Density Estimation for Shift-Invariant Multidimensional Distributions

Anindya De[1]
EECS Department, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA
anindya@eecs.northwestern.edu

Philip M. Long
Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
plong@google.com

Rocco A. Servedio[2]
Department of Computer Science, Columbia University, 500 W. 120th Street, Room 450, New York, NY 10027, USA
rocco@cs.columbia.edu

Abstract

We study density estimation for classes of *shift-invariant* distributions over $\mathbb{R}^d$. A multidimensional distribution is "shift-invariant" if, roughly speaking, it is close in total variation distance to a small shift of it in any direction. Shift-invariance relaxes smoothness assumptions commonly used in non-parametric density estimation to allow jump discontinuities. The different classes of distributions that we consider correspond to different rates of tail decay.

For each such class we give an efficient algorithm that learns any distribution in the class from independent samples with respect to total variation distance. As a special case of our general result, we show that d-dimensional shift-invariant distributions which satisfy an exponential tail bound can be learned to total variation distance error ε using $\tilde{O}_d(1/\varepsilon^{d+2})$ examples and $\tilde{O}_d(1/\varepsilon^{2d+2})$ time. This implies that, for constant d, multivariate log-concave distributions can be learned in $\tilde{O}_d(1/\varepsilon^{2d+2})$ time using $\tilde{O}_d(1/\varepsilon^{d+2})$ samples, answering a question of [29]. All of our results extend to a model of *noise-tolerant* density estimation using Huber's contamination model, in which the target distribution to be learned is a $(1-\varepsilon, \varepsilon)$ mixture of some unknown distribution in the class with some other arbitrary and unknown distribution, and the learning algorithm must output a hypothesis distribution with total variation distance error $O(\varepsilon)$ from the target distribution. We show that our general results are close to best possible by proving a simple $\Omega\left(1/\varepsilon^d\right)$ information-theoretic lower bound on sample complexity even for learning bounded distributions that are shift-invariant.

2012 ACM Subject Classification Theory of computation → Unsupervised learning and clustering

Keywords and phrases Density estimation, unsupervised learning, log-concave distributions, non-parametrics

Digital Object Identifier 10.4230/LIPIcs.ITCS.2019.28

Related Version A full version of the paper is available at [20], `https://arxiv.org/abs/1811.03744`.

[1] Supported by NSF grant CCF-1814706.
[2] Supported by NSF grants CCF-1563155, IIS-1838154 and CCF-1814873.

1 Introduction

In multidimensional density estimation, an algorithm has access to independent draws from an unknown target probability distribution over $\mathbb{R}^d$, which is typically assumed to belong to or be close to some class of "nice" distributions. The goal is to output a hypothesis distribution which with high probability is close to the target distribution. A number of different distance measures can be used to capture the notion of closeness; in this work we use the total variation distance (also known as the "statistical distance" and equivalent to the L_1 distance). This is a well studied framework which has been investigated in detail, see e.g. the books [23, 24].

Multidimensional density estimation is typically attacked in one of two ways. In the first general approach a parameterized hypothesis class is chosen, and a setting of parameters is chosen based on the observed data points. This approach is justified given the belief that the parameterized class contains a good approximation to the distribution generating the data, or even that the parameterized class actually contains the target distribution. See [14, 33, 39] for some well-known multidimensional distribution learning results in this line.

In the second general approach a hypothesis distribution is constructed by "smoothing" the empirical distribution with a kernel function. This approach is justified by the belief that the target distribution satisfies some smoothness assumptions, and is more appropriate when studying distributions that do not have a parametric representation. The current paper falls within this second strand.

The most popular smoothness assumption is that the distribution has a density that belongs to a Sobolev space [42, 6, 30, 24]. The simplest Sobolev space used in this context corresponds to having a bound on the average of the partial first "weak derivatives" of the density; other Sobolev spaces correspond to bounding additional derivatives. A drawback of this approach is that it does not apply to distributions whose densities have jump discontinuities. Such jump discontinuities can arise in various applications, for example, when objects under analysis must satisfy hard constraints.

To address this, some authors have used the weaker assumption that the density belongs to a Besov space [7, 22, 38, 43, 3]. In the simplest case, this allows jump discontinuities as long as the function does not change very fast on average. The precise definition, which is quite technical (see [22]), makes reference to the effect on a distribution of shifting the domain by a small amount.

The densities we consider. In this paper we analyze a clean and simple smoothness assumption, which is a continuous analog of the notion of shift-invariance that has recently been used for analyzing the learnability of various types of discrete distributions [5, 16, 21]. The assumption is based on the *shift-invariance of f in direction v at scale κ*, which, for a density f over $\mathbb{R}^d$, a unit vector $v \in \mathbb{R}^d$, and a positive real value κ, we define to be

$$\mathrm{SI}(f, v, \kappa) \stackrel{\text{def}}{=} \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} |f(x + \kappa' v) - f(x)| \, dx.$$

We define the quantity $\mathrm{SI}(f, \kappa)$ to be the worst case of $\mathrm{SI}(f, v, \kappa)$ over all directions v, i.e. $\mathrm{SI}(f, \kappa) \stackrel{\text{def}}{=} \sup_{v:\|v\|_2=1} \mathrm{SI}(f, v, \kappa)$. For any constant c, we define the class of densities $\mathcal{C}_{\mathrm{SI}}(c, d)$ to consist of all d-dimensional densities f with the property that $\mathrm{SI}(f, \kappa) \leq c$ for all $\kappa > 0$.

Our notion of shift-invariance provides a quantitative way of capturing the intuition that the density f changes gradually on average in every direction. Several natural classes fit nicely into this framework; for example, we note that d-dimensional standard normal distributions are easily shown to belong to $\mathcal{C}_{\mathrm{SI}}(1, d)$. As another example, we will show later that any d-dimensional isotropic log-concave distribution belongs to $\mathcal{C}_{\mathrm{SI}}(O_d(1), d)$.

Many distributions arising in practice have light tails, and distributions with light tails can in general be learned more efficiently. To analyze learning shift-invariant distributions in a manner that takes advantage of light tails when they are available, while accommodating heavier tails when necessary, we define classes with different combinations of shift-invariant and tail behavior. Given a nonincreasing function $g : \mathbb{R}^+ \to [0,1]$ which satisfies $\lim_{t\to+\infty} g(t) = 0$, we define the class of densities $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ to consist of those $f \in \mathcal{C}_{\mathrm{SI}}(c,d)$ which have the additional property that for all $t > 0$, it holds that $\mathbf{Pr}_{\boldsymbol{x}\leftarrow f}\left[||\boldsymbol{x} - \mu|| > t\right] \leq g(t)$, where $\mu \in \mathbb{R}^d$ is the mean of the distribution f.

As motivation for its study, we feel that $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ is a simple and easily understood class that exhibits an attractive tradeoff between expressiveness and tractability. As we show, it is broad enough to include distributions of central interest such as multidimensional isotropic log-concave distributions, but it is also limited enough to admit efficient density estimation algorithms.

Our density estimation framework. We recall the standard notion of density estimation with respect to total variation distance. Given a class $\mathcal{C}$ of densities over $\mathbb{R}^d$, a density estimation algorithm for $\mathcal{C}$ is given access to i.i.d. draws from f, where $f \in \mathcal{C}$ is the unknown *target density* to be learned. For any $f \in \mathcal{C}$, given any parameter $\varepsilon > 0$, after making some number of draws depending on d and ε the density estimation algorithm must output a description of a hypothesis density h over $\mathbb{R}^d$ which, with high probability over the draws from f, satisfies $d_{\mathrm{TV}}(f,h) \leq \varepsilon$. It is of interest both to bound the *sample complexity* of such an algorithm (the number of draws from f that it makes) and its running time. In the full version of this paper [20], we show that our learning results can be extended to a challenging model of *noise-tolerant* density estimation for a class $\mathcal{C}$.

1.1 Results

Our main positive result is a general algorithm which efficiently learns any class $\mathcal{C}_{\mathrm{SI}}(c,d,g)$. Given a constant c and a tail bound g, we show that any distribution in the class $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ can be learned to any error $O(\varepsilon)$ with a sample complexity that depends on c, g, ε and d. The running time of our algorithm is roughly quadratic in the sample complexity, and the sample complexity is $O_{c,d,g}(1) \cdot \left(\frac{1}{\varepsilon}\right)^{d+2}$ (see Theorem 12 in Section 4 for a precise statement of the exact bound). These bounds on the number of examples and running time do not depend on which member of $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ is being learned.

Application: Learning multivariate log-concave densities. A multivariate density function f over $\mathbb{R}^d$ is said to be *log-concave* if there is an upper semi-continuous concave function $\phi : \mathbb{R}^d \to [-\infty, \infty)$ such that $f(x) = e^{\phi(x)}$ for all x. Log-concave distributions arise in a range of contexts and have been well studied; see [11, 12, 3, 1, 9, 25] for work on density estimation of univariate (discrete and continuous) log-concave distributions. In the multivariate case, [35] gave a sample complexity lower bound (for squared Hellinger distance) which implies that $\Omega(1/\varepsilon^{(d+1)/2})$ samples are needed to learn d-dimensional log-concave densities to error ε. More recently, [29] established the first finite sample complexity upper bound for multivariate log-concave densities, by giving an algorithm that learns any d-dimensional log-concave density using $\tilde{O}_d(1/\varepsilon^{(d+5)/2})$ samples. The algorithm of [29] is not computationally efficient, and indeed, Diakonikolas et al. ask if there is an algorithm with running time polynomial in the sample complexity, referring to this as "a challenging and important open question." A subsequent (and recent) work of Carpenter et al. [10] showed that the maximum likelihood estimator (MLE) is statistically efficient (i.e., achieves near optimal sample complexity).

We show that multivariate log-concave densities can be learned in polynomial time as a special case of our main algorithmic result. We establish that any d-dimensional near-isotropic log-concave density is $O_d(1)$-shift-invariant. Together with well-known tail bounds on d-dimensional log-concave densities, this easily yields that any d-dimensional near-isotropic log-concave density belongs to $\mathcal{C}_{\mathrm{SI}}(c, d, g)$ where the tail bound function g is inverse exponential. Theorem 12 then immediately gives a $\tilde{O}_d(1/\varepsilon^{2d+2})$-time algorithm for learning near-isotropic log-concave densities. Adding a preprocessing step to reduce to the near-isotropic case yields an algorithm that works for all log-concave densities.

While our sample complexity is quadratically larger than the optimal sample complexity for learning log-concave distributions (from [29]), such *computational-statistical* tradeoffs are in fact quite common (see, for example, the work of [8] which gives a faster algorithm for learning Gaussian mixture models by using more samples).

A lower bound. We also prove a simple lower bound, showing that any algorithm that learns shift-invariant d-dimensional densities with bounded support to error ε must use $\Omega\left(1/\varepsilon^d\right)$ examples. These densities may be thought of as satisfying the strongest possible rate of tail decay as they have zero tail mass outside of a bounded region (corresponding to $g(t) = 0$ for t larger than some absolute constant). This lower bound shows that a sample complexity of at least $1/\varepsilon^d$ is necessary even for very structured special cases of our multivariate density estimation problem.

1.2 Our approach

For simplicity, and because it is a key component of our general algorithm, we first describe how our algorithm learns an ε-error hypothesis when the target distribution belongs to $\mathcal{C}_{\mathrm{SI}}(c, d)$ and also has *bounded support*: all its mass is on points in the origin-centered ball of radius $1/2$.

In this special case, analyzed in Section 3, our algorithm has two conceptual stages. First, we smooth the density that we are to learn through convolution – this is done in a simple way by randomly perturbing each draw. This convolution uses a kernel that damps the contributions to the density coming from high-frequency functions in its Fourier decomposition; intuitively, the shift-invariance of the target density ensures that the convolved density (which is an average over small shifts of the original density) is close to the original density. In the second conceptual stage, the algorithm approximates relatively few Fourier coefficients of the smoothed density. We show that an inverse Fourier transformation using this approximation still provides an accurate approximation to the target density.[3]

Next, in Section 4, we consider the more general case in which the target distribution belongs to the class $\mathcal{C}_{\mathrm{SI}}(c, d, g)$. Here the high-level idea of our approach is very straightforward: it is essentially to reduce to the simpler special case (of bounded support and good shift-invariance in every direction) described above. (A crucial aspect of this transformation algorithm is that it uses only a small number of draws from the original shift-invariant distribution; we return to this point below.) We can then use the algorithm for the special case to obtain a high-accuracy hypothesis, and perform the inverse transformation to obtain

[3] We note that a simpler version of this approach, which only uses a smoothing kernel and does not employ Fourier analysis, can be shown to give a similar, but quantitatively worse, results, such as a sample complexity of essentially $1/\varepsilon^{2d}$ when $g(t)$ is zero outside of a bounded region. However, this is worse than the lower bound of $\Omega(1/\varepsilon^d)$ by a quadratic factor, whereas our algorithm essentially achieves this optimal sample complexity.

a high-accuracy hypothesis for the original general distribution. We remark that while the conceptual idea is thus very straightforward, there are a number of technical challenges that must be met to implement this approach. One of these is that it is necessary to truncate the tails of the original distribution so that an affine transformation of it will have bounded support, and doing this changes the shift-invariance of the original distribution. Another is that the transformation procedure only succeeds with non-negligible probability, so we must run this overall approach multiple times and perform hypothesis selection to actually end up with a single high-accuracy hypothesis.

In Section 5 we apply the above results to establish efficient learnability of log-concave densities over $\mathbb{R}^d$. To apply our results, we need to have (i) bounds on the rate of tail decay, and (ii) shift-invariance bounds. As noted earlier, exponential tail bounds on d-dimensional log-concave densities are well known, so it remains to establish shift-invariance. Using basic properties of log-concave densities, in Section 5 we show that any d-dimensional isotropic log-concave density is $O_d(1)$-shift-invariant. Armed with this bound, by applying our learning result (Theorem 12) we get that any d-dimensional isotropic log-concave density can be learned in time $\tilde{O}_d(1/\varepsilon^{2d+2})$, using $\tilde{O}_d(1/\varepsilon^{d+2})$ samples. Log-concave distributions are shift-invariant even if they are only approximately isotropic. We show that general log-concave distributions may be learned by bringing them into approximately isotropic position with a preprocessing step, borrowing techniques from [37].

1.3 Related work

The most closely related work that we are aware of was mentioned above: Holmström and Klemelä [30] obtained bounds similar to ours for using kernel methods to learn densities that belong to various Sobolev spaces. As mentioned above, these results do not directly apply for learning densities in $\mathcal{C}_{\mathrm{SI}}(c, d, g)$ because of the possibility of jump discontinuities. Holmström and Klemelä also proved a lower bound on the sample complexity of algorithms that compute kernel density estimates. In contrast our lower bound holds for any density estimation algorithm, kernel-based or otherwise.

The assumption that the target density belongs to a Besov space (see [36]) makes reference to the effect of shifts on the distribution, as does shift-invariance. We do not see any obvious containments between classes of functions defined through shift-invariance and Besov spaces, but this is a potential topic for further research.

Another difference with prior work is the ability of our approach to succeed in the challenging noise-tolerant learning model. We are not aware of analyses for density estimation of densities belonging to Sobolev or Besov spaces that extend to the noise-tolerant setting in which the target density is only assumed to be close to some density in the relevant class.

As mentioned above, shift-invariance was used in the analysis of algorithms for learning discrete probability distributions in [5, 16]. Likewise, both the discrete and continuous Fourier transforms have been used in the past to learn discrete probability distributions [26, 27, 15].

2 Preliminaries

We write $B(r)$ to denote the radius-r ball in $\mathbb{R}^d$, i.e. $B(r) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \leq r^2\}$. If f is a probability density over $\mathbb{R}^d$ and $S \subset \mathbb{R}^d$ is a subset of its domain, we write f_S to denote the density of f conditioned on S.

2.1 Shift-invariance

Roughly speaking, the shift-invariance of a distribution measures how much it changes (in total variation distance) when it is subjected to a small translation. The notion of shift-invariance has typically been used for discrete distributions (especially in the context of proving discrete limit theorems, see e.g. [13] and many references therein). We give a natural continuous analogue of this notion below.

▶ **Definition 1.** Given a probability density f over $\mathbb{R}^d$, a unit vector v, and a positive real value κ, we say that the *shift-invariance of f in direction v at scale κ*, denoted $\mathrm{SI}(f, v, \kappa)$, is

$$\mathrm{SI}(f, v, \kappa) \stackrel{\text{def}}{=} \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} |f(x + \kappa' v) - f(x)|\, dx. \tag{1}$$

Intuitively, if $\mathrm{SI}(f, v, \kappa) = \beta$, then for any direction (unit vector) v the variation distance between f and a shift of f by κ' in direction v is at most $\kappa\beta$ for all $0 \le \kappa' \le \kappa$. The factor $\frac{1}{\kappa}$ in the definition means that $\mathrm{SI}(f, v, \kappa)$ does not necessarily go to zero as κ gets small; the effect of shifting by κ is measured relative to κ.

Let $\mathrm{SI}(f, \kappa) \stackrel{\text{def}}{=} \sup\{\mathrm{SI}(f, v, \kappa) : v \in \mathbb{R}^d, \|v\|_2 = 1\}$. For any constant c we define the class of densities $\mathcal{C}_{\mathrm{SI}}(c, d)$ to consist of all d-dimensional densities f with the property that $\mathrm{SI}(f, \kappa) \le c$ for all $\kappa > 0$.

We could obtain an equivalent definition if we removed the factor $\frac{1}{\kappa}$ from the definition of $\mathrm{SI}(f, v, \kappa)$, and required that $\mathrm{SI}(f, v, \kappa) \le c\kappa$ for all $\kappa > 0$. This could of course be generalized to enforce bounds on the modified $\mathrm{SI}(f, v, \kappa)$ that are not linear in κ. We have chosen to focus on linear bounds in this paper to have cleaner theorems and proofs.

We include "sup" in the definition due to the fact that smaller shifts can sometimes have bigger effects. For example, a sinusoid with period ξ is unaffected by a shift of size ξ, but profoundly affected by a shift of size $\xi/2$. Because of possibilities like this, to capture the intuitive notion that "small shifts do not lead to large changes", we seem to need to evaluate the worst case over shifts of at most a certain size.

As described earlier, given a nonincreasing "tail bound" function $g : \mathbb{R}^+ \to (0, 1)$ which is absolutely continuous and satisfies $\lim_{t\to+\infty} g(t) = 0$, we further define the class of densities $\mathcal{C}_{\mathrm{SI}}(c, d, g)$ to consist of those $f \in \mathcal{C}_{\mathrm{SI}}(c, d)$ which have the additional property that f has *g-light tails*, meaning that for all $t > 0$, it holds that $\mathbf{Pr}_{\boldsymbol{x}\leftarrow f}[||\boldsymbol{x} - \mu|| > t] \le g(t)$, where $\mu \in \mathbb{R}^d$ is the mean of f.

▶ Remark. It will be convenient in our analysis to consider only tail bound functions g that satisfy $\min\{r \in \mathbb{R} : g(r) \le 1/2\} \ge 1/10$ (the constants $1/2$ and $1/10$ are arbitrary here and could be replaced by any other absolute positive constants). This is without loss of generality, since any tail bound function g which does not meet this criterion can simply be replaced by a weaker tail bound function g^* which does meet this criterion, and clearly if f has g-light tails then f also has g^*-light tails.

We will (ab)use the notation $g^{-1}(\varepsilon)$ to mean $\inf\{t : g(t) \le \varepsilon\}$.

The complexity of learning with a tail bound g will be expressed in part using $I_g \stackrel{\text{def}}{=} \int_0^\infty g(\sqrt{z})\, dz$. We remark that the quantity I_g is the "right" quantity in the sense that the integral I_g is finite as long as the density has "non-trivial decay". More precisely, note that by Chebyshev's inequality, $g(\sqrt{z}) = O(z^{-1})$. Since the integral $\int O(z^{-1})dz$ diverges, this means that if I_g is finite, then the density f has a decay sharper than the trivial decay implied by Chebyshev's inequality.

2.2 Fourier transform of high-dimensional distributions

In this subsection we gather some helpful facts from multidimensional Fourier analysis.

While it is possible to do Fourier analysis over $\mathbb{R}^d$, in this paper, we will only do Fourier analysis for functions $f \in L_1([-1,1]^d)$.

▶ **Definition 2.** For any function $f \in L_1([-1,1]^d)$, we define $\widehat{f} : \mathbb{R}^d \to \mathbb{C}$ by $\widehat{f}(\xi) = \int_{x \in \mathbb{R}^d} f(x) \cdot e^{\pi i \cdot \langle \xi, x \rangle} dx$.

Next, we recall the following standard claims about Fourier transforms of functions, which may be found, for example, in [41].

▶ **Claim 3.** *For $f, g \in L_1([-1,1]^d)$ let $h(x) = \int_{y \in \mathbb{R}^d} f(y) \cdot g(x-y) dy$ denote the convolution $h = f * g$ of f and g. Then for any $\xi \in \mathbb{R}^n$, we have $\widehat{h}(\xi) = \widehat{f}(\xi) \cdot \widehat{g}(\xi)$.*

Next, we recall Parseval's identity on the cube.

▶ **Claim 4** (Parseval's identity). *For $f : [-1,1]^d \to \mathbb{R}$ such that $f \in L_2([-1,1]^d)$, it holds that $\int_{[-1,1]^d} f(x)^2 dx = \frac{1}{2^d} \cdot \sum_{\xi \in \mathbb{Z}^d} |\widehat{f}(\xi)|^2$.*

The next claim says that the Fourier inversion formula can be applied to any sequence in $\ell_2(\mathbb{Z}^d)$ to obtain a function whose Fourier series is identical to the given sequence.

▶ **Claim 5** (Fourier inversion formula). *For any $g : \mathbb{Z}^d \to \mathbb{C}$ such that $\sum_{\xi \in \mathbb{Z}^d} |g(\xi)^2| < \infty$, the function $h(x) = \sum_{\xi \in \mathbb{Z}^d} \frac{1}{2^d} \cdot g(\xi) \cdot e^{\pi i \cdot \langle \xi, x \rangle}$, is well defined and satisfies $\widehat{h}(\xi) = g(\xi)$ for all $\xi \in \mathbb{Z}^d$.*

We will also use Young's inequality:

▶ **Claim 6** (Young's inequality). *Let $f \in L_p([-1,1]^d)$, $g \in L_q([-1,1]^d)$, $1 \le p, q, r \le \infty$, such that $1 + 1/r = 1/p + 1/q$. Then $\|f * g\|_r \le \|f\|_p \cdot \|g\|_q$.*

2.3 A useful mollifier

Our algorithm and its analysis require the existence of a compactly supported distribution with fast decaying Fourier transform. Since the precise rate of decay is not very important, we use the C^∞ function $b : [-1,1] \to \mathbb{R}^+$ as follows:

$$b(x) = \begin{cases} c_0 \cdot e^{-\frac{x^2}{1-x^2}} & \text{if } |x| < 1 \\ 0 & \text{if } |x| = 1. \end{cases} \tag{2}$$

Here $c_0 \approx 1.067$ is chosen so that b is a pdf; by symmetry, its mean is 0. (This function has previously been used as a mollifier [34, 28].) The following fact can be found in [32] (while it is proved only for $\xi \in \mathbb{Z}$, it is easy to see that the same proof holds if $\xi \in \mathbb{R}$).

▶ **Fact 7.** *For $b : [-1,1] \to \mathbb{R}^+$ defined in (2) and $\xi \in \mathbb{Z} \setminus \{0\}$, we have that $|\widehat{b}(\xi)| \le e^{-\sqrt{|\xi|}} \cdot |\xi|^{-3/4}$.*

Let us now define the function $b_{d,\gamma} : \mathbb{R}^d \to \mathbb{R}^+$ as $b_{d,\gamma}(x_1, \ldots, x_d) = \frac{1}{\gamma^d} \cdot \prod_{j=1}^d b(x_j/\gamma)$. Combining this definition and Fact 7, we have the following claim:

▶ **Claim 8.** *For $\xi \in \mathbb{Z}^d$ with $\|\xi\|_\infty \ge t$, we have $|\widehat{b_{d,\gamma}}(\xi)| \le e^{-\sqrt{\gamma \cdot t}} \cdot (\gamma \cdot t)^{-3/4}$.*

The next fact is immediate from (2) and the definition of $b_{d,\gamma}$:

▶ **Fact 9.** *$\|b_{d,\gamma}\|_\infty = (c_0/\gamma)^d$ and as a consequence, $\|b_{d,\gamma}\|_2^2 \le (c_0/\gamma)^{2d}$.*

3 A restricted problem: learning shift-invariant distributions with bounded support

As sketched in Section 1.2, we begin by presenting and analyzing a density estimation algorithm for densities that, in addition to being shift-invariant, have support bounded in $B(1/2)$. Our analysis also captures the fact that, to achieve accuracy ε, an algorithm often only needs the density to be learned to have shift invariance at a scale slightly finer than ε.

▶ **Lemma 10.** *There is an algorithm learn-bounded with the following property: For all constant d, for all $\varepsilon, \delta > 0$, all $0 < \kappa < \varepsilon < 1/2$, and all d-dimensional densities f with support in $B(1/2)$ such that $\kappa\mathrm{SI}(f, \kappa) \le \varepsilon/2$, given access to independent draws from f, the algorithm runs in $O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^{2d}\log^{4d}\left(\frac{1}{\kappa}\right)\log\left(\frac{1}{\kappa\delta}\right)\right)$ time uses $O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^{d}\log^{2d}\left(\frac{1}{\kappa}\right)\log\left(\frac{1}{\kappa\delta}\right)\right)$ samples, and with probability $1-\delta$, outputs a hypothesis $h : [-1,1]^d \to \mathbb{R}^+$ such that $\int_{x\in\mathbb{R}^d} |f(x) - h(x)| \le \varepsilon$.*

Further, given any point $z \in [-1,1]^d$, $h(z)$ can be computed in time $O_d\left(\frac{\log^{2d}(1/\kappa)}{\kappa^d}\right)$ and satisfies $h(z) \le O_d\left(\frac{\log^{2d}(1/\kappa)}{\kappa^d}\right)$.

Proof. Let $0 < \gamma := \frac{\kappa}{\sqrt{d}}$, and let us define $q = f * b_{d,\gamma}$. (Here $*$ denotes convolution and $b_{d,\gamma}$ is the mollifier defined in Section 2.3.) We make a few simple observations about q:

- **(i)** Since $\gamma \le 1/2$, we have that q is a density supported on $B(1)$.
- **(ii)** Since d is a constant, a draw from $b_{d,\gamma}$ can be generated in constant time. Thus given a draw from f, one can generate a draw from q in constant time, simply by generating a draw from $b_{d,\gamma}$ and adding it to the draw from f.
- **(iii)** By Young's inequality (Claim 6), we have that $\|q\|_2 \le \|f\|_1 \cdot \|b_{d,\gamma}\|_2$. Noting that f is a density and thus $\|f\|_1 = 1$ and applying Fact 9, we obtain that $\|q\|_2$ is finite. As a consequence, the Fourier coefficients of q are well-defined.

Preliminary analysis. We first observe that because $b_{d,\gamma}$ is supported on $[-\gamma,\gamma]^d$, the distribution q may be viewed as an average of different shifts of f where each shift is by a distance at most $\gamma\sqrt{d} \le \kappa$. Fix any direction v and consider a shift of f in direction v by some distance at most $\gamma\sqrt{d} \le \kappa$. Since $\kappa\mathrm{SI}(f,\kappa) \le \varepsilon/2$, we have that the variation distance between f and this shift in direction v is at most $\varepsilon/2$. Averaging over all such shifts, it follows that $d_{\mathrm{TV}}(q, f) \le \varepsilon/2$.

Next, we observe that by Claim 3, for any $\xi \in \mathbb{Z}^d$, we have $\widehat{q}(\xi) = \widehat{f}(\xi) \cdot \widehat{b_{d,\gamma}}(\xi)$. Since f is a pdf, $|\widehat{f}(\xi)| \le 1$, and thus we have $|\widehat{q}(\xi)| \le |\widehat{b_{d,\gamma}}(\xi)|$. Also, for any parameter $k \in \mathbb{Z}^+$, define $C_k = \{\xi \in \mathbb{Z}^d : \|\xi\|_\infty = k\}$. Let us fix another parameter T (to be determined later). Applying Claim 8, we obtain

$$\sum_{\xi:\|\xi\|_\infty > T} |\widehat{q}(\xi)|^2 \le \sum_{\xi:\|\xi\|_\infty > T} |\widehat{b_{d,\gamma}}(\xi)|^2 \le \sum_{k>T}\sum_{\xi:\|\xi\|_\infty = k} |\widehat{b_{d,\gamma}}(\xi)|^2$$
$$\le \sum_{k>T} |C_k| \cdot e^{-2\cdot\sqrt{\gamma\cdot k}} \cdot (\gamma\cdot k)^{-3/2} \le \sum_{k>T} (2k+1)^d \cdot e^{-2\cdot\sqrt{\gamma\cdot k}} \cdot (\gamma\cdot k)^{-3/2}.$$

An easy calculation shows that if $T \ge \frac{4d^2}{\gamma}\cdot\ln^2\left(\frac{d}{\gamma}\right)$, then $\sum_{\xi:\|\xi\|_\infty>T}|\widehat{q}(\xi)|^2 \le 2(2T+1)^d \cdot e^{-2\cdot\sqrt{\gamma\cdot T}} \cdot (\gamma\cdot T)^{-3/2}$. If we now set T to be $\frac{4d^2}{\gamma}\cdot\ln^2\left(\frac{d}{\gamma}\right) + \frac{1}{\gamma}\cdot\ln^2\left(\frac{8}{\varepsilon}\right)$, then $\sum_{\xi:\|\xi\|_\infty>T}|\widehat{q}(\xi)|^2 \le \frac{\varepsilon^2}{8}$.

The algorithm. We first observe that for any $\xi \in \mathbb{Z}^d$, the Fourier coefficient $\widehat{q}(\xi)$ can be estimated to good accuracy using relatively few draws from q (and hence from f, recalling (ii) above). More precisely, as an easy consequence of the definition of the Fourier transform, we have:

▶ **Observation 11.** *For any $\xi \in \mathbb{Z}^d$, the Fourier coefficient $\widehat{q}(\xi)$ can be estimated to within additive error of magnitude at most η with confidence $1-\beta$ using $O(1/\eta^2 \cdot \log(1/\beta))$ draws from q.*

Let us define the set Low of low-degree Fourier coefficients as $\mathsf{Low} = \{\xi \in \mathbb{Z}^d : \|\xi\|_\infty \leq T\}$. Thus, $|\mathsf{Low}| \leq (2T+1)^d$. Thus, using $S = O(\eta^{-2} \cdot \log(T/\delta))$ draws from f, by Observation 11, with probability $1-\delta$, we can compute a set of values $\{\widehat{u}(\xi)\}_{\xi \in \mathsf{Low}}$ such that

$$\text{For all } \xi \in \mathsf{Low},\ |\widehat{u}(\xi) - \widehat{q}(\xi)| \leq \eta. \tag{3}$$

Recalling (ii), the sequence $\{\widehat{u}(\xi)\}_{\xi \in \mathsf{Low}}$ can be computed in $O(|S| \cdot |\mathsf{Low}|)$ time. Define $\widehat{u}(\xi) = 0$ for $\xi \in \mathbb{Z}^d \setminus \mathsf{Low}$. Combining (3) with this, we get

$$\begin{aligned}\sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 &\leq \sum_{\xi \in \mathsf{Low}} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 + \sum_{\xi \notin \mathsf{Low}} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 \\ &\leq \sum_{\xi \in \mathsf{Low}} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 + \frac{\varepsilon^2}{8}|\mathsf{Low}| \cdot \eta^2 + \frac{\varepsilon^2}{8} \leq (2T+1)^d \cdot \eta^2 + \frac{\varepsilon^2}{8}.\end{aligned}$$

Thus, setting η as $\eta^2 = (2T+1)^{-d} \cdot \frac{\varepsilon^2}{8}$, we get that $\sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 \leq \frac{\varepsilon^2}{4}$. Note that by definition $\widehat{u} : \mathbb{Z}^d \to \mathbb{C}$ satisfies $\sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi)|^2 < \infty$. Thus, we can apply the Fourier inversion formula (Claim 5) to obtain a function $u : [-1,1]^d \to \mathbb{C}$ such that

$$\int_{[-1,1]^d} |u(x) - q(x)|^2 dx = \frac{1}{2^d} \cdot \Big(\sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 \Big) \leq \frac{\varepsilon^2}{4 \cdot 2^d}, \tag{4}$$

where the first equality follows by Parseval's identity (Claim 4). By the Cauchy-Schwarz inequality, $\int_{[-1,1]^d} |u(x) - q(x)| dx \leq \sqrt{2^d} \cdot \sqrt{\int_{[-1,1]^d} |u(x) - q(x)|^2 dx}$. Plugging in (4), we obtain $\int_{[-1,1]^d} |u(x) - q(x)| dx \leq \frac{\varepsilon}{2}$. Let us finally define h (our final hypothesis), $h : [-1,1]^d \to \mathbb{R}^+$, as follows: $h(x) = \max\{0, \mathsf{Re}(u(x))\}$. Note that since $q(x)$ is a non-negative real value for all x, we have

$$\int_{[-1,1]^d} |h(x) - q(x)| dx \leq \int_{[-1,1]^d} |u(x) - q(x)| dx \leq \frac{\varepsilon}{2}. \tag{5}$$

Finally, recalling that we previously proved $d_{\mathrm{TV}}(f,q) \leq \frac{\varepsilon}{2}$, it follows that $\int_{[-1,1]^d} |h(x) - f(x)| dx \leq \varepsilon$.

Complexity analysis. We now analyze the time and sample complexity of this algorithm as well as the complexity of computing h. First of all, observe that plugging in the value of γ and recalling that d is a constant, we get that $T = \frac{4d^2}{\gamma} \cdot \ln^2\left(\frac{d}{\gamma}\right) + \frac{1}{\gamma} \cdot \ln^2\left(\frac{8}{\varepsilon}\right) = O\left(\frac{\log^2(1/\kappa)}{\kappa}\right)$. Combining this with the choice of η we get that the algorithm uses

$$S = O\left(\frac{1}{\eta^2} \cdot \log\left(\frac{|\mathsf{Low}|}{\delta}\right)\right) = O\left(\frac{1}{\eta^2} \cdot \log\left(\frac{T}{\delta}\right)\right) = O\left(\frac{(2T+1)^d \cdot \log\left(\frac{T}{\delta}\right)}{\varepsilon^2}\right)$$

$$= O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^d \log^{2d}\left(\frac{1}{\kappa}\right) \log\left(\frac{1}{\kappa\delta}\right)\right)$$

draws from p. Next, as we have noted before, computing the sequence $\{\widehat{u}(\xi)\}$ takes time

$$O(S \cdot |\mathsf{Low}|) = O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^d \log^{2d}\left(\frac{1}{\kappa}\right) \log\left(\frac{1}{\kappa\delta}\right) T^d\right)$$
$$= O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^{2d} \log^{4d}\left(\frac{1}{\kappa}\right) \log\left(\frac{1}{\kappa\delta}\right)\right).$$

To compute the function u (and hence h) at any point $x \in [-1,1]^d$ takes time $O(|\mathsf{Low}|) = O_d\left(\frac{\log^{2d}(1/\kappa)}{\kappa^d}\right)$. This is because the Fourier inversion formula (Claim 5) has at most $O(|\mathsf{Low}|)$ non-zero terms.

Finally, we prove the upper bound on h. If the training examples are $x_1, ..., x_S$, then for any $z \in [-1,1]^d$, we have

$$h(z) \le |u(z)| = \left|\sum_{\xi \in \mathsf{Low}} \frac{1}{2^d} \cdot \widehat{u}(\xi) \cdot e^{\pi i \cdot \langle \xi, z\rangle}\right| = \left|\sum_{\xi \in \mathsf{Low}} \frac{1}{2^d} \cdot \left(\frac{1}{S}\sum_{t=1}^{S} e^{\pi i \langle \xi, x_t\rangle}\right) \cdot e^{\pi i \cdot \langle \xi, z\rangle}\right|$$
$$\le \frac{|\mathsf{Low}|}{2^d} = O_d\left(\frac{\log^{2d}(1/\kappa)}{\kappa^d}\right),$$

completing the proof. ◀

4 Density estimation for densities in $\mathcal{C}_{\mathrm{SI}}(c,d,g)$

Fix any nonincreasing tail bound function $g : \mathbb{R}^+ \to [0,1]$ which satisfies $\lim_{t\to+\infty} g(t) = 0$ and $\min\{r \in \mathbb{R} : g(r) \le 1/2\} \ge 1/10$ and any constant $c \ge 1$. In this section we prove the following theorem which gives a density estimation algorithm for the class of distributions $\mathcal{C}_{\mathrm{SI}}(c,d,g)$:

▶ **Theorem 12.** *For any c, g as above and any $d \ge 1$, there is an algorithm with the following property: Let f be any target density (unknown to the algorithm) which belongs to $\mathcal{C}_{\mathrm{SI}}(c,d,g)$. Given any error parameter $0 < \varepsilon < 1/2$ and confidence parameter $\delta > 0$ and access to independent draws from f, the algorithm with probability $1 - O(\delta)$ outputs a hypothesis $h : [-1,1]^d \to \mathbb{R}^{\ge 0}$ such that $\int_{x\in\mathbb{R}^d} |f(x) - h(x)| \le O(\varepsilon)$.*

The algorithm runs in $O_{c,d}\left(\left((g^{-1}(\varepsilon))^{2d}\left(\frac{1}{\varepsilon}\right)^{2d+2} \log^{4d}\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right) \log\left(\frac{g^{-1}(\varepsilon)}{\varepsilon\delta}\right) + I_g\right) \log\frac{1}{\delta}\right)$ time and uses $O_{c,d}\left(\left((g^{-1}(\varepsilon))^{d}\left(\frac{1}{\varepsilon}\right)^{d+2} \log^{2d}\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right) \log\left(\frac{g^{-1}(\varepsilon)}{\varepsilon\delta}\right) + I_g\right) \log\frac{1}{\delta}\right)$ samples.

4.1 Outline of the proof

Theorem 12 is proved by a reduction to Lemma 10. The main ingredient in the proof of Theorem 12 is a "transformation algorithm" with the following property: given as input access to i.i.d. draws from any density $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$, the algorithm constructs parameters which enable draws from the density f to be transformed into draws from another density, which we denote r. The density r is obtained by approximating f after conditioning on a non-tail sample, and scaling the result so that it lies in a ball of radius $1/2$.

Given such a transformation algorithm, the approach to learn f is clear: we first run the transformation algorithm to get access to draws from the transformed distribution r. We then use draws from r to run the algorithm of Lemma 10 to learn r to high accuracy. (Intuitively, the error relative to f of the final hypothesis density is $O(\varepsilon)$ because at most

$O(\varepsilon)$ comes from the conditioning and at most $O(\varepsilon)$ from the algorithm of Lemma 10.) We note that while this high-level approach is conceptually straightforward, a number of technical complications arise; for example, our transformation algorithm only succeeds with some non-negligible probability, so we must run the above-described combined procedure multiple times and perform hypothesis testing to identify a successful final hypothesis from the resulting pool of candidates.

The rest of this section is organized as follows: In Section 4.2 we give various necessary technical ingredients for our transformation algorithm. We state and prove the key results about the transformation algorithm in Section 4.3, and we use the transformation algorithm to prove Theorem 12 in Section 4.4.

4.2 Technical ingredients for the transformation algorithm

As sketched earlier, our approach will work with a density obtained by conditioning $f \in \mathrm{SI}(c,d)$ on lying in a certain ball that has mass close to 1 under f. While we know that the original density $f \in \mathrm{SI}(c,d)$ has good shift-invariance, we will further need the conditioned distribution to also have good shift-invariance in order for the learn-bounded algorithm of Section 3 to work. Thus we require the following simple lemma, which shows that conditioning a density $f \in \mathrm{SI}(c,d)$ on a region of large probability cannot hurt its shift invariance too much.

▶ **Lemma 13.** *Let $f \in \mathrm{SI}(c,d)$ and let B be a ball such that $\Pr_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B] \geq 1-\delta$ where $\delta < 1/2$. If f_B is the density of f conditioned on B, then, for all $\kappa > 0$, $\mathrm{SI}(f_B,\kappa) \leq \frac{4\delta}{\kappa} + 2c$.*

Proof. Let v be any unit vector in $\mathbb{R}^d$. Note that f can be expressed as $(1-\delta)f_B + \delta \cdot f_{err}$ where f_{err} is some other density. As a consequence, for any $\kappa > 0$, using the triangle inequality we have that

$$\int_x |f(x) - f(x+\kappa v)|dx \geq (1-\delta)\int_x |f_B(x) - f_B(x+\kappa v)|dx - \delta\int_x |f_{err}(x) - f_{err}(x+\kappa v)|dx.$$

Since $f \in \mathcal{C}_{\mathrm{SI}}(c,d)$ the left hand side is at most $c\kappa$, whereas the subtrahend on the right hand side is trivially at most 2δ. Thus, we get $\int_x |f_B(x) - f_B(x+\kappa v)|dx \leq \frac{2\delta}{1-\delta} + \frac{c\kappa}{1-\delta}$, completing the proof. ◀

If f is an unknown target density then of course its mean is also unknown, and thus we will need to approximate it using draws from f. To do this, it will be helpful to convert our condition on the tails of f to bound the variance of $||\boldsymbol{x} - \mu||$, where $\boldsymbol{x} \sim f$.

▶ **Lemma 14.** *For any $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$, we have $\mathbf{E}_{\boldsymbol{x}\sim f}[||\boldsymbol{x}-\mu||^2] \leq I_g$.*

Proof. We have $\mathbf{E}_{\boldsymbol{x}\sim f}[||\boldsymbol{x}-\mu||^2] = \int_0^\infty \Pr_{\boldsymbol{x}\sim f}[||\boldsymbol{x}-\mu||^2 \geq z]\, dz \leq \int_0^\infty g(\sqrt{z})\, dz = I_g$. ◀

The following easy proposition gives a guarantee on the quality of the empirical mean:

▶ **Lemma 15.** *For any $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$, if $\mu \in \mathbb{R}^d$ is the mean of f and $\widehat{\boldsymbol{\mu}}$ is its empirical estimate based on M samples, then for any $t > 0$ we have $\Pr\left[||\mu - \widehat{\boldsymbol{\mu}}||^2 \geq t\right] \leq \frac{I_g}{Mt}$.*

Proof. If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ are independent draws from f, then

$$\mathbf{E}[||\mu - \widehat{\boldsymbol{\mu}}||^2] = \mathbf{E}\left[\left|\left|\mu - \frac{\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_M}{M}\right|\right|^2\right] = \sum_{i=1}^M \frac{1}{M^2}\mathbf{E}\left[\left|\left|\mu - \boldsymbol{x}_i\right|\right|^2\right] = \frac{I_g}{M},$$

where the last inequality is by Lemma 14. Applying Markov's inequality on the left hand side, we get the stated claim. ◀

4.3 Transformation algorithm

▶ **Lemma 16.** *There is an algorithm* compute-transformation *such that given access to samples from $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$ and an error parameter $0 < \varepsilon < 1/2$, the algorithm takes $O(I_g)$ samples from f and with probability at least $9/10$ produces a vector $\tilde{\mu} \in \mathbb{R}^d$ and a real number t with the following properties: (1) For $B_t = \{x : ||x - \tilde{\mu}|| \le \sqrt{t}\}$, we have $\Pr_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t] \ge 1 - \varepsilon$; (2) $t = O(g^{-1}(\varepsilon)^2)$; (3) For all $\kappa > 0$, the density f_{B_t} satisfies $\mathrm{SI}(f_{B_t}, \kappa) \le \frac{4\varepsilon}{\kappa} + 2c$.*

Proof. For $M = 100 I_g$, the algorithm compute-transformation simply works as follows: set $\tilde{\mu}$ to be the empirical mean of the M samples, and $t = 2((g^{-1}(\varepsilon))^2 + 1/10)$. (Note that since $\min\{r \in \mathbb{R} : g(r) \le 1/2\} \ge 1/10$ we have $t = \Theta(g^{-1}(\varepsilon)^2)$.). Let μ denote the true mean of f. First, by Lemma 15, with probability at least 0.9, the empirical mean $\widehat{\boldsymbol{\mu}}$ will satisfy $||\mu - \widehat{\boldsymbol{\mu}}||^2 \le \frac{1}{10}$. Let us assume for the rest of the proof that this happens; fix any such outcome and denote it $\tilde{\mu}$.

We have $||x - \tilde{\mu}||^2 \le 2(||x - \mu||^2 + ||\mu - \tilde{\mu}||^2) \le 2(||x - \mu||^2 + 1/10)$ and so

$$\Pr_{\boldsymbol{x}\in f}[||\boldsymbol{x} - \tilde{\mu}||^2 > t] \le \Pr_{\boldsymbol{x}\in f}[2(||\boldsymbol{x} - \mu||^2 + 1/10) > t] = \Pr[\|\boldsymbol{x} - \mu\|^2 \ge g^{-1}(\varepsilon)] \le \varepsilon.$$

Applying Lemma 13 completes the proof. ◀

The following proposition elaborates on the properties of the output of the transformation algorithm.

▶ **Lemma 17.** *Let $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$, $\varepsilon > 0$, $\tilde{\mu} \in \mathbb{R}^d$, and $t \in \mathbb{R}$ satisfy the properties stated in Lemma 16. Consider the density f_{scond} defined by $f_{\mathrm{scaled}}(x) \stackrel{\mathrm{def}}{=} 2\sqrt{t} \cdot f\big(2\sqrt{t} \cdot (x + \tilde{\mu})\big)$ and $f_{\mathrm{scond}}(x) \stackrel{\mathrm{def}}{=} f_{\mathrm{scaled},B(1/2)}(x)$ where $f_{\mathrm{scaled},B(1/2)}$ is the result of conditioning f_{scaled} on membership in $B(1/2)$. Then the density $f_{\mathrm{scond}}(x)$ satisfies the following properties: (1) The density f_{scond} is supported in the ball $B(1/2)$; (2) For all $\varepsilon < 1/2$ and $\kappa > 0$, the density f_{scond} satisfies $\mathrm{SI}(f_{\mathrm{scond}}, \kappa) \le \frac{4\varepsilon}{\kappa} + 4c\sqrt{t}$.*

Proof. First, it is easy to verify that function f_{scond} defined above is indeed a density. Item 1 is enforced by fiat. Now, for any direction v, we have

$$\begin{aligned}
\mathrm{SI}(f_{\mathrm{scaled}}, v, \kappa) &= \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} |f_{\mathrm{scaled}}(x + \kappa' v) - f_{\mathrm{scaled}}(x)|\, dx \\
&= \frac{2\sqrt{t}}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} \Big| f(2\sqrt{t}(x + \kappa' v)) - f(2\sqrt{t}x) \Big|\, dx.
\end{aligned}$$

Using a change of variables, $u = 2\sqrt{t}x$, we get

$$\begin{aligned}
\mathrm{SI}(f_{\mathrm{scaled}}, v, \kappa) &= \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} \Big| f(u + \kappa' 2\sqrt{t} v) - f(u) \Big|\, du \\
&= \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,2\sqrt{t}\kappa]} \int_{\mathbb{R}^d} |f(u + \kappa' v) - f(u)|\, du \\
&= 2\sqrt{t} \cdot \mathrm{SI}(f, v, 2\sqrt{t}\kappa) \le 2c\sqrt{t}. \qquad (6)
\end{aligned}$$

The last inequality uses that $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$. Inequality (6) implies that $f_{\mathrm{scaled}} \in \mathcal{C}_{\mathrm{SI}}(2c\sqrt{t}, d, g)$. Now, $\Pr_{\boldsymbol{x}\sim f_{\mathrm{scaled}}}(\boldsymbol{x} \in B(1/2)) = \Pr_{\boldsymbol{x}\sim f}(\boldsymbol{x} \in B_t) \ge 1 - \varepsilon$, so applying Lemma 13 completes the proof. ◀

4.4 Proof of Theorem 12

We are now ready to prove Theorem 12. Consider the following algorithm, which we call construct-candidates:

1. Run the transformation algorithm compute-transformation $D := O(\ln(1/\delta))$ many times (with parameter ε each time). Let $(\tilde{\mu}^{(i)}, t)$ be the output that it produces on the i-th run, where $t = O(g^{-1}(\varepsilon)^2)$.
2. For each $i \in [D]$, let $B_t^{(i)} = \{x : ||x - \tilde{\mu}|| \leq \sqrt{t}\}$ and $f_{\text{scond}}^{(i)}$ be the density defined from $(\tilde{\mu}^{(i)}, t)$ as in Lemma 17.

Before describing the third step of the algorithm, we observe that given the pair $(\tilde{\mu}^{(i)}, t)$ it is easy to check whether any given $x \in \mathbb{R}^d$ belongs to $B_t^{(i)}$. If $\Pr_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] \geq 1/2$, then with probability at least $1/2$ a draw from f can be used as a draw from $f_{B_t^{(i)}}$. In this case, via rejection sampling, it is easy to very efficiently simulate draws from $f_{\text{scond}}^{(i)}$ given access to samples from f (the average slowdown is at most a factor of 2). Note that if $(\tilde{\mu}^{(i)}, t)$ satisfies the properties of Lemma 16, then $\Pr_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] \geq 1 - \varepsilon$ and we fall into this case. On the other hand, if $\Pr_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] < 1/2$, then it may be inefficient to simulate draws from $f_{\text{scond}}^{(i)}$. But any such i will not satisfy the properties of Lemma 16, so if rejection sampling is inefficient to simulate draws from $f_{\text{scond}}^{(i)}$ then we can ignore such an i in what follows. With this in mind, the third and fourth steps of the algorithm are as follows:

3. For each $i \in [D]$,[4] run the algorithm learn-bounded using m samples from $f_{\text{scond}}^{(i)}$, where $m = m(\varepsilon, \delta, d)$ is the sample complexity of learn-bounded from Lemma 10. Let $h_{\text{scond}}^{(i)}$ be the resulting hypothesis that learn-bounded outputs.
4. Finally, for each $i \in [D]$ output the hypothesis obtained by inverting the mapping of Lemma 17, i.e.

$$h^{(i)}(x) \stackrel{\text{def}}{=} \frac{1}{2\sqrt{t}} \cdot h_{\text{scond}}^{(i)}\left(\frac{1}{2\sqrt{t}} \cdot (x - \tilde{\mu}^{(i)})\right). \tag{7}$$

Thus the output of construct-candidate is a D-tuple of hypotheses $(h^{(1)}, \ldots, h^{(D)})$.

We now analyze the construct-candidate algorithm. Given Lemma 16 and Lemma 17, it is not difficult to show that with high probability at least one of the hypotheses that it outputs has error $O(\varepsilon)$ with respect to f:

▶ **Lemma 18.** *With probability at least $1 - O(\delta)$, at least one $h^{(i)}$ has $\int_x |h^{(i)}(x) - f(x)|dx \leq O(\varepsilon)$.*

Proof. It is immediate from Lemma 16 and the choice of D that with probability $1 - \delta$ at least one triple $(\tilde{\mu}^{(i)}, t)$ satisfies the properties of Lemma 16. Fix i' to be an i for which this holds.

Given any $i \in [D]$, it is easy to carry out the check for whether rejection sampling is too inefficient in simulating $f_{\text{scond}}^{(i)}$ in such a way that algorithm learn-bounded will indeed be run to completion (as opposed to being terminated) on $f_{\text{scond}}^{(i')}$ with probability at least $1 - \delta$, so we henceforth suppose that indeed learn-bounded is actually run to completion on $f_{\text{scond}}^{(i')}$. Since $(\tilde{\mu}^{(i')}, t)$ satisfies the properties of Lemma 16, by Lemma 17, taking

[4] Actually, as described above, this and the fourth step are done only for those i for which rejection sampling is not too inefficient in simulating draws from $f_{\text{scond}}^{(i)}$ given draws from f; for the other i's, the run of learn-bounded is terminated.

$\kappa = \min\{\varepsilon/2, \varepsilon/(4g^{-1}(\varepsilon)c)\}$) the density $f^{(i')}_{\mathrm{scond}}$ satisfies the required conditions for Lemma 10 to apply with that choice of κ. The following simple proposition, proved in the long version of this paper [20], implies that $h^{(i)}$ is likewise $O(\varepsilon)$-close to f_{B_t}:

▶ **Proposition 19.** *Let f and g be two densities in $\mathbb{R}^d$ and let $x \mapsto A(x-z)$ be any invertible linear transformation over $\mathbb{R}^d$. Let $f_A(x) = \det(A) \cdot f(A(x-z))$ and $g_A(x) = \det(A) \cdot g(A(x-z))$ be the densities from f and g under this transformation. Then $d_{\mathrm{TV}}(f,g) = d_{\mathrm{TV}}(f_A, g_A)$.*

It remains only to observe that by property 1 of Lemma 16 the density f_{B_t} is ε-close to f, and then by the triangle inequality we have that $h^{(i)}$ is $O(\varepsilon)$-close to f. This gives Lemma 18. ◀

Tracing through the parameters, it is straightforward to verify that the sample and time complexities of construct-candidates are as claimed in the statement of Theorem 12. These sample and time complexities dominate the sample and time complexities of the remaining portion of the algorithm, the hypothesis selection procedure discussed below.

All that is left is to identify a good hypothesis from the pool of D candidates. This can be carried out rather straightforwardly using well-known tools for hypothesis selection. Many variants of the basic hypothesis selection procedure have appeared in the literature, see e.g. [44, 18, 2, 17, 19]). The following is implicit in the proof of Proposition 6 from [19]:

▶ **Proposition 20.** *Let $\mathbf{D}$ be a distribution with support contained in a set W and let $\mathcal{D}_\varepsilon = \{\mathbf{D}_j\}_{j=1}^M$ be a collection of M hypothesis distributions over W with the property that there exists $i \in [M]$ such that $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_i) \le \varepsilon$. There is an algorithm $\mathrm{Select}^{\mathbf{D}}$ which is given ε and a confidence parameter δ, and is provided with access to (i) a source of i.i.d. draws from $\mathbf{D}$ and from $\mathbf{D}_i$, for all $i \in [M]$; and (ii) a $(1+\beta)$ "approximate evaluation oracle" $\mathrm{eval}_{\mathbf{D}_i}(\beta)$, for each $i \in [M]$, which, on input $w \in W$, deterministically outputs $\tilde{D}_i^\beta(w)$ such that the value $\frac{\mathbf{D}_i(w)}{1+\beta} \le \tilde{D}_i^\beta(w) \le (1+\beta)\cdot \mathbf{D}_i(w)$. Further, $(1+\beta)^2 \le (1+\varepsilon/8)$. The $\mathrm{Select}^{\mathbf{D}}$ algorithm has the following behavior: It makes $m = O\left((1/\varepsilon^2)\cdot(\log M + \log(1/\delta))\right)$ draws from $\mathbf{D}$ and from each $\mathbf{D}_i$, $i \in [M]$, and $O(m)$ calls to each oracle $\mathrm{eval}_{\mathbf{D}_i}$, $i \in [M]$. It runs in time $\mathrm{poly}(m, M)$ (counting each call to an $\mathrm{eval}_{\mathbf{D}_i}$ oracle and draw from a $\mathbf{D}_i$ distribution as unit time), and with probability $1-\delta$ it outputs an index $i^\star \in [M]$ that satisfies $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_{i^\star}) \le 6\varepsilon$.*

As suggested above, the remaining step is to apply Proposition 20 to the list of candidate hypothesis $h^{(i)}$ which satisfies the guarantee of Lemma 18. However, to bound the sample and time complexity of running the procedure Proposition 20, we need to bound the complexity both of sampling from $\{h^{(i)}\}_{i\in[D]}$ as well as of constructing approximate evaluation oracles for these measures.[5] In fact, we will first construct densities out of the measures $\{h^{(i)}\}_{i\in[D]}$ and show how to both efficiently sample from these measures as well as construct approximate evaluation oracles for these densities.

Towards this, let us now define $H_{\max}$ as follows: $H_{\max} = \max_{i\in[D]} \max_{z\in[-1,1]^n} h^{(i)}_{\mathrm{scond}}(z)$. From Lemma 10 (recall that Lemma 10 was applied with $\kappa = \min\{\varepsilon/2, \varepsilon/(4g^{-1}(\varepsilon)c)\}$) we get that $H_{\max} = O_{c,d}\left(\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right)^d \log^{2d}\frac{g^{-1}(\varepsilon)}{\varepsilon}\right)$. We will carry out the rest of our calculations in terms of $H_{\max}$.

[5] Note that while $h^{(i)}$ are forced to be non-negative and thus can be seen as measures, they need not integrate to 1 and thus need not be densities.

▶ **Observation 21.** *For any $i \in [D]$, $\int_{x\in[-1,1]^d} h^{(i)}_{\text{scond}}(x)dx$ can be estimated to additive accuracy $\pm\varepsilon$ and confidence $1-\delta$ in time $O_d\left(\frac{H^2_{\max}}{\varepsilon^2}\cdot\log(1/\delta)\right)$.*

Proof. First note that it suffices to estimate the quantity $\mathbf{E}_{x\in[-1,1]^d}[h^{(i)}_{\text{scond}}(x)]$ to additive error $\varepsilon/2^d$. However, this can be estimated using the trivial random sampling algorithm. In particular, as $h^{(i)}_{\text{scond}}(x) \in [0, H_{\max}]$, the variance of the simple unbiased estimator for $\mathbf{E}_{x\in[-1,1]^d}[h^{(i)}_{\text{scond}}(x)]$ is also bounded by $H^2_{\max}$. This finishes the proof. ◀

Note that, while the algorithm of Observation 21 does random sampling, this sampling is not from f, so it adds nothing to the sample complexity of the learning algorithm.

Next, for $i \in [D]$, let us define the quantity Z_i to be $Z_i = \int_x h^{(i)}(x)dx$. Since the functions $h^{(i)}$ and $h^{(i)}_{\text{scond}}$ are obtained from each other by linear transformations (recall (7)), we get that $2\sqrt{t}Z_i = \int_x h^{(i)}_{\text{scond}}\left(\frac{1}{2\sqrt{t}}\cdot(x-\tilde{\mu}^{(i)})\right)dx$. We now define the functions $H^{(i)}$ and $H^{(i)}_{\text{scond}}$ as $H^{(i)}(x) = \frac{h^{(i)}(x)}{Z_i}$ and $H^{(i)}_{\text{scond}}(x) = \frac{h^{(i)}_{\text{scond}}(\frac{1}{2\sqrt{t}}\cdot(x-\tilde{\mu}^{(i)}))}{Z_i}\cdot\frac{1}{2\sqrt{t}}$. Observe that the functions $H^{(i)}$ and $H^{(i)}_{\text{scond}}$ are densities (i.e. they are non-negative and integrate to 1). First, we will show that it suffices to run the procedure Select$^{\mathbf{D}}$ on the densities $H^{(i)}$. To see this, note that Lemma 18 says that there exists $i \in [D]$ such that $h^{(i)}$ satisfies $\int_x |h^{(i)}(x) - f(x)| = O(\varepsilon)$. For such an i, $Z_i \in [1-O(\varepsilon), 1+O(\varepsilon)]$. Thus, we have the following corollary.

▶ **Corollary 22.** *With probability at least $1-\delta$, at least one $H^{(i)}$ satisfies $\int_x |H^{(i)}(x)-f(x)| = O(\varepsilon)$. Further, for such an i, $Z_i \in [1-O(\varepsilon), 1+O(\varepsilon)]$.*

Thus, it suffices to run the procedure Select$^{\mathbf{D}}$ on the candidate distributions $\{H^{(i)}\}_{i\in[D]}$. The next proposition shows that the densities $\{H^{(i)}\}_{i\in[D]}$ are samplable.

▶ **Proposition 23.** *A draw from the density $H^{(i)}(x)$ can be sampled in time $O(H_{\max}/Z_i)$.*

Proof. First of all, note that it suffices to sample from $H^{(i)}_{\text{scond}}$ since $H^{(i)}$ and $H^{(i)}_{\text{scond}}$ are linear transformations of each other. However, sampling from $H^{(i)}_{\text{scond}}$ is easy using rejection sampling. More precisely, the distribution $H^{(i)}_{\text{scond}}$ is supported on $[-1,1]^d$. We sample from $H^{(i)}_{\text{scond}}$ as follows:

1. Let $C = [-1,1]^d \times [0, H_{\max}]$. Sample a uniformly random point $z' = (z_1, \ldots, z_{d+1})$ from C.
2. If $z_{d+1} \le h^{(i)}_{\text{scond}}(z_1, \ldots, z_d)$, then return the point $z = (z_1, \ldots, z_d)$.
3. Else go to Step 1 and repeat.

Now note that conditioned on returning a point in step 2, the point z is returned with probability proportional to $h^{(i)}_{\text{scond}}(z)$. Thus, the distribution sampled by this procedure is indeed $H^{(i)}_{\text{scond}}(z)$. To bound the probability of success, note that the total volume of C is $2^d \times H_{\max}$. On the other hand, step 2 is successful only if z' falls in a region of volume Z_i. This finishes the proof. ◀

The next proposition says that if $Z_i \ge 1/2$, then there is an approximate evaluation oracle for the density $H^{(i)}$.

▶ **Proposition 24.** *Suppose $Z_i \ge 1/2$. Then there is a $(1+O(\varepsilon))$- approximate evaluation oracle for $H^{(i)}$ which can be computed at any point w in time $O\left(\frac{H^2_{\max}}{\varepsilon^2}\right)$.*

Proof. Note that we can evaluate $h^{(i)}$ at any point w exactly and thus the only issue is to estimate the normalizing factor Z_i. Note that since $Z_i \geq 1/2$, estimating Z_i to within an additive $O(\varepsilon)$ gives us a $(1 + O(\varepsilon))$ multiplicative approximation to Z_i and hence to $H^{(i)}(w)$ at any point w. However, by Observation 21, this takes time $O\left(\frac{H_{\max}^2}{\varepsilon^2}\right)$, concluding the proof. ◀

We now apply Proposition 20 as follows.

1. For all $i \in [D]$, estimate Z_i using Observation 21 up to an additive error ε. Let the estimates be $\widehat{Z}_i$.
2. Let us define $S_{\text{feas}} = \{i \in [D] : \widehat{L}_i \geq 1/2\}$.
3. We run the routine $\mathsf{Select}^{\mathbf{D}}$ on the densities $\{H^{(i)}\}_{i \in S_{\text{feas}}}$. To sample from a density $H^{(i)}$, we use Proposition 23. We also construct a $\beta = \varepsilon/32$ approximation oracle for each of the densities $H^{(i)}$ using Proposition 24. Return the output of $\mathsf{Select}^{\mathbf{D}}$.

The correctness of the procedure follows quite easily. Namely, note that Corollary 22 implies that there is one i such that both $Z_i \in [1 - O(\varepsilon), 1 + O(\varepsilon)]$ and $\int_x |H^{(i)}(x) - f(x)| = O(\varepsilon)$. Thus such an i will be in S_{feas}. Thus, by the guarantee of $\mathsf{Select}^{\mathbf{D}}$, the output hypothesis is $O(\varepsilon)$ close to f.

We now bound the sample complexity and time complexity of this hypothesis selection portion of the algorithm. First of all, the number of samples required from f for running $\mathsf{Select}^{\mathbf{D}}$ is $O((1/\varepsilon^2) \cdot (\log(1/\delta) + d^2 \log d + \log\log(1/\delta)) = O((1/\varepsilon^2) \cdot (\log(1/\delta) + d^2 \log d)$. This is clearly dominated by the sample complexity of the previous parts. To bound the time complexity, note that the time complexity of invoking the sampling oracle for any $H^{(i)}$ $(i \in S_{\text{feas}})$ is dominated by the time complexity of the approximate oracle which is $2^{O(d)} \cdot H_{\max}^2/\varepsilon^2$. The total number of calls to the sampling as well as evaluation oracle is upper bounded by $\frac{1}{\varepsilon^2}(D \log D + D \log(1/\delta))$. Plugging in the value of $H_{\max}$ as well as D, we see that the total time complexity is dominated by the bound in the statement of Theorem 12. This finishes the proof.

5 Efficiently learning multivariate log-concave densities

In this section we present our main application, which is an efficient algorithm for learning d-dimensional log-concave densities. We prove the following:

▶ **Theorem 25.** *There is an algorithm with the following property: Let f be a unknown log-concave density over $\mathbb{R}^d$ Given any error parameter $\varepsilon > 0$ and confidence parameter $\delta > 0$ and access to independent draws from f, the algorithm with probability $1 - \delta$ outputs a hypothesis density $h : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ such that $\int_{x \in \mathbb{R}^d} |f(x) - h(x)| \leq O(\varepsilon)$. The algorithm runs in time $O_d\left(\left(\frac{1}{\varepsilon}\right)^{2d+2} \log^{7d}\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon\delta}\right) \log\frac{1}{\delta}\right)$ and uses $O_d\left(\left(\frac{1}{\varepsilon}\right)^{d+2} \log^{4d}\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon\delta}\right) \log\frac{1}{\delta}\right)$ samples.*

We will establish Theorem 25 in two stages. First, we will show that any log-concave f that is nearly isotropic in fact belongs to a suitable class $\mathcal{C}_{\text{SI}}(c, d)$; given this, the theorem follows immediately from Theorem 12 and a straightforward tracing through of the resulting time and sample complexity bounds. Then, we will reduce to the near-isotropic case, similarly to what was done in [37, 4].

First, let us state the theorem for the well-conditioned case. For this, the following definitions will be helpful.

▶ **Definition 26.** Let Σ and $\tilde{\Sigma}$ be two positive semidefinite matrices. We say that Σ and $\tilde{\Sigma}$ are *C-approximations* of each other (denoted by $\Sigma \approx_C \tilde{\Sigma}$) if for every $x \in \mathbb{R}^n$ such that $x^T\tilde{\Sigma}x \neq 0$, we have $\frac{1}{C} \leq \frac{x^T\Sigma x}{x^T\tilde{\Sigma}x} \leq C$.

▶ **Definition 27.** Say that the probability distribution is C-nearly-isotropic if its covariance matrix C-approximates I, the d-by-d identity matrix.

▶ **Theorem 28.** *There is an algorithm with the following property: Let f be a unknown C-nearly-isotropic log-concave density over $\mathbb{R}^d$, where C and d are constants.*

Given any error parameter $\varepsilon > 0$ and confidence parameter $\delta > 0$ and access to independent draws from f, the algorithm with probability $1-\delta$ outputs a hypothesis density $h : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ such that $\int_{x\in\mathbb{R}^d} |f(x) - h(x)| \leq O(\varepsilon)$. The algorithm runs in time $O_{C,d}\left(\left(\frac{1}{\varepsilon}\right)^{2d+2} \log^{7d}\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon\delta}\right) \log\frac{1}{\delta}\right)$ and uses $O_{C,d}\left(\left(\frac{1}{\varepsilon}\right)^{d+2} \log^{4d}\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon\delta}\right) \log\frac{1}{\delta}\right)$ samples.

By Theorem 12, Theorem 28 is an immediate consequence of the following theorem on the shift-invariance of near-isotropic log-concave distributions.

▶ **Theorem 29.** *Let f be a C-nearly-isotropic log-concave density in $\mathbb{R}^d$, for constants C and d. Then, for $g(t) = e^{-\Omega(t)}$, there is a constant $c_1 = O_{C,d}(1)$ such that $f \in \mathcal{C}_{\mathrm{SI}}(c_1, d, g)$.*

Proof. The fact that f has $e^{-\Omega(t)}$-light tails directly follows from Lemma 5.17 of [37], so it remains to prove that there is a constant c_1 such that $f \in \mathcal{C}_{\mathrm{SI}}(c_1, d)$. Because membership in $\mathcal{C}_{\mathrm{SI}}(c_1, d)$ requires that a condition be satisfied for all directions v, rotating a distribution does not affect its membership in $\mathcal{C}_{\mathrm{SI}}(c_1, d)$.

Choose a unit vector v and $\kappa > 0$. By rotating the distribution if necessary, we may assume that $v = e_1$, and our goal of showing that $\mathrm{SI}(f, e_1, \kappa) \leq c_1$ is equivalent to showing that $\int |f(x) - f(x + \kappa' e_1)|dx \leq c_1\kappa$ for all $\kappa' \leq \kappa$.

We bound the integral of the LHS as follows. Fix some value of $x' \stackrel{\text{def}}{=} (x_2, \ldots, x_d)$. Let us define $L_{x'} \stackrel{\text{def}}{=} \{(x_1, x_2, \ldots, x_d) : x_1 \in \mathbb{R}\}$ to be the line through $(0, x_2, \ldots, x_d)$ and $(1, x_2, \ldots, x_d)$. Since the restriction of a concave function to a line is concave, the restriction of a log-concave distribution to a line is log-concave. Since

$$\int |f(x) - f(x + \kappa' e_1)|\, dx = \int_{x'} \int_{x_1} |f(x_1, x_2, ..., x_d) - f(x_1 + \kappa', x_2, ..., x_d)|\, dx_1 dx' \quad (8)$$

we are led to examine the one-dimensional log-concave measure $f(\cdot, x_2, ..., x_d)$. The following will be useful for that.

▶ **Claim 30.** *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a log-concave measure. Then, $\int |\ell(t) - \ell(t+h)|dt \leq 3h \cdot \max_{t\in\mathbb{R}} \ell(t)$.*

Proof. Log-concave measures are unimodal (see [31]). Let z be the mode of ℓ, so that ℓ is non-decreasing on the interval $[-\infty, z]$ and non-increasing in $[z, \infty]$. We have

$$\begin{aligned}
&\int |\ell(t) - \ell(t+h)|\, dt \\
&= \int_{-\infty}^{z-h} |\ell(t) - \ell(t+h)|\, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)|\, dt + \int_{z}^{\infty} |\ell(t) - \ell(t+h)|\, dt
\end{aligned}$$

$$= \int_{-\infty}^{z-h} \ell(t+h) - \ell(t)\, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)|\, dt + \int_{z}^{\infty} \ell(t) - \ell(t+h)\, dt$$

(since z is the mode of ℓ)

$$= \int_{z-h}^{z} \ell(t)\, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)|\, dt + \int_{z}^{z+h} \ell(t)\, dt \le 3h \max_{t\in\mathbb{R}} \ell(t). \quad \blacktriangleleft$$

Returning to the proof of Theorem 29, applying Claim 30 with (8), we get

$$\int |f(x) - f(x+\kappa' e_1)|\, dx \le 3\kappa' \int_{x'} \left(\max_{x_1 \in L_{x'}} f(x_1, x') \right) dx'. \tag{9}$$

Now, since an isotropic log-concave distribution g satisfies $g(x) \le K \exp(-\|x\|)$ for an absolute constant K (see Theorem 5.1 of [40]), our C-nearly-isotropic log-concave distribution f satisfies $f(x) \le C^d K \exp(-\|x\|) = O_{C,d}(\exp(-\|x\|))$. Plugging this into (9), we get

$$\begin{aligned} \int |f(x) - f(x+\kappa' e_1)|\, dx &\le O_{C,d}(\kappa') \int_{x'} \left(\max_{x_1 \in L_{x'}} \exp(-\|(x_1, x')\|) \right) dx' \\ &\le O_{C,d}(\kappa') \int_{x'} \exp(-\|x'\|)\, dx'. \end{aligned}$$

Since the integral converges, this finishes the proof. $\blacktriangleleft$

To learn log-concave distributions that are not C-nearly-isotropic, using techniques from [37], we preprocess the data to bring it into isotropic position, and then apply Theorem 29. The details are in the long version of this paper [20].

6 Learning shift-invariant densities over $\mathbb{R}^d$ with bounded support requires $\Omega(1/\varepsilon^d)$ samples

The following lower bound is proved in the long version of this paper [20].

▶ **Theorem 31.** *Given $d \ge 1$, there is a constant $c_d = \Theta(\sqrt{d})$ such that the following holds: For all sufficiently small ε, let A be an algorithm with the following property: given access to m i.i.d. samples from an arbitrary (and unknown) finitely supported density $f \in \mathcal{C}_{\mathrm{SI}}(c_d, d)$, with probability at least 99/100, A outputs a hypothesis density h such that $d_{\mathrm{TV}}(f, h) \le \varepsilon$. Then $m \ge \Omega((1/\varepsilon)^d)$.*

References

1. J. Acharya, C. Daskalakis, and G. Kamath. Optimal Testing for Properties of Distributions. In *NIPS*, pages 3591–3599, 2015.
2. J. Acharya, A. Jafarpour, A. Orlitsky, and A.T. Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures, 2014. http://arxiv.org/abs/1402.4746.
3. Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.
4. Maria-Florina Balcan and Philip M Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.
5. Andrew D Barbour and Aihua Xia. Poisson perturbations. *ESAIM: Probability and Statistics*, 3:131–150, 1999.
6. Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

7 OV Besov. On a family of function spaces. Embedding and extension theorems. *Dokl. Akad. Nauk SSSR*, 126(6):1163–1165, 1959.

8 Aditya Bhaskara, Ananda Suresh, and Morteza Zadimoghaddam. Sparse solutions to nonnegative linear systems and applications. In *Artificial Intelligence and Statistics*, pages 83–92, 2015.

9 C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing Shape Restrictions of Discrete Distributions. In *STACS*, pages 25:1–25:14, 2016.

10 T. Carpenter, I. Diakonikolas, A. Sidiropoulos, and A. Stewart. Near-Optimal Sample Complexity Bounds for Maximum Likelihood Estimation of Multivariate Log-concave Densities. *CoRR*, abs/1802.10575, 2018.

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

12 S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient Density Estimation via Piecewise Polynomial Approximation. In *STOC*, pages 604–613, 2014.

13 L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

14 S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

15 C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for Poisson multinomials and its applications. In *STOC*, pages 1074–1086, 2016.

16 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.

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

18 C. Daskalakis and G. Kamath. Faster and Sample Near-Optimal Algorithms for Proper Learning Mixtures of Gaussians. In *COLT*, pages 1183–1213, 2014.

19 A. De, I. Diakonikolas, and R. Servedio. Learning from Satisfying Assignments. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 478–497, 2015.

20 A. De, P. M. Long, and R. A. Servedio. Density estimation for shift-invariant multidimensional distributions, 2018. `arXiv:1811.03744`.

21 A. De, P. M. Long, and R. A. Servedio. Learning Sums of Independent Random Variables with Sparse Collective Support. *FOCS*, 2018.

22 Ronald A DeVore and Robert C Sharpley. Besov spaces on domains in $\Re^d$. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.

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

24 Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

25 I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient Robust Proper Learning of Log-concave Distributions. *CoRR*, abs/1606.03077, 2016.

26 I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the Fourier transform for sums of independent integer random variables. In *COLT*, pages 831–849, 2016.

27 I. Diakonikolas, D. M. Kane, and A. Stewart. The Fourier transform of Poisson multinomial distributions and its algorithmic applications. In *STOC*, pages 1060–1073, 2016.

28 Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, 2010.

29 Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning multivariate log-concave distributions. In *Conference on Learning Theory (COLT)*, pages 711–727, 2016.

30 Lasse Holmström and Jussi Klemelä. Asymptotic bounds for the expected L1 error of a multivariate kernel density estimator. *Journal of multivariate analysis*, 42(2):245–266, 1992.

31 Il'dar Abdullovich Ibragimov. On the composition of unimodal distributions. *Theory of Probability & Its Applications*, 1(2):255–260, 1956.

32 S. Johnson. Saddle-point integration of C-infinity "bump" functions. arXiv preprint, 2015. `arXiv:1508.04376`.

33 A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

34 Daniel M Kane, Jelani Nelson, and David P Woodruff. On the exact space complexity of sketching and streaming small norms. In *SODA*, 2010.

35 A.K.H. Kim and R.J. Samworth. Global rates of convergence in log-concave density estimation. Available at arXiv, 2014. `arXiv:1404.2298`.

36 Jussi Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley Publishing, 2009.

37 László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.

38 Elias Masry. Multivariate probability density estimation by wavelet methods: Strong consistency and rates for stationary time series. *Stochastic processes and their applications*, 67(2):177–193, 1997.

39 A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

40 A. Saumard and J.A. Wellner. Log-Concavity and Strong Log-Concavity: a review. Technical report, ArXiV, 23 April 2014. `arXiv:1404.5886`.

41 Winthrop W Smith and Joanne M Smith. Handbook of real-time fast Fourier transforms. *IEEE, New York*, 1995.

42 Sergej Lvovich Sobolev. On a theorem of functional analysis. *Am. Math. Soc. Transl.*, 34:39–68, 1963.

43 Rebecca M Willett and Robert D Nowak. Multiscale Poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.

44 Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.