Journal of Machine Learning Research 21 (2020) 1-16 Submitted 3/19; Published 9/20

Recovery of a Mixture of Gaussians by Sum-of-norms Clustering

Tao Jiang TJ293@CORNELL.EDU
School of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14850, USA

Stephen Vavasis VAVASIS@UWATERLOO.CA
Department of Combinatorics & Optimization
University of Waterloo
Waterloo, ON N2L 3G1, Canada

Chen Wen Zhai ZHAI@MIT.EDU
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Inderjit Dhillon

Abstract

Sum-of-norms clustering is a method for assigning n points in $\mathbf{R}^d$ to K clusters, $1 \leq K \leq n$, using convex optimization. Recently, Panahi et al. (2017) proved that sum-of-norms clustering is guaranteed to recover a mixture of Gaussians under the restriction that the number of samples is not too large. The purpose of this note is to lift this restriction, that is, show that sum-of-norms clustering can recover a mixture of Gaussians even as the number of samples tends to infinity. Our proof relies on an interesting characterization of clusters computed by sum-of-norms clustering that was developed inside a proof of the agglomeration conjecture by Chiquet et al. (2017). Because we believe this theorem has independent interest, we restate and reprove the Chiquet et al. (2017) result herein.

Keywords: Sum-of-norms Clustering, Mixture of Gaussians, Recovery Guarantees, Unsupervised Learning

1. Introduction

Clustering is perhaps the most central problem in unsupervised machine learning and has been studied for over 60 years (Shalev-Shwartz and Ben-David, 2014). The problem may be stated informally as follows. One is given n points, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ lying in $\mathbf{R}^d$. One seeks to partition $\{1, \ldots, n\}$ into K sets $C_1, \ldots, C_K$ such that the $\boldsymbol{a}_i$'s for $i \in C_m$ are closer to each other than to the $\boldsymbol{a}_i$'s for $i \in C_{m'}$, $m' \neq m$.

Clustering is usually posed as a nonconvex optimization problem, and therefore prone to nonoptimal local minimizers, but Pelckmans et al. (2005), Hocking et al. (2011), and Lindsten et al. (2011) proposed the following convex formulation for the clustering problem:

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathbf{R}^d} \frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{a}_i\|^2+\lambda\sum_{1\le i<j\le n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|. \tag{1}$$

This formulation is known in the literature as sum-of-norms clustering, convex clustering, or clusterpath clustering. Let $\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_n^*$ be the optimizer. (Note: (1) is strongly convex, hence the optimizer exists and is unique.) The assignment to clusters is given by the $\boldsymbol{x}_i^*$'s: for i,i', if $\boldsymbol{x}_i^*=\boldsymbol{x}_{i'}^*$ then i,i' are assigned to the same cluster, else they are assigned to different clusters. It is apparent that for $\lambda=0$, each $\boldsymbol{a}_i$ is assigned to a different cluster (unless $\boldsymbol{a}_i=\boldsymbol{a}_{i'}$ exactly), whereas for λ sufficiently large, the second summation drives all the $\boldsymbol{x}_i$'s to be equal (and hence there is one big cluster). Thus, the parameter λ controls the number of clusters produced by the formulation.

Throughout this paper, we assume that all norms are Euclidean, although (1) has also been considered for other norms. In addition, some authors insert nonnegative weights in front of the terms in the above summations. Most of our results, however, require all weights identically 1, but we revisit the question of general weights in Sections 5 and 6.

Recently, there have been various attempts to provide recovery guarantees for sum-of-norms clustering with uniform weights (1). Zhu et al. (2014) showed that if a data set is generated by two well-separated cubes, then sum-of-norms clustering recovers the two clusters perfectly. The separation condition is rather strict: the distance between two cubes must be larger than a threshold dependent on the number of data points and the sizes of two cubes. Tan and Witten (2015) studied the statistical properties of sum-of-norms clustering. Panahi et al. (2017) developed several recovery theorems as well as a first-order optimization method for solving (1). Other authors, for example, Sun et al. (2018) have since extended these results. One of Panahi et al.'s results pertains to a mixture of spherical Gaussians, which is the following generative model for producing the data $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$. The parameters of the model are K means $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\in\mathbf{R}^d$, K variances $\sigma_1^2,\ldots,\sigma_K^2$, and K probabilities $w_1,\ldots,w_K$, all positive and summing to 1. One draws n i.i.d. samples as follows. First, an index $m\in\{1,\ldots,K\}$ is selected at random according to probabilities $w_1,\ldots,w_K$. Next, a point $\boldsymbol{a}$ is chosen according to the spherical Gaussian distribution $N(\boldsymbol{\mu}_m,\sigma_m^2 I)$.

Panahi et al. proved that for the appropriate choice of λ, sum-of-norms clustering formulation (1) will exactly recover a mixture of Gaussians (that is, each point will be labeled with m if it was selected from $N(\boldsymbol{\mu}_m,\sigma_m^2 I)$) provided that for all m,m', $1\le m<m'\le K$,

$$\|\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\|\ge\frac{CK\sigma_{\max}}{w_{\min}}\mathrm{polylog}(n). \tag{2}$$

One issue with this bound is that as the number of samples n tends to infinity, the bound seems to indicate that distinguishing the clusters becomes increasingly difficult (that is, the $\boldsymbol{\mu}_m$'s have to be more distantly separated as $n\to\infty$).

The reason for this aspect of their bound is that their proof technique requires a gap of positive width (that is, a region of $\mathbf{R}^d$ containing no sample points) between $\{\boldsymbol{a}_i : i\in C_m\}$ and $\{\boldsymbol{a}_i : i\in C_{m'}\}$ whenever $m\ne m'$. Clearly, such a gap cannot exist in the mixture-of-Gaussians distribution as the number of samples tends to infinity.

The purpose of this note is to prove that (1) can recover a mixture of Gaussians even as $n\to\infty$. This is the content of Theorem 3 in Section 4 below. Naturally, under this

hypothesis we cannot hope to correctly label all samples since, as $n \to \infty$, some of the samples associated with one mean will be placed arbitrarily close to another mean. Therefore, we are content in showing that (1) can correctly cluster the points lying within some fixed number of standard-deviations for each mean.

A related result by Radchenko and Mukherjee (2017) analyzed the special case of a mixture of Gaussians with $K = 2$, $d = 1$ under slightly different hypotheses. Also, Mixon et al. (2017) showed that semidefinite relaxation of clustering Peng and Wei (2007) can recover a mixture of Gaussians as $n \to \infty$, but this result requires nontrivial postprocessing of the semidefinite solution to recover the clusters.

Our proof technique requires a cluster characterization theorem for sum-of-norms clustering derived by Chiquet et al. (2017). This theorem is not stated by these authors as a theorem, but instead appears as a sequence of steps inside a larger proof in a "supplementary material" appendix to their paper. Because we believe that this theorem is of independent interest, we restate it below and for the sake of completeness provide the proof (which is the same as the proof appearing in Chiquet et al.'s supplementary material). This material appears in Section 2. We conclude with some experimental results in Section 6.

2. Cluster Characterization Theorem

The following theorem is due to Chiquet et al. (2017) appearing as a sequence of steps in a proof of the agglomeration conjecture. Refer to the next section for a discussion of the agglomeration conjecture. We restate the theorem here because it is needed for our analysis and because we believe it is of independent interest.

Theorem 1 *Let $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$ denote the optimizer of (1). For notational ease, let $\boldsymbol{x}^*$ denote the concatenation of these vectors into a single vector in $\mathbf{R}^{nd}$. Suppose that C is a nonempty subset of $\{1, \ldots, n\}$.*

(a) Necessary condition: If for some $\hat{\boldsymbol{x}} \in \mathbf{R}^d$, $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$ and $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$ (that is, C is exactly one cluster determined by (1)), then there exist $\boldsymbol{z}_{ij}^*$ for $i, j \in C$, $i \neq j$, which solve

$$\begin{aligned} \boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l = \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* \quad & \forall i \in C, \\ \left\| \boldsymbol{z}_{ij}^* \right\| \leq 1 \quad & \forall i, j \in C, i \neq j, \\ \boldsymbol{z}_{ij}^* = -\boldsymbol{z}_{ji}^* \quad & \forall i, j \in C, i \neq j. \end{aligned} \tag{3}$$

(b) Sufficient condition: Suppose there exists a solution $\boldsymbol{z}_{ij}^*$ for $j \in C - \{i\}$, $i \in C$ to the conditions (3). Then there exists an $\hat{\boldsymbol{x}} \in \mathbf{R}^d$ such that the minimizer $\boldsymbol{x}^*$ of (1) satisfies $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$.

Note: This theorem is an almost exact characterization of clusters that are determined by formulation (1). The only gap between the necessary and sufficient conditions is that the necessary condition requires that C be exactly all the points in a cluster, whereas the sufficient condition is sufficient for C to be a subset of the points in a cluster. The sufficient condition is notable because it does not require any hypothesis about the other $n - |C|$ points occurring in the input.

Proof (Chiquet et al.) Proof for Necessity (a)
As $\boldsymbol{x}^*$ is the minimizer of the problem (1), and this objective function, call it $f(\boldsymbol{x})$, is convex, it follows that $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$, where $\partial f(\boldsymbol{x}^*)$ denotes the subdifferential, that is, the set of subgradients of f at $\boldsymbol{x}^*$. (See, for example, Hiriart-Urruty and Lemaréchal, 2012, for background on convex analysis). Written explicitly in terms of the derivative of the squared-norm and subdifferential of the norm, this means that $\boldsymbol{x}^*$ satisfies the following condition:

$$\boldsymbol{x}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij}^* = \mathbf{0} \qquad \forall i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{w}_{ij}^*$, $i = 1, \ldots, n$, $j = 1, \ldots, n$, $i \neq j$, are subgradients of the Euclidean norm function satisfying

$$\boldsymbol{w}_{ij}^* = \begin{cases} \frac{\boldsymbol{x}_i^* - \boldsymbol{x}_j^*}{\|\boldsymbol{x}_i^* - \boldsymbol{x}_j^*\|}, & \text{for } \boldsymbol{x}_i^* \neq \boldsymbol{x}_j^*, \\ \text{arbitrary point in } B(\mathbf{0}, 1), & \text{for } \boldsymbol{x}_i^* = \boldsymbol{x}_j^*, \end{cases}$$

with the requirement that $\boldsymbol{w}_{ij}^* = -\boldsymbol{w}_{ji}^*$ in the second case. Here, $B(\boldsymbol{c}, r)$ is notation for the closed Euclidean ball centered at $\boldsymbol{c}$ of radius r. Since $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$, $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$, the KKT condition for $i \in C$ is rewritten as

$$\hat{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\right\|} + \lambda \sum_{j \in C - \{i\}} \boldsymbol{w}_{ij}^* = \mathbf{0}, \tag{5}$$

Define $\boldsymbol{z}_{ij}^* = \boldsymbol{w}_{ij}^*$ for $i, j \in C$, $i \neq j$. Then

$$\left\|\boldsymbol{z}_{ij}^*\right\| \leq 1, \boldsymbol{z}_{ij}^* = -\boldsymbol{z}_{ji}^*, \forall i, j \in C, i \neq j.$$

Substitute $\boldsymbol{w}_{ij}^* = \boldsymbol{z}_{ij}^*$ into the equation (5) to obtain

$$\hat{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\right\|} + \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* = \mathbf{0}, \tag{6}$$

Sum the preceding equation over $i \in C$, noticing that the last term cancels out, leaving

$$|C|\hat{\boldsymbol{x}} - \sum_{i \in C} \boldsymbol{a}_i + \lambda |C| \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\right\|} = \mathbf{0},$$

which is rearranged to (renaming i to l):

$$\lambda \sum_{j \notin C} \frac{\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*}{\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}_j^*\right\|} = -\hat{\boldsymbol{x}} + \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l. \tag{7}$$

Subtract (7) from (6), simplify and rearrange to obtain

$$\boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l = \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* \qquad \forall i \in C, \tag{8}$$

as desired.

Proof for Sufficiency (b)

We will show that at the solution of (1), all the $\boldsymbol{x}_i^*$'s for $i \in C$ have a common value under the hypothesis that $\boldsymbol{z}_{ij}^*$ is a solution to the equation (3) for $i, j \in C$, $i \neq j$.

First, define the following intermediate problem. Let $\tilde{\boldsymbol{a}}$ denote the centroid of $\boldsymbol{a}_i$ for $i \in C$:

$$\tilde{\boldsymbol{a}} = \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l.$$

Consider the weighted problem sum-of-norms clustering problem with unknowns as follows: one unknown $\boldsymbol{x} \in \mathbf{R}^d$ is associated with C, and one unknown $\boldsymbol{x}_j$ is associated with each $j \notin C$ (for a total of $n - |C| + 1$ unknown vectors):

$$\min_{\boldsymbol{x}; \boldsymbol{x}_j} \frac{|C|}{2} \cdot \|\boldsymbol{x} - \tilde{\boldsymbol{a}}\|^2 + \frac{1}{2} \sum_{j \notin C} \|\boldsymbol{x}_j - \boldsymbol{a}_j\|^2 + \lambda |C| \sum_{j \notin C} \|\boldsymbol{x} - \boldsymbol{x}_j\| + \lambda \sum_{\substack{i,j \notin C \\ i<j}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \quad (9)$$

This problem, being strongly convex, has a unique optimizer; denote the optimizing vectors $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}_j$ for $j \notin C$.

The optimality conditions for (9) are:

$$|C|(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{a}}) + \lambda |C| \sum_{j \notin C} \boldsymbol{g}_j = \mathbf{0}, \quad (10)$$

$$\tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i - \lambda |C| \boldsymbol{g}_i + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} = \mathbf{0} \qquad \forall i \notin C, \quad (11)$$

with subgradients defined as follows:

$$\boldsymbol{g}_j = \begin{cases} \frac{\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_j}{\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_j\|}, & \text{for } \tilde{\boldsymbol{x}}_j \neq \tilde{\boldsymbol{x}}, \\ \text{arbitrary in } B(\mathbf{0}, 1), & \text{for } \tilde{\boldsymbol{x}}_j = \tilde{\boldsymbol{x}}, \end{cases} \qquad \forall j \notin C,$$

and

$$\boldsymbol{y}_{ij} = \begin{cases} \frac{\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j}{\|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\|}, & \text{for } \tilde{\boldsymbol{x}}_i \neq \tilde{\boldsymbol{x}}_j, \\ \text{arbitrary in } B(\mathbf{0}, 1), & \text{for } \tilde{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_j, \end{cases} \qquad \forall i, j \notin C, i \neq j,$$

with the proviso that in the second case, $\boldsymbol{y}_{ij} = -\boldsymbol{y}_{ji}$.

We claim that the solution for (1) given by defining $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}$ for $i \in C$ while keeping the $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$ for $j \notin C$, where $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}_j$ are the optimizers for (9) as in the last few paragraphs, is optimal for (1), which proves the main result. To show that this solution is optimal for (1), we need to provide subgradients to establish the necessary condition. Define $\boldsymbol{w}_{ij}$ to be the subgradients of $\boldsymbol{x}_i \mapsto \left\| \boldsymbol{x}_i - \tilde{\boldsymbol{x}}_j^* \right\|$ evaluated at $\tilde{\boldsymbol{x}}_i^*$ as follows:

$$\begin{aligned} \boldsymbol{w}_{ij} &= \boldsymbol{g}_j && \text{for } i \in C, j \notin C, \\ \boldsymbol{w}_{ij} &= \boldsymbol{y}_{ij} && \text{for } i, j \notin C,\ i \neq j, \\ \boldsymbol{w}_{ij} &= \boldsymbol{z}_{ij}^* && \text{for } i, j \in C,\ i \neq j, \end{aligned}$$

Before confirming that the necessary condition is satisfied, we first need to confirm that these are all valid subgradients. In the case that $i \in C$, $j \notin C$, we have constructed $\boldsymbol{g}_j$ to be a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated at $\tilde{\boldsymbol{x}}$, and we have taken $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}$, $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$.

In the case that $i, j \notin C$, we have construct $\boldsymbol{y}_{ij}$ to be a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated at $\tilde{\boldsymbol{x}}_i$, and we have taken $\boldsymbol{x}_i^* = \tilde{\boldsymbol{x}}_i$, $\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j$.

In the case that $i, j \in C$, by construction $\boldsymbol{x}_i^* = \boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}$, so any vector in $B(\boldsymbol{0}, 1)$ is a valid subgradient of $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_j\|$ evaluated $\tilde{\boldsymbol{x}}_i$. Note that since $\boldsymbol{z}_{ij}^* \in B(\boldsymbol{0}, 1)$, then $\boldsymbol{w}_{ij}$ defined above also lies in $B(\boldsymbol{0}, 1)$.

Now we check the necessary conditions for optimality in (1). First, consider an $i \in C$:

$$\begin{aligned}
\tilde{\boldsymbol{x}}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij} &= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \in C - \{i\}} \boldsymbol{w}_{ij} + \lambda \sum_{j \notin C} \boldsymbol{w}_{ij} \\
&= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \lambda \sum_{j \in C - \{i\}} \boldsymbol{z}_{ij}^* + \lambda \sum_{j \notin C} \boldsymbol{g}_j \\
&= \tilde{\boldsymbol{x}} - \boldsymbol{a}_i + \boldsymbol{a}_i - \frac{1}{|C|} \sum_{l \in C} \boldsymbol{a}_l + \lambda \sum_{j \notin C} \boldsymbol{g}_j \qquad \text{(by (3))} \\
&= \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{a}} + \lambda \sum_{j \notin C} \boldsymbol{g}_j \\
&= \boldsymbol{0} \qquad \text{(by (10)).}
\end{aligned}$$

Then we check for $i \notin C$:

$$\begin{aligned}
\tilde{\boldsymbol{x}}_i^* - \boldsymbol{a}_i + \lambda \sum_{j \neq i} \boldsymbol{w}_{ij} &= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i + \lambda \sum_{j \in C} \boldsymbol{w}_{ij} + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{w}_{ij} \\
&= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i + \lambda \sum_{j \in C} (-\boldsymbol{g}_i) + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} \\
&= \tilde{\boldsymbol{x}}_i - \boldsymbol{a}_i - \lambda |C| \boldsymbol{g}_i + \lambda \sum_{j \notin C \cup \{i\}} \boldsymbol{y}_{ij} \\
&= \boldsymbol{0} \qquad \text{(by (11)).}
\end{aligned}$$

■

3. Agglomeration Conjecture

Recall that when $\lambda = 0$, each $\boldsymbol{a}_i$ is in its own cluster in the solution to (1) (provided the $\boldsymbol{a}_i$'s are distinct), whereas for sufficiently large λ, all the points are in one cluster. Hocking et al. (2011) conjectured that sum-of-norms clustering with equal weights has the following agglomeration property: as λ increases, clusters merge with each other but never break up. This means that the solutions to (1) as λ ranges over $[0, \infty)$ induce a tree of hierarchical clusters on the data.

This conjecture was proved by Chiquet et al. (2017) using Theorem 1. Consider a $\bar{\lambda} \geq \lambda$ and its corresponding sum-of-norms cluster model:

$$\min_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n} \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{a}_i\|^2 + \bar{\lambda} \sum_{1 \leq i < j \leq n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{12}$$

Corollary 2 *(Chiquet et al.) If there is a C such that minimizer $\boldsymbol{x}^*$ of (1) satisfies $\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}$ for $i \in C$, $\boldsymbol{x}_i^* \neq \hat{\boldsymbol{x}}$ for $i \notin C$ for some $\hat{\boldsymbol{x}} \in \mathbf{R}^d$, then there exists an $\hat{\boldsymbol{x}}' \in \mathbf{R}^d$ such that the minimizer of (12), $\bar{\boldsymbol{x}}^*$, satisfies $\bar{\boldsymbol{x}}_i^* = \hat{\boldsymbol{x}}'$ for $i \in C$.*

The corollary follows from Theorem 1. If C is a cluster in the solution of (1), then by the necessary condition, there exist multipliers $\boldsymbol{z}_{ij}^*$ satisfying (3) for λ. If we scale each of these multipliers by $\lambda/\bar{\lambda}$, we now obtain a solution to (3) with λ replaced by $\bar{\lambda}$, and the theorem states that this is sufficient for the points in C to be in the same cluster in the solution to (12).

It should be noted that Hocking et al. (2011) construct an example of unequally-weighted sum-of-norms clustering in which the agglomeration property fails. It is still mostly an open question to characterize for which norms and for which families of unequal weights the agglomeration property holds. Refer to Chi and Steinerberger (2018) for some recent progress.

4. Mixture of Gaussians

In this section, we present our main result about recovery of a mixture of Gaussians. As noted in the introduction, a theorem stating that every point is labeled correctly is not possible in the setting of $n \to \infty$, so we settle for a theorem stating that points within a constant number of standard deviations from the means are correctly labeled.

Theorem 3 *Let the vertices $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbf{R}^d$ be generated from a mixture of K Gaussian distributions with parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, $\sigma_1^2, \ldots, \sigma_K^2$, and $w_1, \ldots, w_K$. Let $\theta > 0$ be given, and let*

$$V_m = \{i : \|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \leq \theta\sigma_m\}, \quad m = 1, \ldots, K.$$

Let $\epsilon > 0$ be arbitrary. Then for any $m = 1, \ldots, K$, with probability exponentially close to 1 (and depending on ϵ; see (15)) as $n \to \infty$, for the solution $\boldsymbol{x}^*$ to (1), the points indexed by V_m are in the same cluster provided

$$\lambda \geq \frac{2\theta\sigma_m}{(F(\theta, d)w_m - \epsilon)n}. \tag{13}$$

Here, $F(\theta, d)$ denotes the cumulative density function of the chi distribution with d degrees of freedom (which tends to 1 rapidly as θ increases). Furthermore, the cluster associated with V_m is distinct from the cluster associated with $V_{m'}$, $1 \leq m < m' \leq K$ with probability exponentially close to 1 as $n \to \infty$ (see (16)), provided that

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n-1)}. \tag{14}$$

Proof Let $\epsilon > 0$ be fixed. Fix an $m \in \{1, \ldots, K\}$. First, we show that all the points indexed by V_m are in the same cluster. The usual technique for proving a recovery result is to find subgradients to satisfy the sufficient condition, which in this case is Theorem 1 taking C in the theorem to be V_m. Observe that conditions (3) involve equalities and norm inequalities. A standard technique in the literature (see, for example, Candès and Recht,

2009) is to find the least-squares solution to the equalities and then prove that it satisfies the inequalities. This is the technique we adopt herein. The conditions (3) are in sufficiently simple form that we can write down the least-squares solution in closed form; it turns out to be:

$$\boldsymbol{z}_{ij}^* = \frac{1}{\lambda|V_m|}(\boldsymbol{a}_i - \boldsymbol{a}_j) \quad \forall i, j \in V_m, i \neq j.$$

It follows by construction (and is easy to check) that this formula satisfies the equalities in (3), so the remaining task is to show that the norm bound $\left\|\boldsymbol{z}_{ij}^*\right\| \leq 1$ is satisfied. By definition of V_m, $\|\boldsymbol{a}_i - \boldsymbol{a}_j\| \leq 2\theta\sigma_m$. The probability that an arbitrary sample $\boldsymbol{a}_i$ is associated with mean $\boldsymbol{\mu}_m$ is w_m. Furthermore, with probability $F(\theta, d)$, this sample satisfies $\|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \leq \theta\sigma_m$, that is, $i \in V_m$. Since the second choice in the mixture of Gaussians is conditionally independent from the first, the overall probability that $i \in V_m$ is $F(\theta, d)w_m$. Therefore, $\mathbb{E}[|V_m|] = F(\theta, d)w_m n$. It follows that the probability that $|V_m| \geq (F(\theta, d)w_m - \epsilon)n$ is exponentially close to 1 as $n \to \infty$ for a fixed $\epsilon > 0$. Specifically,

$$\text{Prob}\left[|V_m| \geq (F(\theta, d)w_m - \epsilon)n\right] \geq 1 - \exp(-2\epsilon^2 n), \tag{15}$$

by Hoeffding's inequality (1963) for the binomial distribution. Thus, provided

$$\lambda \geq 2\theta\sigma_m/((F(\theta, d)w_m - \epsilon)n),$$

we have constructed a solution to (3) with probability exponentially close to 1 as $n \to \infty$.

For the second part of the theorem, suppose $1 \leq m < m' \leq K$. For each sample $\boldsymbol{a}_i$ associated with $\boldsymbol{\mu}_m$ satisfying $\|\boldsymbol{a}_i - \boldsymbol{\mu}_m\| \leq \theta\sigma_m$ (that is, lying in V_m), the probability is $1/2$ that

$$(\boldsymbol{a}_i - \boldsymbol{\mu}_m)^T(\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m) \leq 0,$$

by the fact that the spherical Gaussian distribution has mirror-image symmetry about any hyperplane through its mean. Therefore, with probability exponentially close to 1 as $n \to \infty$, we can assume that at least one $i \in V_m$ satisfies the above inequality. In particular,

$$\text{Prob}\left[\exists i \in V_m \text{ s.t. } (\boldsymbol{a}_i - \boldsymbol{\mu}_m)^T(\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m) \leq 0\right] \geq 1 - 2^{-|V_m|}, \tag{16}$$

(Note that, as noted above, $|V_m|$ grows linearly with n with probability exponentially close to 1 as $n \to \infty$.) Similarly, with probability exponentially close to 1, at least one sample $i' \in V_{m'}$ satisfies

$$(\boldsymbol{a}_{i'} - \boldsymbol{\mu}_{m'})^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) \leq 0.$$

Then

$$\begin{aligned}
\|\boldsymbol{a}_i - \boldsymbol{a}_{i'}\|^2 &= \|\boldsymbol{a}_i - \boldsymbol{\mu}_m - \boldsymbol{a}_{i'} + \boldsymbol{\mu}_{m'} + \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2 \\
&= \|\boldsymbol{a}_i - \boldsymbol{\mu}_m - \boldsymbol{a}_i + \boldsymbol{\mu}_{m'}\|^2 + 2(\boldsymbol{a}_i - \boldsymbol{\mu}_m)^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) \\
&\quad - 2(\boldsymbol{a}_{i'} - \boldsymbol{\mu}_{m'})^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}) + \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2 \\
&\geq \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|^2,
\end{aligned} \tag{17}$$

where, in the final line, we used the two inequalities derived earlier in this paragraph.

Consider the first-order optimality conditions for equation (1), which are given by (4). Apply the triangle inequality to the summation in (4) to obtain,

$$\|\boldsymbol{x}_i^* - \boldsymbol{a}_i\| \leq \lambda(n-1), \text{ and} \tag{18}$$
$$\|\boldsymbol{x}_{i'}^* - \boldsymbol{a}_{i'}\| \leq \lambda(n-1). \tag{19}$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i'}^*\| &= \|\boldsymbol{a}_i - \boldsymbol{a}_{i'} + \boldsymbol{x}_i^* - \boldsymbol{a}_i - \boldsymbol{x}_{i'}^* + \boldsymbol{a}_{i'}\| \\
&\geq \|\boldsymbol{a}_i - \boldsymbol{a}_{i'}\| - \|\boldsymbol{x}_i^* - \boldsymbol{a}_i\| - \|\boldsymbol{x}_{i'}^* - \boldsymbol{a}_{i'}\| \qquad \text{(by the triangle inequality)} \\
&\geq \|\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m\| - 2\lambda(n-1) \qquad \text{(by (17), (18), and (19)).}
\end{aligned}$$

Therefore, we conclude that $\boldsymbol{x}_i^* \neq \boldsymbol{x}_{i'}^*$, ., that V_m and $V_{m'}$ are not in the same cluster, provided that the right-hand side of the preceding inequality is positive, that is,

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n-1)}.$$

This concludes the proof of the second statement. ■

In order to state a simpler bound, we can fix some values. For example, let us take $\theta = 2d^{1/2}$ and let $c_d = F(2d^{1/2}, d)$. The Chernoff bound implies that $c_d \to 1$ exponentially fast in d. Let $w_{\min} = \min_{m=1,\ldots,K} w_m$ and $\sigma_{\max} = \max_{m=1,\ldots,K} \sigma_m$. Finally, let us take $\epsilon = c_d w_{\min}/2$. Then the above theorem states there is a λ such that with probability tending to 1 exponentially fast in n, the points in V_m, for any $m = 1, \ldots, K$ are each in the same cluster, and these clusters are distinct, provided that

$$\min_{1 \leq m < m' \leq K} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \frac{16\sqrt{d}\sigma_{\max}}{c_d w_{\min}}. \tag{20}$$

Compared to the Panahi et al. (2017) bound (2), we have removed the dependence of the right-hand side on n as well as the factor of K. (The dependence of the Panahi et al. bound on d is not made explicit so we cannot compare the two bounds' dependence on d. Note that there is still an implicit dependence on K in (20) since necessarily $w_{\min} \leq 1/K$.)

5. Extension to Other Weights

Several authors, for example, Sun et al. (2018) have introduced weights into either the first or second or both summations in (1). One purpose for introducing weights is to be able to eliminate many of the terms in the second summation (that is, use a weight of 0 on those terms) in order to reduce the number of terms in the objective function to $o(n^2)$ for the purpose of efficient computation. For example, Sun et al. use exponentially decaying weights as in (24) below that are zeroed out for $\boldsymbol{a}_i$'s sufficiently far apart. The Chiquet et al. characterization theorem, however, does not extend to fully general weights. (The obstacle is that the left-hand side of (7) does not cancel out the third term on the left-hand side of (6) for general weights.) The most general class of weights for which the theorem

applies is *multiplicative weights*, which are as follows. Each data point $\boldsymbol{a}_i$ for $i = 1, \ldots, n$ is associated with a positive weight r_i. Then both terms in (1) are weighted as follows:

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in \mathbf{R}^d} \frac{1}{2} \sum_{i=1}^{n} r_i \|\boldsymbol{x}_i - \boldsymbol{a}_i\|^2 + \lambda \sum_{1 \le i < j \le n} r_i r_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{21}$$

Therefore, our recovery theorem also extends to multiplicative weights, which is the subject of the rest of this section. A small computational experiment reported in Section 6 suggests recovery of a mixture of Gaussians may also be possible with exponentially decaying weights.

We can draw the same conclusions as Theorem 1 when (3) in the necessary and sufficient conditions is replaced by the following system of equations and inequalities:

$$\begin{aligned} \boldsymbol{a}_i - \sum_{l \in C} \frac{r_l}{\sum_{l' \in C} r_{l'}} \boldsymbol{a}_l = \lambda \sum_{j \in C - \{i\}} r_j \boldsymbol{z}_{ij}^* \quad & \forall i \in C, \\ \left\| \boldsymbol{z}_{ij}^* \right\| \le 1 \quad & \forall i, j \in C, i \ne j, \\ \boldsymbol{z}_{ij}^* = -\boldsymbol{z}_{ji}^* \quad & \forall i, j \in C, i \ne j. \end{aligned} \tag{22}$$

The proof of this generalization is analogous to the proof of Theorem 1, which we omit.

An analogous agglomeration conjecture for this setting was shown by Chiquet et al., that is, the path of solutions to (21) as λ ranges over $[0, \infty)$ contains no splits for multiplicative weights.

With the new theorem of cluster characterization, we can derive the conditions about recovery of a mixture of Gaussians in the case of multiplicative weights, as an extension to Theorem 3. This requires a further modeling assumption on the distribution of the weights. As before, assume each data item $\boldsymbol{a}_i$, $i = 1, \ldots, n$ is chosen from a mixture of K Gaussians. Assume that the weight r_i associated with data item $\boldsymbol{a}_i$ is chosen independently at random according to $r_i \sim \Omega_m$. Here, $m \in \{1, \ldots, K\}$ denotes the specific Gaussian associated with $\boldsymbol{a}_i$. The distributions $\Omega_1, \ldots, \Omega_K$ are all assumed to be supported in a single bounded interval $[0, R]$. Denote the mean of Ω_m as $\bar{r}_m$, $m = 1, \ldots, K$. Assume these means are all positive: $0 < \bar{r}_m \le R$.

The main result is that for any $m = 1, \ldots, K$, with probability exponentially close to 1 (and depending on ϵ) as $n \to \infty$, for the solution $\boldsymbol{x}^*$ computed by (21), the points in V_m are in the same cluster provided that

$$\lambda \ge \frac{2\theta\sigma_m}{(F(\theta, d)w_m - \epsilon) n \bar{r}_m},$$

and the cluster associated with V_m is distinct from the cluster associated with $V_{m'}$, $1 \le m < m' \le K$, provided that

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n-1)(\bar{r} - \epsilon)},$$

where $\bar{r}$ is the overall mean of the r_i's, that is, $\bar{r} = w_1 \bar{r}_1 + \cdots + w_K \bar{r}_K$.

Similar techniques from the proof of Theorem 3 are used to prove the recovery of the multiplicative-weight problem. First, we can construct a solution to (22) as follows

$$\boldsymbol{z}_{ij}^* = \frac{1}{\lambda r'_m} (\boldsymbol{a}_i - \boldsymbol{a}_j) \quad \forall i, j \in V_m, i \ne j,$$

where $r'_m = \sum_{l \in V_m} r_l$. Our task is to prove that the norm bound $\left\| \boldsymbol{z}^*_{ij} \right\| \leq 1$ holds. By definition of V_m, $\|\boldsymbol{a}_i - \boldsymbol{a}_j\| \leq 2\theta\sigma_m$. As before, the probability that $|V_m| \geq (F(\theta, d)w_m - \epsilon_1)n$ is exponentially close to 1 as $n \to \infty$ for a fixed $\epsilon_1 > 0$. Furthermore, the probability that $r'_m \geq (\bar{r}_m - \epsilon_2)|V_m|$ is exponentially close to 1 by Hoeffding's inequality (1963) as $n \to \infty$ for fixed ϵ_2. Thus, provided

$$\lambda \geq \frac{2\theta\sigma_m}{(F(\theta, d)w_m\bar{r}_m - \epsilon)n},$$

we have constructed a solution to (22) with probability exponentially close to 1, which implies that all points in V_m are in the same cluster.

Turn now to the analysis of the upper bound on λ. The first-order optimality conditions of (21) imply the following inequalities by applying the triangle inequality to the summation of subgradients

$$\|\boldsymbol{x}^*_i - \boldsymbol{a}_i\| \leq \lambda \sum_{j \neq i} r_j \quad \forall i. \tag{23}$$

By the same argument in the proof of Theorem 3, there exist at least one $i \in V_m, i' \in V_{m'}$ satisfying the following inequality with probability exponentially close to 1

$$\|\boldsymbol{x}^*_i - \boldsymbol{x}^*_{i'}\| \geq \|\boldsymbol{\mu}_{m'} - \boldsymbol{\mu}_m\| - \lambda \sum_{j \neq i} r_j - \lambda \sum_{j \neq i'} r_j \quad \text{(by (17), (23))}.$$

Therefore, we conclude that $\boldsymbol{x}^*_i \neq \boldsymbol{x}^*_{i'}$, that is, that V_m and $V_{m'}$ are not in the same cluster, provided that for all $i \in V_m, i' \in V_{m'}$

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{\sum_{j \neq i} r_j + \sum_{j \neq i'} r_j}.$$

Applying Hoeffding's bound again, we can claim that for any $\epsilon > 0$, with probability tending to 1 exponentially fast with n, this inequality will hold provided that

$$\lambda < \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{2(n-1)(\bar{r} - \epsilon)}.$$

6. Computational Experiments

In this section, we perform experiments in which a solver for sum-of-norms clustering is applied to a set of points drawn from a mixture of Gaussians. Four experiments are performed to address four questions: (1) How flexibly can λ be chosen? (2) How does the recovery depend on d, the space dimension? (3) How does the recovery degrade as σ (the standard deviation of the Gaussians) increases? and (4) Does the result hold for general weights?

Note that there is no attempt in this section to test sum-of-norms clustering on more general data sets nor to compare it to other clustering algorithms since those topics are outside the scope of this work. For performance of sum-of-norms clustering on more general data sets, we refer the reader to Chi and Lange (2015); Hocking et al. (2011).

In all cases, the code used is our own Julia (Bezanson et al., 2017) implementation of the ADMM solver by Chi and Lange (2015). Each iteration of this solver requires $O(n^2 d)$ operations since the objective function contains $O(n^2)$ terms, each involving vectors

of length d. We observed that the number of iterations to reach a fixed tolerance scales linearly with n. This means that the overall running time scales cubically with n. Our convergence tolerance ϵ_{tol} was taken to be 10^{-6} in all cases. This tolerance corresponds to the quantities ϵ^{pri} and ϵ^{rel} in the supplemental material of Chi and Lange (2015). These parameters correspond to the absolute and relative precisions, which control the primal and dual precisions. The algorithm terminates when the primal and dual residuals are bounded by the precisions respectively. With this tolerance, the runs described below took approximately 27 hours total on an Intel Xeon processor single-threaded.

After termination, clusters were recovered from the approximately converged solution $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n$ as follows. An i is selected arbitrarily from $\{1, \ldots, n\}$. Then all vectors j such that $\|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\| \leq \sqrt{\epsilon_{\text{tol}}}$ are assigned to a cluster. These j's (including i itself) are then deleted from the list of nodes, and the process is repeated until all nodes are used up. Call these recovered clusters $R_1, \ldots, R_{K'}$. The question of how to best retrieve clusters from an approximate solution of (1) is nontrivial, and the first and second authors studied the problem extensively in Jiang and Vavasis (2020).

Then V_m, $m = 1, \ldots, K$, are mapped to one of these recovered clusters, that is, a mapping $\ell : \{1, \ldots, K\} \to \{1, \ldots, K'\}$ is computed such that $R_{\ell(m)}$ contains the most number of elements of V_m. In other words,

$$\ell(m) := \operatorname*{argmax}_{m'=1,\ldots,K'} \#(V_m \cap R_{m'}),$$

for each $m = 1, \ldots, K$, with ties broken arbitrarily. Here, $\#(\cdot)$ denotes set-cardinality. This mapping $\ell(\cdot)$ is not necessarily injective.

Then three scores are computed:

$$s_1 = \frac{1}{\#(V_1 \cup \cdots \cup V_m)} \sum_{m=1}^{K} \#(V_m \cap R_{\ell(m)}),$$

which is the fraction of entries in $V_1 \cup \cdots \cup V_m$ correctly clustered,

$$s_2 = \frac{1}{n} \sum_{m=1}^{K} \#\{i \in \{1, \ldots, n\} : \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{\mu}_m, \sigma_m^2 I) \text{ and } i \in R_{\ell(m)}\},$$

the fraction of entries of all n data points correctly clustered, and

$$s_3 = \frac{\#\ell(\{1, \ldots, K\})}{K},$$

the number of distinct recovered clusters divided by the true number. Note that as λ increases, one would expect s_1 and s_2 to increase while s_3 decreases, since clusters expand as λ increases.

The first experiment is meant to determine whether choices λ outside the range specified by Theorem 3 can still recover clusters. For this experiment we chose $n = 1000$, $d = K = 6$, $w_i = 1/6$ and $\boldsymbol{\mu}_i = \boldsymbol{e}_i$ (ith column of the identity matrix) for $i = 1, \ldots, 6$, $\sigma = 0.0094$, and $\theta = 2.0$. This choice of σ is made so that the upper and lower bounds on λ in Theorem 3 are nearly equal to a single value $\lambda^* = 7.0 \cdot 10^{-4}$. Then we tested recovery for $\lambda = \kappa\lambda^*$ with $\kappa = 1/4, 1/2, 1, 2, 4$, as shown in Table 6.

λ	s_1 (% of V_m recovered)	s_2 (total % recovered)	s_3 (% distinct clusters)
$\lambda^*/4$	38/304	39/1000	6/6
$\lambda^*/2$	304/304	1000/1000	6/6
λ^*	304/304	1000/1000	6/6
$2\lambda^*$	304/304	1000/1000	6/6
$4\lambda^*$	304/304	1000/1000	1/6

Table 1: Recovery for varying λ. Value λ^* is the essentially unique value satisfying the two inequalities of Theorem 3.

The data in Table 6 indicates that the recovery is perfect between $\lambda^*/2$ and $2\lambda^*$. As the theorem predicts, as λ increases, a greater number of V_m's is recovered, but a smaller number of V_m's are distinct. This table suggests that a strengthening may exist of our main theorem in which both inequalities are less restrictive, but not by orders of magnitude.

In the second experiment, we varied d and K. Note that as d and K get larger for fixed n, we move away from the asymptotic range in which Theorem 3 applies since the size of each cluster shrinks. On the other hand, as n is fixed while d and k get larger, we are closer to the range of parameters for which the Panahi et al. result applies. For these tests, we fixed $n = 1000$, looped over $d = K = 4, 16, 64$ and $\theta = \sqrt{d}$ (so that $\theta = 2, 4, 8$). Note that this variation of θ with respect to d is chosen so that $F(\theta, d)$ is about the same value (between .5 and .6) for all three trials. As in the previous experiment, we chose $w_i = 1/K$ and $\boldsymbol{\mu}_i = \boldsymbol{e}_i$ (ith column of the identity matrix) for $i = 1, \ldots, K$. Finally, we chose σ so that the upper and lower bounds in Theorem 3 were equal, and we chose λ to be this unique value of λ. (Note that σ shrinks like $d^{3/2}$ for this variation of parameters.)

We found that in all three cases, all 1000 points were clustered correctly into K distinct clusters (so no table is presented). This robust behavior is not predicted by our theorem, since the arguments in the theorem are weak if n/K is small. See further comments on this matter in Section 7.

The next experiment considers the effect of increasing σ. For this experiment we fixed $d = 1$, $K = 2$, $n = 1000$, $\mu_1 = 0$, $\mu_2 = 1$, $w_1 = w_2 = 1/2$, $\theta = 1$. Let $\lambda_{\max}$ be the value appearing on the right-hand side of (14). In all trials, we fixed $\lambda = \lambda_{\max}$, which does not depend on σ. We chose σ^* to be the value of σ that makes the right-hand sides of (13) and (14) equal. Then we increased σ by factors of $\sqrt{2}$ to observe the effect on recovery. The results appear in Table 2. Note that the method continues to be robust for values of σ modestly outside the range that we have established, but then the behavior quickly decays. It is likely that we could have gotten better performance by carefully tuning λ. The last experiment is a study of exponentially decaying weights, which is a case in which our theory does not apply. Similar to Yuan et al. (2018), we used the following weighting:

$$\min \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{a}_i\|^2 + \lambda \sum_{1 \le i < j \le n} \exp(-\phi \|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2) \|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \tag{24}$$

where $\phi > 0$ is a tuning parameter. Note that for ϕ close to 0, this formulation recovers equal weights, whereas as $\phi \to \infty$, the weights in the second term tend to 0 and hence

σ	s_1 (% of V_m recovered)	s_2 (total % recovered)	s_3 (% distinct clusters)
σ^*	700/700	996/1000	2/2
$2^{1/2}\sigma^*$	700/700	950/1000	2/2
$2\sigma^*$	700/700	742/1000	2/2
$2^{3/2}\sigma^*$	108/700	108/1000	2/2
$4\sigma^*$	46/700	46/1000	2/2

Table 2: Recovery for varying σ. Here, σ^* is the unique value that makes the right-hand sides of (13) and (14) equal.

ϕ	s_1 (% of V_m recovered)	s_2 (total % recovered)	s_3 (% distinct clusters)
500	304/304	999/1000	6/6
1000	304/304	901/1000	6/6
1500	92/304	144/1000	6/6
2000	14/304	14/1000	6/6

Table 3: Recovery for varying ϕ.

each $\boldsymbol{a}_i$ will end up in its own cluster. In the case of Yuan et al. (2018), the exponentially decaying weights are truncated to 0 for points sufficiently far apart in order to improve computational efficiency (by removing most of the terms from the second summation of (1)). However, since our study here does not concern efficiency, we did not truncate any terms. We chose $n = 1000$, $d = K = 6$, $\sigma = .0094$, λ as the unique value that satisfies (13) and (14) if ϕ were zero (equal weights), $\theta = 2$, The results in Table 3 show that for exponentially decaying weights, the correct clusters are recovered provided that ϕ is not too large, that is, the weights do not fall to 0 too quickly.

7. Discussion

The analysis of the mixture of Gaussians in Section 4 used only standard bounds and simple properties of the normal distribution, so it should be apparent to the reader that many extensions of this result (for example, Gaussians with a more general covariance matrix, uniform distributions, many kinds of deterministic distributions) are possible. The key technique is Theorem 1, which essentially decouples the clusters from each other so that each can be analyzed in isolation. Such a theorem does not apply to most other clustering algorithms, or even to sum-of-norm clustering in the case of non-multiplicative weights, so obtaining similar results for other algorithms remains a challenge.

An interesting question concerns the ranges of parameters for which the Panahi et al. result (which requires an upper bound on n), or its extension due to Sun et al. applies versus our bound (which assumes $n \to \infty$). Our result, stated loosely, is that the probability of correct labeling of points a fixed number of standard deviations from the means goes to 1 exponentially fast in n, whereas the other result states that all points are correctly labeled

with probability that goes to 1 exponentially fast in the ratio

$$\frac{\min_{1\leq m<m'\leq K} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\|}{\max_{1\leq m\leq K} \sigma_m}.$$

Is it possible to stitch the two results together into a theorem that encompasses all values of n? One of our computational experiments suggests that this may be possible.

Acknowledgments

We would like to acknowledge support for Research by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada.

References

J. Bezanson, A. Edelman, S. Karpinski, and V.B. Shah. Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1):65–98, 2017.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009.

C. C. Chi and K. Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015. doi: 10.1080/10618600.2014.948181. URL `https://doi.org/10.1080/10618600.2014.948181`. PMID: 27087770.

E. Chi and S. Steinerberger. Recovering trees with convex clustering. 2018. URL `https://arxiv.org/abs/1806.11096`.

J. Chiquet, P. Gutierrez, and G. Rigaill. Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26:205–216, 2017.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2012.

T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: An algorithm for clustering using convex fusion penalties. In *International Conference on Machine Learning*, 2011.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Stat. Assoc*, 58:13–30, 1963.

T. Jiang and S. Vavasis. On identifying clusters from sum-of-norms clustering computation. 2020. URL `https://arxiv.org/abs/2006.11355`.

F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2011.

D. G. Mixon, S. Villar, and R. Ward. Clustering Subgaussian Mixtures by Semidefinite Programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 03 2017. ISSN 2049-8764. doi: 10.1093/imaiai/iax001. URL `https://doi.org/10.1093/imaiai/iax001`.

A. Panahi, D. Dubhashi, F. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. *Journal of Machine Learning Research*, 70, 2017.

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Convex cluster shrinkage. 2005. URL `ftp://ftp.esat.kuleuven.ac.be/sista/kpelckma/ccs_pelckmans2005.pdf`.

J. Peng and Y. Wei. Approximating K-means-type clustering via semidefinite programming. *SIAM J. Optim.*, 18(1):186–205, 2007.

P. Radchenko and G. Mukherjee. Convex clustering via l1 fusion penalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1527–1546, 2017.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantees and efficient algorithm. 2018. URL `https://arxiv.org/abs/1810.02677`.

K.-M. Tan and D. Witten. Statistical properties of convex clustering. *Electron. J. Statist.*, 9(2):2324–2347, 2015. doi: 10.1214/15-EJS1074. URL `https://doi.org/10.1214/15-EJS1074`.

Y. Yuan, D. Sun, and K.-C. Toh. An efficient semismooth Newton based algorithm for convex clustering. 2018. URL `https://arxiv.org/abs/1802.07091`.

C. Zhu, H. Xu, C. Leng, and S. Yan. Convex optimization procedure for clustering: Theoretical revisit. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1619–1627. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5307-convex-optimization-procedure-for-clustering-theoretical-revisit.pdf`.