Finding low error clusterings

Maria-Florina Balcan

Microsoft Research, New England
One Memorial Drive, Cambridge, MA
mabalcan@microsoft.com

Mark Braverman

Microsoft Research, New England
One Memorial Drive, Cambridge, MA
markbrav@microsoft.com

Abstract

A common approach for solving clustering problems is to design algorithms to approximately optimize various objective functions (e.g., k-means or min-sum) defined in terms of some given pairwise distance or similarity information. However, in many learning motivated clustering applications (such as clustering proteins by function) there is some unknown target clustering; in such cases the pairwise information is merely based on heuristics and the real goal is to achieve low error on the data. In these settings, an arbitrary c-approximation algorithm for some objective would work well only if any c-approximation to that objective is close to the target clustering. In recent work, Balcan et. al [7] have shown how both for the k-means and k-median objectives this property allows one to produce clusterings of low error, even for values c such that getting a c-approximation to these objective functions is provably NP-hard.

In this paper we analyze the min-sum objective from this perspective. While [7] also considered the min-sum problem, the results they derived for this objective were substantially weaker. In this work we derive new and more subtle structural properties for min-sum in this context and use these to design efficient algorithms for producing accurate clusterings, both in the transductive and in the inductive case. We also analyze the correlation clustering problem from this perspective, and point out interesting differences between this objective and k-median, k-means, or min-sum objectives.

1 Introduction

Problems of clustering data from pairwise distance or similarity information are ubiquitous in science. A common approach for solving such problems is to view the data points as nodes in a weighted graph (with the weights based on the given pairwise information), and then to design algorithms to optimize various objective functions such as k-means or min-sum. For example, in the min-sum clustering approach the goal is to produce a partition into a given number of clusters k that minimizes the sum of the intracluster distances. Many of the optimization problems corresponding to commonly analyzed objectives (including k-means, min-sum, k-median, or correlation clustering) are NP-hard and so the focus in the theory community has been in designing approximation algorithms for these objectives.[1] For example the best known approximation algorithm for the k-median problem is a $(3+\epsilon)$-approximation [6], while the best approximation for the min-sum problem in general metric spaces is a $O(\log^{1+\delta} n)$-approximation. For many of these problems the approximation guarantees do not match the known hardness results, and significant effort is spent on obtaining tighter approximation guarantees and hardness results [3, 6, 9, 11, 13, 15, 12, 17, 21, 25, 28].

Standard clustering settings used to motivate much of this effort include problems such as clustering proteins by function, images by subject, or documents by topic. In many of these settings there is some unknown correct "target" clustering and the implicit hope is that approximately optimizing objective functions such as those mentioned above will in fact produce a clustering of low error, i.e. a clustering which agrees with the truth on most of the points. In other words, implicit in taking the approximation-algorithms approach is the hope that any c-approximation to our given objective will be pointwise close to the true answer, and our motivation for improving a c_2-approximation to a c_1-approximation (for $c_1 < c_2$) is that perhaps this closeness property holds for c_1 but not c_2.

In recent work, Balcan et. al [7] have shown that if we make this implicit assumption explicit, then one can get accurate clusterings even in cases where getting a good approximation to these objective functions is provably NP-hard. In particular, say that a data set satisfies the (c,ϵ) property for some objective function Φ if any c-approximation to Φ on this data must be ϵ-close to the target clustering. [7] show that for any $c = 1+\alpha > 1$, if data satisfies the (c,ϵ) property for the k-median (or k-means) objectives, then one can produce clusterings that are $O(\epsilon)$-close to the target, *even for values c for which obtaining a c-approximation is NP-hard*. [7] also consider the min-sum objective, however the results they present work only for values of $c > 2$ and under the

[1] A β-approximation algorithm for objective ϕ is an algorithm that runs in polynomial time and returns a solution whose value is within a multiplicative β factor of the optimal solution for the given objective ϕ.

assumption that all the target clusters are large.

1.1 Our Results

In this work we solve the problem of getting accurate clusterings for the min-sum objective under the (c, ϵ)-assumption, improving on the results of Balcan et. al [7] for this objective in multiple respects. In particular, we show it is possible to deal with any constant $c = 1 + \alpha > 1$ (and not only $c > 2$). More importantly we are also able to deal with the presence of small target clusters. To achieve this we derive new and much more subtle structural properties implied by the (c, ϵ)-assumption. In the case where k is small compared to $\log n / \log \log n$ we output a single clustering which is $O(\epsilon/\alpha)$-close to the target, while in the general case our algorithm outputs a small list of clusterings with the property that the target clustering is close to one of those in the list.

We show that the algorithm we develop for the min-sum objective is robust, which allows us to extend it to the *inductive* model. In the inductive model S is merely a small random subset of points from a much larger abstract instance space X, and our goal is to produce a hypothesis $h : X \to Y$ which implicitly represents a clustering of the whole space X and which has low error on the whole X. An appealing characteristic of the algorithm we obtain for the inductive case is that the insertion of new points (which arrive online) is extremely efficient: we only need $O(k)$-comparisons for assigning a new point x to one of the clusters.

We further show that if we do require the clusters to be large, we can reduce the approximation error from $O(\epsilon/\alpha)$ down to $O(\epsilon)$ – the best one can hope for. We thus affirmatively answer several open questions in [7].

We also analyze the correlation clustering problem in this framework. In correlation clustering, the input is a graph with edges labeled $+1$ or -1 and the goal is to find a partition of the nodes that best matches the signs of the edges. This clustering formulation was introduced by Bansal et. al in [11] and it has been extensively studied in a series of follow-up papers both in the theoretical computer science and in the machine learning community [3, 13, 22]. In the original paper Bansal et al. [11] considered two versions of the correlation clustering problem, minimizing disagreements and maximizing agreements.[2] In this paper we focus on the minimizing disagreements objective function. (The maximizing agreement version of correlation clustering is less interesting in our framework since it admits a PTAS[3].) We show that this objective behaves much better than objectives such as k-median, k-means, and min-sum in terms of error rate. More specifically, we show that for this objective, the $(1 + \alpha, \epsilon)$ property implies a $(2.5, O(\epsilon/\alpha))$ property, so one can use a state-of-the-art 2.5-approximation algorithm for minimizing disagreements in order to get an accurate clustering. This contrasts sharply with the previous results proven in this context for objectives such as min-sum, k-median, or k-means.

Our work shows how for a clustering objective such as min-sum we can obtain results comparable to what one could obtain by being able to approximate the objective to an arbitrary small constant. In other words if what we really want is to obtain a clustering of low error, then by making implicit assumptions explicit we can obtain low error clusterings even in cases where getting a c-approximation to the min-sum objective is NP-hard. This points out how one can get much better results than those obtained so far in the approximation algorithms literature by wisely using all the available information for the problem at hand.

1.2 Related Work

Work on approximation algorithms: We review in the following state of the art results on approximation algorithms for the two clustering objectives we discuss in this paper.

Min-sum k-clustering on general metric spaces admits a PTAS for the case of constant k by Fernandez de la Vega et al. [17] (see also [20]). For the case of arbitrary k there is an $O(\delta^{-1} \log^{1+\delta} n)$-approximation algorithm that runs in time $n^{O(1/\delta)}$ due to Bartal et al. [9]. The problem has also been studied in geometric spaces for constant k by Schulman [28] who gave an algorithm for $(\mathbb{R}^d, \ell_2^2)$ that either outputs a $(1 + \epsilon)$-approximation, or a solution that agrees with the *optimum* clustering on $(1 - \epsilon)$-fraction of the points (but could have much larger cost than optimum); the runtime is $O(n^{\log \log n})$ in the worst case and linear for sublogarithmic dimension d. More recently, Czumaj and Sohler have developed a $(4 + \epsilon)$-approximation algorithm for the case when k is small compared to $\log n / \log \log n$ [15].

Correlation Clustering was introduced by Bansal et. al in [11]. In the original paper Bansal et al. [11] have considered two versions of the correlation clustering problem, minimizing disagreements and maximizing agreements, focusing mainly on the case when the graph G is complete. They gave a polynomial time approximation scheme (PTAS) for the maximizing agreements version on complete graphs, while for the minimizing disagreements versions, they gave an approximation algorithm with a constant performance ratio. The constant was a rather large one, and it has subsequently improved to 4 in [13] and then to 2.5 in [3]. In the case when the graph is not complete, the best known approximation is $O(\log n)$ [13].

Other work on Clustering: Our work is most relevant for settings where there is a target clustering and it is motivated by results in [8] which have investigated the goal of approximating a desired target clustering without making any probabilistic assumptions. In addition to this, there has been significant work in machine learning and theoretical computer science on clustering or learning with mixture models [1, 5, 19, 18, 23, 29, 16]. That work, like ours, has an explicit notion of a correct ground-truth clustering of the data points; however, it makes very specific probabilistic assumptions about the data.

There is a large body of other work which does not assume the existence of a target clustering. For example there

[2] In the former case, the goal is to minimize the number of -1 edges inside clusters plus the number of $+1$ edges between clusters, while in the latter case the goal is to maximize the number of $+1$ edges inside the cluster plus the number of -1 edges between. These are equivalent at optimality but differ in their difficulty of approximation.

[3] A PTAS (polynomial-time approximation scheme) is an algorithm that for any given fixed ϵ runs in polynomial time and returns an approximation within a $1 + \epsilon$ factor. Running time may depend exponentially (or worse) on $1/\epsilon$, however

has been work on axiomatizing clustering (in the sense of postulating what natural axioms should a good clustering algorithm or quality measure satisfy), both with possibility [2] and impossibility [24] results, on comparing clusterings [26, 27], and on efficiently testing if a given data set has a clustering satisfying certain properties [4]. The main difference between this type of work and our work is that we have an explicit notion of a correct ground-truth clustering of the data points, and indeed the results we are trying to prove are quite different.

Inductive Setting: In the inductive setting, where we imagine our given data is only a small random sample of the entire data set, our framework is close in spirit to recent work done on sample-based clustering (e.g., [10, 14]) in the context of clustering algorithms designed to optimize a certain objective. Based on such a sample, these algorithms have to output a clustering of the full domain set, that is evaluated with respect to the underlying distribution.

2 Definitions and Preliminaries

The clustering problems in this paper fall into the following general framework. We are given a set S of n points which we want to cluster. We are also given a pairwise similarity and/or dissimilarity information expressed through a weighted graph (G, d) on S. A k-clustering $\mathcal{C}$ is a partition of S into k sets $C_1, C_2, \ldots, C_k$. In this paper, we always assume that there is a *true* or *target* k-clustering $\mathcal{C}_T$ for the point set S.

A natural notion of distance between two k-clusterings $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ and $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_k\}$ which we use throughout the paper is the fraction of points on which they disagree under the optimal matching of clusters in $\mathcal{C}$ to clusters in $\mathcal{C}'$; i.e., we define

$$dist(\mathcal{C}, \mathcal{C}') = \min_{\sigma \in S_k} \frac{1}{n} \sum_{i=1}^{k} |C_i - C'_{\sigma(i)}|,$$

where S_k is the set of bijections $\sigma : [k] \to [k]$. We say that two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are ϵ-*close* if $dist(\mathcal{C}, \mathcal{C}') < \epsilon$ and we say that a clustering has *error* ϵ if it is ϵ-close to the target. We can also define the distance $dist(\mathcal{C}, \mathcal{C}')$ between two clusterings $\mathcal{C} = \{C_1, C_2, \ldots, C_{k_1}\}$ and $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_{k_2}\}$ with a different number of clusters k_1 and k_2 where $k_1 > k_2$ by simply extending the clustering $\mathcal{C}'$ with a few empty clusters and then using the notion of distance defined above. We will now state a useful fact about the distance between two clusterings which we use throughout the paper and which is a simple consequence of the definition:

Fact 1 *Given two clusterings $\mathcal{C}$ and $\mathcal{C}'$, if we produce a list L of disjoint subsets $S_1, S_2, \ldots$, such that for each i, all points in S_i are in the* same *cluster in one of $\mathcal{C}$ or $\mathcal{C}'$ and they are all in* different *clusters in the other, then $\mathcal{C}$ and $\mathcal{C}'$ must have distance at least $\frac{1}{n}\sum_i(|S_i| - 1)$.*

In many cases we will use Fact 1 on sets $\{S_i\}$ of size 2.

We consider two commonly used clustering algorithms which seek to minimize some objective function or "score".

Min-sum clustering The first one is the *min-sum clustering* problem [17, 9]. Here $d : \binom{S}{2} \to \mathbb{R}_{\geq 0}$ is a distance function and the goal is to find a clustering that minimizes

$$\Phi_\Sigma := \sum_{i=1}^{k} \sum_{x,y \in C_i} d(x, y).$$

In this paper we focus on the case where d satisfies the triangle inequality, and we also discuss a few extensions of this condition.

Correlation clustering The second clustering setup we analyze is correlation clustering introduced in [11]. In this setting the graph G is fully connected with edges (x, y) labeled $d(x, y) = +1$ (similar) or $d(x, y) = -1$ (different). The goal is to find a partition of the vertices into clusters that agrees as much as possible with the edge labels. In particular, the Min-Disagreement correlation clustering objective (Min-Disagreement CC) asks to find a clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_{k'}\}$ to minimize the objective function – the number of disagreements (the number of -1 edges inside clusters plus the number of $+1$ edges between clusters):

$$\begin{aligned} \Phi_{CC} &:= \#\{x, y \in C_i : d(x, y) = -\} \\ &+ \#\{x \in C_i,\ y \in C_j,\ i \neq j : d(x, y) = +\}. \end{aligned}$$

Note that in the correlation clustering setting, the target number of clusters is not specified as part of the input.

Given a function Φ (such as k-median or min-sum) and instance (S, d), let

$$\mathrm{OPT}_\Phi = \min_{\mathcal{C}} \Phi(\mathcal{C}),$$

where the minimum is over all k-clusterings of (S, d).

The (c, ϵ)-property The following notion originally introduced in [7] is central to our discussion:

Definition 2 *Given an objective function Φ (such as k-median or min-sum), and $c = 1 + \alpha > 1$, $\epsilon > 0$, we say that instance (S, d)* satisfies the (c, ϵ)-property for Φ *if all clusterings $\mathcal{C}$ with $\Phi(\mathcal{C}) \leq c \cdot \mathrm{OPT}_\Phi$ are ϵ-close to the target clustering $\mathcal{C}_T$ for (S, d).*

Note that for any $c > 1$, the (c, ϵ)-property does not require that the target clustering $\mathcal{C}_T$ exactly coincide with the optimal clustering $\mathcal{C}^*$ under objective Φ. However, it does imply the following simple facts:

Fact 3 *If (S, d) satisfies the (c, ϵ)-property for Φ, then:*

(a) The target clustering $\mathcal{C}_T$, and the optimal clustering $\mathcal{C}^*$ are ϵ-close.

(b) The distance between k-clusterings is a metric, and hence a (c, ϵ) property with respect to the target clustering $\mathcal{C}_T$ implies a $(c, 2\epsilon)$ property with respect to the optimal clustering $\mathcal{C}^*$.

Thus, we can act as if the optimal clustering is indeed the target up to a constant factor loss in the error rate. For simplicity, we will assume throughout the paper (except in Section 4) that $\mathcal{C}_T$ is indeed the optimal clustering $\mathcal{C}^*$.

3 The Min-sum Clustering Problem

Recall that the min-sum k-clustering problem asks to find a k-clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ to minimize the objective function:

$$\Phi(\mathcal{C}) = 2\sum_{i=1}^{k} \sum_{x,y \in C_i} d(x,y).$$

We focus here on the case where $d : \binom{S}{2} \to \mathbb{R}_{\geq 0}$ is a distance function satisfying the triangle inequality. As shown in [7] we have the following:

Theorem 4 *[7] For any $1 \leq c_1 < c_2$, any $\epsilon, \delta > 0$, there exists a family of metric spaces G and target clusterings that satisfy the (c_1, ϵ) property for the min-sum objective and yet do not satisfy even the $(c_2, 1/2 - \delta)$ property for that objective.*

So, in the min-sum objective case it is *not* the case that if the data satisfies the (c_1, ϵ) property, then then we can use c_2 approximation algorithm in order to get a clustering of small error rate, for some $c_2 > c_1$. [7] also shows the following:

Theorem 5 *For the min-sum objective the problem of finding a c-approximation can be reduced to the problem of finding a c-approximation under the (c, ϵ) assumption. Therefore, the problem of finding a c-approximation under the (c, ϵ) assumption is as hard as the problem of finding a c-approximation in general.*

Theorem 5 means that, generally speaking, the (c, ϵ) assumption does not make optimizing min-sum easier.

General overview of our construction. The general idea for our construction is to obtain various structural properties for instances that satisfy the (c, ϵ) assumption, and then to use these properties to give an efficient algorithm for achieving low error clusterings. These structural properties are essential since as mentioned above the general min-sum clustering problem is APX-hard.

The structural properties stem from the fact that under the the (c, ϵ) assumption the optimal solution is fairly "stable": changing it a little increases the cost substantially. The first key property (Lemma 6) we prove using this stability is that most pairs of clusters are quite expensive to merge. Using a vertex-cover argument on a specially designed graph on clusters, we show that we can remove few ($O(\epsilon n)$) points such that no two clusters among the remaining points are cheap to merge.

Next we show that for most of the remaining points x we can draw a ball B_x of an appropriate radius that essentially covers the cluster $C^*(x)$. Such balls B_x and B_y usually do not overlap when $C^*(x) \neq C^*(y)$ (Lemma 9) since such an overlap would mean that x and y are sufficiently close to merge $C^*(x)$ with $C^*(y)$ cheaply, leading to a contradiction. We designate a class of "good" points for which the above is true. All but $O(\epsilon n/\alpha)$ points are good.

We introduce subsets $\tilde{\mathrm{B}}_x \subset \mathrm{B}_x$ to make sure that the ball around each x (whether good or bad) contains only good points from at most one cluster. At the same time, $\tilde{\mathrm{B}}_x$ centered around a good point x still covers the bulk of the cluster $C^*(x)$. The algorithm then uses a greedy covering using the $\{\tilde{\mathrm{B}}_x\}_{x \in S}$ to perform the actual clustering. The analysis of the clustering produced is done using a careful charging argument. It shows that the final clustering is $O(\epsilon n/\alpha)$ close to the optimal.

While the analysis is quite involved, the clustering algorithm itself is simple, robust, and efficient. This simplicity and robustness allows us to extend it to the inductive setting.

3.1 Properties of Min-Sum Clustering

We start by deriving a few structural properties implied by the $(1+\alpha, \epsilon)$-property for min-sum. We emphasize that all the constructions in this subsection are for the analysis purpose; our algorithmic results are described in Section 3.2.

Recall that $\mathcal{C}^*$ denotes the optimal clustering. For $x \in C_i^*$, define $w(x) = \sum_{y \in C_i^*} d(x,y)$ and let

$$w = \mathrm{avg}_x w(x) = \frac{\mathrm{OPT}}{n}.$$

We start by creating a badness graph $\mathcal{G} = (V, E)$ on the set of clusters, connecting pairs of clusters that are not too expensive to merge. Formally, V is the set $\{C_1^*, \ldots, C_k^*\}$ and for any two clusters, C_i^* and C_j^* we add an edge e between them if the additional cost incurred for merging them is at most

$$(|C_i^*| + |C_j^*|)\frac{w\alpha}{2\epsilon}.$$

Lemma 6 *Assume that the min-sum instance (S, d) satisfies the $(1+\alpha, \epsilon)$-property with respect to the target clustering. Then we can remove $< 3\epsilon n$ points such that the remaining set of clusters form an independent set in $\mathcal{G}$.*

Proof: We start by showing that if the min-sum instance (S, d) satisfies the $(1+\alpha, \epsilon)$-property with respect to the target clustering, then there cannot be a collection of disjoint cluster pairs $(C_{i_1}^*, C_{j_1}^*), (C_{i_2}^*, C_{j_2}^*), \ldots, (C_{i_r}^*, C_{j_r}^*)$ such that $C_{i_l}^*$ and $C_{j_l}^*$ are connected in $\mathcal{G}$ and

$$\sum_l (|C_{i_l}^*| + |C_{j_l}^*|) \geq 3\epsilon n.$$

Let $C_{i_l}^*$ and $C_{j_l}^*$ be two clusters such that the additional cost incurred for merging them is at most $(|C_{i_l}^*| + |C_{j_l}^*|)\frac{w\alpha}{2\epsilon}$. Assume w.l.o.g. that $|C_{i_l}^*| \leq |C_{j_l}^*|$. We show now that for any set size s there exists a subset A_s of $C_{j_l}^*$ of size s which we can move from $C_{j_l}^*$ to $C_{i_l}^*$ at an additional cost in the min-sum objective of at most $|A_s|\frac{w\alpha}{\epsilon}$. First note that

$$\sum_{x \in C_{j_l}^*} \sum_{y \in C_{i_l}^*} d(x,y) \leq |C_{j_l}^*|\frac{\alpha w}{\epsilon}.$$

Let $\alpha(x) = \sum_{y \in C_{i_l}^*} d(x,y)$. So

$$\frac{\sum_{x \in C_{j_l}^*} \alpha(x)}{|C_{j_l}^*|} \leq \frac{\alpha w}{\epsilon}.$$

Hence, for any set size $s = \epsilon_l n$ we can select a subset A_s of of $C_{j_l}^*$ (namely the first s elements with the smallest values of $\alpha(x)$) such that

$$\frac{\sum_{x \in A_s} \alpha(x)}{|A_s|} \leq \frac{\alpha w}{\epsilon}.$$

This implies that for any set size s there exists a subset A_s of $C^*_{j_l}$ which we can move from $C^*_{j_l}$ to $C^*_{i_l}$ at an additional cost in the min-sum objective of at most $|A_s|\frac{w\alpha}{\epsilon}$, as desired.

Assume that there exists r and a collection of disjoint pairs $(C^*_{i_1}, C^*_{j_1}), (C^*_{i_2}, C^*_{j_2}), \ldots, (C^*_{i_r}, C^*_{j_r})$ such that $|C^*_{i_l}| \leq |C^*_{j_l}|$, $C^*_{i_l}$ and $C^*_{j_l}$ are connected in $\mathcal{G}$, and

$$\sum_l (|C^*_{i_l}| + |C^*_{j_l}|) \geq 3\epsilon n.$$

Let $A_l \subseteq C^*_{j_l}$ of size

$$\epsilon_l n = \min(\epsilon n - \sum_{s<l} \epsilon_s n, \max(|C^*_{i_l}|, |C^*_{j_l}|/2)).$$

Since

$$\max(|C^*_{i_l}|, |C^*_{j_l}|/2) \geq \frac{1}{3}(|C^*_{i_l}| + |C^*_{j_l}|)$$

and

$$\sum_l (|C^*_{i_l}| + |C^*_{j_l}|) \geq 3\epsilon n$$

we have that

$$\sum_l |A_l| = \sum_l \epsilon_l n = \epsilon n.$$

Let $\mathcal{C}'$ be the clustering obtained by moving A_l from $C^*_{j_l}$ to $C^*_{i_l}$ for $l = 1, \ldots, r$. Each movement of a set A_l from $C^*_{j_l}$ to $C^*_{i_l}$ increases the distance between $\mathcal{C}$ and $\mathcal{C}'$ by ϵ_l. To prove this we use Fact 1 and do a case analysis. If $\epsilon_l n = |C^*_{i_l}|$, then we match up each point x_i in $C^*_{i_l}$ with point y_i in A_l and we define the set S_i as $\{x_i, y_i\}$. If $\epsilon_l n = |C^*_{j_l}|/2$ then we split $C^*_{j_l}$ into two sets $C^{*1}_{j_l}$ and $C^{*2}_{j_l}$ of equal size and match up a each point in x_i in $C^{*1}_{i_l}$ with a point y_i in $C^{*2}_{i_l}$ and then define the set S_i as $\{x_i, y_i\}$. If $\epsilon_l n < \max(|C^*_{i_l}|, |C^*_{j_l}|/2)$, then we apply either of the constructions above. In all cases we produce a list of disjoint subsets $S_1, S_2, \ldots$, such that for each i, all points in S_i are in the *same* cluster in one of $\mathcal{C}$ or $\mathcal{C}'$ and they are all in *different* clusters in the other. Using Fact 1 we obtain that by moving the set A_l from $C^*_{j_l}$ to $C^*_{i_l}$ we increase the distance between $\mathcal{C}$ and $\mathcal{C}'$ by $\frac{1}{n}\sum_i(|S_i| - 1) = \epsilon_l$. Overall we get $dist(\mathcal{C}, \mathcal{C}') = \epsilon = \sum_l \epsilon_l$. We also have

$$\begin{aligned}\Phi(\mathcal{C}') &= \Phi(\mathcal{C}) + \sum_l (\alpha w/\epsilon) \cdot (\epsilon_l n) \\ &\leq \Phi(\mathcal{C}) + \alpha w n \\ &\leq \Phi(\mathcal{C}) + \alpha \mathrm{OPT}.\end{aligned}$$

We thus obtain that $\mathcal{C}'$ which is ϵ-far from the target and whose min-sum cost is within a $1+\alpha$ factor of OPT, contradicting the $(1+\alpha, \epsilon)$-property.

To finish the argument let $\mathcal{L}$ be the output of the greedy vertex cover on the graph $\mathcal{G}$. Specifically, let $\mathcal{L}$ be the list of clusters constructed as follows: pick an arbitrary edge e in $\mathcal{G}$, add both vertices incident to edge e in the list $\mathcal{L}$, delete any edge sharing a vertex with e, and repeating until the graph is out of edges. Note that $\mathcal{L}$ is a vertex cover in $\mathcal{G}$; this is because taking both of the vertices incident to a given each edge in the list $\mathcal{L}$ and we only delete edges incident to one of these vertices, and eventually delete all the edges. Since $\mathcal{L}$ is a vertex cover in $\mathcal{G}$, we have that for any pair (C', C'') which forms an edge in $\mathcal{G}$, either $C' \in \mathcal{L}$ or $C'' \in \mathcal{L}$, so the remaining set of clusters $\mathcal{C} \setminus \mathcal{L}$ is an independent set of $\mathcal{G}$. Since $\mathcal{L}$ is a collection of disjoint edges we also have (according to what we have proved above) $\sum_{C\in\mathcal{L}} |C| \leq 3\epsilon n$. This concludes the proof. ■

If $C^*_i \in \mathcal{L}$ in the above proof, let $H^1_i = C^*_i$. Else let $H^1_i = \emptyset$. By Lemma 6 we have that $\sum_{i=1}^k |H^1_i| \leq 3\epsilon n$ and that the cost of merging two clusters that have not been removed is low. This last condition implies the following:

Lemma 7 *For any two $x \in C^*_i \setminus H^1_i$, $y \in C^*_j \setminus H^1_j$, $w(x) \leq \frac{\alpha w}{15\epsilon}$, $w(y) \leq \frac{\alpha w}{15\epsilon}$, we have*

$$d(x,y) \geq \frac{\alpha w}{3\epsilon} \cdot \frac{1}{\min\left(|C^*_i|, |C^*_{j_l}|\right)}.$$

Proof: Assume there exist $x \in C^*_i$, $y \in C^*_j$, $w(x) \leq \frac{\alpha w}{15\epsilon}$, $w(y) \leq \frac{\alpha w}{15\epsilon}$, s.t.

$$d(x,y) \leq \frac{\alpha w}{3\epsilon} \cdot \frac{1}{\min\left(|C^*_i|, |C^*_j|\right)}.$$

Note the additional cost incurred in the minsum objective by merging C^*_i and C^*_j is at most

$$\begin{aligned}&\leq \sum_{x'\in C^*_i} \sum_{y'\in C^*_j} d(x', y') \\ &\leq \sum_{x'\in C^*_i} \sum_{y'\in C^*_j} (d(x', x) + d(x, y) + d(y, y')).\end{aligned}$$

Therefore the additional cost incurred in the minsum objective by merging C^*_i and C^*_j is at most

$$\begin{aligned}&\leq |C^*_j| w(x) + |C^*_i| w(y) + \frac{\alpha w}{3\epsilon} \cdot \frac{|C^*_i| \cdot |C^*_j|}{\min\left(|C^*_i|, |C^*_{j_l}|\right)} \\ &= |C^*_j| w(x) + |C^*_i| w(y) + \frac{\alpha w}{3\epsilon} \cdot \max\left(|C^*_i|, |C^*_j|\right) \\ &\leq (|C^*_j| + |C^*_i|)\left(\frac{\alpha w}{15\epsilon} + \frac{\alpha w}{3\epsilon}\right) \\ &< (|C^*_i| + |C^*_j|)\frac{w\alpha}{2\epsilon},\end{aligned}$$

which contradicts Lemma 6 and the definition of H^1_i. ■

For all x, let us now define τ_x and B_x that will be used in Algorithm 1. To obtain τ_x, we start $\tau = 0$ and gradually increase it until $|\mathrm{B}(x,\tau)| \geq \frac{1}{20}\frac{\alpha w}{\epsilon\tau}$; once this happens we set $\tau_x = \tau$ and $\mathrm{B}_x = \mathrm{B}(x, \tau_x)$. We can now show the following.

Lemma 8 *For any point x such that $w(x) \leq \frac{\alpha w}{15\epsilon}$ we have $\tau_x \leq \frac{\alpha w}{6\epsilon|C^*_i|}$.*

Proof: Since $w(x) = \sum_{y\in C^*_i} d(x,y) \leq \frac{\alpha w}{15\epsilon}$, we have that at least $|C^*_i|/2$ points in a $\tau = \frac{\alpha w}{6\epsilon|C^*_i|}$ neighborhood of x. This implies

$$|\mathrm{B}(\tau, x)| \cdot \tau > \frac{\alpha w}{12\epsilon} > \frac{\alpha w}{20\epsilon},$$

so $\tau_x \leq \tau$ as desired. ■

Lemma 9 *For any two points $x \in C_i^* \setminus H_i^1$, $y \in C_j^* \setminus H_j^1$, such that $w(x) \leq \frac{\alpha w}{15\epsilon}$, $w(y) \leq \frac{\alpha w}{15\epsilon}$, we have $\mathrm{B}_x \cap \mathrm{B}_y = \emptyset$.*

Proof: By Lemmas 7 and 8 we have

$$\tau_x + \tau_y \leq \frac{\alpha w}{6\epsilon}\left(\frac{1}{|C_i^*|} + \frac{1}{|C_j^*|}\right) \leq \frac{\alpha w}{3\epsilon} \cdot \frac{1}{\min\left(|C_i^*|, |C_j^*|\right)} < d(x,y),$$

which together with Lemma 8 implies the desired result. ■

Let us denote by $H_i^2 = \left\{x \in C_i^* \setminus H_i^1 : w(x) > \frac{\alpha w}{15\epsilon}\right\}$ and $H^2 = \cup_i H_i^2$. Since $\mathbf{E}[w(x)] = w$, by Markov inequality, we have

$$|H^2| \leq \frac{15\epsilon}{\alpha} n.$$

3.2 Algorithm for Min-Sum Clustering

In this section, we show that if our data satisfies the $(1+\alpha, \epsilon)$-property for the min-sum objective, then we can find a clustering that is $O(\epsilon/\alpha)$-close to the target $\mathcal{C}_T$. We start by considering the case where we know the value of OPT or $w = \mathrm{OPT}/n$ and we then show how to get rid of this assumption in Theorem 10.

For the case of known w we show in the following that Algorithm 1 can be used to produces a clustering that is $O(\epsilon/\alpha)$-close to the target. In this algorithm we define *critical thresholds* $\tau_0, \tau_1, \tau_2, \ldots$ as: $\tau_0 = 0$ and τ_i is the ith smallest distinct distance $d(x,y)$ for $x, y \in S$. We can show the following.

Algorithm 1 Min-Sum Algorithm

Input: (S,d), w, $\epsilon \leq 1$, $\alpha > 0$, k.

For all x do:

- Let the initial threshold $\tau = \tau_0$.
- Construct the ball $\mathrm{B}(x,\tau)$ by including all points within distance τ of x.
- **If** $|\mathrm{B}(x,\tau)| \geq \frac{1}{20}\frac{\alpha w}{\epsilon \tau}$ **then** let $\tau_x = \tau$ and $\mathrm{B}_x = \mathrm{B}(x,\tau_x)$ **else** increase τ to the next critical threshold

For all x, let $\tilde{\mathrm{B}}_x := \{y : x \in \mathrm{B}_y, y \in \mathrm{B}_x\}$; set $\mathcal{L} = \emptyset$.

For $i = 1 \ldots k$ do

- Let C_i^o be the largest $\tilde{\mathrm{B}}_x$.
- Add C_i^o to $\mathcal{L}$.
- For all $x' \neq x$, set $\tilde{\mathrm{B}}_{x'} = \tilde{\mathrm{B}}_{x'} \setminus C_i^o$.

Output: Clustering $\mathcal{L}$.

Note that in the B_x construction phase one can alternatively sort the points by their distance from x and add them to $\mathrm{B}(x,\tau)$ one-by-one instead of using critical thresholds.

Theorem 10 *If the min-sum instance (S,d) satisfies the $(1+\alpha,\epsilon)$-property and we are given the value of w, then Algorithm 1 produces a clustering that is $O(\epsilon/\alpha)$-close to the target.*

Proof: We first note that the sets B_x in Algorithm 1 are well defined since for small τ the condition $|\mathrm{B}(x,\tau)| \geq \frac{1}{20}\frac{\alpha w}{\epsilon\tau}$ is obviously false and for very large τ the condition is obviously true because $|\mathrm{B}(x,\tau)| \geq 1$.

For all i, let c_i^* be a point in C_i^* that minimizes $\sum_{x \in C_i^*} d(x, c_i^*)$. By triangle inequality, all $x \in C_i^*$ satisfy

$$w(x) \geq |C_i^*| d(x, c_i^*) - w(c_i^*).$$

Moreover, if $x \in C_i^*$ and $d(x, c_i^*) \geq \frac{\alpha w}{60\epsilon |C_i^*|}$ then

$$w(x) \geq \frac{\alpha w}{60\epsilon} - w(c^*) \geq \frac{\alpha w}{60\epsilon} - w(x),$$

which implies $w(x) \geq \frac{\alpha w}{120\epsilon}$.

Let

$$G_i = \left\{x \in C_i^* \setminus (H_i^1 \cup H_i^2), d(x, c_i^*) < \frac{\alpha w}{60\epsilon |C_i^*|}\right\},$$

and let $G = \cup_i G_i$. Let

$$H_i^3 = \left\{x \in C_i^* : d(x, c_i^*) \geq \frac{\alpha w}{60\epsilon |C_i^*|}\right\}$$

and $H^3 = \cup_i H_i^3$. Thus $G_i = C_i^* \setminus (H_i^1 \cup H_i^1 \cup H_i^1)$ and $G = S \setminus (H^1 \cup H^2 \cup H^3)$. By Markov inequality we have

$$|H^3| \leq 120(\epsilon/\alpha)n.$$

We say that the points of G are "good" and the points of $H := H^1 \cup H^2 \cup H^3 = S \setminus G$ are "bad". As we have seen so far there are not too many bad points: $|H| = O(\frac{\epsilon}{\alpha} n)$ – a fact that we will use later.

Let

$$B_i = \bigcup_{x \in G_i} B_x \setminus C_i^*.$$

Clearly, for all $x \in G_i$ we have

$$\mathrm{B}_x \subseteq C_i^* \cup B_i. \tag{1}$$

From Lemma 9 we know that if $x \in G_i$ and $y \in G_j$ for $i \neq j$ then $\mathrm{B}_x \cap \mathrm{B}_y = \emptyset$. This implies that $B_i \cap B_j = \emptyset$ for $i \neq j$ as well as that if $x \in G_i$ then B_x intersects only G_i no other G_j. Let

$$\tilde{\mathrm{B}}_x = \{y : x \in \mathrm{B}_y, y \in \mathrm{B}_x\}.$$

We now show that for all points x, $\tilde{\mathrm{B}}_x$ intersects at most one set G_i and no other G_j for $j \neq i$. For $x \in G_i$, since $\tilde{\mathrm{B}}_x \subset \mathrm{B}_x$, we get the desired claim. For $z \in S \setminus \cup_i G_i$ we might have B_z intersect two different G_i and G_j. However from Lemma 9 we have that for any two $x \in G_i$ and $y \in G_j$ there is no z such that $z \in \mathrm{B}_x$ and $z \in \mathrm{B}_y$. This implies that there is no z such that we have both $x \in \tilde{\mathrm{B}}_z$ and $y \in \tilde{\mathrm{B}}_z$, so for $z \in S \setminus \cup_i G_i$, $\tilde{\mathrm{B}}_z$can intersect only one G_i.

From above we also have:

$$|\cup_i B_i| \leq |H^1| + |H^2| + |H^3| = O\left(\frac{\epsilon}{\alpha} n\right).$$

We now claim that for any i there exists an x such that

$$|\tilde{\mathrm{B}}_x \cap G_i| \geq |G_i| - 2(|B_i| + |C_i^* \setminus G_i|). \tag{2}$$

We first prove that for all $x \in G_i$ we have $|\mathrm{B}_x| \geq |G_i|$. If $\tau_x > \frac{\alpha w}{30\epsilon |C_i^*|}$, then $\mathrm{B}_x \supseteq G_i$. Else, if $\tau_x < \frac{\alpha w}{30\epsilon |C_i^*|}$ then

$$|\mathrm{B}_x| \geq \frac{1}{20}\frac{\alpha w}{\epsilon}\frac{30|C_i^*|\epsilon}{\alpha w} = 1.5|C_i^*| > |G_i|.$$

So for every $x \in G_i$, we have by (1),

$$|\mathrm{B}_x \cap G_i| \geq |G_i| - |B_i| - |C_i^* \setminus G_i|.$$

This implies that there exists an x^* such that

$$|\{x \in G_i : x^* \in \mathrm{B}_x\}| \geq |G_i| - |B_i| - |C_i^* \setminus G_i|.$$

So,

$$|\{x \in G_i : x^* \in \mathrm{B}_x\} \cap \mathrm{B}_{x^*}| \geq |G_i| - 2|B_i| - 2|C_i^* \setminus G_i|.$$

Since

$$\{x \in G_i : x^* \in \mathrm{B}_x\} \cap B_{x^*} \subseteq \tilde{\mathrm{B}}_x^*,$$

we get relation (2), as desired.

To finish the argument we need to argue that greedy covering on $\tilde{\mathrm{B}}_x$ works well. Let us think of each cluster G_i as initially "unmarked", and then marking it the first time we ever choose a group that intersects it. We now consider a few cases. If the jth C_j^o intersects an *unmarked* G_i, we will assign $\sigma(j) = i$. Note that if this group misses α_i points from G_i, then since we were greedy, according to relation (2), we must have picked at least $\alpha_i - 2(|B_i| + |C_i^* \setminus G_i|)$ elements from H in this group. Overall, we must have

$$\sum_i (\alpha_i - 2(|B_i| + |C_i^* \setminus G_i|)) \leq |H|,$$

which together with

$$\sum_i |B_i| \leq |H| \text{ and } \sum_i |C_i^* \setminus G_i| \leq |H|$$

implies $\sum_i \alpha_i \leq 5|H|$. Thus total error incurred in this way w.r.t. the good set G is given by the number of points missed from G_i, so it is at most $\sum_i \alpha_i \leq 5|H|$. The other case is when the jth group C_j^o intersects a *marked* G_i. In this case we assign $\sigma(j)$ to any arbitrary cluster $C_i^{*\prime}$ not marked by the end of the process. The error incurred from these cases is at most $|H| + \sum_i \alpha_i \leq 6|H|$, since this is an upper bound on the number of points left that aren't in unmarked clusters. Finally, we need to also consider the error with respect to the bad set H. Adding all these up, we obtain that the total error is bounded by $5|H| + 6|H| + |H| = 12|H| = O(\epsilon/\alpha n)$. ■

In the case of unknown w, we show the following:

Theorem 11 *If $k \leq \log n / \log\log n$ and if the min-sum instance (S, d) satisfies the $(1+\alpha, \epsilon)$-property even if we are not given w, we can use Algorithm 1 as a subroutine to produce a clustering that is $O(\epsilon/\alpha)$-close to the target. For the case of general k, we can use Algorithm 1 as a subroutine to produce a list of $\log\log n$ clusterings such that one of them is $O(\epsilon/\alpha)$-close to the target.*

Proof: It is not difficult to verify that the argument in Theorem 10 holds (with only a constant factor loss in the final guarantee on the error rate), even if we use a constant factor approximation for w instead of using the exact value of w in Algorithm 1. If $k \leq \log n / \log\log n$, then we can use the results in [15] for finding a constant factor approximation for w, and thus we are able produce a clustering that is $O(\epsilon/\alpha)$-close to the target.

For the case of general k, we use the fact that there exists an $O(\delta^{-1}\log^{1+\delta} n)$-approximation algorithm in time $n^{O(1/\delta)}$ for the case of arbitrary k [9]. The main idea is to use the algorithm in [9] with $\delta = 1$ to find a lower bound l an upper bound L for w that are within a multiplicative $O(\log^2 n)$ factor of each other. We then try all the values of $l, 2\cdot l, \ldots, 2^i \cdot l, \ldots$ and run Algorithm 1 for each of them. One of the values $2^i \cdot l$ will be a 2-approximation for w and an argument similar to the one in Theorem 10 shows that in that case we get a clustering which is $O(\epsilon/\alpha)$-close to the target. ■

Note: All our arguments above can be extended (with an appropriate loss in the final accuracy guarantees) to the case where the given dissimilarity function d satisfies only the following $d(x, y) \leq \gamma(d(x, z) + d(z, y))$ for some $\gamma > 1$.

Theorem 12 *If the min-sum instance (S, d) satisfies the $(1+\alpha, \epsilon)$-property , then so long as then so long as the smallest correct cluster has size greater than $100\epsilon n/\alpha^2$ we can efficiently find a clustering that is $O(\epsilon)$-close to the target.*

Proof Sketch: Assume that we are given the value of w. We first use the construction in Theorem 10 to produce a clustering $C_1', \ldots, C_k'$ with the property that the target clustering is $O(\epsilon/\alpha)$-close to the target. For each cluster C_i' we compute the center $\tilde{c}_i$ be a point in C_i' that minimizes $\sum_{x \in C_i'} d(x, \tilde{c}_i)$. We then define the set

$$C_s^i = \left\{ x \in C_i, d(x, \tilde{c}_i) < \frac{\alpha w}{60\epsilon |C_i'|} \right\}.$$

The fact that the clusters are $\gg \epsilon n/\alpha^2$ means that each C_s^i captures at least a $(1 - O(\alpha))$-fraction of the corresponding C_i^*.

We now construct a new clustering $C_1'', \ldots, C_k''$ as follows: for each point x and each cluster C_j' , we compute the weight $w_s(x, j)$ as $\sum_{y \in C_s^i} d(x, y)$. We finally insert x into the cluster C_i'' with $i = \mathrm{argmin}_j w_s(x, j)$. The main steps in the correctness proof are the following. We first show that (up to re-indexing of the clusters) $C_s^i \subset G_i$ and that $\left|\cup_i C_s^i\right| = n - O((\epsilon/\alpha)n)$. We then use these facts together fact that each C_s^i is a $(1 - O(\alpha))$-approximation to C_i^* in order to show that all but $O(\epsilon n)$ points will make the right choice.

In the case where we do not know w, we use the technique in [7] of trying increasing values of w: we then stop the first time when we output k clusters that cover at least $n - O((\epsilon/\alpha)n$ of the points in S. ■

3.3 Inductive Setting

In this section we consider an *inductive* model in which the set S is merely a small random subset of points of size n from a much larger abstract instance space X, $|X| = N$, $N \gg n$ and the clustering we output clustering is represented *implicitly* through a hypothesis $h : X \to Y$. In the case where $k \leq \log n / \log\log n$ we produce a clustering of error at most $O(\epsilon/\alpha)$. In the case where $k > \log n / \log\log n$ we produce a list of hypotheses, $\{h_1, \ldots, h_{??}\}$ such that at least one of them has error at most ϵ/α.

We can adapt the algorithm in Theorem 10 to the inductive setting as shown in Algorithm 2. The main idea is to show that our algorithm from the transductive setting is

pretty robust, and it can survive eliminating small clusters, making $\tilde{\mathrm{B}}$ and the set size estimates fuzzy. Specifically, in the case of known w we can show the following:

Theorem 13 *Assume that the min-sum instance (X, d) satisfies the $(1+\alpha, \epsilon)$-property and that we are given the value of w. If we draw a sample S of size $n = O\left(\frac{k^2}{\epsilon^2}\ln\left(\frac{kN}{\delta}\right)\right)$ then we can use Algorithm 2 to produce a clustering which is $O(\epsilon/\alpha)$-close to the target with probability $> 1-\delta$.*

Moreover, inserting a new element only takes $O(k)$ time.

Proof Sketch: The proof works in two phases. In the first stage we redo the analysis of Theorem 10 to show that Algorithm 2 works as well as Algorithm 1 (up to a loss of multiplicative constants) in producing the approximate clustering. The difference is that Algorithm 2 is "fuzzier" than Algorithm 1 in several respects – the comparisons need not be exact, and set-size estimates are only needed within a constant precision.

In the second phase we observe that Algorithm 2 can be executed in the inductive setting with high probability. In particular, set sizes can be estimated within the required precision from few samples, and for each sufficiently large cluster there is a suitable center x_i in the cluster such that $|\tilde{\mathrm{B}}^S_{x_i}| > (1-\gamma)\cdot|\tilde{\mathrm{B}}^{max}_x|$. This implies that the result of the execution of Algorithm 2 on the sample is actually the projection of a valid execution of the algorithm on the entire input to the sample. Thus by the correctness of the algorithm in the transductive setting we obtain its correctness in the inductive mode.

Finally, the correctness of the testing phase follows from the structural properties of the clustering we proved in Theorem 10. ■

Theorem 13 also works if we are given a constant factor approximation rather than an exact value for w.

We now state our main result for the case of unknown w. In the following, we denote by D the diameter of the metric space, i.e, $D = \max_{x,y} d(x,y)$. Using results from [14, 15] on estimating the value of the optimal min-sum based on the sample, we obtain the following theorem.

Theorem 14 *Assume that the minsum instance (X, d) satisfies the $(1+\alpha, \epsilon)$-property and that we are not given thew value of w. If we draw a sample S of size satisfying both $n = O\left(\frac{k^2}{\epsilon^2}\ln\left(\frac{kN}{\delta}\right)\right)$ and $n = \tilde{O}\left(D(k+\ln(1/\delta))(\log n + Dk^2)\right)$, and if $k \le \log n/\log\log n$ then we can use Algorithm 2 as a subroutine to produce a clustering that is $O(\epsilon/\alpha)$-close to the target. For the case of general k, we can use Algorithm 2 as a subroutine to produce a list of $\log\log n$ clusterings such that one of them is $O(\epsilon/\alpha)$-close to the target.*

4 The Correlation Clustering Problem

The correlation clustering setup introduced in [11] is as follows. We are given a fully-connected graph G with edges labeled $+1$ (similar) or -1 (different), and the goal is to find a partition of the vertices into clusters that agrees as much as possible with the edge labels.[4]

[4]Note that the problem is not trivial since we might have inconsistences. In particular, it is possible to have x, y, z such that the edge $(x,y)$is labeled $+1$, the edge (y,z) is labeled $+1$ and the (x,z) is labeled -1.

Algorithm 2 Fuzzy Min-Sum Algorithm

Input: (S,d), w, $\epsilon \le 1$, $\alpha > 0$, k, n, N.

Training phase

Set $w' = wn/N$, $I = \emptyset$, $N = \emptyset$, $\gamma = \epsilon/k$.

For all x do:

- Let the initial threshold $\tau = \tau_0$.
- Construct the ball $\mathrm{B}^S(x,\tau)$ by including all points within distance τ of x.
- **If** $\frac{1}{17}\frac{\alpha w'}{\epsilon\tau} \ge |\mathrm{B}^S(x,\tau)| \ge \frac{1}{18}\frac{\alpha w'}{\epsilon\tau}$ and $|\mathrm{B}^S(x,\tau)| \ge \frac{\epsilon n}{2k}$
  then let $\tau^S_x = \tau$ and $\mathrm{B}^S_x = \mathrm{B}^S(x,\tau^S_x)$; add x to I
  else increase τ to the next critical threshold

For all x, let $\left\{y : x \in \mathrm{B}^S_y, |\mathrm{B}^S_y| \ge \frac{\epsilon n}{k}\right\} \subseteq \tilde{\mathrm{B}}^S_x \subseteq \left\{y : x \in \mathrm{B}^S_y, |\mathrm{B}^S_y| \ge \frac{\epsilon n}{8k}\right\}$. Set $\mathcal{L} = \emptyset$.

For $i = 1 \ldots k$ do

- Let C^o_i be a cluster $\tilde{\mathrm{B}}^S_{x_i}$ of size of at least $(1-\gamma)$ of the largest $|\tilde{\mathrm{B}}^S_x|$ with $x_i \in I$.
- Add C^o_i to $\mathcal{L}$.
- For $x' \ne x$, set $\tilde{\mathrm{B}}^S_{x'} = \tilde{\mathrm{B}}^S_{x'} \setminus C^o_i$.

Testing Phase: . When a new point z arrives, assign it to the cluster C^o_i which minimizes $d(z, x_i)$.

In particular, the Min-Disagreement correlation clustering objective (Min-Disagreement CC) asks to find a clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ the minimizes the number of disagreements: the number of -1 edges inside clusters plus the number of $+1$ edges between clusters. In this clustering formulation one does not need to specify the number of clusters k as a separate parameter, as in measures such as k-median or min-sum clustering. Instead, in correlation clustering, the optimal number of clusters can take any value between 1 and n, depending on the edge labels. The currently best approximation algorithm for minimizing disagreements is a 2.5 approximation [3] and the problem is known to be APX-hard [13].

We can show that the (c,ϵ) assumption does not make optimizing Min-Disagreement CC objective easier.

Theorem 15 *For the Min-Disagreement CC objective the problem of finding a c-approximation can be reduced to the problem of finding a c-approximation under the (c,ϵ) assumption. Therefore, the problem of finding a c-approximation under the (c,ϵ) assumption is as hard as the problem of finding a c-approximation in general.*

We show now that if our input satisfies the $(1+\alpha,\epsilon)$-property for the Min-Disagreement CC objective, then the data satisfies the $(2.5, (49/\alpha+1)\epsilon)$ property as well. Specifically:

Theorem 16 *For the Min-Disagreement CC objective, if the instance (S,d) satisfies the $(1+\alpha,\epsilon)$-property with respect to the target clustering $\mathcal{C}_T$, then the instance (S,d) also satisfies the $(2.5,(49/\alpha+1)\epsilon)$ property with respect to the target clustering $\mathcal{C}_T$.*

Interpretation: This means that under the $(1+\alpha,\epsilon)$-property we can use a state of the art 2.5 approximation algorithm for minimizing disagreements in order to get a $(49/\alpha+1)\epsilon$ accurate clustering.

Proof: We prove the contrapositive. We show that if the instance that does not satisfy the $(2.5,(49/\alpha+1)\epsilon)$ property with respect to the target clustering, then the instance does not satisfy the $(1+\alpha,\epsilon)$ property with respect to the target clustering. Recall that $\mathcal{C}^*$ is the optimal Min-Disagreement CC clustering.

Assume that the instance (S,d) that does not satisfy the $(2.5,(49/\alpha+1)\epsilon)$ with respect to the target clustering. This means that there exists a clustering $\mathcal{C}'=\{C'_1,C'_2,\ldots,C'_{k'}\}$ such that $cost(\mathcal{C}')\leq 2.5\mathrm{OPT}$ and $dist(\mathcal{C}_T,\mathcal{C}')\geq(49/\alpha+1)\epsilon$; since $dist(\mathcal{C}^*,\mathcal{C}_T)\leq\epsilon$ we have $dist(\mathcal{C}^*,\mathcal{C}')\geq\frac{49\epsilon}{\alpha}$. For $x\in S$ we denote by $\mathcal{C}^*(x)$ its cluster in $\mathcal{C}^*$ and by $\mathcal{C}'(x)$ its cluster in $\mathcal{C}'$. We will call a point uninteresting if it does not change too many neighbors between the two clusterings $\mathcal{C}'$ and $\mathcal{C}^*$; formally, x is uninteresting if

$$|\mathcal{C}^*(x)\triangle\mathcal{C}'(x)|<|\mathcal{C}^*(x)\cap\mathcal{C}'(x)|.$$

We show in the following that there are at least $49\frac{\epsilon n}{\alpha}$ interesting points. In order to do this we exhibit a partial matching of the clusters in $\mathcal{C}^*$ and $\mathcal{C}'$; specifically, we connect two clusters C^*_i and C'_j if $|C^*_i\triangle C'_j|<|C^*_i\cap C'_j|$ and we let $\pi(i)=j$. We prove now that this is a partial matching of the clusters in $\mathcal{C}^*$ and $\mathcal{C}'$. Assume by contradiction that this is not the case; i.e. assume that there exist i, j, k such that $C^*_i\cap C^*_k=\emptyset$ and $|C^*_i\triangle C'_j|<|C^*_i\cap C'_j|$ and $|C^*_k\triangle C'_j|<|C^*_k\cap C'_j|$, which implies

$$|C^*_i\triangle C'_j|+|C^*_k\triangle C'_j|<|C^*_i\cap C'_j|+|C^*_k\cap C'_j|. \quad (3)$$

However since $C^*_i\cap C^*_k=\emptyset$ we have both

$$|C^*_i\cap C'_j|+|C^*_k\cap C'_j|\leq|C'_j|$$

and

$$|C^*_i\triangle C'_j|+|C^*_k\triangle C'_j|\geq|C'_j|,$$

which implies

$$|C^*_i\cap C'_j|+|C^*_k\cap C'_j|\leq|C^*_i\triangle C'_j|+|C^*_k\triangle C'_j|,$$

thus contradicting (3). This proves that π is a partial matching of the clusters in $\mathcal{C}^*$ and $\mathcal{C}'$. Let σ be an arbitrary permutation of the whole set S that matches all uninteresting points according to π; i.e., σ is defined as $\sigma(i)=j$ such that if x is uninteresting and $C(x)=C^*_i$, then $C'(x)=C'_{\pi(i)}=C'_j$. By definition we have

$$dist_\sigma(\mathcal{C}^*,\mathcal{C}')\geq dist(\mathcal{C}^*,\mathcal{C}')\geq 49\frac{\epsilon n}{\alpha},$$

which implies that there exists a set I with at least $49\epsilon n/\alpha$ interesting points.

We now compute the cost of isolating an interesting point x. Let us denote by $w(x)$ the contribution of x to the Min-Disagreement CC in $\mathcal{C}^*$ and by $w'(x)$ the contribution of x to the Min-Disagreement CC in $\mathcal{C}'$. We clearly have

$$w(x)+w'(x)\geq|\mathcal{C}^*(x)\triangle\mathcal{C}'(x)|\geq|\mathcal{C}^*(x)\cap\mathcal{C}'(x)|,$$

which implies:

$$2(w(x)+w'(x))\geq\max(|\mathcal{C}^*(x)|,|\mathcal{C}'(x)|).$$

So, for an interesting point x we get that

$$|\{y:R(x,y)=+\}|\leq|\mathcal{C}^*(x)|+w(x)\leq 3(w(x)+w'(x)).$$

So, the cost of isolating an interesting point x is at most $3(w(x)+w'(x))$.

Since $cost(\mathcal{C}')\leq 2.5\mathrm{OPT}$, $cost(\mathcal{C}^*)=\mathrm{OPT}$ and $|I|\geq 49\frac{\epsilon n}{\alpha}$ we have:

$$\frac{1}{|I|}\sum_{x\in I}(w(x)+w'(x))\leq\frac{3.5\mathrm{OPT}}{|I|}\leq\frac{3.5\mathrm{OPT}\alpha}{49\epsilon n}.$$

This implies that for any set size $s\leq|I|$ there exist a set $A\subseteq I$ of size s such that

$$\frac{1}{|A|}\sum_{x\in A}(w(x)+w'(x))\leq\frac{3.5\mathrm{OPT}\alpha}{49\epsilon n}.$$

Note also that for any interesting point x we have

$$w(x)+w'(x)\geq 1,$$

therefore

$$49\epsilon n/\alpha\leq|I|\leq\sum_{x\in I}(w(x)+w'(x))\leq 3.5\mathrm{OPT},$$

which implies

$$\mathrm{OPT}\geq 14\epsilon n/\alpha \quad (4)$$

Let $A\subseteq I$ of size $s=4\epsilon n$ such that

$$\frac{1}{|A|}\sum_{x\in A}(w(x)+w'(x))\leq\frac{3.5\mathrm{OPT}\alpha}{49\epsilon n}.$$

Let $A_s\subseteq A$ be the set of singleton points in the target clustering, i.e, $x\in A_s$ if $\mathcal{C}(x)=\{x\}$, and let $A_{ns}=A\setminus A_s$. We produce a new clustering $\mathcal{C}''$ from $\mathcal{C}^*$ by isolating the points in A_{ns} and by pairing up the points in A_s and merging any two points in the same pair. By Fact 1 have get $dist(\mathcal{C}^*,\mathcal{C}'')\geq 2\epsilon$, so $dist(\mathcal{C}_T,\mathcal{C}'')\geq\epsilon$. Also as shown above, the cost of isolating all the points in A_{ns} is at most $(10.5\alpha\mathrm{OPT}/(49\epsilon n))|A|\leq\alpha(42/49)\cdot\mathrm{OPT}$; also the total cost of merging the singleton interesting pairs is at most $|A_s|/2\leq 2\epsilon n$ which by (4) is at most $\alpha(7/49)\cdot\mathrm{OPT}$. This implies that the cost of isolating all the points in A_{ns} plus the cost of merging the singleton interesting pairs is at most $\alpha\mathrm{OPT}$. So the Min-Disagreement CC cost of $\mathcal{C}''$ is within a $1+\alpha$ factor of OPT, and yet $\mathcal{C}''$ which is ϵ-far from the target. Thus our clustering instance does not satisfy the $(1+\alpha,\epsilon)$ property with respect to the target clustering, which is a contradiction. This completes the proof. ■

Note: The other correlation clustering objective is of maximizing agreements, the number of $+1$ edges inside clusters plus the number of -1 edges between clusters. For maximizing agreements there exists a PTAS [11], so this objective is not interesting in our framework.

4.1 The Non Complete Graph Case

In the case where the graph G is not fully-connected, we do not get the strong result as in Theorem 16. On the contrary, we can show the following:

Theorem 17 *For any $\alpha, \beta < 1/6$, there exists a family of graphs G and target clusterings that satisfy the $(1+\alpha, 0)$ property for the the Min-Disagreement CC objective and yet do not satisfy even the $(1+\alpha+\beta, 1/2)$ property for that objective.*

Proof: Consider a set of n points such that the target clustering consists of one cluster C_1 with $n/2$ points and one cluster C_2 with $n/2$ points. we set both C_1 and C_2 to be fully connected with all the edges inside C_1 and C_2 labeled $+$. Also we designate a single vertex in C_1 which is connected with $n/3$ vertices in C_2 with edges all labeled as $+$, and a vertex in C_2 connected with $(n/3)\cdot(1+\alpha+\beta)$ edges in C_1, all labeled as $-$. It's easy to verify that the instance satisfies the $(1+\alpha, 0)$ property; we have $\mathrm{OPT} = \frac{n}{3}$ and any other solution has cost greater than $n/3(1+\alpha+\beta)$. However the solution does not even satisfy the $(1+\alpha+\beta, 1/2)$ property. The clustering with all the points in one big cluster has cost $1+\alpha+\beta$ and yet, it's distance from the target is $1/2$. ■

5 Conclusions and Open Questions

In this work we get around inherent inapproximability results for the min-sum objective in the case where good approximation to the min-sum objective indeed implies an accurate clustering. We derive strong structural properties from this assumption, and use them to give an efficient algorithm that produce accurate clusterings.

In the minimizing disagreements setting for correlation clustering we show that the same assumption allows us to find an accurate clustering using existing approximation algorithms. One concrete open question remaining is dealing with a non-complete graph in the context correlation clustering for the minimizing disagreements objective.

More generally, it would be interesting to further explore and analyze in this framework other natural classes of commonly used clustering objective functions. It would also be interesting to consider an agnostic version of model, where the (c, ϵ) property is satisfied only after some small number of outliers or ill-behaved data points have been removed.

Acknowledgments: We thank Avrim Blum for numerous useful discussions.

References

[1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Prooceedings of the Eighteenth Annual Conference on Learning Theory*, 2005.

[2] M. Ackerman and S. Ben-David. Which data sets are clusterable? - a theoretical study of clusterability. In *NIPS*, 2008.

[3] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In *Proceedings of the 37th ACM Symposium on Theory of Computing*, 2005.

[4] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. In *Proceedings of STOC*, 2000.

[5] S. Arora and R. Kannan. Learning mixtures of arbitrary gaussians. In *STOC*, 2005.

[6] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3), 2004.

[7] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2009.

[8] M.-F. Balcan, A. Blum, and S. Vempala. A discrimantive framework for clustering via similarity functions. In *STOC*, 2008.

[9] Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k-clustering in metric spaces. In *Proceedings on 33rd Annual ACM Symposium on Theory of Computing*, 2001.

[10] S. Ben-David. A framework for statistical clustering with constant time approximation for k-median and k-means clustering. *Machine Learning*, 66(2-3), 2007.

[11] A. Blum, N. Bansal, and S. Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

[12] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoy. A constant-factor approximation algorithm for the k-median problem. In *STOC*, 1999.

[13] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 524–533, 2003.

[14] A. Czumaj and C. Sohler. Sublinear-time approximation for clustering via random samples. In *Proceedings of the 31st International Colloquium on Automata, Languages and Programming (ICALP)*, 2004.

[15] A. Czumaj and C. Sohler. Small space representations for metric min-sum k-clustering and their applications. In *Proceedings of the 24th International Symposium on Theoretical Aspects of Computer Science*, 2007.

[16] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, 1999.

[17] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *In Proceedings for the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, 2003.

[18] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[19] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.

[20] P. Indyk. Sublinear time algorithms for metric space problems. In *STOC*, 1999.

[21] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *34th STOC*, 2002.

[22] T. Joachims and J. Hopcroft. Error bounds for correlation clustering. In *Proceedings of the International Conference on Machine Learning*, 2005.

[23] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *18th COLT*, 2005.

[24] J. Kleinberg. An impossibility theorem for clustering. In *NIPS*, 2002.

[25] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th FOCS*, 2004.

[26] M. Meila. Comparing clusterings by the variation of information. In *COLT*, 2003.

[27] M. Meila. Comparing clusterings – an axiomatic view. In *International Conference on Machine Learning*, 2005.

[28] L.J. Schulman. Clustering for edge-cost minimization. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing,*, 2000.

[29] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *JCSS*, 68(2):841–860, 2004.