Data Stability in Clustering: A Closer Look

Lev Reyzin

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607
lreyzin@math.uic.edu

Abstract

We consider the model introduced by Bilu and Linial [14], who study problems for which the optimal clustering does not change when distances are perturbed. They show that even when a problem is NP-hard, it is sometimes possible to obtain efficient algorithms for instances resilient to certain multiplicative perturbations, e.g. on the order of $O(\sqrt{n})$ for max-cut clustering. Awasthi et al. [7] consider center-based objectives, and Balcan and Liang [10] analyze the k-median and min-sum objectives, giving efficient algorithms for instances resilient to certain constant multiplicative perturbations.

Here, we are motivated by the question of to what extent these assumptions can be relaxed while allowing for efficient algorithms. We show there is little room to improve these results by giving NP-hardness lower bounds for both the k-median and min-sum objectives. On the other hand, we show that multiplicative resilience parameters, even only on the order of $\Theta(1)$, can be so strong as to make the clustering problem trivial, and we exploit these assumptions to present a simple one-pass streaming algorithm for the k-median objective. We also consider a model of additive perturbations and give a correspondence between additive and multiplicative notions of stability. Our results provide a close examination of the consequences of assuming, even constant, stability in data.

1. Introduction

Clustering is one of the most widely-used techniques in statistical data analysis. The need to partition, or cluster, data into meaningful categories naturally arises in virtually every domain where data is abundant. Unfortunately, most of the natural clustering objectives, including k-median, k-means, and min-sum, are NP-hard to optimize [19, 21]. It is, therefore, unsurprising that many of the clustering algorithms used in practice come with few guarantees.

Motivated by overcoming the hardness results, Bilu and Linial [14] consider a perturbation **resilience assumption** that they argue is often implicitly made when choosing a clustering objective: that the optimum clustering to the desired objective Φ is preserved under multiplicative perturbations up to a factor $\alpha > 1$ to the distances between the points. They reason that if the optimum clustering to an objective Φ is not

resilient, as in, if small perturbations to the distances can cause the optimum to change, then Φ may have been the wrong objective to be optimizing in the first place. Bilu and Linial [14] show that for max-cut clustering, instances resilient to perturbations of $\alpha = O(\sqrt{n})$ have efficient algorithms for recovering the optimum itself.

Continuing that line of research, Awasthi et al. [7] give a polynomial time algorithm that finds the optimum clustering for instances resilient to multiplicative perturbations of $\alpha = 3$ for center-based[1] clustering objectives when centers must come from the data (we call this the **proper** setting), and $\alpha = 2 + \sqrt{3}$ when when the centers do not need to (we call this the **Steiner** setting). Their method relies on a **stability** property implied by perturbation resilience (see Section 2). For the Steiner case, they also prove an NP-hardness lower bound of $\alpha = 3$. Subsequently, Balcan and Liang [10] consider the proper setting and improve the constant past $\alpha = 3$ by giving a new polynomial time algorithm for the k-median objective for $\alpha = 1 + \sqrt{2} \approx 2.4$ stable instances.

1.1. Our results

Our work further delves into the proper setting, for which no lower bounds have previously been shown for the stability property. In Section 3 we show that even in the proper case, where the algorithm is restricted to choosing its centers from the data, for any $\epsilon > 0$, it is NP-hard to optimally cluster $(2 - \epsilon)$-stable instances, both for the **k-median** and **min-sum** objectives (Theorems 5 and 7). To prove this for the min-sum objective, we define a new notion of stability that is implied by perturbation resilience, a notion that may be of independent interest.

Then in Section 4, we look at the implications of assuming resilience or stability in the data, even for a constant perturbation parameter α. We show that for even fairly small constants, the data begins to have very strong structural properties, as to make the clustering task fairly trivial. When α approaches ≈ 5.7, the data begins to show what is called **strict separation**, where each point is closer to points in its own cluster than to points in other clusters (Theorem 9). We show that with strict separation, optimally clustering in the very restrictive one-pass streaming model becomes possible (Theorem 11).

Finally, in Section 5, we look at whether the picture can be improved for clustering data that is stable under additive, rather than multiplicative, perturbations. One hope would be that **additive stability** is a more useful assumption, where a polynomial time algorithm for ϵ-stable instances might be possible. Unfortunately, this is not the case. We consider a natural additive model and show that severe lower bounds hold for the additive notion as well (Theorems 16 and 20). On the positive side, we show via reductions that algorithms for multiplicatively stable data also work for additively stable data for a different but related parameter.

[1] For center-based clustering objectives, the clustering is defined by a choice of centers, and the objective is a function of the distances of the points to their closest center.

Our results demonstrate that on the one hand, it is hard to improve the algorithms to work for low stability constants, and that on the other hand, higher stability constants can be quite strong, to the point of trivializing the problem. Furthermore, switching from a multiplicative to an additive stability assumption does not help to circumvent the hardness results, and perhaps makes matters worse. These results, taken together, narrow the range of interesting parameters for theoretical study and highlight the strong role that the choice of constant plays in stability assumptions.

One thing to note that there is some difference between the very related resilience and stability properties (see Section 2), stability being weaker and more general [7]. Some of our results apply to both notions, and some only to stability. This still leaves open the possibility of devising polynomial-time algorithms that, for a much smaller α, work on all the α-perturbation resilient instances, but not on all α-stable ones.

1.2. Previous work

We examine previous work on stability, both as a data dependent assumption in clustering and in other settings.

1.2.1. Stability as a data assumption in clustering

The classical approach in theoretical computer science to dealing with the worst-case NP-hardness of clustering has been to develop efficient approximation algorithms for the various clustering objectives [3, 4, 11, 15, 22, 17], and significant efforts have been exerted to improve approximation ratios and to prove lower bounds. In particular, for metric k-median, the best known guarantee is a $(3+\epsilon)$-approximation [4], and the best known lower bound is $(1+1/e)$-hardness of approximation [19, 21]. For metric min-sum, the best known result is a $O(\text{polylog}(n))$-approximation to the optimum [11].

In contrast, a more recent direction of research has been to characterize under what conditions we can find a desirable clustering efficiently. Perturbation resilience/stability are such conditions, but they are related to other stability notions in clustering. Ostrovsky et al. [27] demonstrate the effectiveness of Lloyd-type algorithms [24] on instances with the stability property that the cost of the optimal k-means solution is small compared to the cost of the optimal $(k-1)$-means solution, and their guarantees have later been improved by Awasthi et al. [6].

In a different line of work, Balcan et al. [9] consider what stability properties of a similarity function, with respect to the ground truth clustering, are sufficient to cluster well. In a related direction, Balcan et al. [8] argue that, for a given objective Φ, approximation algorithms are most useful when the clusterings they produce are structurally close to the optimum originally sought in choosing to optimize Φ in the first place. They then show that, for many objectives, if one makes this assumption explicit – that all c-approximations to the objective yield a clustering that is ϵ-close to the optimum – then one can recover an ϵ-close clustering

in polynomial time, even for values of c below the hardness of approximation constant. The assumptions and algorithms of Balcan et al. [8] have subsequently been carefully analyzed by Schalekamp et al. [28].

Ackerman and Ben-David [1] also study various notions of resilience, and among their results, introduce a notion of stability similar to the one studied herein, except only the positions of cluster centers are perturbed. Their notion is strictly weaker – i.e. any perturbation resilient instance is also stable in their framework. They show that Euclidean instances stable to perturbations of cluster centers will have polynomial algorithms for finding near-optimal clusterings. Our results, however, hold for more general metric spaces, which evidently is a harder setting for perturbation-resilient clustering.

1.2.2. Stability in other settings

Just as the Bliu and Linial [14] notion of stability gives conditions under which efficient clustering is possible, similar concepts have been studied in game theory. Lipton et al. [23] propose a notion of stability for solution concepts of games. They define a game to be stable if small perturbations to the payoff matrix do not significantly change the value of the game, and they show games are generally not stable under this definition. Then, in a similar spirit to the work of Bilu and Linial, Awasthi et al. [5] propose a related stability condition for a game, which can be leveraged in finding its approximate Nash equilibria.

The Bilu and Linial [14] notion of stability has also been studied in the context of the metric traveling salesman problem, for which Mihalák et al. [25] give efficient algorithms for 1.8-perturbation resilient instances, illustrating another case where a stability assumption can circumvent NP-hardness.

From a different direction, Ben-David et al. [13] consider the stability of clustering algorithms, as opposed to instances. They say an algorithm is stable if it produces similar clusterings for different inputs drawn from the same distribution. They argue that stability is not as useful a notion as had been previously thought in determining various parameters, such as the optimal number of clusters.

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 clusters which we want to optimize over the metric, i.e. Φ assigns a score to every clustering. The optimal clustering with respect to Φ is denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$.

The **k-median objective** requires S to be partitioned into k disjoint subsets $\{S_1, \ldots, S_k\}$ and each subset S_i to be assigned a center $s_i \in S$. The goal is to minimize Φ_{med}, measured by

$$\phi_{\text{med}}(S_1, \ldots, S_k) \doteq \sum_{i=1}^{k} \sum_{p \in S_i} d(p, s_i).$$

The centers in the optimal clustering are denoted as $c_1, \ldots, c_k$. In an optimal solution, each point is assigned to its nearest center.

For the **min-sum objective**, S is partitioned into k disjoint subsets, and the objective is to minimize $\Phi_{\mathrm{m-s}}$, measured by

$$\phi_{\mathrm{m-s}}(S_1, \ldots, S_k) \doteq \sum_{i=1}^{k} \sum_{p,q \in S_i} d(p,q).$$

Now, we define the perturbation resilience notion introduced by Bilu and Linial [14].

Definition 1. *For $\alpha > 1$, a clustering instance (S,d) is α-**perturbation resilient** to a given objective Φ if for any function $d' : S \times S \to \mathbb{R}_{\geq 0}$ such that $\forall p, q \in S$,*

$$d(p,q) \leq d'(p,q) \leq \alpha d(p,q),$$

there is a unique optimal clustering $\mathcal{C}'$ for Φ under d' and this clustering is equal to the optimal clustering $\mathcal{C}$ for Φ under d.

In this paper, we consider the k-median and min-sum objectives, and we thereby investigate the following definitions of stability, which are implied by perturbation resilience, as shown in Sections 3.1 and 3.2. The following definition is adapted from Awasthi et al. [7].

Definition 2. *A clustering instance (S,d) is α-**center stable** for the k-median objective 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*

$$\alpha d(p, c_i) < d(p, c_j).$$

Next, we define a new analogous notion of stability for the min-sum objective, and we show in Section 3.2 that for the min-sum objective, perturbation resilience implies min-sum stability. To help with exposition for the min-sum objective, we define the distance from a point p to a set of points A,

$$d(p, A) \doteq \sum_{q \in A} d(p,q).$$

Definition 3. *A clustering instance (S,d) is α-**min-sum stable** for the min-sum objective if for all optimal clusters $C_i, C_j \in \mathcal{C}$ $(j \neq i)$, any point $p \in C_i$ satisfies*

$$\alpha d(p, C_i) < d(p, C_j).$$

This is a useful generalization because, as we shall see, known algorithms working under the perturbation resilience assumption can also be made to work under the weaker notion of min-sum stability.

3. Lower bounds

3.1. The k-median objective

Awasthi et al. [7] prove the following connection between perturbation resilience and stability. Both their algorithms and the algorithms of Balcan and Liang [10] crucially use this stability assumption.

Lemma 4. *Any clustering instance that is α-perturbation resilient for the k-median objective also satisfies the α-center stability.*

Awasthi et al. [7] proved that for $\alpha < 3 - \epsilon$, k-median clustering α-center stable instances is NP-hard when Steiner points are allowed in the data. Afterwards, Balcan and Liang [10] circumvented this lower bound and achieved a polynomial time algorithm for $\alpha = 1 + \sqrt{2}$ by assuming the algorithm must choose cluster centers from within the data.

In the theorem below, we prove a lower bound for the center stable property in this more restricted setting, showing there is little hope of progress even for data where each point is nearly twice closer to its own center than to any other.

Theorem 5. *For any $\epsilon > 0$, the problem of solving $(2 - \epsilon)$-center stable k-median instances is NP-hard.*

Proof. We reduce from the perfect dominating set promise problem, which we prove to be NP-hard (see Appendix), where we are promised that the input graph $G = (V, E)$ is such that all of its smallest dominating sets D are perfect, and we are asked to find a dominating set of size at most d. The reduction is simple. We take an instance of the NP-hard problem PDS-PP on $G = (V, E)$ on n vertices and reduce it to an $\alpha = 2 - \epsilon$-center stable instance. Our distance metric as follows. Every vertex $v \in V$ becomes a point in the k-center instance. For any two vertices $(u, v) \in E$ we define $d(u, v) = 1/2$. When $(u, v) \notin E$, we set $d(u, v) = 1$. This trivially satisfies the triangle inequality for any graph G, as the sum of the distances along any two edges is at least 1. We set $k = d$.

We observe that a k-median solution of cost $(n - k)/2$ corresponds to a dominating set of size d in the PDS-PP instance, and is therefore NP-hard to find. We also observe that because all solutions of size $\leq d$ in the PDS-PP instance are perfect, each (non-center) point in the k-median solution has distance $1/2$ to exactly one (its own) center, and a distance of 1 to every other center. Hence, this instance is $\alpha = (2 - \epsilon)$-center stable, completing the proof. □

3.2. The min-sum objective

Analogously to Lemma 4, we can show that α-perturbation resilience implies our new notion of α-min-sum stability.

Lemma 6. *If a clustering instance is α-perturbation resilient, then it is also α-min-sum stable.*

Proof. Assume to the contrary that the instance is α-perturbation resilient but is not α-min-sum stable. Then, there exist clusters C_i, C_j in the optimal solution $\mathcal{C}$ and a point $p \in C_i$ such that $\alpha d(p, C_i) \geq d(p, C_j)$. We perturb d as follows. We define d' such that for all points $q \in C_i$, $d'(p, q) = \alpha d(p, q)$, and for the remaining distances, $d' = d$. Clearly d' is an α-perturbation of d.

We now note that $\mathcal{C}$ is not optimal under d'. Namely, we can create a cheaper solution $\mathcal{C}'$ that assigns point p to cluster C_j, and leaves the remaining clusters unchanged, which contradicts optimality of $\mathcal{C}$. This shows that $\mathcal{C}$ is not the optimum under d' which contradicts the instance being α-perturbation resilient. Therefore we can conclude that if a clustering instance is α-perturbation resilient, then must also be α-min-sum stable. □

Moreover, we show in the Appendix that the min-sum algorithm of Balcan and Liang [10], which requires α to be bounded from below by $3\left(\frac{\max_{C\in\mathcal{C}}|C|}{\min_{C\in\mathcal{C}}|C|-1}\right)$, works with this more general condition. This further motivates following bound.

Theorem 7. *For any $\epsilon > 0$, the problem of finding an optimal min-sum k clustering in $(2-\epsilon)$-min-sum stable instances is NP-hard.*

Proof. Consider the **triangle partition problem**. Let graph $G = (V, E)$ and $|V| = n = 3k$, and let each vertex have maximum degree of $d = 4$. The problem of whether the vertices of G can be partitioned into sets $V_1, V_2, \ldots, V_k$ such that each V_i contains a triangle in G is NP-complete [18], even with the degree restriction [29].

We reduce the triangle partition problem to an $\alpha = (2-\epsilon)$-min-sum stable clustering instance. The metric is as follows. Every vertex $v \in V$ becomes a point in the min-sum instance. For any two vertices $(u, v) \in E$ we define $d(u, v) = 1/2$. When $(u, v) \notin E$, we set $d(u, v) = 1$. This satisfies the triangle inequality for any graph, as the sum of the distances along any two edges is at least 1.

Now we show that we can cluster this instance into k clusters such that the cost of the min-sum objective is exactly n if and only if the original instance is a YES instance of triangle partition. This follows from two facts.

1. A YES instance of triangle partition maps to a clustering into $k = n/3$ clusters of size 3 with pairwise distances 1/2, for a total cost of n
2. A cost of n is the best achievable because a balanced clustering with all minimum pairwise intra-cluster distances is optimal.

In the clustering from our reduction, each point has a sum-of-distances to its own cluster of 1. Now we examine the sum-of-distances of any point to other clusters. A point has two distances of 1/2 (edges) to its own cluster, and because $d = 4$, it can have at most two more distances of 1/2 (edges) into any other

cluster, leaving the third distance to the other cluster to be 1, yielding a total cost of ≥ 2 into any other cluster. Hence, it is $\alpha = (2-\epsilon)$-min-sum stable. □

We note that it is tempting to restrict the degree bound to 3 in order to further improve the lower bound. Unfortunately, the triangle partition problem on graphs of maximum degree 3 is polynomial-time solvable [29], and we cannot improve the factor of $2-\epsilon$ by restricting to graphs of degree 3 in this reduction.

4. Strong consequences of stability

In Section 3, we showed that k-median clustering even $(2-\epsilon)$-center stable instances is NP-hard. In this section we show that even for resilience to constant multiplicative perturbations of $\alpha > \frac{1}{2}(5+\sqrt{41}) \approx 5.7$, the data obtains a property referred to as **strict separation**, where all points are closer to all other points in their own cluster than to points in any other cluster; this property is known to be helpful in clustering [9]. Then we show that this property renders center-based clustering fairly trivial even in the difficult one-pass streaming model.

4.1. Strict separation

We will make use of the following lemma, whose proof follows from the triangle inequality. A similar observation appears in [10].

Lemma 8. *For any two points p and p' belonging to different centers c_i and c_j, respectively, in the optimal clustering of an α-center stable instance,*

$$d(c_i, p') > \frac{\alpha(\alpha-1)}{\alpha+1} d(c_i, p).$$

Proof. By triangle inequality, we have $d(c_i, c_j) \leq d(c_i, p') + d(p', c_j)$ and also $d(c_j, p) \leq d(c_j, c_i) + d(c_i, p)$. Combining the two inequalities, we get

$$d(c_j, p) - (c_i, p) \leq d(c_i, p') + d(p', c_j).$$

Applying the definition α-center stability to each side separately, we get

$$(\alpha-1)d(c_i, p) < \left(1 + \frac{1}{\alpha}\right) d(c_i, p'),$$

finishing the proof. □

Now we can prove the following theorem, which shows that even for relatively small multiplicative constants for α, center stable, and therefore perturbation resilient, instances exhibit strict separation.

Theorem 9. *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the optimal clustering of a $\frac{1}{2}(5+\sqrt{41})$-center stable instance. Let $p, p' \in C_i$ and $q \in C_j$ $(i \neq j)$, then $d(p, q) > d(p, p')$.*

Proof. Let $\{c_1, \ldots, c_k\}$ be the centers of clusters $\{C_1, \ldots, C_k\}$. Define

$$p_f \doteq \arg\max_{r \in C_i} d(p, r).$$

By Lemma 8 we have

$$d(c_i, q) > \frac{\alpha(\alpha-1)}{\alpha+1} d(c_i, p)$$

and also

$$d(c_i, q) > \frac{\alpha(\alpha-1)}{\alpha+1} d(c_i, p_f).$$

Adding the two gives us

$$\frac{\alpha(\alpha-1)}{\alpha+1} d(c_i, p) + \frac{\alpha(\alpha-1)}{\alpha+1} d(c_i, p_f) < 2d(c_i, q),$$

and by the triangle inequality, we get

$$\frac{\alpha(\alpha-1)}{\alpha+1} d(p, p_f) < 2d(c_i, q). \tag{1}$$

We also have

$$d(c_i, q) \leq d(p, c_i) + d(p, q). \tag{2}$$

Combining Equations 1 and 2, and by the definition of p_f, we have

$$\begin{aligned} \frac{\alpha(\alpha-1)}{\alpha+1} d(p, p_f) &< 2d(p, c_i) + 2d(q, p) \\ &\leq 2d(p, p_f) + 2d(q, p). \end{aligned}$$

From the RHS and LHS of the above, it follows from the definitions of p_f and p' that

$$\begin{aligned} d(p, q) &> \left(\frac{\alpha(\alpha-1)}{2(\alpha+1)} - 1 \right) d(p, p_f) \\ &\geq \left(\frac{\alpha(\alpha-1)}{2(\alpha+1)} - 1 \right) d(p, p'). \end{aligned}$$

Finally, the statement of the Lemma follows by setting $\alpha \geq \frac{1}{2}(5 + \sqrt{41}) \approx 5.7$. □

4.2. Clustering in the streaming model

Here, we turn to the restrictive **one-pass streaming** model. In the natural streaming model for center-based objectives, the learner sees the data $p_1, p_2, \ldots$ in one pass, and must, using limited memory and time, implicitly cluster the data by retaining k points to use as centers.

The clustering is then the one induced by placing each point in the cluster to the closest center produced by the algorithm. We note that a streaming algorithm can be used for the general batch problem, as one can present the data to the algorithm in a streaming fashion.

Streaming models have been extensively studied in the context of clustering objectives [2, 16, 20, 26], where the known approximation guarantees are weaker than in the standard offline model. We, however, show that an α-center stability assumption can make the problem of finding the optimum tractable for center-based objectives, in only one pass. We view this not so much as an advance in the state-of-the-art in clustering, but rather as an illustration of how powerful stability assumptions can be, even for constant parameter values.

For our result, we can use Theorem 9 to immediately give us the following.

Corollary 10. *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the optimal clustering of a $\frac{1}{2}(5+\sqrt{41})$-center stable instance. Any algorithm that chooses centers $\{c'_1, \ldots, c'_k\}$ such that $c'_i \in C_i$ induces the partition $\mathcal{C}$ when points are assigned to their closest centers.*

This leads to an algorithm that easily and efficiently finds the optimal clustering.

Theorem 11. *For $\frac{1}{2}(5+\sqrt{41})$-center stable instances, we can recover the optimal clustering for the k-median objective, even in one pass in the streaming model.*

Proof. Consider Algorithm 1. It proceeds as follows: it keeps k candidate centers, and whenever a new point comes in, it adds it as a candidate center and arbitrarily (choosing from at least two points) removes any point that realizes the argmin distance among the current candidate centers.

Algorithm 1 A streaming algorithm for $\frac{1}{2}(5+\sqrt{41})$-center stable instances

let $p_1, p_2, \ldots$ be the stream of points
let C be a set of candidate centers, initialized $C = \{p_1, \ldots, p_k\}$
while there is more data in stream **do**
  receive point p_i
  $C = C \cup p_i$
  let $p \in \arg\min_{\{p_j, p_k\} \in C} d(p_j, p_k)$
  $C = C \setminus p$
end while
return C (thereby inducing a clustering $\mathcal{C}$)

The correctness of this algorithm follows from two observations:

1. By the pigeonhole principle, some pair from any set of $k+1$ points must belong to the same cluster.
2. It follows from Theorem 9 that two points in different clusters cannot realize the argmin distance.

Hence, whenever a point is removed as a candidate center, it has a partner in the same optimal cluster that remains. Once the algorithm sees a point from each cluster, by Corollary 10, we get the optimal partition. □

5. Additive stability

So far, in this paper our notions of stability were defined with respect to multiplicative perturbations. Similarly, we can imagine an instance being resilient with respect to additive perturbations. Consider the following definition.

Definition 12. *Let $d : S \times S \to [0, 1]$, and let $0 < \beta \leq 1$. A clustering instance (S, d) is* ***additive β-perturbation*** *resilient to a given objective Φ if for any function $d' : S \times S \to R \geq 0$ such that $\forall p, q \in S$,*

$$d(p, q) \leq d'(p, q) \leq d(p, q) + \beta,$$

there is a unique optimal clustering $\mathcal{C}'$ for Φ under d' and this clustering is equal to the optimal clustering $\mathcal{C}$ for Φ under d.

We note that in the definition above, we require all pairwise distances between points to be at most 1. Otherwise, resilience to additive perturbations would be a very weak notion, as the distances in most instances could be scaled as to be resilient to arbitrary additive perturbations.

Especially in light of positive results for other additive stability notions [1, 12], one possible hope is that our hardness results might only apply to the multiplicative case, and that we might be able to get polynomial time clustering algorithms for instances resilient to arbitrarily small additive perturbations. We show that this is unfortunately not the case – we introduce notions of additive stability, similar to Definitions 2 and 3, and for the k-median and min-sum objectives, we show correspondences between multiplicative and additive stability.

5.1. The k-median objective

Analogously to Definition 2, we can define a notion of additive β-center stability.

Definition 13. *Let $d : S \times S \to [0, 1]$, and let $0 \leq \beta \leq 1$. A clustering instance (S, d) is* ***additive β-center stable*** *to the k-median objective 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*

$$d(p, c_i) + \beta < d(p, c_j).$$

We can now prove that perturbation resilience implies center stability.

Lemma 14. *The proof is similar to that of Lemmas 4. Any clustering instance satisfying additive β-perturbation resilience for the k-median objective also satisfies additive β-center stability.*

Proof. We prove that for every point p and its center c_i in the optimal clustering of an additive β-perturbation resilient instance, it holds that $d(p, c_j) > d(p, c_i) + \beta$ for any $j \neq i$.

Consider an additive β-perturbation resilient clustering instance. Assume we blow up all the pairwise distances within cluster C_i by an additive factor of β. As this is a legitimate perturbation of the distance function, the optimal clustering under this perturbation is the same as the original one. Hence, p is still assigned to the same cluster. Furthermore, since the distances within C_i were all changed by the same constant factor, c_i will remain the center of the cluster. The same holds for any other optimal clusters. Since the optimal clustering under the perturbed distances is unique it follows that even in the perturbed distance function, p prefers c_i to c_j, which implies the lemma. □

We now consider center stability, as in the multiplicative case. We first prove that additive center stability implies multiplicative center stability, and this gives us the property that any algorithm for $\left(\frac{1}{1-\beta}\right)$-center stable instances will work for additive β-center stable instances.

Lemma 15. *Any additive β-center stable clustering instance for the k-median objective is also (multiplicative) $\left(\frac{1}{1-\beta}\right)$-center stable.*

Proof. Let the optimal clustering be $C_1, \ldots, C_k$, with centers $c_1, \ldots, c_k$, of an additive β-center stabile clustering instance. Let $p \in C_i$ and let $i \neq j$. From the stability property,

$$d(p, c_j) > d(p, c_i) + \beta \geq \beta. \tag{3}$$

We also have $d(p, c_i) < d(p, c_j) - \beta$, from which we can see

$$\frac{1}{d(p, c_j) - \beta} < \frac{1}{d(p, c_i)}.$$

This gives us

$$\frac{d(p, c_j)}{d(p, c_i)} > \frac{d(p, c_j)}{d(p, c_j) - \beta} \geq \frac{1}{1 - \beta}. \tag{4}$$

Equation 4 is derived as follows. The middle term, for $d(p, c_j) \geq \beta$ (which we have from Equation 3), is monotonically decreasing in $d(p, c_j)$. Using $d(p, c_j) \leq 1$ bounds it from below. Relating the LHS to the RHS of Equation 4 gives us the needed stability property. □

Now we prove a lower bound that shows that the task of clustering additive $(1/2 - \epsilon)$-center stable instances with respect to the k-median objective remains NP-hard.

Theorem 16. *For any $\epsilon > 0$, the problem of finding an optimal k-median clustering in additive $(1/2 - \epsilon)$-center stable instances is NP-hard.*

Proof. We use the reduction in Theorem 5, in which the metric satisfies the needed property that $d : S \times S \to [0, 1]$. We observe that the instances from the reduction are additive $(1/2 - \epsilon)$-center stable. Hence, an algorithm for solving k-median on a $(1/2 - \epsilon)$-center stable instance can decide whether a PDS-PP instance contains a dominating set of a given size, completing the proof. □

5.2. The min-sum objective

Here we define additive min-sum stability and prove the analogous theorems for the min-sum objective.

Definition 17. *Let $d : S \times S \to [0,1]$, and let $0 \leq \beta \leq 1$. A clustering instance is* ***additive β-min-sum stable*** *for the min-sum objective if for every point p in any optimal cluster C_i, it holds that*

$$d(p, C_i) + \beta(|C_i| - 1) < d(p, C_j).$$

Lemma 18. *If a clustering instance is additive β-perturbation resilient, then it is also additive β-min-sum stable.*

Proof. Assume to the contrary that the instance is β-perturbation resilient but is not β-min-sum stable. Then, there exist clusters C_i, C_j in the optimal solution $\mathcal{C}$ and a point $p \in C_i$ such that $d(p, C_i)+\beta(|C_i|-1) \geq d(p, C_j)$. Then, we perturb d as follows. We define d' such that for all points $q \in C_i$, $d'(p, q) = d(p, q) + \beta$, and for the remaining distances $d' = d$. Clearly d' is a valid additive β-perturbation of d.

We now note that C is not optimal under d'. Namely, we can create a cheaper solution $\mathcal{C}'$ that assigns point p to cluster C_j, and leaves the remaining clusters unchanged, which contradicts optimality of $\mathcal{C}$. This shows that $\mathcal{C}$ is not the optimum under d' which is contradictory to the fact that the instance is additive β-perturbation resilient. Therefore we conclude that if a clustering instance is additive β-perturbation resilient, then it is also additive β-min-sum stable. □

As we did for the k-median objective, we can also reduce additive stability to multiplicative stability for the min-sum objective.

Lemma 19. *Let $t = \frac{\max_{C \in \mathcal{C}} |C|}{\min_{C \in \mathcal{C}} |C|-1}$. Any additive β-min-sum stabile clustering instance for the min-sum objective is also (multiplicative) $\left(\frac{1}{1-\beta/t}\right)$-min-sum stable.*

Proof. Let the optimal clustering be $C_1, \ldots, C_k$ and let $p \in C_i$. Let $i \neq j$. From the stability property, we have

$$\begin{aligned} d(p, C_j) &> d(p, C_i) + \beta(|C_i| - 1) \\ &\geq \beta(|C_i| - 1). \end{aligned} \tag{5}$$

We also have

$$d(p, C_i) < d(p, C_j) - \beta(|C_i| - 1).$$

Taking reciprocals and multiplying by $d(p, C_j)$, we get

$$\begin{aligned}
\frac{d(p,C_j)}{d(p,C_i)} &> \frac{d(p,C_j)}{d(p,C_j) - \beta(|C_i|-1)} \\
&\geq \frac{|C_j|}{|C_j| - \beta(|C_i|-1)} \qquad (6) \\
&\geq \frac{\max_{C\in\mathcal{C}} |C|}{\max_{C\in\mathcal{C}} |C_j| - \beta(\min_{C\in\mathcal{C}} |C| - 1)} \\
&\geq \frac{1}{1-\beta/t}. \qquad (7)
\end{aligned}$$

Equation 6 is derived as follows: $d(p, C_j) \geq \beta(|C_i|-1)$ (which we have from Equation 5), is monotonically decreasing in $d(p, C_j)$. Observing $d(p, c_j) \leq |C_j|$ bounds it from below. Equation 7 gives us the needed property. □

Finally, as with the k-median objective, we show that additive min-sum stability exhibits similar lower bounds as in the multiplicative case.

Theorem 20. *For any $\epsilon > 0$, the problem of finding an optimal min-sum clustering in additive $(1/2-\epsilon)$-min-sum stable instances is NP-hard.*

Proof. We use the reduction in Theorem 7, in which the metric satisfies the property that $d : S \times S \to [0, 1]$. The instances from the reduction are additive $(1/2-\epsilon)$-min-sum stable. Hence, an algorithm for clustering a $(1/2-\epsilon)$-min-sum stable instance can solve the triangle partition problem. □

6. Discussion

Our lower bounds, together with the structural properties implied by fairly small constants, illustrate the importance parameter settings play in stability assumptions. These results make us wonder the degree to which the assumptions studied herein hold in practice; empirical study of real datasets is warranted.

Another interesting direction is to relax the assumptions. Awasthi et al. [7] suggest considering stability under random, and not worst-case, perturbations. Balcan and Liang [10] also study a relaxed version of the assumption, where perturbations can change the optimal clustering, but not by much. It is open to what extent, and on what data, any of these approaches will yield practical improvements.

Acknowledgements

We thank Maria-Florina Balcan and Yingyu Liang for helpful discussions, Avrim Blum and Santosh Vempala for feedback on the writing, and Shai Ben-David for useful pointers.

This work was supported in part by a Simons Postdoctoral Fellowship in Theoretical Computer Science while the author was at the Georgia Institute of Technology.

References

[1] Ackerman, M., and Ben-David, S. Clusterability: A theoretical study. *Journal of Machine Learning Research - Proceedings Track 5* (2009), 1–8.

[2] Ailon, N., Jaiswal, R., and Monteleoni, C. Streaming k-means approximation. In *NIPS* (2009).

[3] Arora, S., Raghavan, P., and Rao, S. Approximation schemes for euclidean -medians and related problems. In *STOC* (1998), pp. 106–113.

[4] Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., and Pandit, V. Local search heuristics for k-median and facility location problems. *SIAM J. Comput. 33*, 3 (2004), 544–562.

[5] Awasthi, P., Balcan, M. F., Blum, A., Sheffet, O., and Vempala, S. On nash-equilibria of approximation-stable games. In *SAGT* (2010), pp. 78–89.

[6] Awasthi, P., Blum, A., and Sheffet, O. Stability yields a ptas for k-median and k-means clustering. In *FOCS* (2010), pp. 309–318.

[7] Awasthi, P., Blum, A., and Sheffet, O. Center-based clustering under perturbation stability. *Inf. Process. Lett. 112*, 1-2 (2012), 49–54.

[8] Balcan, M. F., Blum, A., and Gupta, A. Approximate clustering without the approximation. In *SODA* (2009).

[9] Balcan, M. F., Blum, A., and Vempala, S. A discriminative framework for clustering via similarity functions. In *STOC* (2008), pp. 671–680.

[10] Balcan, M.-F., and Liang, Y. Clustering under perturbation resilience. In *ICALP (1)* (2012), pp. 63–74.

[11] Bartal, Y., Charikar, M., and Raz, D. Approximating min-sum -clustering in metric spaces. In *STOC* (2001), pp. 11–20.

[12] Ben-David, S. Alternative measures of computational complexity with applications to agnostic learning. In *TAMC* (2006), pp. 231–235.

[13] Ben-David, S., von Luxburg, U., and Pál, D. A sober look at clustering stability. In *COLT* (2006), pp. 5–19.

[14] Bilu, Y., and Linial, N. Are stable instances easy? In *Innovations in Computer Science* (2010), pp. 332 – 341.

[15] Charikar, M., Guha, S., Tardos, É., and Shmoys, D. B. A constant-factor approximation algorithm for the k-median problem. *J. Comput. Syst. Sci. 65*, 1 (2002), 129–149.

[16] Charikar, M., O'Callaghan, L., and Panigrahy, R. Better streaming algorithms for clustering problems. In *STOC* (2003), pp. 30–39.

[17] de la Vega, W. F., Karpinski, M., Kenyon, C., and Rabani, Y. Approximation schemes for clustering problems. In *STOC* (2003), pp. 50–58.

[18] Garey, M. R., and Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, 1979.

[19] Guha, S., and Khuller, S. Greedy strikes back: Improved facility location algorithms. *J. Algorithms 31*, 1 (1999), 228–248.

[20] Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callaghan, L. Clustering data streams: Theory and practice. *IEEE Trans. Knowl. Data Eng. 15*, 3 (2003), 515–528.

[21] Jain, K., Mahdian, M., and Saberi, A. A new greedy approach for facility location problems. In *STOC* (2002), ACM, pp. 731–740.

[22] Kumar, A., Sabharwal, Y., and Sen, S. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM 57*, 2 (2010).

[23] Lipton, R. J., Markakis, E., and Mehta, A. On stability properties of economic solution concepts. Manuscript, 2006.

[24] Lloyd, S. Least squares quantization in pcm. *Information Theory, IEEE Transactions on 28*, 2 (Mar. 1982), 129–137.

[25] Mihalák, M., Schöngens, M., Srámek, R., and Widmayer, P. On the complexity of the metric tsp under stability considerations. In *SOFSEM* (2011), pp. 382–393.

[26] Muthukrishnan, S. Data streams: algorithms and applications. In *SODA* (2003), pp. 413–413.

[27] Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. The effectiveness of lloyd-type methods for the k-means problem. In *FOCS* (2006), pp. 165–176.

[28] Schalekamp, F., Yu, M., and van Zuylen, A. Clustering with or without the approximation. In *COCOON* (2010), pp. 70–79.

[29] van Rooij, J. M. M., van Kooten Niekerk, M. E., and Bodlaender, H. L. Partition into triangles on bounded degree graphs. In *SOFSEM* (2011).

Appendix A. Dominating set promise problem

A **dominating set** in a unweighted graph $G = (V, E)$ is a subset $D \subseteq V$ of vertices such that each vertex in $V \setminus D$ has a neighbor in D. A dominating set is **perfect** if each vertex in $D \setminus V$ has exactly one neighbor in D. The problems of finding the smallest dominating set and smallest perfect dominating set are NP-hard.

We introduce a related problem, called the **perfect dominating set promise problem**. In this problem we are promised that the input graph is such that all its dominating sets of size less at most d are perfect, and we are asked to find a set of cardinality at most d.

First, we prove the following.

Theorem 21. *The **perfect dominating set promise problem** (PDS-PP) is NP-hard.*

Proof. The **3d matching problem** (3DM) is as follows: let X, Y, Z be finite disjoint sets with $m = |X| = |Y| = |Z|$. Let T contain triples (x, y, z) with $x \in X, y \in Y, z \in Z$ with $L = |T|$. $M \subseteq T$ is a perfect $3d$-matching if for any two triples $(x_1, y_1, z_1), (x_2, y_2, z_2) \in M$, we have $x_1 \neq x_2, y_1 \neq y_2, z_1 \neq z_2$. We notice that M is a disjoint partition. Determining whether a perfect $3d$-matching exists (YES vs. NO instance) in a $3d$-matching instance is known to be NP-complete.

Now we reduce an instance of the 3DM problem to PDS-PP on $G = (V, E)$. For 3DM elements X, Y, and Z we construct vertices V_X, V_Y, and V_Z, respectively. For each triple in T we construct a vertex in set V_T. Additionally, we make an extra vertex v. This gives $V = V_X \cup V_Y \cup V_Z \cup V_T \cup \{v\}$. We make the edge set E as follows. Every vertex in V_T (which corresponds to a triple) has an edge to the vertices that it contains in the corresponding 3DM instance (one in each of V_X, V_Y, and V_Z). Every vertex in V_T also has an edge to v.

Now we will examine the structure of the smallest dominating set D in the constructed PDS-PP instance. The vertex v must belong to D so that all vertices in V_T are covered. Then, what remains is to optimally cover the vertices in $V_X \cup V_Y \cup V_Z$ – the cheapest solution is to use m vertices from V_T , and this is precisely the 3DM problem, which is NP-hard. Hence, any solution of size $d = m + 1$ for the PDS-PP instance gives a solution to the $3DM$ instance.

We also observe that such a solution makes a perfect dominating set. Each vertex in $V_T \setminus D$ has one neighbor in D, namely v. Each vertex in $V_X \cup V_Y \cup V_Z$ has a unique neighbor in D, namely the vertex in V_T corresponding to its respective set in the 3DM instance. □

Appendix B. Average linkage for min-sum stability

Here, we further support the claim that algorithms designed for α-perturbation resilient instances with respect to the min-sum objective can often be made to work for data satisfying the more general α-min-sum stability property.

Algorithm 2 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 only one cluster remains in $\mathcal{C}'$: merge clusters C, C' in $\mathcal{C}'$ which minimize $d_{avg}(C, C')$.
- Let T be the tree with components as leaves and internal nodes corresponding to the merges performed.

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

Output: Output $\tilde{\mathcal{C}}$.

One such algorithm is Algorithm Appendix B, the Average Linkage algorithm appearing in [10]. The algorithm requires the condition in Lemma 22 to hold, which we can prove indeed holds for α-min-sum stable instances (their proof of the lemma holds for the more restricted class of perturbation-resilient instances). To state the lemma, we first define the distance between two point sets, A and B:

$$d(A, B) \doteq \sum_{p\in A}\sum_{q\in B} d(p, q).$$

Lemma 22. *Assume the optimal clustering is α-min-sum stable. For any two different clusters C and C' in $\mathcal{C}$ and every $A \subset C$, $\alpha d(A, \bar{A}) < d(A, C')$.*

Proof. From the definition of $\alpha d(A, \bar{A})$, we have

$$\begin{aligned}
\alpha d(A, \bar{A}) &= \alpha \sum_{p\in A}\sum_{q\in \bar{A}} d(p, q) \\
&\leq \alpha \sum_{p\in A}\sum_{q\in C} d(p, q) \\
&= \sum_{p\in A} \alpha \sum_{q\in C} d(p, q) \\
&< \sum_{p\in A}\sum_{q\in C'} d(p, q) \\
&= d(A, C').
\end{aligned}$$

The first inequality comes from $\bar{A} \subset C$ and the second by definition of min-sum stability. $\square$

This, in addition to Lemma 6, can be used to show their algorithm can be employed for min-sum stable instances.