arXiv:1605.08188v1 [cs.LG] 26 May 2016

Learning Multivariate Log-concave Distributions

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

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

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

May 27, 2016

Abstract

We study the problem of estimating multivariate log-concave probability density functions. We prove the first sample complexity upper bound for learning log-concave densities on $\mathbb{R}^d$, for all $d \geq 1$. Prior to our work, no upper bound on the sample complexity of this learning problem was known for the case of $d > 3$.

In more detail, we give an estimator that, for any $d \geq 1$ and $\epsilon > 0$, draws $\tilde{O}_d\left((1/\epsilon)^{(d+5)/2}\right)$ samples from an unknown target log-concave density on $\mathbb{R}^d$, and outputs a hypothesis that (with high probability) is ϵ-close to the target, in total variation distance. Our upper bound on the sample complexity comes close to the known lower bound of $\Omega_d\left((1/\epsilon)^{(d+1)/2}\right)$ for this problem.

1 Introduction

1.1 Background and Motivation

The estimation of a probability density function based on observed data is a classical and paradigmatic problem in statistics [Pea95] with a rich history (see e.g., [BBBB72, DG85, Sil86, Sco92, DL01]). This inference task is known as *density estimation* or *distribution learning* and can be informally described as follows: Given a set of samples from an unknown distribution f that is believed to belong to (or be well-approximated by) a given family $\mathcal{D}$, we want to output a hypothesis distribution h that is a good approximation to the target distribution f.

*Part of this work was performed while the author was at the University of Edinburgh. Supported in part by a Marie Curie Career Integration grant.

†Some of this work was performed while visiting the University of Edinburgh.

‡Part of this work was performed while the author was at the University of Edinburgh. Supported in part by a Marie Curie Career Integration grant.

The first and arguably most fundamental goal in density estimation is to characterize the *sample complexity* of the problem in the minimax sense, i.e., the number of samples *inherently* required to obtain a desired accuracy (in expectation or with high probability). In other words, for a given distribution family $\mathcal{D}$ and desired accuracy $\epsilon > 0$, we are interested in obtaining an estimator for $\mathcal{D}$ with a sample complexity upper bound of $N = N(\mathcal{D}, \epsilon)$, and an information-theoretic lower bound showing that *no* estimator for $\mathcal{D}$ can achieve accuracy ϵ with fewer than $\Omega(N)$ samples. The sample complexity of this unsupervised learning problem depends on the *structure* of the underlying family $\mathcal{D}$. Perhaps surprisingly, while density estimation has been studied for several decades, the sample complexity of learning is not yet well-understood for various natural and fundamental distribution families.

In this work, we study the problem of density estimation for the family of log-concave distributions on $\mathbb{R}^d$. A distribution on $\mathbb{R}^d$ is log-concave if the logarithm of its probability density function is a concave function (see Definition 1). Log-concave distributions constitute a rich and attractive non-parametric family that is particularly appealing for modeling and inference [Wal09]. They encompass a range of interesting and well-studied distributions, including uniform, normal, exponential, logistic, extreme value, Laplace, Weibull, Gamma, Chi and Chi-Squared, and Beta distributions (see e.g., [BB05]). Log-concave distributions have been studied in a range of different contexts including economics [An95], statistics and probability theory (see [SW14] for a recent survey), theoretical computer science [LV07], and algebra, combinatorics and geometry [Sta89].

1.2 Our Results and Comparison to Prior Work

The problem of density estimation for log-concave distributions is of central importance in the area of non-parametric shape constrained inference. As such, this problem has received significant attention in the statistics literature, see [CS10, DR09, DW16, CS13, KS14, BD14, HW16] and references therein, and, more recently, in theoretical computer science [CDSS13, CDSS14a, ADLS15, ADK15, CDGR16, DKS16a]. In Section 1.3 we provide a detailed summary of related work. In this subsection, we confine ourselves to describing the prior work that is most relevant to the results of this paper.

We study the following fundamental question:

How many samples are information-theoretically required to learn an arbitrary log-concave density on $\mathbb{R}^d$, up to total variation distance ϵ?

Despite significant amount of work on log-concave density estimation, our understanding of this fundamental question for general dimension d remains surprisingly poor. The only prior work that addresses the $d > 1$ case in the finite sample regime is [KS14]. Specifically, Kim and Samworth [KS14] study this estimation problem with respect to the squared Hellinger distance and obtain the following results:

(1) an information-theoretic sample complexity lower bound of $\Omega_d\left((1/\epsilon)^{(d+1)/2}\right)$ for any $d \in \mathbb{Z}_+$, and

(2) a sample complexity upper bound that is tight (up to logarithmic factors) for $d \leq 3$.

In particular, prior to our work, no sample complexity upper bound was known for $d > 3$.

In this paper, we obtain a sample complexity upper bound of $O_d\left((1/\epsilon)^{(d+5)/2}\right)$, for any $d \in \mathbb{Z}_+$, under the total variation distance. By using the known relation between the total variation and squared Hellinger distances, our sample complexity upper bound immediately implies the same upper bound under the squared Hellinger distance. Moreover, the aforementioned lower bound of [KS14] also directly applies to the total variation distance. Hence, our upper bound is tight up to an $\tilde{O}_d(\epsilon^{-2})$ multiplicative factor.

To formally state our results, we will need some terminology.

Notation and Definitions. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lebesgue measurable function. We will use $f(A)$ to denote $\int_{x \in A} f(x)dx$. A Lebesgue measurable function $f : \mathbb{R}^d \to \mathbb{R}$ is a probability density function (pdf) if $f(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)dx = 1$. The *total variation distance* between pdfs $f, g : \mathbb{R}^d \to \mathbb{R}_+$ is defined as $d_{\mathrm{TV}}(f, g) \stackrel{\mathrm{def}}{=} (1/2) \cdot \|f - g\|_1 = (1/2) \cdot \int_{\mathbb{R}^d} |f(x) - g(x)|dx$.

Definition 1. 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 $\mathbb{R}^d$.

We use the following definition of learning under the total variation distance. We remark that our learning model incorporates adversarial model misspecification, and our proposed estimators are robust in this sense.

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

The main result of this paper is the following theorem:

Theorem 3 (Main Result)**.** *There exists an agnostic learning algorithm for the family $\mathcal{F}_d$ of log-concave densities on $\mathbb{R}^d$ with the following performance guarantee: For any $d \in \mathbb{Z}_+$, $\epsilon > 0$, and any probability density function $f : \mathbb{R}^d \to \mathbb{R}_+$, the algorithm draws $O(d/\epsilon)^{(d+5)/2} \log^2(1/\epsilon)$ samples from f and, with probability at least $9/10$, outputs a hypothesis density $h : \mathbb{R}^d \to \mathbb{R}_+$ such that $d_{\mathrm{TV}}(h, f) \leq 3 \cdot \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{g \in \mathcal{F}_d} d_{\mathrm{TV}}(f, g)$.*

To the best of our knowledge, our estimator provides the first sample complexity guarantees for $\mathcal{F}_d$ for any $d > 3$. With the exception of [KS14], prior work on this problem that provides finite sample guarantees has been confined to the $d = 1$ case. As previously mentioned, Kim and Samworth [KS14] study the case of general dimension d focusing on the squared Hellinger distance. Recall that the squared Hellinger distance is defined as $h^2(f, g) \stackrel{\mathrm{def}}{=} \int_{\mathbb{R}^d} (f^{1/2} - g^{1/2})^2 dx$ and that for any two densities f, g it holds $h^2(f, g) \leq d_{\mathrm{TV}}(f, g) \leq h(f, g)$. Therefore, the sample lower bound of [KS14] also holds under the total variation distance, and our sample upper bound immediately applies under the squared Hellinger distance.

Our proposed estimator establishing Theorem 3 is robust to arbitrary model misspecification under the total variation distance. It should be noted that our estimator does not rely on maximum likelihood, as opposed to most of the statistics literature on this problem. Indeed, as is well-known and easy to show, the maximum likelihood estimator (MLE) is not robust under the total variation distance. In contrast, our estimator relies on the VC inequality [VC71, DL01], a classical result in empirical process theory (see Theorem 4). We remark that the VC inequality has been recently used [CDSS13, CDSS14a, ADLS15] to obtain sharp learning upper bounds for a wide range of *one-dimensional* distribution families, including univariate log-concave densities. As far as we know, ours is the first use of the VC inequality to obtain learning upper bounds for structured distributions in higher dimensions.

1.3 Related Work

The area of density estimation under shape constraints is a classical topic in statistics starting with the pioneering work of Grenander [Gre56] on monotone distributions (see [BBBB72] for an early and [GJ14] for a recent book on the topic). Various structural restrictions have been studied in the literature, starting from monotonicity, unimodality, and concavity [Gre56, Bru58, Rao69, Weg70, HP76, Gro85, Bir87a, Bir87b, Fou97, JW09]. In recent years, there has been a body of work in computer science on this topic with a focus on both sample and computational efficiency [DDS12a, DDS12b, DDO$^+$13, CDSS13, CDSS14a, CDSS14b, ADLS15, DHS15, DKS15a, DKS15b, DDKT16, DKS16b].

During the past decade, density estimation of log-concave densities has been extensively investigated. A line of work in statistics [CS10, DR09, DW16, CS13, BD14] has obtained a complete understanding of the global consistency properties of the maximum likelihood estimator (MLE) for any dimension d. In terms of finite sample bounds, the sample complexity of log-concave density estimation has been characterized for $d = 1$, e.g., it is $\Theta(\epsilon^{-5/2})$ under the variation distance [DL01]. Moreover, it is known [KS14, HW16] that the MLE is sample-efficient in the univariate setting. For general dimension d, [KS14] show that the MLE is nearly-sample optimal under the squared Hellinger distance for $d \leq 3$, and also prove bracketing entropy lower bounds suggesting that the MLE may be sub-optimal for $d > 3$.

A recent line of work in theoretical computer science [CDSS13, CDSS14a, ADLS15, ADK15, CDGR16, DKS16a] studies the $d = 1$ case and obtains efficient estimators under the total variation distance. Specifically, [CDSS14a, ADLS15] design sample-optimal robust estimators for log-concave distributions (among others) based on the VC inequality.

1.4 Technical Overview

In this subsection, we provide a high-level overview of our techniques establishing Theorem 3. Our approach is inspired by the framework introduced in [CDSS13, CDSS14a]. Given a family of structured distributions $\mathcal{D}$ that we want to learn, we proceed as follows: We find an "appropriately structured" distribution family $\mathcal{C}$ that approximates $\mathcal{D}$, in the sense that every density in $\mathcal{D}$ is ϵ-close, in total variation distance, to a density in $\mathcal{C}$. By choosing the family $\mathcal{C}$ appropriately, we can obtain (nearly-)tight sample upper bounds for $\mathcal{D}$ from sample upper bounds for $\mathcal{C}$. Our estimator to achieve this goal (see Lemma 6) leverages the VC inequality.

The aforementioned approach was used in [CDSS13, CDSS14a, ADLS15] to obtain sample-optimal (and computationally efficient) estimators for various *one-dimensional* structured distribution families. In particular, for the family $\mathcal{F}_1$ of univariate log-concave densities, [CDSS13] chooses $\mathcal{C}$ to be the family of densities that are piecewise constant with $O(1/\epsilon)$ interval pieces. Similarly, [CDSS14a, ADLS15] take $\mathcal{C}$ to be the family of densities that are piecewise linear with $O(\epsilon^{-1/2})$ interval pieces.

Our structural approximation result for the high-dimensional case can be viewed as an appropriate generalization of the above one-dimensional results. Specifically, we show that any log-concave density f on $\mathbb{R}^d$ can be ϵ-approximated, in total variation distance, by a function g that is essentially defined by $\tilde{O}_d\left((1/\epsilon)^{(d+1)/2}\right)$ hyperplanes. Once such an approximation has been established, roughly speaking, we exploit the fact that families of sets defined by a small number of hyperplanes have small VC dimension. This allows us to use the VC inequality to learn an approximation to g (and, thus, an approximation to f) from an appropriate number of samples. If V is an upper bound on the corresponding VC dimension, the number of samples needed for this learning task will be $O(V/\epsilon^2)$.

To prove our structural approximation result for log-concave densities f on $\mathbb{R}^d$ we proceed as follows: First, we make use of concentration results for log-concave densities implying that a negligible fraction of f's probability mass comes from points at which f is much smaller than its maximum value. This will allow us to approximate f by a function h that takes only $\tilde{O}_d(1/\epsilon)$ distinct values. Furthermore, the level sets $h^{-1}([x,\infty))$ will be given by the corresponding level sets for f, which are convex. We then use results from convex geometry to approximate each of these convex sets (with respect to volume) by inscribed polytopes with $O_d\left((1/\epsilon)^{(d-1)/2}\right)$ facets. Applying this approximation to each level set of h gives us our function g.

1.5 Organization

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

2 Preliminaries

The VC inequality. For $n \in \mathbb{Z}_+$, we will denote $[n] \stackrel{\text{def}}{=} \{1,\ldots,n\}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lebesgue measurable function. Given a family $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$, we define the *$\mathcal{A}$-norm of f* by $\|f\|_{\mathcal{A}} \stackrel{\text{def}}{=} \sup_{A\in\mathcal{A}} |f(A)|$. We say that a set $X \subseteq \mathbb{R}^d$ is shattered by $\mathcal{A}$ if for every $Y \subseteq X$ there exists $A \in \mathcal{A}$ that satisfies $A \cap X = Y$. The *VC dimension* of a family of sets $\mathcal{A}$ over $\mathbb{R}^d$ 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 ∞.

Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be a probability density function. The empirical distribution $\widehat{f}_n$, corresponding to n independent samples $X_1,\ldots,X_n$ drawn from f, is the probability measure defined by $\widehat{f}_n(A) = (1/n)\cdot\sum_{i=1}^n \mathbf{1}_{X_i\in A}$, for all $A \subseteq \mathbb{R}^d$. The well-known *Vapnik-Chervonenkis (VC) inequality* states the following:

Theorem 4 (VC inequality, [DL01, p.31]). *Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be a probability density function and $\widehat{f}_n$ 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, $\mathbb{E}[\|f - \widehat{f}_n\|_{\mathcal{A}}] \leq C\sqrt{V/n}$, where C is a universal constant.*

Approximation of Convex Sets by Polytopes. There is a large literature on approximating convex sets by polytopes (see, e.g., the surveys [Gru93, Bro08]). We will make essential use of the following theorem that provides a volume approximation by an inscribed polytope with a bounded number of facets:

Theorem 5 ([GMR94, GMR95]). *For any convex body $K \subseteq \mathbb{R}^d$, and n sufficiently large, there exists a convex polytope $P \subseteq K$ with at most n facets such that $\mathrm{vol}(K \setminus P) \leq \frac{Cd}{n^{2/(d-1)}}\mathrm{vol}(K)$, where $C > 0$ is a universal constant.*

3 Proof of Theorem 3

To prove our theorem, we will make essential use of the following general lemma, establishing the existence of a sample-efficient estimator using the VC inequality:

Lemma 6. *Let $\mathcal{D}$ be a family of probability density functions over $\mathbb{R}^d$. Suppose there exists a family $\mathcal{A}$ of subsets of $\mathbb{R}^d$ with VC-dimension V such that the following holds: For any pair of densities $f_1, f_2 \in \mathcal{D}$ we have that $d_{\mathrm{TV}}(f_1, f_2) \leq \|f_1 - f_2\|_{\mathcal{A}} + \epsilon/2$. Then, there exists an agnostic learning algorithm for $\mathcal{D}$ with error guarantee $3 \cdot \mathrm{OPT} + \epsilon$ that succeeds with probability $9/10$ using $O(V/\epsilon^2)$ samples.*

We note that the above generic lemma is implicit in [CDSS14a], and we include a proof in Appendix A.1 for the sake of completeness. The estimator is extremely simple: It draws $n = O(V/\epsilon^2)$ samples from the underlying density f, and output the density $h \in \mathcal{D}$ that minimizes the objective $\|g - \widehat{f}_n\|_{\mathcal{A}}$ over $g \in \mathcal{D}$. The correctness of the estimator relies on the VC inequality.

In view of Lemma 6, to prove Theorem 3 we will establish the existence of a family $\mathcal{A}$ of sets in $\mathbb{R}^d$ whose VC dimension is at most $V = O(d/\epsilon)^{(d+1)/2}\log^2(1/\epsilon)$ that satisfies the condition of the lemma, i.e., for any pair of densities $f_1, f_2 \in \mathcal{F}_d$ it holds that $d_{\mathrm{TV}}(f_1, f_2) \leq \|f_1 - f_2\|_{\mathcal{A}} + \epsilon/2$. Then, Lemma 6 implies that there exists an agnostic estimator for $\mathcal{F}_d$ with sample complexity

$$O(V/\epsilon^2) = O\left(d/\epsilon\right)^{(d+5)/2}\log^2(1/\epsilon) \, .$$

The proof has two main steps. In the first step, we define an appropriately structured family of functions $\mathcal{C}_{d,\epsilon}$ so that an arbitrary log-concave density $f \in \mathcal{F}_d$ can be ϵ-approximated by a function $g \in \mathcal{C}_{d,\epsilon}$. More specifically, each function $g \in \mathcal{C}_{d,\epsilon}$ takes at most $L = O_d((1/\epsilon)\log(1/\epsilon))$ distinct values, and for each $y \geq 0$, the level sets $g^{-1}([y, \infty))$ are a union of intersections of $H = O_d(\epsilon^{-(d-1)/2})$ many halfspaces. We then produce a family $\mathcal{A}_{d,\epsilon}$ of sets so that for $f, g \in \mathcal{C}_{d,\epsilon}$, $d_{\mathrm{TV}}(f, g) = \|f - g\|_{\mathcal{A}_{d,\epsilon}}$ and so that the VC-dimension of $\mathcal{A}_{d,\epsilon}$ is $\tilde{O}(d \cdot L \cdot H)$, which yields the desired result. We proceed with the details below.

We start by formally defining the family of functions $\mathcal{C}_{d,\epsilon}$:

Definition 7. Given $\epsilon > 0$, let $\mathcal{C}_{d,\epsilon}$ be the set of all functions $g : \mathbb{R}^d \to \mathbb{R}$ of the following form:

- We set $L = L(d, \epsilon) \stackrel{\text{def}}{=} \Theta((1/\epsilon)\log(1/\epsilon) + d \log d)$.
- For $i \in [L]$, let $y_i > 0$ and P_i be an intersection of $H \stackrel{\text{def}}{=} \Theta(d/\epsilon)^{(d-1)/2}$ halfspaces in $\mathbb{R}^d$.
- Given $\{(y_i, P_i)\}_{i=1}^L$, we define the function g by

$$g(x) = \begin{cases} \max\{y_i \mid i \in [L] : x \in P_i\} & \text{if } x \in \cup_{j=1}^L P_j \\ 0 & \text{if } x \notin \cup_{j=1}^L P_j \ . \end{cases} \tag{1}$$

Furthermore, we assume that the asymptotic constants used in defining L and H are sufficiently large.

We are now ready to state and prove our first important lemma:

Lemma 8. *For any $f \in \mathcal{F}_d$, and any $\epsilon > 0$, there exists $g \in \mathcal{C}_{d,\epsilon}$ so that $\|f - g\|_1 = O(\epsilon)$.*

Proof. For $y \in \mathbb{R}_+$ and a function $f : \mathbb{R}^d \to \mathbb{R}_+$ we will denote by

$$L_f(y) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d \mid f(x) \geq y\}$$

its level sets. We note that, since f is log-concave, $L_f(y)$ is a convex set for all $y \in \mathbb{R}_+$.

We define the desired approximation in a natural way, by constructing appropriate polyhedral approximations to the level sets $L_f(y)$ for a finite set of y's in a geometric series with ratio $(1 + \epsilon)$. Concretely, given $f \in \mathcal{F}_d$ and $\epsilon > 0$, we define the function $g \in \mathcal{C}_{d,\epsilon}$ as follows: For $i \in [L]$, we set $y_i \stackrel{\text{def}}{=} M_f \cdot (1 - \epsilon)^i$, where M_f will denote the maximum value of f. We then consider the collection of convex sets $L_f(y_i)$, $i \in [L]$, and apply Theorem 5 to approximate each such set by a polytope with an appropriate number of facets. For each $i \in [L]$, Theorem 5, applied for $n = O(d/\epsilon)^{(d-1)/2}$, prescribes that there exists a polytope, P_i, that is the intersection of $H = O(d/\epsilon)^{(d-1)/2}$ many halfspaces in $\mathbb{R}^d$, so that:

(i) $P_i \subseteq L_f(y_i)$, and

(ii) $\text{vol}(P_i) \geq \text{vol}\left(L_f(y_i)\right) \cdot (1 - \epsilon)$.

This defines our function g. It remains to prove that $\|f - g\|_1 = O(\epsilon)$.

We first point out that, by the definition of g, we have that $f(x) \geq g(x)$ for all $x \in \mathbb{R}^d$. This is because, if $g(x) = y_i$, it must be the case that $x \in P_i \subseteq L_f(y_i)$, by condition (i) above, and therefore $f(x) \geq y_i = g(x)$. So, to prove the lemma, it suffices to show that $\int_{\mathbb{R}^d} g(x)dx = 1 - O(\epsilon)$. We start by noting that

$$1 = \int_{\mathbb{R}^d} f(x)dx = \text{vol}\left(\left\{(x, y) \in \mathbb{R}^{d+1} \mid 0 \leq y \leq f(x)\right\}\right) = \int_{\mathbb{R}_+} \text{vol}\left(L_f(y)\right) dy.$$

Similarly, if we denote $L_g(y) = \{x \in \mathbb{R}^d \mid g(x) \geq y\}$, we have that

$$\int_{\mathbb{R}^d} g(x)dx = \int_{\mathbb{R}_+} \text{vol}\left(L_g(y)\right) dy \ .$$

The following claim establishes that the contribution to $\int_{\mathbb{R}^d} f(x)dx$ from the points $x \in \mathbb{R}^d$ with $f(x) \leq y_{L-1}$ is small:

Claim 9. *It holds that* $\int_0^{y_{L-1}} \mathrm{vol}\left(L_f(y)\right) dy \leq \epsilon$.

Proof. We assume without loss of generality that f attains its maximum value, M_f, at $x = \mathbf{0}$. Let $R = L_f\left(\frac{M_f}{e}\right)$. Notice that

$$1 = \int_{\mathbb{R}_+} \mathrm{vol}\left(L_f(y)\right) dy \geq \int_{0 \leq y \leq M_f/e} \mathrm{vol}\left(L_f(y)\right) dy \geq \int_{0 \leq y \leq M_f/e} \mathrm{vol}\left(R\right) dy = \frac{M_f}{e} \cdot \mathrm{vol}\left(R\right) \ ,$$

where we used the fact that $R \subseteq L_f(y)$ since $y \leq M_f/e$. Hence, we have that

$$\mathrm{vol}(R) \leq e/M_f \ .$$

Moreover, we claim that, for $z \geq 1$, by the log-concavity of f we have that

$$L_f(M_f e^{-z}) \subseteq zR \ .$$

Indeed, for $f(x) \geq M_f e^{-z}$, then $f(x/z) \geq f(0)^{(z-1)/z} f(x)^{1/z} \geq M_f/e$. Therefore $x/z \in R$, so $x \in zR$. Hence,

$$\mathrm{vol}\left(L_f(M_f e^{-z})\right) \leq O(z^d/M_f). \tag{2}$$

Recall that, by our definition of L, if we choose sufficiently large asymptotic constants, it holds $y_{L-1} \leq \delta M_f$ for $\delta = \epsilon^2/O(d)^{2d}$. We now have the following sequence of inequalities:

$$\begin{aligned}
\int_0^{y_{L-1}} \mathrm{vol}\left(L_f(y)\right) dy &= \\
&= \int_{\ln(M_f/y_{L-1})}^{\infty} \mathrm{vol}\left(L_f(M_f e^{-z})\right) M_f e^{-z} dz \\
&\qquad \text{(by the change of variable } y = M_f e^{-z}\text{)} \\
&\leq \int_{\ln(1/\delta)}^{\infty} O(z^d e^{-z}) dz \quad \text{(by (2) and the assumption } M_f/y_{L-1} \geq 1/\delta\text{)} \\
&\leq \int_{\ln(1/\delta)}^{\infty} O(d)^d e^{-z/2} dz \qquad \text{(since } e^{z/2} \geq (z/2)^d/d!\text{)} \\
&= O(d)^d \delta^{1/2} \\
&\leq \epsilon \ . \qquad \text{(using the definition of } \delta\text{)}
\end{aligned}$$

This completes the proof of Claim 9. □

We now establish the following crucial claim:

Claim 10. *For* $y_L \leq y \leq y_1$, *we have that* $\mathrm{vol}\left(L_g(y)\right) \geq (1-\epsilon)\mathrm{vol}\left(L_f\left(\frac{y}{1-\epsilon}\right)\right)$.

Proof. Recall that $y_i \stackrel{\text{def}}{=} M_f(1-\epsilon)^i$, $i \in [L]$. Since $y_1 > y_2 > \ldots > y_L$, we can equivalently write (1) as follows:

$$g(x) = \begin{cases} y_i, \text{ where } i = \min\{j \in [L] : x \in P_j\} & \text{if } x \in \cup_{j=1}^L P_j \\ 0 & \text{if } x \notin \cup_{j=1}^L P_j \ . \end{cases} \tag{3}$$

We claim that $L_g(y_i) = \bigcup_{1 \le j \le i} P_j$, $i \in [L]$. Indeed, we can write

$$L_g(y_i) = \{x \in \mathbb{R}^d \mid g(x) \ge y_i\} = \bigcup_{1 \le j \le i} \{x \in \mathbb{R}^d \mid g(x) = y_j\} = \bigcup_{1 \le j \le i} (P_j \setminus \cup_{k<j} P_k) = \bigcup_{1 \le j \le i} P_j \ ,$$

where the second and third equalities follow from (3).

For $y = y_1$, we thus have that

$$\mathrm{vol}(L_g(y_1)) = \mathrm{vol}(P_1) \ge (1-\epsilon)\mathrm{vol}(L_f(y_1)) \ge (1-\epsilon)\mathrm{vol}\left(L_f\left(\frac{y_1}{1-\epsilon}\right)\right) \ ,$$

where the first inequality is implied by (ii), and the second inequality follows from the fact $L_f(y) \supseteq L_f(y')$ whenever $y \le y'$.

For $y_L \le y < y_1$, consider the index $i \in [L-1]$ such that $y_{i+1} \le y < y_i = \frac{y_{i+1}}{1-\epsilon}$. By definition, we have that

$$L_g(y_i) \subseteq L_g(y) \subseteq L_g(y_{i+1}) \ .$$

Recalling that $L_g(y_i) = \bigcup_{1 \le j \le i} P_j$, we obtain $L_g(y) \supseteq P_i$, and therefore

$$\mathrm{vol}(L_g(y)) \ge \mathrm{vol}(P_i) \ge (1-\epsilon)\mathrm{vol}(L_f(y_i)) = (1-\epsilon)\mathrm{vol}\left(L_f\left(\frac{y_{i+1}}{1-\epsilon}\right)\right) \ge (1-\epsilon)\mathrm{vol}\left(L_f\left(\frac{y}{1-\epsilon}\right)\right) \ ,$$

where the second inequality is implied by (ii) and the third inequality uses the fact that $y \ge y_{i+1}$ and the fact $L_f(y) \supseteq L_f(y')$ whenever $y \le y'$. This completes the proof of Claim 10. □

We are now ready to complete the proof. We have the following:

$$\begin{aligned}
\int_{\mathbb{R}^d} g(x)dx &= \int_{y_L}^{y_1} \mathrm{vol}\left(L_g(y)\right) dy \\
&\ge (1-\epsilon) \int_{y_L}^{y_1} \mathrm{vol}\left(L_f\left(\frac{y}{1-\epsilon}\right)\right) dy && \text{(by Claim 10)} \\
&= (1-\epsilon)^2 \cdot \int_{y_L/(1-\epsilon)}^{M_f} \mathrm{vol}\left(L_f(y')\right) dy' \\
&= (1-\epsilon)^2 \cdot \left(\int_0^{M_f} \mathrm{vol}\left(L_f(y)\right) dy - \int_0^{y_{L-1}} \mathrm{vol}\left(L_f(y)\right) dy\right) \\
&\ge (1-\epsilon)^2 \cdot \left(\int_{\mathbb{R}^d} f(x)dx - \epsilon\right) && \text{(by Claim 9)} \\
&= (1-\epsilon)^2 \cdot (1-\epsilon) \\
&= 1 - O(\epsilon) \ .
\end{aligned}$$

The proof of Lemma 8 is now complete. □

We now proceed to define the family of subsets $\mathcal{A}$ and bound from above its VC dimension. In particular, we define $\mathcal{A}$ to be the family of sets that exactly express the differences between two elements of $\mathcal{C}_{d,\epsilon}$:

Definition 11. Define the family $\mathcal{A}_{d,\epsilon}$ of sets in $\mathbb{R}^d$ to be the collection of all sets of the form $\{x \in \mathbb{R}^d : g(x) \geq g'(x)\}$ for some $g, g' \in \mathcal{C}_{d,\epsilon}$. Notice that if $g, g' \in \mathcal{C}_{d,\epsilon}$ then $d_{\mathrm{TV}}(g, g') = \|g - g'\|_{\mathcal{A}_{d,\epsilon}}$.

We show the following lemma:

Lemma 12. *The VC dimension of $\mathcal{A}_{d,\epsilon}$ is at most $O(d/\epsilon)^{(d+1)/2}\log^2(1/\epsilon)$. Furthermore, for $f, f' \in \mathcal{F}_d$, and $c > 0$ is a sufficiently small constant, we have that $d_{\mathrm{TV}}(f, f') \leq \|f - f'\|_{\mathcal{A}_{d,c\epsilon}} + \epsilon/2$.*

Proof. Note that a $g \in \mathcal{C}_{d,\epsilon}$ is determined completely by $L = O((d/\epsilon)\log(d/\epsilon))$ values y_i and $LH = O(d/\epsilon)^{(d+1)/2}\log(1/\epsilon)$ halfspaces used to define the L convex polytopes P_i. We will show that if $g' \in \mathcal{C}_{d,\epsilon}$ is defined by L values y'_i and another set of LH halfspaces, and if $x \in \mathbb{R}^d$, then it is possible to determine whether or not $g(x) \geq g'(x)$ based solely on:

- The relative ordering of the y_i and y'_i.
- Which of the $2LH$ halfspaces x belongs to.

Now consider an arbitrary set T of n 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 $\mathcal{A}_{d,\epsilon}$. By the above, the intersection will be determined by:

- The relative ordering of the $2L$ elements given by the y_i and y'_i.
- The intersections of each of the $2LH$ halfspaces defining g and g' with T.

Note that the number of orderings in question is at most $(2L)!$. Formally, we can write $P_i = \bigcap_{j=1}^{H} P_{i,j}$ for halfspaces $P_{i,j}$, and similarly $P'_i = \bigcap_{j=1}^{H} P'_{i,j}$, where P_i and P'_i appear in the definition of g and g' respectively. We have the following:

Claim 13. *There exist at most $(2L)!$ different $2L$-ary set functions F_k such that for any $g, g' \in \mathcal{C}_{d,\epsilon}$ the set $\{x : g(x) \geq g'(x)\}$ is given by $F_k(P_{1,1}, \ldots, P_{L,H}, P'_{1,1}, \ldots, P'_{L,H})$ for some k. Furthermore, these functions are distributive over intersection, i.e., for all k and $T, S_1, \ldots, S_{2LH} \subseteq \mathbb{R}^d$, we have that $F_k(S_1, \ldots, S_{2LH}) \cap T = F_k(S_1 \cap T, \ldots, S_{2LH} \cap T)$,*

Proof. Note that for a given $x \in \mathbb{R}^d$, we have that $g(x) \geq g'(x)$ if and only if there is an i such that $x \in P_i$ and for all i' with $y'_{i'} \geq y_i$ we have $x \notin P'_{i'}$. That is, $\{x : g(x) \geq g'(x)\} = \bigcup_{i=1}^{L}\left(P_i \setminus \bigcup_{i':y'_{i'}>y_i} P'_{i'}\right)$. In terms of halfspaces, this can be equivalently written as follows:

$$\{x : g(x) \geq g'(x)\} = \bigcup_{i=1}^{L}\left(\bigcap_{j=1}^{H} P_{i,j} \setminus \bigcup_{i':y'_{i'}>y_i}\bigcap_{j=1}^{H} P'_{i',j}\right).$$

Note that, viewed as a function of the halfspaces, the above expression only depends on the relative ordering of the y_i and y'_i. Thus, we can express this as one of at most $(2L)!$ functions of these halfspaces.

Since these functions are defined using only unions, intersections and differences (which all distribute over intersections), so do the F_k. ∎

It is well-known that for any halfspace the number of possible intersections with a set T of size n is at most $O(n)^d$. By Claim 13, for any $A \in \mathcal{A}_{d,\epsilon}$ we have that $A \cap T = F_k(S_1 \cap T, \ldots, S_{2LH} \cap T)$ for halfspaces $S_1, \ldots, S_{2LH}$. There are $(O(n)^d)^{2LH}$ different $2LH$-tuples of intersections of halfspaces with T and at most $(2L)!$ different F_k. Therefore, the number of possible intersections of an element of $\mathcal{A}_{d,\epsilon}$ with T is at most

$$(2L)!O(n)^{2dLH} = \exp(O(d/\epsilon)^{(d+1)/2} \log(1/\epsilon) \log(n)) \; . \tag{4}$$

On the other hand, if $\mathcal{A}_{d,\epsilon}$ has VC dimension n, (4) must be at least 2^n. Therefore, if n is the VC-dimension of $\mathcal{A}_{d,\epsilon}$, we have that

$$n/\log(n) = O(d/\epsilon)^{(d+1)/2} \log(1/\epsilon) \; ,$$

and therefore,

$$n = O(d/\epsilon)^{(d+1)/2} \log^2(1/\epsilon) \; .$$

For the claim comparing the variation distance to $\|\cdot\|_{\mathcal{A}_{d,c\epsilon}}$, note that, by Lemma 8, if c is chosen to be sufficiently small, there exist $g, g' \in \mathcal{C}_{d,c\epsilon}$ so that $d_{\mathrm{TV}}(f,g), d_{\mathrm{TV}}(f',g') \leq \epsilon/8$. We then have that

$$\begin{aligned} d_{\mathrm{TV}}(f,f') &\leq d_{\mathrm{TV}}(f,g) + d_{\mathrm{TV}}(f',g') + d_{\mathrm{TV}}(g,g') \\ &\leq \epsilon/4 + \|g - g'\|_{\mathcal{A}_{d,c\epsilon}} \\ &\leq \epsilon/4 + \|f - f'\|_{\mathcal{A}_{d,c\epsilon}} + d_{\mathrm{TV}}(f,g) + d_{\mathrm{TV}}(f',g') \\ &\leq \|f - f'\|_{\mathcal{A}_{d,c\epsilon}} + \epsilon/2 \; . \end{aligned}$$

This completes the proof of Lemma 12. □

4 Conclusions

In this paper, we gave the first sample complexity upper bound for learning log-concave densities on $\mathbb{R}^d$. Our upper bound agrees with the previously known lower bound up to a multiplicative factor of $\tilde{O}_d(\epsilon^{-2})$. No sample complexity upper bound was previously known for this problem for any $d > 3$.

Our result is a step towards understanding the learnability of log-concave densities in multiple dimensions. A number of interesting open problems remain. We outline the two immediate ones here:

- What is the *optimal* sample complexity of log-concave density estimation? It is a plausible conjecture that the correct answer, under the total variation distance, is $\Theta_d\left((1/\epsilon)^{d/2+2}\right)$. We believe that a more sophisticated version of our structural approximation results could give such an upper bound. On the other hand, it seems likely that an adaptation of the construction in [KS14] could yield a matching lower bound.

- Is there a *polynomial time algorithm* (as a function of the sample complexity) to learn log-concave densities on $\mathbb{R}^d$? The estimator underlying this work (Lemma 6) has been previously exploited [CDSS13, CDSS14a, ADLS15] to obtain computationally efficient learning algorithms for $d = 1$ – in fact, running in sample near-linear time [ADLS15]. Obtaining a computationally efficient algorithm for the case of general dimension is a challenging and important open question.

References

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

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

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

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

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

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

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

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

[Bro08] E. M. Bronstein. Approximation of convex sets by polytopes. *Journal of Mathematical Sciences*, 153(6):727–762, 2008.

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

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

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

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

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

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

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

[DDKT16] 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.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

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

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

[DHS15] I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. In *NIPS*, pages 2566–2574, 2015.

[DKS15a] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKS15b] I. Diakonikolas, D. M. Kane, and A. Stewart. Properly learning poisson binomial distributions in almost polynomial time. *CoRR*, 2015.

[DKS16a] I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient Robust Proper Learning of Log-concave Distributions. Arxiv report, 2016.

[DKS16b] 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.

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

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

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

[Fou97] A.-L. Fougères. Estimation de densités unimodales. *Canadian Journal of Statistics*, 25:375–387, 1997.

[GJ14] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

[GMR94] Y. Gordon, M. Meyer, and S. Reisner. Volume approximation of convex sets by polytopes – a constructive method. *Stud. Math.*, 111:81–95, 1994.

[GMR95] Y. Gordon, M. Meyer, and S. Reisner. Constructing a polytope to approximate a convex body. *Geometriae Dedicata*, 57(2):217–222, 1995.

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

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

[Gru93] P. M. Gruber. Aspects of approximation of convex bodies. *Handbook of Convex Geometry*, 1993.

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

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

[JW09] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

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

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

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

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

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

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

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

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

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

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

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

A Omitted Proofs

A.1 Proof of Lemma 6

The estimator is as follows:

(1) Draw $n = O(V/\epsilon^2)$ samples from f;

(2) Output the density $h \in \mathcal{D}$ that minimizes the objective function $\|g - \widehat{f}_n\|_{\mathcal{A}}$ over $g \in \mathcal{D}$.

We now show that the above estimator is an agnostic learning algorithm for $\mathcal{D}$. Let $f^* = \mathrm{argmin}\{d_{\mathrm{TV}}(f, g) \mid g \in \mathcal{D}\}$, i.e., $\mathrm{OPT} = d_{\mathrm{TV}}(f, f^*)$. Note that for any pair of densities f_1, f_2 and any collection of subsets $\mathcal{A}$ we have that $\|f_1 - f_2\|_{\mathcal{A}} \leq d_{\mathrm{TV}}(f_1, f_2)$. By Theorem 4 and Markov's inequality, it follows that with probability at least $9/10$ over the samples drawn from f we have that

$$\|f - \widehat{f}_n\|_{\mathcal{A}} \leq \epsilon/4 \, .$$

Conditioning on this event, we have that

$$\begin{aligned}
d_{\mathrm{TV}}(h, f) &\leq d_{\mathrm{TV}}(f, f^*) + d_{\mathrm{TV}}(f^*, h) \\
&\leq \mathrm{OPT} + \|f^* - h\|_{\mathcal{A}} + \epsilon/2 && (\text{since } f^*, h \in \mathcal{D}) \\
&\leq \mathrm{OPT} + \|f^* - \widehat{f}_n\|_{\mathcal{A}} + \|h - \widehat{f}_n\|_{\mathcal{A}} + \epsilon/2 \\
&\leq \mathrm{OPT} + 2 \cdot \|f^* - \widehat{f}_n\|_{\mathcal{A}} + \epsilon/2 && (\text{since } \|h - \widehat{f}_n\|_{\mathcal{A}} \leq \|f^* - \widehat{f}_n\|_{\mathcal{A}}) \\
&\leq \mathrm{OPT} + 2 \cdot \|f^* - f\|_{\mathcal{A}} + 2 \cdot \|f - \widehat{f}_n\|_{\mathcal{A}} + \epsilon/2 \\
&\leq \mathrm{OPT} + 2 \cdot d_{\mathrm{TV}}(f^*, f) + 2 \cdot \|f - \widehat{f}_n\|_{\mathcal{A}} + \epsilon/2 \\
&\leq \mathrm{OPT} + 2 \cdot \mathrm{OPT} + 2 \cdot \epsilon/4 + \epsilon/2 \\
&= 3\mathrm{OPT} + \epsilon \, .
\end{aligned}$$

This completes the proof of the lemma.