Clustering with Spectral Norm and the k-means Algorithm

Amit Kumar*
Dept. of Computer Science and Engg.
IIT Delhi
email : amitk@cse.iitd.ac.in

Ravindran Kannan
Microsoft Research India Lab.
email : kannan@microsoft.com

***Abstract*—There has been much progress on efficient algorithms for clustering data points generated by a mixture of k probability distributions under the assumption that the means of the distributions are well-separated, i.e., the distance between the means of any two distributions is at least $\Omega(k)$ standard deviations. These results generally make heavy use of the generative model and particular properties of the distributions. In this paper, we show that a simple clustering algorithm works without assuming any generative (probabilistic) model. Our only assumption is what we call a "proximity condition": the projection of any data point onto the line joining its cluster center to any other cluster center is $\Omega(k)$ standard deviations closer to its own center than the other center. Here the notion of standard deviations is based on the spectral norm of the matrix whose rows represent the difference between a point and the mean of the cluster to which it belongs. We show that in the generative models studied, our proximity condition is satisfied and so we are able to derive most known results for generative models as corollaries of our main result. We also prove some new results for generative models - e.g., we can cluster all but a small fraction of points only assuming a bound on the variance. Our algorithm relies on the well known k-means algorithm, and along the way, we prove a result of independent interest – that the k-means algorithm converges to the "true centers" even in the presence of spurious points provided the initial (estimated) centers are close enough to the corresponding actual centers and all but a small fraction of the points satisfy the proximity condition. Finally, we present a new technique for boosting the ratio of inter-center separation to standard deviation. This allows us to prove results for learning certain mixture of distributions under weaker separation conditions.**

I. Introduction

Clustering is in general a hard problem. But, there has been a lot of research (see Section III for references) on proving that if we have data points generated by a mixture of k probability distributions, then one can cluster the data points into the k clusters, one corresponding to each component, provided the means of the different components are well-separated. There are different notions of well-separated, but mainly, the (best known) results can be qualitatively stated as:

"If the means of every pair of densities are at least $\mathrm{poly}(k)$ times standard deviations apart, then we can learn the mixture in polynomial time."

These results generally make heavy use of the generative model and particular properties of the distributions (indeed, many of them specialize to Gaussians or independent Bernoulli trials). In this paper, we make no assumptions on the generative model of the data. We are still able to derive essentially the same result (loosely stated for now as):

"If the projection of any data point onto the line joining its cluster center to any other cluster center is $\Omega(k)$ times standard deviations closer to its own center than the other center (we call this the "proximity condition"), then we can cluster correctly in polynomial time."

First, if the n points to be clustered form the rows of an $n \times d$ matrix A and C is the corresponding matrix of cluster centers (so each row of C is one of k vectors, namely the centers of k clusters) then note that the maximum directional variance (no probabilities here, the variance is just the average squared distance from the center) of the data in any direction is just

$$\frac{1}{n} \cdot \mathrm{Max}_{v:|v|=1} |(A-C) \cdot v|^2 = \frac{||A-C||^2}{n},$$

where $||A-C||$ is the spectral norm. So, spectral norm scaled by $1/\sqrt{n}$ will play the role of standard deviation in the above assertion. To our knowledge, this is the first result proving that clustering can be done in polynomial time in a general situation with only deterministic assumptions. It settles an open question raised in [1].

We will show that in the generative models studied, our proximity condition is satisfied and so we are able to derive all known results for generative models as corollaries of our theorem (with one qualification: whereas our separation is in terms of the whole data variance, often, in the case of Gaussians, one can make do with separations depending only on individual densities' variances – see Section III.)

Besides Gaussians, the planted partition model (defined later) has also been studied; both these distributions have very "thin tails" and a lot of independence, so one can appeal to concentration results. In section VI-C, we give a clustering algorithm for a mixture of general densities for which we only assume bounds on the variance (and no further concentration). Based on our algorithm, we show how to classify all but an ε fraction of points in this model. Section III has references to recent work dealing with

*This work was done while the author was visiting Microsoft India Research Lab.

distributions which may not even have bounded variance, but these results are for the special class of product densities, with additional constraints.

One crucial technical result we prove (Theorem 5.5) may be of independent interest. It shows that the good old $k-$means algorithm [2] converges to the "true centers" even in the presence of spurious points provided the initial (estimated) centers are close enough to the corresponding actual centers and all but an ε fraction of the points satisfy the proximity condition. Convergence (or lack of it) of the $k-$means algorithm is again well-studied ([3], [4], [5], [6]). The result of [3] (one of the few to formulate sufficient conditions for the $k-$means algorithm to provably work) assumes the condition that the optimal clustering with k centers is substantially better than that with fewer centers and shows that one iteration of the $k-$means algorithm yields a near-optimal solution. Note that if C represents the optimal k-means clustering, then its cost is $||A-C||_F^2$. We show in section VI-D that their condition implies a stronger condition than proximity, which involves $||A-C||_F$ instead of generally much smaller $||A-C||$, for all but an ε fraction of the points. This allows us to prove that our algorithm, which is again based on the $k-$means algorithm, gives a PTAS.

The proof of Theorem 5.5 is based on Theorem 5.4 which shows that if current centers are close to the true centers, then misclassified points (whose nearest current center is not the one closest to the true center) are far away from true centers and so there cannot be too many of them. This is based on a clean geometric argument shown pictorially in Figure 2. Our main theorem in addition allows for an ε fraction of "spurious" points which do not satisfy the proximity condition. Such errors have often proved difficult to account for.

As indicated, all results on generative models assume a lower bound on the inter-center separation in terms of the spectral norm. In section VII, we describe a construction (when data is from a generative model – a mixture of distributions) which boosts the ratio of inter-center separation to spectral norm. The construction is the following: we pick two sets of samples $A_1, A_2, \ldots A_n$ and $B_1, B_2, \ldots B_n$ independently from the mixture. We define new points $X_1, X_2, \ldots X_n$, where X_i is defined as $(A_i' \cdot B_1', A_i' \cdot B_2', \ldots A_i' \cdot B_n')$, where $'$ denotes that we have subtracted the mean (of the mixture.) Using this, we are able to reduce the dependence of inter-center separation on the minimum weight of a component in the mixture that all models generally need. This technique of boosting is likely to have other applications.

II. Preliminaries and the Main Theorem

For a matrix A, we shall use $||A||$ to denote its spectral norm. For a vector v, we use $|v|$ to denote its length. We are given n points in $\Re^d$ which are divided into k clusters – $T_1, T_2, \ldots, T_k$. Let μ_r denote the mean of cluster T_r and n_r denote $|T_r|$. Let A be the $n \times d$ matrix with rows corresponding to the points. Let C be the $n \times d$ matrix where $C_i = \mu_r$, for all $i \in T_r$. We shall use A_i to denote the i^{th} row of A. Let

$$\Delta_{rs} = \left(\frac{ck}{\sqrt{n_r}} + \frac{ck}{\sqrt{n_s}} \right) ||A-C||,$$

where c is a large enough constant.

Definition 2.1: We say a point $A_i \in T_r$ satisfies the *proximity condition* if for any $s \neq r$, the projection of A_i onto the μ_r to μ_s line is at least Δ_{rs} closer to μ_r than to μ_s. We let G (for good) be the set of points satisfying the proximity condition.

Note that the proximity condition implies that the distance between μ_r and μ_s must be at least Δ_{rs}. We are now ready to state the theorem.

Theorem 2.2: If $|G| \geq (1-\varepsilon) \cdot n$, then we can correctly classify all but $O(k^2 \varepsilon \cdot n)$ points in polynomial time. In particular, if $\varepsilon = 0$, all points are classified correctly.

III. Previous Work

Learning mixture of distributions is one of the central problems in machine learning. There is vast amount of literature on learning mixture of Gaussian distributions. One of the most popular methods for this is the well known EM algorithm which tries to maximize the log likelihood function [7]. However, there are few results which demonstrate that it converges to the optimal solution. Dasgupta [8] introduced the problem of learning distributions under suitable *separation conditions*, i.e., we assume that the distance between the means of the distributions in the mixture is large, and the goal is to recover the original clustering of points (perhaps with some error).

We first summarize known results for learning mixtures of Gaussian distributions under separation conditions. We assume there are k Gaussians $F_1, \ldots, F_k$ with mixing weights $w_1, \ldots, w_r$ and σ_r denotes the maximum variance along any direction of the distribution F_r. We ignore logarithmic factors in separation conditions. We also ignore the minimum number of samples required by various algorithms – they are often bounded by a polynomial in the dimension and the mixing weights. Dasgupta [8] gave an algorithm based on random projection to learn mixture of Gaussians provided mixing weights of all distributions are about the same, and $|\mu_i - \mu_j|$ is $\Omega((\sigma_i + \sigma_j) \cdot \sqrt{n})$. Dasgupta and Schulman [9] gave an EM based algorithm provided $|\mu_i - \mu_j|$ is $\Omega((\sigma_i + \sigma_j) \cdot n^{\frac{1}{4}})$. Arora and Kannan [10] generalized this to overlapping distributions as well as in other directions. Vempala and Wang [11] were the first to demonstrate the effectiveness of spectral techniques for learning mixture of distributions. For spherical Gaussians, their algorithm worked with a much weaker separation condition of $\Omega((\sigma_i + \sigma_j) \cdot k^{\frac{1}{4}})$ between μ_i and μ_j. Achlioptas and McSherry [12] extended this to arbitrary Gaussians with separation between μ_i and μ_j being at least $\Omega\left(\left(k + \frac{1}{\sqrt{\min(w_i,w_j)}}\right)\cdot(\sigma_i+\sigma_j)\right)$. Kannan et. al. [13] also gave an algorithm for arbitrary Gaussians with the corresponding separation being $\Omega\left(\frac{k^{\frac{3}{2}}}{w_{\min}^2}\cdot(\sigma_i+\sigma_j)\right)$. Recently, Brubaker and Vempala [14] gave a learning algorithm where the separation only depends on the variance perpendicular to a hyperplane separating two Gaussians, the so called "parallel pancakes problem" (see also the recent result of Vempala [15]).

Much less is known about learning mixtures of heavy tailed distributions. Most of the known results here assume that each distribution is a product distribution, i.e., projection along co-ordinate axes are independent. Often, they also assume some *slope condition* on the line joining any two means. Dasgupta et. al. [16] considered the problem of learning product distributions of heavy tailed distributions when each component distribution satisfied the following mild condition : $P[|X-\mu|\geq \alpha R]\leq \frac{1}{2\alpha}$. Here R is the half-radius of the distribution (these distributions can have unbounded variance). Their algorithm could classify at least $(1-\varepsilon)$ fraction of the points provided the distance between any two means is at least $\Omega\left(\frac{R\cdot k^{\frac{5}{2}}}{\varepsilon^2}\right)$. Under even milder assumptions on the distributions and a slope condition, they could correctly classify all but ε fraction of the points provided the corresponding separation was $\Omega\left(R\cdot\sqrt{\frac{k}{\varepsilon}}\right)$. Their algorithm, however, requires exponential (in d and k) amount of time. This problem was resolved by Chaudhuri and Rao [17]. Dasgupta et. al. [18] considered the problem of classifying samples from a mixture of arbitrary distributions with bounded variance σ in every direction. They showed that if the separation between the means is $\Omega(\sigma k)$ and a suitable slope condition holds, then all the samples can be correctly classified. Their paper also gives a general method for bounding the spectral norm of a matrix when the rows are independent (and some additional conditions hold). We will mention this condition formally in Section VI and make heavy use of it.

Finally, we discuss the *planted partition model* [19]. In this model, an instance consists of a set of n points, and there is an implicit partition of these n points into k groups. Further, there is an (unknown) $k\times k$ matrix of probabilities A. We are given a graph G on these n points, where an edge between two vertices from groups i and j is present with probability A_{ij}. The goal is to recover the actual partition of the points (and hence, an approximation to the matrix A as well). Mnemonically, this is the problem of finding $\mathbf{E}[A]$ given A, where $\mathbf{E}[A]$ denotes component-wise expectation. McSherry[19] showed that if there is a separation of at least

$$c\cdot\sigma^2\cdot k\cdot\left(\frac{1}{w_{\min}}+\log\frac{n}{\delta}\right), \qquad (1)$$

between any two distinct rows of $\mathbf{E}[A]$, then one can recover the actual partition of the vertex set with probability at least $1-\delta$. Here c is a large constant, $w_{\min}$ is such that every group has size at least $w_{\min}\cdot n$, and σ^2 denotes $\max_{i,j} A_{ij}$.

There has been some recent work on finding the *true* clustering based on deterministic assumptions. Balcan et. al. [20] consider this problem when the true clustering is fairly robust with respect to some objective function for clustering, i.e., any good approximation algorithm for this objective function will yield a clustering which is very close to the true clustering. This assumption, however, seems somewhat strong for application to learning mixture of distributions. Balcan et. al. [21] consider the problem of finding similarlity measures on points which can then yield the true clusering (or a small set of candidate clusterings). For us, however, the points are already embedded in the Euclidean space.

There is a rich body of work on the k-means problem and heuristic algorithms for this problem (see for example [22], [3] and references therein). One of the most widely used algorithms for this problem was given by Lloyd [2]. In this algorithm, we start with an arbitrary set of k candidate centers. Each point is assigned to the closest candidate center – this clusters the points into k clusters. For each cluster, we update the candidate center to the mean of the points in the cluster. This gives a new set of k candidate centers. This process is repeated till we get a local optimum. This algorithm may take superpolynomial time to converge [4]. However, there is a growing body of work on proving that this algorithm gives a good clustering in polynomial time if the initial choice of centers is good [23], [24], [3]. Ostrovsky et. al. [3] showed that a modification of the Lloyd's algorithm gives a PTAS for the k-means problem if there is a sufficiently large separation between the means. Our result also fits in this general theme – the k-means algorithm on a choice of centers obtained from a simple spectral algorithm classifies the point correctly.

IV. Our Contributions

Our main contribution is to show that a set of points satisfying the proximity condition can be correctly classified (Theorem 2.2). The algorithm is described in Figure 1. It has two main steps – first find an initial set of centers based on SVD, and then run the standard k-means algorithm with these initial centers as seeds. In Section V, we show that after each iteration of the k-means algorithm, the set of centers come a factor of two close to the true centers. Although both steps of our algorithm – SVD and the k-means algorithm – have been well studied, ours is the first result which shows that *combining* the two leads to a provably good algorithm. In Section VI, we give several applications of Theorem 2.2. We have the following results for learning mixture of distriutions (we ignore poly-logarithmic factors below):

- Arbitrary Gaussian Distributions with separation $\Omega\left(\frac{\sigma k}{\sqrt{w_{\min}}}\right)$: as mentioned above, this matches known results [12], [13] except for the fact that the separation condition between two distributions depends on the maximum standard deviation (as compared to standard deviations of these distributions only).

- Planted distribution model with separation $\Omega\left(\frac{k\sigma}{\sqrt{w_{\min}}}\right)$: this matches the result of McSherry [19] except for a $\sqrt{k}$ factor which we can also remove with a more careful analysis.
- Distributions with bounded variance along every direction : we can classify all but an ε fraction of points if the separation between means is at least $\Omega\left(\frac{k\sigma}{\sqrt{\varepsilon}}\right)$. Although results are known for classifying (all but a small fraction) points from mixtures of distributions with unbounded variance [16], [17], such results work for the special case of product distributions.
- PTAS using the k-means algorithm : We show that the separation condition of Ostrovsky et. al. [3] is stronger than the proximity condition. Using this fact, we are also able to give a PTAS based on the k-means algorithm.

Further, ours is the first algorithm which applies to all of the above settings. In Section VII, we give a general technique for working with weaker separation conditions (for learning mixture of distributions). Under certain technical conditions described in Section VII, we give a construction which increases the spectral norm of $A - C$ at a much faster rate than the increase in inter-mean distance as we increase the number of samples. As applications of this technique, we have the following results :

- Arbitrary Gaussians with separation $\Omega\left(\sigma k \cdot \log \frac{d}{w_{\min}}\right)$: this is the first result for arbitrary Gaussians where the separation depends only logarithmically on the minimum mixture weight.
- Power-law distributions with sufficiently large (but constant) exponent γ (defined in equation (11)) : We prove that we can learn all but ε fraction of samples provided the separation between means is $\Omega\left(\sigma k \cdot \left(\log \frac{d}{w_{\min}} + \frac{1}{\varepsilon^{\frac{1}{\gamma}}}\right)\right)$. For large values of γ, it significantly reduces the dependence on ε.

We expect this technique to have more applications.

V. Proof of Theorem 2.2

Our algorithm for correctly classifying the points will run in several iterations. At the beginning of each iteration, it will have a set of k candidate points. By a Lloyd like step, it will replace these points by another set of k points. This process will go on for polynomial number of steps.

The iterative procedure is described in Figure 1. In the first step, we can use any constant factor approximation algorithm for the k-means problem, e.g., the 9-approximation algorithm of Kanungo et. al. [25]. Note that our overall algorithm is same as Lloyd's algorithm, but we start with a special set of initial points. We now prove that after the first step (the base case), the estimated centers are close to the actual ones – this case follows from [1], but we prove it below for sake of completeness.

Lemma 5.1: (**Base Case**) After the first step of the algorithm `Cluster`, for every μ_r, there exists a center ν_r satisfying

$$|\mu_r - \nu_r| \leq 20\sqrt{k} \cdot \frac{||A - C||}{\sqrt{n_r}}.$$

Proof: Suppose, for sake of contradiction, that there exists an r such that all the centers $\nu_1, \ldots, \nu_k$ are at least $\frac{20\sqrt{k}\cdot||A-C||}{\sqrt{n_r}}$ distance away from μ_r. Consider the points in T_r. Suppose $A_i \in T_r$ is assigned to the center $\nu_{c(i)}$ in this solution. The assignment cost for these points in this k-means solution (which is a 9-approximation to the optimal solution) is

$$\sum_{i \in T_r} |\hat{A}_i - \nu_{c(i)}|^2 = \sum_{i \in T_r} |(\mu_r - \nu_{c(i)}) - (\mu_r - \hat{A}_i)|^2$$

$$\geq \quad \frac{|T_r|}{2} \cdot \left(\frac{20\sqrt{k} \cdot ||A - C||}{\sqrt{n_r}}\right)^2 - \sum_{i \in T_r} |\mu_r - \hat{A}_i|^2 \quad (2)$$

$$\geq \quad 195k||A - C||^2 \quad (3)$$

where inequality (2) follows from the fact that for any two numbers a, b, $(a-b)^2 \geq \frac{a^2}{2} - b^2$; and inequality (3) follows from the fact that $||\hat{A} - C||_F^2 \leq 5k \cdot ||A - C||^2$. But this is a contradiction, because one feasible solution to the k-means problem is to assign points in $\hat{A}_i, i \in T_s$ to μ_s for $s = 1, \ldots, k$ – the cost of this solution is $||\hat{A} - C||_F^2 \leq 5k||A - C||^2$. ■

Observe that the lemma above implies that there is a unique center ν_r associated with each μ_r (because the proximity conditions implies that there is enough separation between any two μ_r and μ_s). We now prove a useful lemma which states that removing small number of points from a cluster T_r can move the mean of the remaining points by only a small distance.

Lemma 5.2: Let X be a subset of T_r. Let $m(X)$ denote the mean of the points in X. Then

$$|m(X) - \mu_r| \leq \frac{||A - C||}{\sqrt{|X|}}.$$

Proof: Let u be unit vector along $m(X) - \mu_r$. Now,

$$|(A - C) \cdot u| \geq \left(\sum_{i \in X} ((A_i - \mu_r) \cdot u)^2\right)^{\frac{1}{2}}$$

$$\geq \frac{1}{\sqrt{|X|}} \left(\sum_{i \in X} |(A_i - \mu_r) \cdot u|\right) \geq \sqrt{|X|} \cdot |m(X) - \mu_r|$$

But, $|(A - C) \cdot u| \leq ||A - C||$. This proves the lemma. ■

Corollary 5.3: Let $Y \subseteq T_s$. Let $m(Y)$ denote the mean of the points in Y. Then, $|m(Y) - \mu_s| \leq \frac{\sqrt{T_s - |Y|} \cdot ||A-C||}{|Y|}$.

```
1) (Base case) Let $\hat{A}_i$ be the projection of the points on the best $k$-dimensional subspace found by computing SVD
   of $A$. Let $\nu_r, 1 \le r \le k$, denote the centers of a (near)-optimal solution to the $k$-means problem for the points $\hat{A}_i$.
2) Repeat
   (i) Assign each point $A_i$ to the closest point among $\nu_r, r = 1, \dots, k$. Let $S_r$ be the set of points assigned to $\nu_r$.
   (ii) Define $\eta_r$ as the mean of the points $S_r$. Update $\eta_r$, $r = 1, \dots, k$, as the new centers, i.e., set $\nu_r = \eta_r$ for the
        next iteration.
```

Figure 1. Algorithm `Cluster`

Proof: Let X denote $T_s - Y$. We know that $\mu_s \cdot |T_s| = |X| \cdot m(X) + |Y| \cdot m(Y)$. So we get

$$|m(Y) - \mu_s| \;=\; \frac{|X|}{|Y|} \cdot |m(X) - \mu_s| \;\le\; \frac{\sqrt{X}}{|Y|} \cdot ||A - C||$$

■

Now we show that if the estimated centers are close to the actual centers, then one iteration of the second step in the algorithm will reduce this separation by at least half.

Notation :

- $\nu_1, \nu_2, \dots \nu_k$ denote the current centers at the beginning of an iteration in the second step of the algorithm, where ν_r is the current center closest to μ_r. Let $\delta_r = |\mu_r - \nu_r|$.
- S_r denotes the set of points A_i for which the closest current center is ν_r.
- η_r denotes the mean of points in S_r; so η_r are the new centers.

The theorem below shows that the set of misclassified points (which really belong to T_r, but have $\nu_s, s \neq r$, as the closest current center) are not too many in number. The proof first shows that any misclassified point must be far away from μ_r and since the sum of squared distances from μ_r for all points in T_r is bounded, there cannot be too many.

Theorem 5.4: Assume that $\delta_r + \delta_s \le \Delta_{rs}/16$ for all $r \neq s$. Then,

$$|T_r \cap S_s \cap G| \le \frac{6ck \cdot ||A - C||^2(\delta_r^2 + \delta_s^2)}{\Delta_{rs}^2 |\mu_r - \mu_s|^2} \tag{4}$$

Further, for any $W \subseteq T_r \cap S_s$,

$$|m(W) - \mu_s| \le \frac{4 \cdot ||A - C||}{\sqrt{|W|}} \tag{5}$$

Proof: Let $\bar{v}$ denote the projection of vector v to the affine space V spanned by $\mu_1, \dots \mu_k, \nu_1, \dots, \nu_k$ and $\eta_1, \eta_2, \dots \eta_k$. Assume $A_i \in T_r \cap S_s \cap G$. Splitting $\bar{A}_i$ into its projection along the line μ_r to μ_s and the component orthogonal to it, we can write

$$\bar{A}_i = \frac{1}{2}(\mu_r + \mu_s) + \lambda(\mu_r - \mu_s) + u,$$

where u is orthogonal to $\mu_r - \mu_s$. Since $\bar{A}_i$ is closer to ν_s than to ν_r, we have

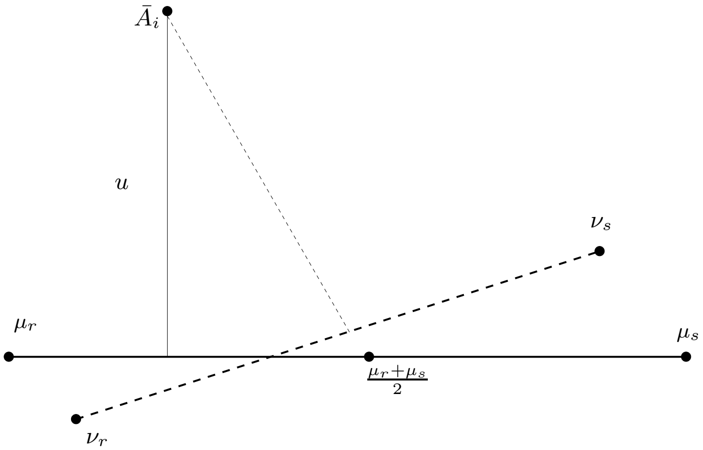

Figure 2. Misclassified A_i (note that the two lines joining μ_r, μ_s and ν_r, ν_s are not necessarily co-planar)

$$\bar{A}_i \cdot (\nu_s - \nu_r) \ge \frac{1}{2}(\nu_s - \nu_r) \cdot (\nu_r + \nu_s)$$

$$\text{i.e.,} \quad \frac{1}{2}(\mu_r + \mu_s) \cdot (\nu_s - \nu_r) + \lambda(\mu_r - \mu_s) \cdot (\nu_s - \nu_r)$$
$$+ u \cdot (\nu_s - \nu_r) \ge \frac{1}{2}(\nu_s - \nu_r) \cdot (\nu_s + \nu_r).$$

We have $u \cdot (\nu_s - \nu_r) = u \cdot ((\nu_s - \mu_s) - (\nu_r - \mu_r))$ since u is orthogonal to $\mu_r - \mu_s$. The last quantity is at most $|u|\delta$, where $\delta = \delta_r + \delta_s$. Substituting this we get

$$\frac{\delta^2}{2} + \frac{\delta}{2}|\mu_r - \mu_s| - \lambda|\mu_r - \mu_s|^2 + \lambda\delta|\mu_r - \mu_s| + |u|\delta \ge 0. \tag{6}$$

Now,

$$|\bar{A}_i - \mu_r| = \left| \left(\frac{1}{2} - \lambda\right) \cdot (\mu_s - \mu_r) + u \right|$$
$$\ge \quad |u| \;\ge\; \frac{\lambda}{\delta} \cdot |\mu_r - \mu_s|^2 - \lambda|\mu_r - \mu_s| - \frac{\delta}{2} - \frac{|\mu_r - \mu_s|}{2}$$
$$\ge \quad \frac{\Delta_{rs}|\mu_r - \mu_s|}{64\delta},$$

where the second last inequality follows from using (6) and the last inequality follows from the fact that $\lambda \ge \frac{\Delta_{rs}}{2|\mu_r - \mu_s|}$ (proximity condition) and the assumption that $\delta \le \Delta_{rs}/16$. Therefore, we have

$$|T_r \cap S_s \cap G| \cdot \frac{\Delta_{rs}^2|\mu_r - \mu_s|^2}{c\delta^2} \;\le\; \sum_{i \in T_r \cap S_s \cap G} |\bar{A}_i - \mu_r|^2$$
$$\le\; \sum_{i \in T_r} |\bar{A}_i - C_i|^2.$$

If we take a basis $u_1, u_2, \ldots u_p$ of V, we see that $\sum_{i \in T_r} |\bar{A}_i - C_i|^2 = \sum_{t=1}^{p} \sum_{i \in T_r} |(\bar{A}_i - C_i) \cdot u_t|^2 = \sum_{t=1}^{p} ||A - C||^2 \leq 3k||A - C||^2$, which proves the first statement of the theorem.

For the second statement, we can write $m(W)$ as

$$m(W) \quad = \tfrac{1}{2}(\mu_r + \mu_s) + \lambda(\mu_r - \mu_s) + u,$$

where, u is orthogonal to $\mu_r - \mu_s$. Since $m(W)$ is the average of points in S_s, we get (arguing as for (6)):

$$|u| \geq \frac{\lambda}{10\delta} |\mu_r - \mu_s|^2.$$

Now, we have

$$\begin{aligned} |m(W) - \mu_r|^2 &= |u|^2 + \left(\lambda - \frac{1}{2}\right)^2 |\mu_r - \mu_s|^2, \text{ and} \\ |m(W) - \mu_s|^2 &= |u|^2 + \left(\lambda + \frac{1}{2}\right)^2 |\mu_r - \mu_s|^2 \end{aligned}$$

If $\lambda \leq 1/4$, then clearly, $|m(W) - \mu_s| \leq 4|m(W) - \mu_r|$. If $\lambda > 1/4$, then we have $|u| \geq \frac{\lambda}{10\delta} |\mu_r - \mu_s|^2 \geq \frac{1}{2} \cdot \left(\lambda + \frac{1}{2}\right) \cdot |\mu_r - \mu_s|$ because $\frac{|\mu_r - \mu_s|}{\delta} \geq 16$. Again, we have $|m(W) - \mu_s| \leq 4|u| \leq 4|m(W) - \mu_r|$. Now, Lemma 5.2 implies that $|m(W) - \mu_r| \leq \frac{||A-C||}{\sqrt{|W|}}$, so the second statement in the theorem. ■

We are now ready to prove the main theorem of this section which will imply Theorem 2.2. This shows that $k-$means converges if the starting centers are close enough to the corresponding true centers. To gain intuition, it is best to look at the case $\varepsilon = 0$, when all points satisfy the proximity condition. Then the theorem says that if $|\nu_s - \mu_s| \leq \frac{\gamma ||A-C||}{\sqrt{n_s}}$ for all s, then $|\eta_s - \mu_s| \leq \frac{\gamma ||A-C||}{2\sqrt{n_s}}$, thus halving the upper bound of the distance to μ_s in each iteration.

Theorem 5.5: Suppose

$$\delta_s \leq \max\left(\frac{\gamma \cdot ||A - C||}{\sqrt{n_s}}, 40\sqrt{k\varepsilon n} \cdot \frac{||A - C||}{n_s}\right),$$

for all s and a parameter $\gamma \leq ck/50$. Then, for all s

$$|\eta_s - \mu_s| \leq \max\left(\frac{\gamma \cdot ||A - C||}{2\sqrt{n_s}}, 40\sqrt{k\varepsilon n} \cdot \frac{||A - C||}{n_s}\right).$$

Proof: Let n_{rs}, μ_{rs} denote the number and mean respectively of $T_r \cap S_s \cap G$ and n'_{rs}, μ'_{rs} of $(T_r \setminus G) \cap S_s$. Similarly, define n_{ss} and μ_{ss} as the size and mean of the points in $T_s \cap S_s$. We get

$$|S_s|\eta_s \quad = n_{ss}\mu_{ss} + \textstyle\sum_{r \neq s} n_{rs}\mu_{rs} + \sum_r n'_{rs}\mu'_{rs}.$$

We have

$$|\mu_{ss} - \mu_s| \leq \tfrac{\sqrt{|S_s| - n_{ss}}}{n_{ss}} ||A - C|| \; ; \; |\mu_{rs} - \mu_s| \leq \tfrac{4||A-C||}{\sqrt{n_{rs}}}$$
$$|\mu'_{rs} - \mu_s| \leq \tfrac{4||A-C||}{\sqrt{n'_{rs}}},$$

where the first one is from Corollary 5.3 (applied to the cluster S_s) and the last two are from the second statement in Theorem 5.4.

Now using the fact that length is a convex function, we have $|\eta_s - \mu_s|$ is at most

$$\begin{aligned} & \frac{n_{ss}}{|S_s|} |\mu_{ss} - \mu_s| + \sum_{r \neq s} \frac{n_{rs}}{|S_s|} |\mu_{rs} - \mu_s| + \sum_r \frac{n'_{rs}}{|S_s|} |\mu'_{rs} - \mu_s| \\ \leq \quad & 4||A - C|| \left(\frac{\sqrt{|S_s| - n_{ss}}}{n_s} + \sum_{r \neq s} \frac{\sqrt{n_{rs}}}{n_s} + \sum_r \frac{\sqrt{n'_{rs}}}{n_s}\right) \\ \leq \quad & 8||A - C|| \left(\sum_{r \neq s} \frac{\sqrt{n_{rs}}}{n_s} + \sum_r \frac{\sqrt{n'_{rs}}}{n_s}\right) \end{aligned}$$

since $|S_s| - n_{ss} = \sum_{r \neq s} n_{rs} + \sum_s n'_{rs}$. Let us look at each of the terms above. Note that $n_{rs} \leq \frac{24ck||A-C||^2 \max(\delta_r, \delta_s)^2}{\Delta_{rs}^2 |\mu_r - \mu_s|^2}$ (using Theorem 5.4). So

$$\begin{aligned} \sum_{r \neq s} \frac{\sqrt{n_{rs}}}{n_s} & \leq \frac{5\sqrt{ck}||A - C||}{n_s} \sum_{r \neq s} \frac{\max(\delta_r, \delta_s)}{\Delta_{rs} \cdot |\mu_r - \mu_s|} \\ \leq \quad & \frac{5\sqrt{ck}||A - C||^2}{n_s} \sum_{r \neq s} \frac{1}{\Delta_{rs} \cdot |\mu_r - \mu_s|} \cdot \left(\frac{\gamma}{\sqrt{\min(n_r, n_s)}} \right. \\ & \left. + \frac{40\sqrt{k\varepsilon n}}{\min(n_r, n_s)}\right) \\ \leq \quad & \frac{5\sqrt{ck}}{n_s} \sum_{r \neq s} \frac{\min(n_r, n_s)}{c^2 k^2} \cdot \left(\frac{\gamma}{\sqrt{\min(n_r, n_s)}} + \frac{40\sqrt{k\varepsilon n}}{\min(n_r, n_s)}\right) \\ \leq \quad & \frac{\gamma}{c\sqrt{n_s}} + \frac{\sqrt{k\varepsilon n}}{n_s} \end{aligned}$$

Note that $\sum_r n'_{rs} \leq \varepsilon n$. So we get $\sum_r \frac{\sqrt{n'_{rs}}}{n_s} \leq \frac{\sqrt{k\varepsilon n}}{n_s}$. Assuming c to be large enough constant proves the theorem. ■

Now we can easily finish the proof of Theorem 2.2. Observe that after the base case in the algorithm, the statement of Theorem 5.5 holds with $\gamma = 20\sqrt{k}$. So after enough number of iterations of the second step in our algorithm, γ will become very small and so we will get

$$\delta_s \leq 40\sqrt{k\varepsilon n} \cdot \frac{||A - C||}{n_s},$$

for all s. Now substituting this is Theorem 5.4, we get

$$|T_r \cap S_s \cap G| \leq \frac{40^2 \cdot 12ck \cdot ||A - C||^4 \cdot \varepsilon nk}{\Delta_{rs}^2 |\mu_r - \mu_s|^2 \cdot \min(n_r, n_s)^2} \leq \varepsilon n$$

Summing over all pairs r, s implies Theorem 2.2.

VI. Applications

We now give applications of Theorem 2.2 to various settings. One of the main technical steps here would be to bound the spectral norm of a random $n \times d$ matrix Y whose rows are chosen independently. We use the following result from [18]. Let D denote the matrix $E[Y^TY]$. Also assume that $n \geq d$.

Fact 6.1: Let γ be such that $\max_i |Y_i| \leq \gamma\sqrt{n}$ (with high probability) and $||D|| \leq \gamma^2 n$. Then $||Y|| \leq \gamma \cdot \sqrt{n} \cdot \text{polylog}(n)$ with high probability.

A. Learning in the planted distribution model

In this section, we show that McSherry's result [19] can be derived as a corollary of our main theorem. Consider an instance of the planted distribution model which satisfies the conditon (1). Each row of A can be thought of as a point in $\Re^n$. So we get a partition of the n points into k clusters. Fix a point $A_i \in T_r$. We will show that the probability that it does not satisfy the proximity condition is at most $\frac{\delta}{n}$. Using union bound, it will then follow that all points satisfy the proximity condition with probability at least $1 - \delta$.

Let $s \neq r$. Let v denote the unit vector along $\mu_r - \mu_s$. Let L_{rs} denote the line joining μ_r and μ_s, and $\hat{A}_i$ be the projection of A_i on L_{rs}. The following result shows that the distance between $\hat{A}_i$ and μ_r is small with high probability. The proof uses standard concentration bounds and we defer it to the full version.

Lemma 6.2: Assume $\sigma \geq \frac{3\log n}{n}$, where $\sigma = \max_{i,j} \sqrt{P_{ij}}$. With probability at least $1 - \frac{\delta}{n \cdot k}$,

$$|\hat{A}_i - \mu_r| \leq ck \cdot \sigma \cdot \left(\log\left(\frac{n}{\delta}\right) + \frac{1}{\sqrt{w_{\min}}}\right),$$

where c is a large constant.
Assuming

$$|\mu_r - \mu_s| \geq 4ck \cdot \sigma \cdot \left(\log\left(\frac{n}{\delta}\right) + \frac{1}{\sqrt{w_{\min}}}\right),$$

we see that

$$|\hat{A}_i - \mu_s| - |\hat{A}_i - \mu_r| \geq \frac{ck||A - C||}{\sqrt{n_r}} + \frac{ck||A - C||}{\sqrt{n_s}},$$

with probability at least $1 - \frac{\delta}{nk}$. Here, we have used the fact that $||A - C|| \leq c' \cdot \sigma\sqrt{n}$ with high probability (Wigner's theorem). Now, using union bound, we get that all the points satisfy the proximity condition with probability at least $1 - \delta$.
Remark : Here we have used C as the matrix whose rows are the actual means μ_r. But while applying Theorem 2.2, C should represent the means of the samples in A belonging to a particular cluster. The error incurred here can be made very small and will not affect the results. So we shall assume that μ_r is the actual mean of points in T_r. Similar comments apply in other applications described next.

B. Learning Mixture of Gaussians

We are given a mixture of k Gaussians $F_1, \ldots, F_k$ in d dimensions. Let the mixture weights of these distributions be $w_1, \ldots, w_k$ and $\mu_1, \ldots, \mu_k$ denote their means respectively.

Lemma 6.3: A set of $n = \text{poly}\left(\frac{d}{w_{\min}}\right)$ samples from the mixture distribution satisfy the proximity condition with high probability if

$$|\mu_r - \mu_s| \geq \frac{ck\sigma_{\max}}{\sqrt{w_{\min}}}\text{polylog}\left(\frac{d}{w_{\min}}\right),$$

for all $r, s, r \neq s$. Here $\sigma_{\max}$ is the maximum variance in any direction of any of the distributions F_r.

Proof: It can be shown that $||A - C||$ is $O\left(\sigma_{\max}\sqrt{n} \cdot \text{polylog}\left(\frac{d}{w_{\min}}\right)\right)$ with high probability (see [18]). Further, let p be a point drawn from the distribution F_r. Let L_{rs} be the line joining μ_r and μ_s. Let $\hat{p}$ be the projection of p on this line. Then the fact that F_r is Gaussian implies that $|\hat{p} - \mu_r| \leq \sigma_{\max}\text{polylog}(n)$ with probability at least $1 - \frac{1}{n^2}$. It is also easy to check that the number of points from F_r in the sample is close to $w_r n$ with high probability. Thus, it follows that all the points satisfy the proximity condition with high probability. ■

The above lemma and Theorem 2.2 imply that we can correctly classify all the points. Since we shall sample at least $\text{poly}(d)$ points from each distribution, we can learn each of the distribution to high accuracy.

C. Mixture of Distributions with Bounded Variance

We consider a mixture of distributions $F_1, \ldots, F_k$ with mixing weights $w_1, \ldots, w_k$. Let σ be an upper bound on the variance along any direction of a point sampled from one of these distributions, i.e., $\sigma \geq E\left[((x - \mu_r) \cdot v)^2\right]$, for all distributions F_r and each unit vector v.

Theorem 6.4: Suppose we are given a set of $n = \text{poly}\left(\frac{d}{w_{\min}}\right)$ samples from the mixture distribution. Assume that $\sigma \geq \frac{\text{polylog}(n)}{\sqrt{d}}$. Then there is an algorithm to correctly classify at least $1 - \varepsilon$ fraction of the points provided

$$|\mu_r - \mu_s| \geq \frac{40k\sigma}{\sqrt{\varepsilon}}\text{polylog}\left(\frac{d}{\varepsilon}\right),$$

for all $r, s, r \neq s$. Here ε is assumed to be less than $w_{\min}$.

Proof: The algorithm is described in Figure 3. We now prove that this algorithm has the desired properties. Let A denote the $n \times d$ matrix of points and C be the corresponding matrix of means. We first bound the spectral norm of $A - C$. The bound obtained is quite high, but suffices for our purpose. The proof uses Fact 6.1, and is omitted here.

Lemma 6.5: With high probability, $||A - C|| \leq \sigma\sqrt{dn} \cdot \text{polylog}(n)$.
The above lemma allows us to bound the distance between μ_r and the nearest mean obtained in Step 1 of the algorithm. The proof proceeds along the same lines as that of Lemma 5.1, and is omitted here.

Lemma 6.6: For each μ_r, there exists a center ν_r which is not removed in Step 2 and $|\mu_r - \nu_r| \leq \frac{10\sigma\sqrt{dk}}{\sqrt{\varepsilon}} \cdot \text{polylog}(n)$.

Note that in the lemma above, ν_r may not be unique for different means μ_r. Call a point $A_i \in T_r$ *bad* if $|A_i - \mu_r| \geq$

```
1) Run the first step of Algorithm Cluster on the set of points, and let $\nu_1, \dots, \nu_k$ denote the centers obtained.
2) Remove centers $\nu_r$ (and the points in the corresponding cluster) to which less than $d^2 \log d$ points are assigned.
   Let $\nu_1, \dots, \nu_{k'}$ be the remaining centers.
3) Remove any point whose distance from the nearest center in $\nu_1, \dots, \nu_{k'}$ is more than $\frac{\sigma\sqrt{n}}{\sqrt{d}}$.
4) Run the algorithm Cluster on the remaining set of points and output the clustering obtained.
```

Figure 3. Algorithm for Clustering points from mixture of distributions with bounded variance.

$\frac{\sigma\sqrt{n}}{2}$. Call a point $A_i \in T_r$ *nice* if $|A_i - \mu_r| \le \frac{\sigma\sqrt{n}}{2\sqrt{d}}$. The following lemma is again easy to prove.

Lemma 6.7: The number of bad points is at most $d \cdot \log d$ with high probability. The number of points which are not nice is at most $d^2 \log d$ with high probability. The number of nice points that are removed is at most $4kd^2 \log d$.

Corollary 6.8: With high probability the following event happens : suppose ν_r does not get removed in Step 2. Then there is a mean μ_r such that $|\mu_r - \nu_r| \le \frac{2\sigma\sqrt{n}}{\sqrt{d}}$.

Proof: Since ν_r is not removed, it has at least one nice point A_i assigned to it (otherwise it will have at most $d^2 \log d$ points assigned to it and will be removed). The distance of A_i to the nearest mean μ_s is at most $\frac{\sigma\sqrt{n}}{2\sqrt{d}}$, and Lemma 6.6 implies that there is a center ν_s which is not removed and for which $|\mu_s - \nu_s| \le \frac{\sigma\sqrt{n}}{2\sqrt{d}}$. So, $|\nu_s - A_i| \le \frac{\sigma\sqrt{n}}{\sqrt{d}}$. Since ν_r is the closest center to A_i, $|\nu_r - A_i| \le \frac{\sigma\sqrt{n}}{\sqrt{d}}$ Now, $|\nu_r - \mu_s| \le |\nu_r - A_i| + |A_i - \mu_s|$ and the result follows. ■

Let A' be the set of points which are remaining after the third step of our algorithm. We now define a new clustering $T'_1, \dots, T'_k$ of points in A'. This clustering will be very close to the actual clustering $T_1, \dots, T_k$ and so it will be enough to correctly cluster a large fraction of the points according to this new clustering. We define T'_r as the following set : $\{A_i \in T_r : A_i$ is not bad and does not get removed $\} \cup \{A_i : A_i$ is a bad point which does not get removed and the nearest center among the actual centers is μ_r $\}$.
Let μ'_r be the mean of T'_r and C' be the corresponding matrix of means. We omit the proof of the following lemma.

Lemma 6.9: With high probability, for all r, $||A' - C'|| \le O(\sigma \cdot \sqrt{n} \cdot \text{polylog}(n))$, and $|\mu_r - \mu_{r'}| \le \frac{10\sigma k d^2 \log d}{\varepsilon\sqrt{n}}$.

We are now ready to prove the main theorem. We would like to recover the clustering C' (since C and C' agree on all but the bad points). We argue that at least $(1-\varepsilon)$ fraction of the points satisfy the proximity condition. Indeed, it is easy to check that at least $(1-\varepsilon)$ fraction of the points A_i are at distance at most $\frac{4\sigma\sqrt{d}}{\sqrt{\varepsilon}}$ from the corresponding mean μ_r and satisfy the proximity condition. Since the distance between μ_s and μ'_s is *very small* (dependent inversely on n), and A_i is only $\frac{4\sigma\sqrt{d}}{\sqrt{\varepsilon}}$ far from μ_r, it will satisfy the proximity condition for A', C' as well (provided n is large enough). ■

D. Sufficient conditions for convergence of $k-$means

As mentioned in Section III, Ostrovsky et. al. [3] provided the first sufficient conditions under which they prove effectiveness of (a suitable variant of) the $k-$means algorithm. Here, we show that their conditions are (much) stronger than the proximity condition. We first describe their conditions. Let Δ_k be the optimal cost of the k-means problem (i.e., sum of distance squared of each point to nearest center) with k centers. They require: $\Delta_k \le \varepsilon\Delta_{k-1}$.

Claim 6.10: The condition above implies the proximity condition for all but ε fraction of the points.

Proof: Suppose the above condition is true. One way of getting a solution with $k-1$ centers is to remove a center μ_r and move all points in T_r to the nearest other center μ_s. Now, their condition implies

$$|\mu_r - \mu_s|^2 \ge \frac{1}{n_r \cdot \varepsilon} ||A - C||_F^2 \quad \forall s \ne r.$$

If some ε fraction of T_r do not satisfy the proximity condition, then the distance squared of each such point to μ_r is at least the distance squared along the line μ_r to μ_s which is at least $(1/4)|\mu_r - \mu_s|^2$ which is at least $\Omega(||A - C||_F^2 / n_r\epsilon)$. So even the assignment cost of such points exceeds $||A - C||_F^2$, the total cost, a contradiction. ■

We now show that our algorithm gives a PTAS for the $k-$means problem.

Getting a PTAS: Let $T_1, \dots, T_k$ be the optimal clustering and $\mu_1, \dots, \mu_k$ be the corresponding means. As before, n_r denotes the size of T_r. Let G be the set of points which satisfy the proximity condition (the good points). The above claim shows that $|G| \ge (1-\varepsilon)n$. For simplicity, assume that *exactly* ε fraction of the points do not satisfy the proximity condition. Let $S_1, \dots, S_r$ be the clustering output by our algorithm. Let μ'_r be the mean of S_r. First observe that Theorem 5.5 implies that when our algorithm stops,

$$|\mu_r - \mu'_r| \le \frac{c \cdot \sqrt{k\varepsilon n}}{n_r} \cdot ||A - C||. \tag{7}$$

for some constant c. For a point A_i, let $\alpha(A_i)$ the square of its distance to the closest mean among $\mu_1, \dots, \mu_k$. Define $\alpha'(A_i)$ for the solution output by our algorithm similarly. We now state the following claims without proof.

Claim 6.11: If $A_i \notin G$, then $\alpha'(A_i) \le (1+O(\varepsilon)) \cdot \alpha(A_i)$.

Claim 6.12: If $A_i \in G$, but is mis-classified by our algorithm, then $\alpha'(A_i) \le (1 + O(\varepsilon)) \cdot \alpha(A_i)$.

Claim 6.13: For all r, $\sum_{A_i \in G \cap T_r \cap S_r} \alpha'(A_i)$ is at most

$$(1+O(\sqrt{\varepsilon})) \cdot \sum_{A_i \in G \cap T_r \cap S_r} \alpha(A_i) + O(\sqrt{\varepsilon}) \cdot \sum_{A_i \notin G} \alpha(A_i).$$

Now summing over all r in Claim 6.13 and using Claims 6.11, 6.12 implies that our algorithm is also a PTAS.

VII. Boosting

Recall that the proximity condition requires that the distance between the means be polynomially dependent on $\frac{1}{w_{\min}}$ – this could be quite poor when one of the clusters is considerably smaller that the others. In this section, we try to overcome this obstacle for a special class of distributions.

Let $F_1, \ldots, F_k$ be a mixture of distributions in d dimensions. Let A be the $n \times d$ matrix of samples from the mixture and C be the corresponding matrix of centers. Let $D_{\min}$ denote $\min_{r,s,r\neq s} |\mu_r - \mu_s|$. The key properties that we desire from the samples are as follows :

1) For all r, s, $r \neq s$,

$$|\mu_r - \mu_s| \geq \frac{10k||A-C||}{\sqrt{n}} \tag{8}$$

2) For all i,

$$|A_i - C_i| \leq D_{\min} \cdot \sqrt{d} n^{\alpha} \text{polylog}(n), \tag{9}$$

where α is a small enough constant (something like 0.1 will suffice).

3) For all $r, s, r \neq s$,

$$\sum_{i \in T_r} [(A_i - \mu_r) \cdot v]^2 \leq \frac{|\mu_r - \mu_s|^2}{16} \cdot |T_r| \tag{10}$$

where v is the unit vector joining μ_r and μ_s. This condition is essentially saying that the average variance of points in T_r along v is bounded by $\frac{|\mu_r - \mu_s|}{4}$.

The number of samples n will be a polynomial in $\frac{d}{w_{\min}}$. To simplify the presentation, we assume that $|\mu_r - \mu_s| \leq D_{\min} \cdot \left(\frac{d}{w_{\min}}\right)^{\beta}$ for a constant β for all pairs r, s. The general case is dealt with in the full version of the paper. We now sample two sets of n points from this distribtion – call these A and B. Assume that both A and B satisfy the condtions (9) and (10). For all r, we assume that the mean of $A_i, i \in T_r$ is μ_r and $T_r \cap A$ has size $w_r \cdot n$. We assume the same for the points in B. The error caused by removing this assumption will not change our results. Let μ denote the overall mean of the points in A (or B). Note that $\mu = \sum_r w_r \mu_r$. We translate the points so that the overall mean is 0. In other words, define a translation f as $f(x) = x - \mu$. Let A'_i denote $f(A_i)$. Define B'_i similarly. We now define a set X of n points in n dimensions. The point X_i is defined as

$$(A'_i \cdot B'_1, \ldots, A'_i \cdot B'_n).$$

The correspondence between X_i and A_i naturally defines a partitioning of X into k clusters. Let $\mathcal{T}_r, r = 1, \ldots, k$, denote these clusters. Note that the mean θ_r of $\mathcal{T}_r$ is $(C'_r \cdot B'_1, \ldots, C'_r \cdot B'_n)$, where $C'_r = C_r - \mu$. The algorithm is described in Figure 4. Here, n is equal to $\left(\frac{d}{w_{\min}}\right)^{8(\beta+1)}$. We now briefly explain why the algorithm works. Let $\phi_r, r = 1, \ldots, k$ be the k centers output by the first step of the algorithm Cluster-Boost. The key result is that this clustering has small classification errors.

Lemma 7.1: For every cluster $\mathcal{T}_r$, there is a (unique) center ϕ_r such that the number of points in $\mathcal{T}_r$ which are not assigned to ϕ_r after the first step of the algorithm is at most $\left(\frac{w_{\min}}{d}\right)^{2\beta} \cdot n$.

The proof of this result is similar to that of Lemma 5.1 and uses the following two lemmas which we state without proof. Let Z denote the matrix of means of X, i.e., $Z_i = \theta_r$ if $X_i \in \mathcal{S}_r$. These lemmas show that the process of constructing X *amplifies* the distance between the means θ_r by a much bigger factor than that for $\frac{||X-Z||}{\sqrt{n}}$.

Lemma 7.2: For all $r, s, r \neq s$,

$$|\theta_r - \theta_s| \geq \frac{|\mu_r - \mu_s|^2}{4} \cdot \sqrt{w_{\min}} \sqrt{n}.$$

Lemma 7.3: With high probability,

$$\frac{||X - Z||}{\sqrt{n}} \leq D_{\min}^2 \cdot d \cdot n^{2\alpha} \cdot \left(\frac{d}{w_{\min}}\right)^{\beta} \cdot \text{polylog}(n)$$

Now, let ν_r denote the center of S_r. Starting with Lemma 7.1, and using ideas similar to those in Theorem 5.5, we can prove that for every s, $|\nu_s - \mu_s| \leq \frac{||A-C||}{\sqrt{n}}$. Using these initial centers ν_s, we run the second step of algorithm Cluster. Then, we have the following analogue of Theorem 2.2 in this setting.

Theorem 7.4: Suppose a mixture of distribution satisfies the conditions (8–10) above and at least $(1 - \varepsilon)$ fraction of sampled points satisfy the proximity condition. Then we can correctly classify all but $O(k^2 \varepsilon)$ fraction of the points.

We now give some applications of Theorem 7.4. We state these results without proof.

1) Learning Gaussian Distributions: Suppose we are given a mixture of Gaussian distribution where the means satisfy the following condition for all $r, s, r \neq s$:

$$|\mu_r - \mu_s| \geq \Omega\left(\sigma k \cdot \log \frac{d}{w_{\min}}\right).$$

Here σ denotes the maximum variance in any direction of the Gaussian distributions.

Lemma 7.5: Given a mixture of k Gaussians satisfying the separation condition above, we can correctly classify a set n samples, where $n = \text{poly}\left(\frac{d}{w_{\min}}\right)$.

2) Learning Power Law Distributions: Consider a mixture of distributions $F_1, \ldots, F_k$ where each of the distributions F_r satisfies the following condition for every unit vector v :

$$P_{X \in F_r}\left[|(X - \mu_r) \cdot v| > \sigma t\right] \leq \frac{1}{t^{\gamma}} \tag{11}$$

```
1) Run the first step of Algorithm Cluster on $X$. Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be the clustering obtained.
2) Let $S_1, \ldots, S_k$ be the clustering of $A$ corresponding to $\mathcal{S}_1, \ldots, \mathcal{S}_k$.
3) Run second step of Algorithm Cluster on $A$ with $S_1, \ldots, S_k$ (and the corresponding means) as the initial solution.
4) Output the clustering obtained.
```

Figure 4. Algorithm `Cluster-Boost`

where $\gamma \geq 2$ is a large enough constant. Suppose the means satisfy the following separation condition for every $r, s, r \neq s$:

$$|\mu_r - \mu_s| \geq \Omega\left(\sigma k \cdot \left(\log \frac{d}{w_{\min}} + \frac{1}{\varepsilon^{\frac{1}{\gamma}}}\right)\right).$$

Then, we have

Theorem 7.6: Given a sample of n points from a mixture of distributions satsifying the conditions above, we can cluster at least $(1 - \varepsilon)$ fraction of the points with high probability. Here, $n = \text{poly}\left(\frac{d}{w_{\min}}\right)$.

References

[1] R. Kannan and S. Vempala, "Spectral algorithms," *Foundations and Trends in Theoretical Computer Science*, vol. 4, no. 3-4, pp. 157–288, 2009.

[2] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[3] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of lloyd-type methods for the k-means problem," in *Proc. 47th IEEE FOCS*, 2006, pp. 165–176.

[4] D. Arthur and S. Vassilvitskii, "How slow is the *k*-means method?" in *Proc. 22nd Annual Symposium on Computational Geometry*, 2006, pp. 144–153.

[5] S. Dasgupta, "How fast is -means?" in *Proc. 16th Annual Conference on Learning Theory*, 2003, p. 735.

[6] S. Har-Peled and B. Sadri, "How fast is the *k*-means method?" in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2005, pp. 877–885.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] S. Dasgupta, "Learning mixtures of gaussians," in *Proc. 40th IEEE FOCS*, 1999, pp. 634–644.

[9] S. Dasgupta and L. J. Schulman, "A probabilistic analysis of em for mixtures of separated, spherical gaussians," *Journal of Machine Learning Research*, vol. 8, pp. 203–226, 2007.

[10] S. Arora and R. Kannan, "Learning mixtures of arbitrary gaussians," in *Proc. 33rd Annual ACM Symposium on Theory of Computing*, 2001, pp. 247–257.

[11] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *J. Comput. Syst. Sci.*, vol. 68, no. 4, pp. 841–860, 2004.

[12] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *Proc. 18th Annual Conference on Learning Theory*, 2005, pp. 458–469.

[13] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," *SIAM J. Comput.*, vol. 38, no. 3, pp. 1141–1156, 2008.

[14] S. C. Brubaker and S. Vempala, "Isotropic PCA and affine-invariant clustering," in *Proc 49th IEEE FOCS*, 2008, pp. 551–560.

[15] S. Vempala, "Learning convex concepts from gaussian distributions with pca," *To appear in IEEE FOCS, 2010.*

[16] A. Dasgupta, J. E. Hopcroft, J. M. Kleinberg, and M. Sandler, "On learning mixtures of heavy-tailed distributions," in *Proc. 46th IEEE FOCS*, 2005, pp. 491–500.

[17] K. Chaudhuri and S. Rao, "Beyond gaussians: Spectral methods for learning mixtures of heavy-tailed distributions," in *Proc. 24th Annual Conference on Learning Theory*, 2008, pp. 21–32.

[18] A. Dasgupta, J. E. Hopcroft, R. Kannan, and P. P. Mitra, "Spectral clustering with limited independence," in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1036–1045.

[19] F. McSherry, "Spectral partitioning of random graphs," in *Proc. 42nd IEEE FOCS*, 2001, pp. 529–537.

[20] M.-F. Balcan, A. Blum, and A. Gupta, "Approximate clustering without the approximation," in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2009, pp. 1068–1077.

[21] M.-F. Balcan, A. Blum, and S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proc. 40th Annual ACM Symposium on Theory of Computing*, 2008, pp. 671–680.

[22] A. Kumar, Y. Sabharwal, and S. Sen, "Linear-time approximation schemes for clustering problems in any dimensions," *J. ACM*, vol. 57, no. 2, 2010.

[23] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.

[24] A. Aggarwal, A. Deshpande, and R. Kannan, "Adaptive sampling for k-means clustering," in *APPROX-RANDOM*, 2009, pp. 15–28.

[25] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "A local search approximation algorithm for k-means clustering," in *Proc. 18th Annual Symposium on Computational Geometry*, 2002, pp. 10–18.