the family of *multivariate* log-concave distributions. A distribution on $\mathbb{R}^d$ is log-concave if the logarithm of its probability density function is concave (see Definition 1). Log-concave distributions constitute a rich non-parametric family encompassing a range of fundamental distributions, including uniform, normal, exponential, logistic, extreme value, Laplace, Weibull, Gamma, Chi and Chi-Squared, and Beta distributions (see, e.g., Bagnoli and Bergstrom (2005)). Due to their fundamental nature and appealing properties, log-concave distributions have been studied in a range of fields including economics An (1995), probability theory Saumard and Wellner (2014), computer science Lovász and Vempala (2007), and geometry Stanley (1989).

The problem of *density estimation* for log-concave distributions is of central importance in the area of non-parametric shape constrained estimation Walther (2009); Saumard and Wellner (2014); Samworth (2017) and has received significant attention during the past decade in statistics Cule et al. (2010); Dumbgen and Rufibach (2009); Doss and Wellner (2016); Chen and Samworth (2013); Kim and Samworth (2016); Balabdaoui and Doss (2018); Han and Wellner (2016) and theoretical computer science Chan et al. (2013, 2014a); Acharya et al. (2017); Canonne et al. (2016); Diakonikolas et al. (2016d, 2017).

1.2. Our Results and Comparison to Prior Work

In this work, we analyze the global convergence rate of the maximum likelihood estimator (MLE) of a multivariate log-concave density. Formally, we study the following fundamental question:

How many samples are information-theoretically sufficient so that the MLE of an arbitrary log-concave density on $\mathbb{R}^d$ learns the underlying density, within squared Hellinger loss ϵ?

Perhaps surprisingly, despite significant effort within the statistics community on analyzing the log-concave MLE, our understanding of its finite sample performance in constant dimension has remained poor. The only result prior to this work that addressed the sample complexity of the MLE in more than one dimensions is by Kim and Samworth (2016). Specifically, Kim and Samworth (2016) obtained the following results:

(1) a sample complexity *lower bound* of $\Omega_d\left((1/\epsilon)^{(d+1)/2}\right)$ that applies to *any* estimator for all $d \geq 2$, and

(2) a sample complexity *upper bound* for the log-concave MLE, that is near-optimal (within logarithmic factors) for $d \leq 3$.

Prior to our work, no finite sample upper bound was known for the log-concave MLE even for $d = 4$.

In recent related work, Diakonikolas et al. (2017) established a finite sample complexity upper bound for learning multivariate log-concave densities under global loss functions. Specifically, the estimator analyzed in Diakonikolas et al. (2017) uses $\tilde{O}_d\left((1/\epsilon)^{(d+5)/2}\right)$[1] samples and learns a log-concave density on $\mathbb{R}^d$ within squared Hellinger loss ϵ, with high probability. We remark that the upper bound of Diakonikolas et al. (2017) was obtained by analyzing an estimator that is *substantially different* than the log-concave MLE. Moreover, the analysis in Diakonikolas et al. (2017) has no implications on the performance of the MLE. Interestingly, some of the technical tools employed in Diakonikolas et al. (2017) will be useful in our current setting.

1. The $\tilde{O}(\cdot)$ notation hides logarithmic factors in its argument.

Due to the fundamental nature of the MLE, understanding its performance merits investigation in its own right. In particular, the log-concave MLE has an intriguing geometric structure that is a topic of current investigation Cule et al. (2010); Robeva et al. (2017). The output of the log-concave MLE satisfies several desirable properties that may not be automatically satisfied by surrogate estimators. These include the log-concavity of the hypothesis, the paradigm of log-concave projections and their continuity in Wasserstein distance, affine equivariance, one-dimensional characterization, and adaptation (see, e.g., Samworth (2017)). An additional motivation comes from a recent conjecture (see, e.g., Wellner (2015)) that for 4-dimensional log-concave densities the MLE may have sub-optimal sample complexity. These facts provide strong motivation for characterizing the sample complexity of the log-concave MLE in any dimension.

To formally state our results, we will need some terminology. The *squared Hellinger distance* between two density functions $f, g : \mathbb{R}^d \to \mathbb{R}_+$ is defined as $h^2(f,g) = (1/2) \cdot \int_{\mathbb{R}^d} (\sqrt{f(x)} - \sqrt{g(x)})^2 dx$.

We now define our two main objects of study:

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 $\mathbb{R}^d$.*

Definition 2 (Log-concave MLE) *Let $f_0 \in \mathcal{F}_d$ and $X_1, \ldots, X_n$ be iid samples from f_0. The* maximum likelihood estimator*, $\hat{f}_n$, is the density $\hat{f}_n \in \mathcal{F}_d$ which maximizes $\frac{1}{n} \sum_{i=1}^n \log(f(X_i))$ over all $f \in \mathcal{F}_d$.*

We can now state our main result:

Theorem 3 (Main Result) *Fix $d \in \mathbb{Z}_+$ and $\epsilon \in (0,1)$. Let $n = \tilde{\Omega}_d\left((1/\epsilon)^{(d+3)/2}\right)$. For any $f_0 \in \mathcal{F}_d$, with probability at least $9/10$ over the n samples from f_0, we have that $h^2(\hat{f}_n, f_0) \leq \epsilon$.*

See Theorem 7 for a more detailed statement. The aforementioned lower bound of Kim and Samworth (2016) implies that our upper bound is tight up to an $\tilde{O}_d(\epsilon^{-1})$ multiplicative factor.

1.3. Related Work

Shape constrained density estimation is a vibrant research field within mathematical statistics. Statistical research in this area started in the 1950s and has seen a recent surge of research activity, in part due to the ubiquity of structured distributions in various domains. The standard method used in statistics to address density estimation problems of this form is the MLE. See Brunk (1958); Rao (1969); Wegman (1970); Hanson and Pledger (1976); Groeneboom (1985); Birgé (1987a,b); Fougères (1997); Chan and Tong (2004); Balabdaoui and Wellner (2007); Jankowski and Wellner (2009); Dumbgen and Rufibach (2009); Balabdaoui et al. (2009); Gao and Wellner (2009); Balabdaoui and Wellner (2010); Koenker and Mizera (2010); Walther (2009); Chen and Samworth (2013); Kim and Samworth (2016); Balabdaoui and Doss (2018); Han and Wellner (2016) for a partial list of works analyzing the MLE for various distribution families. During the past decade, there has been a large body of work on shape constrained density estimation in computer science with a focus on both sample and computational efficiency Daskalakis et al. (2012a,b, 2013); Chan

et al. (2013, 2014a,b); Acharya et al. (2015, 2017); Diakonikolas et al. (2016a,b); Daskalakis et al. (2016); Diakonikolas et al. (2016c); Valiant and Valiant (2016); Diakonikolas et al. (2017).

Density estimation of log-concave densities has been extensively investigated. The univariate case is by now well understood Devroye and Lugosi (2001); Chan et al. (2014a); Acharya et al. (2017); Kim and Samworth (2016); Han and Wellner (2016). For example, it is known Kim and Samworth (2016); Han and Wellner (2016) that $\Theta(\epsilon^{-5/4})$ samples are necessary and sufficient to learn an arbitrary log-concave density over $\mathbb{R}$ within squared Hellinger loss ϵ. Moreover, the MLE is sample-efficient Kim and Samworth (2016); Han and Wellner (2016) and attains certain adaptivity properties Kim et al. (2016). A recent line of work in computer science Chan et al. (2013, 2014a); Acharya et al. (2017); Canonne et al. (2016); Diakonikolas et al. (2016d) gave efficient algorithms for log-concave density estimation under the total variation distance.

Density estimation of multivariate log-concave densities has been systematically studied as well. A line of work Cule et al. (2010); Dumbgen and Rufibach (2009); Doss and Wellner (2016); Chen and Samworth (2013); Balabdaoui and Doss (2018) has obtained a complete understanding of the global consistency properties of the MLE for any dimension. However, both the rate of convergence of the MLE and the minimax rate of convergence remain unknown for $d \geq 4$. For $d \leq 3$, Kim and Samworth (2016) show that the MLE is sample near-optimal (within logarithmic factors) under the squared Hellinger distance. Kim and Samworth (2016) also prove bracketing entropy lower bounds suggesting that the MLE may be sub-optimal for $d > 3$ (also see Wellner (2015)).

1.4. Technical Overview

Here we provide a brief overview of our proof in tandem with a comparison to prior work. We start by noting that the previously known sample complexity upper bound of the log-concave MLE for $d \leq 3$ Kim and Samworth (2016) was obtained by bounding from above the bracketing entropy of the class. As we explain below, our argument is more direct making essential use of the VC inequality (Theorem 4), a classical result from empirical process theory. In contrast to prior work on log-concave density estimation Kim and Samworth (2016); Diakonikolas et al. (2017) which relied on approximations to (log)-concave *functions*, we start by considering approximations to *convex sets*. Let f_0 be the target log-concave density. We show (Lemma 10) that given sufficiently many samples from f_0, with high probability, for any convex set C the empirical mass of C and the probability mass of C under f_0 are close to each other. We then leverage this structural lemma to analyze the error in the log-likelihood of log-concave densities, using the fact that the superlevel sets of a log-concave density are convex.

We remark that our aforementioned structural result (Lemma 10) crucially requires the assumption of the log-concavity of f_0. Naively, one may think that this lemma follows directly from the VC inequality. Recall however that the VC-dimension of the family of convex sets is infinite, even in the plane. For example, for the uniform distribution over the unit circle, a similar result does *not* hold for *any* finite number of samples (the intersection of the convex hull of any subset S of the unit circle with the unit circle is S itself, so we would need uniform convergence on all subsets of the unit circle), and so we need to use the fact that f_0 is log-concave. To prove our lemma, we consider judicious approximations of the convex set C with convex polytopes using known results from convex geometry. In more detail, we consider approximations to the convex set C on the inside and outside with close probabilities under f_0 to the convex set from a family with a bounded VC-dimension.

For any log-concave density f, the probabilities of any superlevel set are close under the empirical distribution and f_0. If $\log f$ were bounded, then that would mean that the empirical log-likelihood of f and the log-likelihood of f under f_0 were close. Unfortunately, for any density f, $\log f$ is unbounded from below. To deal with this issue, we instead consider $\log(\max(f, p_{\min}))$, for some carefully chosen probability value $p_{\min}$ such that we could ignore the contribution of the density below $p_{\min}$ if f is close to f_0. If we can bound the range of $\log(\max(f, p_{\min}))$, we can show that its expectation under f_0 and its empirical version are close to each other (see Lemma 13). To bound the range, we show that if the maximum value of f is much larger than the maximum of f_0, then f has small probability mass outside a set A of small volume; since A has small volume, we see many samples outside it, and so the empirical log-likelihood of f is smaller than the empirical log-likelihood of f_0. Using this fact, we can show that for the MLE $\hat{f}_n$ the expectation of $\log(\max(\hat{f}_n, p_{\min}))$ is large under f_0 and then that $\hat{f}_n$ is close in Hellinger distance to f_0.

1.5. Organization

After setting up the required preliminaries in Section 2, in Section 3 we present the proof of our main result, modulo the proof of our main lemma (Lemma 10). In Section 4, we give a slightly weaker version of Lemma 10 that has a significantly simpler proof. In Section A, we present the proof of Lemma 10. Finally, we conclude with a few open problems in Section 5.

2. Preliminaries

Notation and Definitions. For $m \in \mathbb{Z}_+$, we denote $[m] \stackrel{\text{def}}{=} \{1, \ldots, m\}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lebesgue measurable function. We will use $f(A)$ to denote $\int_A f(x)dx$. A Lebesgue measurable function $f : \mathbb{R}^d \to \mathbb{R}$ is a probability density function (pdf) if $f(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)dx = 1$. Let $f, g : \mathbb{R}^d \to \mathbb{R}_+$ be probability density functions. The *squared Hellinger distance* between f, g is defined as $H^2(f, g) = \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$. The *total variation distance* between f, g is defined as $d_{\mathrm{TV}}(f, g) = \sup_S |f(S) - g(S)|$, where the supremum is over all Lebesgue measurable subsets of the domain. We have that $d_{\mathrm{TV}}(f, g) = (1/2) \cdot \|f - g\|_1 = (1/2) \cdot \int_{\mathbb{R}^d} |f(x) - g(x)|dx$. The *Kullback-Leibler (KL) divergence from g to f* is defined as $\mathrm{KL}(f||g) = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)} dx$.

For $f : A \to B$ and $A' \subseteq A$, the restriction of f to A' is the function $f|_{A'} : A' \to B$. For $y \in [0, \infty)$ and $f : \mathbb{R}^d \to [0, \infty)$ we denote by $L_f(y) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d \mid f(x) \geq y\}$ its *superlevel sets*. If f is log-concave, $L_f(y)$ is a convex set for all $y \in \mathbb{R}_+$. For a function $f : \mathbb{R}^d \to [0, \infty)$, we will denote by M_f its maximum value.

The VC inequality. We start by recalling the notion of VC dimension. We say that a set $X \subseteq \mathbb{R}^d$ is *shattered* by a collection $\mathcal{A}$ of subsets of $\mathbb{R}^d$, if for every $Y \subseteq X$ there exists $A \in \mathcal{A}$ such that $A \cap X = Y$. The *VC dimension* of a family $\mathcal{A}$ of subsets of $\mathbb{R}^d$ is defined to be the maximum cardinality of a subset $X \subseteq \mathbb{R}^d$ that is shattered by $\mathcal{A}$. If there is a shattered subset of size s for all $s \in \mathbb{Z}_+$, then we say that the VC dimension of $\mathcal{A}$ is ∞.

The empirical distribution, f_n, corresponding to a density $f : \mathbb{R}^d \to \mathbb{R}_+$ is the discrete probability measure defined by $f_n(A) = (1/n) \cdot \sum_{i=1}^n \mathbf{1}_A(X_i)$, where the X_i are iid samples drawn from f and $\mathbf{1}_S$ is the characteristic function of the set S. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lebesgue mea-

surable function. Given a family $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$, we define the $\mathcal{A}$-norm of f by $\|f\|_{\mathcal{A}} = \sup_{A\in\mathcal{A}} |f(A)|$. The VC inequality states the following:

Theorem 4 (VC inequality, see Devroye and Lugosi (2001), p. 31) *Let $f : \mathbb{R}^d \to [0,\infty)$ be a probability density function and f_n be the empirical distribution obtained after drawing n samples from f. Let $\mathcal{A}$ be a family of subsets over $\mathbb{R}^d$ with VC dimension V. Then $\mathbf{E}[\|f - f_n\|_{\mathcal{A}}] \le C\sqrt{V/n}$, for some universal constant $C > 0$.*

We will also require a high probability version of the VC inequality which can be obtained using the following standard uniform convergence bound:

Theorem 5 (see Devroye and Lugosi (2001), p. 17) *Let $\mathcal{A}$ be a family of subsets over $\mathbb{R}^d$ and f_n be the empirical distribution of n samples from the density $f : \mathbb{R}^d \to [0,\infty)$. Let X be the random variable $\|f - f_n\|_{\mathcal{A}}$. Then for all $\delta > 0$, we have that $\Pr[X - \mathbf{E}[X] > \delta] \le e^{-2n\delta^2}$.*

Approximating Convex Sets by Polytopes. We make use of the following quantitative bounds of Gordon et al. (1995) that provide volume approximation for any convex body by an inscribed and a circumscribed convex polytope respectively with a bounded number of facets:

Theorem 6 *For any convex body $K \subseteq \mathbb{R}^d$, and n sufficiently large, there exists a convex polytope $P \subseteq K$ with at most ℓ facets such that $\mathrm{vol}(K \setminus P) \le \frac{\kappa d}{\ell^{2/(d-1)}}\mathrm{vol}(K)$, where $\kappa > 0$ is a universal constant. Similarly, there exists a convex polytope P' where $K \subseteq P'$ with at most ℓ facets such that $\mathrm{vol}(P' \setminus K) \le \frac{\kappa d}{\ell^{2/(d-1)}}\mathrm{vol}(K)$.*

3. Main Result: Proof of Theorem 3

The following theorem is a more detailed version of Theorem 3 and is the main result of this paper:

Theorem 7 *Fix $d \in \mathbb{Z}_+$ and $\epsilon, \tau \in (0,1)$. Let $n = \Omega\left((d^2/\epsilon)\ln^3(d/(\epsilon\tau))\right)^{(d+3)/2}$. For any $f_0 \in \mathcal{F}_d$, with probability at least $1-\tau$ over the n samples from f_0, we have that $h^2(\hat{f}_n, f_0) \le \epsilon$.*

This section is devoted to the proof of Theorem 7, which follows from Lemma 19. We will require a sequence of intermediate lemmas and claims.

We summarize the notation that will appear throughout this proof. We use $f_0 \in \mathcal{F}_d$ to denote the target log-concave density. We denote by f_n the empirical distribution obtained after drawing n iid samples $X_1, \ldots, X_n$ from f_0 and by $\hat{f}_n$ the corresponding MLE. Given $d \in \mathbb{Z}_+$ and $0 < \epsilon, \tau < 1$, for concreteness, we will denote:

$$N_1 \stackrel{\text{def}}{=} \Theta\left((d^2/\epsilon)\ln^3(d/(\epsilon\tau))\right)^{(d+3)/2} ,$$

for a sufficiently large universal constant in the big-Θ notation. We will establish that N_1 is an upper bound on the desired sample complexity of the MLE. Moreover, we will denote

$$z \stackrel{\text{def}}{=} \ln(100n^4/\tau^2) , \delta \stackrel{\text{def}}{=} \epsilon/(32z) ,$$

$$p_{\min} \stackrel{\text{def}}{=} M_{f_0} e^{-z} ,$$

and

$$S \stackrel{\text{def}}{=} L_{f_0}(p_{\min}) \ .$$

We start by establishing an upper bound on the volume of superlevel sets:

Lemma 8 (see, e.g., Diakonikolas et al. (2017), p. 8) *Let $f \in \mathcal{F}_d$ with maximum value M_f. Then for all $w \geq 1$, we have $\mathrm{vol}(L_f(M_f e^{-w})) \leq w^d / M_f$, and $\Pr_{X \sim f}[f(X) \leq M_f e^{-w}] \leq O(d)^d e^{-w/2}$.*

We defer this proof to Appendix B. We use Lemma 8 to get a bound on the volume of the superlevel set that contains all the samples with high probability:

Corollary 9 *For $n \geq N_1$, we have that:*

(a) $\mathrm{vol}(S) \leq z^d / M_{f_0}$, *and*

(b) $\Pr_{X \sim f_0}[f_0(X) \leq M_{f_0}/(100n^4/\tau^2)] \leq \tau/(10n)$. *In particular, with probability at least $1 - \tau/10$, all samples $X_1, \ldots, X_n$ from f_0 are in S.*

Proof From Lemma 8, we have that $\mathrm{vol}(S) = \mathrm{vol}(L_{f_0}(M_{f_0} e^{-z})) \leq O(z^d / M_{f_0})$. Also from Lemma 8, we have that $\Pr_{X \sim f_0}[f_0(X) \leq M_{f_0}/(100n^4/\tau^2)] \leq \tau/(10n)$, if we assume a sufficiently large constant is selected in the definition of N_1. Taking a union bound over all samples, we get that with probability at least $1 - \tau/10$, all of the n samples are in S, as required. ■

We can now state our main lemma establishing an upper bound on the error of approximating the probability of every convex set:

Lemma 10 *For $n \geq N_1$, we have that with probability at least $1 - \tau/3$ over the choice of $X_1, \ldots, X_n$ drawn from f_0, for any convex set $C \subseteq \mathbb{R}^d$ it holds that $|f_0(C) - f_n(C)| \leq \delta$.*

The proof of Lemma 10 is deferred to Section A. In Section 4, we establish a weaker version of this lemma that requires more samples but has a simpler proof. Combining Lemma 10 with the observation that for any log-concave density f and $t > 0$ we have that $L_f(t)$ is convex, we obtain the following corollary:

Corollary 11 *Let $n \geq N_1$. Conditioning on the event of Lemma 10, we have that for any $f \in \mathcal{F}_d$ and for any $t \geq 0$ it holds $|\Pr_{X \sim f_0}[f(X) \geq t] - \Pr_{X \sim f_n}[f(X) \geq t]| < \delta$.*

We will require the following technical claim, which follows from standard properties of Lebesgue integration (see Appendix B):

Lemma 12 *Let $g, h : \mathbb{R}^d \to \mathbb{R}$ be probability distributions, and $\phi : \mathbb{R} \to \mathbb{R}$. If $\mathbf{E}_{Y \sim g}[\phi(Y)]$, $\mathbf{E}_{Y \sim h}[\phi(Y)]$ are both finite, then $|\mathbf{E}_{Y \sim g}[\phi(Y)] - \mathbf{E}_{Y \sim h}[\phi(Y)]| \leq \int_{-\infty}^{\infty} |\Pr_{Y \sim g}[\phi(Y) < x] - \Pr_{Y \sim h}[\phi(Y) < x]| dx$.*

Our next lemma establishes a useful upper bound on the empirical error of the truncated likelihood of any log-concave density:

Lemma 13 *Let $n \geq N_1$ and $f \in \mathcal{F}_d$ with maximum value M_f. For all $\rho \in (0, M_f]$, conditioning on the event of Corollary 11, we have*

$$|\mathbf{E}_{X\sim f_0}[\ln(\max(f(X),\rho))] - \mathbf{E}_{X\sim f_n}[\ln(\max(f(X),\rho))]| \leq \delta \cdot \ln(M_f/\rho) \ .$$

Proof Letting $h = f_0$, $g = f_n$, and $\phi(x) = \ln(\max(f(x),\rho))$, by Lemma 12 we have

$$\begin{aligned}
&|\mathbf{E}_{X\sim f_0}[\ln(\max(f(X),\rho))] - \mathbf{E}_{X\sim f_n}[\ln(\max(f(X),\rho))]| \\
&\quad\leq \int_{-\infty}^{\infty} |\Pr_{X\sim f_0}[\ln(\max(f(X),\ln\rho)) < t] - \Pr_{X\sim f_n}[\ln(\max(f(X),\rho)) < t]|\, dt \\
&\quad= \int_{-\infty}^{\ln M_f} |\Pr_{X\sim f_0}[\max(\ln f(X),\ln\rho)) < t] - \Pr_{X\sim f_n}[\max(\ln f(X),\ln\rho)) < t]|\, dt \\
&\quad= \int_{\ln\rho}^{\ln M_f} |\Pr_{X\sim f_0}[\ln(f(X)) < t] - \Pr_{X\sim f_n}[\ln(f(X)) < t]|\, dt \\
&\quad= \int_{\ln\rho}^{\ln M_f} \left|\Pr_{X\sim f_0}[f(X) < e^t] - \Pr_{X\sim f_n}[f(X) < e^t]\right| dt \\
&\quad= \int_{\ln\rho}^{\ln M_f} \left|\Pr_{X\sim f_0}[f(X) \geq e^t] - \Pr_{X\sim f_n}[f(X) \geq e^t]\right| dt.
\end{aligned}$$

Since we conditioned on the event of Corollary 11, we have $|\Pr_{X\sim f_0}[f(X) \geq t] - \Pr_{X\sim f_n}[f(X) \geq t]| \leq \delta$ for all $t \geq 0$. Therefore, we have that

$$|\mathbf{E}_{X\sim f_0}[\ln(\max(f(X),\rho))] - \mathbf{E}_{X\sim f_n}[\ln(\max(f(X),\rho))]| \leq \int_{\ln\rho}^{\ln M_f} \delta dt = \delta \cdot (\ln M_f - \ln\rho) \ ,$$

which concludes the proof. ■

For f_0 itself, we can use Hoeffding's inequality to get a bound on the empirical error of its likelihood:

Lemma 14 *Let $n \geq N_1$. Conditioning on the event of Corollary 9, with probability at least $1-\tau/3$ over $X_1, \ldots, X_n$, we have that*

$$\left|\frac{1}{n}\sum_{i=1}^{n} \ln f_0(X_i) - \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right]\right| \leq \epsilon/8 \ .$$

We defer this proof to Appendix B. The following simple lemma shows that the MLE is supported in the convex hull of the samples:

Lemma 15 *Let $n \geq 1$. Let $X_1, \ldots, X_n$ be samples drawn from f_0, and C be the convex hull of these samples. Then, for all $x \in \mathbb{R}^d \setminus C$, we have $\hat{f}_n(x) = 0$.*

We defer this proof to Appendix B. We need to truncate the likelihood at a density small enough to be ignored for f close to f_0. This motivates the following definition:

Definition 16 *We define $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ such that $\tilde{f}(x) \stackrel{\text{def}}{=} \max\{p_{\min}, \hat{f}_n(x)\}$.*

We show that this truncation and renormalization does not affect the MLE $\hat{f}_n$ by much:

Lemma 17 *Let $n \geq N_1$. Let $g(x) \stackrel{\text{def}}{=} \alpha \tilde{f}(x)\mathbf{1}_S(x)$, $\alpha \in [0, \infty)$, be such that $\int_S g(x)dx = 1$. Conditioning on the event of Corollary 9, we have the following:*

(a) $1 - \epsilon/32 \leq \alpha \leq 1$, *and*

(b) $\mathrm{d_{TV}}(g, \hat{f}_n) \leq 3\epsilon/64$.

Proof We start by proving (a). By the definition of g and Lemma 15, we have $\alpha = \alpha \int_S \hat{f}_n(x)dx \leq \alpha \int_S f'(x)dx = \int_S g(x)dx = 1$, i.e., $\alpha \leq 1$. Furthermore, by the definition of $p_{\min}$ and Corollary 9, we have

$$p_{\min} \cdot \mathrm{vol}(S) \leq \frac{M_{f_0}}{(100n^4/\tau^2)} \cdot \frac{O((\ln(100n^4/\tau^2))^d)}{M_{f_0}} \leq \epsilon/32, \tag{1}$$

and therefore

$$1 = \int_S g(x)dx \leq \alpha \left(\int_S p_{\min}\, dx + \int_S \hat{f}_n(x)dx \right) \leq \alpha(p_{\min} \cdot \mathrm{vol}(S) + 1) \leq \alpha(\epsilon/32 + 1).$$

From this it follows that $\alpha \geq 1/(1 + \epsilon/32) \geq 1 - \epsilon/32$. We have

$$\mathrm{d_{TV}}(g, \hat{f}_n) = \frac{1}{2} \int_{\mathbb{R}^d} |g(x) - \hat{f}_n(x)|dx = \frac{1}{2} \int_S |g(x) - \hat{f}_n(x)|dx\ , \tag{2}$$

since $g(x) = 0$ for $x \notin S$ and $\hat{f}_n$ is supported in S by Lemma 15. We can then write

$$\begin{aligned}
\frac{1}{2} \int_S |g(x) - \hat{f}_n(x)|dx &= \frac{1}{2} \int_S |\alpha f'(x) - \hat{f}_n(x)|dx \\
&\leq \frac{1}{2} \int_S |\alpha - 1| \cdot \hat{f}_n(x)dx + p_{\min} \cdot \mathrm{vol}(S) \\
&\leq \frac{|\alpha - 1|}{2} \int_S \hat{f}_n(x)dx + \epsilon/32 \qquad \text{(from (1))} \\
&\leq \frac{|1 - \alpha|}{2} + \epsilon/32 \leq 3\epsilon/64\ ,
\end{aligned}$$

which completes the proof. ■

To deal with the dependence on the maximum value of f in Lemma 13, we need to bound the maximum value of the MLE.

Lemma 18 *Let $n \geq N_1$. Let $X_1, \ldots, X_n$ be samples drawn from f_0. Then conditioning on the events of Corollary 11 and Lemma 14, for any $f \in \mathcal{F}_d$ with maximum value M_f such that $\ln(M_f / p_{\min}) \geq 4 \ln(100n^4/\tau^2)$, we have $\frac{1}{n}\sum_{i=1}^n \ln f(X_i) < \frac{1}{n}\sum_{i=1}^n \ln f_0(X_i)$.*

This holds because a density f with a large M_f is small outside on a set of small volume, which most of the samples will be outside. We defer this proof to Appendix B.

We have now reached the final result of this section, from which Theorem 7 directly follows. Combining previous lemmas, we show that the likelihood under f_0 of the truncated MLE is close to that of f_0 and so they are close in KL divergence, which leads to a bound in the Hellinger distance of the MLE itself:

Lemma 19 *Let $n \geq N_1$. Let $X_1, \ldots, X_n$ be samples drawn from f_0. With probability at least $1 - \tau$, we have that $h^2(f_0, \hat{f}_n) \leq \epsilon$.*

Proof In this lemma, we will apply Lemmas 13, 14, 17, and 18. By examining the conditions of these lemmas, it is easy to see that with probability at least $1 - \tau$ they all hold. We henceforth condition on this event.

Let $X_1, \ldots, X_n$ be samples drawn from f_0, let $\hat{f}_n$ be as in Definition 2. Let g and $\tilde{f}$ be as defined in Lemma 17 and Definition 16. Let S be as defined in Corollary 9 Then we have that

$$\begin{aligned}
\mathbf{E}_{X\sim f_0}[\ln g(X)] &= \mathbf{E}_{X\sim f_0}[\ln(\alpha \tilde{f}(X))] \\
&\geq \mathbf{E}_{X\sim f_0}[\ln \tilde{f}(X)] - \epsilon/16 && \text{(since } \alpha > 1 - \epsilon/32\text{)} \\
&= \mathbf{E}_{X\sim f_0}[\ln(\max\{\hat{f}_n(X), p_{\min}\})] - \epsilon/16 \\
&\geq \mathbf{E}_{X\sim f_n}[\ln(\max\{\hat{f}_n(X), p_{\min}\})] - 3\epsilon/16 && \text{(by Lemmas 13 and 18)} \\
&\geq \frac{1}{n}\sum_i \ln \hat{f}_n(X_i) - 3\epsilon/16 \\
&\geq \frac{1}{n}\sum_i \ln f_0(X_i) - 3\epsilon/16 \\
&\geq \mathbf{E}_{X\sim f_0}[\ln f_0(X)] - 5\epsilon/16. && \text{(using Lemma 14)}
\end{aligned}$$

Thus, we obtain that

$$\mathrm{KL}(f_0||g) = \mathbf{E}_{X\sim f_0}[\ln f_0(X)] - \mathbf{E}_{X\sim f_0}[\ln g(X)] \leq 5\epsilon/16. \tag{3}$$

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 $h^2(k_1, k_2) \leq \mathrm{d_{TV}}(k_1, k_2)$ and $h^2(k_1, k_2) \leq \mathrm{KL}(k_1||k_2)$. Therefore, we have that

$$\begin{aligned}
h(f_0, \hat{f}_n) &\leq h(f_0, g) + h(g, \hat{f}_n) \\
&\leq \mathrm{KL}(f_0||g)^{1/2} + \mathrm{d_{TV}}(g, \hat{f}_n)^{1/2} \\
&= (5\epsilon/16)^{1/2} + (3\epsilon/64)^{1/2} && \text{(by (3) and Lemma 17)} \\
&\leq \epsilon^{1/2}\,,
\end{aligned}$$

concluding the proof. ■

4. Warmup for the Proof of Lemma 10

For the sake of exposition of the main ideas used in the proof of Lemma 10, we first prove Lemma 21, which achieves a weaker bound on the sample complexity, but has a significantly simpler proof. Let us first give a brief, and somewhat imprecise, overview of the proof of Lemma 21. The high-level goal is to approximate some convex set $C \subseteq \mathbb{R}^d$ by some set, belonging to a family of low VC dimension. We then can obtain the desired bound using Theorem 4. To that end, we compute inner and outer approximations, C^{in} and C^{out}, of C via polyhedral sets with a small number of facets. By Lemma 20, we can argue that the VC dimension of this family is low. We therefore obtain that

f_0 and f_n are close on the inner and outer approximations of C. It remains to argue that the total difference between f_0 and f_n in $C^{\text{out}} \setminus C^{\text{in}}$ is also small. It thus suffices to bound the volume of $C^{\text{out}} \setminus C^{\text{in}}$. This can be achieved by first defining some set $S \subseteq \mathbb{R}^d$ that excludes the tail of f_0. Since f_0 is logconcave, we can show that S has small volume. The final bound is obtained by restricting the above argument on $C \cap S$.

Throughout this section, we define $N_2 \stackrel{\text{def}}{=} \Theta\left(2^{O(d)}(d^{(2d+3)}/\epsilon)(\ln(d^{(d+1)}/(\epsilon\tau)))^{(d+1)}\right)^{(d+5)/2}$.

We will require the following simple fact:

Lemma 20 (see Alon et al. (1992)) *Let $h, d \in \mathbb{Z}_+$, and let $\mathcal{A}$ be the set of all convex polytopes in $\mathbb{R}^d$ with at most h facets. Then, the VC dimension of $\mathcal{A}$ is at most $2(d+1)h\log((d+1)h)$.*

The main result of this section is the following:

Lemma 21 *Let $n \geq N_2$. With probability at least $1 - \frac{3\tau}{10}$ over the choice of $X_1, \ldots, X_n$, for any convex set $C \subseteq \mathbb{R}^d$ it holds that $|f_0(C) - f_n(C)| < \delta$.*

Proof Recall that $z = \ln(100n^4/\tau^2)$ and $S = L_{f_0}(M_{f_0}e^{-z})$. Let $\mathcal{C}$ be the family of convex sets on $\mathbb{R}^d$. For any $C \in \mathcal{C}$, let $C' = C \cap S$. Since f_0 is log-concave, it follows that S is convex, and thus C' is also convex.

Let $\mathcal{E}_1$ be the event that all samples $X_1, \ldots, X_n$ lie in S. Let $\mathcal{X} = X_1, \ldots, X_n$. By Corollary 9, we have

$$\Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \geq 1 - \tau/10. \tag{4}$$

Conditioned on $\mathcal{E}_1$ occurring, we have with probability 1, for any $C \in \mathcal{C}$, $f_n(C) = f_n(C')$. In other words,

$$\Pr_{\mathcal{X}\sim f_0}[\forall C \in \mathcal{C}, f_n(C \setminus C') = 0 | \mathcal{E}_1] = 1. \tag{5}$$

From Corollary 9, we have $\Pr_{X\sim f_0}[f_0(X) \leq M_{f_0}/(100n^4/\tau^2)] \leq \tau/(10n)$, and therefore

$$f_0(C \setminus C') \leq f_0(\mathbb{R}^d \setminus S) \leq \tau/(10n) \leq \delta/5. \tag{6}$$

Combining (4), (5), (6), and letting $Q = \sup_{C\in\mathcal{C}} |f_0(C \setminus C') - f_n(C \setminus C')|$, we have that

$$\begin{aligned}
\Pr_{\mathcal{X}\sim f_0}[Q \leq \delta/5] &\geq \Pr_{\mathcal{X}\sim f_0}[Q \leq \delta/5 | \mathcal{E}_1] \cdot \Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \\
&\geq \Pr_{\mathcal{X}\sim f_0}\left[\forall C \in \mathcal{C}, f_n(C \setminus C') = 0 | \mathcal{E}_1\right] \cdot \Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \\
&\geq 1 - \tau/10.
\end{aligned} \tag{7}$$

Let $\mathcal{A}$ be the set of convex polytopes in $\mathbb{R}^d$ with at most $H = (10\kappa d z^d/\delta)^{(d-1)/2}$ facets, where κ is the universal constant in Theorem 6. By Theorem 6, there exist convex polytopes $T, T' \in \mathcal{A}$, with $T \subseteq C' \subseteq T'$, such that $\text{vol}(C' \setminus T) \leq \frac{\delta}{10z^d}\text{vol}(S) \leq \frac{\delta}{10M_{f_0}}$ and $\text{vol}(T' \setminus C') \leq \frac{\delta}{10z^d}\text{vol}(S) \leq \frac{\delta}{10M_{f_0}}$. Therefore, since M_{f_0} is the maximum value of f_0, we have

$$f_0(C' \setminus T) \leq \text{vol}(C' \setminus T) \cdot M_{f_0} \leq \delta/10, \tag{8}$$

and

$$f_0(T' \setminus C') \leq \mathrm{vol}(T' \setminus C') \cdot M_{f_0} \leq \delta/10. \tag{9}$$

Noting that $\mathbf{E}[|f_0(T) - f_n(T)|] \leq \mathbf{E}[||f_0 - f_n||_{\mathcal{A}}]$, by Theorem 4 we have for some universal constant α that $\mathbf{E}[|f_0(T) - f_n(T)|] \leq \sqrt{\alpha V/n}$. The following claim is obtained via a simple calculation (see Appendix B):

Claim 22 *For $n \geq N_2$, we have that $\sqrt{\alpha V/n} \leq \delta/10$.*

Let $\mathcal{E}_2$ be the event that $||f_0 - f_n||_{\mathcal{A}} \leq 3\delta/10$. By Claim 22 and Theorem 5 we have

$$\begin{aligned}
\Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_2] &= 1 - \Pr_{\mathcal{X}\sim f_0}[||f_0 - f_n||_{\mathcal{A}} > 3\delta/10] \\
&\geq 1 - \Pr_{\mathcal{X}\sim f_0}[||f_0 - f_n||_{\mathcal{A}} - \mathbf{E}[||f_0 - f_n||_{\mathcal{A}}] > \delta/5] \\
&\geq 1 - e^{-2n(\delta/5)^2} \\
&\geq 1 - \tau/5.
\end{aligned} \tag{10}$$

For any choice of samples $X_1, \ldots, X_n$, we have

$$\begin{aligned}
f_n(C') &\geq f_n(T) && \text{(since } T \subseteq C'\text{)} \\
&\geq f_0(C') - f_0(C' \setminus T) - |f_0(T) - f_n(T)| \\
&\geq f_0(C') - \frac{\delta}{10} - |f_0(T) - f_n(T)|. && \text{(by (8))}
\end{aligned} \tag{11}$$

In a similar way, using that $C' \subseteq T'$, we have

$$f_n(C') \leq f_0(C') + \frac{\delta}{10} + |f_0(T') - f_n(T')|. \qquad \text{(by (9))} \tag{12}$$

By (11) and (12) and the union bound, we obtain

$$|f_n(C') - f_0(C')| \leq \frac{\delta}{10} + \max\left\{|f_0(T) - f_n(T)|, |f_0(T') - f_n(T')|\right\}. \tag{13}$$

Combining (7), (10), (13), and letting $Q' = \sup_{C\in\mathcal{C}} |f_n(C) - f_0(C)|$, we get

$$\begin{aligned}
\Pr_{\mathcal{X}\sim f_0}[Q' \leq 2\delta/5] &\geq \Pr_{\mathcal{X}\sim f_0}[(\sup_{C\in\mathcal{C}} |f_n(C \setminus C') - f_0(C \setminus C')| \leq \delta/5) \wedge Q \leq 3\delta/10)] \\
&\geq \Pr_{\mathcal{X}\sim f_0}[(\sup_{C\in\mathcal{C}} |f_n(C \setminus C') - f_0(C \setminus C')| \leq \delta/5) \wedge (||f_n - f_0||_{\mathcal{A}} \leq 3\delta/10)] \\
&\geq 1 - 3\tau/10,
\end{aligned}$$

which concludes the proof. ■

5. Conclusions

In this paper, we gave the first sample complexity upper bound for the MLE of multivariate log-concave densities on $\mathbb{R}^d$, for any $d \geq 4$. Our upper bound agrees with the previously known lower bound up to a multiplicative factor of $\tilde{O}_d(\epsilon^{-1})$.

A number of open problems remain: What is the *optimal* sample complexity of the multivariate log-concave MLE? In particular, is the log-concave MLE sample-optimal for $d \geq 4$? Does the multivariate log-concave MLE have similar adaptivity properties as in one dimension? And is there a polynomial time algorithm to compute it?

References

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.

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.

N. Alon, J. Spencer, and P. Erdos. *The Probabilistic Method*. Wiley-Interscience, New York, 1992.

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.

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.

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.

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.

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.

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.

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

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

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

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

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

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

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

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

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, 2014b.

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

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.

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

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

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.

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.

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

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, 2016a. Full version available at https://arxiv.org/abs/1505.00662.

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, 2016b. Full version available at https://arxiv.org/abs/1511.04066.

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

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

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.

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.

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.

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

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.

Y. Gordon, M. Meyer, and S. Reisner. Constructing a polytope to approximate a convex body. *Geometriae Dedicata*, 57(2):217–222, 1995.

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

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

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

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

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

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

A. Kim, A. Guntuboyina, and R. J. Samworth. Adaptation in log-concave density estimation. *ArXiv e-prints*, 2016. Available at http://arxiv.org/abs/1609.00861.

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.

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

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

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

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.

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

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

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.

G. Valiant and P. Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 142–155, 2016.

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

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

J. A. Wellner. Nonparametric estimation of s-concave and log-concave densities: an alternative to maximum likelihood. *Talk given at European Meeting of Statisticians, Amsterdam*, 2015. Available at https://www.stat.washington.edu/jaw/RESEARCH/TALKS/EMS-2015.1-rev1.pdf.

Appendix A. Proof of Lemma 10

We are now ready to prove the main technical part of our work, which is Lemma 10. The proof builds upon the argument used in the proof of Lemma 21, which achieves a weaker sample complexity bound. Recall that in the proof of Lemma 21 we use inner and outer polyhedral approximations of C, restricted on some appropriate bounded $S \subseteq \mathbb{R}^d$. The main difference in the proof of Lemma 10 is that we now use roughly $O(\log n)$ inner and outer polyhedral approximations of intersections of C with different super-levelsets of f_0. We need slightly more samples due to the higher number of facets, and consequently higher VC dimension of the resulting approximations. However, since we use a finer discretization of the values of f_0, we incur lower error in total.

The following Lemma is implicit in Diakonikolas et al. (2017). We reproduce its proof for completeness in Appendix B.

Lemma 23 *Let $L, H \in \mathbb{Z}_+$. We define the set $\mathcal{A}_{H,L}$, elements of which are defined by the following process: Starting with L convex polytopes each with at most H facets, all combinations of intersection, difference, and union of these polytopes are elements of $\mathcal{A}_{H,L}$. If V is the VC dimension of $\mathcal{A}_{H,L}$, then $V/\log(V) = O(dLH)$.*

We are now prepared to present the proof of Lemma 10. Let

$$S_i = L_{f_0}(M_{f_0} e^{-i})$$

and let $S_0 = \emptyset$. Let $L = \ln(100n^4/\tau)$. Note that by Lemma 8, we have that $\Pr_{X\sim f_0}[f_0(X) \leq M_{f_0} e^{-z}] = O(d)^d e^{-z/2}$ and thus

$$\Pr_{X\sim f_0}[X \notin S_L] = \Pr_{X\sim f_0}[f_0(X) < M_{f_0} e^{-L}] \leq \frac{\tau}{10n}.$$

Let $\mathcal{E}_1$ be the event that all samples $X_1, \ldots, X_n$ lie in S_L. Let $\mathcal{X} = X_1, \ldots, X_n$. We have that

$$\Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \geq 1 - \tau/10. \tag{14}$$

Let $\mathcal{C}$ be the set of convex sets in $\mathbb{R}^d$. For any $C \in \mathcal{C}$, for all $i \in [L]$, let

$$C_i = C \cap S_i.$$

Note that, conditioned on $\mathcal{E}_1$ occurring, we have with probability 1 that, for all $C \in \mathcal{C}$, $f_n(C) = f_n(C_L)$. In other words,

$$\Pr_{\mathcal{X}\sim f_0}[\forall C \in \mathcal{C}, f_n(C \setminus C_L) = 0 | \mathcal{E}_1] = 1. \tag{15}$$

Furthermore, by our choice of L we have $f_0(\mathbb{R}^d \setminus S_L) \leq \frac{\tau}{10n}$, and therefore

$$f_0(C \setminus C_L) \leq \frac{\tau}{10n} \leq \delta/5. \tag{16}$$

Combining 14, 15, 16, and letting $Q = \sup_{C\in\mathcal{C}} |f_0(C \setminus C_L) - f_n(C \setminus C_L)|$, we have

$$\begin{aligned}
\Pr_{\mathcal{X}\sim f_0}[Q \leq \delta/5] &\geq \Pr_{\mathcal{X}\sim f_0}\left[Q \leq \delta/5 \middle| \mathcal{E}_1\right] \cdot \Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \\
&\geq \Pr_{\mathcal{X}\sim f_0}[\forall C \in \mathcal{C}, f_n(C \setminus C_L) = 0 | \mathcal{E}_1] \cdot \Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_1] \\
&\geq 1 - \tau/10.
\end{aligned} \tag{17}$$

Using Theorem 6, for $i \in [L]$ let $P_i^{\text{in}}, P_i^{\text{out}}$ be convex polytopes with $H = (10\kappa d/\delta)^{(d-1)/2}$ facets, where κ is the universal constant from Theorem 6, such that $P_i^{\text{in}} \subseteq C_i \subseteq P_i^{\text{out}}$,

$$\text{vol}(C_i \setminus P_i^{\text{in}}) \leq \delta \cdot \text{vol}(C_i)/10 \leq \delta \cdot \text{vol}(S_i)/10, \tag{18}$$

and

$$\text{vol}(P_i^{\text{out}} \setminus C_i) \leq \delta \cdot \text{vol}(C_i)/10 \leq \delta \cdot \text{vol}(S_i)/10. \tag{19}$$

Let

$$C^{\text{in}} = \bigcup_{i\in[L]} P_i^{\text{in}}.$$

For any $i \in [L]$, let P_i^S be a convex polytope with at most H facets such that $P_i^S \subseteq S_i$ and $\text{vol}(S_i \setminus P_i^S) \leq \delta \cdot \text{vol}(S_i)/10$.

Let

$$S_i' = \bigcup_{1\leq j\leq i} P_j^S$$

and $S_0' = \emptyset$. Let

$$C^{\text{out}} = \bigcup_{i\in[L]} (P_i^{\text{out}} \setminus S_{i-1}').$$

We will now show that C^{in} and C^{out} satisfy the following conditions:

1. $C^{\text{in}} \subseteq C_L \subseteq C^{\text{out}}$.

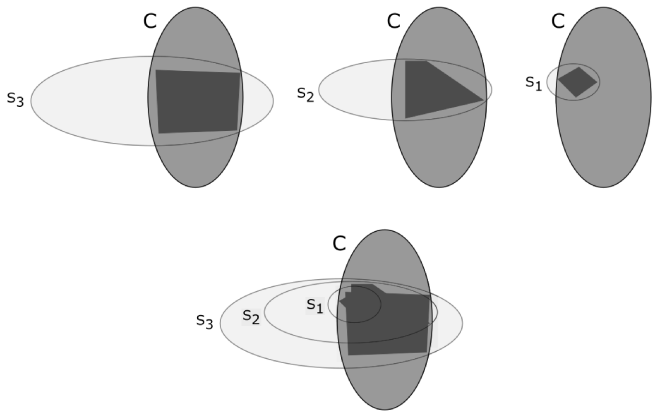

Figure 1: Constructing C^{in}. For each set S_i, a convex polytope approximating $C \cap S_i$ from the inside is found, and C^{in} is formed by taking the union of these convex polytopes.

2. $f_0(C^{\text{out}} \setminus C_L) < \delta/2$.

3. $f_0(C_L \setminus C^{\text{in}}) < \delta/2$.

First, we consider C^{in}. Since $P_i^{\text{in}} \subseteq C_i \subseteq C_L$ for all $i \in [L]$, it follows that $\bigcup_{i\in[L]} P_i^{\text{in}} = C^{\text{in}} \subseteq C_L$. Observe that by the above definitions, we have that

$$(C_L \setminus C^{\text{in}}) \cap (S_i \setminus S_{i-1}) \subseteq (C_L \setminus C^{\text{in}}) \setminus S_{i-1} \subseteq (C_L \setminus P_i^{\text{in}}) \setminus S_{i-1}. \tag{20}$$

From (20), we therefore have

$$(C_L \setminus C^{\text{in}}) = \bigcup_{i\in[L]} \left[(C_L \setminus C^{\text{in}}) \cap (S_i \setminus S_{i-1})\right] \subseteq \bigcup_{i\in[L]} (C_i \setminus P_i^{\text{in}}) \setminus S_{i-1}, \tag{21}$$

and so

$$\begin{aligned}
f_0(C_L \setminus C^{\text{in}}) &\leq \sum_{i\in[L]} f_0((C_i \setminus P_i^{\text{in}}) \setminus S_{i-1}) && \text{(by (21))} \\
&\leq \sum_{i\in[L]} \text{vol}((C_i \setminus P_i^{\text{in}}) \setminus S_{i-1}) M_{f_0} e^{-(i-1)} \\
&\leq \sum_{i\in[L]} \text{vol}(C_i \setminus P_i^{\text{in}}) M_{f_0} e^{-(i-1)} \\
&\leq \sum_{i\in[L]} (\delta/10)\text{vol}(S_i) M_{f_0} e^{-(i-1)} && \text{(by (18))} \\
&\leq (\delta/10) \sum_{i\in[L]} \text{vol}(L_{f_0}(M_{f_0} e^{-i})) M_{f_0} e^{-(i-1)} \\
&\leq (\delta/10) \int_0^{M_{f_0}} \text{vol}(L_{f_0}(y)) dy < \delta/2. && (22)
\end{aligned}$$

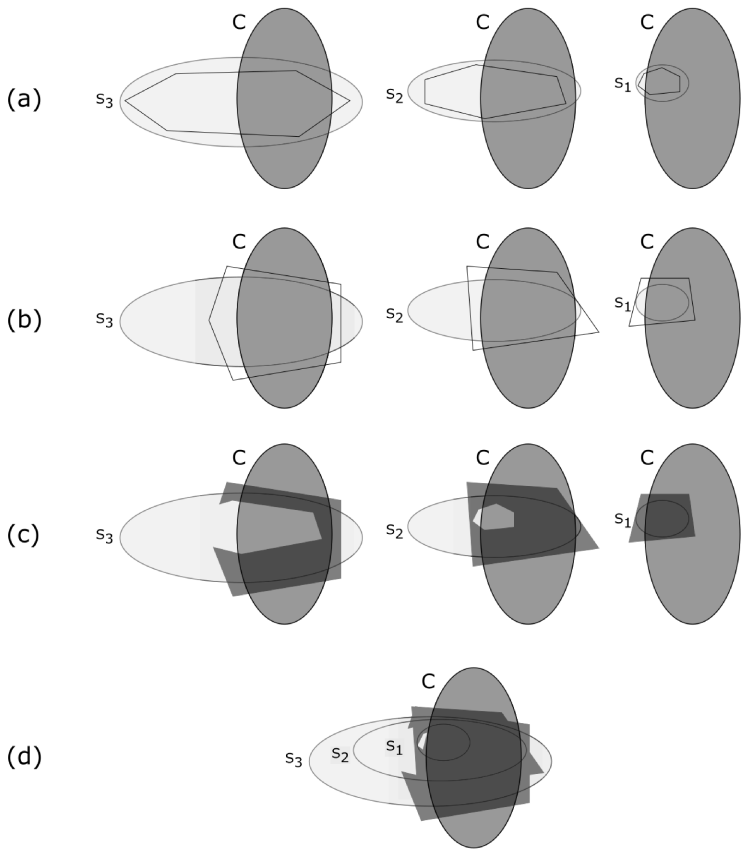

Figure 2: Constructing C^{out}. For each set S_i, a convex polytope approximating S_i from the inside is found (P_i^S, see row **(a)**), and a convex polytope approximating $C \cap S_i$ from the outside is found (P_i^{out}, see row **(b)**). For each i, the set $P_i^{\text{out}} \setminus (\cup_{j=1}^{i-1} P_j^S)$ is constructed (see row **(c)**), and the union of these sets finish the construction of C^{out} (see row **(d)**).

Now we consider C^{out}. Let $x \in C_L$. Then there exists $i \in [L]$ such that $x \in S_i$ and $x \notin S_{i-1}$. Thus $x \in P_i^{\text{out}}$ and $x \notin S'_{i-1}$, from which we have that $x \in C^{\text{out}} = \bigcup_{i\in[L]}(P_i^{\text{out}} \setminus S'_{i-1})$. Therefore $C_L \subseteq C^{\text{out}}$. Let $y \in C^{\text{out}} \setminus C_L$. From the definition of C^{out}, there must exist some $i \in [L]$ such that $y \in P_i^{\text{out}} \setminus S'_{i-1}$. If $y \in P_i^{\text{out}} \setminus C_i$, we are done. Suppose that $y \notin P_i^{\text{out}} \setminus C_i$. Since we have that $y \in P_i^{\text{out}}$, we must also have that $y \in C_i$. But $C_i \subseteq C_L$, and we began with $y \in C^{\text{out}} \setminus C_L$, which makes a contradiction. Therefore,

$$C^{\text{out}} \setminus C_L \subseteq \cup_{i\in[L]} \left(P_i^{\text{out}} \setminus C_i\right). \tag{23}$$

Thus, we have that

$$\begin{aligned}
f_0(C^{\text{out}} \setminus C_L) &\leq \sum_{i\in[L]} f_0(P_i^{\text{out}} \setminus C_i) && \text{(by (23))} \\
&\leq \sum_{i\in[L]} \text{vol}(P_i^{\text{out}} \setminus C_i) M_{f_0} e^{-(i-1)} \\
&\leq \sum_{i\in[L]} (\delta/10)\text{vol}(S_i) M_{f_0} e^{-(i-1)} && \text{(by (19))} \\
&\leq (\delta/10) \sum_{i\in[L]} \text{vol}(L_{f_0}(M_{f_0}e^{-i})) M_{f_0} e^{-(i-1)} \\
&\leq (\delta/10) \int_0^{M_{f_0}} \text{vol}(L_{f_0}(y))dy < \delta/2. && (24)
\end{aligned}$$

We define the set $\mathcal{A}$, elements of which are defined by the following process: Starting with $2L$ convex polytopes each with at most H facets, all combinations of intersection, difference, and union of these convex polytopes are elements of $\mathcal{A}$. Then for any convex set C with $C^{\text{in}}, C^{\text{out}}$ as defined above, we have that $C^{\text{out}}, C^{\text{in}} \in \mathcal{A}$. From Lemma 23, we have that if V is the VC dimension of $\mathcal{A}$, then

$$V/\ln(V) = O(dLH).$$

Using Theorem 4, we have for some universal constant α that

$$\mathbf{E}[|f_0(C^{\text{in}}) - f_n(C^{\text{in}})|] \leq \mathbf{E}[||f_0 - f_n||_{\mathcal{A}}] = \sqrt{\frac{\alpha V}{n}}. \tag{25}$$

The following claim is obtained via a simple calculation (see Appendix B):

Claim 24 *For $n \geq N_1$ we have that $\sqrt{\frac{\alpha V}{n}} \leq \delta/10$.*

Let $\mathcal{E}_2$ be the event that $||f_0 - f_n||_{\mathcal{A}} \leq \delta/2$. Then by (25), Claim 24, and Theorem 5, we have that

$$\begin{aligned}
\Pr_{\mathcal{X}\sim f_0}[\mathcal{E}_2] &= 1 - \Pr_{\mathcal{X}\sim f_0}[||f_0 - f_n||_{\mathcal{A}} > \delta/2] \\
&\geq 1 - \Pr_{\mathcal{X}\sim f_0}[||f_0 - f_n||_{\mathcal{A}} - \mathbf{E}[||f_0 - f_n||_{\mathcal{A}}] > \delta/10] \\
&\geq 1 - e^{-2n(\delta/10)^2} \\
&\geq 1 - \tau/10. && (26)
\end{aligned}$$

This next claim follows from (22) and (24). The full proof can be found in Appendix B.

Claim 25 *If $\mathcal{E}_1$ and $\mathcal{E}_2$ hold, we have that* $\sup_{C\in\mathcal{C}} |f_n(C_L) - f_0(C_L)| \le 7\delta/10$.

Combining (17), (26), Claim 25, and letting $Q' = \sup_{C\in\mathcal{C}} |f_n(C) - f_0(C)|$, we get

$$\begin{aligned}\Pr_{\mathcal{X}\sim f_0}[Q' \le \delta] &\ge \Pr_{\mathcal{X}\sim f_0}[\sup_{C\in\mathcal{C}} |f_n(C\setminus C_L) - f_0(C\setminus C_L)| \le \delta/5) \wedge \sup_{C\in\mathcal{C}} |f_n(C_L) - f_0(C_L)| \le 7\delta/10)] \\ &\ge 1 - \frac{\tau}{10} - \frac{\tau}{5} \\ &\ge 1 - 3\tau/10,\end{aligned}$$

which concludes the proof.

Appendix B. Deferred Proofs

B.1. Proof of Lemma 8

W.l.o.g. we may assume that $f(0) = M_f$. We let $R = L_f(M_f/e)$. Then using the fact that if $y \le M_f/e$ then $R \subseteq L_f(y)$, we have that

$$1 = \int_{\mathbb{R}_+} \mathrm{vol}(L_f(y))dy \ge \int_{0\le y\le M_f/e} \mathrm{vol}(L_f(y))dy \ge \int_{0\le y\le M_f/e} \mathrm{vol}(R)dy = \frac{M_f}{e}\cdot \mathrm{vol}(R) \tag{27}$$

Suppose that $f(x) \ge M_f e^{-w}$, for some $x \in \mathbb{R}^d$. By the definition of log-concavity we have $f(x/w) \ge f(0)^{(w-1)/w} f(x)^{1/w}$. By the assumption we get $f(x/w) \ge M_f^{(w-1)/w}(M_f/e^w)^{1/w} = M_f^{(w-1)/w} M_f^{1/w}/e = M_f/e$. Thus $x/w \in R$, and so $x \in wR$. Therefore $L_f(M_f e^{-w}) \subseteq wR$. Thus by (27) we get

$$\mathrm{vol}(L_f(M_f e^{-w})) \le \mathrm{vol}(wR) \le w^d \cdot \mathrm{vol}(R) = w^d/M_f, \tag{28}$$

which proved the first part of the assertion.

It remains to prove the second part. We have

$$\begin{aligned}\Pr_{X\sim f}[f(X) \le M_f e^{-z}] &\le \int_0^{M_f e^{-z}} \mathrm{vol}(L_f(y))dy \\ &= \int_z^\infty \mathrm{vol}(L_f(M_f e^{-x}))M_f e^{-x}dx && \text{(setting } y = M_f e^{-x}\text{)} \\ &\le \int_z^\infty O(x^d/M_f)M_f e^{-x}dx && \text{(by (28))} \\ &= \int_z^\infty O(x^d e^{-x})dx \\ &\le \int_z^\infty O(d)^d e^{-x/2}dx && \text{(since } e^{x/2} \ge (x/2)^d/d!\text{)} \\ &= O(d)^d e^{-z/2},\end{aligned}$$

which concludes the proof.

B.2. Proof of Lemma 12

We begin with a few common definitions and observations. If X is a random variable defined on a probability space (Ω, Σ, P), then the expected value $\mathbf{E}[X]$ of X is defined as the Lebesgue integral

$$\mathbf{E}[X] = \int_{\Omega} X(\omega) dP(\omega).$$

Next, we define two functions

$$X_+(\omega) = \max(X(\omega), 0)$$

and

$$X_-(\omega) = -\min(X(\omega), 0).$$

We observe that these functions are both measurable (and therefore also random variables), and that $\mathbf{E}[X] = \mathbf{E}[X_+] - \mathbf{E}[X_-]$. Finally, we observe that if $X : \Omega \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ is a non-negative random variable then

$$\mathbf{E}[X] = \int_0^{\infty} \Pr[X > x] dx.$$

Similarly, if $X : \Omega \to \mathbb{R}_{\geq 0} \cup \{-\infty\}$ is a non-positive random variable then

$$\mathbf{E}[X] = -\int_{-\infty}^{0} \Pr[X < x] dx.$$

Applying the definitions and observations of the previous paragraph, we have the following derivation:

$$\begin{aligned}
&\mathbf{E}_{Y\sim g}[\phi(Y)]-\mathbf{E}_{Y\sim h}[\phi(Y)] \\
&\quad= (\mathbf{E}_{Y\sim g}[\phi(Y)_+] - \mathbf{E}_{Y\sim g}[\phi(Y)_-]) - (\mathbf{E}_{Y\sim h}[\phi(Y)_+] - \mathbf{E}_{Y\sim h}[\phi(Y)_-]) \\
&\quad= (\mathbf{E}_{Y\sim g}[\phi(Y)_+] + \mathbf{E}_{Y\sim g}[-\phi(Y)_-]) - (\mathbf{E}_{Y\sim h}[\phi(Y)_+] + \mathbf{E}_{Y\sim h}[-\phi(Y)_-]) \\
&\quad= \left(\int_0^\infty \Pr_{Y\sim g}[\phi(Y)_+ > x]dx + \int_{-\infty}^0 \Pr_{Y\sim g}[-\phi(Y)_- < x]dx\right) \\
&\qquad - \left(\int_0^\infty \Pr_{Y\sim h}[\phi(Y)_+ > x]dx + \int_{-\infty}^0 \Pr_{Y\sim h}[-\phi(Y)_- < x]dx\right) \\
&\quad= \left(\int_0^\infty \Pr_{Y\sim g}[\phi(Y) > x]dx + \int_{-\infty}^0 \Pr_{Y\sim g}[\phi(Y) < x]dx\right) \\
&\qquad - \left(\int_0^\infty \Pr_{Y\sim h}[\phi(Y) > x]dx + \int_{-\infty}^0 \Pr_{Y\sim h}[\phi(Y) < x]dx\right) \\
&\quad= \int_0^\infty \Pr_{Y\sim g}[\phi(Y) > x] - \Pr_{Y\sim h}[\phi(Y) > x]dx \\
&\qquad + \int_{-\infty}^0 \Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]dx \\
&\quad= \int_0^\infty (1 - \Pr_{Y\sim g}[\phi(Y) < x]) - (1 - \Pr_{Y\sim h}[\phi(Y) < x])dx \\
&\qquad + \int_{-\infty}^0 \Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]dx \\
&\quad= \int_0^\infty \Pr_{Y\sim h}[\phi(Y) < x] - \Pr_{Y\sim g}[\phi(Y) < x])dx \\
&\qquad + \int_{-\infty}^0 \Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]dx \\
&\quad\le \int_0^\infty |\Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]|\, dx \\
&\qquad + \int_{-\infty}^0 |\Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]|\, dx \\
&\quad= \int_{-\infty}^\infty |\Pr_{Y\sim g}[\phi(Y) < x] - \Pr_{Y\sim h}[\phi(Y) < x]|\, dx.
\end{aligned}$$

A symmetric argument shows that

$$\mathbf{E}_{Y\sim h}[\phi(Y)] - \mathbf{E}_{Y\sim g}[\phi(Y)] \le \int_{-\infty}^\infty |\Pr_{Y\sim h}[\phi(Y) < x] - \Pr_{Y\sim g}[\phi(Y) < x]|\, dx,$$

concluding the proof.

B.3. Proof of Lemma 14

Recall that $z = \ln(100n^4/\tau^2)$, $S = L_{f_0}(M_{f_0}e^{-z})$, and $p_{\min} = M_{f_0}/(100n^4/\tau^2)$. Note that for any $x \in S$, we have $f_0(x) \ge p_{\min}$ by construction. Since we have conditioned on the event of

Corollary 9 holding, it follows that for each $i \in [n]$, $f_0(X_i) \geq p_{\min}$. Therefore, letting $\rho \stackrel{\text{def}}{=} p_{\min}$, we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\ln(f_0(X_i)) - \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right]\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\ln(\max(f_0(X_i),\rho)) - \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\ln(\max(f_0(X_i),\rho)) - \mathbf{E}_{X\sim f_0}\left[\ln(\max(f_0(X),\rho))\right]\right| \\
&\quad + \left|\mathbf{E}_{X\sim f_0}\left[\ln(\max(f_0(X),\rho))\right] - \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\ln(\max(f_0(X_i),\rho)) - \mathbf{E}_{X\sim f_0}\left[\ln(\max(f_0(X),\rho))\right]\right| \\
&\quad + \int_{-\infty}^{\ln\rho}\Pr[\ln f_0(X)\leq T]dT\,. \qquad (29)
\end{aligned}$$

By Hoeffding's inequality we have

$$\begin{aligned}
&\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}\ln(\max(f_0(X_i),\rho)) - \mathbf{E}_{X\sim f_0}\left[\ln(\max(f_0(X),\rho))\right]\right| > \frac{\epsilon}{16}\right] \\
&\leq 2\exp\left(\frac{-2n^2(\epsilon/16)^2}{n\cdot(\ln M_{f_0} - \ln\rho)^2}\right) \\
&\leq 2\exp\left(\frac{-n\epsilon^2/16^2}{(\ln(100n^4/\tau^2))^2}\right) \\
&\leq \tau/3\,. \qquad \text{(since } n \geq N_1\text{)} \qquad (30)
\end{aligned}$$

Next we have

$$\begin{aligned}
\int_{-\infty}^{\ln\rho}\Pr_{X\sim f_0}[\ln f_0(X)\leq T]dT &\leq \int_0^{\infty}\Pr_{X\sim f_0}[\ln f_0(X)\leq \ln\rho - y]dy \qquad \text{(setting } y = \ln\rho - T\text{)} \\
&\leq \int_0^{\infty} O(d)^d(\rho/M_{f_0})^{1/2}e^{-y/2}dy \qquad \text{(by Lemma 8)} \\
&= \int_0^{\infty} O(d)^d\frac{\tau}{10n^2}e^{-y/2}dy \qquad (\rho = M_{f_0}/(100n^4/\tau^2)) \\
&\leq 2\cdot O(d)^d\frac{\tau}{10n^2} \\
&\leq \epsilon/16. \qquad \text{(since } n \geq N_1\text{)} \qquad (31)
\end{aligned}$$

By applying (30) and (31) to bound (29) from above, with probability at least $1-\tau/3$ we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\ln f_0(X_i) - \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right]\right| \leq \epsilon/8\,,$$

which concludes the proof.

B.4. Proof of Lemma 15

Suppose there exists $x \in \mathbb{R}^d \setminus C$ such that $\hat{f}_n(x) > 0$. Then, we have that $L_{\hat{f}_n}(\hat{f}_n(x)) \setminus C \neq \emptyset$ and thus $\int_{\mathbb{R}^d \setminus C} \hat{f}_n(x)dx > 0$. From this, it follows that $\int_C \hat{f}_n(x)dx < 1$, and so there exists some $\alpha > 1$ such that $\alpha \int_C \hat{f}_n(x)dx = 1$. Let $\hat{g}_n : C \to \mathbb{R}$ be such that $\hat{g}_n = \alpha \cdot \hat{f}_n|_C$. Since C is a convex set and $\int_C \hat{g}_n(x)dx = 1$, we have that $\hat{g}_n$ is a log-concave density. Observe that

$$\frac{1}{n}\sum_{i=1}^{n} \log(\hat{g}_n(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \log(\alpha \hat{f}_n(X_i)) > \frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_n(X_i)) \,, \tag{32}$$

where we used that $\alpha > 1$. By definition, $\hat{f}_n$ maximizes $\frac{1}{n}\sum_{i=1}^{n} \log(f(X_i))$ over all log-concave densities f, which contradicts (32). Therefore, for all $x \in \mathbb{R}^d \setminus C$, we have that $\hat{f}_n(x) = 0$.

B.5. Proof of Lemma 18

This lemma holds because for a density f with a large maximum value M_f, f is small outside a set of small volume, and most of the samples drawn from f_0 will be outside this set. Let

$$\gamma = \exp\left(2\left(\frac{1}{n}\sum_{i=1}^{n} \ln f_0(X_i) - \frac{1}{2}\ln M_f - 1\right)\right)$$

and

$$A = L_f(\gamma).$$

If we have that $\mathrm{vol}(A) \cdot M_{f_0} \leq 1/3$, then it follows that $f_0(A) \leq 1/3$. Since f is log-concave, A is a convex set, and since we condition on Corollary 11 holding, we have with probability 1 that $|f_0(A) - f_n(A)| < \delta < 1/6$. Therefore, we have that $f_n(A) < 1/2$, in which case at least $1/2$ of the samples $X_1, \ldots, X_n$ are not contained within A. Thus, we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \ln f(x) &\leq \frac{1}{2}\ln\gamma + \frac{1}{2}\ln M_f \\
&= \frac{1}{2}\cdot 2\left(\frac{1}{n}\sum_{i=1}^{n} \ln f_0(X_i) - \frac{1}{2}\ln M_f - 1\right) + \frac{1}{2}\ln M_f \\
&< \frac{1}{n}\sum_{i=1}^{n} \ln f_0(X_i).
\end{aligned}$$

Now we check to see how large M_f must be to ensure that $\mathrm{vol}(A) \cdot M_{f_0} \leq 1/3$. We have that

$$\begin{aligned}
\mathrm{vol}(A) \cdot M_{f_0} &= \mathrm{vol}(L_f(\gamma)) \cdot M_{f_0} \\
&= \mathrm{vol}\left(L_f\left(M_f \cdot \exp\left(\frac{2}{n}\sum_{i=1}^{n} \ln f_0(X_i) - 2 - 2\ln M_f\right)\right)\right) \cdot M_{f_0} \\
&\leq \frac{M_{f_0}}{M_f} \cdot O\left(\left(2 - \frac{2}{n}\sum_{i=1}^{n} \ln f_0(X_i) + 2\ln M_f\right)^d\right) \,. \qquad \text{(by Lemma 8)}
\end{aligned}$$

Since we condition on the event of Lemma 14 holding, we have with probability 1 that

$$\frac{1}{n}\sum_{i=1}^{n} \ln f_0(X_i) \geq \mathbf{E}_{X\sim f_0}\left[\ln f_0(X)\right] - \epsilon \geq \ln p_{\min} - \epsilon,$$

and so we have that

$$\begin{aligned}\mathrm{vol}(A)\cdot M_{f_0} &\leq \frac{M_{f_0}}{M_f}\cdot O\left(\left(2+2\ln M_f - 2\ln p_{\min} + 2\epsilon\right)^d\right)\\ &< \frac{M_{f_0}}{M_f}\cdot O\left(\left(2\ln M_f - 2\ln M_{f_0} + 3\ln(n^4 100/\tau^2)\right)^d\right).\end{aligned}$$

The following claim follows by a simple calculation:

Claim 26 *If* $\ln(M_f/M_{f_0}) \geq 3\ln(100n^4/\tau^2)$*, then* $\mathrm{vol}(A)\cdot M_{f_0} \leq 1/3$.

Proof Recall that

$$\begin{aligned}\mathrm{vol}(A)\cdot M_{f_0} &\leq \frac{M_{f_0}}{M_f}\cdot O\left(\left(2+2\ln M_f - 2\ln p_{\min} + 2\epsilon\right)^d\right)\\ &< \frac{M_{f_0}}{M_f}\cdot O\left(\left(2\ln M_f - 2\ln M_{f_0} + 3\ln(n^4 100/\tau^2)\right)^d\right).\end{aligned}$$

We search for M_f such that $\mathrm{vol}(A)\cdot M_{f_0} \leq 1/3$. It is sufficient for M_f to satisfy, for some constant $c > 1$,

$$\begin{aligned}M_{f_0}/M_f \cdot c\left(2\ln M_f - 2\ln M_{f_0} + 3\ln(n^4 100/\tau^2)\right)^d &\leq 1/3\\ \ln\left((M_{f_0}/M_f)\cdot c\left(2\ln(M_f/M_{f_0}) + 3\ln(n^4 100/\tau^2)\right)^d\right) &\leq \ln(1/3)\\ \ln\left(M_{f_0}/M_f\right) + \ln c + \ln\left(\left(2\ln(M_f/M_{f_0}) + 3\ln(n^4 100/\tau^2)\right)^d\right) &\leq \ln(1/3)\\ \ln\left(\left(2\ln(M_f/M_{f_0}) + 3\ln(n^4 100/\tau^2)\right)^d\right) + \ln(3c) &\leq \ln\left(M_f/M_{f_0}\right)\\ d\ln\left(2\ln(M_f/M_{f_0}) + 3\ln(n^4 100/\tau^2)\right) + \ln(3c) &\leq \ln\left(M_f/M_{f_0}\right). \qquad (33)\end{aligned}$$

If we have M_f such that $\ln(M_f/M_{f_0}) \geq 3\ln(n^4 100/\tau^2)$, and a sufficiently large constant is chosen for N_1 so that $\ln(3c) \leq \ln(n^4 100/\tau^2)$, then (33) becomes

$$d\ln\left(3\ln(M_f/M_{f_0})\right) \leq 2\ln\left(M_f/M_{f_0}\right). \qquad (34)$$

The next inequality is equivalent to (34) :

$$\left(3\ln(M_f/M_{f_0})\right)^{d/2} \leq M_f/M_{f_0}$$

We note that the derivative of $(3\ln x)^{d/2}$ is

$$\frac{3^{d/2} d(\ln x)^{d/2-1}}{2x}.$$

We also note that for $x = (3d)^{d/2+1}(\ln(9d))^{d/2+1}$ we have that

$$\frac{3^{d/2}d(\ln x)^{d/2-1}}{2x} \leq 1.$$

and

$$\begin{aligned}(3\ln x)^{d/2} &= [3(d/2+1)\ln(9d\ln(9d)]^{d/2} \\ &= 3^{d/2}\cdot d^{d/2}\cdot[2\ln(9d)]^{d/2} \\ &\leq x.\end{aligned}$$

Therefore, assuming sufficiently large constants are chosen in the definition of N_1, if

$$\ln(M_f/M_{f_0}) \geq 3\ln(n^4 100/\tau^2)$$

then $\mathrm{vol}(A)\cdot M_{f_0} \leq 1/3$. ■

Therefore, for $\ln(M_f/M_{f_0}) \geq 3\ln(100n^4/\tau^2)$ we have that $\frac{1}{n}\sum_{i=1}^n \ln f(X_i) < \frac{1}{n}\sum_i \ln f_0(x_i)$ and

$$\ln M_f - \ln p_{\min} = \ln(M_f/M_{f_0}) + \ln(100n^4/\tau^2) \geq 4\ln(100n^4/\tau^2)\ ,$$

concluding the proof.

B.6. Proof of Claim 22

By Lemma 20 we have that the VC dimension of $\mathcal{A}$ is $V \leq 2(d+1)H\ln((d+1)H)$, and so $V \leq (10\kappa)^{(d+1)/2}d^{(d+5)/2}(\ln(100n^4/\tau^2))^d/\delta)^{(d+1)/2}$. Noting that $\mathbf{E}[|f_0(T)-f_n(T)|] \leq \mathbf{E}[||f_0-f_n||_{\mathcal{A}}]$, by Theorem 4 we get that

$$\begin{aligned}\mathbf{E}[|f_0(T)-f_n(T)|] &\leq \sqrt{\frac{O(V)}{n}} \\ &\leq \sqrt{\frac{O\left((10\kappa)^{(d+1)/2}d^{(d+5)/2}(\ln(100n^4/\tau^2))^d/\delta)^{(d+1)/2}\right)}{n}}.\end{aligned}$$

For the next part we want that $\mathbf{E}[|f_0(T)-f_n(T)|] \leq \delta/10$. This holds when

$$n = \Omega\left((d/\epsilon)(\ln(100n^4/\tau^2))^{(d+1)}\right)^{(d+5)/2}$$

If $n \geq b\left(c^{d+1}(d^{(2d+3)}/\epsilon)(\ln(d^{(d+1)}/(\epsilon\tau)))^{(d+1)}\right)^{(d+5)/2}$ for some constants $b > 1, c \geq 100\ln c$, then we have

$$\begin{aligned}(d/\epsilon)(\ln(100n^4/\tau^2))^{(d+1)} &\leq (d^{(2d+3)}/\epsilon)(100\ln c)^{(d+1)}\left(\ln(d^{(d+1)}/(\epsilon\tau))\right)^{(d+1)} \\ &\leq c^{d+1}(d^{(2d+3)}/\epsilon)(\ln(d^{(d+1)}/(\epsilon\tau)))^{(d+1)}\end{aligned}$$

and therefore $n = \Omega\left((d/\epsilon)(\ln(100n^4/\tau^2)^{(d+1)})\right)^{(d+5)/2}$ as desired. Therefore, for $n \geq N_2$ we have

$$\mathbf{E}[|f_0(T)-f_n(T)|] \leq \delta/10. \tag{35}$$

B.7. Proof of Lemma 23

Consider an arbitrary set T of t points in $\mathbb{R}^d$. We wish to bound the number of possible distinct sets that can be obtained by the intersection of T with a set in $\mathcal{A}_{H,L}$. We note that $\mathcal{A}_{H,L}$ can also be constructed in the following manner: Take an arrangement consisting of at most $H \cdot L$ hyperplanes. This arrangement partitions $\mathcal{R}^d$ into a set of components. Then, the union of subsets of these components are elements of $\mathcal{A}_{H,L}$. Any halfspace can be perturbed, without changing its intersection with T, so that its boundary intersects $d'+1$ points in T, where $d' \leq d$ is the dimension of the affine subspace spanned by T. Any such subset uniquely determines the intersection of the halfspace with T. Therefore, the number of possible intersections with a set of size t is at most $O(t)^d$. It follows then that the number of possible intersections of any $A \in \mathcal{A}_{H,L}$ and any set of size t is at most $(O(t)^d)^{LH} \leq O(t)^{dLH}$. If $\mathcal{A}$ has VC dimension t, then is must be that $O(t)^{dLH} \geq 2^t$, and therefore $t/\log(t) = O(dLH)$.

B.8. Proof of Claim 24

Recalling that $L = \ln(100n^4/\tau^2)$ and $H = (10\kappa d/\delta)^{(d-1)/2}$, we have that

$$\begin{aligned} V/\ln(V) &= O\left(d \cdot \ln(100n^4/\tau^2) \cdot (10\kappa d/\delta)^{(d-1)/2}\right) \\ &= O\left((10\kappa)^{(d-1)/2} d^{(d+1)/2} \ln(100n^4/\tau^2)/\delta^{(d-1)/2}\right). \end{aligned} \tag{36}$$

We note that

$$\begin{aligned} \ln\left((10\kappa)^{(d-1)/2} d^{(d+3)/2} (\ln(100n^4/\tau^2))^2/\delta^{(d-1)/2}\right) &\leq \frac{d-1}{2} \ln\left((10\kappa) d^3 (\ln(100n^4/\tau^2))^6/\delta\right) \\ &\leq d \ln\left((10\kappa) d^3 (\ln(100n^4/\tau^2))^6/\delta\right) \\ &\leq cd \ln(\ln(100n^4/\tau^2)) \end{aligned}$$

for some sufficiently large constant c. Therefore, letting

$$V = O\left((10\kappa)^{(d-1)/2} d^{(d+3)/2} (\ln(100n^4/\tau^2))^2/\delta^{(d-1)/2}\right)$$

satisfies (36). Therefore, we have that

$$\sqrt{\frac{\alpha V}{n}} = \sqrt{\frac{\alpha \cdot O\left((10\kappa)^{(d-1)/2} d^{(d+3)/2} (\ln(100n^4/\tau^2))^2/\delta^{(d-1)/2}\right)}{n}},$$

and thus when

$$n = \Omega\left((10\kappa)^{(d-1)/2} d^{(d+3)/2} (\ln(100n^4/\tau^2))^{(d+7)/2}/\epsilon^{(d+3)/2}\right) \tag{37}$$

we have that $\sqrt{\frac{\alpha V}{n}} \leq \delta/10$. To simplify (37), we note that the

$$d^{(d+3)/2} (\ln(100n^4/\tau^2))^{(d+7)/2}/\epsilon^{(d+3)/2} \leq \left((d/\epsilon)(\ln(100n^4/\tau^2))^2\right)^{(d+3)/2}.$$

Thus, if we let $n = (c(d^2/\epsilon)(\ln(d/(\epsilon\tau)))^3)^{(d+3)/2}$ for some large constant c, then we have that

$$\begin{aligned}\ln(100n^4/\tau^2) &= \frac{d+3}{2}\ln(100c(d^2/\epsilon)(\ln(d/(\epsilon\tau)))^3) + \ln(1/\tau^2)\\ &\leq c'd\ln(d/(\epsilon\tau))\end{aligned}$$

for some large constant c'. Thus, assuming a sufficiently large constant is chosen, for $n \geq N_1$ we have that (37) holds, and therefore

$$\sqrt{\frac{\alpha V}{n}} \leq \delta/10. \tag{38}$$

B.9. Proof of Claim 25

For any choice of the samples $X_1, \ldots, X_n$, we have

$$\begin{aligned}f_n(C_L) &\geq f_n(C^{\text{in}}) && \text{(since } C^{\text{in}} \subseteq C_L\text{)}\\ &\geq f_0(C^{\text{in}}) - |f_0(C^{\text{in}}) - f_n(C^{\text{in}})|\\ &= f_0(C_L) - f_0(C_L \setminus C^{\text{in}}) - |f_0(C^{\text{in}}) - f_n(C^{\text{in}})|\\ &\geq f_0(C_L) - \frac{\delta}{2} - |f_0(C^{\text{in}}) - f_n(C^{\text{in}})|. && \text{(by (22))}\end{aligned} \tag{39}$$

Similarly, we have

$$\begin{aligned}f_n(C_L) &\leq f_n(C^{\text{out}}) && \text{(since } C_L \subseteq C^{\text{out}}\text{)}\\ &\leq f_0(C^{\text{out}}) + |f_0(C^{\text{out}}) - f_n(C^{\text{out}})|\\ &= f_0(C_L) + f_0(C^{\text{out}} \setminus C_L) - |f_0(C^{\text{out}}) - f_n(C^{\text{out}})|\\ &\leq f_0(C_L) + \frac{\delta}{2} + |f_0(C^{\text{out}}) - f_n(C^{\text{out}})|. && \text{(by (24))}\end{aligned} \tag{40}$$

Combining (39) and (40), we therefore have that

$$|f_n(C_L) - f_0(C_L)| \leq \frac{\delta}{2} + \max\left\{|f_0(C^{\text{in}}) - f_n(C^{\text{in}})|, |f_0(C^{\text{out}}) - f_n(C^{\text{out}})|\right\}$$

From this, we therefore have that

$$\sup_{C \in \mathcal{C}} |f_n(C_L) - f_0(C_L)| \leq 7\delta/10,$$

concluding the proof.