Near-optimal Sample Complexity Bounds for Robust Learning of Gaussian Mixtures via Compression Schemes

HASSAN ASHTIANI, McMaster University, Canada and Vector Institute, Canada
SHAI BEN-DAVID, University of Waterloo, Canada
NICHOLAS J. A. HARVEY and CHRISTOPHER LIAW, University of British Columbia, Canada
ABBAS MEHRABIAN, McGill University, Canada
YANIV PLAN, University of British Columbia, Canada

We introduce a novel technique for distribution learning based on a notion of *sample compression*. Any class of distributions that allows such a compression scheme can be learned with few samples. Moreover, if a class of distributions has such a compression scheme, then so do the classes of *products* and *mixtures* of those distributions.

As an application of this technique, we prove that $\widetilde{\Theta}(kd^2/\varepsilon^2)$ samples are necessary and sufficient for learning a mixture of k Gaussians in $\mathbb{R}^d$, up to error ε in total variation distance. This improves both the known upper bounds and lower bounds for this problem. For mixtures of axis-aligned Gaussians, we show that $\widetilde{O}(kd/\varepsilon^2)$ samples suffice, matching a known lower bound. Moreover, these results hold in an agnostic learning (or robust estimation) setting, in which the target distribution is only approximately a mixture of Gaussians. Our main upper bound is proven by showing that the class of Gaussians in $\mathbb{R}^d$ admits a small compression scheme.

CCS Concepts: • **Mathematics of computing** → **Density estimation**; Multivariate statistics; • **Theory of computation** → **Sample complexity and generalization bounds**; Unsupervised learning and clustering;

Additional Key Words and Phrases: Compression schemes, density estimation, distribution learning, mixtures of Gaussians

1 INTRODUCTION

Estimating distributions from observed data is a fundamental task in statistics, which has been
studied for over a century. This task frequently arises in applied machine learning, commonly as-
suming that the distribution can be modeled approximately by a mixture of Gaussians. Popular
software packages have implemented heuristics, such as the expectation maximization (EM) al-
gorithm, for learning a mixture of Gaussians. The theoretical machine learning community has a
rich literature on distribution learning as well; for example, the recent survey of Reference [20]
considers learning structured distributions, and the survey of Reference [27] focuses on mixtures
of Gaussians.

This article develops a general technique for distribution learning, then employs this technique
in the canonical setting of Gaussian mixtures. The learning model we adopt is *density estimation*:
Given i.i.d. samples from an unknown target distribution, find a distribution that is close to the
target in *total variation (TV) distance*. Our analysis focuses on sample complexity rather than com-
putational complexity. That is, we seek a learning algorithm that obtains a good estimate of the
target distribution using as few samples as possible, but we do not worry about its running time.
For background on this model, see, e.g., References [17, 20].

Our new technique for proving upper bounds on the sample complexity involves a novel form
of *sample compression*: If it is possible to "encode" each member of a class of distributions using
a carefully chosen subset of its samples, then we obtain an upper bound on the sample complex-
ity of distribution learning for that class. In particular, by constructing compression schemes for
mixtures of axis-aligned Gaussians and general Gaussians, we obtain new upper bounds on sam-
ple complexities of learning with respect to these classes. Furthermore, we prove that these new
bounds are optimal up to polylogarithmic factors.

1.1 The Distribution Learning Framework

A *distribution learning method* or *density estimation method* is an algorithm that takes as input a
sequence of i.i.d. samples generated from a distribution g and outputs (a description of) a distri-
bution $\widehat{g}$ as an estimate for g. We work with absolutely continuous distributions in this article (i.e.,
distributions that have a density with respect to the Lebesgue measure), so we identify a proba-
bility distribution with its probability density function. The *total variation (TV) distance* between
two probability distributions f_1 and f_2 over $\mathbb{R}^d$ is defined as

$$\mathrm{TV}\,(f_1, f_2) \;:=\; \sup_{B \subseteq \mathbb{R}^d} \int_B \left(f_1(x) - f_2(x) \right) \mathrm{d}x \;=\; \frac{1}{2} \|f_1 - f_2\|_1, \tag{1}$$

where $\|f\|_1 := \int_{\mathbb{R}^d} |f(x)|\,\mathrm{d}x$ is the L^1 *norm* of f, and $\|f_1 - f_2\|_1$ is the L^1 *distance* between f_1 and
f_2. (Strictly speaking, the supremum in Equation (1) should be over the Borel sigma algebra on
$\mathbb{R}^d$; however, we do not worry about such measure-theoretic issues in this article.) Sometimes we
write $\mathrm{TV}\,(X, Y)$, where X and Y are random variables rather than distributions. In the following
definitions, $\mathcal{F}$ is a class of probability distributions, and g is a distribution not necessarily in $\mathcal{F}$.

Definition 1.1 (ε-approximation, ε-close, (ε, C)-approximation). We say a distribution $\widehat{g}$ is an ε-
approximation for g, or $\widehat{g}$ is ε-*close* to g, if $\|\widehat{g} - g\|_1 \le \varepsilon$. A distribution $\widehat{g}$ is an (ε, C)-*approximation*

for g with respect to $\mathcal{F}$ if

$$\|\widehat{g} - g\|_1 \leq C \cdot \inf_{f \in \mathcal{F}} \|f - g\|_1 + \varepsilon.$$

Definition 1.2 (PAC-learning Distributions, Realizable Setting). A distribution learning method is called a *(realizable) PAC-learner* for $\mathcal{F}$ with sample complexity $m_{\mathcal{F}}(\varepsilon, \delta)$ if, for all distributions $g \in \mathcal{F}$ and all $\varepsilon, \delta \in (0, 1)$, given ε, δ, and an i.i.d. sample of size $m_{\mathcal{F}}(\varepsilon, \delta)$ from g, outputs an ε-approximation of g, with probability at least $1 - \delta$ (over the samples and the algorithm's randomness).

Definition 1.3 (PAC-learning Distributions, Agnostic Setting). A distribution learning method is called a *C-agnostic PAC-learner for $\mathcal{F}$* with sample complexity $m_{\mathcal{F}}^C(\varepsilon, \delta)$ if, for all distributions g and all $\varepsilon, \delta \in (0, 1)$, given ε, δ, and a sample of size $m_{\mathcal{F}}^C(\varepsilon, \delta)$ generated i.i.d. from g, outputs an (ε, C)-approximation of g with respect to $\mathcal{F}$, with probability at least $1 - \delta$.

The statement that a class can be "*C*-learned in the agnostic setting" means there exists a *C*-agnostic PAC-learner for the class. The case $C > 1$ is sometimes called *semi*-agnostic learning in the learning theory literature. Note that minimizing the L^1 distance is equivalent to minimizing the TV distance, as the former is twice the latter. Let $\Delta_n := \{ (w_1, \ldots, w_n) \in \mathbb{R}^n : w_i \geq 0, \sum w_i = 1 \}$ denote the n-dimensional simplex.

Definition 1.4 (k-mix($\mathcal{F}$)). Let $\mathcal{F}$ be a class of probability distributions. Then the class of k-mixtures of $\mathcal{F}$, written k-mix($\mathcal{F}$), is defined as

$$k\text{-mix}(\mathcal{F}) := \{ \textstyle\sum_{i=1}^k w_i f_i : (w_1, \ldots, w_k) \in \Delta_k, f_1, \ldots, f_k \in \mathcal{F} \}.$$

1.2 Sample Compression Schemes

In this article, we introduce a method for learning distributions via a novel form of *compression*. Given a class $\mathcal{F}$ of distributions, suppose there is a method for "compressing" information about any distribution $f \in \mathcal{F}$ using a subset of samples from f and some additional bits. Further, suppose there exists a fixed, deterministic *decoder* for $\mathcal{F}$, which, given the subset of samples and the additional bits, approximately recovers f. If the size of the subset and the number of bits is guaranteed to be small, we show that the sample complexity of learning $\mathcal{F}$ is small as well.

More precisely, we say class $\mathcal{F}$ *admits (τ, t, m) compression* if there exists a *decoder function* such that, upon generating m i.i.d. samples from any $f \in \mathcal{F}$, we are guaranteed, with probability at least 2/3, to have τ data points from the sample and a sequence of at most t bits on which the decoder outputs a distribution that is within total variation distance of ε from f. Note that τ, t, and m may be functions of ε, the accuracy parameter. The decoder function is specific to the class $\mathcal{F}$ but does not depend on the particular f.

This definition is further generalized to a stronger notion of *robust compression*, formally defined in Definition 4.2, where the target distribution is to be encoded using samples that are not necessarily generated from the target itself but are generated from a distribution close to the target. More precisely, class $\mathcal{F}$ *admits (τ, t, m) robust compression* if there exists a *decoder function*, such that for any $f \in \mathcal{F}$, upon generating m i.i.d. samples from any distribution that is "close" to f, we are guaranteed, with probability at least 2/3, to have τ data points from the sample and a sequence of at most t bits on which the decoder outputs a distribution that is within total variation distance of ε from f. We prove that robust compression implies agnostic learning. In particular, if $\mathcal{F}$ admits (τ, t, m) robust compression, then the sample complexity of agnostic learning with respect to $\mathcal{F}$ is bounded by $\widetilde{O}(m + (\tau + t)/\varepsilon^2)$ (Theorem 4.5). ($\widetilde{O}$ allows for polylogarithmic factors.)

To illustrate the compression technique, it is instructive to compare it with another density estimation technique, called the *cover method* in Reference [20, Section 1.5]. Let $(\mathcal{D}, d)$ be a metric

space and let $\mathcal{F} \subseteq \mathcal{D}$. Given $\varepsilon > 0$, a subset $C \subseteq \mathcal{D}$ is called an ε-*net* for $\mathcal{F}$ if for any $f \in \mathcal{F}$ there exists some $c \in C$ such that $d(f, c) \leq \varepsilon$. The size of the smallest ε-net is called the *covering number* of $\mathcal{F}$ and its logarithm is called the *metric entropy* of $\mathcal{F}$. (In this informal discussion, we are ignoring the dependence on ε for brevity.) The metric entropy is a measure of "dimension" for the set $\mathcal{F}$.

The cover method for learning a class $\mathcal{F}$ of distributions over domain Z works as follows: View $\mathcal{F}$ as a subset of the class of all distributions over Z equipped with the L^1 metric. Consider a small ε-net of $\mathcal{F}$ and then choose the distribution in the net that is closest to the target distribution in the L^1 distance. If the net is finite, then this last step is equivalent to learning the closest distribution among finitely many candidates to a target distribution, which has an algorithm whose sample complexity scales logarithmically with the net size (see Theorem 4.4).

The issue with using the cover method for learning the class of Gaussians is that the covering number and the metric entropy of the class are infinite with respect to the L^1 metric. In fact, even for the class of mean-zero Gaussians with entry-wise bounded covariance matrices, the metric entropy is infinite unless we assume a bound on the condition number of the covariance matrices (consider two-dimensional singular Gaussians each supported on a different one-dimensional subspace with unit variance on that subspace).

The power of compression schemes is that they take a data-dependent approach: A first round of sampling is used to shrink the space of feasible distributions significantly, making the metric entropy of the resulting feasible set finite. This allows us to achieve bounds for Gaussians that do not depend on the condition number or the size of the parameters (Lemma 5.3).

An additional attractive property of compression is that it enjoys two closure properties. Specifically, if a base class admits compression, then the class of products of the base class, as well as the class of mixtures of that base class, are compressible (Lemmata 4.6 and 4.8).

As an application of this technique, we prove tight (up to logarithmic factors) sample complexity upper bounds of $\widetilde{O}(kd^2/\varepsilon^2)$ for learning mixtures of k Gaussians over $\mathbb{R}^d$ (Theorem 1.5) and $\widetilde{O}(kd/\varepsilon^2)$ for learning mixtures of k axis-aligned Gaussians (Theorem 1.7).

In light of the closure properties of compression schemes, we need only provide a compression scheme for the class of Gaussian distributions to obtain a compression scheme (and a sample complexity bound) for mixtures of Gaussians. We prove that the class of d-dimensional Gaussian distributions admits $(\widetilde{O}(d), \widetilde{O}(d^2), \widetilde{O}(d))$ robust compression (Lemma 5.3). The high-level idea is that by generating $\widetilde{O}(d)$ samples from a Gaussian, one can get a rough sketch of the geometry of the Gaussian. In particular, the points drawn from a Gaussian concentrate around an ellipsoid centered at the mean and whose principal axes are the eigenvectors of the covariance matrix. Using ideas from convex geometry and random matrix theory, we show one can encode the center of the ellipsoid *and* the principal axes using linear combinations of these samples. Then, we discretize the coefficients and obtain an approximate encoding.

Our compression framework is quite flexible and can be used to prove sample complexity upper bounds for other distribution classes as well. This is left for future work.

1.3 Main Results

Our first main result is an upper bound for learning mixtures of multivariate Gaussians. This bound is tight up to logarithmic factors. Let k denote the number of mixture components and d denote the dimension. Henceforth, the notations $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ suppress $\mathrm{polylog}(kd/\varepsilon\delta)$ factors; there are no hidden dependencies on any other parameters (such as the condition number) when we use the O, Ω, $\widetilde{O}$, and $\widetilde{\Omega}$ notation.

THEOREM 1.5. *The class of k-mixtures of d-dimensional Gaussians can be learned in the realizable setting, and can be 12-learned in the agnostic setting, using $\widetilde{O}(kd^2/\varepsilon^2)$ samples.*

We emphasize that the $\widetilde{O}(\cdot)$ notation has *no dependence whatsoever on the scaling, condition number, separation*, or any other structural property of the distribution. Previously, the best known upper bounds on the sample complexity of this problem were $\widetilde{O}(kd^2/\varepsilon^4)$, due to Reference [6], and $O(k^4d^4/\varepsilon^2)$, based on a VC dimension bound that we discuss below. For the case of a single Gaussian (i.e., $k = 1$), a bound of $O(d^2/\varepsilon^2)$ is known, again using a VC dimension bound discussed below.

Our second main result is a lower bound matching Theorem 1.5 up to logarithmic factors.

THEOREM 1.6. *Any method for learning the class of k-mixtures of d-dimensional Gaussians in the realizable setting has sample complexity $\widetilde{\Omega}(kd^2/\varepsilon^2)$.*

Note that this is a worst-case (i.e., minimax) lower bound: For any estimation method, *there exists at least one distribution* that requires that many samples. Previously, the best known lower bound on the sample complexity was $\widetilde{\Omega}(kd/\varepsilon^2)$ [44]. Even for a single Gaussian (i.e., $k = 1$), an $\widetilde{\Omega}(d^2/\varepsilon^2)$ lower bound was not known prior to this work.

Our third main result is an upper bound for learning mixtures of *axis-aligned* Gaussians, i.e., Gaussians with diagonal covariance matrices. This bound is also tight up to logarithmic factors.

THEOREM 1.7. *The class of k-mixtures of axis-aligned d-dimensional Gaussians can be learned in the realizable setting, and can be 12-learned in the agnostic setting, using $\widetilde{O}(kd/\varepsilon^2)$ samples.*

A matching lower bound of $\widetilde{\Omega}(kd/\varepsilon^2)$ was proved in Reference [44]. Previously, the best known upper bounds were $\widetilde{O}(kd/\varepsilon^4)$, due to Reference [6], and $O((k^4d^2 + k^3d^3)/\varepsilon^2)$, based on a VC dimension bound that we discuss below.

In the agnostic results of Theorem 1.5 and Theorem 1.7, the constant 12 can be decreased to any constant larger than 9. One may verify this statement through a detailed inspection of our proofs. We omit a full derivation of this improved constant to avoid tedious details in the proofs.

Our techniques. The upper bounds are proved using the compression technique discussed in Section 1.2. Next, we discuss the main ideas used in the proof of our lower bound, Theorem 1.6. To prove our lower bound for mixtures of Gaussians, we first prove a lower bound of $\widetilde{\Omega}(d^2/\varepsilon^2)$ for learning a single Gaussian. The main step is to construct a large family, of size $2^{\Omega(d^2)}$, of covariance matrices such that the associated Gaussian distributions are well-separated in terms of their total variation distance, while simultaneously ensuring that their mutual Kullback-Leibler divergences are small. Once this is established, we can then apply a generalized form of Fano's inequality to complete the proof.

To construct this family of covariance matrices, we sample $2^{\Omega(d^2)}$ matrices from the following probabilistic process: Start with an identity covariance matrix; then choose a uniformly random subspace of dimension $d/9$ and slightly increase the eigenvalues corresponding to this eigenspace. It is easy to bound the KL divergences between the constructed Gaussians. Quantifying this gap will then give the desired lower bound on the total variation distance.

Minimax estimation rates. Our results are stated in terms of sample complexity, which is the terminology mostly used in the machine learning literature. It is also possible to state our results in terms of minimax estimation rates, which are often used in the statistics literature. There is a direct connection between sample complexity bounds and estimation rates, although translating between them sometimes incurs logarithmic factors.

We recall some definitions from the minimax estimation framework (see, e.g., Reference [45, Chapter 2]). Let $\mathcal{F}$ be a class of probability distributions defined on domain $\mathcal{X}$. The *risk* of a density

estimation method $\widehat{f} : \mathcal{X}^n \to \mathcal{F}$ for this class is defined as

$$\mathcal{R}_n(\widehat{f}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathbf{E} \, \mathrm{TV}(\widehat{f}(X_1, \ldots, X_n), f),$$

where the expectation is over the i.i.d. samples $X_1, \ldots, X_n$ from f, and possible randomization of the estimator. The *minimax estimation rate* for $\mathcal{F}$ is the smallest risk over all possible estimators $\widehat{f} : \mathcal{X}^n \to \mathcal{F}$, i.e.,

$$\mathcal{R}_n(\mathcal{F}) := \inf_{\widehat{f}} \mathcal{R}_n(\widehat{f}, \mathcal{F}) = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbf{E} \, \mathrm{TV}(\widehat{f}(X_1, \ldots, X_n), f).$$

Let $\mathcal{G}_{d,k}$ denote the class of k-mixtures of d-dimensional Gaussian distributions, and let $\mathcal{A}_{d,k}$ denote the class of k-mixtures of d-dimensional axis-aligned Gaussian distributions. Then Theorem 1.5 implies the minimax estimation rate of $\mathcal{G}_{d,k}$ is $\widetilde{O}(\sqrt{kd^2/n})$; indeed, the theorem states that to get error $\leq \varepsilon$, we need some $n \leq \mathrm{polylog}(kd/\varepsilon)kd^2/\varepsilon^2$ many samples, hence, solving for ε, we find that given n samples, we obtain an estimator with error $\varepsilon \leq \mathrm{polylog}(kdn)\sqrt{kd^2/n}$. Similarly, Theorem 1.7 implies the minimax estimation rate of $\mathcal{A}_{d,k}$ is $\widetilde{O}(\sqrt{kd/n})$. Note that these theorems indeed imply stronger statements than these risk bounds, since they give guarantees for the case where the target distribution does not necessarily belong to the known class $\mathcal{F}$, a setting not captured by the minimax framework.

Finally, the proof of Theorem 1.6 (see Theorem 6.8 below) implies that the minimax rate of $\mathcal{G}_{d,k}$ is $\Omega(\sqrt{kd^2/n})$, improving the $\Omega(\sqrt{kd^2}/\sqrt{n} \log n)$ lower bound proved in the preliminary version of this article [4].

Computational efficiency. Although our approach for proving sample complexity upper bounds is algorithmic, our focus is not on computational efficiency. The resulting algorithms have nearly optimal sample complexities, but their running times are exponential in the dimension d and the number of mixture components k. More precisely, the running time is $2^{kd^2 \, \mathrm{polylog}(d,k,1/\varepsilon,1/\delta)}$ for mixtures of general Gaussians, and $2^{kd \, \mathrm{polylog}(d,k,1/\varepsilon,1/\delta)}$ for mixtures of axis-aligned Gaussians. The existence of an algorithm for density estimation that runs in time $\mathrm{poly}(k, d)$ is unknown even for the class of mixtures of axis-aligned Gaussians (see Reference [23, Question 1.1]).

Even for the case of a single Gaussian, the published proofs of the $O(d^2/\varepsilon^2)$ bound, of which we are aware, are not algorithmically efficient, e.g., Reference [6, Theorem 13]. Adopting ideas from our proof of Theorem 1.5, an algorithmically efficient learner for a single Gaussian can be obtained simply by computing the empirical mean and (an appropriate estimate of the) covariance matrix using $O(d^2/\varepsilon^2)$ samples. The details appear in our technical report [5].

Article outline. Next, we review some related work. Then, we introduce our notation and recall some standard facts in Section 2. In Section 3, we provide justification for our learning model. In Section 4, we formally define compression schemes for distributions, prove their closure properties, and show their connection with density estimation. Theorems 1.5 and 1.7 are proved in Section 5. Theorem 1.6 is proven in Section 6. All omitted proofs appear in the appendices.

1.4 Related Work

Distribution learning is a vast topic, and many approaches have been considered in the literature. This section reviews the approaches that are particularly relevant to our work.

For parametric families of distributions, a common approach is to estimate the parameters of the distribution, in a maximum likelihood sense, or aiming to approximate the true parameters. For mixtures of Gaussians, there is a rich theoretical literature on algorithms that approximate

the mixing weights, means and covariances (e.g., References [3, 8, 13, 38]); see Reference [27] for a survey. The strictness of this objective cuts both ways. On the one hand, a successful learner uncovers substantial structure of the target distribution. On the other hand, this objective is impossible when the means and covariances are extremely close. Thus, algorithms for parameter estimation of Gaussian mixtures necessarily require some separation assumptions on the target parameters.

Density estimation has a long history in statistics, where the focus is on the sample complexity question; see References [16, 17, 42] for general background. Density estimation was first studied in the computational learning theory community under the name *PAC learning of distributions* in Reference [28], whose focus is on the computational complexity of the learning problem.

Various measures of dissimilarity between distributions have been considered in existing density estimation schemes. One natural measure is the TV distance, which has been adopted by several papers on learning mixtures of Gaussians [6, 11, 14]. Another natural measure, which has also been considered for mixtures of Gaussians, is the Kullback-Leibler (KL) divergence [24]. Some prior work has also used the L^2 distance for density estimation [1, 21]. This article focuses on the TV distance (i.e., the L^1 distance), and we provide justification for this choice in Section 3.

A popular method for distribution learning in practice is kernel density estimation (see, e.g., Reference [17, Chapter 9]). The rigorously proven sample complexity/estimation rate upper bounds for this method require either smoothness assumptions (e.g., Reference [17, Theorem 9.5]) or boundedness assumptions (e.g., Reference [26, Theorem 2.2]) on the class of densities. The class of Gaussians is not universally Lipschitz or universally bounded, so those results do not apply to the problems we consider. Moreover, numerical calculations demonstrate that the number of samples required to estimate a standard Gaussian (within L^2 distance 0.1 using Gaussian kernels) grow exponentially with the dimension (see Reference [42, Table 4.2 on page 94]), which hints that this method suffers from the curse of dimensionality.

Another approach for deriving sample complexity upper bounds for distribution learning, called the *minimum distance estimate* in Reference [17, Chapter 6], is based on the uniform convergence theory and the notion of *Vapnik-Chervonenkis dimension* (Definition 2.5). It is proved in Reference [17] that an upper bound for a class of distributions can be obtained by bounding the VC dimension of an associated set system, called the *Yatracos family* (Definition 2.6). For example, Reference [22] used this method to bound the sample complexity of learning high-dimensional log-concave distributions. For learning d-dimensional Gaussians, this approach leads to the optimal sample complexity upper bound of $O(d^2/\varepsilon^2)$. However, for mixtures of Gaussians and axis-aligned Gaussians, the best known VC dimension bounds (see Reference [2, Theorem 8.14] and Reference [17, Section 8.5]) result in loose upper bounds of $O(k^4d^4/\varepsilon^2)$ and $O((k^4d^2 + k^3d^3)/\varepsilon^2)$, respectively.

Another approach is to first approximate the Gaussian mixture class using a more manageable class such as piecewise polynomials and then study the associated Yatracos family; see, e.g., Reference [11]. However, piecewise polynomials do a poor job in approximating d-dimensional Gaussians, resulting in an exponential dependence on d.

For density estimation of mixtures of Gaussians using the TV distance, the best known sample complexity upper bounds (in terms of k and d) are $\widetilde{O}(kd^2/\varepsilon^4)$ for general Gaussians and $\widetilde{O}(kd/\varepsilon^4)$ for axis-aligned Gaussians, both due to Reference [6]. For the general Gaussian case, their method takes an i.i.d. sample of size $\widetilde{O}(kd^2/\varepsilon^2)$ and partitions this sample in every possible way into k subsets. Based on those partitions, $k^{\widetilde{O}(kd^2/\varepsilon^2)}$ "candidate distributions" are generated. The problem is then reduced to learning with respect to this finite class of candidates. Their sample complexity has a suboptimal factor of $1/\varepsilon^4$, of which $1/\varepsilon^2$ arises in their approach for choosing the best candidate, and another factor $1/\varepsilon^2$ is due to the exponent in the number of candidates.

Table 1. Bounds on the Sample Complexities of Learning Gaussian Mixtures and Their Subclasses

	Number of Gaussians	Dimension	Axis-aligned	Sample complexity	Reference
Upper Bounds	1	d	no	$O(d^2/\varepsilon^2)$	standard
	1	d	yes	$O(d/\varepsilon^2)$	standard
	k	1	n/a	$\widetilde{O}(k/\varepsilon^2)$	[11]
	k	d	no	$\widetilde{O}(kd^2/\varepsilon^2)$	this article
	k	d	yes	$\widetilde{O}(kd/\varepsilon^2)$	this article
Lower Bounds	1	d	no	$\widetilde{\Omega}(d^2/\varepsilon^2)$	this article
	1	d	yes	$\widetilde{\Omega}(d/\varepsilon^2)$	[44]
	k	1	n/a	$\widetilde{\Omega}(k/\varepsilon^2)$	[44]
	k	d	no	$\widetilde{\Omega}(kd^2/\varepsilon^2)$	this article
	k	d	yes	$\widetilde{\Omega}(kd/\varepsilon^2)$	[44]

The lower bounds are minimax (i.e., worst-case). The bounds in the first two rows are well known; proofs can be found in Reference [6].

Our approach via compression schemes also ultimately reduces the problem to learning with respect to finite classes, although yielding a more refined bound than Reference [6]. In our sample complexity upper bounds, one factor of $1/\varepsilon^2$ is again incurred due to learning with respect to finite classes. The key is that the number of compressed samples does not depend on ε, so the overall sample complexity bound has only an $\widetilde{O}(1/\varepsilon^2)$ dependence on ε.

As for lower bounds on the sample complexity for learning mixtures of Gaussians under the TV distance, much fewer results are known. The only lower bound prior to this work is due to Reference [44], which shows a bound of $\widetilde{\Omega}(kd/\varepsilon^2)$ for learning mixtures of axis-aligned Gaussians (and hence for general Gaussians as well). This bound is tight for the axis-aligned case, as we show in Theorem 1.7, but loose in the general case, as we show in Theorem 1.6. After the preliminary version of this article was completed [4], an alternative construction was provided in Reference [19] giving the same lower bound as ours using a deterministic construction.

A summary of known bounds on sample complexities for learning Gaussian mixtures and their subclasses is presented in Table 1.

2 PRELIMINARIES

Basic notation. The notation $\log(\cdot)$ denotes logarithm in the natural base, $[M]$ denotes $\{1, 2, \ldots, M\}$, and A^c denotes the complement of a set A. Throughout the article, a/bc always means $a/(bc)$. For any $p \geq 1$, the L^p distance between functions f and g over $\mathbb{R}^d$ is defined as $\|f - g\|_p := (\int_{\mathbb{R}^d} |f(x) - g(x)|^p \, dx)^{1/p}$, and the ℓ_p distance between two vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ is defined as $(\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$, and their ℓ_∞ distance is defined as $\max_i |x_i - y_i|$.

Probability terminology. For random variables X and Y, the notation $X \overset{d}{=} Y$ means that X and Y have the same distribution. For a distribution g, we write $X \sim g$ to mean X is a random variable with distribution g, and $S \sim g^m$ means that S is an i.i.d. sample of size m generated from g.

PROPOSITION 2.1. *Let X and Y be random variables taking values in the same set. For any function f, we have* $\mathrm{TV}(f(X), f(Y)) \leq \mathrm{TV}(X, Y)$.

PROOF. For any set A, we have

$$\mathbf{Pr}\left[f(X) \in A\right] - \mathbf{Pr}\left[f(Y) \in A\right] = \mathbf{Pr}[X \in f^{-1}(A)] - \mathbf{Pr}[Y \in f^{-1}(A)] \leq \mathrm{TV}(X, Y).$$

Taking the supremum of the left-hand side proves the proposition. □

Matrix terminology. We use $\|v\|_2$ to denote the Euclidean norm of a vector v, $\|A\|_s$ to denote the *spectral norm*, or the *operator norm*, of a matrix A, that is, $\|A\|_s := \sup_{\|v\|_2=1} \|Av\|_2$, and $\|A\|_F := \sqrt{\mathrm{Tr}(A^\mathsf{T} A)}$ to denote the *Frobenius norm* of a matrix A.

Gaussian and χ_d distributions. Let d denote the dimension. A Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is denoted by $\mathcal{N}(\mu, \Sigma)$. If Σ is a diagonal matrix, then $\mathcal{N}(\mu, \Sigma)$ is called an *axis-aligned* Gaussian. If Σ is full-rank, then $\mathcal{N}(\mu, \Sigma)$ is called a *full-rank* Gaussian. The $\mathcal{N}(0, I)$ distribution is called the standard Gaussian distribution, and an $\mathcal{N}(0, 1)$ random variable is called standard normal. It is easy to check that if $g \sim \mathcal{N}(0, I)$ then $\mu + \Sigma^{1/2} g \sim \mathcal{N}(\mu, \Sigma)$. Let $g_1, \ldots, g_n$ be i.i.d. standard normal; then $\sum_{i=1}^{n} g_i^2$ is said to have the *chi-squared distribution* with parameter n and is denoted by χ_n. Observe that $\mathbf{E}\chi_n = n$.

LEMMA 2.2 ([30, LEMMA 1]). *For any positive integer d, $\Pr[\chi_d - d \geq 2\sqrt{dt} + 2t] \leq \exp(-t)$.*

COROLLARY 2.3. *For any positive integer d, $\Pr[\chi_d \geq 16d] \leq \exp(-3)$.*

THEOREM 2.4 (CHERNOFF BOUND, SEE THEOREM 4.5 IN REFERENCE [37]). *Let $X = \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are independent $\{0, 1\}$ random variables. For any $0 < \delta < 1$, we have $\Pr[X \leq (1 - \delta)\mathbf{E}X] \leq \exp(-\delta^2 \mathbf{E}X/2)$.*

Definition 2.5 (Vapnik-Chervonenkis (VC) dimension [46]). Let $\mathcal{A} \subseteq 2^X$ be a family of subsets of a set X. The *VC dimension* of $\mathcal{A}$, denoted by VC-dim$(\mathcal{A})$, is the size of the largest set $X \subseteq \mathcal{X}$ such that for each $Y \subseteq X$ there exists $B \in \mathcal{A}$ with $X \cap B = Y$.

For examples and applications of the VC dimension, see, e.g., Reference [17, Chapter 4].

Definition 2.6 (Yatracos family [48]). For a class of densities $\mathcal{F}$ over $\mathbb{R}^d$, the associated *Yatracos family* is the following family of subsets of $\mathbb{R}^d$:

$$\{\{x \in \mathbb{R}^d : f(x) > g(x)\} : f, g \in \mathcal{F}\}.$$

2.1 KL Divergence, Log-det Divergence, and Total Variation Distance

Definition 2.7 (Kullback-Leibler (KL) divergence [29]). The *Kullback-Leibler (KL) divergence* between densities f_1 and f_2 is defined by

$$\mathrm{KL}(f_1 \parallel f_2) := \int_{\mathbb{R}^d} f_1(x) \log \frac{f_1(x)}{f_2(x)} dx,$$

where we define $\mathrm{KL}(f_1 \parallel f_2) = +\infty$ if the set $\{x : f_2(x) = 0 < f_1(x)\}$ has positive Lebesgue measure.

The KL divergence is a measure of distance between distributions, which is asymmetric and does not satisfy the triangle inequality. However, it is always nonnegative (see, e.g., Reference [12, Theorem 2.6.3]) and can take value $+\infty$.

Definition 2.8 (Log-det Divergence). Let A and B be symmetric positive definite matrices of the same size. The *log-det divergence* of A and B is defined as LD $(A, B) := \mathrm{Tr}(B^{-1}A - I) - \log \det(B^{-1}A)$.

From the definition, it is apparent that LD (A, B) only depends on the spectrum of $B^{-1}A$. The log-det divergence is an asymmetric measure of distance between matrices and is closely related to the KL divergence between their corresponding Gaussian distributions, as can be seen from Lemma 2.10 below, which illustrates that the log-det divergence is always nonnegative.

LEMMA 2.9. *Let A and B be positive definite $d \times d$ matrices satisfying $\|B^{-1/2}AB^{-1/2} - I\|_s \leq \alpha$ for some $\alpha \in [0, 1/2]$. Then LD $(A, B) \leq d\alpha^2$.*

PROOF. Since $B^{-1/2}AB^{-1/2}$ is symmetric, its eigenvalues $\lambda_1, \ldots, \lambda_d$ are real, and, since we have $\|B^{-1/2}AB^{-1/2} - I\|_s \le \alpha$, each $\lambda_i \in [1-\alpha, 1+\alpha]$. So,

$$\text{LD}(A, B) = \text{Tr}(B^{-1}A - I) - \log\det(B^{-1}A) = \sum_{i=1}^d (\lambda_i - 1) - \log\prod_{i=1}^d \lambda_i$$

$$= \sum_{i=1}^d (\lambda_i - 1 - \log(\lambda_i)) \le \sum_{i=1}^d (\lambda_i - 1)^2 \le d\alpha^2.$$

The first inequality follows from $x - 1 - \log x \le (x-1)^2$, valid for any $x \ge 1/2$.   □

LEMMA 2.10. *For any two full-rank Gaussians $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$, we have*

$$2\,\text{TV}\left(\mathcal{N}(\mu_0, \Sigma_0), \mathcal{N}(\mu_1, \Sigma_1)\right)^2 \le \text{KL}\left(\mathcal{N}(\mu_0, \Sigma_0 \parallel \mathcal{N}(\mu_1, \Sigma_1)\right)$$

$$= \frac{1}{2}\left(\text{LD}(\Sigma_0, \Sigma_1) + (\mu_0 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_0 - \mu_1)\right).$$

PROOF. The inequality is Pinsker's inequality (see, e.g., Reference [45, Lemma 2.5]) applied to the distributions $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$, and the equality is a known formula for the KL divergence between Gaussians (see, e.g., Reference [40, Equation A.23]).   □

LEMMA 2.11. *For any $\mu, \sigma, \widehat{\mu}, \widehat{\sigma} \in \mathbb{R}$ with $|\widehat{\mu} - \mu| \le \varepsilon\sigma$ and $|\widehat{\sigma} - \sigma| \le \varepsilon\sigma$ and $\varepsilon \in [0, 2/3]$, we have*

$$\|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)\|_1 \le 2\varepsilon.$$

PROOF. By Lemma 2.10,

$$4\,\text{TV}(\mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2), \mathcal{N}(\mu, \sigma^2))^2 \le \frac{\widehat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\widehat{\sigma}^2}{\sigma^2}\right) + \frac{|\mu - \widehat{\mu}|^2}{\sigma^2} \le \left(\frac{\widehat{\sigma}}{\sigma}\right)^2 - 1 - \log\left(\left(\frac{\widehat{\sigma}}{\sigma}\right)^2\right) + \varepsilon^2.$$

Since $z := \widehat{\sigma}/\sigma \in [1 - \varepsilon, 1 + \varepsilon]$ and $\varepsilon \le 2/3$, using the inequality $x^2 - 1 - \log(x^2) \le 3(x-1)^2$ valid for all $x \in [1/3, 5/3]$, we find

$$\text{TV}(\mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2), \mathcal{N}(\mu, \sigma^2))^2 \le \frac{1}{4}(3(z-1)^2 + \varepsilon^2) \le \frac{1}{4}(4\varepsilon^2) = \varepsilon^2.$$

The lemma follows, since the L^1 distance is symmetric and is equal to twice the TV distance.   □

2.2 Bounds on Net Sizes

Definition 2.12 (ε-net). Let $\varepsilon \ge 0$. We say $N \subseteq X$ is an ε-net for X in metric d if for each $x \in X$ there exists some $y \in N$ such that $d(x, y) \le \varepsilon$.

The following lemma is Corollary 4.2.13 in Reference [47]; we include a proof for completeness. The notation B_2^d denotes the d-dimensional Euclidean ball, $B_2^d := \{ x \in \mathbb{R}^d : \|x\|_2 \le 1 \}$.

LEMMA 2.13. *For any $\varepsilon \in (0, 1]$, there exists an ε-net for B_2^d in ℓ_2 metric of size $(3/\varepsilon)^d$.*

PROOF. Take a maximal set of points in the unit ball B_2^d that are ε-separated, i.e., the distance between every pair of points is greater than ε. Such a set must be an ε-net by maximality. Moreover, by the triangle inequality, the balls of radius $\varepsilon/2$ centered at the points in the ε-net are disjoint and hence the sum of their volumes is not more than the volume of a ball of radius $1 + \varepsilon/2$. The volume of a d-dimensional ball of radius r is $c_d r^d$ for some constant c_d, thus the size of this ε-net is at most $\frac{c_d(1+\varepsilon/2)^d}{c_d(\varepsilon/2)^d} \le (3/\varepsilon)^d$, as required.   □

LEMMA 2.14. *For any $\varepsilon \in (0, 1]$ there exists an ε-net for Δ_d in ℓ_∞ metric of size ε^{-d}.*

PROOF. We give an algorithm to construct the net: Partition $[0, 1]^d$ into ε^{-d} cubes of side-length ε; for each cube that intersects Δ_d, put one arbitrary point of the intersection in the net. □

3 JUSTIFICATION FOR OUR MODEL

Some of the existing models for learning mixtures of Gaussians need structural assumptions on the target distribution. For example, learning under the parameter estimation model requires that the means are sufficiently separated and that the mixing weights are not too small; see the discussion after Reference [27, Definition 1].

A key motivation for our work is to study a model for learning mixtures of Gaussians that requires no structural assumptions at all. Specifically, we would like to identify a model in which Gaussian mixtures can be learned up to error ε with sample complexity depending only on k, d, and ε, and then derive optimal sample complexity bounds in that model. Density estimation under the TV distance is one such model: Reference [6, Theorem 14] and Theorem 1.5 in this article show that mixtures of Gaussians can be learned up to error ε with sample complexity depending on k, d, and ε only. In this section, we provide further justification for using this particular model. In Section 3.1, we argue that the TV distance is not an arbitrary choice. If, instead, we had used the KL divergence or any L^p distance, with $p > 1$, then the sample complexity must necessarily depend on the structural properties of the distribution. Thus, TV distance is a natural choice.

It is also natural to wonder whether some of our results could be derived from existing results on parameter estimation. In Section 3.2, we show that this is not the case: Entry-wise estimation of the covariance matrices is quite unrelated to density estimation under the TV distance. Thus, our model is natural and our results are not subsumed by previous work.

3.1 Comparison to KL Divergence and L^p Distances

In this section, we consider the problem of density estimation, for mixtures of Gaussians, using a distance measure that is either the KL divergence or an L^p distance with $p > 1$. Under these distance measures, we show that the sample complexity of this problem must necessarily depend on the structural properties of the distribution—that is, it cannot be bounded purely as a function of k, d and ε.

First, we consider using the KL divergence. Recall that KL divergence is not symmetric. We consider using KL divergence only in one direction and show that no algorithm can guarantee, after receiving a uniformly bounded number of samples from the true distribution, that the KL divergence between the true distribution and the output distribution is smaller than any finite number. In fact, this holds even for mixtures of two one-dimensional Gaussians with unit variances.

THEOREM 3.1. *Let $\mathcal{F}$ be the class of mixtures of two Gaussians in $\mathbb{R}$, both of which have unit variance. Let $\mathcal{A}$ be any algorithm whose input is a finite sequence of real numbers and whose output is a (Lebesgue) measurable density function. Then, for every $m \in \mathbb{N}$ and every $\tau > 0$, there exists a density $f \in \mathcal{F}$ such that if $X'_1, \ldots, X'_m \sim f$ then $\mathrm{KL}(f \parallel \mathcal{A}(X'_1, \ldots, X'_m)) \geq \tau$ with probability at least 0.98.*

We present the proof idea here, leaving the formal argument to Appendix A. Let $a \in \mathbb{N}$ and consider the set of distributions $(1 - \delta) \cdot \mathcal{N}(0, 1) + \delta \cdot \mathcal{N}(a, 1)$, where $\delta \ll 1/m$. Any algorithm that draws m samples from such a distribution will likely have all of its samples come from $\mathcal{N}(0, 1)$. However, the only way for the KL divergence to be small is if the distribution returned by $\mathcal{A}$ has non-negligible mass near the $\mathcal{N}(a, 1)$ distribution, which is impossible, since the samples provide no information about a.

Next, we consider L^p distances and prove a result analogous to Theorem 3.1. The main difference is that the argument uses Gaussians with different variances, which can strongly influence the L^p distance.

THEOREM 3.2. *Let $\mathcal{F}$ be the class of mixtures of two Gaussians in $\mathbb{R}$. Let $\mathcal{A}$ be any algorithm whose input is a finite sequence of real numbers and whose output is a (Lebesgue) measurable density function. Then, for every $p > 1$, every $m \in \mathbb{N}$, and every $\tau > 0$, there exists a density $f \in \mathcal{F}$ such that if $X'_1, \ldots, X'_m \sim f$, then $\|f - \mathcal{A}(X'_1, \ldots, X'_m)\|_p \geq \tau$ with probability at least 0.98.*

The proofs of Theorems 3.1 and 3.2 appear in Appendix A. Note that the theorems hold even for randomized algorithms.

3.2 Comparison to Parameter Estimation

In this section, we observe that neither our upper bound (Theorem 1.5) nor our lower bound (Theorem 1.6) can directly follow from results about parameter estimation for Gaussian mixtures. First, recall that our sample complexity upper bound in Theorem 1.5 has no dependence on the structural properties of the Gaussians in the mixture. Next, consider an algorithm that learns a single d-dimensional Gaussian and provides a proximity guarantee on the entries of the covariance matrix. If we use this entrywise guarantee to infer closeness in either KL divergence or TV distance, then we argue that the error must depend on the condition number of the covariance matrix.

The *condition number* of a matrix Σ, i.e., the ratio of its maximum and minimum eigenvalues, is denoted by $\kappa(\Sigma)$.

PROPOSITION 3.3. *Set $\varepsilon = \frac{2}{\kappa(\Sigma)+1}$. There exist two covariance matrices Σ and $\widehat{\Sigma}$ that are good entrywise approximations (both additively and multiplicatively):*

$$|\Sigma_{i,j} - \widehat{\Sigma}_{i,j}| \leq \varepsilon \qquad and \qquad \widehat{\Sigma}_{i,j} \in [1, 1 + 2\varepsilon] \cdot \Sigma_{i,j} \qquad \forall i,j,$$

but the corresponding Gaussian distributions are as far as they can get, i.e.,

$$\mathrm{KL}\left(\mathcal{N}(0, \Sigma) \parallel \mathcal{N}(0, \widehat{\Sigma})\right) = \infty \qquad and \qquad \mathrm{TV}\left(\mathcal{N}(0, \Sigma), \mathcal{N}(0, \widehat{\Sigma})\right) = 1.$$

PROOF. Define

$$\Sigma = \begin{bmatrix} 1 & -(1 - \varepsilon) \\ -(1 - \varepsilon) & 1 \end{bmatrix} \qquad and \qquad \widehat{\Sigma} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

where $\varepsilon \in (0, 1/2)$. The eigenvalues of Σ are $2 - \varepsilon$ and ε, so $\kappa(\Sigma) = \frac{2}{\varepsilon} - 1$, satisfying the stated condition for ε. Observe that Σ and $\widehat{\Sigma}$ satisfy the entrywise approximation guarantees in the statement of the theorem. However, Σ is non-singular and $\widehat{\Sigma}$ is singular. Thus, $\mathrm{TV}(\mathcal{N}(0, \Sigma), \mathcal{N}(0, \widehat{\Sigma})) = 1$ (consider the event that a random variable lies in the range of $\widehat{\Sigma}$), and $\mathrm{KL}(\mathcal{N}(0, \Sigma) \parallel \mathcal{N}(0, \widehat{\Sigma})) = \infty$ (recall the definition of KL divergence). □

Thus, given a black-box algorithm that provides an entrywise ε-approximation to the true covariance matrix Σ, if $\kappa(\Sigma) \geq 2/\varepsilon$, it might output $\widehat{\Sigma}$, which does not approximate Σ in KL divergence or total variation distance. Thus, Theorem 1.5 is not a direct consequence of any parameter estimation algorithm with entrywise covariance guarantees.

Remark 3.4. The matrix $\widehat{\Sigma}$ in the above construction is singular, so $\mathcal{N}(0, \widehat{\Sigma})$ is a singular Gaussian. The point of this proposition is that an algorithm that provides a good entrywise approximation for the covariance matrix does not necessarily provide a good statistical approximation, as it may output a singular Gaussian for example. Note that our algorithm handles singular Gaussians as is.

One might wonder instead if our lower bound is a direct consequence of existing lower bounds on parameter estimation. We show that this is also not the case: The next proposition shows that there exist Gaussians that are close in TV distance but whose covariance matrices do not satisfy any (multiplicative) entrywise guarantee. Thus, even if a lower bound concludes that a class of algorithms cannot provide entrywise covariance guarantees, it is still possible that an algorithm in that class can provide guarantees on the TV distance.

PROPOSITION 3.5. *For any $\varepsilon \in (0, 1/2)$, there exist two covariance matrices Σ and $\widehat{\Sigma}$ such that $\mathrm{TV}(\mathcal{N}(0, \widehat{\Sigma}), \mathcal{N}(0, \Sigma)) \leq \varepsilon$, but there exist i, j such that, for any $c \geq 1$, $\widehat{\Sigma}_{i,j} \notin [1/c, c] \cdot \Sigma_{i,j}$.*

PROOF. Define

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \widehat{\Sigma} = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix},$$

where $\varepsilon \in [0, 1/2]$. By Lemma 2.10,

$$2\,\mathrm{TV}\left(\mathcal{N}(0, \Sigma_0), \mathcal{N}(0, \Sigma_1)\right)^2 \;\leq\; \frac{\mathrm{LD}\left(\Sigma_0, \Sigma_1\right)}{2} = \frac{-\log(1 - \varepsilon^2)}{2} \leq \varepsilon^2,$$

so $\mathrm{TV}(\mathcal{N}(0, \widehat{\Sigma}), \mathcal{N}(0, \Sigma)) \leq \varepsilon$. However, $\widehat{\Sigma}$ is not an entrywise multiplicative approximation of Σ. □

4 COMPRESSION SCHEMES

The main technique introduced in this article is using compression for density estimation. An overview of this technique was given in Section 1.2. In this section, we provide formal definitions of compression schemes and their usage.

4.1 Definition of a Compression Scheme

Let $\mathcal{F}$ be a class of distributions over a domain Z. Intuitively, a compression scheme for $\mathcal{F}$ involves two agents: an *encoder* and a *decoder*.

- The encoder knows a distribution $g \in \mathcal{F}$ and receives m samples from this distribution. She uses her knowledge of g to construct and send a small *message* to the decoder, which will suffice for him to construct a distribution that is close to g. From the m samples, the encoder selects a subset of size τ, which are somehow representative of g. This subset, together with t additional bits, constitutes the message sent to the decoder.
- The decoder receives the message (the τ data points and the t bits) and constructs a distribution that is close to g.

Of course, there is some probability that the samples are not representative of the distribution g, in which case the compression scheme will fail. Thus, we only require that the decoding succeed with constant probability.

We emphasize that including samples in the message is critical and they cannot be omitted. Using samples in addition to bits is the main conceptual novelty of our approach and allows learning distribution classes with infinite metric entropy.

The formal definition of a decoder follows:

Definition 4.1 (Decoder). A *decoder* for $\mathcal{F}$ is a deterministic function $\mathcal{J} : \bigcup_{n=0}^{\infty} Z^n \times \bigcup_{n=0}^{\infty} \{0, 1\}^n \to \mathcal{F}$, which takes a finite sequence of elements of Z and a finite sequence of bits, and outputs a member of $\mathcal{F}$.

The formal definition of a compression scheme follows:

Definition 4.2 (Robust Compression Schemes). Let $\tau, t, m : (0, 1) \to \mathbb{Z}_{\geq 0}$ be functions, and let $r \geq 0$. We say $\mathcal{F}$ admits (τ, t, m) r-robust compression if there exists a decoder $\mathcal{J}$ for $\mathcal{F}$ such that for any distribution $g \in \mathcal{F}$ and any distribution q on Z with $\|g - q\|_1 \leq r$, the following holds:

> For any $\varepsilon \in (0, 1)$, if a sample S is drawn from $q^{m(\varepsilon)}$, then, with probability at least $2/3$, there exists a sequence L of at most $\tau(\varepsilon)$ elements of S, and a sequence B of at most $t(\varepsilon)$ bits, such that $\|\mathcal{J}(L, B) - g\|_1 \leq \varepsilon$.

Note that S and L are sequences rather than sets; in particular, they can contain repetitions. Last, note that $m(\varepsilon)$ is a *lower bound* on the number of samples needed, whereas $\tau(\varepsilon), t(\varepsilon)$ are *upper bounds* on the size of compression and the number of bits.

To summarize, the definition asserts that with probability $2/3$, there is a (short) sequence L of elements from S and a (short) sequence B of additional bits, from which g can be approximately reconstructed. We emphasize that L and B depend on g. Intuitively, the samples in L together with the bits B nearly represent the distribution g. This is a notion similar to that of a sufficient statistic, but it is not equivalent, since here the representation is only approximate. We say that the distribution g is *encoded* by the message (L, B). This compression scheme is called "robust," since it requires g to be approximately reconstructed from a sample generated from q rather than g itself. A 0-robust compression scheme is called a (non-robust) compression scheme.

Remark 4.3. In the preceding definition, we required that L and B exist with probability only $2/3$. Naturally, one can boost this probability to $1 - \delta$ by generating a sample of size $m(\varepsilon)\lceil \log_3(1/\delta) \rceil$: We partition this sample into $\lceil \log_3(1/\delta) \rceil$ parts of size $m(\varepsilon)$. Each part has a desirable subsequence L with probability at least $2/3$, so the probability that none of the parts has a desirable subsequence L is not more than $(1/3)^{\lceil \log_3(1/\delta) \rceil} \leq \delta$.

4.2 Connection between Compression and Learning

We now show that if a class of distributions has a (robust) compression scheme, then it can be learned in the (agnostic) density estimation model.

The main idea is as follows: An encoder cannot be implemented in the density estimation model, because she requires knowledge of the target distribution g. However, since her interaction with the decoder only amounts to sending a short message, we can explore all possible behaviors of the encoder by a brute-force search over all possible messages that she could have sent. When any such message is provided as input to the decoder, he will output some distribution f. Moreover, for *at least one* such message, namely, the one that would have been produced by the encoder, the decoder will output an f that is guaranteed to be close to g. Thus, if we collect all possible distributions produced by the decoder on all possible input messages, then the only remaining task is to select the distribution from that collection that is closest to g. Fortunately, this task has a known solution: The following result states that a finite class of size M can be 3-learned in the agnostic setting using $O(\log(M/\delta)/\varepsilon^2)$ samples.

THEOREM 4.4. *There exists a deterministic algorithm that, given candidate distributions* $f_1, \ldots, f_M$, *a parameter* $\varepsilon > 0$, *and* $\lceil \log(3M^2/\delta)/2\varepsilon^2 \rceil$ *i.i.d. samples from an unknown distribution* g, *outputs an index* $j \in [M]$ *such that*

$$\|f_j - g\|_1 \leq 3 \min_{i \in [M]} \|f_i - g\|_1 + 4\varepsilon,$$

with probability at least $1 - \delta/3$.

This result is essentially proven in Reference [48, Theorem 1]. It immediately follows from Reference [17, Theorem 6.3] and Hoeffding's inequality [17, Theorem 2.1].

Our approach for relating compression schemes and density estimation, described informally above, is made formal by the following theorem. It uses Theorem 4.4 to select the best distribution that the decoder could output. Note that we assume the learner knows all the problem parameters, such as $k, d, \varepsilon, \delta, \tau, t, m$, and r, but is oblivious to the target distribution.

THEOREM 4.5 (COMPRESSION IMPLIES LEARNING). *Suppose $\mathcal{F}$ admits (τ, t, m) r-robust compression. Let $\tau'(\varepsilon) := \tau(\varepsilon) + t(\varepsilon)$. Then $\mathcal{F}$ can be $\max\{3, 2/r\}$-learned in the agnostic setting using*

$$O\left(m\left(\frac{\varepsilon}{6}\right) \log\left(\frac{1}{\delta}\right) + \frac{\tau'(\varepsilon/6) \log(m(\varepsilon/6) \log_3(1/\delta)) + \log(1/\delta)}{\varepsilon^2}\right) = \widetilde{O}\left(m\left(\frac{\varepsilon}{6}\right) + \frac{\tau'(\varepsilon/6) \log m(\varepsilon/6)}{\varepsilon^2}\right)$$

samples. If $\mathcal{F}$ admits (τ, t, m) non-robust compression, then $\mathcal{F}$ can be learned in the realizable setting using the same number of samples.

PROOF. We give the proof for the agnostic case; the proof for the realizable case is similar. Let q be the target distribution from which the samples are being generated. Let $\alpha := \inf_{f \in \mathcal{F}} \|f - q\|_1$ be the approximation error of q with respect to $\mathcal{F}$. The goal of the learner is to find a distribution $\widehat{h}$ such that $\|\widehat{h} - q\|_1 \leq \max\{3, 2/r\} \cdot \alpha + \varepsilon$.

First, consider the case $\alpha < r$. In this case, we develop a learner that finds a distribution $\widehat{h}$ such that $\|\widehat{h} - q\|_1 \leq 3\alpha + \varepsilon$. Let $g \in \mathcal{F}$ be a distribution such that

$$\|g - q\|_1 \leq \min\left\{\alpha + \frac{\varepsilon}{12}, r\right\}. \tag{2}$$

Such a g exists by the definition of α. By assumption, $\mathcal{F}$ admits (τ, t, m) r-robust compression. Let $\mathcal{J}$ denote the corresponding decoder. Given ε, the learner first asks for an i.i.d. sample $S \sim q^{m(\varepsilon/6) \cdot \log_3(2/\delta)}$. Recall the definition of robust compression and Remark 4.3, which allows us to amplify the success probability of the decoder. Then, with probability at least $1 - \delta/2$, there exist $L \in S^{\tau(\varepsilon/6)}$ and $B \in \{0, 1\}^{t(\varepsilon/6)}$ satisfying the following guarantee: letting $h^* := \mathcal{J}(L, B)$, we have

$$\|h^* - g\|_1 \leq \frac{\varepsilon}{6}. \tag{3}$$

The learner is of course unaware of L and B. However, given the sample S, it can try all of the possibilities for L and B and create a candidate set of distributions. More concretely, let

$$H = \{ \mathcal{J}(L', B') : L' \in S^{\tau(\varepsilon/6)}, B' \in \{0, 1\}^{t(\varepsilon/6)} \}.$$

Note that

$$|H| \leq \left(m(\varepsilon/6) \log_3(2/\delta)\right)^{\tau(\varepsilon/6)} 2^{t(\varepsilon/6)} \leq \left(m(\varepsilon/6) \log_3(2/\delta)\right)^{\tau'(\varepsilon/6)}.$$

Since H is finite, we will use the algorithm of Theorem 4.4 to find a good candidate $\widehat{h}$ from H. In particular, we set the accuracy parameter in Theorem 4.4 to be $\varepsilon/16$ and the confidence parameter to be $\delta/2$. In this case, Theorem 4.4 requires

$$\frac{\log(6|H|^2/\delta)}{2(\varepsilon/16)^2} = O\left(\frac{\tau'(\varepsilon/6) \log(m(\frac{\varepsilon}{6}) \log_3(\frac{1}{\delta})) + \log(\frac{1}{\delta})}{\varepsilon^2}\right) = \widetilde{O}\left(\frac{\tau'(\varepsilon/6) \log m(\varepsilon/6)}{\varepsilon^2}\right)$$

additional samples, and its output $\widehat{h}$ satisfies the following guarantee:

$$\|\widehat{h} - q\|_1 \leq 3\|h^* - q\|_1 + 4\frac{\varepsilon}{16} \qquad\qquad \text{(by Theorem 4.4)}$$

$$\leq 3(\|h^* - g\|_1 + \|g - q\|_1) + \frac{\varepsilon}{4} \qquad\qquad \text{(by the triangle inequality)}$$

$$\leq 3\left(\frac{\varepsilon}{6} + \left(\alpha + \frac{\varepsilon}{12}\right)\right) + \frac{\varepsilon}{4} \qquad\qquad \text{(by (2) and (3))}$$

$$= 3\alpha + \varepsilon.$$

Note that the above procedure uses $\widetilde{O}(m(\varepsilon/6) + \tau'(\varepsilon/6)\log m(\varepsilon/6)/\varepsilon^2)$ samples and its failure probability is at most δ: The probability of either H not containing a good h^* or the failure of Theorem 4.4, in choosing a good candidate among H, is bounded by $\delta/2 + \delta/2 = \delta$.

The other case, $\alpha \geq r$, is trivial: The learner outputs some distribution $\widehat{h}$. Since $\widehat{h}$ and q are density functions, we have $\|\widehat{h} - q\|_1 \leq 2 \leq \frac{2}{r} \cdot \alpha \leq \max\{3, 2/r\} \cdot \alpha + \varepsilon$. □

4.3 Combining Compression Schemes

In the rest of this section, we prove a few lemmata showing that compression schemes can be combined in useful ways. These results concern product distributions (which will be useful for axis-aligned Gaussians) and mixture distributions (which will be useful for mixtures of Gaussians).

First, Lemma 4.6 below states that if a class $\mathcal{F}$ of distributions can be robustly compressed, then the class of distributions that are formed by taking products of members of $\mathcal{F}$ can also be robustly compressed. If $p_1, \ldots, p_d$ are distributions over domains $Z_1, \ldots, Z_d$, then $\prod_{i=1}^d p_i$ denotes the standard product distribution over $\prod_{i=1}^d Z_i$. For a class $\mathcal{F}$ of distributions, define

$$\mathcal{F}^d := \left\{ \prod_{i=1}^d p_i \ : \ p_1, \ldots, p_d \in \mathcal{F} \right\}.$$

LEMMA 4.6 (COMPRESSING PRODUCT DISTRIBUTIONS). *For any τ, t, m, r, d,*

$$\text{if } \mathcal{F} \text{ admits } \left(\quad \tau(\varepsilon), \qquad t(\varepsilon), \qquad m(\varepsilon) \quad \right) \text{ r-robust compression,}$$

$$\text{then } \mathcal{F}^d \text{ admits } \left(d \cdot \tau(\varepsilon/d),\ d \cdot t(\varepsilon/d),\ \log_3(3d) \cdot m(\varepsilon/d) \right) \text{ r-robust compression.}$$

For the proof, we need the following standard proposition, which can be proved, e.g., using the coupling characterization of the total variation distance.

PROPOSITION 4.7 (LEMMA 3.3.7 IN REFERENCE [41]). *For $i \in [d]$, let p_i and q_i be probability distributions over the same domain Z. Then $\|\Pi_{i=1}^d p_i - \Pi_{i=1}^d q_i\|_1 \leq \sum_{i=1}^d \|p_i - q_i\|_1$.*

PROOF OF LEMMA 4.6. Let Z be the domain of $\mathcal{F}$ and $G = \Pi_{i=1}^d g_i$ be an arbitrary element of $\mathcal{F}^d$, with all $g_i \in \mathcal{F}$. Let Q be an arbitrary distribution over Z^d subject to $\|G - Q\|_1 \leq r$. Let $q_1, \ldots, q_d$ be the marginal distributions of Q on the d components. Observe that $\|q_j - g_j\|_1 \leq r$ for each $j \in [d]$, since Proposition 2.1 implies that projection onto a coordinate cannot increase the total variation distance.

The lemma's hypothesis is that $\mathcal{F}$ admits (τ, t, m) r-robust compression. Let $\mathcal{J}$ denote the corresponding decoder, let $m_0 := m(\varepsilon/d)\log_3(3d)$ and $S \sim Q^{m_0}$. To prove the lemma, we must encode an ε-approximation of G using $d \cdot \tau(\varepsilon/d)$ elements of S and $d \cdot t(\varepsilon/d)$ bits.

Since S contains m_0 samples, each of which is a d-dimensional vector, we may think of S as a $d \times m_0$ matrix over Z. Let S_i denote the ith row of this matrix. That is, for $i \in [d]$, let $S_i \in Z^{m_0}$ be

the vector of the ith components of all elements of S. By definition of q_i, we have $S_i \sim q_i^{m_0}$ for each i. As observed above, we have $\|q_i - g_i\|_1 \le r$.

Apply Remark 4.3 with parameters ε/d and $\delta = 1/3d$ for each $i \in [d]$. Then, for each i, the following statement holds with probability at least $1 - 1/3d$: There exists a sequence L_i of at most $\tau(\varepsilon/d)$ elements of S_i and a sequence B_i of at most $t(\varepsilon/d)$ bits such that $\|\mathcal{J}(L_i, B_i) - g_i\|_1 \le \varepsilon/d$. By the union bound, this statement holds simultaneously for all $i \in [d]$ with probability at least $2/3$. We encode these $L_1, \ldots, L_d, B_1, \ldots, B_d$ using $d \cdot \tau(\varepsilon/d)$ samples from S and $d \cdot t(\varepsilon/d)$ bits. Our decoder for $\mathcal{F}^d$ then extracts $L_1, \ldots, L_d, B_1, \ldots, B_d$ from these samples and bits, and then outputs $\prod_{i=1}^d \mathcal{J}(L_i, B_i) \in \mathcal{F}^d$. Finally, Proposition 4.7 gives

$$\left\| \Pi_{i=1}^d \mathcal{J}(L_i, B_i) - G \right\|_1 \le \sum_{i=1}^d \|\mathcal{J}(L_i, B_i) - g_i\|_1 \le d \cdot \varepsilon/d = \varepsilon,$$

completing the proof. □

Our next lemma states that if a class $\mathcal{F}$ of distributions can be compressed, then the class of distributions that are formed by taking mixtures of members of $\mathcal{F}$ can also be compressed.

LEMMA 4.8 (COMPRESSING MIXTURES, NON-ROBUSTLY). *For any τ, t, m, r, d, if class $\mathcal{F}$ admits $(\tau(\varepsilon), t(\varepsilon), m(\varepsilon))$ non-robust compression, then the class of its k-mixtures, i.e., k-mix($\mathcal{F}$), admits $(k \cdot \tau(\frac{\varepsilon}{3}), \ k \cdot t(\frac{\varepsilon}{3}) + k \log_2(\frac{3k}{\varepsilon}), \ \frac{48k \log(6k)}{\varepsilon} \cdot m(\frac{\varepsilon}{3}))$ non-robust compression.*

PROOF. Consider any $g \in k$-mix($\mathcal{F}$), so $g = \sum_{i \in [k]} w_i f_i$ for some distributions $f_1, \ldots, f_k \in \mathcal{F}$ and mixing weights $w_1, \ldots, w_k$. Define $m_0 := 48m(\varepsilon/3)k \log(6k)/\varepsilon$, and draw $S \sim g^{m_0}$. Then S has the same distribution as the process that performs m_0 independent trials as follows: Select a component i proportional to the weights $w_1, \ldots, w_k$, then draw a sample from f_i. In the latter process, we define S_i to be the sequence of samples that were generated using f_i. Our encoder for g will discretize the mixing weights and use the compression scheme for $\mathcal{F}$ to separately encode each f_i.

Encoding the mixing weights. We encode $w_1, \ldots, w_k$ using bits as follows: Consider an $(\varepsilon/3k)$-net in ℓ_∞ for Δ_k of size $(3k/\varepsilon)^k$ (see Lemma 2.14). Let $(\widehat{w}_1, \ldots, \widehat{w}_k)$ be an element in the net that has

$$\|(\widehat{w}_1, \ldots, \widehat{w}_k) - (w_1, \ldots, w_k)\|_\infty \ \le \ \varepsilon/3k. \tag{4}$$

Encoding the element $(\widehat{w}_1, \ldots, \widehat{w}_k)$ from the net requires only $k \log_2(3k/\varepsilon)$ bits.

Encoding $f_1, \ldots, f_k$. For any $i \in [k]$, we say that index i is *negligible* if $w_i \le \varepsilon/(6k)$. For any negligible index, we will approximate f_i by an arbitrary distribution $\widehat{f}_i$. For any non-negligible index, we will likely have enough samples from f_i to use the compression scheme for $\mathcal{F}$ to encode a distribution $\widehat{f}_i$ that approximates f_i.

Define $m_1 = m(\varepsilon/3) \log(6k)$. For each non-negligible index i, by the Chernoff bound (Theorem 2.4), with probability at least $1 - 1/6k$, we have $|S_i| \ge m_1$. By a union bound, this statement holds simultaneously for all non-negligible $i \in [k]$ with probability at least $5/6$.

Apply Remark 4.3 with parameters $\varepsilon/3$ and $\delta = 1/6k$ for each non-negligible index i. Then, for each such i, the following statement holds with probability at least $1 - 1/6k$: There exist $\tau(\varepsilon/3)$ samples from S_i and $t(\varepsilon/3)$ bits from which the decoder for $\mathcal{F}$ constructs a distribution $\widehat{f}_i$ with

$$\|f_i - \widehat{f}_i\|_1 \ \le \ \varepsilon/3. \tag{5}$$

By the union bound, this statement holds simultaneously for all non-negligible indices with probability at least $5/6$. The encoding consists of these samples and bits for each non-negligible i, whereas for negligible i, we use the same number of samples and bits, chosen arbitrarily.

By a union bound, the failure probability of the encoding is at most $2 \cdot (1 - 5/6) = 1/3$.

Complexity of the encoding. The discretized weights require $k \log_2(3k/\varepsilon)$ bits. For each index $i \in [k]$, we use $\tau(\varepsilon/3)$ samples and $t(\varepsilon/3)$ bits. Thus, the total number of bits is $k \cdot t(\varepsilon/3) + k \log_2(3k/\varepsilon)$, and the total number of samples is $k \cdot \tau(\varepsilon/3)$.

Decoding. The decoder for k-mix($\mathcal{F}$) is given the discretized weights $\widehat{w}_1, \ldots, \widehat{w}_k$. It is also given, for each index i, $\tau(\varepsilon/3)$ samples and $t(\varepsilon/3)$ bits, which it provides to the decoder for $\mathcal{F}$, yielding the distribution $\widehat{f}_i$. (Recall that, for a negligible index i, the distribution $\widehat{f}_i$ is arbitrary.) The decoder outputs the distribution $\sum_i \widehat{w}_i \widehat{f}_i$.

To complete the proof of the lemma, we will show that $\| \sum_i w_i f_i - \sum_i \widehat{w}_i \widehat{f}_i \|_1 \leq \varepsilon$ with probability at least $2/3$. Let $N \subseteq [k]$ denote the set of negligible components. Recall that the encoder succeeds with probability at least $2/3$, in which case the decoded distributions $\widehat{f}_i$ will satisfy Inequality (5) for each $i \notin N$. So, we have

$$
\left\| \sum_{i \in [k]} (\widehat{w}_i \widehat{f}_i - w_i f_i) \right\|_1 \leq \left\| \sum_{i \in [k]} w_i(\widehat{f}_i - f_i) \right\|_1 + \left\| \sum_{i \in [k]} (\widehat{w}_i - w_i)\widehat{f}_i \right\|_1
$$

$$
\leq \left\| \sum_{i \in N} w_i(\widehat{f}_i - f_i) \right\|_1 + \left\| \sum_{i \notin N} w_i(\widehat{f}_i - f_i) \right\|_1 + \sum_{i \in [k]} |\widehat{w}_i - w_i| \cdot \|\widehat{f}_i\|_1
$$

$$
\leq \sum_{i \in N} w_i \cdot 2 + \sum_{i \notin N} w_i \cdot \frac{\varepsilon}{3} + \sum_{i \in [k]} \frac{\varepsilon}{3k} \cdot 1 \qquad \text{(by (4) and (5))}
$$

$$
\leq k \cdot \frac{\varepsilon}{6k} \cdot 2 + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \qquad \text{(by definition of } N\text{)}.
$$

This completes the analysis of the compression scheme for k-mix($\mathcal{F}$).   □

The preceding lemma shows that non-robust compression of $\mathcal{F}$ implies non-robust compression of k-mix($\mathcal{F}$). We do not know whether an analogous statement holds for robust compression. That is, does robust compression of $\mathcal{F}$ imply robust compression of k-mix($\mathcal{F}$), for a general class $\mathcal{F}$? Nevertheless, in the next lemma, we show that if $\mathcal{F}$ can be robustly compressed, then k-mix($\mathcal{F}$) can be *learned in the agnostic setting.*

LEMMA 4.9 (LEARNING MIXTURES, ROBUSTLY). *Suppose $\mathcal{F}$ admits $(\tau(\varepsilon), t(\varepsilon), m(\varepsilon))$ r-robust compression, and let $\tau'(\varepsilon) := \tau(\varepsilon) + t(\varepsilon)$. Then k-mix($\mathcal{F}$) admits $3(1 + 2/r)$-agnostic learning with sample complexity*

$$
\widetilde{O}\left(\frac{km(\varepsilon/10)}{\varepsilon} + \frac{k\tau'(\varepsilon/10) \log m(\varepsilon/10)}{\varepsilon^2} \right).
$$

We first give a sketch of the proof. Let g be the target distribution and suppose there exists $\rho \geq 0$ and $f \in k$-mix($\mathcal{F}$) such that $\|g - f\|_1 \leq \rho$. Since $f \in k$-mix($\mathcal{F}$), we can write $f = \sum_{i \in [k]} w_i f_i$, where $f_i \in \mathcal{F}$, $w_i \geq 0$, and $\sum_{i \in [k]} w_i = 1$. A first attempt would be to try to write $g = \sum_{i \in [k]} w_i g_i$ such that each $\|g_i - f_i\|_1 \leq r$; if this were true, then given a sufficient number of samples from g, we would have sufficient samples from each g_i, and then we could use an r-robust compression scheme for $\mathcal{F}$ to encode, for each i, some $\widehat{f}_i$ close to f_i. Alas, it is not clear whether we can ensure that $\|g_i - f_i\|_1 \leq r$ for all i. However, Lemma 4.10 below asserts that we can write $g = \sum_{i \in [k]} w_i g_i$ in such a way that, for each i, either $\|g_i - f_i\|_1 \leq r$ or w_i is small (in fact, the *sum* of all such weights is small) and, hence, their contribution to the TV distance is small. Thus, we need only deal with the case where $\|g_i - f_i\|_1 \leq r$, a task for which r-robust compression is well-suited.

LEMMA 4.10. *Let $f = \sum_{i \in [k]} w_i f_i$ be a density with $(w_1, \ldots, w_k) \in \Delta_k$ and each $f_i \in \mathcal{F}$. Let g be a density such that $\|g - f\|_1 \leq \rho$. Then, we can write $g = \sum_{i \in [k]} w_i g_i$ such that each g_i is a density and for any $r > 0$,*

$$\sum_{i \,:\, \|g_i - f_i\|_1 > r} w_i \;<\; \rho/r.$$

The proof of this lemma is cumbersome and appears in Appendix B. We now prove Lemma 4.9.

PROOF OF LEMMA 4.9. Let g be the target distribution, and let $f \in k\text{-mix}(\mathcal{F})$ be such that $\|f - g\|_1 \leq \rho$. To prove the lemma, we need to describe a learning algorithm that outputs a distribution whose L^1 distance to g is bounded by $3\rho(1 + 2/r) + \varepsilon$.

Let $g = \sum_{i \in [k]} w_i g_i$ be the representation given by Lemma 4.10. The learner first takes $M = 160m(\varepsilon/10)\log_3(3k/\delta)k/\varepsilon$ samples from g. Let S be the set of these samples. We view g as a mixture of the g_i, so S can be partitioned into k subsets such that the ith subset has distribution g_i. We learn each of the components individually. The learner does not know which sample point comes from which component, but it can try all possible ways of partitioning S into k subsets, hence generating several candidate distributions, such that at least one of them is close to g. Moreover, the learner "guesses" the weights w_i: Let W be an $(\varepsilon/10k)$-net in ℓ_∞ for Δ_k of size $(10k/\varepsilon)^k$ (see Lemma 2.14). So, there exists some point $(\widehat{w}_1, \ldots, \widehat{w}_k) \in W$ such that

$$\max_i |w_i - \widehat{w}_i| \leq \varepsilon/10k. \tag{6}$$

For each $i \in [k]$, component i is called *tiny* if $w_i < \varepsilon/20k$, *far* if $\|g_i - f_i\|_1 > r$, and *nice* otherwise. The sum of weights of tiny components is at most $\varepsilon/20$, and the sum of weights of far components is at most ρ/r by Lemma 4.10.

The number of samples from component i is binomial with mean Mw_i. By the Chernoff bound (Theorem 2.4) and a union bound over nice components, with probability at least $1 - \delta/3$, there are at least $m(\varepsilon/10)\log_3(3k/\delta)$ points from each nice component. If this is the case, then the definition of robust compression implies that, for each nice component g_i, with probability at least $1 - \delta/3k$, there exists a sequence $L_i \in S^{\tau(\varepsilon/10)}$ and a sequence $B_i \in \{0, 1\}^{t(\varepsilon/10)}$ such that $\|\mathcal{J}(L_i, B_i) - f_i\|_1 \leq \varepsilon/10$, where $\mathcal{J}$ is the decoder for $\mathcal{F}$. By a union bound over nice components, this is simultaneously true for all nice components, with probability at least $1 - \delta/3$.

Thus far, we have proved that, with probability at least $1 - 2\delta/3$, there exist sequences $L_1, \ldots, L_k \in S^{\tau(\varepsilon/10)}$ and $B_1, \ldots, B_k \in \{0, 1\}^{t(\varepsilon/10)}$ such that

$$\|\mathcal{J}(L_i, B_i) - f_i\|_1 \leq \varepsilon/10 \text{ for each nice component } i. \tag{7}$$

The learner builds the following set of candidate distributions:

$$C := \left\{ \sum_{i=1}^{k} w_i' \mathcal{J}(L_i', B_i') \;:\; L_1', \ldots, L_k' \in S^{\tau(\varepsilon/10)}, \; B_1', \ldots, B_k' \in \{0, 1\}^{t(\varepsilon/10)}, \; (w_1', \ldots, w_k') \in W \right\}.$$

We claim that, with probability at least $1 - 2\delta/3$, at least one of the distributions in C is $(3\varepsilon/10 + 2\rho/r + \rho)$-close to g. This corresponds to the "correct" sequences L, B, and $\widehat{w}$; that is, when $L_i' = L_i, B_i' = B_i$, and $w_i' = \widehat{w}_i$ for all $i \in [k]$. To prove the claim, let T, F, and N denote the set of tiny,

far, and nice components, respectively. Then, we have

$$\left\| \sum_{i \in [k]} \widehat{w}_i \mathcal{J}(L_i, B_i) - w_i g_i \right\|_1 \leq \left\| \sum_{i \in [k]} w_i (\mathcal{J}(L_i, B_i) - f_i) \right\|_1 + \left\| \sum_{i \in [k]} (\widehat{w}_i - w_i) \mathcal{J}(L_i, B_i) \right\|_1 + \| f - g \|_1$$

$$\leq \sum_{i \in T \cup F} w_i \| \mathcal{J}(L_i, B_i) - f_i \|_1 + \sum_{i \in N} w_i \| \mathcal{J}(L_i, B_i) - f_i \|_1$$

$$+ \sum_{i \in [k]} |\widehat{w}_i - w_i| \cdot \| \mathcal{J}(L_i, B_i) \|_1 + \rho$$

$$\leq \sum_{i \in T \cup F} w_i \cdot 2 + \sum_{i \in N} w_i \cdot (\varepsilon/10) + \sum_{i \in [k]} (\varepsilon/10k) + \rho$$

$$\leq (\varepsilon/10 + 2\rho/r) + \varepsilon/10 + \varepsilon/10 + \rho$$

$$= 3\varepsilon/10 + 2\rho/r + \rho,$$

where the first two inequalities follow from the triangle inequality and, since $\| f - g \|_1 \leq \rho$, the third inequality follows from inequalities (7) and (6), and the fourth inequality follows from the definition of tiny and Lemma 4.10. This proves the claim.

Next, the learner applies the algorithm of Theorem 4.4, with error parameter $\varepsilon/40$, to obtain a member of C whose distance from g is bounded by $3 \cdot (3\varepsilon/10 + 2\rho/r + \rho) + 4(\varepsilon/40) \leq \varepsilon + 3\rho(1 + 2/r)$, as required. The overall failure probability is bounded by $2\delta/3$ (failure probability of the claim) plus $\delta/3$ (failure probability of the algorithm of Theorem 4.4).

The sample complexity of the algorithm is bounded as follows: The number of candidate distributions can be bounded by

$$|C| \leq \left(M^{\tau(\varepsilon/10)} 2^{t(\varepsilon/10)} \right)^k \cdot (10k/\varepsilon)^k \leq M^{k\tau'(\varepsilon/10)} \cdot (10k/\varepsilon)^k,$$

whence the total sample complexity can be bounded by

$$M + \frac{\log(3|C|^2/\delta)}{2\varepsilon^2}$$

$$= O \left(m\left(\frac{\varepsilon}{10}\right) \log\left(\frac{k}{\delta}\right) \frac{k}{\varepsilon} + \frac{\log(1/\delta) + k\log(k/\varepsilon) + k\tau'(\varepsilon/10)\log\left(m(\varepsilon/10)\log(k/\delta)k/\varepsilon\right)}{\varepsilon^2} \right)$$

$$= \widetilde{O} \left(\frac{km(\varepsilon/10)}{\varepsilon} + \frac{k\tau'(\varepsilon/10)\log m(\varepsilon/10)}{\varepsilon^2} \right). \qquad \square$$

5 UPPER BOUNDS: LEARNING GAUSSIAN MIXTURES VIA COMPRESSION SCHEMES

The main positive results of this article are sample complexity bounds for learning mixtures of Gaussians (Theorems 1.5 and 1.7). In this section, we prove these results by describing a compression scheme for a single Gaussian and then applying the tools developed in the previous section. To begin, we illustrate the technique by analyzing the simpler problem of learning mixtures of axis-aligned Gaussians in the realizable setting. To avoid the technical issue of $\log(1/\varepsilon)$ getting extremely small, henceforth, we will assume that ε is bounded away from 1, say, $\varepsilon \leq 0.99$.

5.1 Warm-up: Mixtures of Axis-aligned Gaussians, Non-robustly

In this short section, we give an illustrative use of our compression framework to prove an upper bound of $\widetilde{O}(kd/\varepsilon^2)$ for the sample complexity of learning mixtures of k axis-aligned Gaussians in

the realizable setting. The next section gives a general argument for arbitrary Gaussians in the agnostic setting.

LEMMA 5.1. *The class of one-dimensional Gaussians admits a* $(3, O(\log(1/\varepsilon)), 3)$ *non-robust compression scheme.*

PROOF. Let $0 < c < 1 < C$ be such that $\mathbf{Pr}_{X \sim \mathcal{N}(0,1)}[c < |X| < C] \geq 0.99$. Let $\mathcal{N}(\mu, \sigma^2)$ be the target distribution. We first show how to encode σ. Let $g_1, g_2 \sim \mathcal{N}(\mu, \sigma^2)$. Then $g \coloneqq \frac{1}{\sqrt{2}}(g_1 - g_2) \sim \mathcal{N}(0, \sigma^2)$. So, with probability at least 0.99, we have $\sigma c < |g| < \sigma C$. Conditioned on this event, we have $\lambda \coloneqq \sigma/g \in [-1/c, 1/c]$. We now choose $\widehat{\lambda} \in \{0, \pm\varepsilon/2C^2, \pm 2\varepsilon/2C^2, \pm 3\varepsilon/2C^2 \ldots, \pm 1/c\}$ satisfying $|\widehat{\lambda} - \lambda| \leq \varepsilon/4C^2$, and we encode the standard deviation by $(g_1, g_2, \widehat{\lambda})$. The decoder then estimates $\widehat{\sigma} \coloneqq \widehat{\lambda}(g_1 - g_2)/\sqrt{2}$. Note that $|\widehat{\sigma} - \sigma| \leq |\widehat{\lambda} - \lambda||g| \leq \sigma\varepsilon/4C$ and that the encoding requires two sample points and $O(\log(C^2/c\varepsilon)) = O(\log(1/\varepsilon))$ bits for encoding $\widehat{\lambda}$.

Now, we turn to encoding μ. Let $g_3 \sim \mathcal{N}(\mu, \sigma^2)$. Then $|g_3 - \mu| \leq C\sigma$ with probability at least 0.99. We will condition on this event, which implies the existence of some $\eta \in [-C, C]$ satisfying $g_3 + \sigma\eta = \mu$. We choose $\widehat{\eta} \in \{0, \pm\varepsilon/2, \pm 2\varepsilon/2, \pm 3\varepsilon/2 \ldots, \pm C\}$ such that $|\widehat{\eta} - \eta| \leq \varepsilon/4$, and encode the mean by $(g_3, \widehat{\eta})$. The decoder estimates $\widehat{\mu} \coloneqq g_3 + \widehat{\sigma}\widehat{\eta}$. Again, note that $|\widehat{\mu} - \mu| = |\sigma\eta - \widehat{\sigma}\widehat{\eta}| \leq |\sigma\eta - \sigma\widehat{\eta}| + |\sigma\widehat{\eta} - \widehat{\sigma}\widehat{\eta}| \leq \sigma\varepsilon/2$. Moreover, encoding the mean requires one sample point and $O(\log(1/\varepsilon))$ bits.

To summarize, the decoder has $|\widehat{\mu} - \mu| \leq \sigma\varepsilon/2$ and $|\widehat{\sigma} - \sigma| \leq \sigma\varepsilon/2$. Plugging these bounds into Lemma 2.11 gives $\|\mathcal{N}(\mu, \sigma^2) - \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)\|_1 \leq \varepsilon$, as required. □

Remark 5.2. In the above argument, the samples are not "compressed" in the usual sense of this verb. Nevertheless, our formal definition of compression, Definition 4.2, allows $m = \tau$.

To complete the proof of Theorem 1.7 in the realizable setting, we note that Lemma 5.1 combined with Lemma 4.6 implies that the class of axis-aligned Gaussians in $\mathbb{R}^d$ admits an

$$\left(O(d),\ O(d\log(d/\varepsilon)),\ O(\log(3d)) \right)$$

non-robust compression scheme. (By noting that any axis-aligned Gaussian is a product of one-dimensional Gaussians.) Then, by Lemma 4.8, the class of mixtures of k axis-aligned Gaussians admits an

$$\left(O(kd),\ O(kd\log(d/\varepsilon) + k\log(k/\varepsilon)),\ O(k\log(k)\log(d)/\varepsilon) \right)$$

non-robust compression scheme. Theorem 4.5 now implies that the class of k-mixtures of axis-aligned Gaussians in $\mathbb{R}^d$ can be learned using $\widetilde{O}(kd/\varepsilon^2)$ samples in the realizable setting.

5.2 Learning Axis-aligned and General Gaussian Mixtures in the Agnostic Setting

We now turn to the general case and prove an upper bound of $\widetilde{O}(kd^2/\varepsilon^2)$ for the sample complexity of learning mixtures of k Gaussians in d dimensions, and an upper bound of $\widetilde{O}(kd/\varepsilon^2)$ for the sample complexity of learning mixtures of k axis-aligned Gaussians, both in the agnostic sense. The heart of the proof is to show that Gaussians have robust compression schemes in any dimension.

LEMMA 5.3. *For any positive integer d, the class $\mathcal{G}_{d,1}$ of d-dimensional Gaussians admits an*

$$\left(O(d),\ O(d^2\log(d/\varepsilon)),\ O(d) \right)$$

2/3-robust compression scheme.

Remark 5.4. In the special case $d = 1$, there also exists a $(4, 1, O(1/\varepsilon))$ 0.773-robust compression scheme using completely different ideas. The proof appears in our technical report [5]. Surprisingly, this compression scheme has constant size, as the value of $\tau + t$ is independent of ε (unlike

Lemma 5.3). This scheme could be used instead of Lemma 5.3 for proving Theorem 1.7, although it would not improve the sample complexity bound asymptotically.

Remark 5.5. The proof of Lemma 5.3 can be amended to give an r-robust compression scheme for any $r < 1$, which will change the constant 12 in the agnostic results of Theorem 1.5 and Theorem 1.7 to any constant larger than 9, at the expense of worse constants for τ, t and m. This is straightforward but entails additional cumbersome notation, hence, we omit the details.

Before proving Lemma 5.3, we show how to combine it with previous lemmata to prove our main upper bounds.

PROOF OF THEOREM 1.5. Combining Lemma 5.3 and Lemma 4.9 gives that the class of k-mixtures of d-dimensional Gaussians is 12-agnostically learnable with sample complexity $\widetilde{O}(kd^2/\varepsilon^2)$.   □

PROOF OF THEOREM 1.7. Recall that $\mathcal{A}_{d,k}$ denotes the class of k-mixtures of d-dimensional axis-aligned Gaussian distributions. Observe that any distribution in $\mathcal{A}_{d,1}$ is a product of d one-dimensional Gaussians, hence applying Lemma 5.3 for the case $d = 1$ and then Lemma 4.6 shows that $\mathcal{A}_{d,1}$ admits $\big(O(d), O(d \log(d/\varepsilon)), O(\log(3d))\big)$ 2/3-robust compression. Lemma 4.9 then implies that $\mathcal{A}_{d,k}$ is 12-agnostically learnable with sample complexity $\widetilde{O}(kd/\varepsilon^2)$, completing the proof.   □

5.3 Proof of Lemma 5.3

We first provide an overview of the proof. For simplicity, first assume that we are in the non-robust, zero-mean, full-rank scenario, i.e., we want to encode the distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ has rank d. Let $v_1, \ldots, v_d$ be an orthogonal set of vectors that satisfy $\Sigma = \sum_{i=1}^{d} v_i v_i^\mathsf{T}$. (Such a representation can be obtained from the eigendecomposition of Σ and noting that all its eigenvalues are positive. Note that the vectors v_i are not normalized.) Let $g_1, \ldots, g_d$ be i.i.d. samples from $\mathcal{N}(0, \Sigma)$. As $\mathrm{span}\{g_1, \ldots, g_d\} = \mathbb{R}^d$ with probability 1, a natural idea is to find, for each i, coefficients $\lambda_{i,1}, \ldots, \lambda_{i,d}$ such that $v_i = \sum_{j=1}^{d} \lambda_{i,j} g_j$. The encoder sends $g_1, \ldots, g_d$ and a discretization of the values $\{\lambda_{i,j}\}_{i,j \in [d]}$ as her message and the decoder will recover Σ approximately. If the discretization of each $\lambda_{i,j}$ is accomplished with b bits, then we would obtain a $(d, d^2 b, d)$ compression scheme.

The main difficulty is to control the bit complexity of a suitable discretization of $\lambda_{i,j}$—to achieve the optimal sample complexity bound, the bit complexity must be polylog(d/ε) per coefficient. The key to achieve a suitable discretization is the following fact from geometric functional analysis (Lemma 5.6, cf. Reference [34, Corollary 4.1]): Denote the convex hull of a set T is by conv(T). Given a sequence $g_1, \ldots, g_m$ of $m = O(d)$ i.i.d. samples from $\mathcal{N}(0, \Sigma)$, with high probability, we have $\frac{1}{20} \cdot \mathcal{E} \subseteq \mathrm{conv}\{\pm g_1, \ldots, \pm g_m\}$, where $\mathcal{E}$ is the ellipsoid centered at zero with principal axes $v_1, \ldots, v_d$. This enables us to express each v_i as $\sum_{j=1}^{m} \lambda_{i,j} g_j$, where each $\lambda_{i,j} \in [-20, 20]$; we then discretize by building an poly(ε/d)-net of size poly(d/ε) on this interval, achieving the desired polylog(d/ε) per coefficient bit complexity.

Next, suppose the mean is not zero, say we want to encode the distribution $\mathcal{N}(\mu, \Sigma)$. Note that if $g_1, g_2 \sim \mathcal{N}(\mu, \Sigma)$, then $\frac{g_1 - g_2}{\sqrt{2}} \sim \mathcal{N}(0, \Sigma)$, and thus, we can use the same compression scheme as above to encode $v_1, \ldots, v_d$. To encode μ, the idea is that a *single* sample $g \sim \mathcal{N}(0, \Sigma)$ is unlikely to be too far from μ. Specifically, if $\mathcal{E}$ is the ellipsoid defined by Σ, centered at zero, then, with high probability, $\mu \in g + O(\sqrt{d}) \cdot \mathcal{E}$. Thus, we build a net of the set $g + O(\sqrt{d}) \cdot \mathcal{E}$ and the encoder sends g as well as the identity of the point in the net closest to μ.

We now proceed to the formal proof. We first prove a lemma that is similar to known results in random matrix theory (cf. Reference [34, Corollary 4.1]) but is tailored for our purposes. Let us denote $a \cdot B_2^d := \{ y \in \mathbb{R}^d : \|y\|_2 \le a \}$.

LEMMA 5.6. *Suppose that $q_1, \ldots, q_m \in \mathbb{R}^d$ are i.i.d. samples from some distribution Q that satisfies* TV $(Q, \mathcal{N}(0, I_d)) \le 2/3$. *Let*

$$T := \{ \pm q_i : \|q_i\|_2 \le 4\sqrt{d} \}.$$

For a large enough absolute constant C, if $m \ge Cd$ then

$$\mathbf{Pr}\left[\frac{1}{20} B_2^d \subseteq \mathrm{conv}(T) \right] \ge 5/6.$$

PROOF. Let $S^{d-1} := \{ y \in \mathbb{R}^d : \|y\|_2 = 1 \}$. Consider the following statement:

$$\max_{q \in T} |\langle y, q \rangle| \ge \frac{1}{20} \qquad \forall y \in S^{d-1}. \tag{8}$$

We first show that Inequality (8) implies that $\frac{1}{20} B_2^d \subseteq \mathrm{conv}(T)$, which is the event that we wish to analyze. To see this, suppose, for the sake of contradiction, that $\frac{1}{20} B_2^d \not\subseteq \mathrm{conv}(T)$, so there exists some point $z \in \frac{1}{20} B_2^d \setminus \mathrm{conv}(T)$. By the strict hyperplane separation theorem (e.g., Reference [10, Example 2.20]), there exists some direction $y \in S^{d-1}$ such that $\langle y, z \rangle > \langle y, q \rangle$ for all $q \in T$. Since $\|z\|_2 \le 1/20$, we have $1/20 \ge \langle y, z \rangle$. This means $1/20 > \max_{q \in T} \langle y, q \rangle = \max_{q \in T} |\langle y, q \rangle|$, contradicting Inequality (8).

Next, for each $y \in S^{d-1}$, let

$$H_y := \left\{ x \in \mathbb{R}^d : \|x\|_2 \le 4\sqrt{d}, |\langle x, y \rangle| \ge \frac{1}{20} \right\},$$

and let $\mathcal{H} := \{ H_y : y \in S^{d-1} \}$. Let $U := \{ q_1, \ldots, q_m \}$ and observe that Inequality (8) is equivalent to the event

$$U \cap H \ne \emptyset \qquad \forall H \in \mathcal{H}.$$

So, to complete the proof, we need only show that this event happens with probability at least $5/6$. Recalling that U is an i.i.d. sample of size m from Q, by the Vapnik-Chervonenkis inequality (see, e.g., Reference [47, Theorem 8.3.23]), for some absolute constant c,

$$\mathbf{E} \sup_{H \in \mathcal{H}} \left| Q(H) - \frac{|U \cap H|}{|U|} \right| \le c\sqrt{\mathrm{VC\text{-}dim}(\mathcal{H})/m}. \tag{9}$$

Define $p := \inf_{H \in \mathcal{H}} Q(H)$. We claim that, to complete the proof of the lemma, it suffices to show that VC-dim$(\mathcal{H}) = O(d)$ and $p = \Omega(1)$. Indeed, if these statements are true, then we can choose $m = 144c^2 \cdot \mathrm{VC\text{-}dim}(\mathcal{H})/p^2 = O(d)$ and the right-hand-side of Inequality (9) becomes $p/12$. Markov's inequality would then imply that with probability at least $5/6$, for all $H \in \mathcal{H}$, we have $Q(H) - \frac{|U \cap H|}{|U|} \le p/2$, which implies

$$\frac{|U \cap H|}{|U|} \ge Q(H) - p/2 \ge p/2 > 0,$$

that is, $U \cap H$ is nonempty for all $H \in \mathcal{H}$. Next, we show that VC-dim$(\mathcal{H}) = O(d)$ and $p = \Omega(1)$.

Since restricting the domain cannot increase the VC dimension, the VC dimension of $\mathcal{H}$ is not more than that of $\{\{ x \in \mathbb{R}^d : |\langle x, y \rangle| \ge \frac{1}{20} \} : y \in S^{d-1} \}$. Every set in this family is a union of two half-spaces (corresponding to $\langle x, y \rangle \ge \frac{1}{20}$ and $\langle x, y \rangle \le -\frac{1}{20}$). The VC dimension of the family of half-spaces is $d + 1$ (see, e.g., Reference [17, Corollary 4.2]), so, by Reference [9, Lemma 3.2.3], the VC dimension of the family of pairwise unions of half-spaces is bounded by $4(d + 1) \log_2(6) = O(d)$.

Finally, to show that $p = \Omega(1)$, let $g \sim \mathcal{N}(0, I_d)$ and note that for any $y \in S^{d-1}$, by the union bound,

$$\Pr[g \in H_y] \geq \Pr[\|g\|_2 \leq 4\sqrt{d}] - \Pr[|\langle g, y \rangle| < 1/20] \geq (1 - e^{-3}) - 1/10 > 0.85, \tag{10}$$

where we have used that $\Pr[\|g\|_2 > 4\sqrt{d}] \leq \exp(-3)$ (by Corollary 2.3) and that $\langle g, y \rangle$ is a standard normal random variable so its probability density function is bounded by 1. Since $\mathrm{TV}(Q, \mathcal{N}(0, I_d)) \leq 2/3$, from Inequality (10), we obtain $Q(H_y) \geq 0.85 - 2/3 > 0.18$, completing the proof of the lemma. □

To prove Lemma 5.3, we need to describe a 2/3-robust compression scheme for d-dimensional Gaussians. Accordingly, we consider a target distribution Q such that there exists a Gaussian $\mathcal{N}(\mu, \Sigma)$ satisfying $\|Q - \mathcal{N}(\mu, \Sigma)\|_1 \leq 2/3$. Recall, from Equation (1), that this implies $\mathrm{TV}(Q, \mathcal{N}(\mu, \Sigma)) \leq 1/3$.

We may assume that Σ has full rank, since there is a reduction from the case of rank-deficient Σ: If the rank of Σ is $\rho < d$, then any $X \sim \mathcal{N}(\mu, \Sigma)$ lies in some affine subspace $\mathcal{S}$ of dimension ρ. Thus, since $\mathrm{TV}(Q, \mathcal{N}(\mu, \Sigma)) \leq 1/3$, any $X \sim Q$ lies in $\mathcal{S}$ with probability at least 2/3. By the Chernoff bound (Theorem 2.4), with high probability, after seeing $10d$ samples from Q, at least $\rho + 1$ points from $\mathcal{S}$ will appear in the sample. We encode $\mathcal{S}$ using these samples, and for the rest of the process we work in this affine subspace and discard outside points.

Definition of $v_1, \ldots, v_d$, and Ψ. Since Σ has full rank, there exist an orthogonal set of vectors $v_1, \ldots, v_d$ satisfying $\Sigma = \sum_{i=1}^d v_i v_i^\mathsf{T}$. For convenience, let $\Psi = \Sigma^{1/2}$ be the unique positive definite square root of Σ. Observe that

$$\Psi = \sum_{i=1}^d \frac{v_i v_i^\mathsf{T}}{\|v_i\|_2}, \qquad \Sigma^{-1} = \sum_{i=1}^d \frac{v_i v_i^\mathsf{T}}{\|v_i\|_2^4}, \qquad \text{and} \qquad \Psi^{-1} = \sum_{i=1}^d \frac{v_i v_i^\mathsf{T}}{\|v_i\|_2^3}. \tag{11}$$

We next show how to encode the mean and the eigenvectors.

LEMMA 5.7. *Let C be a sufficiently large absolute constant, and let S be an i.i.d. sample of size $2m = 2Cd$ from Q, where $\mathrm{TV}(Q, \mathcal{N}(\mu, \Sigma)) \leq 1/3$. Then, with probability at least 2/3, one can encode vectors $\widehat{v}_1, \ldots, \widehat{v}_d, \widehat{\mu} \in \mathbb{R}^d$ satisfying*

$$\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 \leq \varepsilon/24d^2 \qquad \forall j \in [d], \text{ and} \tag{12}$$

$$\|\Psi^{-1}(\widehat{\mu} - \mu)\|_2 \leq \varepsilon/2, \tag{13}$$

using $O(d^2 \log(d/\varepsilon))$ bits and the points in S.

PROOF. The samples in S will be denoted $X_1, \ldots, X_{2m}$.

Encoding $\widehat{v}_1, \ldots, \widehat{v}_d$. Define the "standardized" samples

$$Y_i := \frac{1}{\sqrt{2}} \Psi^{-1}(X_{2i} - X_{2i-1}) \qquad \forall i \in [m].$$

We claim that, for each $i \in [m]$, Y_i has TV distance at most 2/3 from $\mathcal{N}(0, I)$. Indeed, If X_{2i} and X_{2i-1} had distribution $\mathcal{N}(\mu, \Sigma)$, then Y_i would have distribution $\mathcal{N}(0, I)$. Instead, both X_{2i} and X_{2i-1} have TV distance at most 1/3 from $\mathcal{N}(\mu, \Sigma)$, so, by Lemma 4.7 and Proposition 2.1, Y_i has TV distance at most 2/3 from $\mathcal{N}(0, I)$. Define the event

$$\mathcal{E} := \left\{ \frac{1}{C} B_2^d \subseteq \mathrm{conv}\{\pm Y_i : i \in \mathcal{I}\} \right\}, \text{ where } \mathcal{I} := \left\{ i \in [m] : \|Y_i\|_2 \leq 4\sqrt{d} \right\}.$$

Since C is large, and in particular $C \geq 20$, by Lemma 5.6, we have $\Pr[\mathcal{E}] \geq 5/6$. Our encoding will assume $\mathcal{E}$ happens.

Fix some $j \in [d]$. Referring to Equations (11), we see that $\Psi^{-1} v_j = v_j / \|v_j\|_2$ has unit norm. Since $\mathcal{E}$ occurs, we can write

$$\frac{\Psi^{-1} v_j}{C} = \sum_{i \in [m]} \theta_{j,i} Y_i$$

for some vector $\theta_j \in [-1, 1]^m$ supported on $\mathcal{I}$. Applying Ψ to both sides, we obtain

$$v_j = \frac{C}{\sqrt{2}} \sum_{i \in \mathcal{I}} \theta_{j,i} (X_{2i} - X_{2i-1}).$$

To discretize θ_j, consider the natural $(\varepsilon/96Cmd^3)$-net for $[-1, 1]^m$ in the ℓ_∞ distance formed by the Cartesian product of one-dimensional nets, namely, $\{n\varepsilon/48Cmd^3 : -\lfloor 48Cmd^3/\varepsilon \rfloor \le n \le \lfloor 48Cmd^3/\varepsilon \rfloor\}^m$. This net has size at most $(97Cmd^3/\varepsilon)^m$. Recalling that $m = O(d)$, it follows that any element of the net can be described using $O(d \log(d/\varepsilon))$ bits. Let $\widehat{\theta}_j$ be an element in the net that is closest to θ_j. Since each θ_j is supported on $\mathcal{I}$ and due to the structure of the net, we may choose $\widehat{\theta}_j$ also to be supported on $\mathcal{I}$. Define

$$\widehat{v}_j := \frac{C}{\sqrt{2}} \sum_{i \in \mathcal{I}} \widehat{\theta}_{j,i} (X_{2i} - X_{2i-1}).$$

The vectors $\widehat{v}_1, \ldots, \widehat{v}_d$ are encoded using the points in S and $\widehat{\theta}_1, \ldots, \widehat{\theta}_d$. Encoding each $\widehat{\theta}_i$ requires $O(d \log(2d/\varepsilon))$ bits, so encoding all of them requires $O(d^2 \log(2d/\varepsilon))$ bits.

The error of this encoding is

$$\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 = \frac{C}{\sqrt{2}} \left\| \sum_{i \in \mathcal{I}} (\theta_{j,i} - \widehat{\theta}_{j,i}) \Psi^{-1}(X_{2i} - X_{2i-1}) \right\|_2$$

$$\le \frac{C}{\sqrt{2}} |\mathcal{I}| \left(\max_{i \in \mathcal{I}} |\theta_{j,i} - \widehat{\theta}_{j,i}| \right) \left(\max_{i \in \mathcal{I}} \sqrt{2} \|Y_i\|_2 \right).$$

$$\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 \le \frac{C}{\sqrt{2}} m \left(\frac{\varepsilon}{96Cmd^3} \right) \left(4\sqrt{2}\sqrt{d} \right) \le \frac{\varepsilon}{24d^2},$$

By the definition of $\widehat{\theta}_j$, we have $\|\widehat{\theta}_j - \theta_j\|_\infty \le \varepsilon/96Cmd^3$. By the definition of $\mathcal{I}$, we have $\|Y_i\|_2 \le 4\sqrt{d}$, leading to the bound

$$\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 \le \frac{C}{\sqrt{2}} m \left(\frac{\varepsilon}{96Cmd^3} \right) \left(4\sqrt{2}\sqrt{d} \right) \le \frac{\varepsilon}{24d^2}, \tag{14}$$

establishing Inequality (12).

Encoding $\widehat{\mu}$. Let $Z_i := \Psi^{-1}(X_i - \mu)$ and observe that Z_i has a distribution with TV distance at most $1/3$ to $\mathcal{N}(0, I)$. Define the event

$$\mathcal{E}' := \left\{ \min\{\|Z_1\|_2, \|Z_2\|_2\} \le 4\sqrt{d} \right\}.$$

Corollary 2.3 implies that

$$\mathbf{Pr}[\|Z_i\|_2 \ge 4\sqrt{d}] \le \exp(-3) + 1/3 < \sqrt{1/6}.$$

Thus, $\mathbf{Pr}[\mathcal{E}'] \ge 5/6$. Our encoding will assume that $\mathcal{E}'$ happens.

By symmetry, assume that $\|Z_1\|_2 \le 4\sqrt{d}$, and suppose that $Z_1 = \sum_{j \in [d]} \lambda_j v_j / \|v_j\|_2$. Thus, we have $(\lambda_1, \ldots, \lambda_d) \in 4\sqrt{d} B_2^d$. Also, by the definition of Ψ, we have

$$\mu = X_1 - \Psi Z_1 = X_1 - \left(\sum_{i=1}^{d} \frac{v_i v_i^\mathsf{T}}{\|v_i\|_2} \right) \left(\sum_{j=1}^{d} \lambda_j \frac{v_j}{\|v_j\|_2} \right) = X_1 - \sum_{j \in [d]} \lambda_j v_j.$$

To discretize $(\lambda_1, \ldots, \lambda_d)$, consider an $(\varepsilon/3d)$-net for $4\sqrt{d} B_2^d$ of size $O(d^{1.5}/\varepsilon)^d$ (see Lemma 2.13). Let $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ be the closest element to $(\lambda_1, \ldots, \lambda_d)$ in this net. The encoding is

$$\widehat{\mu} := X_1 - \sum_{j \in [d]} \widehat{\lambda}_j \widehat{v}_j.$$

The error of this encoding is

$$
\begin{aligned}
\|\Psi^{-1}(\mu - \widehat{\mu})\|_2 &= \left\| \sum_{j \in [d]} \Psi^{-1}(\lambda_j v_j - \widehat{\lambda}_j \widehat{v}_j) \right\|_2 \\
&\le \sum_{j \in [d]} \|\widehat{\lambda}_j (\Psi^{-1} v_j - \Psi^{-1}\widehat{v}_j) + (\lambda_j - \widehat{\lambda}_j)\Psi^{-1} v_j\|_2 \\
&\le d \cdot \max_{j \in [d]} \{ |\widehat{\lambda}_j| \cdot \|\Psi^{-1} v_j - \Psi^{-1}\widehat{v}_j\|_2 + |\lambda_j - \widehat{\lambda}_j| \cdot \|\Psi^{-1} v_j\|_2 \}.
\end{aligned}
$$

By the definition of $\widehat{\lambda}$, we have $\max_j |\widehat{\lambda}_j| \le 4\sqrt{d}$ and $\max_j |\lambda_j - \widehat{\lambda}_j| \le \varepsilon/3d$. From Equations (11), we have $\|\Psi^{-1} v_j\|_2 = 1$. Last, using Equation (14), we have $\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 \le \varepsilon/24d^2$, leading to the bound

$$\|\Psi^{-1}(\mu - \widehat{\mu})\|_2 \le d \cdot \left(4\sqrt{d} \cdot \frac{\varepsilon}{24d^2} + \frac{\varepsilon}{3d} \cdot 1 \right) \le \varepsilon/2,$$

establishing Inequality (13). The encoding for $\widehat{\mu}$ consists of X_1, the already encoded $\widehat{v}_1, \ldots, \widehat{v}_d$, and $\widehat{\lambda}$. Since $\widehat{\lambda}$ comes from a net of size $O(d^{1.5}/\varepsilon)^d$, the additional number of required bits to encode $\widehat{\mu}$ is $O(d \log(d/\varepsilon))$. Finally, note that all encodings succeed so long as both $\mathcal{E}$ and $\mathcal{E}'$ occur, which happens with probability at least $2/3$.   □

Lemma 5.3 now follows immediately from the following lemma:

LEMMA 5.8. *Suppose that the vectors* $\widehat{v}_1, \ldots, \widehat{v}_d, \widehat{\mu} \in \mathbb{R}^d$ *satisfy*

$$\|\Psi^{-1}(\widehat{v}_j - v_j)\|_2 \le \rho \le 1/6d \qquad \forall j \in [d], \text{ and} \tag{15}$$

$$\|\Psi^{-1}(\widehat{\mu} - \mu)\|_2 \le \zeta. \tag{16}$$

Then

$$\mathrm{TV} \left(\mathcal{N}\left(\mu, \textstyle\sum_{i \in [d]} v_i v_i^\mathsf{T} \right), \mathcal{N}\left(\widehat{\mu}, \textstyle\sum_{i \in [d]} \widehat{v}_i \widehat{v}_i^\mathsf{T} \right) \right) \le \frac{\sqrt{9d^3 \rho^2 + \zeta^2}}{2}.$$

PROOF. Recall the definition of log-det divergence: $\mathrm{LD}(A, B) := \mathrm{Tr}(B^{-1}A - I) - \log \det(B^{-1}A)$. Define $\widehat{\Sigma} := \sum_i \widehat{v}_i \widehat{v}_i^\mathsf{T}$. We will show that

$$\mathrm{LD}(\widehat{\Sigma}, \Sigma) \le 9d^3 \rho^2. \tag{17}$$

If this is true, then Lemma 2.10 and Inequality (16) yield

$$\mathrm{TV}\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma}) \right)^2 \le \frac{1}{4}\left(\mathrm{LD}(\widehat{\Sigma}, \Sigma) + (\mu - \widehat{\mu})^\mathsf{T} \Sigma^{-1}(\mu - \widehat{\mu}) \right) \le \frac{1}{4}(9d^3\rho^2 + \zeta^2),$$

which completes the proof of the lemma.

Thus, we need only prove Inequality (17). Recall from Equations (11) that $\Psi = \Sigma^{1/2}$ is positive definite. Define $B := \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} = \Psi^{-1}\widehat{\Sigma}\Psi$. We will show that $\|B - I\|_s \leq 3d\rho$. Then Lemma 2.9 will imply that $\mathrm{LD}(\widehat{\Sigma}, \Sigma) \leq 9d^3\rho^2$, which establishes Inequality (17).

To complete the proof, note that by the triangle inequality

$$\|B - I\|_s = \left\| \sum_{i=1}^{d} \left(\Psi^{-1}\widehat{v}_i\widehat{v}_i^\mathsf{T}\Psi^{-1} - \Psi^{-1}v_iv_i^\mathsf{T}\Psi^{-1} \right) \right\|_s \leq \sum_{i=1}^{d} \left\| \Psi^{-1}\widehat{v}_i\widehat{v}_i^\mathsf{T}\Psi^{-1} - \Psi^{-1}v_iv_i^\mathsf{T}\Psi^{-1} \right\|_s$$

$$= \sum_{i=1}^{d} \left\| x_ix_i^\mathsf{T} - y_iy_i^\mathsf{T} \right\|_s,$$

with $x_i := \Psi^{-1}\widehat{v}_i$ and $y_i := \Psi^{-1}v_i$. From Equations (11), we have $\|y_i\|_2 = \|\Psi^{-1}v_i\|_2 = 1$. By Inequality (15), $\|x_i - y_i\|_2 \leq \rho$. By applying the following simple lemma, we conclude that $\|B - I\|_s \leq 3d\rho$, completing the proof of the lemma. □

LEMMA 5.9. *Let x and y be vectors satisfying $\|y\|_2 = 1$ and $\|x - y\|_2 \leq \varepsilon \leq 1$. Then, we have $\|xx^\mathsf{T} - yy^\mathsf{T}\|_s \leq 3\varepsilon$.*

PROOF. Suppose $x = y + z$ with $\|z\|_2 \leq \varepsilon$. Then,

$$\|xx^\mathsf{T} - yy^\mathsf{T}\|_s = \|yz^\mathsf{T} + zy^\mathsf{T} + zz^\mathsf{T}\|_s \leq \|yz^\mathsf{T}\|_s + \|zy^\mathsf{T}\|_s + \|zz^\mathsf{T}\|_s \leq \varepsilon + \varepsilon + \varepsilon^2 \leq 3\varepsilon.$$

The first inequality is the triangle inequality for the operator norm. The second inequality uses the facts that $\|AB\|_s \leq \|A\|_s \cdot \|B\|_s$ for any two size-compatible matrices A and B (see, e.g., Reference [25, Fact 7(c) in Section 24.4]) and that, for any vector v, the operator norm of v as a matrix coincides with its Euclidean norm as a vector. □

6 THE LOWER BOUND FOR GAUSSIANS AND THEIR MIXTURES

In this section, we establish a lower bound of $\widetilde{\Omega}(d^2/\varepsilon^2)$ for learning a single Gaussian, and then lift it to obtain a lower bound of $\widetilde{\Omega}(kd^2/\varepsilon^2)$ for learning mixtures of k Gaussians in d dimensions. Both of our lower bounds are for the realizable setting and thus also hold in the agnostic setting.

The high-level strategy for our proof is similar to that adopted in earlier work for mixtures of spherical Gaussians [44]. The idea is to create a large number of distributions that are pairwise close in KL divergence (roughly ε^2) but pairwise far in TV distance (roughly ε). An application of the following lemma will then yield the desired sample complexity lower bound.

LEMMA 6.1. *Let $\kappa : \mathbb{R} \to \mathbb{R}$ be a function and let $\mathcal{F}$ be a class of distributions such that, for all small enough $\varepsilon > 0$, there exist distributions $f_1, \dots, f_M \in \mathcal{F}$ with*

$$\mathrm{KL}(f_i \| f_j) \leq \kappa(\varepsilon) \quad \text{and} \quad \mathrm{TV}(f_i, f_j) > 2\varepsilon \qquad \forall i \neq j \in [M].$$

Then any method that learns $\mathcal{F}$ to within total variation distance ε with success probability at least $2/3$ has sample complexity $\Omega\left(\frac{\log M}{\kappa(\varepsilon)\log(1/\varepsilon)}\right)$.

The preceding lemma is a straightforward consequence of the following result, which is a generalized form of Fano's inequality in information theory [12, Theorem 2.10.1].

LEMMA 6.2 (GENERALIZED FANO'S INEQUALITY [49, LEMMA 3]). *Let the distributions $f_1, \dots, f_M$ satisfy*

$$\mathrm{KL}(f_i \| f_j) \leq \beta \quad \text{and} \quad \|f_i - f_j\|_1 > \alpha \qquad \forall i \neq j \in [M].$$

Consider any density estimation method that has an explicit description of $f_1, \dots, f_M$, receives n i.i.d. samples from some f_i without knowing i, then outputs an estimate $\widehat{f}$ for f_i. For each i, define

$e_i := \mathbf{E}\|f_i - \widehat{f}\|_1$ *for the case in which the method receives samples from f_i. Then*

$$\max_i e_i \ \geq \ \alpha \cdot \left(\frac{\log M - n\beta + \log 2}{2\log M} \right).$$

PROOF OF LEMMA 6.1. Consider a distribution learning method $\mathcal{A}$ that takes $m(\varepsilon)$ samples and learns $\mathcal{F}$ to within total variation distance ε with success probability at least 2/3. Our goal is to prove $m(\varepsilon) = \Omega(\log M / \kappa(\varepsilon) \log(1/\varepsilon))$. Consider M distributions $f_1, \ldots, f_M$ satisfying the hypotheses. We will design an estimator $\widehat{f}$ for the finite class $\{f_1, \ldots, f_M\}$ with sample complexity $n = km(\varepsilon)$ and then apply Lemma 6.2 to this estimator.

Suppose we are given $km(\varepsilon)$ samples from the unknown f_j. We partition the sample into k equal parts and run $\mathcal{A}$ on each part. This gives us k outputs $g_1, \ldots, g_k$. If some f_ℓ is within TV distance ε of more than half of these outputs, then we output $\widehat{f} = f_\ell$; otherwise, we output $\widehat{f} = f_1$. Note that, since $f_1, \ldots, f_M$ are 2ε-separated in the TV distance, there will be no ambiguity. We now analyze the error of this estimator.

By the guarantee of $\mathcal{A}$, for each $i \in [k]$, with probability at least 2/3, we have $\mathrm{TV}(g_i, f_j) \leq \varepsilon$. Let $\mathcal{E}$ be the event that more than half of the g_i satisfy $\mathrm{TV}(g_i, f_j) \leq \varepsilon$. When $\mathcal{E}$ happens, the estimate will be $\widehat{f} = f_j$, i.e., $\|f_j - \widehat{f}\|_1 = 0$. Moreover, by the Chernoff bound (Theorem 2.4), $\Pr[\mathcal{E}] \geq 1 - \exp(-\Omega(k))$. Thus, the expected error is

$$e_j \ = \ \mathbf{E}\|f_j - \widehat{f}\|_1 \ \leq \ \Pr[\mathcal{E}^c] \cdot 2 \ \leq \ \exp(-\Omega(k)) \qquad \forall j \in [M].$$

The total number of samples is $n = km(\varepsilon)$, so Lemma 6.2 gives

$$2\varepsilon \cdot \left(\frac{\log M - (km(\varepsilon))\kappa(\varepsilon) + \log 2}{2\log M} \right) \ \leq \ \exp(-\Omega(k)).$$

Choosing $k = \Theta(\log(1/\varepsilon))$ and rearranging gives $m(\varepsilon) = \Omega(\log M / \kappa(\varepsilon) \log(1/\varepsilon))$, as required. ☐

6.1 The Lower Bound for Learning a Single Gaussian

Our lower bound for learning a single Gaussian is the following theorem:

THEOREM 6.3. *Any algorithm that learns the class of d-dimensional Gaussians in the realizable setting within total variation distance ε and with success probability 2/3 has sample complexity* $\Omega\left(\frac{d^2}{\varepsilon^2 \log(1/\varepsilon)} \right)$.

PROOF. To apply Lemma 6.1, we must create a large number M of Gaussian distributions whose pairwise KL divergences are at most $\kappa(\varepsilon)$ and whose pairwise TV distances are at least 2ε. We will accomplish this with parameters $M = 2^{\Omega(d^2)}$ and $\kappa = O(\varepsilon^2)$, and Lemma 6.1 yields the desired lower bound.

The existence of such M distributions is shown using the probabilistic method. Let us fix the parameters $r = 9$ and $\lambda = \Theta(\varepsilon d^{-1/2})$. Assume, for simplicity, that d/r is an integer. For each $a \in [M]$, we pick U_a to be a uniformly random $d \times d/r$ matrix with orthonormal columns. That is, U_a consists of the first d/r columns of a uniformly random $d \times d$ orthogonal matrix (a random matrix with Haar measure on the group of $d \times d$ orthogonal matrices). From this, we create the distribution

$$f_a := \mathcal{N}(0, \Sigma_a), \qquad \text{where} \qquad \Sigma_a = I_d + \lambda U_a U_a^\top \qquad \forall a \in [M].$$

To apply Lemma 6.1, we must analyze the pairwise KL divergences and TV distances between $f_1, \ldots, f_M$. Observe that, by the construction of U_a, for any fixed $d \times d$ orthogonal matrix W, the matrices U_a and WU_a have the same distribution, that is, U_a is rotationally invariant (see, e.g., Reference [36, Section 1.2]).

Bounding the KL divergences. This analysis is straightforward, since there is a closed-form expression for the KL divergence between Gaussians. First, observe that all Σ_a have the same spectrum: There are d/r eigenvalues equal to $1 + \lambda$ and the remaining eigenvalues equal 1. Consequently,

$$\log \det \left(\Sigma_b \Sigma_a^{-1} \right) \;=\; \log \left(\det \Sigma_b \cdot \det \Sigma_a^{-1} \right) \;=\; 0. \tag{18}$$

Next observe that

$$\Sigma_a^{-1} \;=\; I - \frac{\lambda}{1 + \lambda} U_a U_a^\mathsf{T}, \tag{19}$$

which can be verified simply by multiplying by Σ_a. Thus,

$$
\begin{aligned}
2 \cdot \mathrm{KL}\left(f_a \parallel f_b \right) &= \mathrm{Tr}\left(\Sigma_a^{-1} \Sigma_b - I \right) &&\text{(by Equation (18) and Lemma 2.10)} \\[4pt]
&= \mathrm{Tr}\left(\left(I - \frac{\lambda}{1 + \lambda} U_a U_a^\mathsf{T} \right) \left(I + \lambda U_b U_b^\mathsf{T} \right) - I \right) &&\text{(by Equation (19))} \\[4pt]
&= \mathrm{Tr}\left(\lambda U_b U_b^\mathsf{T} - \frac{\lambda}{1 + \lambda} U_a U_a^\mathsf{T} - \frac{\lambda^2}{1 + \lambda} U_a U_a^\mathsf{T} U_b U_b^\mathsf{T} \right) \\[4pt]
&= \lambda \cdot \frac{d}{r} - \frac{\lambda}{1 + \lambda} \cdot \frac{d}{r} - \frac{\lambda^2}{1 + \lambda} \cdot \left\| U_a^\mathsf{T} U_b \right\|_F^2 \\[4pt]
&\leq \lambda \cdot \frac{d}{r} - \frac{\lambda}{1 + \lambda} \cdot \frac{d}{r} = \frac{\lambda^2 d}{(1 + \lambda) r} \leq \frac{\lambda^2 d}{r} = O(\varepsilon^2).
\end{aligned}
$$

This bound holds with probability 1.

Bounding the TV distances. The remaining step is to show that $\mathrm{TV}\left(f_a, f_b \right) = \Omega(\varepsilon)$ for all $a \neq b$. Then, by scaling ε by a constant factor, Lemma 6.1 completes the proof.

First, we provide some intuition on why such a bound should hold. Let S_a be the subspace spanned by the columns of U_a. A vector drawn from $\mathcal{N}(0, \Sigma_a)$ has a slightly larger projection onto S_a than a vector drawn from $\mathcal{N}(0, \Sigma_b)$. This reveals an event that has slightly higher probability under the former distribution than under the latter. Recalling the definition of the TV distance as a supremum over events (see Equation (1)), such an argument gives the desired lower bound on the TV distances up to logarithmic factors. This approach was used in a preliminary version of this work [5, version 2], but it is fairly technical.

Here, we use a simpler argument, formulated as Lemma 6.6 below, which shows that if $\|U_a^\mathsf{T} U_b\|_F^2 \leq d/2r$, then $\mathrm{TV}\left(f_a, f_b \right) = \Omega(\varepsilon)$. The condition $\|U_a^\mathsf{T} U_b\|_F^2 \leq d/2r$ means that the columns of U_a are nearly pairwise orthogonal to the columns of U_b, which intuitively should hold, since U_a and U_b are chosen randomly. This is formalized in Lemma 6.4 below, which states that, with positive probability, $\|U_a^\mathsf{T} U_b\|_F^2 \leq d/2r$ for all $a \neq b$. Then Lemma 6.6 implies that, by our choice of parameters, for all $a \neq b$, we have $\mathrm{TV}\left(f_a, f_b \right) = \Omega\left(\min\{1, \lambda \sqrt{d/r}\} \right) = \Omega(\varepsilon)$, completing the proof. □

The main technical lemma underlying our lower bound is Lemma 6.4.

LEMMA 6.4. *Suppose $d \geq r \geq 9$. There exists $M = 2^{\Omega(d^2/r^2)}$ such that the following holds. Construct M independent $d \times d/r$ matrices $\{U_a\}_{a=1}^M$ such that the columns of each U_a are the first d/r columns of a uniformly random $d \times d$ orthogonal matrix. Then, with positive probability, we have $\|U_a^\mathsf{T} U_b\|_F^2 \leq d/2r$ for all distinct pairs $a, b \in [M]$.*

For its proof, we will need a basic lemma about Gaussian matrices.

LEMMA 6.5. *Let G be a random matrix with i.i.d. Gaussian entries, and let $G = U \Sigma V^\mathsf{T}$ be its singular value decomposition. (To make the decomposition unique, we assume that $0 < \Sigma_{11} < \Sigma_{22} < \cdots$.) Then the matrices U, Σ, V are mutually independent. Moreover, U is a uniformly random orthogonal matrix.*

Proof. For any orthogonal matrix A, by rotational invariance of the Gaussian distribution (see, e.g., Reference [47, Proposition 3.3.2]), we have $AG \stackrel{d}{=} G$. Thus, letting W, X be size-compatible uniformly random orthogonal matrices (i.e., having Haar measure on the set of orthogonal matrices of appropriate sizes), independent of each other and of G, we have $G \stackrel{d}{=} WGX^\mathsf{T} = (WU)\Sigma(XV)^\mathsf{T}$, where WU, Σ, and XV are mutually independent, and WU is a uniformly random orthogonal matrix, as required. (We have used the fact that, by the definition of Haar measure, e.g., Reference [36, Section 1.2], if W, X are independent uniformly random orthogonal matrices and U, V are orthogonal matrices with compatible dimensions such that the products WU and XV are well-defined, then WU and XV are also independent uniformly random orthogonal matrices.)   □

Proof of Lemma 6.4. We will show that, for any two matrices U_a and U_b constructed independently as described, with probability $1 - O(\exp(-\Omega(d^2/r^2)))$, we have $\|U_a^\mathsf{T} U_b\|_F^2 \le d/2r$. The lemma then follows from the union bound.

Let $s = d/r$ and fix $a, b \in [M]$ with $a \ne b$. By the rotational invariance of the distribution of U_b, we may assume without loss of generality that $U_a = \begin{bmatrix} I \\ 0 \end{bmatrix}$. Thus, $\|U_a^\mathsf{T} U_b\|_F^2 \stackrel{d}{=} \|U^{(s)}\|_F^2$, where $U^{(s)}$ is an $s \times s$ principal submatrix of a uniformly random $d \times d$ orthogonal matrix U. (Alternatively, the columns of $U^{(s)}$ are the first s coordinates of s orthonormal vectors in $\mathbb{R}^d$ chosen uniformly at random.) Hence, it suffices to show that $\|U^{(s)}\|_F^2 \le d/2r$ with probability at least $1 - O(\exp(-\Omega(d^2/r^2)))$. The main difficulty is that $U^{(s)}$ does not have independent entries due to the orthonormality, but intuitively it should behave similarly to a matrix with independent Gaussian entries.

Before proceeding, we review some useful facts about Gaussian matrices. Let $G \in \mathbb{R}^{d \times s}$ be a Gaussian matrix with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Let $G = U_G \Sigma_G V_G^\mathsf{T}$ be its singular value decomposition, where $U_G \in \mathbb{R}^{d \times s}$, $\Sigma_G, V_G \in \mathbb{R}^{s \times s}$, and the diagonal entries of Σ_G are sorted ascendingly. By Lemma 6.5, U_G is a uniformly random orthogonal matrix, so the top s rows of U_G, denoted by $U_G^{(s)}$, has the same distribution as $U^{(s)}$. Let $G^{(s)}$ denote the top s rows of G; then, $G^{(s)} = U_G^{(s)} \Sigma_G V_G^\mathsf{T}$. Moreover, by Lemma 6.5 again, U_G is independent of Σ_G, V_G.

Let $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and largest singular values of a matrix, respectively. Note that $\sigma_{\min}(\Sigma_G) = \sigma_{\min}(G)$. Theorem 2.13 in Reference [15] states that, for any $t > 0$,

$$\Pr\left[\sigma_{\min}(\Sigma_G) \le 1 - 1/\sqrt{r} - t\right] \le \exp(-t^2 d). \tag{20}$$

Finally, we will use that for any two size-compatible matrices A, B, we have

$$\max\{\sigma_{\min}(A)\|B\|_F, \sigma_{\min}(B)\|A\|_F\} \le \|AB\|_F \le \min\{\sigma_{\max}(A)\|B\|_F, \sigma_{\max}(B)\|A\|_F\}, \tag{21}$$

see, e.g., Reference [25, Fact 7(c) in Section 24.4].

We now proceed with the rest of the proof. Recall that our goal is to show that $\|U^{(s)}\|_F^2 \le d/2r$ with probability at least $1 - O(\exp(-\Omega(d^2/r^2)))$, where $U^{(s)}$ is the $s \times s$ principal submatrix of a uniformly random orthogonal $d \times d$ matrix U.

The matrix U is naturally related to G. Similarly, the matrix $U^{(s)}$ is naturally related to the Gaussian matrix $G^{(s)} \in \mathbb{R}^{s \times s}$. More precisely, since U_G is independent of Σ_G, V_G, we have

$$G^{(s)} = U_G^{(s)} \Sigma_G V_G^\mathsf{T} \stackrel{d}{=} U^{(s)} \Sigma_G V_G^\mathsf{T}. \tag{22}$$

Observe that $\mathbf{E}\|G^{(s)}\|_F^2 = s^2/d = d/r^2$, so our goal is to show that $\|U^{(s)}\|_F^2$ is unlikely to exceed this by a multiplicative factor of $r/2$.

By the definition of the singular value decomposition, V_G is orthogonal, so all its singular values are 1, hence, using Inequality (21), we get $\|U^{(s)} \Sigma_G V_G^\mathsf{T}\|_F = \|U^{(s)} \Sigma_G^\mathsf{T}\|_F$. The Frobenius norms

$\|G^{(s)}\|_F$ and $\|U^{(s)}\|_F$ can be related as

$$\|G^{(s)}\|_F \overset{d}{=} \left\|U^{(s)}\Sigma_G V_G^\mathsf{T}\right\|_F = \|U^{(s)}\Sigma_G\|_F \ge \sigma_{\min}(\Sigma_G)\|U^{(s)}\|_F,$$

the first equality is by (22) and the inequality follows from (21).

Since Σ_G and $U^{(s)}$ are independent, for any real α and β, we have

$$\mathbf{Pr}\left[\|G^{(s)}\|_F^2 \ge \alpha\beta\right] \ge \mathbf{Pr}\left[\sigma_{\min}(\Sigma_G)^2 \ge \alpha \text{ and } \|U^{(s)}\|_F^2 \ge \beta\right]$$
$$= \mathbf{Pr}\left[\sigma_{\min}(\Sigma_G) \ge \sqrt{\alpha}\right]\mathbf{Pr}\left[\|U^{(s)}\|_F^2 \ge \beta\right].$$

In particular, setting $\alpha = 1/3$ and $\beta = d/2r$ gives

$$\mathbf{Pr}\left[\|U^{(s)}\|_F^2 \ge d/2r\right] \le \frac{\mathbf{Pr}\left[\|G^{(s)}\|_F^2 \ge d/6r\right]}{\mathbf{Pr}\left[\sigma_{\min}(\Sigma_G) \ge 1/\sqrt{3}\right]}.$$

For the numerator, since $d\|G^{(s)}\|_F^2 \overset{d}{=} \chi_{s^2}$, we have

$$\mathbf{Pr}\left[\|G^{(s)}\|_F^2 \ge d/6r\right] = \Pr[\chi_{s^2} \ge d^2/6r] \le \Pr[\chi_{s^2} \ge 3s^2/2] \le \exp(-s^2/25) = \exp(-\Omega(d^2/r^2)),$$

where we have used $r \ge 9$ for the first inequality and Lemma 2.2 for the second inequality. For the denominator, setting $t = 1 - 1/\sqrt{r} - 1/\sqrt{3} > 0.08$ in Inequality (20) gives

$$\mathbf{Pr}\left[\sigma_{\min}(\Sigma_G) \ge 1/\sqrt{3}\right] \ge 1 - \exp(-t^2 d) > 0.006,$$

hence $\mathbf{Pr}[\|U^{(s)}\|_F^2 \ge d/2r] = O(\exp(-\Omega(d^2/r^2)))$, completing the proof. □

LEMMA 6.6. *Suppose that $\lambda \le 1/4$. If U_a and U_b are $d \times d/r$ matrices satisfying $\|U_a^\mathsf{T}U_b\|_F^2 \le d/2r$, then $\mathrm{TV}(\mathcal{N}(0, I_d + \lambda U_a U_a^\mathsf{T}), \mathcal{N}(0, I_d + \lambda U_b U_b^\mathsf{T})) = \Omega(\min\{1, \lambda\sqrt{d/r}\})$.*

PROOF. We will use the following approximate characterization of the TV distance between two zero-mean Gaussians: For positive definite matrices Σ_a and Σ_b,

$$\mathrm{TV}\left(\mathcal{N}(0, \Sigma_a), \mathcal{N}(0, \Sigma_b)\right) = \Theta\left(\min\left\{1, \left\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\right\|_F\right\}\right).$$

This result appears in Reference [18, Theorem 1.1]; see also Reference [7, Corollary 2]. Hence, to complete the proof it suffices to show that $\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F \ge \frac{4}{5}\lambda\sqrt{d/r}$. Since $\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I = \Sigma_a^{-1/2}(\Sigma_b - \Sigma_a)\Sigma_a^{-1/2}$, applying the left inequality in (21) twice gives

$$\left\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\right\|_F \ge \sigma_{\min}\left(\Sigma_a^{-1/2}\right)^2\|\Sigma_b - \Sigma_a\|_F.$$

Since the eigenvalues of Σ_a are 1 and $1 + \lambda$, we have $\sigma_{\min}(\Sigma_a^{-1/2}) = (1 + \lambda)^{-1/2} \ge \sqrt{4/5}$ as $\lambda \le 1/4$. Moreover, $\|\Sigma_b - \Sigma_a\|_F = \lambda\|U_b U_b^\mathsf{T} - U_a U_a^\mathsf{T}\|_F$, and since $U_b U_b^\mathsf{T} - U_a U_a^\mathsf{T}$ is symmetric, we have

$$\left\|U_b U_b^\mathsf{T} - U_a U_a^\mathsf{T}\right\|_F^2 = \mathrm{Tr}\left(\left(U_b U_b^\mathsf{T} - U_a U_a^\mathsf{T}\right)\left(U_b U_b^\mathsf{T} - U_a U_a^\mathsf{T}\right)\right)$$
$$= \mathrm{Tr}\left(U_b U_b^\mathsf{T} U_b U_b^\mathsf{T}\right) + \mathrm{Tr}\left(U_a U_a^\mathsf{T} U_a U_a^\mathsf{T}\right) - \mathrm{Tr}\left(U_b U_b^\mathsf{T} U_a U_a^\mathsf{T}\right) - \mathrm{Tr}\left(U_a U_a^\mathsf{T} U_b U_b^\mathsf{T}\right)$$
$$= \mathrm{Tr}\left(U_b U_b^\mathsf{T}\right) + \mathrm{Tr}\left(U_a U_a^\mathsf{T}\right) - \mathrm{Tr}\left(U_b^\mathsf{T} U_a U_a^\mathsf{T} U_b\right) - \mathrm{Tr}\left(U_a^\mathsf{T} U_b U_b^\mathsf{T} U_a\right)$$
$$= d/r + d/r - \left\|U_a^\mathsf{T} U_b\right\|_F^2 - \left\|U_a^\mathsf{T} U_b\right\|_F^2$$
$$\ge 2d/r - 2d/2r = d/r,$$

hence $\|\Sigma_a^{-1/2}\Sigma_b\Sigma_a^{-1/2} - I\|_F \ge \frac{4}{5}\lambda\sqrt{d/r}$, completing the proof of the lemma. □

6.2 The Lower Bound for Learning Gaussian Mixtures

For proving our lower bound for mixtures, Theorem 1.6, we will need a standard result. This lemma follows from the Gilbert-Varshamov bound in coding theory (see, e.g., Reference [32, Theorem 5.2.6]); we include a proof for completeness.

LEMMA 6.7. *Let $T \geq 4$ and $k \in \mathbb{N}$. There exists a set of k-tuples $\mathcal{X} \subseteq [T]^k$ such that $|\mathcal{X}| \geq T^{k/4}$ and every distinct $x, y \in \mathcal{X}$ differ in at least $k/4$ components.*

PROOF. For any $x \in [T]^k$, the number of k-tuples in $[T]^k$ that differ with x in at most $k/4$ components is bounded by

$$\sum_{i=0}^{k/4} \binom{k}{i}(T-1)^i < T^{k/4} \sum_{i=0}^{k/4} \binom{k}{i} \leq (4eT)^{k/4} < T^{3k/4},$$

where we have used the inequality $\sum_{i=0}^m \binom{n}{i} \leq (en/m)^m$, valid for all $1 \leq m \leq n$ (see, e.g., Reference [47, Exercise 0.0.5]). We give an iterative algorithm to build $\mathcal{X}$: Choose an arbitrary k-tuple from $[T]^k$, put it in $\mathcal{X}$, remove from $[T]^k$ the k-tuples that differ with the chosen k-tuple in at most $k/4$ components, and repeat. By the above calculation, the size of the final set $\mathcal{X}$ will be at least $T^{k/4}$. □

Theorem 1.6 follows immediately from the following result.

THEOREM 6.8. *Any algorithm that learns the class of mixtures of k Gaussians in $\mathbb{R}^d$ in the realizable setting within total variation distance ε and with success probability at least $2/3$ has sample complexity $\Omega\left(\frac{kd^2}{\varepsilon^2 \log(1/\varepsilon)}\right)$.*

PROOF. Recall that $\mathcal{G}_{d,k}$ denotes the class of k-mixtures of d-dimensional Gaussian distributions. As we will use Lemma 6.1 again to obtain the sample complexity lower bound, it suffices to construct $2^{\Omega(kd^2)}$ distributions in $\mathcal{G}_{d,k}$ with pairwise KL divergences $O(\varepsilon^2)$ and pairwise TV distances $\Omega(\varepsilon)$. Our family of distributions will use the covariance matrices constructed in Theorem 6.3. Some care is required to ensure that the TV distances are large, and we will adopt some ideas used in earlier work for mixtures of spherical Gaussians [44, Appendix C.2].

The proof of Theorem 6.3 shows that there exists a family of symmetric positive definite matrices $\Sigma_1, \ldots, \Sigma_T$ with $T = 2^{\Omega(d^2)}$ satisfying

$$\mathrm{KL}\left(\mathcal{N}(0, \Sigma_i) \parallel \mathcal{N}(0, \Sigma_j)\right) \leq O(\varepsilon^2) \qquad \forall i \neq j \tag{23a}$$

$$\mathrm{TV}\left(\mathcal{N}(0, \Sigma_i), \mathcal{N}(0, \Sigma_j)\right) \geq \Omega(\varepsilon) \qquad \forall i \neq j \tag{23b}$$

$$\|\Sigma_i\|_s \leq 2 \qquad \forall i. \tag{23c}$$

Next, we will create a family of distributions in $\mathcal{G}_{d,k}$, where each Gaussian in each mixture uses one of these Σ_i as its covariance matrix. However, there is a tension. On the one hand, we want any two of these mixture distributions to use disjoint sets of covariance matrices so the TV distance between the mixtures is large. On the other hand, this constraint reduces the number of mixture distributions we can create, while we want many distributions to maximize the lower bound. This tension is resolved by a compromise obtained via error-correcting codes (Lemma 6.7).

The construction proceeds as follows: First, we pick $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, which will serve as the means for the Gaussians. We choose them to be far apart: For some Δ, to be chosen later, we pick them in such a way that $\|\mu_i - \mu_j\|_2 \geq \Delta$ for all $i \neq j$. Each mixture distribution will be a uniform mixture of k Gaussians, for which the ith Gaussian has mean μ_i. The choice of covariance matrices

is determined using the error-correcting code. Specifically, let $X \subset [T]^k$ be a set as in Lemma 6.7 above. The family of mixture distributions is

$$\mathcal{F} := \{ f_x : x \in X \}, \qquad \text{where} \qquad f_x := \frac{1}{k}\Big(\mathcal{N}(\mu_1, \Sigma_{x_1}) + \cdots + \mathcal{N}(\mu_k, \Sigma_{x_k})\Big).$$

As desired, we have $|\mathcal{F}| = T^{\Omega(k)} = 2^{\Omega(kd^2)}$.

To analyze $\mathcal{F}$, the first task is to prove the pairwise KL divergence upper bound. This is straightforward. Fix distinct $x, y \in X$. For each i, Inequality (23a) shows that

$$\mathrm{KL}\Big(\mathcal{N}(\mu_i, \Sigma_{x_i}) \,\|\, \mathcal{N}(\mu_i, \Sigma_{y_i})\Big) = \mathrm{KL}\Big(\mathcal{N}(0, \Sigma_{x_i}) \,\|\, \mathcal{N}(0, \Sigma_{y_i})\Big) \leq O(\varepsilon^2).$$

The convexity of KL divergence [12, Theorem 2.7.2] then shows that $\mathrm{KL}(f_x \,\|\, f_y) \leq O(\varepsilon^2)$.

The remaining task is to prove $\mathrm{TV}(f_x, f_y) \geq \Omega(\varepsilon)$ for all distinct $f_x, f_y \in \mathcal{F}$. The intuition is as follows: Say that index $i \in [k]$ *disagrees* if $x_i \neq y_i$. Whenever i disagrees, the ith Gaussian in f_x and ith Gaussian in f_y have TV distance $\Omega(\varepsilon)$ by Inequality (23b). Moreover, the total mixture weight apportioned to disagreeing indices is at least $1/4$, since Lemma 6.7 ensures that the number of disagreements is at least $k/4$, and each mixture uses uniform weights on its components. Thus, the disagreeing coordinates should suffice to imply that the TV distance is $\Omega(\varepsilon)$. Proving this formally requires some care, because each Gaussian is supported on all of $\mathbb{R}^d$, so there is interaction between all Gaussians involved in the mixtures. However, choosing a large enough Δ ensures that the means are far apart, so the interaction is negligible.

Formally, for each $j \in [k]$, let $A'_j \subseteq \mathbb{R}^d$ be such that

$$\mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{x_j})}[g \in A'_j] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{y_j})}[g \in A'_j] = \mathrm{TV}\Big(\mathcal{N}(\mu_j, \Sigma_{x_j}), \mathcal{N}(\mu_j, \Sigma_{y_j})\Big), \qquad (24)$$

and define

$$A_j = A'_j \cap B_j, \qquad \text{where} \qquad B_j = \Big\{ x \in \mathbb{R}^d : \|x - \mu_j\|_2 < \Delta/2 \Big\}.$$

Note that the separation of $\mu_1, \ldots, \mu_k$ implies that the balls $B_1, \ldots, B_k$ are disjoint. Consequently, the sets $A_1, \ldots, A_k$ are also disjoint.

Several preliminary inequalities are required concerning these events. First, for each $i \in [k]$,

$$\begin{aligned}
\mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \notin B_i] &= \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}\Big[\|g - \mu_i\|_2^2 \geq (\Delta/2)^2\Big] \\
&= \mathbf{Pr}_{g \sim \mathcal{N}(0, \Sigma_{x_i})}\Big[\|g\|_2^2 \geq (\Delta/2)^2\Big] && \text{(translating to zero-mean)} \\
&\leq \mathbf{Pr}_{g \sim \mathcal{N}(0, I_d)}\Big[\|g\|_2^2 \geq \Delta^2/8\Big] && \text{(by Inequality (23c))} \\
&\leq \varepsilon^2/k^2, && (25)
\end{aligned}$$

by applying Lemma 2.2 with $t = 2\ln(k/\varepsilon)$ and choosing Δ to satisfy $\Delta^2/8 = d + 2\sqrt{dt} + 2t$. Inequality (25) also holds replacing x_i with y_i. Since $A'_i \setminus A_i \subseteq B_i^c$, Inequality (25) shows that

$$\left| \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \in A_i] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \in A'_i] \right| \leq \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \notin B_i] \leq \varepsilon^2/k^2. \qquad (26)$$

This inequality also holds using y_i instead of x_i. For $i \neq j$, we have $A_j \subseteq B_i^c$, so

$$\mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{y_i})}[g \in A_j] \leq \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{y_i})}[g \notin B_i] \leq \varepsilon^2/k^2. \qquad (27)$$

Finally, by Equation (24), Inequality (26), and the triangle inequality,

$$\mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{x_j})}[g \in A_j] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{y_j})}[g \in A_j] \geq \mathrm{TV}\Big(\mathcal{N}(\mu_j, \Sigma_{x_j}), \mathcal{N}(\mu_j, \Sigma_{y_j})\Big) - 2\varepsilon^2/k^2. \qquad (28)$$

The total variation distance is lower bounded as follows: Let $A := A_1 \cup \cdots \cup A_k$. Then

$$
\begin{aligned}
&\mathrm{TV}(f_x, f_y) \\
&\geq \mathbf{Pr}_{g \sim f_x}[g \in A] - \mathbf{Pr}_{g \sim f_y}[g \in A] \\
&= \sum_{j=1}^{k} \left(\mathbf{Pr}_{g \sim f_x}[g \in A_j] - \mathbf{Pr}_{g \sim f_y}[g \in A_j] \right) && \text{(by disjointness of the } A_j) \\
&= \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{k} \left(\mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \in A_j] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{y_i})}[g \in A_j] \right) && \text{(expanding } f_x \text{ and } f_y) \\
&= \frac{1}{k} \sum_{j=1}^{k} \left(\mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{x_j})}[g \in A_j] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{y_j})}[g \in A_j] \right) && \text{(summands with } i = j) \\
&\quad + \frac{1}{k} \sum_{j=1}^{k} \sum_{i \neq j} \Big(\underbrace{\mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{x_i})}[g \in A_j]}_{\geq 0} - \underbrace{\mathbf{Pr}_{g \sim \mathcal{N}(\mu_i, \Sigma_{y_i})}[g \in A_j]}_{\leq \varepsilon^2/k^2 \text{ by (27)}} \Big) && \text{(summands with } i \neq j) \\
&\geq \frac{1}{k} \sum_{j=1}^{k} \left(\mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{x_j})}[g \in A_j] - \mathbf{Pr}_{g \sim \mathcal{N}(\mu_j, \Sigma_{y_j})}[g \in A_j] \right) - \varepsilon^2 \\
&\geq \frac{1}{k} \sum_{j=1}^{k} \left(\mathrm{TV}\left(\mathcal{N}(\mu_j, \Sigma_{x_j}), \mathcal{N}(\mu_j, \Sigma_{y_j}) \right) - 2\varepsilon^2/k^2 \right) - \varepsilon^2 && \text{(by (28))} \\
&\geq \frac{1}{k}(k/4)\Omega(\varepsilon) - 3\varepsilon^2 \ = \ \Omega(\varepsilon),
\end{aligned}
$$

where the last inequality is because $\mathrm{TV}(\mathcal{N}(\mu_j, \Sigma_{x_j}), \mathcal{N}(\mu_j, \Sigma_{y_j})) \geq \Omega(\varepsilon)$ whenever $x_j \neq y_j$ (see (23b)), which is the case for at least $k/4$ of the indices j. □

7 DISCUSSION AND OPEN PROBLEMS

We have built a connection between distribution learning and compression. Another concept related to compression is that of *core-sets*. The idea of core-sets is to summarize the training data, using a small subset of them, in a way that *any algorithm* minimizing the error on the subset will have small error on the whole set. Core-sets have been used in maximum likelihood estimation [35] and to solve clustering problems [43]. There is one important distinction between compression and core-sets: Compression can be more powerful, since it can use complex class-specific decoders.

Our work opens several avenues for further research.

Eliminating the polylogarithmic factors. Our sample complexity lower and upper bounds (in both axis-aligned and general cases) differ by multiplicative polylogarithmic factors. Can one remove these factors? In this direction, we propose the following conjecture. (See page 5 for the relevant definitions.)

CONJECTURE 7.1. *The minimax estimation rate for $\mathcal{G}_{d,k}$ and $\mathcal{A}_{d,k}$ is $\Theta(\sqrt{kd^2/n})$ and $\Theta(\sqrt{kd/n})$, respectively.*

Note that this conjecture is true for the case $k = 1$ (see Reference [19]). For general k, the lower bound for $\mathcal{G}_{d,k}$ follows from the proof of Theorem 1.6 (see Theorem 6.8), and the lower bound for $\mathcal{A}_{d,k}$ follows from the lower bound proof in Reference [44]. The upper bounds hold up to polylogarithmic factors (see Theorem 1.5 and Theorem 1.7).

Learning mixtures of sparse Gaussians. Our main results imply that the sample complexity for learning mixtures of *axis-aligned* Gaussians is smaller than that of mixtures of general Gaussians by a factor of $\widetilde{\Theta}(d)$. Can one interpolate between these two extremes by exploiting some notion of sparsity of the target distribution?

Consider the class of d-dimensional Gaussians whose inverse covariance matrices have at most m off-diagonal nonzero entries. Note that m measures the amount of correlation between the Gaussian components: If we build a probabilistic graphical model (also known as a Markov random field) whose nodes are the Gaussian components, then an axis-aligned Gaussian corresponds to an empty graph with no correlation between the components, in which case $m = 0$, and a Gaussian with fully correlated components corresponds to the complete graph, in which case $m = \binom{d}{2}$. In general, m counts the number of edges in this graph (see Reference [31, Proposition 5.2] for a proof).

The sample complexity of learning with respect to the class of d-dimensional Gaussians whose inverse covariance matrices have at most m off-diagonal nonzero entries is $\widetilde{\Theta}((m + d)/\varepsilon^2)$ (see Reference [19]). This result on learning a single Gaussian and the fact that in some applications the underlying Gaussians are sparse motivates the following question: Can one extend the bound of Reference [19] to mixtures of Gaussians, obtaining sample complexity bounds that depend on some notion of sparsity of the mixture components?

Polynomial time algorithms for learning mixtures of Gaussians. The running time of our density estimation algorithm is $2^{kd^2 \, \mathrm{polylog}(d,k,1/\varepsilon,1/\delta)}$, which is not polynomial in the problem parameters. An important open question is whether there exists an algorithm for learning mixtures of Gaussians that runs in time $\mathrm{poly}(k, d, 1/\varepsilon, 1/\delta)$ (see also Reference [20, Open Problem 15.5]). If the covariance matrices for all the Gaussians are multiples of the identity matrix (known as spherical Gaussians), Reference [44] gives an algorithm with running time that is polynomial in d and $1/\varepsilon$ but exponential in k. However, for mixtures of general Gaussians, it is shown in Reference [23] that no polynomial time (in all the parameters) algorithm exists for the case that the learner has access to the distribution only via *statistical queries*. (See Reference [23] for the definition of this model.)

What if k is not known? Our density estimation algorithms assume that k is given as input, while in some applications k might be unknown. One approach is to perform a binary search: Run our algorithm for $k = 1, 2, 4, 8, \ldots$, and stop as soon as the output of our algorithm has total variation distance less than ε from the target distribution. Unfortunately, it is not clear how to approximate this total variation distance. It is also not clear how to apply this approach to the robust learning scenario.

Is robust compression closed under taking mixtures? Lemma 4.8 states that for any distribution class $\mathcal{F}$, non-robust compression of $\mathcal{F}$ implies non-robust compression of k-mix($\mathcal{F}$). Does an analogous statement hold for robust compression? That is, does robust compression of $\mathcal{F}$ imply robust compression of k-mix($\mathcal{F}$) for a general class $\mathcal{F}$?

Sample complexity for learning with respect to the KL divergence. Theorem 3.1 states that there does not exist a function $g(k, d, \varepsilon)$ such that there exists an algorithm that upon receiving $g(k, d, \varepsilon)$ i.i.d. samples from an unknown $f \in \mathcal{G}_{d,k}$, outputs $\widehat{f}$ such that $\mathrm{KL}(f \parallel \widehat{f}) \leq \varepsilon$ with probability more than 0.02. Recalling that the KL divergence is asymmetric, we pose the following question: What is the smallest function $g(k, d, \varepsilon, \delta)$ such that there exists an algorithm that, upon receiving $g(k, d, \varepsilon, \delta)$ i.i.d. samples from an unknown $f \in \mathcal{G}_{d,k}$, outputs $\widehat{f}$ such that $\mathrm{KL}(\widehat{f} \parallel f) \leq \varepsilon$ with probability at least $1 - \delta$?

Characterizing the sample complexity of learning a class of distributions. A central open problem in distribution learning and density estimation is characterizing the sample complexity of learning a distribution class (Reference [20, Open Problem 15.1]). An insight from supervised learning theory is that the sample complexity of learning a class (of concepts, functions, or distributions) is typically proportional, up to logarithmic factors, to (some notion of) intrinsic dimension of that class divided by ε^2, where ε is the error tolerance. For the case of binary classification, the intrinsic dimension is captured by the VC dimension of the concept class (see References [9, 46]). For the case of distribution learning with respect to "natural" parametric classes, we expect this dimension to be equal to the number of parameters. This is indeed true for the class of Gaussians (which have d^2 parameters) and axis-aligned Gaussians (which have d parameters), and we showed in this article that it holds for their mixtures as well (which have kd^2 and kd parameters, respectively).

One may wonder if the VC dimension of the Yatracos family associated with a class of distributions can characterize its sample complexity, but it is not hard to come up with examples where this VC dimension is infinite while the class can be learned with finite samples. Covering numbers do not characterize the sample complexity either: For instance, the class of Gaussians does not have a finite covering number in the TV metric, nevertheless it is learnable with finitely many samples. Thus, we leave characterizing the sample complexity of learning a class of distributions as an important open problem.

Do learnable classes have bounded compression schemes? In binary classification, the combinatorial notion of Littlestone-Warmuth compression has been shown to be sufficient [33] and necessary [39] for learning. In this work, we showed that the new but related notion of distribution compression is sufficient for distribution learning (Theorem 4.5). Whether the existence of compression schemes is necessary for learning a class of distributions remains an intriguing open problem. In this direction, we conjecture the following converse for Theorem 4.5: Let $m_{\mathcal{F}}(\varepsilon, \delta)$ denote the sample complexity function associated with learning the class $\mathcal{F}$ of distributions (see Definition 1.2).

CONJECTURE 7.2. *There exists a universal constant C such that any class $\mathcal{F}$ admits an $(C\varepsilon^2 m(\varepsilon) \log(\varepsilon^2 m(\varepsilon))^C, C\varepsilon^2 m(\varepsilon) \log(\varepsilon^2 m(\varepsilon))^C, Cm(\varepsilon) \log(m(\varepsilon))^C)$ (non-robust) compression, where $m(\varepsilon) := m_{\mathcal{F}}(\varepsilon, \frac{1}{2})$.*

The value $\varepsilon^2 m(\varepsilon)$ is a candidate for the notion of "intrinsic dimension" of the class. We also propose the following weaker conjecture:

CONJECTURE 7.3. *There exist universal polynomials P, Q, and R such that any class $\mathcal{F}$ admits a $(P(\varepsilon^2 m(\varepsilon), \log(1/\varepsilon)), Q(\varepsilon^2 m(\varepsilon), \log(1/\varepsilon)), R(m(\varepsilon), 1/\varepsilon))$ (non-robust) compression, where $m(\varepsilon) := m_{\mathcal{F}}(\varepsilon, \frac{1}{2})$.*

APPENDICES

A   PROOFS OF THEOREM 3.1 AND THEOREM 3.2

In this section, let ν denote the Lebesgue measure on $\mathbb{R}$. We will first prove Theorem 3.1. To that end, we begin with a simple calculation that will be useful later.

LEMMA A.1. *Suppose $I \subseteq \mathbb{R}$ satisfies $\nu(I) \geq \gamma$. Moreover, let $f, h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be measurable density functions such that $f(x) \geq \beta$ and $h(x) \leq \alpha$ for all $x \in I$, and $f(x) > 0$ for all $x \in \mathbb{R}$. Then $\mathrm{KL}\,(f \parallel h) \geq \gamma\beta \log(\beta/\alpha) - 1/e$.*

PROOF. Write

$$\mathrm{KL}\,(f \parallel h) = \int_I f(x) \log \frac{f(x)}{h(x)} \,\mathrm{d}x + \int_{I^c} f(x) \log \frac{f(x)}{h(x)} \,\mathrm{d}x.$$

For the first integral, we have

$$\int_I f(x) \log \frac{f(x)}{h(x)} \, dx \geq \int_I \beta \log \frac{\beta}{\alpha} \, dx \geq \gamma \beta \log(\beta/\alpha).$$

Next, we bound the second integral and show that it has value at least $-1/e$, which completes the proof. Let $F = \int_{I^c} f(x) \, dx$ and $H = \int_{I^c} h(x) \, dx$. Note that $F \geq 0$ as $f(x) > 0$ for all $x \in \mathbb{R}$. If $H = 0$, then $h(x) = 0$ almost everywhere on I^c so the second integral is $+\infty$. So, assume that $H > 0$. Then f/F and h/H are densities on I^c. Hence, we have

$$\int_{I^c} f(x) \log \frac{f(x)}{h(x)} \, dx = \underbrace{F \int_{I^c} \frac{f(x)}{F} \log \frac{f(x)/F}{h(x)/H} \, dx}_{\geq 0} + F \int_{I^c} \frac{f(x)}{F} \log \frac{F}{H} \, dx \geq F \log(F/H),$$

where the inequality is because the KL divergence of two densities is always nonnegative. Since $H \leq 1$, we have $-\log(H) \geq 0$ so $F \log(F/H) = F \log F - F \log H \geq F \log F \geq -1/e$, as required. □

PROOF OF THEOREM 3.1. We allow the algorithm to be randomized. Denote by $\mathcal{A}(X_1, \ldots, X_m, R)$ the output of $\mathcal{A}$ given the sample $X_1, \ldots, X_m$ and an independent source of randomness R. We will first analyze the behavior of the algorithm when the true distribution is $\mathcal{N}(0, 1)$ and show that there exists some $a' \in \mathbb{R}$ for which the algorithm's output puts almost no probability mass on around a'. We then show that if the true distribution is a carefully chosen mixture of $\mathcal{N}(0, 1)$ and $\mathcal{N}(a', 1)$, then the algorithm's output does not change with high probability, so it still puts almost no mass on around a'; hence, the KL divergence of the output and the true distribution is large.

Define the parameters $\delta = \frac{0.01}{m}$, $\beta = \frac{\delta}{\sqrt{2\pi}} \exp(-1/32)$, and $\alpha = \beta \exp\left(\frac{-4\tau - 4/e}{\beta}\right)$. Let $X_1, \ldots, X_m \sim \mathcal{N}(0, 1)$ and set $h = \mathcal{A}(X_1, \ldots, X_m, R)$. Note that h is random. Define the (random) set $H = \{ x \in \mathbb{R} : h(x) \geq \alpha \}$. Then $\nu(H) \leq 1/\alpha$. For any $a \in \mathbb{Z}$, define $I_a = [a - 1/4, a + 1/4]$. Note that the I_a are disjoint intervals. Hence, $\sum_{a \in \mathbb{Z}} \nu(I_a \cap H) \leq 1/\alpha$ deterministically so $\mathbf{E}\left[\sum_{a \in \mathbb{Z}} \nu(I_a \cap H)\right] \leq 1/\alpha$. Note that the left-hand side of the inequality is an infinite sum while the right-hand side is a finite number. Since expectation is linear, we can find $a' \in Z$ such that $\mathbf{E}\left[\nu(I_{a'} \cap H)\right] \leq 1/400$. By Markov's Inequality, $\nu(I_{a'} \cap H) \leq 1/4$ with probability at least 0.99. We condition on this event.

Define $f = (1 - \delta)\mathcal{N}(0, 1) + \delta \cdot \mathcal{N}(a', 1)$ and note that f is positive everywhere, and, for all $x \in I_{a'}$, we have $f(x) \geq \frac{\delta}{\sqrt{2\pi}} \exp(-1/32) = \beta$. Let $J_{a'} = I_{a'} \setminus H$. Then $\nu(J_{a'}) \geq \nu(I_{a'}) - \nu(I_{a'} \cap H) \geq 1/4$, and for all $x \in J_{a'}$, we have $f(x) \geq \beta$ and $h(x) < \alpha$. So,

$$\mathrm{KL}\left(f \parallel h\right) \geq \beta \log(\beta/\alpha)/4 - 1/e = \tau,$$

where the inequality is by Lemma A.1 and the equality is by the definition of α. Hence, $\mathrm{KL}\left(f \parallel h\right) \geq \tau$ with probability at least 0.99.

Note that $\mathrm{TV}\left(f, \mathcal{N}(0, 1)\right) \leq \delta$. If $S = (X_1, \ldots, X_m)$ and $S' = (X_1', \ldots, X_m')$ where $X_i \sim \mathcal{N}(0, 1)$ and $X_i' \sim f$, then Proposition 4.7 gives $\mathrm{TV}\left(S, S'\right) \leq m\delta = 0.01$. Hence, if $h = \mathcal{A}(S, R)$ and $h' = \mathcal{A}(S', R)$, then $\mathrm{TV}\left(h, h'\right) \leq 0.01$ so $\mathbf{Pr}[\mathrm{KL}\left(f \parallel h'\right) \geq \tau] \geq \mathbf{Pr}[\mathrm{KL}\left(f \parallel h\right) \geq \tau] - 0.01 \geq 0.98$, completing the proof. □

Next, we prove Theorem 3.2.

PROOF OF THEOREM 3.2. Define the parameters $\delta = \frac{0.01}{m}$, $\sigma^{p-1} = \frac{\delta^p}{\tau^p 6^p} \sqrt{\ln(9/2\pi)}$, and $M = 4\sigma \sqrt{\ln(9/2\pi)}$.

Let $X_1, \ldots, X_m \sim \mathcal{N}(0, 1)$ and set $h = \mathcal{A}(X_1, \ldots, X_m, R)$, where, as in the proof of Theorem 3.1, R is the algorithm's independent source of randomness. Note that h is random. Next, let $H = \{ x \in \mathbb{R} : h(x) \geq \delta/6\sigma \}$. Then $\nu(H) \leq 6\sigma/\delta$. For $a \in \mathbb{Z}$, define the intervals $I_a = [aM - M/4, aM + M/4]$ and note that I_a are disjoint intervals. Hence, $\sum_{a \in \mathbb{Z}} \nu(I_a \cap H) \leq 6\sigma/\delta$ deterministically so

$\mathbf{E}\left[\sum_{a \in \mathbb{Z}} \nu(I_a \cap H)\right] \leq 6\sigma/\delta$. Note that the left-hand side of the inequality is an infinite sum while the right-hand side of the inequality is a finite number. Since expectation is linear, there exists $a' \in Z$ such that $\mathbf{E}\left[\nu(I_{a'} \cap H)\right] \leq M/400$. By Markov's Inequality, $\nu(I_{a'} \cap H) \leq M/4$ with probability at least 0.99. We condition on this event.

Define $f = (1 - \delta)\mathcal{N}(0, 1) + \delta \cdot \mathcal{N}(a', \sigma^2)$, and note that for all $x \in I_{a'}$, we have $f(x) \geq \delta \frac{1}{\sqrt{2\pi}\sigma} \exp(-(M/4)^2/2\sigma^2) = \delta/3\sigma$.

Let $J_{a'} = I_{a'} \setminus H$. Then $\nu(J_{a'}) \geq M/2 - M/4 = M/4 = \sigma\sqrt{\ln(9/2\pi)}$, and, for all $x \in J_{a'}$, we have $f(x) \geq \delta/3\sigma$ and $h(x) \leq \delta/6\sigma$. So,

$$\|f - h\|_p^p \geq \int_{J_{a'}} |f(x) - h(x)|^p \, \mathrm{d}x \geq \frac{\delta^p}{(6\sigma)^p}\sigma\sqrt{\ln(9/2\pi)} = \frac{\delta^p}{6^p\sigma^{p-1}}\sqrt{\ln(9/2\pi)} = \tau^p,$$

where the last equality is by the definition of σ. Hence, $\|f - h\|_p \geq \tau$ with probability at least 0.99.

Note that $\mathrm{TV}\,(f, \mathcal{N}(0, 1)) \leq \delta$. If $S = (X_1, \ldots, X_m)$ and $S' = (X_1', \ldots, X_m')$ where $X_i \sim \mathcal{N}(0, 1)$ and $X_i' \sim f$, then Proposition 4.7 gives $\mathrm{TV}\,(S, S') \leq m\delta = 0.01$. Hence, if $h = \mathcal{A}(S, R)$ and $h' = \mathcal{A}(S', R)$, then $\mathrm{TV}\,(h, h') \leq 0.01$, so $\mathbf{Pr}[\|f - h\|_p \geq \tau] \geq \mathbf{Pr}[\|f - h\|_p \geq \tau] - 0.01 \geq 0.98$. □

B PROOF OF LEMMA 4.10

We restate the lemma for convenience.

LEMMA B.1. *Let $f = \sum_{i \in [k]} w_i f_i$ be a density with $(w_1, \ldots, w_k) \in \Delta_k$ and each $f_i \in \mathcal{F}$. Let g be a density such that $\|g - f\|_1 \leq \rho$. Then, we can write $g = \sum_{i \in [k]} w_i g_i$ such that each g_i is a density and for any $r > 0$,*

$$\sum_{i \,:\, \|g_i - f_i\|_1 > r} w_i < \rho/r.$$

Let $X := \{x : g(x) < f(x)\}$. Our goal is to "transform" each f_i into another density g_i so that $g = \sum_{i \in [k]} w_i g_i$. Note that X consists of the domain points on which f exceeds g. Hence, to transform each f_i into g_i, we will scale it down multiplicatively on points in X, and scale it up additively on points not in X. These transformations need to be done carefully for each function g_i to end up being nonnegative and integrate to 1.

To that end, we define

$$g_i(x) := \begin{cases} f_i(x)g(x)/f(x) & \text{for } x \in X, \\ f_i(x) + \Delta_i(x) & \text{for } x \notin X, \end{cases}$$

where

$$\Delta_i(x) := \left(g(x) - f(x)\right)\left(\int_X f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \, \mathrm{d}y\right) \Bigg/ \int_X \left(f(y) - g(y)\right) \mathrm{d}y. \tag{29}$$

Recall that Z is the domain of g and the densities in $\mathcal{F}$. We now check that each g_i is a density and that $g = \sum_{i \in [k]} w_i g_i$.

LEMMA B.2. *For all $i \in [k]$, g_i is a density on Z.*

PROOF. We first check that $g_i(x) \geq 0$ for all x. If $x \in X$, then $g_i(x) \geq 0$, because f_i, g, f are all densities and hence nonnegative. If $x \notin X$, then $\Delta_i(x) \geq 0$, because $g(x) - f(x) \geq 0$ on X^c and $f(x) - g(x) \geq 0$ on X. We now check that $\int_Z g_i(x) \, \mathrm{d}x = 1$. Since both g and f are densities, both integrate to 1 over Z, and therefore

$$\int_{X^c} (g(x) - f(x)) \, \mathrm{d}x = \int_X (f(x) - g(x)) \, \mathrm{d}x. \tag{30}$$

The following calculation completes the proof:

$$\int_{X^c} g_i(x)\, dx = \int_{X^c} (\Delta_i(x) + f_i(x))\, dx$$

$$= \frac{\int_{X^c} (g(x) - f(x))\, dx}{\int_X (f(y) - g(y))\, dy} \cdot \int_X \left(f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) dy \qquad \text{(by Equation (29))}$$

$$+ \int_{X^c} f_i(x)\, dx$$

$$= \int_X \left(f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) dy + \int_{X^c} f_i(x)\, dx \qquad \text{(by Equation (30))}$$

$$= \int_X f_i(y) \cdot \left(1 - \frac{g(y)}{f(y)} \right) dy + \int_{X^c} f_i(y)\, dy$$

$$= \int_X f_i(y)\, dy - \int_X f_i(y) \cdot \left(\frac{g(y)}{f(y)} \right) dy + \int_{X^c} f_i(y)\, dy$$

$$= 1 - \int_X f_i(y) \cdot \frac{g(y)}{f(y)}\, dy \qquad (f_i \text{ is a density})$$

$$= 1 - \int_X g_i(y)\, dy \qquad \qquad \square$$

LEMMA B.3. *We have* $g = \sum_{i \in [k]} w_i g_i$.

PROOF. For $x \in X$, since $\sum_{i \in [k]} w_i f_i = f$, we have

$$\sum_{i \in [k]} w_i g_i(x) = \sum_{i \in [k]} w_i f_i(x) \frac{g(x)}{f(x)} = g(x).$$

And for $x \notin X$, we have

$$\sum_{i \in [k]} w_i g_i(x) = \sum_{i \in [k]} w_i \Delta_i(x) + w_i f_i(x) \qquad \text{(definition of } g_i)$$

$$= \sum_{i \in [k]} w_i \left(\frac{(g(x) - f(x))}{\int_X (f(y) - g(y))\, dy} \cdot \int_X \left(f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) dy \right) \quad \text{(by Equation (29))}$$

$$+ \sum_{i \in [k]} w_i f_i(x)$$

$$= \frac{(g(x) - f(x))}{\int_X (f(y) - g(y))\, dy} \cdot \int_X \left(\sum_{i \in [k]} w_i f_i(y) \cdot \frac{f(y) - g(y)}{f(y)} \right) dy$$

$$+ f(x) \qquad \left(\sum_{i \in [k]} w_i f_i = f \right)$$

$$= \frac{(g(x) - f(x))}{\int_X (f(y) - g(y))\, dy} \cdot \int_X (f(y) - g(y))\, dy + f(x) \qquad \left(\sum_{i \in [k]} w_i f_i = f \right)$$

$$= g(x) - f(x) + f(x) = g(x). \qquad \square$$

Let $I := \{ i \in [k] : \|f_i - g_i\|_1 > r \}$. It remains to show that $\sum_{i \in I} w_i < \rho/r$. Observe from the definition of the g_i that we also have $X = \{ x : g_i(x) < f_i(x) \}$ for each $i \in [k]$. Thus, using

Lemma B.3,

$$\|f - g\|_1 \;=\; 2 \int_X \left(f(x) - g(x)\right) \mathrm{d}x \;=\; 2 \sum_{i \in [k]} w_i \int_X \left(f_i(x) - g_i(x)\right) \mathrm{d}x \;=\; \sum_{i \in [k]} w_i \|f_i - g_i\|_1.$$

Thus, from the hypothesis of the lemma,

$$\rho \;\geq\; \|f - g\|_1 \;=\; \sum_{i \in [k]} w_i \|f_i - g_i\|_1 \;\geq\; \sum_{i \in I} w_i \|f_i - g_i\|_1 \;>\; \sum_{i \in I} w_i r,$$

by definition of I. This gives $\sum_{i \in I} w_i < \rho/r$, as required.

ACKNOWLEDGMENTS

REFERENCES

[1] Jayadev Acharya, Ilias Diakonikolas, Chinmay Hegde, Jerry Zheng Li, and Ludwig Schmidt. 2015. Fast and near-optimal algorithms for approximating distributions by histograms. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS'15)*. Association for Computing Machinery, New York, NY, 249–263. DOI : https://doi.org/10.1145/2745754.2745772

[2] Martin Anthony and Peter L. Bartlett. 1999. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK. DOI : https://doi.org/10.1017/CBO9780511624216

[3] Sanjeev Arora and Ravi Kannan. 2005. Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Probab.* 15, 1A (02 2005), 69–92. DOI : https://doi.org/10.1214/105051604000000512

[4] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. 2018. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 3412–3421. Retrieved from https://papers.nips.cc/paper/7601-nearly-tight-sample-complexity-bounds-for-learning-mixtures-of-gaussians-via-sample-compression-schemes.

[5] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. 2020. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. *arxiv:cs.LG/1710.05209* (2020).

[6] Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. 2018. Sample-efficient learning of mixtures. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*. AAAI Publications, 2679–2686. Retrieved from https://arxiv.org/abs/1706.01596.

[7] S. S. Barsov and V. V. Ul'yanov. 1987. Estimates of the proximity of Gaussian measures. *Sov. Math., Dokl.* 34 (1987), 462–466.

[8] Mikhail Belkin and Kaushik Sinha. 2010. Polynomial learning of distribution families. In *Proceedings of the IEEE 51st Symposium on Foundations of Computer Science (FOCS'10)*. IEEE Computer Society, Washington, DC, 103–112. DOI : https://doi.org/10.1109/FOCS.2010.16

[9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* 36, 4 (Oct. 1989), 929–965. DOI : https://doi.org/10.1145/76359.76371

[10] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK. DOI : https://doi.org/10.1017/CBO9780511804441

[11] Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. 2014. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th ACM Symposium on Theory of Computing (STOC'14)*. ACM, New York, NY, 604–613. DOI : https://doi.org/10.1145/2591796.2591848

[12] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley-Interscience (John Wiley & Sons), Hoboken, NJ.

[13] Sanjoy Dasgupta. 1999. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, Los Alamitos, CA, 634–644. DOI: https://doi.org/10.1109/SFFCS.1999.814639

[14] Constantinos Daskalakis and Gautam Kamath. 2014. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Conference on Learning Theory (Proceedings of Machine Learning Research)*, Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (Eds.), Vol. 35. PMLR, Barcelona, Spain, 1183–1213. Retrieved from http://proceedings.mlr.press/v35/daskalakis14.html.

[15] Kenneth R. Davidson and Stanislaw J. Szarek. 2001. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces, Vol. I*. North-Holland, Amsterdam, 317–366. DOI: https://doi.org/10.1016/S1874-5849(01)80010-3

[16] Luc Devroye. 1987. *A Course in Density Estimation* (Progress in Probability and Statistics), Vol. 14. Birkhäuser Boston, Inc., Boston, MA.

[17] Luc Devroye and Gábor Lugosi. 2001. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York. DOI: https://doi.org/10.1007/978-1-4613-0125-7

[18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. 2018. The total variation distance between high-dimensional Gaussians. *arxiv:math.ST/1810.08693* (2018).

[19] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. 2020. The minimax learning rates of normal and Ising undirected graphical models. *Electron. J. Statist.* 14, 1 (2020), 2338–2361. DOI: https://doi.org/10.1214/20-EJS1721

[20] Ilias Diakonikolas. 2016. Learning structured distributions. In *Handbook of Big Data*. CRC Press, Boca Raton, FL, 267–283. Retrieved from http://www.iliasdiakonikolas.org/distribution-learning-survey.pdf.

[21] Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. 2017. Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6391–6401. Retrieved from http://papers.nips.cc/paper/7218-communication-efficient-distributed-learning-of-discrete-distributions.

[22] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2017. Learning multivariate log-concave distributions. In *Proceedings of the Conference on Learning Theory (Proceedings of Machine Learning Research)*, Satyen Kale and Ohad Shamir (Eds.), Vol. 65. PMLR, Amsterdam, Netherlands, 711–727. Retrieved from http://proceedings.mlr.press/v65/diakonikolas17a.html.

[23] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. 2017. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract). In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE Computer Society, Los Alamitos, CA, 73–84. DOI: https://doi.org/10.1109/FOCS.2017.16

[24] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. 2006. PAC learning axis-aligned mixtures of Gaussians with no separation assumption. In *Proceedings of the 19th Conference on Learning Theory (COLT'06)*. Springer-Verlag, Berlin, 20–34. DOI: https://doi.org/10.1007/11776420_5

[25] Leslie Hogben (Ed.). 2014. *Handbook of Linear Algebra* (2nd ed.). CRC Press, Boca Raton, FL.

[26] Ildar Ibragimov. 2001. Estimation of analytic functions. In *State of the Art in Probability and Statistics (Leiden, 1999)* (IMS Lecture Notes–Monograph Series), Vol. 36. Institute of Mathematical Statistics, Beachwood, OH, 359–383. DOI: https://doi.org/10.1214/lnms/1215090078

[27] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. 2012. Disentangling Gaussians. *Commun. ACM* 55, 2 (Feb. 2012), 113–120. DOI: https://doi.org/10.1145/2076450.2076474

[28] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. 1994. On the learnability of discrete distributions. In *Proceedings of the 26th ACM Symposium on Theory of Computing (STOC'94)*. ACM, New York, NY, 273–282. DOI: https://doi.org/10.1145/195058.195155

[29] S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.* 22 (1951), 79–86. DOI: https://doi.org/10.1214/aoms/1177729694

[30] B. Laurent and P. Massart. 2000. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* 28, 5 (2000), 1302–1338. DOI: https://doi.org/10.1214/aos/1015957395

[31] Steffen L. Lauritzen. 1996. *Graphical Models* (Oxford Statistical Science Series), Vol. 17. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

[32] San Ling and Chaoping Xing. 2004. *Coding Theory: A First Course*. Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/CBO9780511755279

[33] Nick Littlestone and Manfred Warmuth. 1986. *Relating Data Compression and Learnability*. Technical Report. University of California, Santa Cruz. Retrieved from https://users.soe.ucsc.edu/~manfred/pubs/T1.pdf.

[34] Alexander E. Litvak, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann. 2005. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.* 195, 2 (2005), 491–523. https://doi.org/10.1016/j.aim.2004.08.004

[35] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. 2018. Training Gaussian mixture models at scale via coresets. *J. Mach. Learn. Res.* 18, 160 (2018), 1–25. Retrieved from http://jmlr.org/papers/v18/15-506.html.

[36] Elizabeth S. Meckes. 2019. *The Random Matrix Theory of the Classical Compact Groups (Cambridge Tracts in Mathematics)*, Vol. 218. Cambridge University Press, Cambridge, UK. DOI:https://doi.org/10.1017/9781108303453.009

[37] Michael Mitzenmacher and Eli Upfal. 2017. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis* (2nd ed.). Cambridge University Press, Cambridge, UK.

[38] Ankur Moitra and Gregory Valiant. 2010. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the IEEE 51st Symposium on Foundations of Computer Science (FOCS'10)*. IEEE Computer Society, Washington, DC, 93–102. DOI:https://doi.org/10.1109/FOCS.2010.15

[39] Shay Moran and Amir Yehudayoff. 2016. Sample compression schemes for VC classes. *J. ACM* 63, 3 (2016). DOI:https://doi.org/10.1145/2890490

[40] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA. Retrieved from http://www.gaussianprocess.org/gpml/chapters/.

[41] Rolf-Dieter Reiss. 1989. *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4613-9620-8

[42] Bernard W. Silverman. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[43] Christian Sohler and David P. Woodruff. 2018. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS'18)*. IEEE Computer Society, Los Alamitos, CA, 802–813. DOI:https://doi.org/10.1109/FOCS.2018.00081

[44] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. 2014. Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1395–1403. Retrieved from http://papers.nips.cc/paper/5251-near-optimal-sample-estimators-for-spherical-gaussian-mixtures.

[45] Alexandre B. Tsybakov. 2009. *Introduction to Nonparametric Estimation*. Springer, New York. DOI:https://doi.org/10.1007/b13794

[46] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probab. Applic.* 16, 2 (1971), 264–280. DOI:https://doi.org/10.1137/1116025

[47] Roman Vershynin. 2018. *High-dimensional Probability: An Introduction with Applications in Data Science (Cambridge Series in Statistical and Probabilistic Mathematics)*, Vol. 47. Cambridge University Press, Cambridge, UK.

[48] Yannis G. Yatracos. 1985. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* 13, 2 (1985), 768–774. DOI:https://doi.org/10.1214/aos/1176349553

[49] Bin Yu. 1997. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*. Springer, New York, 423–435. Retrieved from https://www.stat.berkeley.edu/~binyu/ps/LeCam.pdf.