Learning mixtures of elliptical distributions

UROP+ Final Paper, Summer 2021
Kerri Lu
Mentor: Cole Franks
Project suggested by: Ankur Moitra

September 9, 2021

Abstract

In this paper we consider the problem of density estimation for mixtures of high-dimensional elliptical distributions. We prove that $\tilde{O}\left(\frac{k(d^2+\log M)}{\epsilon^2}\right)$ samples are required for learning the distribution of a mixture of k many d-dimensional elliptical distributions to total variation distance ε with density functions drawn from a known set of size M.

1 Introduction

Learning mixtures of probability distributions is an important problem in unsupervised machine learning. While there are known algorithms for learning mixtures of multivariate Gaussians in polynomial time and sample complexity, less is known about the time and sample complexity for learning the broader class of multivariate elliptical distributions, defined as follows.

Definition 1.1 (Elliptical distributions)**.** A d-dimensional elliptical distribution with center $\boldsymbol{\mu}$, shape matrix Σ, and radial density function $g : \mathbb{R} \to \mathbb{R}$ has density function $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = \frac{\Gamma(d/2)}{2\pi^{d/2}|\Sigma|^{1/2}} g((\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})).$$

Examples of elliptical distributions include multivariate Gaussians, multivariate t distributions, Cauchy distributions, and the multivariate logistic distribution. In particular, elliptical distributions include heavy-tailed distributions, whose tails decay at a slower than exponential rate.

Definition 1.2 (Mixtures)**.** A mixture of k distributions $D_1, D_2, ..., D_k$ can be expressed as $D = \sum_{i=1}^k w_i D_i$ where the mixing weights $w_1, w_2, ..., w_k$ sum to 1. To sample from D, we randomly choose distribution D_i with probability w_i, and then choose a sample $\boldsymbol{x} \sim D_i$.

We will denote by k-mix(F) the set of all possible mixture distributions $D = \sum_{i=1}^k w_i D_i$ where $D_1, D_2, ..., D_k$ are contained in a class F.

We will prove a sample complexity upper bound for density estimation for a mixture of high-dimensional elliptical distributions. Given samples from a mixture $D = \sum_{i=1}^k w_i D_i$ where D_i are elliptical distributions contained in a class F, our goal is to output a distribution that is close in total variation distance to D. Additionally, if we are given samples from a distribution Q that is close to D, our algorithm should output a distribution that is close in total variation distance to Q relative to k-mixtures of distributions in F.

Let $TV(f, g)$ denote the total variation distance between two distributions f and g. We formally define agnostic PAC-learning of a distribution as follows.

Definition 1.3 (Agnostic PAC-learning)**.** A *C-agnostic PAC-learner* for a class F with sample complexity $m_F^C(\varepsilon, \delta)$ is a function which takes as input $m_F^C(\varepsilon, \delta)$ i.i.d. samples from an arbitrary distribution g, and outputs a distribution $\hat{g} \in F$ such that

$$TV(\hat{g}, g) \leq C \cdot \inf_{f \in F} TV(f, g) + \varepsilon$$

with probability at least $1 - \delta$.

Note that if $g \in F$, the *realizable case*, then $\inf_{f \in F} TV(f, g) = TV(g, g) = 0$ so the above inequality becomes

$$TV(\hat{g}, g) \leq \varepsilon.$$

Note that our goal for this paper is only to prove a sample complexity bound for density estimation of the entire mixture. This is an easier task than estimating the weights and density of each component distribution separately. The benefit of this approach is that it does not require any separation conditions between the centers of the distributions, or conditions on the minimal weight or maximal variance for each component distribution. No sample complexity bound was previously known for learning mixtures of elliptical distributions. Our main result is the following theorem.

Theorem 1.4. *Let $F_d(\phi_1, \phi_2, \dots \phi_M)$ be the class of d-dimensional elliptical distributions with characteristic function contained in the finite set $\{\phi_1, \phi_2, \dots \phi_M\}$. Let $f_d(\phi_1), f_d(\phi_2), \dots f_d(\phi_M)$ be the corresponding radial density functions. Let R_1 and R_2 be finite constants (that may depend on d and the functions ϕ_i) such that for all $1 \leq i \leq M$,*

$$P_{x \sim E_d(\mathbf{0}, I_d, \phi_i)}(\|x\|_2^2 \geq R_1 d) \leq 0.025, \tag{1}$$

$$\text{and} \quad P_{x \sim E_1(0,1,\phi_i)}(|x| \leq R_2) \leq 0.05. \tag{2}$$

If $f_d(\phi_i)$ are monotonic decreasing for all d and all $1 \leq i \leq M$, then k-mix$(F_d(\phi_1, \phi_2, \dots \phi_M))$ can be 12-agnostic PAC learned with sample complexity

$$\tilde{O}\left(\frac{k(d^2 \log(R_1/R_2) + \log M)}{\varepsilon^2}\right).$$

The sample complexity is, up to a polylogarithmic factor, the same as the sample complexity for learning a mixture of Gaussians. Unfortunately, the time complexity for our algorithm is exponential.

To prove our main theorem, we rely on the sample compression method developed by Ashtiani et. al. in [2]. In particular, the authors showed that if any distribution in a class F can be encoded using a small number of samples and a short sequence of bits, then k-mix(F) can be learned in time linear in k and the number of samples and bits.

1.1 Related Work

There is a large body of work on learning mixtures of Gaussians. There has been some work in clustering mixtures of heavy-tailed distributions, though no time or sample complexity bounds have been proven for learning mixtures of elliptical distributions in general.

Learning mixtures of Gaussians. Spectral approaches for clustering mixtures of Gaussians [12, 1] relied on principal components analysis, projecting samples onto a lower-dimensional subspace and then clustering the samples using distance concentration. These methods require that the means

of the Gaussians are sufficiently well-separated. Parameter estimation methods seek to find the mean, covariance and weight of each individual component in the mixture. For example, [8] proves an algorithm for learning the mixture of two Gaussians using random projections to one dimension and applying the method of moments to learn the mixture parameters, using polynomial time and polynomial samples. Other algorithms for learning the component Gaussians include the k-means algorithm [11, 3] and expectation maximization [13], but they are not guaranteed to converge at a global optimum. Most relevant to our approach, [2] used robust sample compression to show a sample complexity upper bound of $\tilde{O}(kd^2/\epsilon^2)$ for learning mixtures of k distinct d-dimensional Gaussians, as well as a matching lower bound (up to a polylogarithmic factor).

Learning mixtures of heavy-tailed distributions. [5] showed a sample complexity upper bound of $\tilde{O}(dk)$ for clustering mixtures of heavy-tailed symmetric distributions with independent coordinates under separation conditions between the distribution centers. [4] gave a polynomial-time, polynomial sample complexity algorithm for clustering mixtures of heavy-tailed product distributions. However, elliptical distributions do not have independent coordinates in general. In fact, the only elliptical distributions that are also product distributions are Gaussians.

Another clustering approach uses list decodable mean estimation [7], which views samples from a mixture as samples from a single distribution with a large fraction of the points contaminated. A polynomial time algorithm outputs a list of the possible means for the contaminated distribution and clusters the samples using the means. This approach requires bounded mean and covariance for each mixture component.

There are also iterative reweighting algorithms (related to expectation maximization) that optimize a non-convex potential to find mixture parameters [10]. Recently, [9] proposed using gradient descent on Riemannian manifolds to optimize the parameter estimates. There have been empirical demonstrations for these algorithms, but no sample complexity or convergence guarantees.

1.2 Organization

The rest of the paper is structured as follows. In section 2, we review the basics of elliptical distributions and total variation distance. In section 3, we outline the method in [2] for using sample compression schemes to prove sample complexity upper bounds for learning mixtures of distributions. In section 4, we apply the sample compression method to prove our upper bound for learning mixtures of elliptical distributions.

2 Preliminaries

In this section, we review the basics of elliptical distributions and total variation distance.

2.1 Elliptical distributions

Elliptical distributions are symmetric about their center $\boldsymbol{\mu}$. The mean of an elliptical distribution, if it exists, is equal to $\boldsymbol{\mu}$. The covariance of an elliptical distribution, if it exists, is a scalar multiple of the shape matrix Σ. However, some important classes of elliptical distributions (such as Cauchy distributions) do not have a mean or covariance.

We denote by $E_d(\boldsymbol{\mu}, \Sigma, \phi)$ an elliptical distribution with center $\boldsymbol{\mu}$, shape Σ, and characteristic function ϕ. By definition, ϕ is the Fourier transform of the probability density function of the distribution. Thus, if the radial density function of $E_d(\boldsymbol{\mu}, \Sigma, \phi)$ is f, then $\phi(\boldsymbol{x}) = e^{i\boldsymbol{x}^T\boldsymbol{\mu}}\psi_f(\boldsymbol{x}^T\Sigma\boldsymbol{x})$

where ψ_f is the Fourier transform of f. We will use the following well-known properties of elliptical distributions.

Lemma 2.1 (Affine transformations of elliptical distributions are elliptical)**.** *Let A be a $k \times d$ matrix and let $\boldsymbol{b}$ be a $k \times 1$ vector. If $\boldsymbol{x} \sim E_d(\boldsymbol{\mu}, \Sigma, \phi)$, then $A\boldsymbol{x} + \boldsymbol{b} \sim E_k(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T, \phi)$.*

Lemma 2.2 (Sums of elliptical vectors with the same shape matrix are elliptical)**.** *Suppose $\boldsymbol{x_1} \sim E_d(\boldsymbol{\mu}_1, \Sigma, \phi_1)$ and $\boldsymbol{x_2} \sim (\boldsymbol{\mu}_2, \Sigma, \phi_2)$. If $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are independent, then $\boldsymbol{x_1} + \boldsymbol{x_2} \sim E_d(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma, \phi_1\phi_2)$.*

2.2 Total variation distance

Definition 2.3. The total variation distance between two distributions with density functions functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is

$$TV(f, g) = \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - g(\boldsymbol{x})| d\boldsymbol{x} = \frac{1}{2}\|f - g\|_1.$$

where $\|f - g\|_1 = \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - g(\boldsymbol{x})| d\boldsymbol{x}$ is the L_1 distance between functions f and g.

Given samples from a mixture of elliptical distributions with density f, we wish to construct a mixture of elliptical distributions with density $\hat{f}$ such that f and $\hat{f}$ are close in total variation distance. Density estimation for a mixture of distributions under total variation distance does not require structural assumptions on the minimal mixture weight, minimal separation between distribution centers, or maximal variance of a distribution in any direction. By contrast, under other models such as parameter estimation, density estimation under KL divergence, or density estimation under L_p distances (for $p \geq 1$), even learning a mixture of Gaussian distributions would require additional assumptions on the mixture weights and separation between the Gaussian means [2].

There are known upper and lower bounds for the total variation distance between two multivariate Gaussians in terms of their covariance matrices and the difference between their means [6]. However, we are not aware of such bounds for elliptical distributions in general.

3 Sample compression schemes for learning mixtures of distributions

In this section we describe the sample compression schemes developed by Ashtiani et. al. in [2]. A sample compression scheme for a class of distributions F on domain Z consists of an encoder and a decoder.

- The encoder wishes to encode a distribution $f \in F$. Given a set of m samples from f, the encoder chooses a representative subset of τ samples (in Z^τ) and a sequence of t bits (in $\{0, 1\}^t$) to send to the decoder.
- The decoder receives the τ samples and t bits and outputs a distribution $\hat{f} \in F$ such that $TV(f, \hat{f})$ is small. Thus, the decoder is formally defined as a family of deterministic functions of the form $J_F : \cup_{n=0}^{\tau} Z^n \times \cup_{n=0}^{t} \{0, 1\}^n \to F$ where τ and t can range from 0 to ∞.

The goal of the encoder is to represent the distribution in as few samples and bits as possible while still ensuring that the decoder can output a good approximation for the distribution with high probability.

Under *non-robust* compression, the encoder receives samples from $f \in F$ and sends a sequence of bits and a "representative" subset of the samples. The decoder outputs a distribution that is close to f with high probability.

Definition 3.1 (Non-robust compression schemes)**.** We say that a class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ *non-robust compression* if there exists a decoder J_F such that the following holds for any $f \in F$:

If a sample S is drawn from $f^{m(\varepsilon)}$, then with probability at least $2/3$, there exists a subset $L \subset S$ of at most $\tau(\varepsilon)$ samples and a sequence B of at most $t(\varepsilon)$ bits such that $\|J_F(L, B) - \hat{f}\|_1 \leq \varepsilon$.

Under *r-robust* compression, the encoder receives samples not from $f \in F$ but from another distribution q that is within r total variation distance of f. Now, the encoder must represent f using a subset of the samples from q and a sequence of bits, such that the decoder can output a distribution that is close to f with high probability.

Definition 3.2 (Robust compression schemes)**.** We say that a class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ *r-robust compression* if there exists a decoder J_F such that the following holds for any $f \in F$:

Suppose $\|q - f\|_1 \leq r$. If a sample S is drawn from $q^{m(\varepsilon)}$, then with probability at least $2/3$, there exists a subset $L \subset S$ of at most $\tau(\varepsilon)$ samples and a sequence B of at most $t(\varepsilon)$ bits such that $\|J_F(L, B) - \hat{f}\|_1 \leq \varepsilon$.

[2] showed that if there exists a non-robust sample compression scheme for F, then F is learnable in the realizable setting. If there exists a robust sample compression scheme for F, then F is learnable in the agnostic setting.

Lemma 3.3 (Theorem 4.5 in [2])**.** *If class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ r-robust compression, then F can be $\max\{3, 2/r\}$-learned in the agnostic setting with sample complexity*

$$\tilde{O}\left(m(\varepsilon/6) + \frac{(\tau(\varepsilon) + t(\varepsilon/6)) \log m(\varepsilon/6)}{\varepsilon^2}\right).$$

If F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ non-robust compression, then F can be learned in the realizable setting with the same sample complexity.

Suppose there exists a *non-robust* sample compression scheme (τ, t, m) for F. Then there exists a non-robust sample compression scheme for k-mix(F).

Lemma 3.4 (Lemma 4.8 in [2])**.** *If class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ non-robust compression, then k-mix(F) admits $(k \cdot \tau(\varepsilon/3), k \cdot t(\varepsilon/3) + k \log(3k/\varepsilon), \frac{48k \log(6k)}{\varepsilon} \cdot m(\varepsilon/3))$ non-robust compression.*

Combining Lemmas 3.3 and 3.4, we conclude that there exists an algorithm for learning any finite mixture of distributions from F in the *realizable* setting, with sample complexity log-linear in τ, t, and m.

Lemma 3.5. *If class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ non-robust compression, then k-mix(F) admits learning in the realizable setting with sample complexity*

$$\tilde{O}\left(\frac{km(\epsilon/18)}{\epsilon} + \frac{k(\tau(\epsilon/18) + t(\epsilon/18)) \log m(\epsilon/18)}{\varepsilon^2}\right).$$

Similarly, finding a *robust* sample compression scheme (τ, t, m) for a class F of distributions guarantees an algorithm for learning any finite mixture of distributions from F in the *agnostic* setting, with sample complexity log-linear in τ, t, and m. We will use the following sample complexity bound in our proofs for agnostic learning of mixtures of elliptical distributions.

Lemma 3.6 (Lemma 4.9 in [2]). *If class F admits $(\tau(\epsilon), t(\varepsilon), m(\varepsilon))$ r-robust compression, then k-mix(F) admits $\frac{3}{2}(1+2/r)$-agnostic learning with sample complexity*

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

However, the time complexity of the algorithm in [2] for learning mixtures using sample compression schemes is exponential. The algorithm exhaustively tries every possible way of splitting the set m of samples into k subsets. The algorithm also exhaustively guesses the weights $w_1, w_2, \ldots, w_k$ by constructing a fine mesh over $[0,1]^k$. Then, since there are finitely many sets of samples L and sequences of bits B in $\cup_{n=0}^{\tau} Z^n \times \cup_{n=0}^{t}\{0,1\}^n$, the algorithm can calculate all of the possible candidate distributions $J_F(L_i, B_i)$ for each subset i. It is then shown that one may use this list of candidate distributions to output a mixture of k candidate distributions $J_F(L_i, B_i)$ with weights $\hat{w}_i$ that is close in total variation distance to the actual mixture distribution, with high probability.

4 Upper bound for learning mixture of elliptical distributions

In this section, we prove a sample complexity upper bound of $O(kd^2/\epsilon^2)$ for learning mixtures of k elliptical distributions of dimension d. An outline of the proof is as follows.

First, we give an algorithm for encoding a single elliptical distribution $f \overset{d}{=} E_d(\boldsymbol{\mu}, \Sigma, \phi)$. We initially assume that ϕ is known and fixed. We show that we can encode approximations for $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ using $O(d)$ samples and $O\left(d^2 \log\left(\frac{R_1 d}{R_2 \varepsilon}\right)\right)$ bits. Next, we show that if $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ are close and Σ and $\hat{\Sigma}$ are close, then $E_d(\boldsymbol{\mu}, \Sigma, \phi)$ and $E_d(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \phi)$ are close in total variation distance.

The above results allow us to prove that if ϕ is drawn from a known finite set of candidate characteristic functions, then our sample compression scheme correctly encodes a single elliptical distribution. Finally, we apply Lemma 3.6 to prove our sample complexity bound for learning mixtures of elliptical distributions.

Throughout the proof, we will use the following notation for balls in $\mathbb{R}^d$.

Definition 4.1 (Balls in $\mathbb{R}^d$). Let $B_d(r)$ denote the ball of radius r centered at the origin in $\mathbb{R}^d$. Let $B_d(r, \boldsymbol{\mu})$ denote the ball of radius r centered at $\boldsymbol{\mu}$.

4.1 Encoding a single elliptical distribution

Our goal in this section is to encode $\boldsymbol{\mu}$ and Σ for an elliptical distribution with known characteristic function ϕ. Suppose we are given m samples. An overview of the encoding algorithm is as follows.

Let $\Sigma = \sum_{i=1}^{d} \boldsymbol{w_i} \boldsymbol{w_i}^T$, which is always possible because it is a positive-semidefinite matrix. First, we encode $\hat{\boldsymbol{w_i}}$. In Lemma 4.2, we show that with high probability, the $1/20$ radius ball centered at the origin is contained in the convex hull of samples. In Lemma 4.3, we show that this allows us

to express each vector $\boldsymbol{w_i}$ as a linear combination of samples where each coefficient is in $[-20, 20]$. We construct a net for $[-20, 20]^m$ and encode the point in the net closest to the coefficient vector.

Next, we encode $\hat{\boldsymbol{\mu}}$. With high probability, some sample is close to the actual center $\boldsymbol{\mu}$. We create a discrete net around the sample using the estimated $\hat{\boldsymbol{w}}_{\boldsymbol{i}}$ vectors and encode the element of the net closest to the mean. The next two lemmas are based on Lemmas 5.6 and 5.7 from [2]. Let $\text{conv}(T)$ denote the convex hull of the set T.

Lemma 4.2. *Suppose $\boldsymbol{q_1}, ..., \boldsymbol{q_m}$ are i.i.d. samples from a d-dimensional distribution Q such that $TV(Q, E_d(\mathbf{0}, I_d, \phi^2)) \leq 2/3$ where ϕ is as in Theorem 1.4. Let $T = \{\pm \boldsymbol{q_i} : \|\boldsymbol{q_i}\|_2 \leq \sqrt{R_1 d}\}$. There exists an absolute constant C such that if $m \geq Cd$, then*

$$P\left[B_d\left(2R_2\right) \subseteq \text{conv}(T)\right] \geq 5/6.$$

Proof. For each $\boldsymbol{y} \in S^{d-1}$, let

$$H_{\boldsymbol{y}} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq \sqrt{R_1 d}, |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq 2R_2\}$$

Let $U = \{\boldsymbol{q_1}, \boldsymbol{q_2}, ..., \boldsymbol{q_m}\}$. It is necessary and sufficient to show that with probability at least $5/6$,

$$U \cap H_{\boldsymbol{y}} \neq \emptyset \quad \forall \boldsymbol{y} \in S^{d-1}.$$

Let $\mathcal{H} = \{H_{\boldsymbol{y}} : \boldsymbol{y} \in S^{d-1}\}$. By the Vapnik-Chervonenkis inequality, for some constant c,

$$E\left[\sup_{H \in \mathcal{H}} \left| Q(H) - \frac{|U \cap H|}{m} \right| \right] \leq c\sqrt{\frac{\text{VC-dim}(\mathcal{H})}{m}}. \tag{3}$$

Note that VC-dim$(\mathcal{H})$ is at most the VC dimension of $\{\boldsymbol{x} \in \mathbb{R}^d : |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq 2R_2\}$. The family of pairwise unions of half-spaces has VC dimension at most $4(d+1)\log_2 6$. So VC-dim$(\mathcal{H}) = O(d)$.

Let $\boldsymbol{g} \sim E_d(\mathbf{0}, I_d, \phi^2)$. By Lemma 2.2, we have $\boldsymbol{g} \sim \boldsymbol{x_1} + \boldsymbol{x_2}$ where $\boldsymbol{x_1}, \boldsymbol{x_2}$ are i.i.d. samples from $E_d(\mathbf{0}, I_d, \phi)$. By triangle inequality, $\|\boldsymbol{x_1} + \boldsymbol{x_2}\|_2 \leq \|\boldsymbol{x_1}\|_2 + \|\boldsymbol{x_2}\|_2$. By union bound, we have

$$\begin{aligned} P(\|\boldsymbol{g}\|_2 \geq 2\sqrt{R_1 d}) &\leq P(\|\boldsymbol{x_1}\|_2 + \|\boldsymbol{x_2}\|_2 \geq 2\sqrt{R_1 d}) \\ &\leq P(\|\boldsymbol{x_1}\|_2 \geq \sqrt{R_1 d}) + P(\|\boldsymbol{x_2}\|_2 \geq \sqrt{R_1 d}) \\ &\leq 0.05 \end{aligned}$$

where the last inequality follows from Equation 1. By Lemma 2.1, for any $\boldsymbol{y} \in S^{d-1}$, $\langle \boldsymbol{g}, \boldsymbol{y} \rangle \sim E_1(0, 1, \phi^2)$, so that $\langle \boldsymbol{g}, \boldsymbol{y} \rangle \sim x_1 + x_2$ where x_1 and x_2 are i.i.d. samples from $E_1(0, 1, \phi)$. By the triangle inequality and union bound,

$$\begin{aligned} P\left[|\langle \boldsymbol{g}, \boldsymbol{y} \rangle| < 2R_2\right] &= P\left[|x_1 + x_2| < 2R_2\right] \\ &\leq P\left[|x_1| + |x_2| < 2R_2\right] \\ &\leq P\left[|x_1| < R_2\right] + P\left[|x_2| < R_2\right] \\ &\leq \frac{1}{10}, \end{aligned}$$

where the last inequality follows from Equation 2 in Theorem 1.4.

By union bound, for any $\boldsymbol{y} \in S^{d-1}$ we have

$$\begin{aligned} P[\boldsymbol{g} \in H_{\boldsymbol{y}}] &\geq 1 - P\left[\|\boldsymbol{g}\|_2 \geq \sqrt{R_1 d}\right] - P\left[|\langle \boldsymbol{g}, \boldsymbol{y}\rangle| < 2R_2\right] \\ &\geq 1 - 0.05 - \frac{1}{10} \\ &= 0.85. \end{aligned}$$

Because $TV(Q, E_d(\mathbf{0}, I_d, \phi^2)) \leq 2/3$ we have $Q(H) \geq 0.85 - 2/3 > 0.18 = \Omega(1)$ for any $H \in \mathcal{H}$. Let $p = \inf_{H\in\mathcal{H}} Q(H)$. Let sample size $m = 144c^2 \cdot \text{VC-dim}(\mathcal{H})/p^2$. Note that $m = O(d)$, since we have shown above that $p = \Omega(1)$ and $\text{VC-dim}(\mathcal{H}) = O(d)$. Then Equation 3 becomes

$$E\left[\sup_{H\in\mathcal{H}}\left|Q(H) - \frac{|U \cap H|}{m}\right|\right] \leq p/12.$$

By Markov's inequality, this implies that with probability at least $5/6$,

$$Q(H) - \frac{|U \cap H|}{m} \leq p/2 \quad \forall H \in \mathcal{H}.$$

Then, since $Q(H) \geq p$ for all $H \in \mathcal{H}$,

$$\frac{|U \cap H|}{m} \geq Q(H) - p/2 \geq p/2 > 0.$$

This proves $|U \cap H| \neq \emptyset$ for all $H \in \mathcal{H}$ as desired. □

Suppose Σ is full-rank and $\Sigma = \sum_{i=1}^{d} \boldsymbol{w_i}\boldsymbol{w_i}^T$. As in [2], the case where Σ has rank $p < d$ can be reduced to the full-rank case since a significant fraction of the samples will lie in an affine subspace S of dimension p with high probability. We can encode S using the samples that lie in it.

Lemma 4.3. *Let $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_{2m}}$ be i.i.d. samples from a fixed d-dimensional distribution Q satisfying $TV(Q, E_d(\boldsymbol{\mu}, \Sigma, \phi)) \leq 1/3$. There exists an absolute constant C such that if $m \geq Cd$, then with probability at least $2/3$, one can encode vectors $\hat{\boldsymbol{w}}_{\mathbf{1}}, ..., \hat{\boldsymbol{w}}_{\boldsymbol{d}}, \hat{\boldsymbol{\mu}} \in \mathbb{R}^d$ satisfying*

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{w}}_{\boldsymbol{j}} - \boldsymbol{w_j})\|_2 \leq \frac{\epsilon}{6} d^{-5/2} \quad \forall j \quad \textit{and}$$

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 \leq \frac{\epsilon}{2}$$

using $O(d^2 \log(Rd/\epsilon))$ bits and the $2m$ samples.

Proof. Let $\boldsymbol{z_i} = \Sigma^{-1/2}(\boldsymbol{x_i} - \boldsymbol{\mu})$ for $1 \leq i \leq 2m$. By Lemma 2.1, the distribution of $\boldsymbol{z_i}$ has TV distance at most $1/3$ from $E_d(\mathbf{0}, I_d, \phi)$. Let $\boldsymbol{q_i} = \boldsymbol{z_{2i}} - \boldsymbol{z_{2i-1}}$ for $1 \leq i \leq m$. Then the distribution of $\boldsymbol{q_i}$ has TV distance at most $2/3$ from $E_d(\mathbf{0}, I_d, \phi^2)$.

Let $T = \{\pm \boldsymbol{q_i} : \|\boldsymbol{q_i}\|_2 \leq \sqrt{R_1 d}\}$ and $\mathcal{E}$ denote the event $B_d(2R_2) \subseteq \text{conv}(T)$. By Lemma 4.2, $P(\mathcal{E}) \geq 5/6$. We assume $\mathcal{E}$ occurs for the rest of the proof. The rest of the proof is similar to the proof of Lemma 5.7 in [2], up to a constant factor (since we use $\sqrt{Rd}$ instead of $4\sqrt{d}$). We outline the proof below for completeness.

Encoding $\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, ..., \hat{\boldsymbol{w}}_d$: Let $C \geq \frac{1}{2R_2}$. For each $1 \leq j \leq d$, note that $\Sigma^{-1/2}\boldsymbol{w_j}/C$ has norm $1/C$ and is thus contained in $B_d(2R_2) \subseteq \text{conv}(T)$. Then

$$\frac{\Sigma^{-1/2}\boldsymbol{w_j}}{C} = \sum_{i=1}^{m} \theta_{j,i}\boldsymbol{q_i}$$

for some vector $\boldsymbol{\theta_j} \in [-1,1]^m$. We discretize $\boldsymbol{\theta_j}$ using a $\frac{\epsilon}{6R_1Cmd^3}$-net for $[-1,1]^m$. Since $m = O(d)$, any element of the net can be encoded in $O\left(d\log\left(\frac{R_1Cd}{\epsilon}\right)\right) = O\left(d\log\left(\frac{R_1 d}{R_2\epsilon}\right)\right)$ bits. The encoder chooses the element $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$ in the net that is closest to the true $\boldsymbol{\theta_j}$.

Let $I = \{i : \boldsymbol{q_i} \in T\}$. The decoder can calculate the estimated vectors as

$$\hat{\boldsymbol{w}}_{\boldsymbol{j}} = C\sum_{i\in I} \hat{\boldsymbol{\theta}}_{\boldsymbol{j}}(\boldsymbol{x_{2i}} - \boldsymbol{x_{2i-1}}).$$

Then

$$\begin{aligned}
\|\Sigma^{-1/2}(\hat{\boldsymbol{w}}_{\boldsymbol{j}} - \boldsymbol{w_j})\|_2 &= C\|\sum_{i\in I}(\hat{\theta}_{j,i} - \theta_{j,i})\boldsymbol{q_i}\|_2 \\
&\leq Cm\frac{\epsilon}{6R_1Cmd^3}(\sqrt{R_1 d}) \\
&\leq \frac{\epsilon}{6\sqrt{R_1}}d^{-5/2} \\
&\leq \frac{\epsilon}{6}d^{-5/2}
\end{aligned}$$

where the last inequality is due to $R_1 \geq 1$.

Encoding $\hat{\boldsymbol{\mu}}$: Recall from the beginning of the proof that $\boldsymbol{z_i}$ has TV distance at most $1/3$ from $E_d(\mathbf{0}, I_d, \phi)$. By Equation 1,

$$P(\|\boldsymbol{z_i}\|_2 \geq \sqrt{R_1 d}) \leq 0.05 + 1/3 \leq \sqrt{1/6}.$$

Then $P\left(\min\{\|\boldsymbol{z_1}\|_2, \|\boldsymbol{z_2}\|_2\} \leq \sqrt{R_1 d}\right) = 1 - P\left(\|\boldsymbol{z_i}\|_2 \geq \sqrt{R_1 d}\right)^2 \geq 5/6$. Assume the event $\min\{\|\boldsymbol{z_1}\|_2, \|\boldsymbol{z_2}\|_2\} \leq \sqrt{R_1 d}$ occurs for the rest of the proof.

Assume without loss of generality that $\|\boldsymbol{z_1}\|_2 \leq \sqrt{R_1 d}$. Then $\boldsymbol{z_1} = \sum_{j=1}^{d} \lambda_j \frac{\boldsymbol{w_j}}{\|\boldsymbol{w_j}\|_2}$ for some vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d) \in B_d(\sqrt{R_1 d})$. We have

$$\boldsymbol{\mu} = \boldsymbol{x_1} - \Sigma^{1/2}\boldsymbol{z_1} = \boldsymbol{x_1} - \sum_{j=1}^{d} \lambda_j \boldsymbol{w_j}.$$

We discretize $\boldsymbol{\lambda}$ using a $\frac{\epsilon}{3d}$-net for $B_d(\sqrt{R_1 d})$. An element of this net can be encoded in $O(d\log(d/\epsilon))$. Choose the element $\hat{\lambda}$ closest to λ. The decoder calculates

$$\hat{\boldsymbol{\mu}} = \boldsymbol{x_1} - \sum_{j=1}^{d} \hat{\lambda}_j \hat{\boldsymbol{w}}_{\boldsymbol{j}}.$$

Then

$$\begin{aligned}\|\Sigma^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\|_2 &\leq \sum_{j=1}^{d}\|\hat{\lambda}_j(\Sigma^{-1/2}\boldsymbol{w_j}-\Sigma^{-1/2}\hat{\boldsymbol{w}}_{\boldsymbol{j}})+(\lambda_j-\hat{\lambda}_j)\Sigma^{-1/2}\boldsymbol{w_j}\|_2\\ &\leq d(\sqrt{R_1 d}\cdot\frac{\epsilon}{6\sqrt{R_1}}d^{-5/2}+\frac{\epsilon}{3d}\cdot 1)\\ &\leq \frac{\epsilon}{2}.\end{aligned}$$

□

4.2 Bounding total variation distance under parameter estimates

Next we show that our estimates of $\boldsymbol{\mu}$ and Σ are accurate enough to guarantee that the encoded distribution is close in total variation distance to the original distribution. Our main result for this section is the following.

Lemma 4.4. *Let* $\Sigma=\sum_i \boldsymbol{w_i}\boldsymbol{w_i}^T$ *and* $\hat{\Sigma}=\sum_i \hat{\boldsymbol{w}}_{\boldsymbol{i}}\hat{\boldsymbol{w}}_{\boldsymbol{i}}^T$. *Suppose that*

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{w}}_{\boldsymbol{j}}-\boldsymbol{w_j})\|_2 \leq \frac{\epsilon}{12d^2}\leq\frac{1}{6d}\quad\forall 1\leq j\leq d\quad \textit{and} \tag{4}$$

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\|_2\leq\frac{\epsilon^2}{16d^2}. \tag{5}$$

Then

$$TV(E_d(\hat{\boldsymbol{\mu}},\hat{\Sigma},\phi),E_d(\boldsymbol{\mu},\Sigma,\phi))\leq\epsilon.$$

We express the L_1 distance between two probability density functions as the integral over s of the volume of the symmetric difference between the shadows of the "caps" of the the two density function above height s. The next lemma is easily proved using Fubini's theorem.

Lemma 4.5. *Let* h, g *be two nonnegative functions. Then*

$$\|h-g\|_1=\int_0^\infty \text{vol}\left(h^{-1}([s,\infty))\ \Delta\ g^{-1}([s,\infty))\right)ds,$$

where $A\Delta B$ *denotes the symmetric difference between two sets* A, B.

To find the total variation distance between $E_d(\hat{\boldsymbol{\mu}},\hat{\Sigma},\phi)$ and $E_d(\boldsymbol{\mu},\Sigma,\phi)$, we split into two cases. First, we find the total variation distance between two elliptical distributions with the same shape matrix but different centers. Then, we find the total variation distance between two elliptical distributions with the same center but different shape matrices. We can then use the triangle inequality to bound the total variation distance between two elliptical distributions with different centers and different shape matrices.

We first find the total variation distance between two spherical distributions with the same shape matrix I_d but different centers. We use the following inequality.

Equation 4.6. Suppose $\delta\leq\frac{\epsilon}{2d}$. Then $(1+\delta)^d-1\leq\epsilon$.

Lemma 4.7 (Same shape matrix case)**.** *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a monotone decreasing function, and let g denote the corresponding radial function on $\mathbb{R}^d$. Let $h(x) = g(x - \delta)$ for $\delta \in \mathbb{R}^d$. If $d = 1$, let $\varepsilon \leq \frac{8}{f(0)}$. If $d > 1$, let ε be small enough that $\varepsilon^{d-1} \leq \frac{1}{24 f(0)}$. Finally assume $\|\delta\|_2 \leq \frac{\epsilon^2}{16 d^2}$ Then we have*

$$\|h - g\|_1 \leq \epsilon.$$

Proof. First consider the case $d = 1$. We have

$$\begin{aligned}
\|h - g\|_1 &= \int_0^\infty \operatorname{vol}\left(h^{-1}([s,\infty)) \;\Delta\; g^{-1}([s,\infty))\right) ds \\
&= \int_0^\infty \operatorname{vol}\left([-f^{-1}(s), f^{-1}(s)] \;\Delta\; [-f^{-1}(s) + \delta, f^{-1}(s) + \delta]\right) ds \\
&\leq \int_0^{f(0)} 2\delta ds \\
&\leq 2 \cdot \frac{\epsilon^2}{16} \cdot f(0) \\
&\leq \epsilon.
\end{aligned}$$

Now suppose $d > 1$. We break the integral over $[0, \infty)$ into the intervals $[0, f(\varepsilon)]$ and $(f(\varepsilon), \infty)$. We have

$$\begin{aligned}
\|h - g\|_1 &= \int_0^\infty \operatorname{vol}\left(h^{-1}([s,\infty)) \;\Delta\; g^{-1}([s,\infty))\right) ds \\
&= \int_0^\infty \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d(f^{-1}(s), \boldsymbol{\delta})\right) ds \\
&= \int_0^{f(\varepsilon)} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)\right) ds + \int_{f(\varepsilon)}^{f(0)} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)\right) ds \\
&= \frac{1}{2} \int_0^{f(\varepsilon)} \left[\operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)\right) + \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), -\boldsymbol{\delta}\right)\right)\right] ds \\
&\quad + \int_{f(\varepsilon)}^{f(0)} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)\right) ds \\
&\leq \int_0^{f(\varepsilon)} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s) + \|\boldsymbol{\delta}\|_2\right)\right) ds + \int_{f(\varepsilon)}^{f(0)} \operatorname{vol}\left(B_d(f^{-1}(s)) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)\right) ds.
\end{aligned}$$

To see that the last inequality holds, note that $B_d\left(f^{-1}(s)\right)$, $B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)$, and $B_d\left(f^{-1}(s), -\boldsymbol{\delta}\right)$ are all contained in $B_d\left(f^{-1}(s) + \|\boldsymbol{\delta}\|_2\right)$. This implies that half of $B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), \boldsymbol{\delta}\right)$ is contained in $B_d\left(f^{-1}(s) + \|\boldsymbol{\delta}\|_2\right)$ but not in $B_d\left(f^{-1}(s)\right)$. Similarly, half of $B_d\left(f^{-1}(s)\right) \Delta B_d\left(f^{-1}(s), -\boldsymbol{\delta}\right)$ is contained in $B_d\left(f^{-1}(s) + \|\boldsymbol{\delta}\|_2\right)$ but not in $B_d\left(f^{-1}(s)\right)$.

If $s \in [0, f(\varepsilon)]$, then $s \leq f(\varepsilon) < f(\frac{\varepsilon}{4d})$. Since f is monotone decreasing, this implies $f^{-1}(s) > \frac{\varepsilon}{4d}$. Then $f^{-1}(s) + \|\boldsymbol{\delta}\|_2 \leq f^{-1}(s) + \frac{\varepsilon^2}{16 d^2} < (1 + \frac{\varepsilon}{4d}) f^{-1}(s)$.

If $s \in (f(\varepsilon), f(0))$, then $f^{-1}(s) < \varepsilon$. Then the volume of the symmetric difference of two d-dimensional balls with radius $f^{-1}(s)$ is less than twice the volume of $B_d(\varepsilon)$.

Thus

$$\begin{aligned}
\|h-g\|_1 &< \int_0^{f(\varepsilon)} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right)\Delta B_d((1+\frac{\varepsilon}{4d})f^{-1}(s))\right) ds + \int_{f(\varepsilon)}^{f(0)} 2\operatorname{vol}\left(B_d(\varepsilon)\right) ds \\
&\leq [(1+\frac{\varepsilon}{4d})^d - 1]\int_0^{\infty} \operatorname{vol}\left(B_d\left(f^{-1}(s)\right)\right) ds + 2\varepsilon^d f(0)\cdot \operatorname{vol}\left(B_d(1)\right) \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&\leq \varepsilon.
\end{aligned}$$

□

Next we find the total variation distance between two elliptical distributions with the same center but different shape matrices. We use the following inequality in our proof.

Equation 4.8. Suppose $\lambda \leq \frac{\epsilon}{2d}$. Let $1+\delta = \frac{1}{1-\lambda}$. Then $(1+\delta)^{d/2} - (1-\delta)^{d/2} \leq \epsilon$.

Lemma 4.9 (Same center case)**.** *Let $f : \mathbb{R}^+ \to [0,1]$ be a monotone decreasing function, and let g denote the corresponding radial function on $\mathbb{R}^d$. Let $h(x^T x) = \frac{1}{|\Sigma|^{1/2}} g(x^T\Sigma^{-1}x)$. Suppose the eigenvalues of Σ^{-1} are in $[1-\lambda, 1+\lambda]$. If $\lambda \leq \frac{\epsilon}{4d}$, then $\|h-g\|_1 \leq \epsilon$.*

Proof. Let $g_2(\boldsymbol{x}) = g(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}) = |\Sigma|^{1/2}h(\boldsymbol{x}^T\boldsymbol{x})$. Let $1+\delta = \frac{1}{1-\lambda}$.

Using Inequality 4.8,

$$\begin{aligned}
\|h-g_2\|_1 &= \int_{\mathbb{R}^d} \left|h(\boldsymbol{x}^T\boldsymbol{x}) - |\Sigma|^{1/2}h(\boldsymbol{x}^T\boldsymbol{x})\right| d\boldsymbol{x} \\
&= \left|1-|\Sigma|^{1/2}\right| \int_{\mathbb{R}^d} h(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} \\
&= \left|1-|\Sigma|^{1/2}\right| \\
&\leq \left(\frac{1}{1-\lambda}\right)^{d/2} - 1 \\
&= (1+\delta)^{d/2} - 1 \\
&\leq (1+\delta)^{d/2} - (1-\delta)^{d/2} \\
&\leq \frac{\epsilon}{2}
\end{aligned}$$

where we used $\int_{\mathbb{R}^d} h(\boldsymbol{x}^T\boldsymbol{x})d\boldsymbol{x} = 1$. We also have

$$\begin{aligned}
\|g_2 - g\|_1 &= \int_0^\infty \text{vol}\left(g_2^{-1}([s,\infty))\right) \Delta\, g^{-1}([s,\infty)))ds \\
&\leq \int_0^\infty \text{vol}\left(B_d\left((\frac{1}{1-\lambda})^{1/2}f^{-1}(s)\right)\Delta B_d\left((\frac{1}{1+\lambda})^{1/2}f^{-1}(s)\right)\right)ds \\
&\leq \int_0^\infty \text{vol}\left(B_d\left((1+\delta)^{1/2}f^{-1}(s)\right)\Delta B_d((1-\delta)^{1/2}f^{-1}(s))\right)ds \\
&= \int_0^\infty \text{vol}\left(B_d(1)\right)f^{-1}(s)^d[(1+\delta)^{d/2} - (1-\delta)^{d/2}]ds \\
&\leq \frac{\varepsilon}{2}\int_0^\infty \text{vol}\left(B_d(1)\right)f^{-1}(s)^d ds \\
&= \frac{\varepsilon}{2}
\end{aligned}$$

where the last line is true because $\int_{\mathbb{R}^d} g(x^Tx)dx = \int_0^\infty \text{vol}\left(B_d(1)\right)f^{-1}(s)^d ds = 1$. By triangle inequality, $\|h - g\|_1 \leq \|h - g_2\|_1 + \|g_2 - g\|_1 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$. □

Finally, we can prove Lemma 4.4. We use the following lemma from [2] in our proof.

Lemma 4.10 (Lemmas 5.8 and 5.9 in [2])**.** *Suppose* $\|\Sigma^{-1/2}(\hat{\boldsymbol{w}}_{\boldsymbol{j}} - \boldsymbol{w}_{\boldsymbol{j}})\|_2 \leq \rho$. *Then* $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_d\|_{op} \leq 3d\rho$.

Proof of Lemma 4.4. Let $\boldsymbol{\mu}' = \Sigma^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ and $\Sigma' = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$. By Equation 5, we have $\|\boldsymbol{\mu}'\|_2 \leq \frac{\epsilon^2}{16d^2}$. Thus, Lemma 4.7 implies that

$$TV(E_d(\mathbf{0}, I_d, \phi), E_d(\boldsymbol{\mu}', I_d, \phi)) \leq \epsilon/2.$$

By Equation 4 and Lemma 4.10, we have $\|\Sigma' - I_d\|_{op} \leq 3d \cdot \frac{\epsilon}{12d^2} = \frac{\epsilon}{4d} \leq \frac{1}{2}$. Then the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_d$ of Σ' fall in $[1 - \frac{\epsilon}{4d}, 1 + \frac{\epsilon}{4d}]$. By Lemma 4.9, we have

$$TV(E_d(\boldsymbol{\mu}', I_d, \phi), E_d(\boldsymbol{\mu}', \Sigma', \phi)) = TV(E_d(\mathbf{0}, I_d, \phi), E_d(\mathbf{0}, \Sigma', \phi)) \leq \epsilon/2.$$

By triangle inequality,

$$\begin{aligned}
TV(E_d(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \phi), E_d(\boldsymbol{\mu}, \Sigma, \phi)) &= TV(E_d(\mathbf{0}, I_d, \phi), E_d(\boldsymbol{\mu}', \Sigma', \phi)) \\
&\leq TV(E_d(\mathbf{0}, I_d, \phi), E_d(\boldsymbol{\mu}', I_d, \phi)) + TV(E_d(\boldsymbol{\mu}', I_d, \phi), E_d(\boldsymbol{\mu}', \Sigma', \phi)) \\
&\leq \epsilon/2 + \epsilon/2 \\
&= \varepsilon.
\end{aligned}$$

□

4.3 Robust compression of a single elliptical distribution

Lemma 4.11. *Suppose the characteristic functions* $\phi_1, \phi_2, \ldots \phi_M$ *satisfy the conditions of 1.4. Then* $F_d(\phi_1, \phi_2, \ldots \phi_M)$ *admits a* $\left(O(d), O\left(d^2 \log\left(\frac{R_1 d^3}{R_2 \epsilon^2}\right) + \log M\right), O(d)\right)$ *2/3-robust compression scheme.*

Proof. Let $f \in F$. We can write $f = E_d(\boldsymbol{\mu}, \Sigma, \phi_i)$ where $1 \le i \le M$. Let g be a distribution such that $\|g - f\|_1 \le 2/3$. Then $TV(g, f) = \frac{1}{2}\|g - f\|_1 \le 1/3$ so we can apply Lemma 4.3.

Let $\delta = \frac{\epsilon^2}{8d^2}$. By Lemma 4.3, if the characteristic function ϕ_i is fixed, one can use $O\left(d^2 \log\left(\frac{R_1 d}{R_2 \delta}\right)\right) = O\left(d^2 \log\left(\frac{R_1 d^3}{R_2 \epsilon^2}\right)\right)$ bits and $O(d)$ samples from g to encode vectors $\hat{\boldsymbol{w}}_1, ..., \hat{\boldsymbol{w}}_d, \hat{\boldsymbol{\mu}} \in \mathbb{R}^d$ satisfying

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{w}}_j - \boldsymbol{w}_j)\|_2 \le \frac{\delta}{6d^2} = \frac{\epsilon^2}{48d^4} \quad \forall j \quad \text{and}$$

$$\|\Sigma^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_2 \le \frac{\delta}{2} = \frac{\epsilon^2}{16d^2}$$

with probability at least $2/3$. Thus by Lemma 4.4, we have

$$TV(E_d(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \phi_i), f) = TV(E_d(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \phi_i), E_d(\boldsymbol{\mu}, \Sigma, \phi_i)) \le \epsilon.$$

The desired result follows from the fact that we can use an additional $\log M$ bits to encode the identity of the characteristic function ϕ_i. □

4.4 Agnostically learning a mixture of elliptical distributions

We prove our main theorem.

Proof of Theorem 1.4. Lemma 4.11 gives us $\tau(\varepsilon) = O(d)$, $t(\varepsilon) = O(d^2 \log(Rd^3/\epsilon^2) + \log M)$, $m(\varepsilon) = O(d)$, and $r = 2/3$ for r-robust sample compression of the class $F_d(\phi_1, \phi_2, \dots \phi_M)$. Plugging these values into Lemma 3.6, we have that k-mix$(F_d(\phi_1, \phi_2, \dots, \phi_M))$ admits $\frac{3}{2}(1 + 2/r) = 6$-agnostic learning with sample complexity $\tilde{O}\left(\frac{k(d^2 \log(R_1/R_2) + \log M)}{\varepsilon^2}\right)$. □

5 Conclusion

We have shown a sample complexity upper bound of $\tilde{O}(k(d^2 + \log M)/\epsilon^2)$ for density estimation of k-mixtures of d-dimensional elliptical distributions with radial density functions contained in a finite set of size M. Unlike previously known algorithms for clustering heavy-tailed distributions, we do not require independent coordinates, bounded covariance, conditions on the minimal distribution weight, or separation conditions between distribution centers.

Possible future research directions include the following.

Learning mixture components and weights. We know how to output a mixture of elliptical distributions that is close in total variation distance to the actual mixture we hope to learn. However, it would be a more challenging task to learn the mixture weights and the centers and shape matrices of the k component distributions in the mixture. One approach could be to use list-decodable mean estimation to cluster the sample points with $o(1)$ error, and then use robust estimators to approximate the center and shape matrix for each cluster.

Sample complexity lower bounds for density estimation. [2] proves a matching sample complexity lower bound (up to a poly-logarithmic factor) of $\tilde{\Omega}(kd^2/\varepsilon^2)$ for learning mixtures of Gaussians. Their proof constructs $2^{\Omega(d^2)}$ d-dimensional Gaussians that are pairwise close in KL divergence but pairwise far in total variation distance, and then applies Fano's inequality. We would like to know whether there is a matching sample complexity lower bound for learning mixtures of elliptical

distributions. However, it is more challenging to bound the KL divergence and total variation for pairs of elliptical distributions, since there are no closed-form formulas for these distance metrics in general.

Polynomial time algorithms. Our polynomial sample complexity upper bound corresponds to an exponential-time algorithm for density estimation of mixtures. Polynomial-time algorithms are known for learning mixtures of Gaussians, but none are currently known for learning mixtures of elliptical distributions.

References

[1] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[2] Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3416–3425, 2018.

[3] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009.

[4] Kamalika Chaudhuri and Satish Rao. Beyond gaussians: Spectral methods for learning mixtures of heavy-tailed distributions. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 21–32. Omnipress, 2008.

[5] Anirban Dasgupta, John Hopcroft, Jon Kleinberg, and Mark Sandler. On learning mixtures of heavy-tailed distributions. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 491–500. IEEE, 2005.

[6] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[7] Ilias Diakonikolas, Daniel M Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. Clustering mixture models in almost-linear time via list-decodable mean estimation. *arXiv preprint arXiv:2106.08537*, 2021.

[8] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

[9] Shengxi Li, Zeyang Yu, and Danilo Mandic. A universal framework for learning the elliptical mixture model. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[10] Suvrit Sra and Reshad Hosseini. Geometric optimisation on positive definite matrices for elliptically contoured distributions. *Advances in Neural Information Processing Systems*, 26:2562–2570, 2013.

[11] Ting Su and Jennifer G Dy. In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intelligent Data Analysis*, 11(4):319–338, 2007.

[12] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[13] Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.