# Density estimation for shift-invariant multidimensional distributions

Anindya De[*]
Northwestern University
anindya@eecs.northwestern.edu

Philip M. Long
Google
plong@google.com

Rocco A. Servedio[†]
Columbia University
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 $\varepsilon$ 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 [DKS16b]. 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.

## 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 [DG85, DL12].

[*]Supported by NSF grant CCF-1814706
[†]Supported by NSF grants CCF-1319788 and CCF-1420349

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 [Das99, KMV10, MV10] 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 [Sob63, BC91, HK92, DL12]. 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 [Bes59, DS93, Mas97, WN07, ADLS17]. 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 [DS93]), 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 [BX99, DDO$^+$13, DLS18]. The assumption is based on the *shift-invariance of $f$ in direction $v$ at scale $\kappa$*, which, for a density $f$ over $\mathbb{R}^d$, a unit vector $v \in \mathbb{R}^d$, and a positive real value $\kappa$, we define to be

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

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.

$$\mathrm{SI}(f, \kappa) \stackrel{\mathrm{def}}{=} \sup_{v : \|v\|_2 = 1} \mathrm{SI}(f, v, \kappa).$$

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 $\mathrm{SI}(f, \kappa) \le c$ 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

$$\Pr_{\boldsymbol{x} \leftarrow f}[||\boldsymbol{x} - \mu|| > t] \leq g(t),$$

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 noise-tolerant 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 $\varepsilon$ 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.

Our learning results will hold even in a challenging model of *noise-tolerant* density estimation for a class $\mathcal{C}$. In this framework, the density estimation algorithm is given access to i.i.d. draws from $f'$, which is a mixture $f' = (1 - \varepsilon)f + \varepsilon f_{\mathrm{noise}}$ where $f \in \mathcal{C}$ and $f_{\mathrm{noise}}$ may be any density. (We will sometimes say that such an $f'$ is an *$\varepsilon$-corrupted* version of $f$. This model of noise is sometimes referred to as *Huber's contamination model* [Hub67].) Now the goal of the density estimation algorithm is to output a description of a hypothesis density $h$ over $\mathbb{R}^d$ which, with probability at least (say) 9/10 over the draws from $f'$, satisfies $d_{\mathrm{TV}}(f', h) \leq O(\varepsilon)$. This is a challenging variant of the usual density estimation framework, especially for multidimensional density estimation. In particular, there are simple distribution learning problems (such as learning a single Gaussian or product distribution over $\{0, 1\}^n$) which are essentially trivial in the noise-free setting, but for which computationally efficient noise-tolerant learning algorithms have proved to be a significant challenge [DKK$^+$16, DKK$^+$18, SCV18].

## 1.1 Results

Our main positive result is a general algorithm which efficiently learns any class $\mathcal{C}_{\mathrm{SI}}(c, d, g)$ in the noise-tolerant model described above. 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 noise-tolerantly learned to any error $O(\varepsilon)$ with a sample complexity that depends on $c, g, \varepsilon$ 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 29 in Section 5 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 [CDSS13, CDSS14, ADLS17, ADK15, CDGR16, DKS16a] for work on density estimation of univariate (discrete and continuous) log-concave distributions. In the multivariate case, [KS14] 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 $\varepsilon$. More recently, [DKS16b] established the first finite sample complexity upper bound for multivariate log-concave densities, by giving an algorithm that *semi-agnostically* (i.e. noise-tolerantly in a very strong sense) learns any $d$-dimensional log-concave density using $\tilde{O}_d(1/\varepsilon^{(d+5)/2})$ samples. The algorithm of [DKS16b] 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. [CDSS18] showed that the maximum likelihood estimator (MLE) is statistically efficient (i.e., achieves near optimal sample complexity). However, we note that the MLE is computationally inefficient and thus has no bearing on the question of finding an efficient algorithm for learning log-concave densities.

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 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 log-concave density belongs to $\mathcal{C}_{\mathrm{SI}}(c, d, g)$ where the tail bound function $g$ is inverse exponential. Theorem 29 then immediately implies the following, answering the open question of [DKS16b]:

**Theorem 1.** *There is an algorithm with the following property: Let $f$ be a unknown log-concave density over $\mathbb{R}^d$ and let $f'$ be an $\varepsilon$-corruption of $f$. 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 $\tilde{O}_d(1/\varepsilon^{2d+2}) \cdot \log^2(1/\delta)$ and uses $\tilde{O}_d(1/\varepsilon^{d+2}) \cdot \log^2(1/\delta)$ many samples.*

While our sample complexity is quadratically larger than the optimal sample complexity for learning log-concave distributions (from [DKS16b]), such *computational-statistical* tradeoffs are in fact quite common (see, for example, the work of [BSZ15] 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 $\varepsilon$ 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 $\varepsilon$-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.[1]

[1] We note that a simpler version of this approach, which only uses a smoothing kernel and does not employ Fourier analysis, can

Next, in Section 4, we consider the more general case in which the target distribution belongs to the class $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ (so at this point we are not yet in the noise-tolerant framework). 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 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 consider the most general case of noise-tolerant density estimation for $\mathcal{C}_{\mathrm{SI}}(c,d,g)$. Recall that in this setting the target density $f'$ is some distribution which need not actually belong to $\mathcal{C}_{\mathrm{SI}}(c,d,g)$ but satisfies $d_{\mathrm{TV}}(f',f) \le \varepsilon$ for some density $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$. It turns out that this case can be handled using essentially the same algorithm as the previous paragraph. We show that even in the noise-tolerant setting, our transformation algorithm will still successfully find a transformation as above that would succeed if the target density were $f \in \mathcal{C}_{\mathrm{SI}}(c,d,g)$ rather than $f'$. (This robustness of the transformation algorithm crucially relies on the fact that it only uses a small number of draws from the given distribution to be learned.) We then show that after transforming $f'$ in this way, the original algorithm for the special case can in fact learn the transformed version of $f'$ to high accuracy; then, as in the previous paragraph, performing the inverse transformation gives a high-accuracy hypothesis for $f'$.

In Section 6 we apply the above results to establish efficient noise-tolerant 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 6 we show that any $d$-dimensional isotropic log-concave density is $O_d(1)$-shift-invariant. Armed with this bound, by applying our noise-tolerant learning result (Theorem 29) we get that any $d$-dimensional isotropic log-concave density can be noise-tolerantly 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 [LV07].

**The lower bound.** As is standard, our lower bound (proved in Section 7) is obtained via Fano's inequality. We identify a large set $\mathcal{F}$ of bounded-support shift-invariant $d$-dimensional densities with the following two properties: all pairs of densities from $\mathcal{F}$ have KL-divergence that is not too big (so that they are hard to tell apart), but also have total variation distance that is not too small (so that a successful learning algorithm is required to tell them apart). The members of $\mathcal{F}$ are obtained by choosing functions that take one of two values in each cell of a $d$-dimensional checkerboard. The two possible values are within a small constant factor of each other, which keeps the KL divergence small. To make the total variation distance large, we choose the values using an error-correcting code – this means that distinct members of $\mathcal{F}$ have different

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.

values on a constant fraction of the cells, which leads to large variation distance.

### 1.3 Related work

The most closely related work that we are aware of was mentioned above: [HK92] 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. [HK92] 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 [Kle09]) 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 [BX99, DDO$^+$13]. Likewise, both the discrete and continuous Fourier transforms have been used in the past to learn discrete probability distributions [DKS16c, DKS16d, DDKT16].

## 2 Preliminaries

We write $B(r)$ to denote the radius-$r$ ball in $\mathbb{R}^d$, i.e. $B(r) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \le r^2\}$. 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. [CGS11] and many references therein). We give a natural continuous analogue of this notion below.

**Definition 2.** *Given a probability density $f$ over $\mathbb{R}^d$, a unit vector $v$, and a positive real value $\kappa$, we say that the* shift-invariance of $f$ in direction $v$ at scale $\kappa$, *denoted* $\mathrm{SI}(f, v, \kappa)$*, is*

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

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 $\kappa'$ 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 $\kappa$ gets small; the effect of shifting by $\kappa$ is measured relative to $\kappa$.

Let

$$\mathrm{SI}(f, \kappa) \stackrel{\mathrm{def}}{=} \sup\{\mathrm{SI}(f, v, \kappa) : v \in \mathbb{R}^d, \|v\|_2 = 1\}.$$

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 $\mathrm{SI}(f, \kappa) \leq c$ 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 $\mathrm{SI}(f, v, \kappa) \leq c\kappa$ 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 $\kappa$. 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 $\xi$ is unaffected by a shift of size $\xi$, 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 $\mathbf{Pr}_{\boldsymbol{x} \leftarrow f}\left[||\boldsymbol{x} - \mu|| > t\right] \leq g(t)$, where $\mu \in \mathbb{R}^d$ is the mean of $f$.

**Remark 3.** *It will be convenient in our analysis to consider only tail bound functions $g$ that satisfy* $\min\{r \in \mathbb{R} : g(r) \leq 1/2\} \geq 1/10$ *(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 $\inf\{t : g(t) \leq \varepsilon\}$.

The complexity of learning with a tail bound $g$ will be expressed in part using

$$I_g \stackrel{\mathrm{def}}{=} \int_0^\infty g(\sqrt{z})\, dz.$$

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 4.** *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 [SS95].

**Claim 5.** *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 6** (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 7** (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 8** (Young's inequality)**.** *Let $f \in L_p([-1,1]^d)$, $g \in L_q([-1,1]^d)$, $1 \leq p, q, r \leq \infty$, such that $1 + 1/r = 1/p + 1/q$. Then $\|f * g\|_r \leq \|f\|_p \cdot \|g\|_q$.*

### 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 $C^\infty$ 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 [KNW10, DKN10].) The following fact can be found in [Joh15] (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 9.** *For $b : [-1,1] \to \mathbb{R}^+$ defined in (2) and $\xi \in \mathbb{Z} \setminus \{0\}$, we have that $|\widehat{b}(\xi)| \leq e^{-\sqrt{|\xi|}} \cdot |\xi|^{-3/4}$.*

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 9, we have the following claim:

**Claim 10.** *For $\xi \in \mathbb{Z}^d$ with $\|\xi\|_\infty \geq t$, we have $|\widehat{b_{d,\gamma}}(\xi)| \leq e^{-\sqrt{\gamma \cdot t}} \cdot (\gamma \cdot t)^{-3/4}$.*

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

**Fact 11.** *$\|b_{d,\gamma}\|_\infty = (c_0/\gamma)^d$ and as a consequence, $\|b_{d,\gamma}\|_2^2 \leq (c_0/\gamma)^{2d}$.*

## 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 $\varepsilon$, an algorithm often only needs the density to be learned to have shift invariance at a scale slightly finer than $\varepsilon$.

**Lemma 12.** *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 $\kappa \mathrm{SI}(f, \kappa) \leq \varepsilon/2$, given access to independent draws from $f$, the algorithm runs in time*

$$O_d\left(\frac{1}{\varepsilon^2}\left(\frac{1}{\kappa}\right)^{2d} \log^{4d}\left(\frac{1}{\kappa}\right) \log\left(\frac{1}{\kappa\delta}\right)\right)$$

*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* $\int_{x\in\mathbb{R}^d} |f(x)-h(x)| \leq \varepsilon$.

*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* $h(z) \leq O_d\left(\frac{\log^{2d}(1/\kappa)}{\kappa^d}\right)$.

*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 $\gamma \leq 1/2$, 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 8), we have that $\|q\|_2 \leq \|f\|_1 \cdot \|b_{d,\gamma}\|_2$. Noting that $f$ is a density and thus $\|f\|_1 = 1$ and applying Fact 11, 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 $\gamma\sqrt{d} \leq \kappa$. Fix any direction $v$ and consider a shift of $f$ in direction $v$ by some distance at most $\gamma\sqrt{d} \leq \kappa$. Since $\kappa \mathrm{SI}(f,\kappa) \leq \varepsilon/2$, 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

$$d_{\mathrm{TV}}(q,f) \leq \varepsilon/2. \tag{3}$$

Next, we observe that by Claim 5, 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, $|\widehat{f}(\xi)| \leq 1$, and thus we have $|\widehat{q}(\xi)| \leq |\widehat{b_{d,\gamma}}(\xi)|$. 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 10, we obtain

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

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

*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 13.** *For any* $\xi \in \mathbb{Z}^d$*, the Fourier coefficient* $\widehat{q}(\xi)$ *can be estimated to within additive error of magnitude at most* $\eta$ *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 $\mathsf{Low} = \{\xi \in \mathbb{Z}^d : \|\xi\|_\infty \le T\}$. Thus, $|\mathsf{Low}| \le (2T+1)^d$. Thus, using $S = O(\eta^{-2} \cdot \log(T/\delta))$ draws from $f$, by Observation 13, with probability $1-\delta$, we can compute a set of values $\{\widehat{u}(\xi)\}_{\xi \in \mathsf{Low}}$ such that

$$\text{For all } \xi \in \mathsf{Low},\ |\widehat{u}(\xi) - \widehat{q}(\xi)| \le \eta. \tag{4}$$

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 (4) with this, we get

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

Thus, setting $\eta$ as $\eta^2 = (2T+1)^{-d} \cdot \frac{\varepsilon^2}{8}$, we get that

$$\sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 \le \frac{\varepsilon^2}{4}. \tag{5}$$

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 7) to obtain a function $u : [-1,1]^d \to \mathbb{C}$ such that

$$\int_{[-1,1]^d} |u(x) - q(x)|^2 dx = \frac{1}{2^d} \cdot \Big( \sum_{\xi \in \mathbb{Z}^d} |\widehat{u}(\xi) - \widehat{q}(\xi)|^2 \Big) \le \frac{\varepsilon^2}{4 \cdot 2^d}, \tag{6}$$

where the first equality follows by Parseval's identity (Claim 6). By the Cauchy-Schwarz inequality,

$$\int_{[-1,1]^d} |u(x) - q(x)| dx \le \sqrt{2^d} \cdot \sqrt{\int_{[-1,1]^d} |u(x) - q(x)|^2 dx}.$$

Plugging in (6), we obtain $\int_{[-1,1]^d} |u(x) - q(x)| dx \le \frac{\varepsilon}{2}$. 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

$$\int_{[-1,1]^d} |h(x) - q(x)| dx \le \int_{[-1,1]^d} |u(x) - q(x)| dx \le \frac{\varepsilon}{2}. \tag{7}$$

Finally, recalling that by (3) we have $d_{\mathrm{TV}}(f, q) \le \frac{\varepsilon}{2}$, it follows that $\int_{[-1,1]^d} |h(x) - f(x)| dx \le \varepsilon$.

*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 $\gamma$ 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 $\eta$ (set just above (5)), 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 7) 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. □

With an eye towards our ultimate goal of obtaining noise-tolerant density estimation algorithms, the next corollary says that the algorithm in Lemma 12 is robust to noise. All the parameters have the same meaning and relations as in Lemma 12.

**Corollary 14.** *Let $f'$ be a density supported in $B(1/2)^2$ such that there is a $d$-dimensional density $f$ satisfying the following two properties: (i) $f$ satisfies all the conditions in the hypothesis of Lemma 12, and (ii) $f'$ is an $\varepsilon$-corrupted version of $f$, i.e. $f' = (1-\varepsilon)f + \varepsilon f_{\text{noise}}$ for some density $f_{\text{noise}}$. Then given access to samples from $f'$, the algorithm* learn-bounded *returns a hypothesis $h : [-1,1]^d \to \mathbb{R}^+$ which satisfies $\int_{x\in\mathbb{R}^d} |f'(x) - h(x)| \le 2\varepsilon$. All the other guarantees including the sample complexity and time complexity remain the same as Lemma 12.*

*Proof.* The proof of Lemma 12 can be broken down into two parts:

- $f$ can be approximated by $q$, and

[2] Looking ahead, while in general an "$\varepsilon$-noisy" version of $f$ need not be supported in $B(1/2)$, the reduction we employ will in fact ensure that we only need to deal with noisy distributions that are in fact supported in $B(1/2)$.

- $q$ can be learned.

The argument that $q$ can be learned only used two facts about it:

- it is supported in $[-1,1]^d$, and
- it has few nonzero Fourier coefficients.

So, now consider the distribution $q' = f' * b_{d,\gamma}$ where $b_{d,\gamma}$ is the same distribution as in Lemma 12. Because $q'$ is the result of convolving $f'$ (a density supported in $B(1/2)$ with $b_{d,\gamma}$, it is supported in $[-1,1]^d$, and has the same Fourier concentration property that we used for $q$. Thus, the algorithm will return a hypothesis distribution $h(x)$ such that the analogue of (7) holds, i.e.

$$\int_{[-1,1]^d} |h(x) - q'(x)| dx \leq \frac{\varepsilon}{2}. \tag{8}$$

Recalling that the density $f'$ can be expressed as $(1-\varepsilon)f + \varepsilon f_{\text{noise}}$ where $f_{\text{noise}}$ is some density supported in $B(1/2)$, we now have

$$\begin{aligned} d_{\text{TV}}(q', f') = d_{\text{TV}}(f' * b_{d,\gamma}, f') &= d_{\text{TV}}((1-\varepsilon)f * b_{d,\gamma} + \varepsilon f_{\text{noise}} * b_{d,\gamma}, (1-\varepsilon)f + \varepsilon f_{\text{noise}}) \\ &\leq (1-\varepsilon) d_{\text{TV}}(f * b_{d,\gamma}, f) + \varepsilon d_{\text{TV}}(f_{\text{noise}} * b_{d,\gamma}, f_{\text{noise}}) \\ &\leq \varepsilon/2 + \varepsilon \leq 3\varepsilon/2. \end{aligned}$$

The penultimate inequality uses (3) and the fact that the total variation distance between any two distributions is bounded by 1. Combining the above with (8), the corollary is proved. □

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

Fix any nonincreasing tail bound function $g : \mathbb{R}^+ \to [0,1]$ which satisfies $\lim_{t\to+\infty} g(t) = 0$ and the condition $\min\{r \in \mathbb{R} : g(r) \leq 1/2\} \geq 1/10\}$ of Remark 3 and any constant $c \geq 1$. In this section we prove the following theorem which gives a density estimation algorithm for the class of distributions $\mathcal{C}_{\text{SI}}(c, d, g)$:

**Theorem 15.** *For any $c, g$ as above and any $d \geq 1$, there is an algorithm with the following property: Let $f$ be any target density (unknown to the algorithm) which belongs to $\mathcal{C}_{\text{SI}}(c, d, g)$. 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 $h : [-1,1]^d \to \mathbb{R}^{\geq 0}$ such that $\int_{x\in\mathbb{R}^d} |f(x) - h(x)| \leq O(\varepsilon)$.*

*The algorithm runs in time*

$$O_{c,d}\left(\left((g^{-1}(\varepsilon))^{2d} \left(\frac{1}{\varepsilon}\right)^{2d+2} \log^{4d}\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right) \log\left(\frac{g^{-1}(\varepsilon)}{\varepsilon\delta}\right) + I_g\right) \log\frac{1}{\delta}\right)$$

*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 15 is proved by a reduction to Lemma 12. The main ingredient in the proof of Theorem 15 is a "transformation algorithm" with the following property: given as input access to i.i.d. draws from any density $f \in \mathcal{C}_{\rm SI}(c, d, g)$, 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 12 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 12.) 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 15 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 16.** *Let $f \in \mathrm{SI}(c, d)$ and let $B$ be a ball such that $\mathbf{Pr}_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B] \geq 1 - \delta$ 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 $f \in \mathcal{C}_{\rm SI}(c, d)$ the left hand side is at most $c\kappa$, whereas the subtrahend on the right hand side is trivially at most $2\delta$. Thus, we get

$$\int_x |f_B(x) - f_B(x + \kappa v)|dx \leq \frac{2\delta}{1-\delta} + \frac{c\kappa}{1-\delta}, \tag{9}$$

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

*Proof.* We have $\mathbf{E}_{\boldsymbol{x}\sim f}[||\boldsymbol{x}-\mu||^2] = \int_0^\infty \mathbf{Pr}_{\boldsymbol{x}\sim f}[||\boldsymbol{x}-\mu||^2 \ge z]\, dz \le \int_0^\infty g(\sqrt{z})\, dz = I_g$. □

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

**Lemma 18.** *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*

$$\mathbf{Pr}\left[||\mu - \widehat{\boldsymbol{\mu}}||^2 \ge t\right] \le \frac{I_g}{Mt}.$$

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

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

where the last inequality is by Lemma 17. Applying Markov's inequality on the left hand side, we get the stated claim. □

### 4.3 Transformation algorithm

**Lemma 19.** *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 $\mathbf{Pr}_{\boldsymbol{x}\sim f}[\boldsymbol{x}\in B_t] \ge 1-\varepsilon$.*
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 by Remark 3 we have $t = \Theta(g^{-1}(\varepsilon)^2)$.). Let $\mu$ denote the true mean of $f$. First, by Lemma 18, with probability at least 0.9, the empirical mean $\widehat{\boldsymbol{\mu}}$ will be close to the true mean $\mu$ in the following sense:

$$||\mu - \widehat{\boldsymbol{\mu}}||^2 \le \frac{1}{10}. \tag{10}$$

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

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

Applying Lemma 16 completes the proof. □

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

**Lemma 20.** *Let $f \in \mathcal{C}_{\rm SI}(c, d, g)$, $\varepsilon > 0$, $\tilde{\mu} \in \mathbb{R}^d$, and $t \in \mathbb{R}$ satisfy the properties stated in Lemma 19. Consider the density $f_{\rm scond}$ defined by*

$$f_{\rm scaled}(x) \stackrel{\rm def}{=} 2\sqrt{t} \cdot f\big(2\sqrt{t} \cdot (x + \tilde{\mu})\big)$$
$$f_{\rm scond}(x) \stackrel{\rm def}{=} f_{{\rm scaled}, B(1/2)}(x),$$

*where $f_{{\rm scaled}, B(1/2)}$ is the result of conditioning $f_{\rm scaled}$ on membership in $B(1/2)$. Then the density $f_{\rm scond}(x)$ satisfies the following properties:*

1. *The density $f_{\rm scond}$ is supported in the ball $B(1/2)$.*

2. *For all $\varepsilon < 1/2$ and $\kappa > 0$, the density $f_{\rm scond}$ satisfies*

$$\mathrm{SI}(f_{\rm scond}, \kappa) \le \frac{4\varepsilon}{\kappa} + 4c\sqrt{t}.$$

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

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

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

$$\begin{aligned} \mathrm{SI}(f_{\rm scaled}, v, \kappa) &= \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0,\kappa]} \int_{\mathbb{R}^d} \Big| f(u + \kappa' 2\sqrt{t} v) - f(u) \Big| \, du \\ &= \frac{1}{\kappa} \cdot \sup_{\kappa' \in [0, 2\sqrt{t}\kappa]} \int_{\mathbb{R}^d} \big| f(u + \kappa' v) - f(u) \big| \, du \\ &= 2\sqrt{t} \cdot \mathrm{SI}(f, v, 2\sqrt{t}\kappa) \le 2c\sqrt{t}. \end{aligned} \tag{11}$$

The last inequality uses that $f \in \mathcal{C}_{\rm SI}(c, d, g)$. Inequality (11) implies that $f_{\rm scaled} \in \mathcal{C}_{\rm SI}(2c\sqrt{t}, d, g)$. Now, $\mathbf{Pr}_{\boldsymbol{x} \sim f_{\rm scaled}}(\boldsymbol{x} \in B(1/2)) = \mathbf{Pr}_{\boldsymbol{x} \sim f}(\boldsymbol{x} \in B_t) \ge 1 - \varepsilon$, so applying Lemma 16 completes the proof. □

## 4.4 Proof of Theorem 15

We are now ready to prove Theorem 15. 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 $\varepsilon$ 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 $B_t^{(i)} = \{x : ||x - \tilde{\mu}|| \le \sqrt{t}\}$ and $f_{\rm scond}^{(i)}$ be the density defined from $(\tilde{\mu}^{(i)}, t)$ as in Lemma 20.

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 $B_t^{(i)}$. We further make the following observations:

- If $\mathbf{Pr}_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] \geq 1/2$, 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 19, then $\mathbf{Pr}_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] \geq 1 - \varepsilon$ and we fall into this case.
- On the other hand, if $\mathbf{Pr}_{\boldsymbol{x}\sim f}[\boldsymbol{x} \in B_t^{(i)}] < 1/2$, then it may be inefficient to simulate draws from $f_{\text{scond}}^{(i)}$. But any such $i$ will not satisfy the properties of Lemma 19, 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 $i \in [D]$,[3] 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 12. 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 20, i.e.

$$h^{(i)}(x) \stackrel{\text{def}}{=} \frac{1}{2\sqrt{t}} \cdot h_{\text{scond}}^{(i)}\left(\frac{1}{2\sqrt{t}} \cdot (x - \tilde{\mu}^{(i)})\right). \tag{12}$$

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 19 and Lemma 20, 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 21.** *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 19 and the choice of $D$ that with probability $1 - \delta$ at least one triple $(\tilde{\mu}^{(i)}, t)$ satisfies the properties of Lemma 19. 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 19, by Lemma 20, taking $\kappa = \min\{\varepsilon/2, \varepsilon/(4g^{-1}(\varepsilon)c)\}$) the density $f_{\text{scond}}^{(i')}$ satisfies the required conditions for Lemma 12 to apply with that choice of $\kappa$. The following simple proposition implies that $h^{(i)}$ is likewise $O(\varepsilon)$-close to $f_{B_t}$:

**Proposition 22.** *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* $d_{\text{TV}}(f, g) = d_{\text{TV}}(f_A, g_A)$.

[3] 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.

*Proof.*

$$\begin{aligned} d_{\mathrm{TV}}(f_A, g_A) &= \int_x |f_A(x) - g_A(x)| dx = \int_x \det(A) |f(A(x-z)) - g(A(x-z))| dx \\ &= \int_z |f(z) - g(z)| dz = d_{\mathrm{TV}}(f, g), \end{aligned}$$

where the penultimate equality follows by a linear transformation of variables. □

It remains only to observe that by property 1 of Lemma 19 the density $f_{B_t}$ is $\varepsilon$-close to $f$, and then by the triangle inequality we have that $h^{(i)}$ is $O(\varepsilon)$-close to $f$. This gives Lemma 21. □

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 15. 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. [Yat85, DK14, AJOS14, DDS12, DDS15]). The following is implicit in the proof of Proposition 6 from [DDS15]:

**Proposition 23.** *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 $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_i) \leq \varepsilon$. There is an algorithm $\mathrm{Select}^{\mathbf{D}}$ which is given $\varepsilon$ and a confidence parameter $\delta$, 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 $\frac{\mathbf{D}_i(w)}{1+\beta} \leq \tilde{D}_i^\beta(w) \leq (1+\beta) \cdot \mathbf{D}_i(w)$. Further, $(1+\beta)^2 \leq (1+\varepsilon/8)$. 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 $d_{\mathrm{TV}}(\mathbf{D}, \mathbf{D}_{i^\star}) \leq 6\varepsilon$.*

As suggested above, the remaining step is to apply Proposition 23 to the list of candidate hypothesis $h^{(i)}$ which satisfies the guarantee of Lemma 21. However, to bound the sample and time complexity of running the procedure Proposition 23, 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.[4] 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 12 (recall that Lemma 12 was applied with $\kappa = \min\{\varepsilon/2, \varepsilon/(4g^{-1}(\varepsilon)c)\}$) we get that

$$H_{\max} = O_{c,d}\left(\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right)^d \log^{2d} \frac{g^{-1}(\varepsilon)}{\varepsilon}\right). \tag{13}$$

We will carry out the rest of our calculations in terms of $H_{\max}$.

[4]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 24.** *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 24 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 (12)), we get that 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

$$H^{(i)}(x) = \frac{h^{(i)}(x)}{Z_i} \quad\text{and}\quad H^{(i)}_{\text{scond}}(x) = \frac{h^{(i)}_{\text{scond}}(\frac{1}{2\sqrt{t}}\cdot(x-\tilde{\mu}^{(i)}))}{Z_i}\cdot\frac{1}{2\sqrt{t}}.$$

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 $\text{Select}^{\mathbf{D}}$ on the densities $H^{(i)}$. To see this, note that Lemma 21 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 25.** *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 $\text{Select}^{\mathbf{D}}$ on the candidate distributions $\{H^{(i)}\}_{i\in[D]}$. The next proposition shows that the densities $\{H^{(i)}\}_{i\in[D]}$ are samplable.

**Proposition 26.** *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 $z_{d+1} \leq h^{(i)}_{\text{scond}}(z_1, \ldots, z_d)$, 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 $Z_i \geq 1/2$, then there is an approximate evaluation oracle for the density $H^{(i)}$.

**Proposition 27.** *Suppose $Z_i \geq 1/2$. Then there is a $(1 + O(\varepsilon))$- approximate evaluation oracle for $H^{(i)}$ which can be computed at any point $w$ in time $O\left(\frac{H_{\max}^2}{\varepsilon^2}\right)$.*

*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 24, this takes time $O\left(\frac{H_{\max}^2}{\varepsilon^2}\right)$, concluding the proof. □

We now apply Proposition 23 as follows.

1. For all $i \in [D]$, estimate $Z_i$ using Observation 24 up to an additive error $\varepsilon$. 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 26. We also construct a $\beta = \varepsilon/32$ approximation oracle for each of the densities $H^{(i)}$ using Proposition 27. Return the output of $\mathsf{Select}^{\mathbf{D}}$.

The correctness of the procedure follows quite easily. Namely, note that Corollary 25 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 $S_{\text{feas}}$. 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 15. This finishes the proof.

# 5 Noise-tolerant density estimation for $\mathcal{C}_{\text{SI}}(c, d, g)$

Fix any nonincreasing tail bound function $g : \mathbb{R}^+ \to [0, 1]$ which satisfies $\lim_{t\to+\infty} g(t) = 0$ and the condition of Remark 3 and any constant $c \geq 1$. In this section we prove our main result, Theorem 29, which gives a noise tolerant density estimation algorithm for $\mathcal{C}_{\text{SI}}(c, d, g)$. We first recall the precise model of noise we consider in this paper.

**Definition 28.** *For two densities $f$ and $f' \in \mathbb{R}^d$, we say that $f'$ is an $\varepsilon$-corruption of $f$ if $f'$ can be expressed as $f' = (1 - \varepsilon) \cdot f + \varepsilon \cdot f_{err}$ where $f_{err}$ is a density in $\mathbb{R}^d$.*

This model of noise is sometimes referred to as *Huber's contamination model* [Hub67].

**Theorem 29** (Noise-tolerant density estimation for $\mathcal{C}_{\text{SI}}(c, d, g)$.)**.** *For any $c, g$ as above and any $d \geq 1$, there is an algorithm with the following property: Let $f$ be any density (unknown to the algorithm) which belongs to $\mathcal{C}_{\text{SI}}(c, d, g)$ and let $f'$ be an $\varepsilon$-corruption of $f$. Given $\varepsilon$ and any confidence parameter $\delta > 0$*

*and access to independent draws from $f'$, the algorithm with probability $1 - O(\delta)$ outputs a hypothesis $h : [-1, 1]^d \to \mathbb{R}^{\geq 0}$ such that $\int_{x\in\mathbb{R}^d} |f'(x) - h(x)| \leq O(\varepsilon)$.*

*The algorithm runs in time*

$$O_{c,d}\left(\exp(O(I_g)) \cdot \left((g^{-1}(\varepsilon))^{2d}\left(\frac{1}{\varepsilon}\right)^{2d+2}\log^{4d}\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right)\log\left(\frac{g^{-1}(\varepsilon)}{\varepsilon\delta}\right) + I_g\right)\log\frac{1}{\delta}\right)$$

*and uses*

$$O_{c,d}\left(\exp(O(I_g)) \cdot \left((g^{-1}(\varepsilon))^{d}\left(\frac{1}{\varepsilon}\right)^{d+2}\log^{2d}\left(\frac{g^{-1}(\varepsilon)}{\varepsilon}\right)\log\left(\frac{g^{-1}(\varepsilon)}{\varepsilon\delta}\right) + I_g\right)\log\frac{1}{\delta}\right)$$

*samples.*

Theorem 29 is identical to Theorem 15 except that now the target density from which draws are received is $f'$, which is an $\varepsilon$-corruption of $f \in \mathcal{C}_{\mathrm{SI}}(c, d, g)$, rather than $f$ itself, and the dependence on $I_g$ is exponential. (On the other hand, if $f$ is isotropic, then recalling Lemma 17 the function $g$ can be taken to be such that $I_g = \mathbf{E}_{\boldsymbol{x}\sim f}[||\boldsymbol{x} - \mu||^2] = d$, so that $\exp(O(I_g)) = O(1)$ and the complexity is the same as in the noise-free case.) This requires essentially no changes in the algorithm and fortunately most of the analysis from earlier can also be reused in a fairly black-box way. We briefly explain how the analysis of Section 4 can be augmented to handle having access to draws from $f'$ rather than $f$.

*Proof of Theorem 29.* Without loss of generality, we can assume that $\varepsilon \leq 1/10$, as otherwise any hypothesis is trivially $O(\varepsilon)$-close to the target density. Recall that $f'$ can be expressed as $f' = (1 - \varepsilon) \cdot f + \varepsilon \cdot f_{err}$ for some density $f_{err} \in \mathbb{R}^d$. Since $\varepsilon \leq 1/10$, this means that with probability $9/10$, a random sample from $f'$ is in fact distributed exactly as a random sample from $f$. We now revisit the steps in the algorithm construct-candidates (Proof of Theorem 15) and briefly sketch why sample access to $f'$ instead of $f$ suffices.

**Step 1:** Note that each invocation of the algorithm compute-transformation is supposed to draw $O(I_g)$ samples from $f$. We have access to $f'$ rather than $f$, but since each sample from $f'$, with probability at least $9/10$, is a sample from $f$, with probability at least $\exp(-O(I_g))$, a run of compute-transformation with samples from $f'$ is the same as a run with samples from $f$. So now in Step 1, the algorithm compute-transformation is run $\exp(-O(I_g)) \cdot \ln(1/\delta))$ many times rather than $O(\ln(1/\delta))$ many times as in the original version (this accounts for the additional $\exp(O(I_g))$ factor in the bounds of Theorem 29 versus Theorem 15).

**Step 2:** This goes exactly as before with no changes. In particular, for every $i$, if $f^{(i)}_{\mathrm{trans}}$ is the true density obtained by the $i^{th}$ transformation, then we have sample access to ${f'}^{(i)}_{\mathrm{trans}}$ which is an $\varepsilon$-corruption of $f^{(i)}_{\mathrm{trans}}$.

**Step 3:** In Step 3, we run the routine learn-bounded as usual. In particular, let us assume that $f^{(i)}_{\mathrm{trans}}$ satisfies the conditions of Lemma 12. Then Corollary 14, which established noise-tolerance of learn-bounded, implies that with sample access to ${f'}^{(i)}_{\mathrm{trans}}$, the resulting hypothesis ${h'}^{(i)}_{\mathrm{trans}}$ is $2\varepsilon$-close to $f^{(i)}_{\mathrm{trans}}$.

It is easy to verify that Step 4 and the subsequent steps in hypothesis testing can go on exactly as before with sample access to $f'$ instead of $f$. In particular, the hypothesis testing routine $\mathrm{Select}^D$ will output a hypothesis $h^{(i)}$ which is $2\varepsilon$ close to $f$. ☐

## 6 Efficiently learning multivariate log-concave densities

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

**Theorem 30** (Restatement of Theorem 1). *There is an algorithm with the following property: Let $f$ be a unknown log-concave density over $\mathbb{R}^d$ and let $f'$ be an $\varepsilon$-corruption of $f$. 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 30 in two stages. First, we will show that any log-concave $f$ that is nearly isotropic in fact belongs to a suitable class $\mathcal{C}_{\mathrm{SI}}(c, d)$; given this, the theorem follows immediately from Theorem 29 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 [LV07, BL13].

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

**Definition 31.** *Let $\Sigma$ and $\tilde{\Sigma}$ be two positive semidefinite matrices. We say that $\Sigma$ 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 32.** *Say that the probability distribution is $C$-nearly-isotropic if its covariance matrix $C$-approximates $I$, the $d$-by-$d$ identity matrix.*

**Theorem 33.** *There is an algorithm with the following property: Let $f$ be a unknown $C$-nearly-isotropic log-concave density over $\mathbb{R}^d$ and let $f'$ be an $\varepsilon$-corruption of $f$, 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 29, Theorem 33 is an immediate consequence of the following theorem on the shift-invariance of near-isotropic log-concave distributions.

**Theorem 34.** *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 [LV07], 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

$$\int |f(x) - f(x + \kappa' e_1)| dx \leq c_1 \kappa \tag{14}$$

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

$$\int |f(x) - f(x + \kappa' e_1)| \, dx = \int_{x'} \int_{x_1} |f(x_1, x_2, ..., x_d) - f(x_1 + \kappa', x_2, ..., x_d)| \, dx_1 dx' \tag{15}$$

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 35.** *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 [Ibr56]). Let $z$ be the mode of $\ell$, so that $\ell$ is non-decreasing on the interval $[-\infty, z]$ and non-increasing in $[z, \infty]$. We have

$$\begin{aligned}
&\int |\ell(t) - \ell(t+h)| \, dt \\
&= \int_{-\infty}^{z-h} |\ell(t) - \ell(t+h)| \, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)| \, dt + \int_{z}^{\infty} |\ell(t) - \ell(t+h)| \, dt \\
&= \int_{-\infty}^{z-h} \ell(t+h) - \ell(t) \, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)| \, dt + \int_{z}^{\infty} \ell(t) - \ell(t+h) \, dt \\
&\hspace{20em} \text{(since } z \text{ is the mode of } \ell) \\
&= \int_{z-h}^{z} \ell(t) \, dt + \int_{z-h}^{z} |\ell(t) - \ell(t+h)| \, dt + \int_{z}^{z+h} \ell(t) \, dt \\
&\leq 3h \max_{t \in \mathbb{R}} \ell(t).
\end{aligned}$$

□

Returning to the proof of Theorem 34, applying Claim 35 with (15), we get

$$\int |f(x) - f(x + \kappa' e_1)|\, dx \le 3\kappa' \int_{x'} \left( \max_{x_1 \in L_{x'}} f(x_1, x') \right) dx'. \tag{16}$$

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 [SW14]) , 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 (16), 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. □

Finally, we turn to the problem of learning log-concave distributions that are not $C$-nearly-isotropic. To learn in this case, we need to rescale the axes if necessary to transform the distribution so that it is $C$-nearly-isotropic. To compute this rescaling, we will first assume that there is no noise present in the samples. (The trick to handle noise is simple and will be discussed later.) First, as has often been observed, we may assume without loss of generality that the covariance matrix of the target has full rank, since, otherwise the algorithm can efficiently find the affine span of the entire distribution (possibly up to a negligible amount of probability mass), and the algorithm can be carried out within that lower dimensional subspace. To bring the distribution to nearly isotropic position, we will be using ideas from [LV07]. (We require the additional analysis below, rather than invoking their results as a black box, to cope with the fact that the mean is unknown.)

Our starting point is the following lemma due to Lovász and Vempala [LV07].

**Lemma 36.** *Let $f$ be a zero-mean log-concave density on $\mathbb{R}^d$. For $m = O(d \log^3 d)$, if $\Sigma$ denotes the (population) covariance matrix of $f$ and $\widehat{\mathbf{\Sigma}}$ is the empirical covariance matrix from $m$ samples of $f$, then, with probability $9/10$, $\widehat{\mathbf{\Sigma}}$ is a $11/10$ approximation to $\Sigma$.*

Lemma 36 enables us to estimate the covariance matrix if we know the mean. To apply it when we do not, we appear to need an estimate of the mean that is especially good in directions with low variance. The following is aimed at obtaining such an estimate.

Recall that, for a set $A$ of real-valued functions on a common domain $X$, the *pseudo-dimension* of $A$, which is denoted by $\mathrm{Pdim}(A)$, is the VC-dimension of the set of indicator functions, one for each $a \in A$, for whether $(x, y)$ satisfies $a(x) \ge y$. We will use the following standard VC bound.

**Lemma 37** ([Tal94])**.** *For any set $A$ of functions with a common domain $X$ and ranges contained in $[-M, M]$, for any distribution $D$, $m = O(\frac{M^2(\mathrm{Pdim}(A) + \log(1/\gamma))}{\varepsilon^2})$ suffices for a set $S$ of $m$ examples drawn according to $D$, with probability $1 - \gamma$, to have all $a \in A$ have*

$$\left| \mathbf{E}_{\boldsymbol{x} \sim D}(a(\boldsymbol{x})) - \frac{1}{m} \sum_{x' \in S} a(x') \right| \le \varepsilon.$$

The proof of the following lemma follows a similar lemma in [KLS09].

**Lemma 38.** *Fix a function $b$ from $\mathbb{R}^d$ to $\mathbb{R}^+$. Define $a_u = b(u) \cdot (u \cdot x)$. The pseudo-dimension of $\{a_u : u \in B(1)\}$ is $O(d)$.*

*Proof.* Any $(x, y)$ satisfies $a_u(x) \geq y$ iff

$$b(u)(u \cdot x) \geq y.$$

Thus, the set of indicator functions for $a_u(x) \geq y$ can be embedded into the set of homogeneous halfspaces over $\mathbb{R}^{d+1}$, which is known to have VC-dimension $O(d)$. □

Now we are ready for the result we require on estimating the mean:

**Lemma 39.** *Fix any log-concave distribution $f$ over $\mathbb{R}^d$ and any $\alpha > 0$. For $m = O\left(\frac{d \log^2(d/\alpha)}{\alpha^2}\right)$, with probability at least $3/4$, a multiset $S$ of $m$ samples drawn i.i.d. from $f$ satisfies, for all unit length $u$,*

$$\frac{|\mathbf{E}_{\boldsymbol{x} \sim S}(u \cdot \boldsymbol{x}) - \mathbf{E}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})|}{\sqrt{\mathbf{Var}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})}} \leq \alpha. \tag{17}$$

*Proof.* Translating the distribution $f$ translates both $\mathbf{E}_{\boldsymbol{x} \sim S}(u \cdot \boldsymbol{x})$ and $\mathbf{E}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})$ the same way, and does not affect $\mathbf{Var}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})$, so we may assume without loss of generality that $f$ has zero mean. Let $f_B$ be the distribution obtained from $f$ by conditioning the choice of $x$ on the event that $|u \cdot x| \leq \sqrt{\mathbf{Var}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})} \cdot \frac{\ln(8m)}{c}$ for all unit length $u$, where $c$ is a large constant. Lemma 5.17 of [LV07] implies that, for large enough $c$, the total variation distance between $f$ and $f_B$ is $1/(8m)$, so that the total variation distance between $m$ draws from $f$ and $m$ draws from $f_B$ is at most $1/8$. We henceforth assume that the $m$ draws from $f$ are in fact drawn from $f_B$, and proceed to analyze $f_B$.

For any unit length $u$, define $a_u$ by $a_u(x) = \frac{|u \cdot x|}{\sqrt{\mathbf{Var}_{\boldsymbol{x} \sim f}(u \cdot \boldsymbol{x})}}$. Lemma 38 implies that $\{a_u : u \in B(1)\}$ has pseudo-dimension $O(d)$. Furthermore, when $x$ is chosen from the support of $f_B$, each $a_u$ takes values in an interval of size $O(\log m)$. Thus we may apply Lemma 37 to obtain Lemma 39. □

Now we are ready to present and analyze the transformation.

**Lemma 40.** *There is an algorithm* rescale *such that given access to samples from a log-concave distribution $f$, and an error parameter $\varepsilon > 0$, the algorithm takes $O(d \log^3 d)$ samples from $f$ and with probability at least $1/2$ produces a non-singular positive definite matrix $\tilde{\Sigma} \in \mathbb{R}^{d \times d}$ such that, if $\Sigma$ is the covariance matrix of $f$, for any unit vector $v$,*

$$\frac{1}{2} \leq \frac{v^T \Sigma v}{v^T \tilde{\Sigma} v} \leq 2.$$

*Proof.* For a large constant $C$ and $M = C d \log^3 d$ the algorithm rescale first uses $M$ examples to construct an estimate $\tilde{\boldsymbol{\mu}}$ of the mean of $f$, and then uses $\tilde{\boldsymbol{\mu}}$ to use the examples estimate the covariance matrix.

Lemma 39 implies that, if $C$ is large enough, then with probability $3/4$, for all unit length $v$, we have

$$\frac{|\mu \cdot v - \tilde{\boldsymbol{\mu}} \cdot v|}{\sqrt{\mathbf{Var}_{\boldsymbol{x} \sim f}[v \cdot \boldsymbol{x}]}} \leq \frac{1}{10}. \tag{18}$$

Lemma 36 implies that, with probability $3/4$ over a random i.i.d. draw of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \sim f$, we have

$$9/10 \leq \frac{\frac{1}{M} \sum_{i=1}^{M} (v \cdot \boldsymbol{x}_i - v \cdot \mu)^2}{\mathbf{Var}_{\boldsymbol{x} \sim f}[v \cdot \boldsymbol{x}]} \leq 11/10. \tag{19}$$

We henceforth assume that both (18) and (19) hold (this happens with probability at least $1/2$), and we let $\tilde{\mu}$ and $x_1, \ldots, x_M$ denote the corresponding outcomes.

Let $\Sigma$ be the true co-variance of $f$, and let $\tilde{\Sigma}$ be the estimate that was used (which depends on $\tilde{\mu}$).

We have that

$$\begin{aligned}
\frac{v^T\tilde{\Sigma}v}{v^T\Sigma v} &= \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\tilde{\mu})^2}{v^T\Sigma v} \\
&= \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&= \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu + v\cdot\mu - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&\le \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} + \frac{\frac{2|v\cdot\mu - v\cdot\tilde{\mu}|}{M}\sum_{i=1}^M |v\cdot x_i - v\cdot\mu|}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} + \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot\mu - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&= \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} + \frac{2|v\cdot\mu - v\cdot\tilde{\mu}|}{M\sqrt{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}}\sum_{i=1}^M \frac{|v\cdot x_i - v\cdot\mu|}{\sqrt{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}} + \frac{(v\cdot\mu - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&\le \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} + \frac{2|v\cdot\mu - v\cdot\tilde{\mu}|}{M\sqrt{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}} \times \sqrt{M\sum_{i=1}^M \frac{(v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}} + \frac{(v\cdot\mu - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&= \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} + \frac{2|v\cdot\mu - v\cdot\tilde{\mu}|}{\sqrt{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}} \times \sqrt{\frac{1}{M}\sum_{i=1}^M \frac{(v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]}} + \frac{(v\cdot\mu - v\cdot\tilde{\mu})^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&\le 11/10 + 2(1/10)\sqrt{11/10} + 1/100 \le 2,
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz and the third inequality is by (19) and (18). Similarly,

$$\begin{aligned}
\frac{v^T\tilde{\Sigma}v}{v^T\Sigma v} &\ge \frac{\frac{1}{M}\sum_{i=1}^M (v\cdot x_i - v\cdot\mu)^2}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} - \frac{\frac{2|v\cdot\mu - v\cdot\tilde{\mu}|}{M}\sum_{i=1}^M |v\cdot x_i - v\cdot\mu|}{\mathbf{Var}_{\boldsymbol{x}\sim f}[v\cdot\boldsymbol{x}]} \\
&\ge 9/10 - 2(1/10)\sqrt{11/10} \ge 1/2,
\end{aligned}$$

completing the proof. □

*Proof of Theorem 30.* The basic algorithm (for the noise-free setting) applies the procedure rescale from Lemma 40 to find an estimate of the covariance matrix of $f$, rescales the axes so that the transformed distribution is 2-nearly-isotropic, learns the transformed distribution, and then rescales the axes again to restore their original scales.

In the presence of noise, Lemma 40 succeeds with probability 1/2, if all the examples are not noisy. But since the noise rate is at most $\varepsilon$, and we may assume without loss of generality that $\varepsilon < 1/10$, since the number of examples required in Lemma 40 is independent of $\varepsilon$, any invocation of the method succeeds in the presence of noise with probability $\Omega_d(1)$, which is at least some positive constant (since $d$ is a constant). Thus, if an algorithm performs $O_d(\log(1/\delta))$ many repetitions, with probability at least $1-\delta/2$ one of them will succeed. It can therefore call the algorithm of Theorem 29 $O(\log(1/\delta))$ times, and then applying the hypothesis testing procedure of Proposition 23 to the results, to achieve the claimed result. □

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

In this section we give a simple lower bound which shows that $\Omega(1/\varepsilon^d)$ samples are required for $\varepsilon$-accurate density estimation even of shift-invariant $d$-dimensional densities with bounded support. As discussed in the introduction, densities with bounded support may be viewed as satisfying the strongest possible rate of tail decay as they have zero tail mass outside of a bounded region.

**Theorem 41.** *Given $d \geq 1$, there is a constant $c_d = \Theta(\sqrt{d})$ such that the following holds: For all sufficiently small $\varepsilon$, 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 $d_{\mathrm{TV}}(f, h) \leq \varepsilon$. Then $m \geq \Omega((1/\varepsilon)^d)$.*

Since an algorithm that achieves a small error with high probability can be used to achieve small error in expectation, to prove Theorem 41 it suffices to show that any algorithm that achieves expected error $O(\varepsilon)$ must use $\Omega((1/\varepsilon)^d)$ samples. To establish this we use Lemma 42 (given below), which provides a lower bound on the number of examples needed for small expected error.

To obtain the desired lower bound from Lemma 42, we establish the existence of a family $\mathcal{F}$ of densities $\mathcal{F} = \{f_1, \ldots, f_N\} \in \mathcal{C}_{\mathrm{SI}}(c_d, d)$, where $N = \exp(\Omega((1/\varepsilon)^d))$. These densities will be shown to satisfy the following two properties: for any $i \neq j \in [N]$ we have (1) $d_{\mathrm{TV}}(f_i, f_j) = \Omega(\varepsilon)$, and (2) the Kullback-Leibler divergence $D_{KL}(f_i || f_j)$ is at most $O(1)$, yielding Theorem 41.

## 7.1 Fano's inequality

The main tool we use for our lower bound is Fano's inequality, or more precisely, the following extension of it given by [IH81] and [AB83]:

**Theorem 42** (Generalization of Fano's Inequality.)**.** *Let $f_1, \ldots, f_{N+1}$ be a collection of $N+1$ distributions such that for any $i \neq j \in [N+1]$, we have (i) $d_{\mathrm{TV}}(f_i, f_j) \geq \alpha/2$, and (ii) $D_{KL}(f_i || f_j) \leq \beta$, where $D_{KL}$ denotes Kullback-Leibler divergence. Then for any algorithm that makes $m$ draws from an unknown target distribution $f_i$, $i \in [N+1]$, and outputs a hypothesis distribution $\tilde{f}$, there is some $i \in [t+1]$ such that if the target distribution is $f_i$, then*

$$\mathbf{E}[d_{\mathrm{TV}}(f, \tilde{f})] \geq \frac{\alpha}{2}\left(1 - \frac{m\beta + \ln 2}{\ln N}\right).$$

*In particular, to achieve expected error at most $\alpha/4$, any learning algorithm must have sample complexity $m = \Omega\left(\frac{\ln N}{\beta}\right)$.*

## 7.2 The family of densities we analyze

Let $T$ be a positive integer that is $T = \lceil C/\varepsilon \rceil$ for a large constant $C$. We consider probability densities over $\mathbb{R}^d$ which (i) are supported on $[-T, T)^d$, and (ii) are piecewise constant on each of the $(2T)^d$ many disjoint unit cubes whose union is $[-T, T)^d$. Writing $A$ to denote the set $\{-T, -T+1, \ldots, -1, 0, 1, \ldots, T-1\}^d$, each of the $(2T)^d$ many disjoint unit cubes mentioned above is indexed by a unique element $a = (a_1, \ldots, a_d) \in A$ in the obvious way. We write $\mathrm{cube}(a)$ to denote the unit cube indexed by $a$. Given $x \in [-T, T)^d$ we write $a(x)$ to denote the unique element $a \in A$ such that $x \in \mathrm{cube}(a)$.

For any $z \in \{0,1\}^A$, we define a probability density $f_z$ over $[-T,T)^d$ as $f_z(x) = (T + z_{a(x)})/Z$, where $Z = \Theta((2T)^{(d+1)})$ is a normalizing factor so that $f_z$ is indeed a density (i.e. it integrates to 1).

It is well known (via an elementary probabilistic argument) that there is a subset $S \subset \{0,1\}^A$ of size $2^{\Theta(|A|)}$ such that any two distinct strings $z, z' \in S$ differ in $\Theta(|A|)$ many coordinates. We define the set $\mathcal{F}$ of densities to be $\mathcal{F} = \{f_z : z \in S\}$.

### 7.3 Membership in $\mathcal{C}_{\mathrm{SI}}(c_d, d)$

It is obvious that every density in $\mathcal{F}$ is finitely supported. In this subsection we prove that $\mathcal{F} \subseteq \mathcal{C}_{\mathrm{SI}}(c_d, d)$. First, we bound the variation distance incurred by shifting along a coordinate axis:

**Lemma 43.** *For any $f \in \mathcal{F}$, $i \in [d]$, and $\kappa \in (0,1)$, we have $\int |f(x + \kappa e_i) - f(x)|\, dx \le \Theta(\kappa\varepsilon)$.*

*Proof.* We have

$$\int |f(x+\kappa e_i) - f(x)|\, dx = \int_{\{x:x_i<-T\}} |f(x+\kappa e_i) - f(x)|\, dx + \int_{\{x:x_i>T-\kappa\}} |f(x+\kappa e_i) - f(x)|\, dx + \int_{\{x:x_i\in[-T,T-\kappa]\}} |f(x+\kappa e_i) - f(x)|\, dx.$$

If $x_i < -T - \kappa$, then $f(x+\kappa e_i) = f(x) = 0$. When $-T-\kappa \le x_i < -T$, we have $|f(x+\kappa e_i) - f(x)| \le (T+1)/Z$. Thus

$$\int_{\{x:x_i<-T\}} |f(x+\kappa e_i) - f(x)|\, dx \le (\kappa(2T)^{d-1})(T+1)/Z = \Theta(\kappa\varepsilon).$$

A similar argument gives that $\int_{\{x:x_i>T-\kappa\}} |f(x+\kappa e_i) - f(x)|\, dx \le \Theta(\kappa\varepsilon)$. Finally,

$$\int_{\{x:x_i\in[-T,T-\kappa]\}} |f(x+\kappa e_i) - f(x)|\, dx = \int_{\{x:x_i\in[-T,T-\kappa],\lceil x_i\rceil - x_i\le\kappa\}} |f(x+\kappa e_i) - f(x)|\, dx \le \int_{\{x:x_i\in[-T,T-\kappa],\lceil x_i\rceil - x_i\le\kappa\}} (1/Z)\, dx.$$

The set $\{x : x_i \in [-T, T-\kappa], \lceil x_i\rceil - x_i \le \kappa\}$ is made up of $2T-1$ "slabs" that are each of width $\kappa$, and consequently $\int_{\{x:x_i\in[-T,T-\kappa],\lceil x_i\rceil - x_i\le\kappa\}} \le (2T-1)\kappa(2T)^{d-1}/Z = \Theta(\kappa/T) = \Theta(\kappa\varepsilon)$, recalling that $Z = \Theta((2T)^{d+1})$. This completes the proof. ∎

Given Lemma 43, it is easy to bound the variation distance incurred by shifting in an arbitrary direction:

**Lemma 44.** *For any $f \in \mathcal{F}$, for any unit vector $v$, for any $\kappa < 1$, we have $\int |f(x+\kappa v) - f(x)|\, dx \le c\kappa\varepsilon\sqrt{d}$ for a universal constant $c > 0$.*

*Proof.* Writing the unit vector $v$ as $\sum_{j=1}^{d} v_j e_j$, we have

$$\begin{aligned}\int |f(x+\kappa v) - f(x)|\, dx &= \int \left| \sum_{i=1}^{d} f\left(x + \kappa \sum_{j=1}^{i} v_j e_j\right) - f\left(x + \kappa \sum_{j=1}^{i-1} v_j e_j\right) \right| dx \\ &\leq \sum_{i=1}^{d} \int \left| f\left(x + \kappa \sum_{j=1}^{i} v_j e_j\right) - f\left(x + \kappa \sum_{j=1}^{i-1} v_j e_j\right) \right| dx \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{(triangle inequality)} \\ &= \sum_{i=1}^{d} \int |f\left(x + \kappa |v_i| e_i\right) - f\left(x\right)|\, dx \qquad \text{(variable substitution)} \\ &\leq c\kappa\varepsilon \sum_{i=1}^{d} |v_i| \leq c\kappa\varepsilon\sqrt{d}, \qquad \text{(Lemma 43 and Cauchy-Schwarz)}\end{aligned}$$

completing the proof. □

As an easy consequence of Lemma 44, Lemma 17 and the definition of $\mathrm{SI}(f, v, \kappa)$ we obtain the following:

**Corollary 45.** *There is a constant* $c_d = \Theta(\sqrt{d})$ *such that for any* $f \in \mathcal{F}$ *and any unit vector* $v \in \mathbb{R}^d$*, we have* $\mathrm{SI}(f, v) \leq c_d$*. Hence* $\mathcal{F} \subseteq \mathcal{C}_{\mathrm{SI}}(c_d, d)$*.*

### 7.4 The upper bound on $KL$ divergence and lower bound on variation distance

Recall that if $f$ and $g$ are probability density functions supported on a set $S \subseteq \mathbb{R}^d$, then the *Kullback Leibler divergence* between $f$ and $g$ is defined as $D_{KL}(f||g) = \int_S f(x) \ln \frac{f(x)}{g(x)} dx$. As an immediate consequence of this definition, we have the following claim:

**Claim 46.** *Let* $f, g$ *be two densities such that for some absolute constant* $C > 1$ *we have that every* $x$ *satisfies* $\frac{1}{C} f(x) \leq g(x) \leq C f(x)$*. Then* $D_{KL}(f||g) \leq O(1)$*.*

It is easy to see that any $f_i, f_j$ in the family $\mathcal{F}$ of densities described in Section 7.2 are such that $\frac{1}{C} f_i(x) \leq f_j(x) \leq C f_i(x)$. Thus we have:

**Lemma 47.** $D_{KL}(f_i||f_j) \leq O(1)$ *for all* $i \neq j \in [N]$*.*

Finally, we need a lower bound on the total variation distance between any pair of elements of $\mathcal{F}$:

**Lemma 48.** *There is an absolute constant* $c > 0$ *such that, for any* $f_u, f_v \in \mathcal{F}$*,* $d_{\mathrm{TV}}(f_u, f_v) = \Omega(\varepsilon)$*.*

*Proof.* We have $d_{\mathrm{TV}}(f_u, f_v) = (1/Z)|\{a \in A : u_a \neq v_a\}| = (1/Z)\Omega(|A|) = \Omega(1/T) = \Omega(\varepsilon)$. □

### 7.5 Putting it together

By Lemma 48, each pair of elements of $\mathcal{F}$ are separated by $\Omega(\varepsilon)$ in total variation distance. Since Lemma 47 implies that each pair of elements of $\mathcal{F}$ has KL-divergence $O(1)$, Theorem 42 implies that $\Omega(\ln |\mathcal{F}|) = \Omega((1/\varepsilon)^d)$ examples are needed to achieve expected error at most $O(\varepsilon)$. Since Corollary 45 gives that $\mathcal{F} \subseteq \mathcal{C}_{\mathrm{SI}}(c_d, d)$ for a constant $c_d = \Theta(\sqrt{d})$, this proves Theorem 41.

## References


[AB83] P. Assouad and C. Birge. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I*, 296:1021–1024, 1983.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3591–3599, 2015.

[ADLS17] 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.

[AJOS14] J. Acharya, A. Jafarpour, A. Orlitsky, and A.T. Suresh. Near-optimal-sample estimators for spherical gaussian mixtures. Technical Report http://arxiv.org/abs/1402.4746, ArXiV, 19 Feb 2014.

[BC91] Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

[Bes59] OV Besov. On a family of function spaces. embedding and extension theorems. *Dokl. Akad. Nauk SSSR*, 126(6):1163–1165, 1959.

[BL13] 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.

[BSZ15] Aditya Bhaskara, Ananda Suresh, and Morteza Zadimoghaddam. Sparse solutions to nonnegative linear systems and applications. In *Artificial Intelligence and Statistics*, pages 83–92, 2015.

[BX99] Andrew D Barbour and Aihua Xia. Poisson perturbations. *ESAIM: Probability and Statistics*, 3:131–150, 1999.

[CDGR16] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, 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.

[CDSS14] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDSS18] Timothy Carpenter, Ilias Diakonikolas, Anastasios Sidiropoulos, and Alistair Stewart. Near-Optimal Sample Complexity Bounds for Maximum Likelihood Estimation of Multivariate Log-concave Densities. *CoRR*, abs/1802.10575, 2018.

[CGS11] L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[Das99] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

[DDKT16] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016*, pages 1074–1086, 2016.

[DDO+13] Constantinos Daskalakis, Ilias Diakonikolas, Ryan O'Donnell, Rocco A. Servedio, and Li-Yang Tan. Learning sums of independent integer random variables. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 217–226. IEEE, 2013.

[DDS12] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS15] A. De, I. Diakonikolas, and R. Servedio. Learning from Satisfying Assignments. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 478–497, 2015.

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

[DK14] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *COLT*, pages 1183–1213, 2014.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS*, pages 655–664, 2016.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 2683–2702, 2018.

[DKN10] Ilias Diakonikolas, Daniel M Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, 2010.

[DKS16a] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *CoRR*, abs/1606.03077, 2016.

[DKS16b] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning multivariate log-concave distributions. In *Conference on Learning Theory (COLT)*, pages 711–727, 2016.

[DKS16c] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Optimal learning via the Fourier transform for sums of independent integer random variables. In *Conference on Learning Theory*, pages 831–849, 2016.

[DKS16d] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. The Fourier transform of Poisson multinomial distributions and its algorithmic applications. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1060–1073. ACM, 2016.

[DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[DLS18] A. De, P. M. Long, and R. A. Servedio. Learning sums of independent random variables with sparse collective support. *FOCS*, 2018.

[DS93] 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.

[HK92] 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.

[Hub67] P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

[Ibr56] Ildar Abdullovich Ibragimov. On the composition of unimodal distributions. *Theory of Probability & Its Applications*, 1(2):255–260, 1956.

[IH81] I.A. Ibragimov and R.Z. Has'minskii. *Statistical estimation, asymptotic theory (Applications of Mathematics, vol. 16)*. Springer-Verlag, New York, 1981.

[Joh15] S. Johnson. Saddle-point integration of c-infinity "bump" functions. arXiv preprint arXiv:1508.04376, 2015.

[Kle09] Jussi Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley Publishing, 2009.

[KLS09] A. Klivans, P. Long, and R. Servedio. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[KNW10] Daniel M Kane, Jelani Nelson, and David P Woodruff. On the exact space complexity of sketching and streaming small norms. In *SODA*, 2010.

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

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

[Mas97] 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.

[MV10] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[SCV18] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference, ITCS*, pages 45:1–45:21, 2018.

[Sob63] Sergej Lvovich Sobolev. On a theorem of functional analysis. *Am. Math. Soc. Transl.*, 34:39–68, 1963.

[SS95] Winthrop W Smith and Joanne M Smith. Handbook of real-time fast fourier transforms. *IEEE, New York*, 1995.

[SW14] A. Saumard and J.A. Wellner. Log-concavity and strong log-concavity: a review. Technical Report https://arxiv.org/pdf/1404.5886.pdf, 23 April 2014.

[Tal94] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[WN07] Rebecca M Willett and Robert D Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007.

[Yat85] Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.