Learning Mixtures of Gaussians Using the $k$-Means Algorithm

Kamalika Chaudhuri  
CSE Department, UC San Diego  
kamalika@soe.ucsd.edu

Sanjoy Dasgupta  
CSE Department, UC San Diego  
dasgupta@cs.ucsd.edu

Andrea Vattani  
CSE Department, UC San Diego  
avattani@cs.ucsd.edu

December 1, 2009

Abstract

One of the most popular algorithms for clustering in Euclidean space is the $k$-means algorithm; $k$-means is difficult to analyze mathematically, and few theoretical guarantees are known about it, particularly when the data is well-clustered. In this paper, we attempt to fill this gap in the literature by analyzing the behavior of $k$-means on well-clustered data. In particular, we study the case when each cluster is distributed as a different Gaussian — or, in other words, when the input comes from a mixture of Gaussians.

We analyze three aspects of the $k$-means algorithm under this assumption. First, we show that when the input comes from a mixture of two spherical Gaussians, a variant of the 2-means algorithm successfully isolates the subspace containing the means of the mixture components. Second, we show an exact expression for the convergence of our variant of the 2-means algorithm, when the input is a very large number of samples from a mixture of spherical Gaussians. Our analysis does not require any lower bound on the separation between the mixture components.

Finally, we study the sample requirement of $k$-means; for a mixture of 2 spherical Gaussians, we show an upper bound on the number of samples required by a variant of 2-means to get close to the true solution. The sample requirement grows with increasing dimensionality of the data, and decreasing separation between the means of the Gaussians. To match our upper bound, we show an information-theoretic lower bound on any algorithm that learns mixtures of two spherical Gaussians; our lower bound indicates that in the case when the overlap between the probability masses of the two distributions is small, the sample requirement of $k$-means is near-optimal.
1 Introduction

One of the most popular algorithms for clustering in Euclidean space is the $k$-means algorithm \[\text{[Llo82, For65, Mac67]}\]; this is a simple, local-search algorithm that iteratively refines a partition of the input points until convergence. Like many local-search algorithms, $k$-means is notoriously difficult to analyze, and few theoretical guarantees are known about it.

There has been three lines of work on the $k$-means algorithm. A first line of questioning addresses the quality of the solution produced by $k$-means, in comparison to the globally optimal solution. While it has been well-known that for general inputs the quality of this solution can be arbitrarily bad, the conditions under which $k$-means yields a globally optimal solution on *well-clustered* data are not well-understood. A second line of work \[\text{[AV06, Vat09]}\] examines the number of iterations required by $k$-means to converge. \[\text{[Vat09]}\] shows that there exists a set of $n$ points on the plane, such that $k$-means takes as many as $\Omega(2^n)$ iterations to converge on these points. A smoothed analysis upper bound of $\text{poly}(n)$ iterations has been established by \[\text{[AMR09]}\], but this bound is still much higher than what is observed in practice, where the number of iterations are frequently sublinear in $n$. Moreover, the smoothed analysis bound applies to small perturbations of arbitrary inputs, and the question of whether one can get faster convergence on well-clustered inputs, is still unresolved. A third question, considered in the statistics literature, is the statistical efficiency of $k$-means. Suppose the input is drawn from some simple distribution, for which $k$-means is statistically consistent; then, how many samples is required for $k$-means to converge? Are there other consistent procedures with a better sample requirement?

In this paper, we study all three aspects of $k$-means, by studying the behavior of $k$-means on Gaussian clusters. Such data is frequently modelled as a mixture of Gaussians; a mixture is a collection of Gaussians $D = \{D_1, \ldots, D_k\}$ and weights $w_1, \ldots, w_k$, such that $\sum_i w_i = 1$. To sample from the mixture, we first pick $i$ with probability $w_i$ and then draw a random sample from $D_i$. Clustering such data then reduces to the problem of *learning a mixture*; here, we are given only the ability to sample from a mixture, and our goal is to learn the parameters of each Gaussian $D_i$, as well as determine which Gaussian each sample came from.

Our results are as follows. First, we show that when the input comes from a mixture of two spherical Gaussians, a variant of the 2-means algorithm successfully isolates the subspace containing the means of the Gaussians. Second, we show an exact expression for the convergence of a variant of the 2-means algorithm, when the input is a large number of samples from a mixture of two spherical Gaussians. Our analysis shows that the convergence-rate is logarithmic in the dimension, and decreases with increasing separation between the mixture components. Finally, we address the sample requirement of $k$-means; for a mixture of 2 spherical Gaussians, we show an upper bound on the number of samples required by a variant of 2-means to get close to the true solution. The sample requirement grows with increasing dimensionality of the data, and decreasing separation between the means of the distributions. To match our upper bound, we show an information-theoretic lower bound on any algorithm that learns mixtures of two spherical Gaussians; our lower bound indicates that in the case when the overlap between the probability masses of the two distributions is small, the sample requirement of 2-means is *near-optimal*.

Additionally, we make some partial progress towards analyzing $k$-means in the more general case – we show that if our variant of 2-means is run on a mixture of 2 spherical Gaussians, then, it converges to a vector in the subspace containing the means of $D_i$.

The key insight in our analysis is a novel potential function $\theta_t$, which is the minimum angle between the subspace of the means of $D_i$, and the normal to the hyperplane separator in 2-means. We show that this angle decreases with iterations of our variant of 2-means, and we can characterize convergence rates and sample requirements, by characterizing the rate of decrease of the potential.
Our Results. More specifically, our results are as follows. We perform a probabilistic analysis of a variant of 2-means; our variant is essentially a symmetrized version of 2-means, and it reduces to 2-means when we have a very large number of samples from a mixture of two identical spherical Gaussians with equal weights. In the 2-means algorithm, the separator between the two clusters is always a hyperplane, and we use the angle $\theta_t$ between the normal to this hyperplane and the mean of a mixture component in round $t$, as a measure of the potential in each round. Note that when $\theta_t = 0$, we have arrived at the correct solution.

First, in Section 3, we consider the case when we have at our disposal a very large number of samples from a mixture of $N(\mu_1, (\sigma_1)^2 I_d)$ and $N(\mu_2, (\sigma_2)^2 I_d)$ with mixing weights $\rho_1, \rho_2$ respectively. We show an exact relationship between $\theta_t$ and $\theta_{t+1}$, for any value of $\mu_j, \sigma_j, \rho_j$ and $t$. Using this relationship, we can approximate the rate of convergence of 2-means, for different values of the separation, as well as different initialization procedures. Our guarantees illustrate that the progress of $k$-means is very fast – namely, the square of the cosine of $\theta_t$ grows by at least a constant factor (for high separation) each round, when one is far from the actual solution, and slow when the actual solution is very close.

Next, in Section 4, we characterize the sample requirement for our variant of 2-means to succeed, when the input is a mixture of two spherical Gaussians. For the case of two identical spherical Gaussians with equal mixing weight, our results imply that when the separation $\mu < 1$, and when $\Omega(\frac{1}{\mu^4})$ samples are used in each round, the 2-means algorithm makes progress at roughly the same rate as in Section 3. This agrees with the $\Omega(\frac{1}{\mu^4})$ sample complexity lower bound [Lin96] for learning a mixture of Gaussians on the line, as well as with experimental results of [SSR06]. When $\mu > 1$, our variant of 2-means makes progress in each round, when the number of samples is at least $\tilde{\Omega}(\frac{1}{\mu^2})$.

Then, in Section 5, we provide an information-theoretic lower bound on the sample requirement of any algorithm for learning a mixture of two spherical Gaussians with standard deviation 1 and equal weight. We show that when the separation $\mu > 1$, any algorithm requires $\Omega(\frac{1}{\mu^2})$ samples to converge to a vector within angle $\theta = \cos^{-1}(c)$ of the true solution, where $c$ is a constant. This indicates that $k$-means has near-optimal sample requirement when $\mu > 1$.

Finally, in Section 6, we examine the performance of 2-means when the input comes from a mixture of $k$ spherical Gaussians. We show that, in this case, the normal to the hyperplane separating the two clusters converges to a vector in the subspace containing the means of the mixture components. Again, we characterize exactly the rate of convergence, which looks very similar to the bounds in Section 3.

Related Work. The convergence-time of the $k$-means algorithm has been analyzed in the worst-case [AV06, Vat09], and the smoothed analysis settings [MR09, AMR09, Vat09] shows that the convergence-time of $k$-means may be $\Omega(2^n)$ even in the plane. [AMR09] establishes a $O(n^{30})$ smoothed complexity bound. [ORSS06] analyzes the performance of $k$-means when the data obeys a clusterability condition; however, their clusterability condition is very different, and moreover, they examine conditions under which constant-factor approximations can be found. In statistics literature, the $k$-means algorithm has been shown to be consistent [Mac67]. [Pol81] shows that minimizing the $k$-means objective function (namely, the sum of the squares of the distances between each point and the center it is assigned to), is consistent, given sufficiently many samples. As optimizing the $k$-means objective is NP-Hard, one cannot hope to always get an exact solution. None of these two works quantify either the convergence rate or the exact sample requirement of $k$-means.

There has been two lines of previous work on theoretical analysis of the EM algorithm [DLR77], which is closely related to $k$-means. Essentially, for learning mixtures of identical Gaussians, the only difference between EM and $k$-means is that EM uses partial assignments or soft clusterings,
whereas $k$-means does not. First, [RW84, XJ96] view learning mixtures as an optimization problem, and EM as an optimization procedure over the likelihood surface. They analyze the structure of the likelihood surface around the optimum to conclude that EM has first-order convergence. An optimization procedure on a parameter $m$ is said to have first-order convergence, if,

$$||m_{t+1} - m^*|| \leq R \cdot ||m_t - m^*||$$

where $m_t$ is the estimate of $m$ at time step $t$ using $n$ samples, $m^*$ is the maximum likelihood estimator for $m$ using $n$ samples, and $R$ is some fixed constant between 0 and 1. In contrast, our analysis also applies when one is far from the optimum.

The second line of work is a probabilistic analysis of EM due to [DS00]; they show a two-round variant of EM which converges to the correct partitioning of the samples, when the input is generated by a mixture of $k$ well-separated, spherical Gaussians. For their analysis to work, they require the mixture components to be separated such that two samples from the same Gaussian are a little closer in space than two samples from different Gaussians. In contrast, our analysis applies when the separation is much smaller.

The sample requirement of learning mixtures has been previously studied in the literature, but not in the context of $k$-means. [CHRZ07, Cha07] provides an algorithm that learns a mixture of two binary product distributions with uniform weights, when the separation $\mu$ between the mixture components is at least a constant, so long as $\tilde{\Omega}(d/\mu^2)$ samples are available. (Notice that for such distributions, the directional standard deviation is at most 1.) Their algorithm is similar to $k$-means in some respects, but different in that they use different sets of coordinates in each round, and this is very crucial in their analysis. Additionally, [BCOFZ07] show a spectral algorithm which learns a mixture of $k$ binary product distributions, when the distributions have small overlap in probability mass, and the sample size is at least $\tilde{\Omega}(d/\mu^2)$. [Lin96] shows that at least $\tilde{\Omega}(1/\mu^4)$ samples are required to learn a mixture of two Gaussians in one dimension.

We note that although our lower bound of $\Omega(d/\mu^2)$ for $\mu > 1$ seems to contradict the upper bound of [CHRZ07, Cha07], this is not actually the case. Our lower bound characterizes the number of samples required to find a vector at an angle $\theta = \cos^{-1}(1/10)$ with the vector joining the means. However, in order to classify a constant fraction of the points correctly, we only need to find a vector at an angle $\theta' = \cos^{-1}(1/\mu)$ with the vector joining the means. Since the goal of [CHRZ07] is to simply classify a constant fraction of the samples, their upper bound is less than $O(d/\mu^2)$.

In addition to theoretical analysis, there has been very interesting experimental work due to [SSR06], which studies the sample requirement for EM on a mixture of $k$ spherical Gaussians. They conjecture that the problem of learning mixtures has three phases, depending on the number of samples: with less than about $\frac{d}{\mu}$ samples, learning mixtures is information-theoretically hard; with more than about $\frac{d}{\mu^2}$ samples, it is computationally easy, and in between, computationally hard, but easy in an information-theoretic sense. Finally, there has been a line of work which provides algorithms (different from EM or $k$-means) that are guaranteed to learn mixtures of Gaussians under certain separation conditions – see, for example, [Das99, VW02, AK05, AM05, KSV05, CR08, BV08]. For mixtures of two Gaussians, our result is comparable to the best results for spherical Gaussians [VW02] in terms of separation requirement, and we have a smaller sample requirement.

## 2 The Setting

The $k$-means algorithm iteratively refines a partitioning of the input data. At each iteration, $k$ points are maintained as centers; each input is assigned to its closest center. The center of each
cluster is then recomputed as the empirical mean of the points assigned to the cluster. This procedure is continued until convergence.

Our variant of k-means is described below. There are two main differences between the actual 2-means algorithm, and our variant. First, we use a separate set of samples in each iteration. Secondly, we always fix the cluster boundary to be a hyperplane through the origin. When the input is a very large number of samples from a mixture of two identical Gaussians with equal mixing weights, and with center of mass at the origin, this is exactly 2-means initialized with symmetric centers (with respect to the origin). We analyze this symmetrized version of 2-means even when the mixing weights and the variances of the Gaussians in the mixture are not equal.

The input to our algorithm is a set of samples $S$, a number of iterations $N$, and a starting vector $\hat{u}_0$, and the output is a vector $u_N$ obtained after $N$ iterations of the 2-means algorithm.

**2-means-iterate($S$, $N$, $u_0$)**

1. Partition $S$ randomly into sets of equal size $S_1, \ldots, S_N$.

2. For iteration $t = 0, \ldots, N - 1$, compute:

   $$ C_{t+1} = \{ x \in S_{t+1} | \langle x, u_t \rangle > 0 \} $$

   $$ \bar{C}_{t+1} = \{ x \in S_{t+1} | \langle x, u_t \rangle < 0 \} $$

   Compute: $u_{t+1}$ as the empirical average of $C_{t+1}$.

**Notation.** In Sections 3 and 4, we analyze Algorithm 2-means-iterate, when the input is generated by a mixture $D = \{D_1, D_2\}$ of two Gaussians. We let $D_1 = N(\mu^1, (\sigma^1)^2 I_d)$, $D_2 = N(\mu^2, (\sigma^2)^2 I_d)$, with mixing weights $\rho^1$ and $\rho^2$. We also assume without loss of generality that for all $j$, $\sigma^j \geq 1$. As the center of mass of the mixture lies at the origin, $\rho^1 \mu^1 + \rho^2 \mu^2 = 0$. In Section 6, we study a somewhat more general case.

We define $b$ as the unit vector along $\mu^1$, i.e. $b = \frac{\mu^1}{||\mu^1||}$. Henceforth, for any vector $v$, we use the notation $\hat{v}$ to denote the unit vector along $v$, i.e. $\hat{v} = \frac{v}{||v||}$. Therefore, $\hat{u}_t$ is the unit vector along $u_t$. We assume without loss of generality that $\mu^1$ lies in the cluster $C_{t+1}$. In addition, for each $t$, we define $\theta_t$ as the angle between $\mu^1$ and $u_t$. We use the cosine of $\theta_t$ as a measure of progress of the algorithm at round $t$, and our goal is to show that this quantity increases as $t$ increases. Observe that $0 \leq \cos(\theta_t) \leq 1$, and $\cos(\theta_t) = 1$ when $u_t$ and $\mu^1$ are aligned along the same direction. For each $t$, we define $\tau_t^j = \langle \mu^j, \hat{u}_t \rangle = \langle \mu^j, b \rangle \cos(\theta_t)$. Moreover, from our notation, $\cos(\theta_t) = \frac{\tau_t^j}{||\mu^j||}$. In addition, we define $\rho_{\min} = \min_j \rho^j$, $\rho_{\max} = \min_j ||\mu^j||$, and $\sigma_{\max} = \max_j \sigma^j$. For the special case of two identical spherical Gaussians with equal weights, we use $\mu = ||\mu^1|| = ||\mu^2||$. Finally, for $a \leq b$, we use the notation $\Phi(a,b)$ to denote the probability that a standard normal variable takes values between $a$ and $b$.

### 3 Exact Estimation

In this section, we examine the performance of Algorithm 2-means-iterate when one can estimate the vectors $u_t$ exactly – that is, when a very large number of samples from the mixture is available. Our main result of this section is Lemma 1 which exactly characterizes the behavior of 2-means-iterate at a specific iteration $t$.

For any $t$, we define the quantities $\xi_t$ and $m_t$ as follows:

$$ \xi_t = \sum_j \rho^j \sigma^j e^{-\frac{(\tau_t^j)^2}{2(\sigma^j)^2}} \frac{1}{\sqrt{2\pi}}, \quad m_t = \sum_j \rho^j \langle \mu^j, b \rangle \cdot \Phi(-\frac{\tau_t^j}{\sigma^j}, \infty) $$

5
Figure 1: Here we are depicting the plane defined by the vectors $\mu^1$ and $\tilde{u}_t$. The vector $\tilde{v}_t$ is simply the unit vector along $\mu^1 - \langle \mu^1, \tilde{u}_t \rangle \tilde{u}_t$. Therefore, we have $\tau_t^1 = ||\mu^1|| \cos(\theta_t)$ and $\sqrt{||\mu^1||^2 - (\tau_t^1)^2} = ||\mu^2|| \sin(\theta_t)$.

Now, our main lemma can be stated as follows.

Lemma 1.

$$\cos^2(\theta_{t+1}) = \cos^2(\theta_t) \left( 1 + \tan^2(\theta_t) \frac{2 \cos(\theta_t) \xi_t m_t + m_t^2}{\xi_t^2 + 2 \cos(\theta_t) \xi_t m_t + m_t^2} \right)$$

The proof is in the Appendix. Using Lemma 1 we can characterize the convergence rates and times of 2-means-iterate for different values of $\mu^j$, $\rho^j$ and $\sigma^j$, as well as different initializations of $u_0$.

The convergence rates can be characterized in terms of two natural parameters of the problem, $M = \sum_j \rho^j ||\mu^j||^2 / \sigma_j^2$, which measures how much the distributions are separated, and $V = \sum_j \rho^j \sigma^j$, which measures the average standard deviations of the distributions. We observe that as $\sigma^j \geq 1$, for all $j$, $V \geq 1$ always. To characterize these rates, it is also convenient to look at two different cases, according to the value of $\mu^j$, the separation between the mixture components.

Small $\mu^j$. First, we consider the case when each $||\mu^j||/\sigma^j$ is less than a fixed constant $\sqrt{\ln \frac{9}{2\pi}}$, including the case when $||\mu^j||$ can be much less than 1. In this case, the Gaussians are not even separated in terms of probability mass; in fact, as $||\mu^j||/\sigma^j$ decreases, the overlap in probability mass between the Gaussians tends to 1. However, we show that 2-means-iterate can still do something interesting, in terms of recovering the subspace containing the means of the distributions. Theorem 2 summarizes the convergence rate in this case.

Theorem 2 (Small $\mu^j$). Let $||\mu^j||/\sigma^j < \sqrt{\ln \frac{9}{2\pi}}$, for $j = 1, 2$. Then, there exist fixed constants $a_1$ and $a_2$, such that:

$$\cos^2(\theta_t)(1 + a_1(M/V) \sin^2(\theta_t)) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t)(1 + a_2(M/V) \sin^2(\theta_t))$$

For a mixture of two identical Gaussians with equal mixing weights, we can conclude:

Corollary 3. For a mixture of two identical spherical Gaussians with equal mixing weights, standard deviation 1, if $\mu = ||\mu^1|| = ||\mu^2|| < \sqrt{\ln \frac{9}{2\pi}}$, then,

$$\cos^2(\theta_t)(1 + a_1^j \mu^2 \sin^2(\theta_t)) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t)(1 + a_2^j \mu^2 \sin^2(\theta_t))$$

The proof follows by a combination of Lemma 1 and Lemma 25. From Corollary 3, we observe that $\cos^2(\theta_t)$ grows by a factor of $1 + \Theta(\mu^2)$ in each iteration, except when $\theta_t$ is very close to 0.
This means that when 2-means-iterate is far from the actual solution, it approaches the solution at a consistently high rate. The convergence rate only grows slower, once $k$-means is very close to the actual solution.

**Large $\mu^j$.** In this case, there exists a $j$ such that $|\mu^j|/\sigma^j \geq \sqrt{\ln \frac{9}{2\pi}}$. In this regime, the Gaussians have small overlap in probability mass, yet, the distance between two samples from the same distribution is much greater than the separation between the distributions. Our guarantees for this case are summarized by Theorem 4.

We see from Theorem 4 that there are two regimes of behavior of the convergence rate, depending on the value of $\max_j |\tau^j|/\sigma^j$. These regimes have a natural interpretation. The first regime corresponds to the case when $\theta_t$ is large enough, such that when projected onto $u_t$, at most a constant fraction of samples from the two distributions can be classified with high confidence. The second regime corresponds to the case when $\theta_t$ is close enough to 0 such that when projected along $u_t$, most of the samples from the distributions can be classified with high confidence. As expected, in the second regime, the convergence rate is much slower than in the first regime.

**Theorem 4 (Large $\mu^j$).** Suppose there exists $j$ such that $|\mu^j|/\sigma^j \geq \sqrt{\ln \frac{9}{2\pi}}$. If $|\tau^j|/\sigma^j < \sqrt{\ln \frac{9}{2\pi}}$, for all $j$, then, there exist fixed constants $a_3$, $a_4$, $a_5$ and $a_6$ such that:

$$
\cos^2(\theta_t) \left(1 + \frac{a_3(M/V)^2 \sin^2(\theta_t)}{a_4 + (M/V)^2 \cos^2(\theta_t)}\right) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t) \left(1 + \frac{a_5((M/V) + (M/V)^2) \sin^2(\theta_t)}{a_6 + (M/V)^2 \cos^2(\theta_t)}\right)
$$

On the other hand, if there exists $j$ such that $|\tau^j|/\sigma^j \geq \sqrt{\ln \frac{9}{2\pi}}$, then, there exist fixed constants $a_7$ and $a_8$ such that:

$$
\cos^2(\theta_t)(1 + \frac{a_7 \rho_{\min} \mu_{\min}^2}{a_8 V^2 + \rho_{\min} \mu_{\min}^2} \tan^2(\theta_t)) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t)(1 + \tan^2(\theta_t))
$$

For two identical Gaussians with standard deviation 1, we can conclude:

**Corollary 5.** For a mixture of two identical Gaussians with equal mixing weights, and standard deviation 1, if $\mu = ||\mu^1|| = ||\mu^2|| > \sqrt{\ln \frac{9}{2\pi}}$, and if $|\tau^j| = |\tau^j| \leq \sqrt{\ln \frac{9}{2\pi}}$, then, there exist fixed constants $a_3$, $a_4$, $a_5$, $a_6$ such that:

$$
\cos^2(\theta_t) \left(1 + \frac{a'_3 \mu_4 \sin^2(\theta_t)}{a'_4 + \mu_4 \cos^2(\theta_t)}\right) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t) \left(1 + \frac{a'_5 \mu_4 \sin^2(\theta_t)}{a'_6 + \mu_4 \cos^2(\theta_t)}\right)
$$

On the other hand, if $|\tau^j| = |\tau^j| \geq \sqrt{\ln \frac{9}{2\pi}}$, then, there exists a fixed constant $a'_7$ such that:

$$
\cos^2(\theta_t)(1 + a'_7 \tan^2(\theta_t)) \leq \cos^2(\theta_{t+1}) \leq \cos^2(\theta_t)(1 + \tan^2(\theta_t))
$$

In this case as well, we observe the same phenomenon: the convergence rate is high when we are far away from the solution, and slow when we are close. Using Theorems 2 and 4 we can characterize the convergence times of 2-means-iterate; for the sake of simplicity, we present the convergence time bounds for a mixture of two spherical Gaussians with equal mixing weights and standard deviation 1. We recall that in this case 2-means-iterate is exactly 2-means.

**Corollary 6 (Convergence Time).** If $\theta_0$ is the initial angle between $\mu^1$ and $u_0$, then, $\cos^2(\theta_N) \geq 1 - \epsilon$ after $N = C_0 \cdot \left(\frac{\ln(\cos^2(\theta_0))}{\ln(1+\mu^2)} + \frac{1}{\ln(1+\epsilon)}\right)$ iterations, where $C_0$ is a fixed constant.
Effect of Initialization. As apparent from Corollary\ref{cor:initialization} the effect of initialization is only to ensure a lower bound on the value of cos(θ₀). We illustrate below, two natural ways by which one can select u₀, and their effect on the convergence rate. For the sake of simplicity, we state these bounds for the case in which we have two identical Gaussians with equal mixing weights and standard deviation 1.

- First, one can choose u₀ uniformly at random from the surface of a unit sphere in \( \mathbb{R}^d \); in this case, \( \cos^2(θ₀) = \Theta(1/2) \), with constant probability, and as a result, the convergence time to reach \( \cos^{-1}(1/\sqrt{2}) \) is \( O(\frac{\ln d}{\ln(1+\mu^2)}) \).
- A second way to choose u₀ is to set it to be a random sample from the mixture; in this case, \( \cos^2(θ₀) = \Theta(\frac{1+\mu^2}{d}) \) with constant probability, and the time to reach \( \cos^{-1}(1/\sqrt{2}) \) is \( O(\frac{\ln d}{\ln(1+\mu^2)}) \).

4 Finite Samples

In this section, we analyze Algorithm 2-means-iterate, when we are required to estimate the statistics at each round with a finite number of samples. We characterize the number of samples needed to ensure that 2-means-iterate makes progress in each round, and we also characterize the rate of progress when the required number of samples are available.

The main result of this section is the following lemma, which characterizes \( θ_{t+1} \), the angle between \( µ^1 \) and the hyperplane separator in 2-means-iterate, given \( θ_t \). Notice that now \( θ_t \) is a random variable, which depends on the samples drawn in rounds 1, . . . , \( t-1 \), and given \( θ_t \), \( θ_{t+1} \) is a random variable, whose value depends on samples in round \( t \). Also we use \( u_{t+1} \) as the center of partition \( C_t \) in iteration \( t+1 \), and \( E[u_{t+1}] \) is the expected center. Note that all the expectations in round \( t \) are conditioned on \( θ_t \). In addition, we use \( S_{t+1} \) to denote the quantity \( E[X : 1_X \in C_{t+1}] \), where \( 1_X \in C_{t+1} \) is the indicator function for the event \( X \in C_{t+1} \), and the expectation is taken over the entire mixture. Note that, \( S_{t+1} = E[u_{t+1}] \Pr[X \in C_{t+1}] = Z_{t+1} E[u_{t+1}] \). We use \( \hat{S}_{t+1} \) to denote the empirical value of \( S_{t+1} \).

**Lemma 7.** If we use \( n \) samples in iteration \( t \), then, given \( θ_t \), with probability \( 1-2δ \),

\[
\cos^2(θ_{t+1}) ≥ \cos^2(θ_t) \left( 1 + \tan^2(θ_t) \frac{2 \cos(θ_t)ξ_t m_t + m^2}{ξ_t^2 + 2 \cos(θ_t)ξ_t m_t + m^2 + \Delta_2} \right) - \left( \frac{2 θ_{t+1}}{\xi^2_t + 2 \xi_t m_t \cos(θ_t)} + \Delta_2 \right)
\]

where,

\[
\Delta_1 = \frac{8 \log(4n/δ)(σ_{max} + \max_j ||µ^j||)}{\sqrt{n}}
\]

\[
\Delta_2 = \frac{128 \log^2 (8n/δ)(σ_{max}^2 d + \sum_j ||µ^j||^2)}{n} + \frac{8 \log(n/δ)}{\sqrt{n}}(σ_{max}||S_{t+1}|| + \max_j (⟨S_{t+1}, µ^j⟩))
\]

The main idea behind the proof of Lemma 7 is that we can write \( \cos^2(θ_{t+1}) = \frac{⟨S_{t+1}, µ^1⟩^2}{||S_{t+1}||^2 ||µ^1||^2} \). Next, we can use Lemma 1 and the definition of \( S_{t+1} \) to get an expression for \( \frac{⟨S_{t+1}, µ^1⟩^2}{||S_{t+1}||^2 ||µ^1||^2} \), and Lemmas 8 and 9 to bound \( ⟨S_{t+1} − S_{t+1}, µ^1⟩ \), and \( ||S_{t+1}||^2 − ||S_{t+1}||^2 \). Plugging in all these values gives us a proof of Lemma 7. We also assume for the rest of the section that the number of samples \( n \) is at most some polynomial in \( d \), such that \( \log(n) = Θ(\log(d)) \).

The two main lemmas used in the proof of Lemma 7 are Lemmas 8 and 9. To state them, we need to define some notation. At time \( t \), we use the notation...
Lemma 8. For any $t$, and for any vector $v$ with norm $\|v\|$, with probability at least $1 - \delta$,

$$\|\langle \hat{S}_{t+1} - S_{t+1}, v \rangle\| \leq \frac{8 \log(4n/\delta)(\sigma_{\max}\|v\| + \max_j \|\mu^j, v\|)}{\sqrt{n}}$$

Lemma 9. For any $t$, with probability at least $1 - \delta$,

$$\|\hat{S}_{t+1}\|^2 \leq \|S_{t+1}\|^2 + \frac{128 \log^2(8n/\delta)(\sigma_{\max}^2 d + \sum_j (\mu^j)^2)}{n} + \frac{16 \log(8n/\delta)}{\sqrt{n}}(\sigma_{\max}\|S_{t+1}\| + \max_j \|\mu^j\|)$$

The proofs of Lemmas 8 and 9 are in the Appendix. Applying Lemma 7, we can characterize the number of samples required for success in the first round is $\tilde{\Theta}(\frac{d}{\varepsilon^2})$. This is due to the fact that we use a fresh set of samples in every round, in order to simplify our analysis. In practice, successive iterations of k-means or EM is run on the same data-set.

Corollary 11. Let $\mu = \|\mu^1\| = \|\mu^2\| < \sqrt{\ln \frac{9}{2\pi}}$. If the number of samples drawn in round $t$ is at least $a_9 \sigma_{\max}^2 \log^2(d/\delta)\left(\frac{d}{\mu_2^2 \sin^4(\theta_t)} + \frac{\mu_1^2 \cos^2(\theta_t) \sin^4(\theta_t)}{M^2 \sin^4(\theta_t) \cos(\theta_t)}\right)$, for some fixed constant $a_9$, then, with probability at least $1 - \delta$, $\cos^2(\theta_{t+1}) \geq \cos^2(\theta_t)(1 + a_{10}\mu^2 \sin^2(\theta_t))$, where $a_{10}$ is some fixed constant.

In particular, when we initialize $u_0$ with a vector picked uniformly at random from a $d$-dimensional sphere, $\cos^2(\theta_0) \geq \frac{1}{d}$, with constant probability, and thus the number of samples required for success in the first round is $\tilde{\Theta}(\frac{d}{\varepsilon^2})$. This bound matches with the lower bounds for learning mixtures of Gaussians in one dimension [Lin96], as well as with conjectured lower bounds in experimental work [SSR06]. The following corollary summarizes the total number of samples required to learn the mixture with some fixed precision, for two identical spherical Gaussians with variance 1 and equal mixing weights.

Corollary 12. Let $\mu = \|\mu^1\| = \|\mu^2\| \leq \sqrt{\ln \frac{9}{2\pi}}$. Suppose $u_0$ is chosen uniformly at random, and the number of rounds is $N \geq C_0 \cdot (\ln \frac{d}{\mu_1^2 \sin^4(\theta_t)} + \ln \frac{1}{\mu_2^2 \sin^4(\theta_t)})$, where $C_0$ is the fixed constant in Corollary 8. If the number of samples $|S|$ is at least: $\frac{N \cdot a_{10} \mu^2 \sin^2(\theta_t)}{\mu_2^2 \sin^4(\theta_t)}$, then, with constant probability, after $N$ rounds, $\cos^2(\theta_N) \geq 1 - \varepsilon$.

One can show a very similar corollary when $u_0$ is initialized as a random sample from the mixture. We note that the total number of samples is a factor of $N \approx \frac{\ln d}{\mu^2}$ times greater than the bound in Theorem 10. This is due to the fact that we use a fresh set of samples in every round, in order to simplify our analysis. In practice, successive iterations of k-means or EM is run on the same data-set.

Theorem 13 (Large $\mu^j$). Suppose that there exists some $j$ such that $\|\mu^j\|/\sigma^j \geq \sqrt{\ln \frac{9}{2\pi}}$, and suppose that the number of samples drawn in round $t$ is at least

$$a_{11} \log^2(d/\delta)\left(\frac{d \sigma_{\max}^2}{\mu_2^2 \mu_{\min}^2 \sin^4(\theta_t)} + \frac{\sigma_{\max}^2 \max_j \|\mu^j\|^2}{M^2 \cos^2(\theta_t) \sin^4(\theta_t)} + \frac{\sigma_{\max}^2 \max_j \|\mu^j\|^2}{\mu_2^4 \mu_{\min}^4 \sin^4(\theta_t)}\right)$$
for some constant $a_{11}$. If $|\tau_t^j| \leq \sqrt{\ln \frac{a}{2\pi}}$, for all $j$, then, with probability at least $1 - \delta$, $\cos^2(\theta_{t+1}) \geq \cos^2(\theta_{t}) (1 + a_{12} \min(1, M^2 + M V) \sin^2(\theta_{t}))$; otherwise, with probability at least $1 - \delta$, $\cos^2(\theta_{t+1}) \geq \cos^2(\theta_{t}) (1 + a_{12} \sin^2(\theta_{t}))$, where $a_{12}$ and $a_{13}$ are fixed constants.

For a mixture of two identical Gaussians with equal mixing weights and standard deviation 1, our result implies:

**Corollary 14.** Suppose that $\mu = ||\mu_1|| = ||\mu_2|| \geq \sqrt{\ln \frac{a}{2\pi}}$, and suppose that the number of samples in round $t$ is at least: $a_{11} \log^2(d/\delta) \left( \frac{d}{\mu^2 \sin^4(\theta_{t})} + \frac{1}{\mu^2 \cos^2(\theta_{t}) \sin^4(\theta_{t})} \right)$, for some constant $a_{11}$. If $|\tau_t^j| \leq \sqrt{\ln \frac{a}{2\pi}}$, then, with probability at least $1 - \delta$, $\cos^2(\theta_{t+1}) \geq \cos^2(\theta_{t}) (1 + a_{12} \sin^2(\theta_{t}))$; otherwise, with probability $1 - \delta$, $\cos^2(\theta_{t+1}) \geq \cos^2(\theta_{t}) (1 + a_{13} \tan^2(\theta_{t}))$, where $a_{12}$ and $a_{13}$ are fixed constants.

Again, if we pick $u_0$ uniformly at random, we require about $\Omega(\frac{d}{\mu^2})$ samples for the first round to succeed. When $\mu > 1$, this bound is worse than $\frac{d}{\mu^2}$, but matches with the upper bounds of [BCOFZ07]. The following corollary shows the number of samples required in total for 2-means-iterate to converge.

**Corollary 15.** Let $\mu \geq \sqrt{\ln \frac{a}{2\pi}}$. Suppose $u_0$ is chosen uniformly at random and the number of rounds is $N \geq C_0 \cdot (\ln d + \frac{1}{\ln(1 + \epsilon)})$, where $C_0$ is the constant in Corollary 6. If $|S|$ is at least $\frac{2N^2d \log^2(d)}{\mu^2 \epsilon^2}$, then, with constant probability, after $N$ rounds, $\cos^2(\theta_N) \geq 1 - \epsilon$.

## 5 Lower Bounds

In this section, we prove a lower bound on the sample complexity of learning mixtures of Gaussians, using Fano’s Inequality [Yu97, CT05], stated in Theorem 10. Our main theorem in this section can be summarized as follows.

**Theorem 16.** Suppose we are given samples from the mixture $D(\mu) = \frac{1}{2} \mathcal{N}(\mu, I_d) + \frac{1}{2} \mathcal{N}(-\mu, I_d)$, for some $\mu$, and let $\hat{\mu}$ be the estimate of $\mu$ computed from $n$ samples. If $n < \frac{C d}{||\mu||^2}$ for some constant $C$, and $||\mu|| > 1$, then, there exists $\mu$ such that $\mathbb{E}_{D(\mu)} ||\mu - \hat{\mu}|| \geq C' ||\mu||$, where $C'$ is a constant.

The main tools in the proof of Theorem 16 are the following lemmas, and a generalized version of Fano’s Inequality [CT05, Yu97].

**Lemma 17.** Let $\mu_1, \mu_2 \in \mathbb{R}^d$, and let $D_1$ and $D_2$ be the following mixture distributions: $D_1 = \frac{1}{2} \mathcal{N}(\mu_1, I_d) + \frac{1}{2} \mathcal{N}(-\mu_1, I_d)$, and $D_2 = \frac{1}{2} \mathcal{N}(\mu_2, I_d) + \frac{1}{2} \mathcal{N}(-\mu_2, I_d)$. Then,

$$\text{KL}(D_1, D_2) \leq \frac{1}{\sqrt{2\pi}} \cdot \left( ||\mu_2||^2 - ||\mu_1||^2 + \frac{3\sqrt{2\pi}}{2} \ln 2 + 2 ||\mu_1|| (e^{-||\mu_1||^2/2} + \sqrt{2\pi} ||\mu_1|| \Phi(0, ||\mu_1||)) \right)$$

**Lemma 18.** There exists a set of vectors $V = \{v_1, \ldots, v_K\}$ in $\mathbb{R}^d$ with the following properties:

1. For each $i$ and $j$, $d(v_i, v_j) \geq \frac{1}{5}$, $d(v_i, -v_j) \geq \frac{1}{5}$.
2. $K = e^{d/10}$.
3. For all $i$, $||v_i|| \leq \sqrt{\frac{5}{2}}$.
Theorem 19 (Fano’s Inequality). Consider a class of densities $F$, which contains $r$ densities $f_1, \ldots, f_r$, corresponding to parameter values $\theta_1, \ldots, \theta_r$. Let $d(\cdot)$ be any metric on $\theta$, and let $\hat{\theta}$ be an estimate of $\theta$ from $n$ samples from a density $f$ in $F$. If, for all $i$ and $j$, $d(\theta_i, \theta_j) \geq \alpha$, and $\text{KL}(f_i, f_j) \leq \beta$, then, $\max_j E_j d(\hat{\theta}, \theta_j) \geq \frac{n}{2} \left( 1 - \frac{\alpha^2 + \log 2}{\log(r-1)} \right)$, where $E_j$ denotes the expectation with respect to distribution $j$.

Proof. (Of Theorem 19) We apply Fano’s Inequality. Our class of densities $F$ is the class of all mixtures of the form $\frac{1}{2} \mathcal{N}(\mu', I_d) + \frac{1}{2} \mathcal{N}(\mu, I_d)$. We set the parameter $\theta = \mu'$, and $d(\mu_1, \mu_2) = ||\mu_1 - \mu_2||$. We construct a subclass $\mathcal{F} = \{f_1, \ldots, f_r\}$ of $F$ as follows. We set each $f_i = \frac{1}{2} \mathcal{N}(||\mu||v_i, I_d) + \frac{1}{2} \mathcal{N}(-||\mu||v_i, I_d)$, for each vector $v_i$ in $V$ in Lemma 18. Notice that now $r = e^{d(\mathcal{F})}$. Moreover, for each pair $i$ and $j$, $d(\mu_i, \mu_j) \geq ||\mu||$. The Theorem now follows by an application of Fano’s Inequality 19.

6 More General $k$-means

In this section, we show that when we apply 2-means on an input generated by a mixture of $k$ spherical Gaussians, the normal to the hyperplane which partitions the two clusters in the 2-means algorithm, converges to a vector in the subspace $M$ containing the means of mixture components. We assume that our input is generated by a mixture of $k$ spherical Gaussians, with means $\mu_j$, variances $(\sigma_j)^2$, $j = 1, \ldots, k$, and mixing weights $\rho_1, \ldots, \rho_k$. The mixture is centered at the origin such that $\sum \rho_j \mu_j = 0$. We use $M$ to denote the subspace containing the means $\mu_1, \ldots, \mu_k$. We use Algorithm 2-means-iterate on this input, and our goal is to show that it still converges to a vector in $M$.

Notation. In the sequel, given a vector $x$ and a subspace $W$, we define the angle between $x$ and $W$ as the angle between $x$ and the projection of $x$ onto $W$. We examine the angle $\theta_l$, between $u_l$ and $M$, and our goal is to show that the cosine of this angle grows as $t$ increases. Our main result of this section is Lemma 20 which exactly defines the behavior of 2-means-iterate on a mixture of $k$ spherical Gaussians. Recall that at time $t$, we use $\hat{u}_t$ to partition the input data, and the projection of $\hat{u}_t$ along $M$ is $\cos(\theta_l)$ by definition. Let $b_1^l$ be a unit vector lying in the subspace $M$ such that: $\hat{u}_t = \cos(\theta_l) b_1^l + \sin(\theta_l) v_l$, where $v_l$ lies in the orthogonal complement of $M$, and has norm 1. We define a second vector $\hat{u}_t^\perp$ as follows: $\hat{u}_t^\perp = \sin(\theta_l) b_1^l - \cos(\theta_l) v_l$. We observe that $\langle \hat{u}_t, \hat{u}_t^\perp \rangle = 0$, $||\hat{u}_t^\perp|| = 1$, and the projection of $\hat{u}_t^\perp$ on $M$ is $\sin(\theta_l) b_1^l$. We now extend the set $\{b_1^l\}$ to complete an orthonormal basis $B = \{b_1^l, \ldots, b_{k-1}^l\}$ of $M$. We also observe that $\{b_2^l, \ldots, b_{k-1}^l, \hat{u}_t, \hat{u}_t^\perp\}$ is an orthonormal basis of the subspace spanned by any basis of $M$, along with $v_l$, and can be extended to a basis of $\mathbb{R}^d$.

For $j = 1, \ldots, k$, we define $\tau_j^l$ as follows: $\tau_j^l = \langle \mu_j, \hat{u}_t \rangle = \cos(\theta_l) \langle \mu_j, b_1^l \rangle$. Finally we (re)-define the quantity $\xi_t$, and define $m_t^l$, for $l = 1, \ldots, k - 1$ as:

$$\xi_t = \sum_j \rho_j^l \sigma_j^l e^{-\frac{(\tau_j^l)^2}{2(\sigma_j^l)^2}} \sqrt{2\pi} \quad \text{and} \quad m_t^l = \sum_j \rho_j^l \Phi\left(-\frac{\tau_j^l}{\sigma_j^l}, \infty\right) \langle \mu_j, b_1^l \rangle$$

Our main lemma is stated below. The proof is in the Appendix.

Lemma 20. At any iteration $t$ of Algorithm 2-means-iterate,\[ \cos^2(\theta_{t+1}) = \cos^2(\theta_t) \left( 1 + \tan^2(\theta_t) \right) \]2 cos(\theta_t) \xi_t m_t^l + \sum_l (m_t^l)^2 \xi_t^2 + 2 \cos(\theta_t) \xi_t m_t^l + \sum_l (m_t^l)^2 \right)\]
References

[AK05] S. Arora and R. Kannan. Learning mixtures of separated nonspherical Gaussians. Ann. Applied Prob., 15(1A):69–92, 2005.

[AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In COLT, 2005.

[AMR09] D. Arthur, B. Manthey, and H. Röglin. k-means has polynomial smoothed complexity. In FOCS, 2009. To appear.

[AV06] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In SoCG, 2006.

[BCOFZ07] A. Blum, A. Coja-Oghlan, A. M. Frieze, and S. Zhou. Separating populations with wide data: A spectral analysis. In ISAAC, 2007.

[BV08] S. C. Brubaker and S. Vempala. Isotropic PCA and affine-invariant clustering. In FOCS, 2008.

[Cha07] K. Chaudhuri. Learning Mixtures of Distributions. PhD thesis, University of California, Berkeley, 2007. UCB/EECS-2007-124.

[CHRZ07] K. Chaudhuri, E. Halperin, S. Rao, and S. Zhou. A rigorous analysis of population stratification with limited data. In SODA, 2007.

[CR08] K. Chaudhuri and S. Rao. Learning mixtures of distributions using correlations and independence. In COLT, 2008.

[CT05] T. Cover and J. Thomas. Elements of Information Theory : Second Edition. Wiley, 2005.

[Das99] S. Dasgupta. Learning mixtures of gaussians. In FOCS, 1999.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society B, 39, 1977.

[DS00] S. Dasgupta and L. Schulman. A two-round variant of EM for Gaussian mixtures. In UAI, 2000.

[For65] E. Forgy. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. Biometrics, 1965.

[KSV05] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In COLT, 2005.

[Lin96] B. G. Lindsey. Mixture Models:Theory, Geometry and Applications. IMS, 1996.

[Llo82] S.P. Lloyd. Least squares quantization in PCM. IEEE Trans. Information Theory, 1982.

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In Berkeley Symposium on Mathematical Statistics and Probability, 1967.
Appendix

6.1 Proof of Lemma 1

In this section, we prove Lemma 1. First, we need some additional notation.

Notation. We define, for \( j = 1, 2 \):

\[
\begin{align*}
  w^j_{t+1} &= \Pr[x \sim D_j | x \in C_{t+1}] \\
  u^j_{t+1} &= \mathbb{E}[x | x \sim D_j, x \in C_{t+1}]
\end{align*}
\]

We observe that \( u_{t+1} \) now can be written as:

\[
  u_{t+1} = w^1_{t+1}u^1_{t+1} + w^2_{t+1}u^2_{t+1}
\]

Moreover, we define \( Z_{t+1} = \Pr[x \in C_{t+1}] \).

Proof of Lemma 1. We start by providing exact expressions for \( w^1_{t+1} \) and \( w^2_{t+1} \) with respect to the partition computed in the previous round \( t \). These are used to compute the projections of \( u_{t+1} \) along the vectors \( \hat{u}_t \) and \( \mu_1 - \langle \mu_1, \hat{u}_t \rangle \hat{u}_t \), which finally leads to a proof of Lemma 1.

Lemma 21. In round \( t \), for \( j = 1, 2 \),

\[
  w^j_{t+1} = \frac{\rho^j \Phi(-\frac{\rho^j}{\sigma^2}, \infty)}{Z_{t+1}}.
\]
Proof. We can write:
\[ w^j_{t+1} = \frac{\Pr[ x \in C_{t+1} | x \sim D_j ] \Pr[ x \sim D_j ]}{\Pr[ x \in C_{t+1} ]} \]

We note that \( \Pr[ x \sim D_j ] = \rho^j \), and \( \Pr[ x \in C_{t+1} ] = Z_{t+1} \).

As \( D_j \) is a spherical Gaussian, for any \( x \) generated from \( D_j \), and for any vector \( y \) orthogonal to \( u_t \), \( \langle y, x \rangle \) is distributed independently from \( \langle \tilde{u}_t, x \rangle \). Moreover, we observe that \( \langle \tilde{u}_t, x \rangle \) is distributed as a Gaussian with mean \( \langle \mu^j, \tilde{u}_t \rangle = \tau^j_t \) and standard deviation \( \sigma^j \). Therefore,
\[
\Pr[ x \in C_{t+1} | x \sim D_j ] = \Pr_{x \sim D_j}[ \langle \tilde{u}_t, x \rangle > 0 ] = \Pr[ (\tau^j_t, \sigma^j) \geq 0 ] = \Phi\left( -\frac{\tau^j_t}{\sigma^j} \right)
\]
from which the lemma follows.

Lemma 22. For any \( t \), \( \langle u_{t+1}, \tilde{u}_t \rangle = \frac{\xi_t + m_t \cos(\theta_t)}{Z_{t+1}} \).

Proof. Consider a sample \( x \) drawn from \( D_j \). Then, \( \langle x, \tilde{u}_t \rangle \) is distributed as a Gaussian with mean \( \langle \mu^j, \tilde{u}_t \rangle = \tau^j_t \) and standard deviation \( \sigma^j \). We recall that \( \Pr[ x \in C_{t+1} ] = Z_{t+1} \). Therefore, \( \langle u_{t+1}, \tilde{u}_t \rangle \) is equal to:
\[
\mathbb{E}[ x, x \in C_{t+1} | x \sim D_j ] = \frac{1}{\Pr[ x \in C_{t+1} | x \sim D_j ]} \cdot \int_{y=0}^{\infty} \frac{ye^{-\frac{(y-\tau^j_t)^2}{2\sigma^j}}}{\sigma^j \sqrt{2\pi}} dy
\]
which is, again, equal to:
\[
\frac{1}{\Phi\left( -\frac{\tau^j_t}{\sigma^j} \right)} \left( \frac{\tau^j_t}{\sigma^j} \Phi\left( -\frac{\tau^j_t}{\sigma^j} \right) + \int_{y=0}^{\infty} \frac{(y-\tau^j_t)e^{-\frac{(y-\tau^j_t)^2}{2\sigma^j}}}{\sigma^j \sqrt{2\pi}} dy \right)
\]
We can compute the integral in the equation above as follows.
\[
\int_{y=0}^{\infty} (y-\tau^j_t) e^{-\frac{(y-\tau^j_t)^2}{2\sigma^j}} dy = (\sigma^j)^2 \int_{z=(\tau^j_t)^2/2\sigma^j}^{\infty} z e^{-z} dz = (\sigma^j)^2 e^{-\left(\frac{\tau^j_t}{\sigma^j}\right)^2/2}\]
We can now compute \( \langle u_{t+1}, \tilde{u}_t \rangle \) as follows.
\[
\langle u_{t+1}, \tilde{u}_t \rangle = w^1_{t+1} \langle u^1_{t+1}, \tilde{u}_t \rangle + w^2_{t+1} \langle u^2_{t+1}, \tilde{u}_t \rangle = \frac{1}{Z_{t+1}} \cdot \sum_j \left( \rho^j \frac{\tau^j_t}{\sigma^j} \Phi\left( -\frac{\tau^j_t}{\sigma^j} \right) + \rho^j (\sigma^j)^2 e^{-\left(\frac{\tau^j_t}{\sigma^j}\right)^2/2}\right)
\]
The lemma follows by recalling \( \tau^j_t = \langle \mu^j, b \rangle \cos(\theta_t) \) and plugging in the values of \( m_t \) and \( \xi_t \).

Lemma 23. Let \( \tilde{v}_t \) be a unit vector along \( \mu_t - \langle \mu_1, \tilde{u}_t \rangle \tilde{u}_t \). Then, \( \langle u_{t+1}, \tilde{v}_t \rangle = \frac{m_t \sin(\theta_t)}{Z_{t+1}} \). In addition, for any vector \( z \) orthogonal to \( \tilde{u}_t \) and \( \tilde{v}_t \), \( \langle u_{t+1}, z \rangle = 0 \).
Proof. We observe that for a sample \( x \) drawn from distribution \( D_1 \) (respectively, \( D_2 \)) and any unit vector \( v_1 \), orthogonal to \( \tilde{u}_t \), \( \langle x, v_1 \rangle \) is distributed as a Gaussian with mean \( \langle \mu^1, v_1 \rangle \) (resp. \( \langle \mu^2, v_1 \rangle \)) and standard deviation \( \sigma^1 \) (resp. \( \sigma^2 \)). Therefore, the projection of \( u_{t+1} \) on \( \tilde{v}_t \) can be written as:

\[
\langle u_{t+1}, \tilde{v}_t \rangle = \sum_j w^j_{t+1} \langle \mu^j, \tilde{v}_t \rangle = \frac{1}{Z_{t+1}} \sum_j \rho^j \Phi(-\frac{\tau^j}{\sigma^j}, \infty) \langle \mu^j, \tilde{v}_t \rangle
\]

from which the first part of the lemma follows.

The second part of the lemma follows from the observation that for any vector \( z \) orthogonal to \( \tilde{u}_t \) and \( \langle \mu^3, z \rangle = 0 \), for \( j = 1, 2 \).

**Lemma 24.** For any \( t \),

\[
\langle u_{t+1}, \mu^1 \rangle = \frac{||\mu^1||(|\xi_t \cos(\theta_t) + m_t|)}{Z_{t+1}} \quad ||u_{t+1}||^2 = \frac{\xi_t^2 + m_t^2 + 2\xi_t m_t \cos(\theta_t)}{(Z_{t+1})^2}
\]

*Proof. As we have an infinite number of samples, \( \theta_{t+1} \) lies on the same plane as \( \theta_t \). Therefore, we can write \( \langle u_{t+1}, \mu^1 \rangle = \langle u_{t+1}, \tilde{u}_t \rangle \langle \mu^1, \tilde{u}_t \rangle + \langle u_{t+1}, \tilde{v}_t \rangle \langle \mu^1, \tilde{v}_t \rangle \). Moreover, we can write \( ||u_{t+1}||^2 = \langle u_{t+1}, \tilde{u}_t \rangle^2 + \langle u_{t+1}, \tilde{v}_t \rangle^2 \). Thus, the first two equation follow by using Lemma 22 and 23 and recalling that \( \langle \mu^1, \tilde{u}_t \rangle = \tau_1^1 = ||\mu^1|| \cos(\theta_t) \) and \( \langle \mu^1, \tilde{v}_t \rangle = ||\mu^1|| \sin(\theta_t) \).

We are now ready to complete the proof of Lemma 1.

*Proof. (Of Lemma 1) By definition of \( \theta_{t+1} \), \( \cos^2(\theta_{t+1}) = \frac{\langle u_{t+1}, \mu^1 \rangle^2}{||u_{t+1}||^2 ||\mu^1||^2} \). Therefore,

\[
||\mu^1||^2 \cos^2(\theta_{t+1}) = \frac{\langle u_{t+1}, \mu^1 \rangle^2}{||u_{t+1}||^2} = (\tau_1^1)^2 \left( 1 + \frac{\langle u_{t+1}, \mu^1 \rangle^2 - ||\mu^1||^2 \cos^2(\theta_t) ||u_{t+1}||^2}{||\mu^1||^2 \cos^2(\theta_t) ||u_{t+1}||^2} \right)
\]

\[
= (\tau_1^1)^2 \left( 1 + \frac{||\mu^1||^2 \sin^2(\theta_t)(m_t^2 + 2\xi_t m_t \cos(\theta_t))}{||\mu^1||^2 \cos^2(\theta_t) ||u_{t+1}||^2} \right)
\]

\[
= ||\mu^1||^2 \cos^2(\theta_t) \left( 1 + \tan^2(\theta_t) \frac{m_t^2 + 2\xi_t m_t \cos(\theta_t)}{||u_{t+1}||^2} \right)
\]

where we used Lemma 24 and the observation that \( \cos(\theta_t) = \frac{\tau_1^1}{||\mu^2||} \). The Lemma follows by replacing \( ||u_{t+1}||^2 \) using the expression in Lemma 24.

The next Lemma helps us to derive Theorem 2 from Lemma 1. It shows how to approximate \( \Phi(-\tau, \tau) \) when \( \tau \) is small.

**Lemma 25.** Let \( \tau \leq \sqrt{\ln \frac{9}{2\pi}} \). Then, \( \frac{5}{3\sqrt{2\pi}} \tau \leq \Phi(-\tau, \tau) \leq \frac{2}{\sqrt{2\pi}} \tau \). In addition, \( \frac{2e^{-\tau^2/2}}{\sqrt{2\pi}} \geq \frac{2}{3} \).
6.2 Proofs of Sample Requirement Bounds

For the rest of the section, we prove Lemmas 8 and 9, which lead to a proof of Lemma 7. First, we need to define some notation.

**Notation.** At time $t$, we use the notation $S_{t+1}$ to denote the quantity $E[X \cdot 1_{X \in C_{t+1}}]$, where $1_{X \in C_{t+1}}$ is the indicator function for the event $X \in C_{t+1}$, and the expectation is taken over the entire mixture.

In the sequel, we also use the notation $\hat{S}_{t+1}$ to denote the empirical value of $S_{t+1}$. Our goal is to bound the concentration of certain functions of $\hat{S}_{t+1}$ around their expected values, when we are given only $n$ samples from the mixture. Recall that we define $\theta_{t+1}$ as the angle between $\mu^i$ and the hyperplane separator in 2-means-iterate, given $\theta_t$. Notice that now $\theta_t$ is a random variable, which depends on the samples drawn in rounds $1, \ldots, t-1$, and given $\theta_t$, $\theta_{t+1}$ is a random variable, whose value depends on samples in round $t$. Also we use $u_{t+1}$ as the center of partition $C_t$ in iteration $t+1$, and $E[u_{t+1}]$ is the expected center. Note that all the expectations in round $t$ are conditioned on $\theta_t$.

**Proofs.** We are now ready to prove Lemmas 8 and 9.

**Proof.** (Of Lemma 8) Let $X_1, \ldots, X_n$ be the iid samples from the mixture; for each $i$, we can write the projection of $X_i$ along $v$ as follows:

$$\langle X_i, v \rangle = Y_i + Z_i$$

where $Z_i \sim N(0, \sigma^2)$, if $X_i$ is generated from distribution $D^j$, and $Y_i = \langle \mu^j, v \rangle$, if $X_i$ is generated by $D^j$. Therefore, we can write:

$$\langle \hat{S}_{t+1}, v \rangle = \frac{1}{n} \left( \sum_i Y_i \cdot 1_{X_i \in C_{t+1}} + \sum_i Z_i \cdot 1_{X_i \in C_{t+1}} \right)$$

To determine the concentration of $\langle \hat{S}_{t+1}, v \rangle$ around its expected value, we address the two terms separately.

The first term is a sum of $n$ independently distributed random variables, such that changing one variable changes the sum by at most $\max_j 2\|\mu^j, v\|/n$; therefore, to calculate its concentration, one can apply Hoeffding’s Inequality. It follows that with probability at most $\frac{4}{2}$,

$$\frac{1}{n} \sum_i Y_i \cdot 1_{X_i \in C_{t+1}} - E\left[\frac{1}{n} \sum_i Y_i \cdot 1_{X_i \in C_{t+1}}\right] > \max_j \frac{4\|\mu^j, v\|\sqrt{\log(4n/\delta)}}{\sqrt{n}}$$

We note that, in the second term, each $Z_i$ is a Gaussian with mean 0 and variance $\sigma^2$, scaled by $\|v\|$. For some $0 \leq \delta' \leq 1$, let $E_i(\delta')$ denote the event

$$-\sigma_{\text{max}}\|v\|\sqrt{2\log(1/\delta')} \leq Z_i \cdot 1_{X_i \in C_{t+1}} \leq \sigma_{\text{max}}\|v\|\sqrt{2\log(1/\delta')}$$

As $Z_i \sim N(0, \sigma^2)$, if $X_i$ is generated from distribution $D_j$, and $1_{X_i \in C_{t+1}}$ takes values 0 and 1, for any $i$, for $\delta'$ small enough, $E_i(\delta') \geq 1 - \delta'$. We also observe that, as the Gaussians $Z_i$ are independently distributed, conditioned on the union of the events $E_i$, the Gaussians $Z_i$ are still independent. Therefore, conditioned on the event $\cup_i E_i(\delta')$,
\[ \frac{1}{n} \sum Z_i \cdot 1_{X_i \in C_{t+1}} \] is the sum of \( n \) independent random variables, such that changing one variable changes the sum by at most \( 2\sigma_{\max}||v||\sqrt{2\log(1/\delta')/n} \). We can now apply Hoeffding’s bound to conclude that with probability at least \( 1 - \frac{\delta}{2} \),

\[
\left| \frac{1}{n} \sum Z_i \cdot 1_{X_i \in C_{t+1}} - \mathbb{E} \left[ \frac{1}{n} \sum Z_i \cdot 1_{X_i \in C_{t+1}} \right] \right| \leq \frac{4\sigma_{\max}||v||\sqrt{2\log(1/\delta')/n}}{\sqrt{n}} \leq \frac{8\sigma_{\max}||v||\log(4n/\delta)}{\sqrt{n}}
\]

The lemma now follows by applying an union bound.

**Proof.** (Of Lemma 9) We can write:

\[
||\hat{S}_{t+1}||^2 \leq ||S_{t+1}||^2 + ||\hat{S}_{t+1} - S_{t+1}||^2 + 2||\hat{S}_{t+1} - S_{t+1}, S_{t+1}||
\]

If \( v_1, \ldots, v_d \) is any orthonormal basis of \( \mathbb{R}^d \), then, we can bound the second term as follows. With probability at least \( 1 - \frac{\delta}{2} \),

\[
||\hat{S}_{t+1} - S_{t+1}||^2 = \sum_{i=1}^{d} ((\hat{S}_{t+1} - S_{t+1}, v_i))^2 \leq \frac{128\log^2(8n/\delta)}{n} (\sigma_{\max}^2 ||v_i||^2 + \sum_{i,j} (\mu_j, v_i)^2)
\]

\[
\leq \frac{128\log^2(8n/\delta)}{n} (\sigma_{\max}^2 d + \sum_{j} (\mu_j)^2)
\]

The second step follows by the application of Lemma 8 and the fact that for any \( a \) and \( b \), \( (a + b)^2 \leq 2(a^2 + b^2) \).

Using Lemma 8 with probability at least \( 1 - \frac{\delta}{2} \),

\[
\langle \hat{S}_{t+1} - S_{t+1}, S_{t+1} \rangle \leq \frac{8\log(8n/\delta)}{\sqrt{n}} (\sigma_{\max} ||S_{t+1}|| + \max_j ||S_{t+1}, \mu_j||)
\]

The lemma follows by a union bound over these two above events.

### 6.3 Proofs of Lower Bounds

**Proof.** (Of Lemma 17) Let \( P \) be the plane containing the origin \( O \) and the vectors \( \mu_1 \) and \( \mu_2 \). If \( v \) is a vector orthogonal to \( P \), then, the projection of \( D_1 \) along \( v \) is a Gaussian \( \mathcal{N}(0,1) \), which is distributed independently of the projection of \( D_1 \) along \( P \) (and same is the case for \( D_2 \)). Therefore, to compute the KL-Divergence of \( D_1 \) and \( D_2 \), it is sufficient to compute the KL-Divergence of the projections of \( D_1 \) and \( D_2 \) along the plane \( P \).

Let \( x \) be a vector in \( P \). Then,

\[
\text{KL}(D_1, D_2) = \frac{1}{\sqrt{2\pi}} \int_{x \in P} \left( \frac{1}{2} e^{-||x - \mu_1||^2/2} + \frac{1}{2} e^{-||x + \mu_1||^2/2} \right) \ln \left( \frac{\frac{1}{2} e^{-||x - \mu_1||^2/2} + \frac{1}{2} e^{-||x + \mu_1||^2/2}}{\frac{1}{2} e^{-||x - \mu_2||^2/2} + \frac{1}{2} e^{-||x + \mu_2||^2/2}} \right) dx
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{x \in P} \left( \frac{1}{2} e^{-||x - \mu_1||^2/2} + \frac{1}{2} e^{-||x + \mu_1||^2/2} \right) \ln \left( \frac{e^{-||x + \mu_1||^2/2} \cdot (1 + e^{2(x, \mu_1)})}{e^{-||x + \mu_2||^2/2} \cdot (1 + e^{2(x, \mu_2)})} \right) dx
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{x \in P} \left( \frac{1}{2} e^{-||x - \mu_1||^2/2} + \frac{1}{2} e^{-||x + \mu_1||^2/2} \right) \left( ||x + \mu_2||^2 - ||x + \mu_1||^2 \right) \ln \left( 1 + \frac{e^{2(x, \mu_1)}}{1 + e^{2(x, \mu_2)}} \right) dx
\]
We observe that for any \( x \), \( \| x + \mu_2 \|^2 - \| x + \mu_1 \|^2 = \| \mu_2 \|^2 - \| \mu_1 \|^2 + 2 \langle x, \mu_2 - \mu_1 \rangle \). As the expected value of \( D_1 \) is 0, we can write that:

\[
\int_{x \in \mathbb{P}} \left( \frac{1}{2} e^{-\|x-\mu_1\|^2/2} + \frac{1}{2} e^{-\|x+\mu_1\|^2/2} \right) \langle x, \mu_2 - \mu_1 \rangle = \mathbb{E}_{x \sim D_1} \langle x, \mu_1 - \mu_2 \rangle = 0 \quad (1)
\]

We now focus on the case where \( \| \mu_1 \| >> 1 \). We observe that for any \( \mu_2 \) and any \( x, 1+e^{2\langle x, \mu_2 \rangle} > 1 \). Therefore, combining the previous two equations,

\[
\text{KL}(D_1, D_2) \leq \frac{1}{\sqrt{2\pi}} \left( \| \mu_2 \|^2 - \| \mu_1 \|^2 + \int_{x \in \mathbb{P}} \left( \frac{1}{2} e^{-\|x-\mu_1\|^2/2} + \frac{1}{2} e^{-\|x+\mu_1\|^2/2} \right) \ln(1 + e^{2\langle x, \mu_1 \rangle}) \, dx \right)
\]

Again, since the projection of \( D_1 \) perpendicular to \( \mu_1 \) is distributed independently of the projection of \( D_1 \) along \( \mu_1 \), the above integral can be taken over a one-dimensional \( x \) which varies along the vector \( \mu_1 \). For the rest of the proof, we abuse notation, and use \( \mu_1 \) to denote both the vector \( \mu_1 \) and the scalar \( \| \mu_1 \| \). We can write:

\[
\int_{x=-\infty}^{\infty} \left( \frac{1}{2} e^{-(x-\mu_1)^2/2} + \frac{1}{2} e^{-(x+\mu_1)^2/2} \right) \ln(1 + e^{2\mu_1 x}) \, dx
\]

\[
\leq \sqrt{2\pi} \ln 2 + \int_{x=0}^{\infty} \left( \frac{1}{2} e^{-(x-\mu_1)^2/2} + \frac{1}{2} e^{-(x+\mu_1)^2/2} \right) \ln(1 + e^{2\mu_1 x}) \, dx
\]

\[
\leq \sqrt{2\pi} \ln 2 + \int_{x=0}^{\infty} \left( \frac{1}{2} e^{-(x-\mu_1)^2/2} + \frac{1}{2} e^{-(x+\mu_1)^2/2} \right) \ln(2 + 2x\mu_1) \, dx
\]

\[
\leq \frac{3\sqrt{2\pi}}{2} \ln 2 + 2\mu_1 \int_{x=0}^{\infty} \left( \frac{1}{2} e^{-(x-\mu_1)^2/2} + \frac{1}{2} e^{-(x+\mu_1)^2/2} \right) x \, dx
\]

The first part follows because for \( x < 0 \), \( \ln(1 + e^{2\mu_1 x}) \leq \ln 2 \). The second part follows because for \( x > 0 \), \( \ln(1 + e^{2\mu_1 x}) \leq \ln(2e^{2\mu_1 x}) \). The third part follows from the symmetry of \( D_1 \) around the origin.

Now, for any \( a \), we can write:

\[
\frac{1}{\sqrt{2\pi}} \int_{x=0}^{\infty} xe^{-(x+a)^2/2} \, dx = \frac{1}{\sqrt{2\pi}} \cdot e^{-a^2/2} - a\Phi(a, \infty)
\]

Plugging this in, we can show that,

\[
\text{KL}(D_1, D_2) \leq \frac{1}{\sqrt{2\pi}} \left( \| \mu_2 \|^2 - \| \mu_1 \|^2 + \frac{3\sqrt{2\pi}}{2} \ln 2 + 2\| \mu_1 \| (e^{-\| \mu_1 \|^2/2} + \sqrt{2\pi}\| \mu_1 \| \Phi(0, \| \mu_1 \|)) \right)
\]

from which the lemma follows.

**Proof.** (Of Lemma 18) For each \( i \), let each \( v_i \) be drawn independently from the distribution \( \chi^2d/2 \mathcal{N}(0, I_d) \). For each \( i, j \), let \( P_{ij} = \frac{d}{2} \cdot d(v_i, v_j) \) and \( N_{ij} = \frac{d}{2} \cdot d(v_i, -v_j) \). Then, for each \( i \) and \( j \), \( P_{ij} \) and \( N_{ij} \) are distributed according to the Chi-squared distribution with parameter \( d \). From Lemma 26 it follows that: \( \Pr[P_{ij} < \frac{d}{2}] \leq e^{-3d/10} \). A similar lemma can also be shown to hold for the random variables \( N_{ij} \). Applying the Union Bound, the probability that this holds for \( P_{ij} \) and \( N_{ij} \) for all pairs \( (i, j), i \in V, j \in V \) is at most \( 2K^2e^{-3d/10} \). This probability is at most \( \frac{1}{2} \) when \( K = e^{d/10} \).

In addition, we observe that for each vector \( v_i \), \( d \cdot \| v_i \|^2 \) is also distributed as a Chi-squared distribution with parameter \( d \). From Lemma 26 for each \( i \), \( \Pr[\| v_i \|^2 > 7/5] \leq e^{-2d/15} \). The second part of the lemma now follows by an Union Bound over all \( K \) vectors in the set \( V \).
Lemma 26. Let $X$ be a random variable, drawn from the Chi-squared distribution with parameter $d$. Then,

$$\Pr[X < \frac{d}{10}] \leq e^{-3d/10}$$

Moreover,

$$\Pr[X > \frac{7d}{5}] \leq e^{-2d/15}$$

Proof. Let $Y$ be the random variable defined as follows: $Y = d - X$. Then,

$$\Pr[X < \frac{d}{10}] = \Pr[Y > \frac{9d}{10}] = \Pr[e^{tY} > e^{9dt/10}] \leq \frac{E[e^{tY}]}{e^{9dt/10}}$$

where the last step uses a Markov’s Inequality. We observe that $E[e^{tY}] = e^{td}E[e^{-tX}] = e^{td}(1 - 2t)^{d/2}$, for $t < \frac{1}{2}$. The first part of the lemma follows from the observation that $(1 - 2t)^{d/2} \leq e^{-td}$, and by plugging in $t = \frac{1}{3}$.

For the second part, we again observe that

$$\Pr[X > \frac{7d}{5}] \leq (1 - 2t)^{-d/2}e^{-7dt/5} \leq e^{-2dt/5}$$

The lemma now follows by plugging in $t = \frac{1}{3}$.

6.4 More General $k$-means: Results and Proofs

In this section, we show that when we apply 2-means on an input generated by a mixture of $k$ spherical Gaussians, the normal to the hyperplane which partitions the two clusters in the 2-means algorithm, converges to a vector in the subspace $\mathcal{M}$ containing the means of mixture components. This subspace is interesting because, in this subspace, the distance between the means is as high as in the original space; however, if the number of clusters is small compared to the dimension, the distance between two samples from the same cluster is much smaller. In fact, several algorithms for learning mixture models [VW02, AM05, CR08] attempt to isolate this subspace first, and then use some simple clustering methods in this subspace.

6.4.1 The Setting

We assume that our input is generated by a mixture of $k$ spherical Gaussians, with means $\mu^j$, variances $(\sigma^j)^2$, $j = 1, \ldots, k$, and mixing weights $\rho^1, \ldots, \rho^k$. The mixture is centered at the origin such that $\sum \rho^j \mu^j = 0$. We use $\mathcal{M}$ to denote the subspace containing the means $\mu^1, \ldots, \mu^k$.

We use Algorithm 2-means-iterate on this input, and our goal is to show that it still converges to a vector in $\mathcal{M}$.

In the sequel, given a vector $x$ and a subspace $W$, we define the angle between $x$ and $W$ as the angle between $x$ and the projection of $x$ onto $W$. As in Sections 2 and 3, we examine the angle $\theta_t$, between $u_t$ and $\mathcal{M}$, and our goal is to show that the cosine of this angle grows as $t$ increases. Our main result of this section is Lemma 20, which, analogous to Lemma 1 in Section 3, exactly defines the behavior of 2-means on a mixture of $k$ spherical Gaussians.

Before we can prove the lemma, we need some additional notation.
6.4.2 Notation

Recall that at time $t$, we use $\tilde{u}_t$ to partition the input data, and the projection of $\tilde{u}_t$ along $\mathcal{M}$ is $\cos(\theta_t)$ by definition. Let $b^1_t$ be a unit vector lying in the subspace $\mathcal{M}$ such that:

$$\tilde{u}_t = \cos(\theta_t) b^1_t + \sin(\theta_t) v_t$$

where $v_t$ lies in the orthogonal complement of $\mathcal{M}$, and has norm 1. We define a second vector $\tilde{u}^\perp_t$ as follows:

$$\tilde{u}^\perp_t = \sin(\theta_t) b^1_t - \cos(\theta_t) v_t$$

We observe that $\langle \tilde{u}_t, \tilde{u}^\perp_t \rangle = 0$, $||\tilde{u}^\perp_t|| = 1$, and the projection of $\tilde{u}^\perp_t$ on $\mathcal{M}$ is $\sin(\theta_t) b^1_t$.

We now extend the set $\{b^1_t\}$ to complete an orthonormal basis $\mathcal{B} = \{b^1_t, \ldots, b^{k-1}_t\}$ of $\mathcal{M}$. We also observe that $\{b^2_t, \ldots, b^{k-1}_t, \tilde{u}_t, \tilde{u}^\perp_t\}$ is an orthonormal basis of the subspace spanned by any basis of $\mathcal{M}$, along with $v_t$, and can be extended to a basis of $\mathbb{R}^d$.

For $j = 1, \ldots, k$, we define $\tau^j_t$ as follows:

$$\tau^j_t = \langle \mu^j, \tilde{u}_t \rangle = \cos(\theta_t) \langle \mu^j, b^1_t \rangle$$

Finally we (re)-define the quantity $\xi_t$ as

$$\xi_t = \sum_j \rho^j \sigma^j \frac{e^{-\left(\tau^j_t\right)^2/2\sigma^j_t^2}}{\sqrt{2\pi}}$$

and, for any $l = 1, \ldots, k - 1$, we define:

$$m^l_t = \sum_j \rho^j \Phi\left(-\frac{\tau^j_t}{\sigma^j_t}, \infty\right) \langle \mu^j, b^l_t \rangle$$

6.4.3 Proof of Lemma 20

The main idea behind the proof of Lemma 20 is to estimate the norm and the projection of $u_{t+1}$; we do this in three steps. First, we estimate the projection of $u_{t+1}$ along $\tilde{u}_t$; next, we estimate this projection on $\tilde{u}^\perp_t$, and finally, we estimate its projection along $b^2_t, \ldots, b^l_t$. Combining these projections, and observing that the projection of $u_{t+1}$ on any direction perpendicular to these is 0, we can prove the lemma.

As before, we define

$$Z_{t+1} = \Pr[x \in C_{t+1}]$$

Now we make the following claim.

**Lemma 27.** For any $t$ and any $j$,

$$\Pr[x \sim D_j|x \in C_{t+1}] = \frac{\rho^j}{Z_{t+1}} \Phi\left(-\frac{\tau^j_t}{\sigma^j_t}, \infty\right)$$

**Proof.** Same proof of Lemma 21

Next, we estimate the projection of $u_{t+1}$ along $\tilde{u}_t$.

**Lemma 28.**

$$\langle u_{t+1}, \tilde{u}_t \rangle = \frac{\xi_t + \cos(\theta_t) m^1_t}{Z_{t+1}}$$
Proof. Consider a sample \( x \) drawn from distribution \( D_j \). The projection of \( x \) on \( \tilde{u}_t \) is distributed as a Gaussian with mean \( \tau^j_t \) and standard deviation \( \sigma^j \). The probability that \( x \) lies in \( C_{t+1} \) is \( \Pr[N(\tau^j_t, \sigma^j) > 0] = \Phi(-\frac{\tau^j_t}{\sigma^j}, \infty) \). Given that \( x \) lies in \( C_{t+1} \), the projection of \( x \) on \( \tilde{u}_t \) is distributed as a truncated Gaussian, with mean \( \tau^j_t \) and standard deviation \( \sigma^j \), which is truncated at 0. Therefore,

\[
E[(x, \tilde{u}_t)|x \in C_{t+1}, x \sim D_j] = \frac{1}{\Phi(-\frac{\tau^j_t}{\sigma^j}, \infty)} \left( \int_{y=0}^{\infty} \frac{ye^{-(y-\tau^j_t)^2/2}}{\sigma^j \sqrt{2\pi}} dy + \int_{y=0}^{\infty} \frac{(y - \tau^j_t)e^{-(y-\tau^j_t)^2/2}}{\sigma^j \sqrt{2\pi}} dy \right)
\]

which is again equal to

\[
\frac{1}{\Phi(-\frac{\tau^j_t}{\sigma^j}, \infty)} \left( \tau^j_t \Phi(-\frac{\tau^j_t}{\sigma^j}, \infty) + \int_{y=0}^{\infty} \frac{(y - \tau^j_t)e^{-(y-\tau^j_t)^2/2}}{\sigma^j \sqrt{2\pi}} dy \right)
\]

We can evaluate the integral in the equation above as follows.

\[
\int_{y=0}^{\infty} (y - \tau^j_t)e^{-(y-\tau^j_t)^2/2} dy = (\sigma^j)^2 \int_{z=0}^{\infty} e^{-2} dz = (\sigma^j)^2 e^{-(\tau^j_t)^2/2}
\]

Therefore we can conclude that

\[
E[(x, \tilde{u}_t)|x \in C_{t+1}, x \sim D_j] = \tau^j_t + \frac{1}{\Phi(-\frac{\tau^j_t}{\sigma^j}, \infty)} \cdot \sigma^j \frac{e^{-(\tau^j_t)^2/2}}{\sqrt{2\pi}}
\]

Now we can write

\[
\langle u_{t+1}, \tilde{u}_t \rangle = \sum_j E[(x, \tilde{u}_t)|x \sim D_j, x \in C_{t+1}] \Pr[x \sim D_j|x \in C_{t+1}]
\]

\[
= \frac{1}{Z_{t+1}} \sum_j \rho^j \Phi(-\frac{\tau^j_t}{\sigma^j}, \infty) E[(x, \tilde{u}_t)|x \sim D_j, x \in C_{t+1}]
\]

where we used lemma [27]. The lemma follows by recalling that \( \tau^j_t = \cos(\theta_t)\langle \mu^j, b^j_t \rangle \).

Lemma 29. For any \( t \),

\[
\langle u_{t+1}, \tilde{u}^1_t \rangle = \frac{\sin(\theta_t)m^1_t}{Z_{t+1}}
\]

Proof. Let \( x \) be a sample drawn from distribution \( D_j \). Since \( \tilde{u}^1_t \) is perpendicular to \( \tilde{u}_t \), and \( D_j \) is a spherical Gaussian, given that \( x \in C_{t+1} \), that is, the projection of \( x \) on \( \tilde{u}_t \) is greater than 0, the projection of \( x \) on \( \tilde{u}^1_t \) is still distributed as a Gaussian with mean \( \langle \mu^j, \tilde{u}^1_t \rangle \) and standard deviation \( \sigma^j \). That is,

\[
E[(x, \tilde{u}^1_t)|x \sim D_j, x \in C_{t+1}] = \langle \mu^j, \tilde{u}^1_t \rangle
\]

Also recall that, by definition of \( \tilde{u}^1_t \), \( \langle \mu^j, \tilde{u}^1_t \rangle = \sin(\theta_t)\langle \mu^j, b^1_t \rangle \). To prove the lemma, we observe that \( \langle u_{t+1}, \tilde{u}^1_t \rangle \) is equal to

\[
\sum_j E[(x, \tilde{u}^1_t)|x \sim D_j, x \in C_{t+1}] \Pr[x \sim D_j|x \in C_{t+1}]
\]

The lemma follows by using lemma [27].
Lemma 30. For \( l \geq 2 \),
\[
\langle u_{t+1}, b^l_t \rangle = \frac{m^l_t}{Z_{t+1}}
\]

Proof. Let \( x \) be a sample drawn from distribution \( D_j \). Since \( b^l_t \) is perpendicular to \( \hat{u}_t \), and \( D_j \) is a spherical Gaussian, given that \( x \in C_{t+1} \), that is, the projection of \( x \) on \( \hat{u}_t \) is greater than 0, the projection of \( x \) on \( b^l_t \) is still distributed as a Gaussian with mean \( \langle \mu^l, b^l_t \rangle \) and standard deviation \( \sigma^l \). That is,
\[
E[\langle x, b^l_t \rangle | x \sim D_j, x \in C_{t+1}] = \langle \mu^l, b^l_t \rangle
\]

To prove the lemma, we observe that \( \langle b^l_t, u_{t+1} \rangle \) is equal to
\[
\sum_j E[\langle x, b^l_t \rangle | x \sim D_j, x \in C_{t+1}] \Pr[x \sim D_j | x \in C_{t+1}]
\]
The lemma follows by using lemma 27.

Finally, we show a lemma which estimates the norm of the vector \( u_{t+1} \).

Lemma 31.
\[
||u_{t+1}||^2 = \frac{1}{Z_{t+1}^2}(\xi_t^2 + 2\xi_t \cos(\theta_t)m^1_t + \sum_{l=1}^k (m^l_t)^2)
\]

Proof. Combining Lemmas 28, 29 and 30, we can write:
\[
||u_{t+1}||^2 = \langle \hat{u}_t, u_{t+1} \rangle^2 + \langle \hat{u}^\perp_t, u_{t+1} \rangle^2 + \sum_{l \geq 2} \langle b^l_t, u_{t+1} \rangle^2
\]
\[
= \frac{1}{Z_{t+1}^2}(\xi_t^2 + 2\xi_t \cos(\theta_t)m^1_t + \cos^2(\theta_t)(m^1_t)^2 + \sin^2(\theta_t)(m^1_t)^2 + \sum_{l=2}^k (m^l_t)^2)
\]
The lemma follows by plugging in the fact that \( \cos^2(\theta_t) + \sin^2(\theta_t) = 1 \).

Now we are ready to prove Lemma 20.

Proof. (Of Lemma 20) Since \( b^1_t, \ldots, b^k_t \) form a basis of \( M \), we can write:
\[
\cos^2(\theta_{t+1}) = \frac{\sum_{i=1}^k \langle u_{t+1}, b^i_t \rangle^2}{||u_{t+1}||^2}
\]

\( ||u_{t+1}||^2 \) is estimated in Lemma 31 and \( \langle u_{t+1}, b^i_t \rangle \) is estimated by Lemma 29. Using these lemmas, as \( b^l_t \) lies in the subspace spanned by the orthogonal vectors \( \hat{u}_t \) and \( \hat{u}^\perp_t \), we can write:
\[
\langle u_{t+1}, b^1_t \rangle = \langle \hat{u}_t, u_{t+1} \rangle \langle \hat{u}_t, b^1_t \rangle + \langle \hat{u}^\perp_t, u_{t+1} \rangle \langle \hat{u}^\perp_t, b^1_t \rangle
\]
\[
= \frac{\cos(\theta_t)\xi_t + m^1_t}{Z_{t+1}}
\]
Plugging this in to Equation 2, we get:
\[
\cos^2(\theta_{t+1}) = \frac{\xi_t^2 \cos^2(\theta_t) + 2\xi_t \cos(\theta_t)m^1_t + \sum_l(m^l_t)^2}{\xi_t^2 + 2\xi_t \cos(\theta_t)m^1_t + \sum_l(m^l_t)^2}
\]
The lemma follows by rearranging the above equation, similar to the proof of Lemma 1.