Clustering under Perturbation Resilience

Maria Florina Balcan and Yingyu Liang

School of Computer Science, Georgia Institute of Technology
ninamf@cc.gatech.edu, yliang39@gatech.edu

Abstract. Motivated by the fact that distances between data points in many real-world clustering instances are often based on heuristic measures, Bilu and Linial [6] proposed analyzing objective based clustering problems under the assumption that the optimum clustering to the objective is preserved under small multiplicative perturbations to distances between points. In this paper, we provide several results within this framework. For separable center-based objectives, we present an algorithm that can optimally cluster instances resilient to $(1 + \sqrt{2})$-factor perturbations, solving an open problem of Awasthi et al. [2]. For the k-median objective, we additionally give algorithms for a weaker, relaxed, and more realistic assumption in which we allow the optimal solution to change in a small fraction of the points after perturbation. We also provide positive results for min-sum clustering which is a generally much harder objective than k-median (and also non-center-based). Our algorithms are based on new linkage criteria that may be of independent interest.

Keywords: clustering, perturbation resilience, k-median, min-sum.

1 Introduction

Problems of clustering data from pairwise distance 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-median or min-sum. For example, in the k-median clustering problem the goal is to partition the data into k clusters C_i, giving each a center c_i, in order to minimize the sum of the distances of all data points to the centers of their cluster. In the min-sum clustering approach the goal is to find k clusters C_i that minimize the sum of all intra-cluster pairwise distances. Yet unfortunately, for most natural clustering objectives, finding the optimal solution to the objective function is NP-hard. As a consequence, there has been substantial work on approximation algorithms [1,5,7,8,9] with both upper and lower bounds on the approximability of these objective functions on worst case instances.

Recently, Bilu and Linial [6] suggested an exciting, alternative approach aimed at understanding the complexity of clustering instances which arise in practice. Motivated by the fact that distances between data points in clustering instances are often based on a heuristic measure, they argue that interesting instances should be resilient to small perturbations in these distances. In particular, if small perturbations can cause the optimal clustering for a given objective to change drastically, then that probably is not

A. Czumaj et al. (Eds.): ICALP 2012, Part I, LNCS 7391, pp. 63–74, 2012.

a meaningful objective to be optimizing. They specifically define an instance to be α-perturbation resilient for an objective Φ if perturbing pairwise distances by multiplicative factors in the range $[1, \alpha]$ does not change the optimum clustering under Φ. They consider in detail the case of max-cut clustering and give an efficient algorithm to recover the optimum when the instance is resilient to perturbations on the order of $O(\sqrt{n\Delta})$ where n is the number of points and Δ is the maximum degree of the graph. They also give an efficient algorithm for instance of unweighted max-cut that is resilient to perturbations on the order of $O(\frac{n}{\delta})$ where δ is the minimum degree of the graph.

Two important questions raised by the work of Bilu and Linial [6] are: (1) the degree of resilience needed for their algorithm to succeed is quite high: can one develop algorithms for important clustering objectives that require much less resilience? (2) the resilience definition requires the optimum solution to remain *exactly* the same after perturbation: can one succeed under weaker conditions? In the context of *separable center-based* objectives such as k-median and k-center, Awasthi et al. [2] partially address the first question and show that an algorithm based on the single-linkage heuristic can efficiently find the optimal clustering for α-perturbation-resilient instances for $\alpha = 3$. They also conjecture it to be NP-hard to beat 3 and prove beating 3 is NP-hard for a related notion.

In this work, we address both questions raised by Bilu and Linial [6] and additionally improve over Awasthi et al. [2]. First, for separable center-based objectives we design a polynomial time algorithm for finding the optimum for instances resilient to perturbations of value $\alpha = 1+\sqrt{2}$, thus beating the previously best known factor of 3 of Awasthi et al. [2]. Second, for k-median, we consider a weaker, relaxed, and more realistic notion of perturbation-resilience where we allow the optimal clustering of the perturbed instance to differ from the optimal of the original in a small ϵ fraction of the points. This is arguably a more natural though also more difficult condition to deal with. We give positive results for this case as well, showing for somewhat larger values of α that we can still achieve a near-optimal clustering. We additionally give positive results for min-sum clustering which is a generally much harder objective than k-median (and also non-center-based). For example, the best known guarantee for min-sum clustering on worst-case instances is an $O(\delta^{-1} \log^{1+\delta} n)$-approximation in time $n^{O(1/\delta)}$ due to [5]; by contrast, the best guarantee known for k-median is factor $3 + \epsilon$ due to [1].

Our results are achieved by carefully deriving structural properties of perturbation-resilience. At a high level, all the algorithms we introduce work by first running appropriate linkage procedures to produce a tree, and then running dynamic programming to retrieve the best k-clustering in the tree. To ensure that (under perturbation resilience) the tree output in the first step has a low-cost pruning, we derive new linkage procedures (closure linkage and approximate closure linkage) which are of independent interest.

Our Results: We provide several results for clustering perturbation-resilient instances in the metric space for separable center-based objectives and for the min-sum objective.

In Section 3 we improve on the bounds of Awasthi et al. [2] for α-perturbation resilient instances for separable center-based objectives, giving an algorithm that efficiently [1] finds the optimum for $\alpha = 1 + \sqrt{2}$. Commonly used separable center-based

[1] For clarity, efficient means polynomial in n (number of points) and k (number of clusters).

objectives, such as k-median, are NP-hard to even approximate, yet we can recover the exact solution for perturbation resilient instances. Our algorithm is based on a new linkage procedure using a new notion of distance (closure distance) between sets that may be of independent interest.

In Section 4 we consider the more challenging and more general notion of (α, ϵ)-perturbation resilience for k-median, where we allow the optimal solution after perturbation to be ϵ-close to the original. We provide an efficient algorithm which for $\alpha > 2 + \sqrt{7}$ produces $(1 + O(\epsilon/\rho))$-approximation to the optimum, where ρ is the fraction of the points in the smallest cluster. The key property we derive and exploit is that, except for ϵn bad points, most points are α closer to their own center than to any other center. Using this, we then design an approximate version of the closure linkage criterion that allows us to carefully eliminate the noise introduced by the bad points and construct a tree with a low-cost pruning that is a good approximation to the optimum.

In Section 5 we provide the first efficient algorithm for optimally clustering α-min-sum perturbation resilient instances. Our algorithm is based on an appropriate modification of average linkage that exploits the structure of such instances.

Due to the lack of space we only provide sketches for most proofs in this paper. Full proofs appear in the long version of the paper [4]. In the long version, we also provide sublinear-time algorithms, showing algorithms that can return an implicit clustering from only access to a small random sample.

2 Notation and Preliminaries

In a clustering instance, we are given a set S of n points in a finite metric space, and we denote $d : S \times S \to \mathbb{R}_{\geq 0}$ as the distance function. Φ denotes the objective function over a partition of S into $k < n$ clusters which we want to optimize, i.e. Φ assigns a score to every clustering. The optimal clustering w.r.t. Φ is denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, and its cost is denoted as $\mathcal{OPT}$. The core concept we study in this paper is the perturbation resilience notion introduced by Bilu and Linial [6]. Formally:

Definition 1. *A clustering instance (S, d) is α-**perturbation resilient** to an objective Φ if for any $d' : S \times S \to \mathbb{R}$ s.t. $\forall p, q \in S, d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$, there is a unique optimal clustering $\mathcal{C}'$ for Φ under d' that equals the optimal clustering $\mathcal{C}$ under d.*

In this paper, we focus on center-based and min-sum objectives. For center-based objectives, we consider *separable center-based objectives* defined by Awasthi et al. [2].

Definition 2. *A clustering objective is **center-based** if the solution can be defined by partitioning S into k clusters $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$ and assigning a set of centers $\mathbf{p} = \{p_1, p_2, \ldots, p_k\} \subseteq S$ for the clusters. Such an objective is **separable** if it furthermore satisfies the following two conditions: 1) The objective function value of a given clustering is either a (weighted) sum or the maximum of the individual cluster scores; 2) Given a proposed single cluster, its score can be computed in polynomial time.*

For example, for the k-median objective which we study substantially, the objective is $\Phi(\mathcal{P}, \mathbf{p}) = \sum_{i=1}^{k} \sum_{p \in P_i} d(p, p_i)$. Other examples of center-based objectives include k-means for which $\Phi(\mathcal{P}, \mathbf{p}) = \sum_{i=1}^{k} \sum_{p \in P_i} d^2(p, p_i)$, and k-centers for which

$\Phi(\mathcal{P}, \mathbf{p}) = \max_{i=1}^{k} \max_{p \in P_i} d(p, p_i)$. The centers in the optimal solution are denoted as $\mathbf{c} = \{c_1, \ldots, c_k\}$. Clearly, in an optimal solution, each point is assigned to its nearest center. In such cases, the objective is denoted as $\Phi(\mathbf{c})$.

We also consider a different type of objective function: the *min-sum objective*. For this objective, S is partitioned into k clusters $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$, and the goal is to minimize $\Phi(\mathcal{P}) = \sum_{i=1}^{k} \sum_{p,q \in P_i} d(p, q)$.

In Section 4 we consider a generalization of Definition 1 where we allow a small difference between the original and the new optimum after perturbation. Formally:

Definition 3. *Let $\mathcal{C}$ be the optimal k-clustering and $\mathcal{C}'$ be another k-clustering of a set of n points. We say $\mathcal{C}'$ is ϵ-**close** to $\mathcal{C}$ if $\min_{\sigma \in \mathcal{S}_k} \sum_{i=1}^{k} |C_i \setminus C'_{\sigma(i)}| \leq \epsilon n$, where σ is a matching between indices of clusters of $\mathcal{C}'$ and those of $\mathcal{C}$.*

Definition 4. *A clustering instance (S, d) is (α, ϵ)-**perturbation resilient** to an objective Φ if for any $d' : S \times S \to \mathbb{R}$ s.t. $\forall p, q \in S, d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$, the optimal clustering $\mathcal{C}'$ for Φ under d' is ϵ-close to the optimal clustering $\mathcal{C}$ under d.*

For simplicity, we use shorthand $d(A, B) = \sum_{p \in A} \sum_{q \in B} d(p, q)$ and $d(p, B) = d(\{p\}, B)$. Also, we will sometimes assume that $\min_i |C_i|$ and ϵn is known. (Otherwise, we can simply search over the n possible different values for each parameter.)

3 α-Perturbation Resilience for Center-Based Objectives

In this section we show that, for $\alpha \geq 1 + \sqrt{2}$, if the clustering instance is α-perturbation resilient for separable center-based objectives, then we can efficiently find the optimal clustering. This improves on the $\alpha \geq 3$ bound of Awasthi et al. [2] and stands in sharp contrast to the NP-Hardness results on worst-case instances. Our algorithm succeeds for an even weaker property, the α-center proximity, introduced in Awasthi et al. [2].

Definition 5. *A clustering instance (S, d) satisfies the α-**center proximity** property if for any optimal cluster $C_i \in \mathcal{C}$ with center c_i, $C_j \in \mathcal{C} (j \neq i)$ with center c_j, any point $p \in C_i$ satisfies $\alpha d(p, c_i) < d(p, c_j)$.*

Lemma 1. *([2]) Any clustering instance that is α-perturbation resilient to separable center-based objectives also satisfies the α-center proximity.*

The proof follows by constructing a specific perturbation that blows up all the pairwise distances within C_i by a factor of α. By α-perturbation resilience, the optimal clustering remains the same, which then implies the desired result. In this section, we prove our results for α-center proximity. The results also hold for α-perturbation resilience since it implies α-center proximity. We begin with some key properties.

Lemma 2. *For any points $p \in C_i$ and $q \in C_j (j \neq i)$ in the optimal clustering of an α-center proximity instance, when $\alpha \geq 1 + \sqrt{2}$, we have:*
(1) $d(c_i, q) > d(c_i, p)$, (2) $d(p, c_i) < d(p, q)$.

Proof. (1) By Lemma 1, $d(q, c_i) > \alpha d(q, c_j)$. By triangle inequality, $d(c_i, c_j) \leq d(q, c_j) + d(q, c_i) < (1 + \frac{1}{\alpha})d(q, c_i)$. Also, $d(p, c_j) > \alpha d(p, c_i)$ and thus $d(c_i, c_j) \geq d(p, c_j) - d(p, c_i) > (\alpha - 1)d(p, c_i)$. The result follows by these inequalities.
(2) It also follows from triangle inequality. The proof appears in [2]. □

Lemma 2 implies that for any optimal cluster C_i, the ball of radius $\max_{p \in C_i} d(c_i, p)$ around the center c_i contains *only* points from C_i, and moreover, points inside the ball are each closer to the center than to any point outside the ball. Inspired by this structural property, we define the notion of closure distance between two sets as the radius of the minimum ball that covers the sets and has some margin from points outside the ball. We show that any (strict) subset of an optimal cluster has smaller closure distance to another subset in the same cluster than to any subset or union of other clusters. Using this, we will be able to define an appropriate linkage procedure that produces a tree on subsets that will all be laminar with respect to the optimal clusters. This will then allow us to extract from the tree the optimal solution using dynamic programming. We now define the notion of closure distance and then present our algorithm.

Definition 6. *Let $\mathbb{B}(p, r) = \{q : d(q, p) \leq r\}$. The **closure distance** $d_S(A, A')$ between two disjoint non-empty subsets A and A' of point set S is the minimum $d \geq 0$ such that there is a point $c \in A \cup A'$ satisfying the following requirements:*
(1) coverage: the ball $\mathbb{B}(c, d)$ covers A and A', i.e. $A \cup A' \subseteq \mathbb{B}(c, d)$;
(2) margin: points inside $\mathbb{B}(c, d)$ are closer to the center c than to points outside, i.e. $\forall p \in \mathbb{B}(c, d), q \notin \mathbb{B}(c, d)$, we have $d(c, p) < d(p, q)$.

Note that for any A, A', $d_S(A, A') = d_S(A', A) \leq \max_{p,q \in S} d(p, q)$, and it can be computed in polynomial time.

Algorithm 1. Separable center-based objectives, α perturbation resilience

Input: Data set S, distance function $d(\cdot, \cdot)$ on S.
Phase 1: Begin with n singleton clusters.
• Repeat till only one cluster remains: merge clusters C, C' which minimize $d_S(C, C')$.
• Let T be the tree with single points as leaves and internal nodes corresponding to the merges.

Phase 2: Apply dynamic programming on T to get the minimum cost pruning $\tilde{\mathcal{C}}$.
Output: Clustering $\tilde{\mathcal{C}}$.

Theorem 1. *For $(1 + \sqrt{2})$-center proximity instances, Algorithm 1 outputs the optimal clustering in polynomial time.*

The proof follows from the following key property of the Phase 1 of Algorithm 1.

Theorem 2. *For $(1 + \sqrt{2})$-center proximity instances, Phase 1 of Algorithm 1 constructs a binary tree such that the optimal clustering is a pruning of this tree.*

Proof. We prove correctness by induction. In particular, assume that our current clustering is *laminar* to the optimal clustering – that is, for each cluster A in our current

clustering and each C in the optimal clustering, we have either $A \subseteq C$, or $C \subseteq A$ or $A \cap C = \varnothing$. This is clearly true at the start. To prove that the merge steps preserve the laminarity, we need to show the following: if A is a strict subset of an optimal cluster C_i, A' is a subset of another optimal cluster or the union of one or more other clusters, then there exists B from $C_i \setminus A$ in the current clustering, such that $d_S(A, B) < d_S(A, A')$.

Let $d = \max_{p \in C_i} d(c_i, p), p^* = \arg\max_{p \in C_i} d(c_i, p)$. We first prove that there is a cluster $B \subseteq C_i \setminus A$ in the current clustering such that $d_S(A, B) \leq d$. There are two cases. First, if $c_i \notin A$, then define B to be the cluster in the current clustering that contains c_i. By induction, $B \subseteq C_i \setminus A$. Then we have $d_S(B, A) \leq d$ since there is $c_i \in B$, and (1) for any $p \in A \cup B$, $d(c_i, p) \leq d$, (2) for any $p \in S$ satisfying $d(c_i, p) \leq d$, and any $q \in S$ satisfying $d(c_i, q) > d$, by Lemma 2 we know $p \in C_i$ and $q \notin C_i$, and thus $d(c_i, p) < d(p, q)$. Second, if $c_i \in A$, we pick any $B \subseteq C_i \setminus A$ and a similar argument gives $d_S(A, B) \leq d$.

As a second step, we need to show that $d < \hat{d} = d_S(A, A')$. There are two cases: the center for $d_S(A, A')$ is in A or in A'. In the first case, there is a point $c \in A$ such that c and $\hat{d}$ satisfy the requirements of the closure distance. Pick a point $q \in A'$, and suppose C_j is the optimal cluster that contains q. As $d(c, q) \leq \hat{d}$, and by Lemma 2 $d(c_j, q) < d(c, q)$, we must have $d(c_j, c) \leq \hat{d}$ (otherwise it violates the second requirement of closure distance). Then we have $d = d(p^*, c_i) < d(p^*, c_j)/\alpha \leq (d + d(c_i, c) + d(c, c_j))/\alpha$ from Lemma 1 and triangle inequality. Since $d(c_i, c) < d(c, c_j)/\alpha$, we can combine the above inequalities and compare d and $d(c, c_j)$, and when $\alpha \geq 1 + \sqrt{2}$ we have $d < d(c, c_j) \leq \hat{d}$.

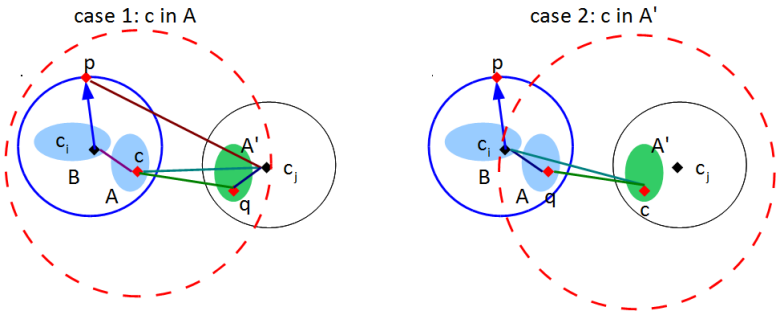

Fig. 1. Illustration for comparing d and $d_S(A, A')$ in Theorem 2

Now consider the second case, when there is a point $c \in A'$ such that c and $\hat{d}$ satisfy the requirements of the closure distance. Pick a point $q \in A$. We have $\hat{d} \geq d(c, q)$ from the first requirement, and $d(c, q) > d(c_i, q)$ by Lemma 2. Then from the second requirement $d(c_i, c) \leq \hat{d}$. So by Lemma 2, $d = d(c_i, p^*) < d(c_i, c) \leq \hat{d}$. □

Note: Our factor of $\alpha = 1 + \sqrt{2}$ beats the NP-hardness *lower bound* of $\alpha = 3$ of [2] for center proximity instances. The reason is that the lower bound requires the addition of Steiner points that can act as centers but are not part of the data to be clustered (though

the upper bound of [2] does not allow such Steiner points). One can also show a lower bound for center proximity instances without Steiner points. In particular one can show that for any $\epsilon > 0$, solving $(2-\epsilon)$-center proximity k-median instances is NP-hard [10].

4 (α, ϵ)-Perturbation Resilience for the k-Median Objective

In this section we consider a natural relaxation of the α-perturbation resilience, the (α, ϵ)-perturbation resilience, that requires the optimal clustering after perturbation to be ϵ-close to the original. We show that for (α, ϵ)-perturbation resilient instances, with $\alpha > 2 + \sqrt{7}$ and $\epsilon = O(\epsilon'\rho)$ where ρ is the fraction of the points in the smallest cluster, we can in polynomial time output a clustering that provides a $(1 + \epsilon')$-approximation to the optimum. Thus this improves over the best worst-case approximation guarantees known when $\epsilon' \leq 2$ and also beats the lower bound of $(1 + 2/e)$ on the best approximation achievable on worst case instances for metric k-median [9] when $\epsilon' \leq 1/e$.

The key idea is to understand and leverage the structure implied by (α, ϵ)-perturbation resilience. We show that perturbation resilience implies that there exists only a small fraction of points that are bad in the sense that their distance to their own center is not α times smaller than their distance to any other centers in the optimal solution. We then use this bounded number of bad points in our clustering algorithm.

4.1 Structure of (α, ϵ)-Perturbation Resilience

To understand (α, ϵ)-perturbation resilience, we need to consider the difference between the optimal clustering $\mathcal{C}$ under d and the optimal clustering $\mathcal{C}'$ under d', defined as $\min_{\sigma \in \mathcal{S}_k} \sum_{i=1}^{k} |C_i \backslash C'_{\sigma(i)}|$. Without loss of generality, we assume in this subsection that $\mathcal{C}'$ is indexed so that the argmin σ is the identity, and the difference is $\sum_{i=1}^{k} |C_i \setminus C'_i|$. We denote by c'_i the center of C'_i.

In the following we call a point *good* if it is α times closer to its own center than to any other center in the optimal clustering; otherwise we call it *bad*. Let B_i be the set of bad points in C_i. That is, $B_i = \{p : p \in C_i, \exists j \neq i, \alpha d(c_i, p) > d(c_j, p)\}$. Let $G_i = C_i \setminus B_i$ be the good points in cluster C_i. Let $B = \cup_i B_i$ and $G = \cup_i G_i$. We show that under perturbation resilience we do not have too many bad points. Formally:

Theorem 3. *Suppose the clustering instance is (α, ϵ)-perturbation resilient to k-median and $\min_i |C_i| > \frac{6\alpha}{\alpha - 1}\epsilon n$. Then $|B| \leq \epsilon n$.*

Here we describe a proof sketch of the theorem. In the full version we provide the detailed proof, and also point out that the bound in Theorem 3 is an optimal bound for the bad points in the sense that for any $\alpha > 1$ and $\epsilon < \frac{1}{5}$, we can construct an (α, ϵ)-perturbation resilient 2-median instance which has ϵn bad points.

Proof Sketch of [Theorem 3] The main idea is to construct a specific perturbation that forces certain selected bad points to move from their original optimal clusters. For technical reasons, we only perturb a selected subset of bad points, and show that they move out after perturbation. Then the (α, ϵ)-perturbation resilience leads to a bound on the number of selected bad points, which can also be proved to be a bound

on all the bad points. The selected bad points $\hat{B}_i$ in cluster C_i are defined by arbitrarily selecting $\min(\epsilon n + 1, |B_i|)$ points from B_i. Let $\hat{B} = \cup_i \hat{B}_i$. For $p \in \hat{B}_i$, let $c(p) = \arg\min_{c_j, j \neq i} d(p, c_j)$ denote its second nearest center; for $p \in C_i \setminus \hat{B}_i$, $c(p) = c_i$. The perturbation we consider blows up all distances by a factor of α except for those distances between p and $c(p)$. Formally, we define d' as $d'(p, q) = d(p, q)$ if $p = c(q)$ or $q = c(p)$, and $d'(p, q) = \alpha d(p, q)$ otherwise.

The key challenge in proving a bound on the selected bad points is to show that $c'_i = c_i$ for all i, i.e., the optimal centers do not change after the perturbation. Then in the optimum under d' each point p is assigned to the center $c(p)$, and therefore the selected bad points ($\hat{B}$) will move from their original optimal clusters. By (α, ϵ)-perturbation resilience property we get an upper bound on the number of selected bad points.

Suppose C'_i is obtained by adding point set A_i and removing point set M_i from C_i, i.e. $A_i = C'_i \backslash C_i$, $M_i = C_i \backslash C'_i$. At a high level, we prove that $c_i = c'_i$ for all i as follows. We first show that for each cluster, its new center is close to its old center, roughly speaking since the new and old clusters have a lot in common (Claim 1). We then show if $c'_i \neq c_i$ for some i, then the weighted sum of the distances $\sum_{1 \leq i \leq k} |C_i| d(c_i, c'_i)$ should be large (Claim 2). However, this contradicts Claim 1, so $c'_i = c_i$ for all i.

Claim 1. *For each i, $d(c_i, (C_i \cap C'_i) \setminus \hat{B}_i) \geq \frac{\alpha+2}{\alpha+1} \frac{|C_i|}{3} d(c_i, c'_i)$.*

Proof Sketch: The key idea is that under d', c'_i is the optimal center, so it has no more cost than c_i on C'_i. Since $\hat{B}_i \setminus M_i$ and A_i are small compared to $(C_i \cap C'_i) \setminus \hat{B}_i$, c'_i cannot save much on $\hat{B}_i \setminus M_i$ and A_i, thus it cannot have much more cost on $(C_i \cap C'_i) \setminus \hat{B}_i$ than c_i. Then c'_i is close to $(C_i \cap C'_i) \setminus \hat{B}_i$, and so is c_i, then c'_i is close to c_i. Formally, we have $d'(c'_i, C'_i) \leq d'(c_i, C'_i)$. We divide C'_i into $(C_i \cap C'_i) \setminus \hat{B}_i$, $\hat{B}_i \setminus M_i$ and A_i, and move terms on $(C_i \cap C'_i) \setminus \hat{B}_i$ to one side (the cost more than c_i on $(C_i \cap C'_i) \setminus \hat{B}_i$), the rest terms to another side (the cost saved on $\hat{B}_i \setminus M_i$ and A_i). After translating from d' to d, we apply triangle inequality and obtain the claim. □

Claim 2. *Let $I_i = 1$ if $c_i \neq c'_i$ and $I_i = 0$ otherwise. Then we have $\sum_{1 \leq i \leq k} I_i d(c_i, (C_i \cap C'_i) \setminus \hat{B}_i) \leq \sum_{1 \leq i \leq k} \frac{|C_i|}{3} d(c_i, c'_i)$.*

Proof Sketch: The key idea is that the clustering that under d' assigns points in $C'_i \setminus \hat{B}_i$ to c_i and points p in $\hat{B}_i \setminus M_i$ to $c(p)$, saves much cost on $(C_i \cap C'_i) \setminus \hat{B}_i$ compared to the optimal clustering $\{C'_i\}$ under d', if $c'_i \neq c_i$. Then $\{C'_i\}$ must save this cost on other parts of points. So $\{c'_i\}$ should be near these points and $\{c_i\}$ should be far away, and the weighted sum of the distances between $\{c'_i\}$ and $\{c_i\}$ should be large. Formally, $\sum_i d'(c'_i, C'_i) \leq \sum_i [d'(c_i, C'_i \setminus \hat{B}_i) + \sum_{p \in \hat{B}_i \setminus M_i} d'(c(p), p)]$ since $\{c'_i\}$ are the optimal centers for C'_i under d'. By dividing C'_i into $A_i, \hat{B}_i \setminus M_i$ and $(C_i \cap C'_i) \setminus \hat{B}_i$, and by the fact $\alpha \sum_i d(c_i, C_i) \leq \alpha \sum_i d(c'_i, C_i)$ since c_i are the optimal centers, we can show that $\{C'_i\}$ should save as much as approximately $(\alpha - 1) \sum_i d(c_i, (C_i \cap C'_i) \setminus \hat{B}_i)$ cost on points other than $(C_i \cap C'_i) \setminus \hat{B}_i$. Then the result follows by triangle inequality. □

These claims lead to $\sum_{1 \leq i \leq k} |C_i| d(c_i, c'_i) [1 - (\alpha + 2) I_i / (\alpha + 1)] \geq 0$. If $I_i = 0$, then $d(c_i, c'_i) = 0$; if $I_i = 1$, the coefficient of $d(c_i, c'_i)$ is negative. So the left hand side is at most 0. Then all terms equal 0, i.e. $d(c_i, c'_i) = 0 (1 \leq i \leq k)$. Then points in $\hat{B}_i$ will

move to other clusters after perturbation, which means that $\hat{B}_i \subseteq M_i$, thus $\hat{B} \subseteq \cup_i M_i$. Then $|\hat{B}| \leq |\cup_i M_i| \leq \epsilon n$. In particular, $|\hat{B}_i| \leq \epsilon n$ for any i. Then $|B_i| \leq \epsilon n$, otherwise $|\hat{B}_i|$ would be $\epsilon n + 1$. So $\hat{B}_i = B_i$, and $\hat{B} = B$ and $|B| = |\hat{B}| \leq \epsilon n$. □

4.2 Approximating the Optimal Clustering

Since (α, ϵ)-perturbation resilient instances have at most ϵn bad points, we can show that for $\alpha > 4$ such instances satisfy the ϵ-strict separation property (the property that after eliminating an ϵ fraction of the points, the remaining points are closer to points in their own cluster than to other points in different clusters). Therefore, we could use the algorithms in [3] to output a tree with a pruning ϵ-close to the optimal clustering. However, this pruning might not have a small cost and it is not clear how to retrieve a small cost clustering from the tree constructed by these generic algorithms. Here we design a new algorithm for obtaining a good approximation for (α, ϵ)-perturbation resilient instances. This algorithm first uses a novel linkage procedure based on an approximate version of the closure condition in Section 3 to construct a tree, and then processes the tree to output a desired clustering. We first define the approximate closure condition.

Definition 7. *Suppose $\mathcal{C}'$ is a clustering of S and $p, q \in S$.*
Let $U_{p,q}$ denote the set of clusters that are nearly contained in the ball $\mathbb{B}(p, d(p,q))$, i.e. $U_{p,q} = \{C | C \in \mathcal{C}', |C \setminus \mathbb{B}(p, d(p,q))| \leq \epsilon n, C \cap \mathbb{B}(p, d(p,q)) \neq \emptyset\}$.
*The ball $\mathbb{B}(p, d(p,q))$ satisfies the **approximate closure** condition with respect to $\mathcal{C}'$ if $|\cup_{C \in U_{p,q}} C| \geq \min_i |C_i| - \epsilon n$ and the following conditions are satisfied:*
(1) approximate coverage: it covers most of $U_{p,q}$, i.e. $|\cup_{C \in U_{p,q}} C_i \setminus \mathbb{B}(p, d(p,q))| \leq \epsilon n$;[2]
(2) approximate margin: after removing a few points outside the ball, points inside are closer to each other than to points outside, i.e. $\exists E \subseteq S \setminus \mathbb{B}(p, d(p,q)), |E| \leq \epsilon n$, s.t. $\forall p_1, p_2 \in \mathbb{B}(p, d(p,q))$, $q_1 \in S \setminus \mathbb{B}(p, d(p,q)) \setminus E$, we have $d(p_1, p_2) < d(p_1, q_1)$.

We are now ready to present our main algorithm for the (α, ϵ)-perturbation resilient instances, Algorithm 2. Informally, it starts with singleton points in their own clusters. It then checks in increasing order of $d(p,q)$ whether the ball $\mathbb{B}(p, d(p,q))$ satisfies the approximate closure condition, and if so it merges all the clusters nearly contained within $\mathbb{B}(p, d(p,q))$. As we show below, the tree produced has a pruning that respects the optimal clustering. However, this pruning may contain more than k-clusters, so in the second phase, we clean the tree so that there is a pruning with k-clusters that coincides with the optimal clustering on the good points. Finally we run dynamic programming to get the minimum cost pruning, which provides a good approximation to the optimum.

Our main result in this section is Theorem 4, which follows from Lemma 3 for Phase 1 of the algorithm and Lemma 4 for Phase 2.

Theorem 4. *For (α, ϵ)-perturbation resilient instances to k-median, if $\alpha > 2 + \sqrt{7}$ and $\epsilon \leq \rho/8$ where $\rho = \min_i |C_i|/n$, then in polynomial time, Algorithm 2 outputs a tree $\tilde{T}$ that contains a pruning ϵ-close to the optimal clustering. Moreover, if $\epsilon \leq \rho\epsilon'/8$ where $\epsilon' \leq 1$, the clustering produced is a $(1 + \epsilon')$-approximation to the optimum.*

[2] Note that in the definition of $U_{p,q}$, each cluster in it has at most ϵn points outside $\mathbb{B}(p, d(p,q))$. But the approximate coverage is stronger: $U_{p,q}$, as a whole, can have at most ϵn outside.

Algorithm 2. k-median, (α, ϵ) perturbation resilience

Input: Data set S, distance function $d(\cdot,\cdot)$ on S, $\min_i |C_i|$, $\epsilon > 0$
Phase 1: Initialize $\mathcal{C}'$ to be the clustering with each singleton point being a cluster.
• Sort all the pairwise distances $d(p,q)$. For $d(p,q)$ in ascending order,
• If $\mathbb{B}(p, d(p,q))$ satisfies approximate closure condition and $|U_{p,q}| > 1$, merge $U_{p,q}$.
• Construct the tree T with points as leaves and internal nodes corresponding to the merges.
Phase 2: If a node has only singleton points as children, delete his children; get T'.
• Assign any singleton node p to the non-singleton leaf of smallest median distance; get $\tilde{T}$.
Phase 3: Apply dynamic programming on $\tilde{T}$ to get the minimum cost pruning $\tilde{\mathcal{C}}$.
Output: Clustering $\tilde{\mathcal{C}}$, (optional) tree $\tilde{T}$.

Lemma 3. *If $\alpha > 2+\sqrt{7}$, $\epsilon \le \rho/8$, then the tree T contains nodes $N_i (1 \le i \le k)$ such that $N_i \setminus B = C_i \setminus B$.*

Proof Sketch: For each i, we let $q_i^* = \arg\max_{q \in C_i \setminus B} d(c_i, q)$. The proof follows from two key facts: (1) If $\mathcal{C}' \setminus B$ is laminar to $\mathcal{C} \setminus B$ right before checking some $d(p,q)$, and $U_{p,q}$ contains both good points from C_i and $C_j (i \neq j)$, then $d(c_i, q_i^*)$ and $d(c_j, q_j^*)$ are checked before $d(p,q)$. (2) If $\mathcal{C}' \setminus B$ is laminar to $\mathcal{C} \setminus B$ right before checking $d(c_i, q_i^*)$, we have that right after checking $d(c_i, q_i^*)$ there is a cluster containing all the good points in cluster i and no other good points.

Consider any merge step s.t. $U_{p,q}$ contains good points from both C_i and $C_j (j \neq i)$. Fact (1) implies both $d(c_i, q_i^*)$ and $d(c_j, q_j^*)$ must have been checked, and then fact (2) implies all good points in C_i and C_j respectively have already been merged. So the laminarity is always satisfied. Then the lemma follows from fact (2).

We now prove fact (1). Suppose that there exist good points from C_i and C_j in $U_{p,q}$. From the laminarity assumption, the fact that clusters in $U_{p,q}$ have only ϵn points outside $\mathbb{B}(p, d(p,q))$ and $|B| \le \epsilon n$, we can show there exist good points $p_i \in C_i$ and $p_j \in C_j$ in $\mathbb{B}(p, d(p,q))$. When $\alpha > 2+\sqrt{7}$ we can show $d(c_i, q_i^*) < d(p_i, p_j)/2$, and by triangle inequality $d(p_i, p_j)/2 \le d(p,q)$, so $d(p,q) > d(c_i, q_i^*)$. The same argument leads to $d(p,q) > d(c_j, q_j^*)$. So $d(c_i, q_i^*)$ and $d(c_j, q_j^*)$ are checked before $d(p,q)$.

We now prove fact (2). It is sufficient to show that $\cup_{C \in U_{c_i,q_i^*}} C \setminus B = C_i \setminus B$ and U_{c_i,q_i^*} satisfies the approximate closure condition. First, U_{c_i,q_i^*} contains no good points outside C_i by fact (1). Second, any C containing good points from C_i is in U_{c_i,q_i^*}. By fact (1), C has no good points outside C_i. Since $\mathbb{B}(c_i, d(c_i, q_i^*))$ contains all good points in C_i, C has only bad points outside the ball, so $C \in U_{c_i,q_i^*}$. We finally show U_{c_i,q_i^*} satisfies the approximate closure condition. Since in addition to all good points in C_i, $\cup_{C \in U_{c_i,q_i^*}} C$ can only contain bad points, it has at most ϵn points outside $\mathbb{B}(c_i, d(c_i, q_i^*))$, so approximate coverage condition is satisfied. And we can show for $\alpha > 2+\sqrt{7}$, $2d(c_i, q_i^*)$ is smaller than the distance between any point in $\mathbb{B}(c_i, d(c_i, q_i^*))$ and any good point outside C_i. Then let $E = B \setminus \mathbb{B}(c_i, d(c_i, q_i^*))$, approximate margin condition is satisfied. We also have $|\cup_{C \in U_{c_i,q_i^*}} C| \ge |C_i \setminus B| \ge \min_i |C_i| - \epsilon n$. □

Lemma 4. *If $\alpha > 2+\sqrt{7}$, $\epsilon \le \epsilon'\rho/8$ where $\epsilon' \le 1$, then $\tilde{\mathcal{C}}$ is a $(1+\epsilon')$-approximation.*

Proof Sketch: By Lemma 3, T has a pruning $\mathcal{P}$ that contains $N_i (1 \le i \le k)$ and possibly some bad points, such that $N_i \setminus B = C_i \setminus B$. Therefore, each non-singleton

leaf in T' has only good points from one optimal cluster and has more good points than bad points. This implies that each singleton good point in T' is assigned to a leaf that has good points from its own optimal cluster.

So after Phase 2, $\mathcal{P}$ in T becomes $\mathcal{P}' = \{N_i'\}$ in $\tilde{T}$ such that $N_i' \setminus B = C_i \setminus B$. It is sufficient to prove the cost of $\mathcal{P}'$ approximates $\mathcal{OPT}$, i.e. to bound the increase of cost caused by a bad point $p_j \in C_j$ ending up in $N_i'(i \neq j)$. There are two cases: p_j belongs to a non-singleton leaf node in T' or p_j is a singleton in T'. In either case, we can find $K = (\min_i |C_i| - \epsilon n)/2 - \epsilon n$ good points p_{it} from C_i in the leaf in which p_j ends up in $\tilde{T}$, and K good points p_{js} from C_j in any other leaf containing only good points from C_j, such that $d(p_j, p_{it}) \leq d(p_j, p_{js})$. Then $d(p_j, c_i) - d(p_j, c_j)$ can be bounded by

$$\frac{1}{K}\left\{\sum_{1\leq t\leq K}\left[d(p_j,p_{it})+d(p_{it},c_i)\right]-\sum_{1\leq s\leq K}\left[d(p_j,p_{js})-d(p_{js},c_j)\right]\right\}\leq\frac{1}{K}\mathcal{OPT}.$$

As $|B| \leq \epsilon n$, the cost of $\mathcal{P}'$ is $\leq (1 + \frac{\epsilon n}{K})\mathcal{OPT}$. Setting $\epsilon' \geq \frac{\epsilon n}{K}$ gives the lemma. □

We note that approximate margin condition in the Definition 7 can be verified in $O(n^3)$ time by enumerating $p_1, p_2 \in \mathbb{B}(p, d(p,q)), q_1 \notin \mathbb{B}(p, d(p,q))$, and checking if there are no more than ϵn such q_1 that there exist p_1, p_2 violating the condition. So the algorithm runs in polynomial time.

5 α-Perturbation Resilience for the Min-Sum Objective

In this section we provide an efficient algorithm for clustering α-perturbation resilient instances for the min-sum k-clustering problem (Algorithm 3). We use the following notations: $d_{avg}(A, B) = d(A, B)/(|A||B|)$ and $d_{avg}(p, B) = d_{avg}(\{p\}, B)$.

Theorem 5. *For* $(3\frac{\max_i |C_i|}{\min_i |C_i|-1})$*-perturbation resilient instances to min-sum, Algorithm 3 outputs the optimal min-sum* k*-clustering in polynomial time.*

Algorithm 3. Min-sum, α perturbation resilience

Input: Data set S, distance function $d(\cdot,\cdot)$ on S, $\min_i |C_i|$.
Phase 1: Connect each point with its $\frac{1}{2}\min_i |C_i|$ nearest neighbors.
- Initialize the clustering $\mathcal{C}'$ with each connected component being a cluster.
- Repeat till one cluster remains in $\mathcal{C}'$: merge clusters C, C' that minimize $d_{avg}(C, C')$.
- Let T be the tree with components as leaves and internal nodes corresponding to the merges.

Phase 2: Apply dynamic programming on T to get the minimum cost pruning $\tilde{\mathcal{C}}$.
Output: Output $\tilde{\mathcal{C}}$.

Proof Sketch: First we show that the α-perturbation resilience property implies that for any two optimal clusters C_i and C_j and any $A \subseteq C_i$, we have $\alpha d(A, C_i \setminus A) < d(A, C_j)$. This follows by considering the perturbation where $d'(p,q) = \alpha d(p,q)$ if $p \in A, q \in C_i \setminus A$ and $d'(p,q) = d(p,q)$ otherwise, and using the fact that the optimum does not change after the perturbation. This can be used to show that

when $\alpha > 3\frac{\max_i |C_i|}{\min_i |C_i|-1}$ we have: (1) for any optimal clusters C_i and C_j and any $A \subseteq C_i$, $A' \subseteq C_j$ s.t. $\min(|C_i \setminus A|, |C_j \setminus A'|) > \min_i |C_i|/2$ we have $d_{avg}(A, A') > \min\{d_{avg}(A, C_i \setminus A), d_{avg}(A', C_j \setminus A')\}$; (2) for any point p in the optimal cluster C_i, twice its average distance to points in $C_i \setminus \{p\}$ is smaller than the distance to any point in other optimal cluster C_j. Fact (2) implies that for any point $p \in C_i$ its $|C_i|/2$ nearest neighbors are in the same optimal cluster, so the leaves of the tree T are laminar to the optimum clustering. Fact (1) can be used to show that the merges preserve the laminarity with the optimal clustering, so the minimum cost pruning of T will be the optimal clustering, as desired. See the full version for the details. □

6 Discussion and Open Questions

In this work, we advance the line of research on perturbation resilience in clustering in multiple ways. For α-perturbation resilient instances, we improve on the known guarantees for center-based objectives and give the first analysis for min-sum. Furthermore, for k-median, we analyze and give the first algorithmic guarantees known for a relaxed but more challenging condition of (α, ϵ)-perturbation resilience, where an ϵ fraction of points are allowed to move after perturbation. We also give sublinear-time algorithms for k-median and min-sum under perturbation resilience in the long version.

A natural direction for future investigation is to explore whether one can take advantage of smaller perturbation factors for perturbation resilient instances in Euclidian spaces. More broadly, it would be interesting to explore other ways in which perturbation resilient instances behave better than worst case instances (e.g., natural algorithms converge faster).

Acknowledgments. This work was supported by NSF grant CCF-0953192, by AFOSR grant FA9550-09-1-0538, by a Microsoft Research Faculty Fellowship, and by a Google Research award.

References

1. Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., Pandit, V.: Local search heuristics for k-median and facility location problems. SIAM J. Comput. 33(3) (2004)
2. Awasthi, P., Blum, A., Sheffet, O.: Center-based clustering under perturbation stability. Inf. Process. Lett. 112(1-2), 49–54 (2012)
3. Balcan, M.F., Gupta, P.: Robust hierarchical clustering. In: COLT (2010)
4. Balcan, M.F., Liang, Y.: Clustering under Perturbation Resilience. CoRR, abs/1112.0826 (2011)
5. Bartal, Y., Charikar, M., Raz, D.: Approximating min-sum -clustering in metric spaces. In: STOC (2001)
6. Bilu, Y., Linial, N.: Are stable instances easy? In: Innovations in Computer Science (2010)
7. Charikar, M., Guha, S., Tardos, É., Shmoys, D.B.: A constant-factor approximation algorithm for the k-median problem. J. Comput. Syst. Sci. 65(1) (2002)
8. de la Vega, W.F., Karpinski, M., Kenyon, C., Rabani, Y.: Approximation schemes for clustering problems. In: STOC (2003)
9. Jain, K., Mahdian, M., Saberi, A.: A new greedy approach for facility location problems. In: STOC (2002)
10. Reyzin, L.: Data stability in clustering: A closer look. CoRR, abs/1107.2379 (2011)