The Computational Complexity of Almost Stable Clustering with Penalties

Farnam Mansouri[1] , **Sandra Zilles**[1,2,3] and **Kamyar Khodamoradi**[2]
[1]Department of Computer Science, University of Waterloo
[2]Department of Computer Science, University of Regina
[3]Alberta Machine Intelligence Institute (Amii)
frnm.mansouri@gmail.com, {sandra.zilles, kamyar.khodamoradi}@uregina.ca

Abstract

We investigate the complexity of stable (or perturbation-resilient) instances of k-MEANS and k-MEDIAN clustering problems in metrics with small doubling dimension. While these problems have been extensively studied under multiplicative perturbation resilience in low-dimensional Euclidean spaces, we adopt a more general notion of stability, termed "almost stable", known from the literature as (α, ε)-perturbation resilience. Additionally, we extend our results to k-MEANS/k-MEDIAN with penalties, where each data point is either assigned to a cluster centre or incurs a penalty.

We show that certain special cases of almost stable k-MEANS/k-MEDIAN (with penalties) are solvable in polynomial time. To complement this, we also examine the hardness of almost stable instances and $(1 + \frac{1}{poly(n)})$-stable instances of k-MEANS/k-MEDIAN (with penalties), proving super-polynomial lower bounds on the runtime of any exact algorithm under the widely believed Exponential Time Hypothesis (ETH).

1 Introduction

A fundamental challenge in statistical data analysis is to organize an unlabelled set of data points into groups of similar objects. Clustering, a prominent technique of unsupervised learning, is widely used across various scientific disciplines to address this challenge. The computational complexity of various clustering objectives has been extensively studied in the literature. While many well-known objectives, such as k-Means, k-Median, and k-Centre, are NP-hard to optimize [Guha and Khuller, 1999; Vazirani, 2001; Drineas *et al.*, 2004], simple heuristics like local search have long been known to perform effectively on real-world data. This disparity has sparked significant interest in beyond worst-case analysis, aiming to identify structural properties of the input data points that enable heuristic algorithms to produce nearly optimal solutions.

[Bilu and Linial, 2012] first introduced the perturbation resilience condition in the context of max-cut clustering. [Awasthi *et al.*, 2010] studied stables instances of centre-based clustering objectives (including k-MEANS and k-MEDIAN) and proved that 3-stable instances of these problems are solvable in polynomial time. Since then, the challenge has been to determine the smallest stability value for which the problem is in P. [Balcan and Liang, 2016] improved the bound to $1 + \sqrt{2}$, and [Makarychev *et al.*, 2012] showed that with a slightly weaker promise of metric perturbation resilience, 2-stable instances of a class of centre-based clustering objectives can be optimized in polynomial time. [Friggstad *et al.*, 2019a] focused on metrics of bounded doubling dimension (which include $\mathbb{R}^d$ for constant d) and proved that for any constant ε, $(1 + \varepsilon)$-stable instances of k-MEANS and k-MEANS are solvable in polynomial time. They complemented this result by showing that when the dimension d is part if the input, a universal constant $\varepsilon_\circ$ exists such that even a PTAS for $(1 + \varepsilon_\circ)$-stable instances of k-MEANS in $\mathbb{R}^d$ would imply NP $=$ RP.

The definition of stable clustering problems is motivated by the intuition that most real-world instances of such problems behave in a "stable" manner for the right objective [Bilu and Linial, 2012], i.e., the structure of the optimum solution is resilient against small perturbations in input values. Here, the objective value of a solution is used as a proxy for its resilience to perturbation, i.e., for stable instances the solution with the optimum objective value must maintain its optimality under a set of valid perturbations. To address cases when this model of stability is too restrictive, the notion of α-stability can be relaxed to ensure that perturbing the metric keeps the optimum solution "β-close" (for some notion of closeness) to the optimum solution of the original metric, for some $\beta \geq 0$. [Ben-David *et al.*, 2006] viewed stability as a property of the clustering algorithm itself, and allowed for various optimum solutions that are structurally similar under Hamming distance. [Balcan and Liang, 2016] introduced (α, β)-stable instances of centre-based clustering (see Section 2 for details) and provided an approximation scheme for (α, β)-stable instances of k-MEDIAN when $\alpha > 2 + \sqrt{3}$ and β is a function of the minimum size of a cluster in the optimal solution.

1.1 Our Contribution

We study the perturbation resilience of k-MEANS and k-MEDIAN in metrics of bounded doubling dimension. Our ultimate goal is to give tight bounds on the (α, β)-stability parameter of efficiently solvable problem instances. While we do not fully close the gap, we make significant progress towards this goal by establishing both upper and lower bounds, thereby providing a more refined picture of the complexity of the problem. Our results also extend to the penalty version of k-MEANS and k-MEDIAN, in which points may not be assigned to a cluster, thus incurring a penalty.

We first consider a special case of (α, β)-stability with $\beta = 0$. This notion of stability, which we call strongly α-stable (see Section 2) still generalizes the definition used in recent works, including the one used in [Friggstad *et al.*, 2019a]. We show that such instances of the k-MEANS and k-MEDIAN (with penalties) are solvable over doubling metrics:

Theorem 1 (Exact Solution for Stable k-MEANS)**.** *For any given $(1+\epsilon')$-stable instance of k-MEANS in doubling metrics, Algorithm 1 finds an optimal solution in polynomial time.*

Theorem 2 (Exact Solution for Stable Penalty k-MEANS)**.** *Fix any $d \geq 1$ and $\varepsilon' > 0$. $(1+\varepsilon')$-stable instances of the k-MEANS problem in metric spaces with doubling dimension d can be solved in polynomial time.*

A PTAS for the Stable k-MEANS problem in doubling dimensions is known in the literature [Friggstad *et al.*, 2019b]. But as noted by [Friggstad *et al.*, 2019a], this result does not imply that the Stable k-MEANS is polynomial-time solvable in doubling dimensions. It is natural to ask if choosing a suitably small error parameter ε for the local search and then using the stability of the optimum solutions together lead to an exact solution. However, this approach may lead to values of ε that are inverse polynomials in n, leading to exponential running times as the local search considers all swaps of size roughly $n^{\mathcal{O}(1/\varepsilon)}$.

Next, we focus on the hardness of (α, β)-stable instances (which we call "almost stable" instances). Assuming ETH, we prove that an optimal solution cannot be obtained, even if we allow the algorithm a super-polynomial runtime, *and* restrict the instance to a bounded-dimensional Euclidean space.

Theorem 3 (Hardness of Almost Stable Penalty k-MEDIAN)**.** *For any $\alpha \in (1, 1.2)$, $\beta > 0$, there is no $f(k)n^{o(\sqrt{k})}$-time algorithm for (α, β)-stable Euclidean k-MEDIAN with penalties in $\mathbb{R}^2$, unless ETH fails.*

Theorem 4 (Hardness of Almost Stable k-MEDIAN)**.** *For any $\alpha \in (1, 1.2)$, $\beta > 0$, there is no $f(k)n^{o(\sqrt{k})}$-time algorithm for (α, β)-stable k-MEDIAN in metrics of bounded doubling constant, unless ETH fails.*

Finally, to complement our algorithmic upper bounds for α-stable instances of k-MEANS/k-MEDIAN (with penalties), we ask the question: *how close to one can α get to ensure the instance is still efficiently solvable?* While we do not obtain APX-hardness, we manage to show that for inverse polynomial values of ε, the problem is not likely to have an FPT solution.

Theorem 5 (Hardness of $(1 + \frac{1}{poly(n)})$-Stable Penalty k-MEDIAN)**.** *There is no $f(k)n^{o(k)}$-time algorithm for $(1 + \mathcal{O}\left(\frac{1}{n^{12}}\right))$-stable Euclidean k-MEDIAN with penalties in $\mathbb{R}^4$, unless ETH fails.*

Theorem 6 (Hardness of $(1 + \frac{1}{poly(n)})$-Stable k-MEDIAN)**.** *There is no $f(k)n^{o(k)}$-time algorithm for $(1+\mathcal{O}\left(\frac{1}{n^{16}}\right))$-stable Euclidean k-MEDIAN in $\mathbb{R}^6$, unless ETH fails.*

Due to page limits, many formal proofs are omitted; the reader is referred to the full version of the paper for more details [Khodamoradi *et al.*, 2025].

2 Preliminaries

Consider a doubling metric space (V, δ), where V is a set of points and δ a distance function over V with doubling dimension d. We focus on stable instances of discrete clustering problems k-MEANS and k-MEDIAN, with and without penalties, over V with metric δ. Let $X \subseteq V$ be a finite set of data points, $C \subseteq V$ to be a finite set of candidate centres, $p : X \to \mathbb{R}$ a penalty function, $\delta : V \times V \to \mathbb{R}$ be a metric, and k be an integer. We denote a k-MEANS/k-MEDIAN instance as a triple (X, C, δ), and a k-MEANS/k-MEDIAN with penalties instance as a four-tuple (X, C, p, δ). A feasible solution to the instance is a set $S \subseteq C$ with $|S| = k$. For k-MEANS(k-MEDIAN) the cost of solution is defined as $\text{cost}(S) := \sum_{j \in X} \delta(j, S)^2$ $(\text{cost}(S) := \sum_{j \in X} \delta(j, S))$, and for k-MEANS(k-MEDIAN) with penalties the cost of solution is defined as $\text{cost}_{\text{pen}}(S) := \sum_{j \in X} \min\left(\delta(j, S)^2, p(j)\right)$ $(\text{cost}_{\text{pen}}(S) := \sum_{j \in X} \min\left(\delta(j, S), p(j)\right))$ where $\delta(j, S)$ denotes $\min_{i \in S} \delta(j, i)$. Throughout the paper, we refer to $n := |C| + |X|$ as the size of the input, since the computational complexity of the algorithms is primarily determined by how the algorithm scales with the number of data points.

Next, we define α-stable clustering. Intuitively, α-stability for k-MEANS/k-MEDIAN ensures that the optimal solution for a perturbed metric remains optimal for the original metric. We follow the concept of multiplicative stability, first introduced by [Awasthi *et al.*, 2010] and [Bilu and Linial, 2012], but we present a more general version here. Let $\alpha \geq 1$. We call an instance $\mathcal{I} = (X, C, \delta)$ of k-MEANS/k-MEDIAN *α-stable* if for any $\delta \leq \delta' \leq \alpha\delta$ every optimum solution of the k-MEANS/k-MEDIAN (with penalties) instance $\mathcal{I}' = (X, C, \delta')$ is also an optimum solution for $\mathcal{I}$. Similarly, an instance $\mathcal{I} = (X, C, \delta, p)$ of k-MEANS/k-MEDIAN with penalties is called *α-stable* if for any $\delta \leq \delta' \leq \alpha\delta$ and any p' with $p \leq p' \leq \alpha^2 \cdot p$ for k-MEANS and $p \leq p' \leq \alpha \cdot p$ for k-MEDIAN, every optimum solution of the k-MEANS/k-MEDIAN with penalties instance $\mathcal{I}' = (X, C, \delta', p')$ is also an optimum solution for $\mathcal{I}$.

Let $\text{cost}(S)$ and $\text{cost}'(S)$ denote the cost of S before and after perturbation. We also study (α, β)-stable instances of k-MEANS and k-MEDIAN, as defined by [Balcan and Liang, 2016]. Recall that in this generalization, β is a measure of proximity between feasible solutions. Various distance metrics can be considered to define β-closeness. We adopt the notion $\delta_{\mathcal{I}}^{bij} : \binom{C}{k} \times \binom{C}{k} \to \mathbb{R}^{\geq 0}$, defined as follows. However, all our results easily extend to other natural notions of distance. For any $S_1, S_2 \subseteq C$ with $|S_1| = |S_2| = k$, define:

$$\delta_{\mathcal{I}}^{bij}(S_1, S_2) := \min\Big\{ \sum\nolimits_{s_1 \in S_1} \delta(s_1, f(s_1)) \mid f \text{ is a bijection between } S_1 \text{ and } S_2 \Big\}. \tag{1}$$

We also generalize this notion to clustering with penalties. Let $\alpha \geq 1$ and $\beta \geq 0$. We call an instance $\mathcal{I} = (X, C, \delta)$ ($\mathcal{I} = (X, C, \delta, p)$) of k-MEANS/k-MEDIAN with penalties) (α, β)-stable if for any optimum solution $O \subseteq C$ of $\mathcal{I}$, and any optimum solution $O' \subseteq C$ of the k-MEANS/k-MEDIAN instance (with penalties) $\mathcal{I}' = (X, C, \delta')$ ($\mathcal{I}' = (X, C, \delta', p')$), where δ' is any symmetric function such that $\delta \leq \delta' \leq \alpha\delta$ (and p' is any function such that $p \leq p' \leq \alpha^2 \cdot p$ for k-MEANS, and $p \leq p' \leq \alpha \cdot p$ for k-MEDIAN), we have $\delta^{bij}_{\mathcal{I}}(O, O') \leq \beta$.

Observe that $\beta = 0$ yields the definition of strongly α-stable instances. If $\beta > 0$, we call the instance of k-MEANS/k-MEDIAN an *almost stable* instance. Below, we study the performance of a local search algorithm for cases where α is close to 1, that is $\alpha = 1 + \varepsilon'$. To avoid confusion, we use ε' to refer to the stability parameter, and ε to refer to a parameter of our local search analysis. It should be mentioned that ε only affects the runtime of the local search and is a function of ε'.

3 Polynomial Algorithms for Stable Clusterings

In this section, we generalize the exact polynomial-time algorithm of [Friggstad *et al.*, 2019a] by incorporating penalties. Most of the results in this section are an adaptation of those in [Friggstad *et al.*, 2019a]. We first show how to obtain optima for k-MEANS without penalties in *doubling metrics* (see Theorem 1). Doubling metrics are those metrics for which the doubling dimension is a constant. A metric is said to have a doubling dimension of d if every ball of radius R in that metric can be covered with at most 2^d balls of radius $R/2$.

Next, we define a polynomial-time algorithm for k-MEANS with penalties (see Theorem 2). Let (X, C, p, δ) be an instance of k-MEANS with penalties, and $\mathcal{F}_k = \{S \subseteq C : |S| = k\}$ the set of feasible solutions. Throughout this section, we use Algorithm 1, a generalization of the local search method introduced by [Friggstad *et al.*, 2019a] to k-MEANS with penalties. Their original algorithm is the standard ρ-swap local search algorithm, modified to perform the swap that provides the best improvement at each step. We set $\rho = d^{O(d)} \cdot \varepsilon^{-O(d/\varepsilon)}$.

Algorithm 1 ρ-Swap Local Search

Input: Let $\mathcal{F}_k$ be the family of feasible sets. Let $S \in \mathcal{F}_k$.
Output: The local optimal set S

```
while there exists a set S' ∈ F_k such that |S − S'| ≤ ρ
   and cost(S') < cost(S) (cost_pen(S') < cost_pen(S)) do
    S ← arg min_{S'∈F_k, |S−S'|≤ρ} cost(S')
      (arg min_{S'∈F_k, |S−S'|≤ρ} cost_pen(S') )
end while
return S
```

To expand Friggstad et al.'s terminology to the case with penalties, for any $S, O \in \mathcal{F}_k$, define:

- For $j \in X$, let $\sigma(j, S)$ be the centre in S nearest to j, breaking ties by a fixed ordering of C.
- $\overline{X}_{S,O} = \{j \in X : \sigma(j, S) \in S - O$ and $\sigma(j, O) \in O - S\}$.
- $\overline{X}^{pen}_{S,O} = \{j \in X : \sigma(j, S) \in S - O$ and $\sigma(j, O) \in O - S$ and $\delta(j, S)^2, \delta(j, O)^2 < p(j)\}$.
- $\Psi(S, O) = \sum_{j \in \overline{X}_{S,O}} \delta(j, \sigma(j, S))^2 + \delta(j, \sigma(j, O))^2$, $\Psi_{pen}(S, O) = \sum_{j \in \overline{X}^{pen}_{S,O}} \delta(j, \sigma(j, S))^2 + \delta(j, \sigma(j, O))^2$.

[Friggstad *et al.*, 2019a] first showed that Algorithm 1 visits a "nearly-good" solution in a polynomial number of rounds and returns such a solution upon termination. Next, they used the properties of stable instances to show that any nearly-good solution must be the unique optimal solution, hence Algorithm 1 is an exact polynomial-time algorithm to solve the problem. Adapting their proof to our generalized definition of stability is technically non-trivial. To this end, we begin by modifying the notion of nearly-good solutions.

Definition 1. *We call $S \in \mathcal{F}_k$ a* nearly-good *solution if for every optimum solution $F \in \mathcal{F}_k$ we have $\mathrm{cost}(S) \leq \mathrm{cost}(F) + 2\varepsilon \cdot \Psi(S, F)$ for k-MEANS and $\mathrm{cost}_{\mathrm{pen}}(S) \leq \mathrm{cost}_{\mathrm{pen}}(O) + 2\varepsilon \cdot \Psi_{pen}(S, F)$ for k-MEANS with penalties.*

Adapting a technique introduced by [Friggstad *et al.*, 2019b], we can show the existence of a desirable partitioning of the set of centres via a probabilistic argument. Fix any $S', O', T \subseteq C$ and $X' \subseteq X$ such that S' and O' are disjoint, $T \subseteq O' \cup S'$, and $|T| = |O'| = |S'|$. Define Δ_j^T for each $j \in X'$ to be $\delta\left(j, \sigma\left(j, S' \triangle T\right)\right)^2 - \delta\left(j, \sigma\left(j, S'\right)\right)^2$. Δ_j^T denotes the change in j's assignment cost when solution S' is replaced by $S' \Delta T$. One obtains the following.

Theorem 7 (Theorem 5 in [Friggstad *et al.*, 2019a])**.** *There is a randomized algorithm that samples a partition π of $S' \cup O'$ such that $|T \cap S'| = |T \cap O'| \leq \rho$ for each $T \in \pi$ and*

$$\mathbf{E}_\pi \left[\sum_{T \in \pi} \sum_{j \in X'} \Delta_j^T\right] \leq \sum_{j \in X'} (1 + \varepsilon) \cdot \delta\left(j, O'\right)^2 - (1 - \varepsilon) \cdot \delta\left(j, S'\right)^2 .$$

Next, we observe that if a feasible solution S is not nearly-good according to our more strict definition, it still holds that the local search algorithm must take a large step on S:

Theorem 8 (Theorem 6 in [Friggstad *et al.*, 2019a]))**.** *For any $S, O \in \mathcal{F}_k$, if $\mathrm{cost}(S) > \mathrm{cost}(O) + \varepsilon \cdot \Psi(S, O)$, then there exists an $S' \in \mathcal{F}_k$ with $|S - S'| \leq \rho$ and $\mathrm{cost}\left(S'\right) \leq \mathrm{cost}(S) + \left(\mathrm{cost}(O) - \mathrm{cost}(S) + \varepsilon \cdot \Psi(S, O)\right)/k$.*

With the new definition of nearly-good solutions not being tied to a unique global optimum, we need a more intricate argument to show that local search always makes significant progress towards a nearly-good solution. To that end, we specify, for each set S produced by the algorithm at an iteration, the global optimum which is a witness to S not being nearly-good and show the existence of such a set S is a contradiction.

Lemma 1. *Given any instance of the k-MEANS problem, Algorithm 1 with $\rho = d^{O(d)} \cdot \varepsilon^{-O(d/\varepsilon)}$ terminates at a nearly-good solution. Also, within $2k \cdot \ln(n\Delta)$ iterations, the algorithm produces a nearly-good solution, where $\Delta = \max_{j \in X, i \in C} \delta^2(i, j)$.*

3.1 Stable k-MEANS without Penalties

Lemma 1 alone does not guarantee that Algorithm 1 finds an optimal solution for the $(1 + \varepsilon')$-stable instances of k-MEANS.

We use the stability condition to show that a nearly-good solution must be one of the optimal ones for the instance. Combined with the fact that the returned solution of Algorithm 1 is nearly-good we get the desired result:

Theorem 1 (Exact Solution for Stable k-MEANS)**.** *For any given $(1+\epsilon')$-stable instance of k-MEANS in doubling metrics, Algorithm 1 finds an optimal solution in polynomial time.*

Proof. We need to modify Lemma 2 in [Friggstad *et al.*, 2019a]. We first perturb the distances as follows:

$$\delta'(i,j) = \begin{cases} (1+\varepsilon')\cdot\delta(i,j)\ , & \text{if } i \neq \sigma(j,S)\ , \\ \delta(i,j)\ , & \text{otherwise}\,. \end{cases} \tag{2}$$

We present a modified version of Lemma 2:

Lemma 2 (Lemma 2 of [Friggstad *et al.*, 2019a], Restatement)**.** *If for feasible solutions $S, O \in \mathcal{F}_k$, we have that if $cost(S) \le cost(O) + 2\varepsilon\Phi(S,O)$, then $cost'(S) \le cost'(O)$.*

In this lemma, we let S be the nearly-good solution of Algorithm 1. Let O be any optimum solution of $\mathcal{I}'$. Using stability condition O should also be an optimum solution for $\mathcal{I}$. We observe that $\mathrm{cost}'(S) = \mathrm{cost}(S) \le \mathrm{cost}'(O)$, where cost' indicates the cost of a solution under δ'. Therefore S is a optimum solution for $\mathcal{I}'$. Using the stability condition one more time, we conclude that S must also be an optimum solution of $\mathcal{I}$ and this $S \in \mathcal{O} = \{O_1, \ldots, O_\ell\}$. Assume that Algorithm 1 encounters S in iteration $t \le \lfloor 2k \cdot \ln(\Delta n) \rfloor$. Since S has the lowest possible cost for the instance $\mathcal{I}$, the local search must terminate at iteration t and return S. $\square$

3.2 Stable k-MEANS with Penalties

We now provide polynomial-time exact ($(1+\mathcal{O}(\varepsilon'))$-approximate) solutions for stable instances of k-MEANS with penalties in doubling metrics. All results in this section extend to k-MEDIAN with minor modification. We first show how to solve $(1+\varepsilon')$-stable instances of k-MEANS exactly in polynomial time when the metric has a bounded doubling dimension. We first focus on $(1+\varepsilon')$-stable instances for which any optimal solution will remain the optimum for the perturbed instance.

Theorem 2 (Exact Solution for Stable Penalty k-MEANS)**.** *Fix any $d \ge 1$ and $\varepsilon' > 0$. $(1+\varepsilon')$-stable instances of the k-MEANS problem in metric spaces with doubling dimension d can be solved in polynomial time.*

The local search in [Friggstad *et al.*, 2019a] is oblivious to penalties and cannot guarantee nearly-good solutions under this new setting. Our core idea for addressing this issue is to capture the penalties for the points of X by adding a dummy centre z^* to V (specifically, to C) and defining $C' = C \cup \{z^*\}$. We extend δ to $V \cup \{z^*\}$ by setting $\delta(z^*, j)^2 = \delta(j, z^*)^2 = p(j)$ for every $j \in X$, and $\delta(z^*, c) = 0$ for all $c \in C'$. Here z^* is introduced solely to provide the intuition that analyzing the local search algorithm for k-MEANS with penalties is equivalent to analyzing a restricted local search over the reduced k-Means problem where z^* cannot be swapped. Also note that δ no longer satisfies the triangle inequality, making the analysis more challenging. For any solution S of the $k+1$-MEANS instance (X, C', δ) that contains z^*, the cost of S equals the cost of $S \backslash \{z^*\}$ for the original k-MEANS with penalties instance (X, C, p, δ). Thus, the task of finding the optimum solution of the (X, C, p, δ) instance is reduced to finding the optimum solution that contains z^* for the $k+1$-MEANS instance (X, C', δ). Our key insight here is to modify the ρ-swap algorithm to ensure it always makes the cheapest swap that does not evict z^*. This eliminates the need to invoke the triangle inequality on z^*.

We first show that our modified local search converges to a nearly-good solution in polynomial time. Then, we prove that the stability condition ensures that any such nearly-good solution is, in fact, the optimal one. While the structure of our proof is similar to Friggstad et al.'s, we need to tackle challenges in each step. We will highlight how we address these challenges in what follows.

Before rigorously proving Theorem 2, we prove that a nearly-good solution can be found in polynomial time for k-MEANS with penalties as well. To do so, it suffices to replicate Theorem 8 for the penalty setting. Consider the original instance (X, C, p, δ) and fix any $S, O \in \mathcal{F}_k$. Define $S' = S \backslash O$ and $O' = O \backslash S$, and fix some $T \subseteq O' \cup S'$ such that $|T| = |O'| = |S'|$. Moreover, for any $j \in \overline{X}^{pen}_{S,O}$ define $\tilde{\Delta}^T_j$ to be $\min(\delta\left(j, \sigma\left(j, S' \triangle T\right)\right)^2, p(j)) - \delta\left(j, \sigma\left(S'\right)\right)^2$. Using Theorem 7 with $X' = \overline{X}^{pen}_{S,O}$, there exists a randomized algorithm that samples a partition π of $S' \cup O'$ with $|T \cap S'| = |T \cap O'| \le \rho$ for each $T \in \pi$ such that

$$\begin{aligned} &\mathbf{E}_\pi\left[\sum\nolimits_{T\in\pi}\sum\nolimits_{j\in\overline{X}^{pen}_{S,O}}\tilde{\Delta}^T_j\right] \\ &\le \mathbf{E}_\pi\left[\sum_{T\in\pi}\sum_{j\in\overline{X}^{pen}_{S,O}}\delta\left(j,\sigma\left(j,S'\triangle T\right)\right)^2 - \delta\left(j,\sigma\left(j,S'\right)\right)^2\right] \\ &\le \sum\nolimits_{j\in\overline{X}^{pen}_{S,O}}(1+\varepsilon)\cdot\delta\left(j,O'\right)^2 - (1-\varepsilon)\cdot\delta\left(j,S'\right)^2 \end{aligned} \tag{3}$$

The following result extends Theorem 6 of [Friggstad *et al.*, 2019a] to handle penalties, via a more involved case analysis.

Theorem 9. *For any $S, O \in \mathcal{F}_k$, if $\mathrm{cost}_{\mathrm{pen}}(S) > \mathrm{cost}_{\mathrm{pen}}(O) + \varepsilon \cdot \Psi_{pen}(S,O)$, then there exists an $S' \in \mathcal{F}_k$ with $|S - S'| \le \rho$ such that*

$$\mathrm{cost}_{\mathrm{pen}}\left(S'\right) \le \mathrm{cost}_{\mathrm{pen}}(S) + \frac{\mathrm{cost}_{\mathrm{pen}}(O) - \mathrm{cost}_{\mathrm{pen}}(S) + \varepsilon\cdot\Psi_{pen}(S,O)}{k}$$

Similarly to deriving Lemma 1, we prove the following.

Lemma 3. *For k-MEANS with penalties, when Algorithm 1 terminates, the returned solution is nearly-good. Also, for at least one of the first $2k \cdot \ln(n\Delta)$ iterations of Algorithm 1, S is a nearly-good solution, where $\Delta = \max_{j\in X, i\in C}\delta^2(i,j)$.*

Proof of Theorem 2 Let ε satisfy $1 + 6\varepsilon = (1+\varepsilon')^2$, i.e., $\varepsilon \approx \varepsilon'/3$ for small ε'. Moreover, let S be the nearly-good solution that Algorithm 1 encounters within $2k \cdot \ln(n\Delta)$ iterations (by Lemma 3). Define a perturbed distance function δ' as in (2) and a perturbed penalty function via

$$p'(j) = \begin{cases} \left(1+\varepsilon'\right)^2 p(j) & ;\text{if } \delta(j,S) < p(j)\,, \\ p(j) & ;\text{otherwise} \end{cases} \tag{4}$$

Moreover, for any $S' \in \mathcal{F}_k$, let $\mathrm{cost}'_{\mathrm{pen}}(S') = \sum_{j \in X} \min(\min_{i \in S'} \delta'(i,j)^2, p'(j))$ be the cost of S' under δ'. Clearly, $\delta \leq \delta' \leq (1+\varepsilon') \cdot \delta$ and $p \leq p' \leq (1+\varepsilon')^2 \cdot p$. Let O_q be any optimum solution w.r.t. $\mathrm{cost}'_{\mathrm{pen}}$, due to stability condition O should be an optimum solution w.r.t. $\mathrm{cost}_{\mathrm{pen}}$. Therefore, S is nearly good for O_q. Partition X as follows:

- $X^1 = \{j \in X : \sigma(j,S) \in S - O_q$ and $\sigma(j,O_q) \in S \cap O_q$ and $\delta(j,S)^2 < p(j)\}$
- $X^2 = \{j \in X : \sigma(j,S) \in S \cap O_q$ and $\sigma(j,O_q) \in O_q - S$ and $\delta(j,O_q)^2 < p(j)\}$
- $X^3 = \{j \in X : \sigma(j,S), \sigma(j,O_q) \in S \cap O_q$ or $\delta(j,S)^2, \delta(j,O_q)^2 \geq p(j)\}$
- $X^4 = \{j \in X : \sigma(j,S) \in S - O_q$ and $\sigma(j,O_q) \in O_q - S$ and $\delta(j,S)^2, \delta(j,O_q)^2 < p(j)\}$.

Observe that $X^4 = \overline{X}^{pen}_{S,O_q}$. As in the proof of Theorem 9, let $c^*_j = \min\{\delta(j,\sigma(j,O_q))^2, p(j)\}$ be the cost incurred by point j in the optimum solution, and analogously $c_j = \min\{\delta(j,\sigma(j,S))^2, p(j)\}$ for S. To calculate $\mathrm{cost}'_{\mathrm{pen}}(O_q)$, we consider each $j \in X$ case by case. (i) If $j \in X^1 \cup X^4$, then $p'(j) = (1+\varepsilon')^2 p(j)$ and $\delta'(i,j) = (1+\varepsilon')\delta(i,j)$ for all $i \in O_q$. Thus, $\min\{\min_{i \in O_q}\{\delta'(i,j)^2\}, p'(j)\} = (1+\varepsilon')^2 c^*_j$. (ii) For $j \in X^2$, only $i = \sigma(j,S) \in O_q$ does not have its $\delta'(i,j)^2$ multiplied by $(1+\varepsilon')$. Thus, if $\delta^2(j,S) \leq p(j)$ we have $\min\{\min_{i \in O_q}\{\delta'(i,j)^2\}, p'(j)\} = \min\{(1+\varepsilon')^2 c^*_j, \delta(j,S)\}$. Otherwise, $\min\{\min_{i \in O_q}\{\delta'(i,j)^2\}, p'(j)\} = \min\{(1+\varepsilon')^2 c^*_j, p(j)\}$. Hence, $\min\{\min_{i \in O_q} \delta'(i,j)^2, p'(j)\} = \min\{(1+\varepsilon')^2 c^*_j, c_j\}$. (iii) For $j \in X^3$ clearly $\min\{\min_{i \in O_q} \delta'(i,j)^2, p'(j)\} = c_j = c^*_j$. Thus we derive

$$\mathrm{cost}'_{\mathrm{pen}}(O_q) = \sum_{j \in X^1} (1+\varepsilon')^2 c^*_j + \sum_{j \in X^3} c^*_j + \sum_{j \in X^2} \min\left\{(1+\varepsilon')^2 c^*_j, c_j\right\} + \sum_{j \in X^4} (1+\varepsilon')^2 c^*_j. \quad (5)$$

Finally, we require Lemma 4.

Lemma 4. *If S is nearly-good for O_q, then* $\sum_{j \in X^4} c_j \leq (1+6\varepsilon) \cdot \left(\sum_{j \in X^1} c^*_j + \sum_{j \in X^4} c^*_j\right)$.

By Lemma 4, we bound $\mathrm{cost}_{\mathrm{pen}}(S)$ in the following way:

$$\begin{aligned}
&\mathrm{cost}_{\mathrm{pen}}(S) \leq \mathrm{cost}_{\mathrm{pen}}(O_q) + 2\varepsilon \cdot \Psi_{pen}(S,O_q) \\
&= \mathrm{cost}_{\mathrm{pen}}(O_q) + 2\varepsilon \sum_{j \in X^4} c^*_j + 2\varepsilon \sum_{j \in X^4} c_j \\
&\leq \mathrm{cost}_{\mathrm{pen}}(O_q) + 2\varepsilon \sum_{j \in X^4} c^*_j + 2\varepsilon(1+6\varepsilon)[\sum_{j \in X^1} c^*_j + \sum_{j \in X^4} c^*_j] \\
&\leq \sum_{j \in X^1} (1+6\varepsilon) c^*_j + \sum_{j \in X^2} c^*_j + \sum_{j \in X^3} c^*_j + \sum_{j \in X^4} (1+6\varepsilon) c^*_j \\
&\leq \sum_{j \in X^1} (1+6\varepsilon) c^*_j + \sum_{j \in X^2} \min\left\{(1+6\varepsilon) c^*_j, c_j\right\} \\
&\quad + \sum_{j \in X^3} c^*_j + \sum_{j \in X^4} (1+6\varepsilon) c^*_j.
\end{aligned} \quad (6)$$

The last bound again uses $c^*_j \leq c_j$ for $j \in X^2$. Recall we chose ε so that $(1+6\varepsilon) = (1+\varepsilon')^2$. Thus, combining Lemma 4, (6) and the simple observation that $\mathrm{cost}'_{\mathrm{pen}}(S) = \mathrm{cost}_{\mathrm{pen}}(S)$, we get that $\mathrm{cost}'_{\mathrm{pen}}(S) = \mathrm{cost}_{\mathrm{pen}}(S) \leq \mathrm{cost}'_{\mathrm{pen}}(O_q)$. Which means that S is an optimum solution for (X, C, p', δ'). This completes the proof.

4 Hardness of (α, β)-stable k-Median

Section 3 discussed steps towards finding optimum and near-optimum solutions of (α,β)-stable k-Means/ k-Median with or without penalties. Next, we complement this with hardness results for (α,β)-stable k-Median, as the question of whether we can obtain the exact optimum solution of (α,β)-stable k-Means/ k-Median with penalties efficiently remains open. In Sections 4.1 and 4.2, we obtain hardness results for (α,β)-stable k-Median, respectively, with and without penalties for any $\alpha \in (1, 1.2)$, $\beta > 0$. While, for ease of presentation, our hardness results are shown for k-Median, every result in this section readily extends to k-Means.

The results of this section are inspired by a reduction of the Grid Tiling Inequality problem to k^2-Median with penalties, which was introduced by [Cohen-Addad *et al.*, 2018]. We here present our adapted version of the reduction from Grid Tiling Inequality to k^2-Median.

Consider any Grid Tiling Inequality problem with integer n and collection S of k^2 non-empty sets. We define a Euclidean k^2-Median instance with penalties similar to the one in Theorem 6.2 of [Cohen-Addad *et al.*, 2018]. Let $\mathcal{I}(S,n) = (X^{grid}, C^{grid}, p^{grid}, \delta)$ denote this instance. Fix $\varepsilon = O(\beta/n^3)$ for any $\beta > 0$. Define a set of data points X^{grid} in the region consisting of a square of side length $2k + \varepsilon(n-1)$, where the lower left corner of the square is on the origin (i.e., $A = \{(x,y) \mid 0 \leq x, y \leq 2k + \varepsilon(n-1)\}$). They are spaced evenly in a grid G, with two consecutive (horizontal or vertical) data points at distance ε from each other; thus there are $\Sigma = (2k/\varepsilon + n)^2$ data points. Each data point $\vec{j}$ has a penalty of $p^{grid}(\vec{j}) = 1$. Thinking of this grid as a discrete approximation of the uniform measure on the square A, we work with the discrete measure μ carried by the data points, where each data point is weighted 1, so that $\iint_A d\mu = \Sigma$.

For each set $S_{i,j}$, we introduce $|S_{i,j}| \leq n^2$ candidate centres, and let $C_{i,j}$ denote the set of such candidate centres (note that there are k^2 such sets), where $C_{i,j} = \{(2i-1, 2j-1) + \varepsilon(u-1, v-1) \mid (u,v) \in S_{i,j}\}$. Note that the candidate centres are also placed on vertices of G, and that if $S_{i,j}$ has all possible pairs so that $S_{i,j} = [n] \times [n]$, then $C_{i,j}$ precisely forms a subgrid of n^2 evenly spaced points in which consecutive points are at distance ε from each other and the lower left point of the subgrid lies at $(2i-1, 2j-1)$. The final set of candidate centres is given by $C^{grid} = \cup_{1 \leq i,j \leq k} C_{i,j}$.

Theorem 10 (Theorem 6.2 of [Cohen-Addad *et al.*, 2018])**.** *There exist a $\nu \geq 0$, such that the* Grid Tiling Inequality *problem with integer n and collection S has a solution if and only if there exists a solution to $\mathcal{I}(S,n)$ with cost at most ν.*

Previously, we mentioned that μ approximates the uniform measure. Let us elaborate on what we mean by this. Define,

$R = [-0.5\varepsilon, 2k + \varepsilon(n-0.5)] \times [-0.5\varepsilon, 2k + \varepsilon(n-0.5)]$ to be a square with bottom left corner $(-0.5\varepsilon, -0.5\varepsilon)$ and top right corner $(2k + \varepsilon(n-0.5), 2k + \varepsilon(n-0.5))$. Denote by u_R the uniform probability measure on R and define $\mu^\star := \Sigma \cdot u_R$. In other words, $\mu^\star$ is a constant measure on R with the property $\mu^\star(R) = \Sigma$ and $\mu^\star(M) = 1$ for any unit square M. Notice that for any measurable set D and measurable function f, $\varepsilon^2 \iint f d\mu$ can be viewed as the two-dimensional Riemann sum for $\iint_D f d\mu^\star$. Therefore, $\iint_D f d\mu$ approximates $\frac{1}{\varepsilon^2} \iint_D f d\mu^\star$. The following Lemma will formalize this approximation for two specific f and D that will be used later.

Lemma 5. *Fix any $\vec{a} \in C$. Define D_r to be the circle with centre $\vec{a}$ and radius $r \leq 1$. Then*

$$(i)\ \frac{1}{\varepsilon^2} \iint_{D_{r-\varepsilon}} d\mu^\star \leq \iint_{D_r} 1 d\mu \leq \frac{1}{\varepsilon^2} \iint_{D_{r+\varepsilon}} d\mu^\star,$$

$$(ii) \iint_{D_r} \delta(\vec{x}, \vec{a}) d\mu \leq \frac{1}{\varepsilon^2} \iint_{D_{r+\varepsilon}} \delta(\vec{x}, \vec{a}) d\mu^\star + \frac{1}{\varepsilon} \iint_{D_{r+\varepsilon}} d\mu^\star,$$

where $\iint_{D_r} 1 d\mu^\star = \pi r^2$ and $\iint_{D_r} \delta(\vec{x}, \vec{a}) d\mu^\star = \frac{2}{3}\pi r^3$.

4.1 Stable Euclidean k-Median with Penalties

We will now prove that for any $1 < \alpha < 1.2$, $0 < \beta$ (α, β)-stable Euclidean k-Median with penalties in $\mathbb{R}^2$ is hard. The proof idea is to show that $\mathcal{I}(S, n)$ is also stable. Since any two centres in any $C_{i,j}$ for $1 \leq i, j \leq k$ are close, for showing $(\alpha, \beta)-$stability it is enough to show that, even if we perturb distances by a constant factor, the optimum solution would be k^2 candidate centres each from a $C_{i,j}$ for $i, j \leq k$. Intuitively, this is not hard to see, as the cost of switching from $\vec{a} \in \mathcal{C}_{i,j}$ to another $\vec{b} \neq \vec{a} \in \mathcal{C}_{i,j}$ is relatively small. Conversely, by adding a centre $\vec{c}$ in a cluster $\mathcal{C}_{i',j'}$ that currently lacks a centre, we can significantly reduce the costs for points located near $\vec{c}$.

Theorem 3 (Hardness of Almost Stable Penalty k-Median)**.** *For any $\alpha \in (1, 1.2)$, $\beta > 0$, there is no $f(k)n^{o(\sqrt{k})}$-time algorithm for (α, β)-stable Euclidean k-Median with penalties in $\mathbb{R}^2$, unless ETH fails.*

Proof. By Theorem 10, it suffices to show that $\mathcal{I}(S, n)$ is also (α, β)-stable. Consider any $\delta \leq \delta' \leq \alpha \cdot \delta$ and $p \leq p' \leq \alpha \cdot p$. Define $\mathrm{cost}'_{\mathrm{pen}}(S) = \sum_{j \in X^{grid}} \min_{i \in S} \{\min \{\delta'(i,j), p'(j)\}\}$ for every $S \subseteq C^{grid}$. First we prove that every optimal solution O of the k^2-Median, $(X^{grid}, C^{grid}, \delta', p')$ cannot have two centres in a $C_{i,j}$ for $1 \leq i, j, \leq k$. Assume that there are $1 \leq i, j, \leq k$ such that there exist $\vec{a}, \vec{b} \in C_{i,j} \cap O$. By the pigeonhole principle, we have i' and j', $1 \leq i', j', \leq k$ such that $C_{i',j'} \cap O = \varnothing$. Let $O' = (O \backslash \{\vec{a}\}) \cup \{\vec{c}\}$ for any $\vec{c} \in C_{i',j'}$. We then argue that $\mathrm{cost}'_{\mathrm{pen}}(O')$ is smaller than $\mathrm{cost}'_{\mathrm{pen}}(O)$, which is a contradiction. Let D be the set of all points in a circle with centre $\vec{c}$ and radius $\frac{1}{\alpha}$. Every $\vec{x} \in D$ satisfies $\delta'(\vec{x}, \vec{c}) \leq \alpha\delta(\vec{x}, \vec{c}) \leq 1$. Also, since every other point $\vec{d} \neq \vec{a}$, $\vec{d} \in O'$, has the property $\delta(\vec{c}, \vec{d}) \geq 2 - (n-1)\varepsilon$, we have $\delta'(\vec{d}, \vec{x}) \geq \delta(\vec{d}, \vec{x}) \geq 2 - \alpha - (n-1)\varepsilon$ which is greater than 1 for large enough n (for n polynomial in $\frac{1}{\alpha - 1}$). Also, $p'(\vec{x}) \geq p^{grid}(\vec{x}) = 1$ which is smaller than $\delta'(\vec{x}, \vec{c})$. Therefore, every point in D will be assigned to $\vec{c}$. Thus, adding $\vec{c}$ to O will decrease $\mathrm{cost}'_{\mathrm{pen}}(O)$ at least by

$$\begin{aligned} \iint_D (1 - \delta'(\vec{x}, \vec{c})) d\mu \ &\geq \iint_D (1 - \alpha\delta(\vec{x}, \vec{c}))\, d\mu \\ &\overset{(i)}{\geq} \frac{1}{\varepsilon^2}(\pi(\frac{1}{\alpha} - \varepsilon)^2 - \pi\alpha(\frac{2}{3} \cdot (\frac{1}{\alpha} + \varepsilon)^3 + \varepsilon(\frac{1}{\alpha} + \varepsilon)^2)), \end{aligned} \tag{7}$$

where (i) is due to Lemma 5. For any $\vec{x} \in A$ to be assigned to $\vec{a}$, $\delta'(\vec{x}, \vec{a})$ should be less than $p'(\vec{x}) \leq \alpha$. Subsequently, all such $\vec{x}$ should have $\delta(\vec{x}, \vec{a}) \leq \alpha$, i.e., they should belong to a circle D' with radius α and centre $\vec{a}$. Also note

$$\begin{aligned} \delta'(\vec{x}, \vec{b}) &\leq \alpha\delta(\vec{x}, \vec{b}) \leq \alpha(\delta(\vec{x}, \vec{a}) + \varepsilon(n-1)) \\ &= \delta(\vec{x}, \vec{a}) + (\alpha - 1)\,\delta(\vec{x}, \vec{a}) + \alpha\varepsilon(n-1) \\ &\leq \delta'(\vec{x}, \vec{a}) + (\alpha - 1)\,\delta(\vec{x}, \vec{a}) + \alpha\varepsilon(n-1). \end{aligned} \tag{8}$$

Thus, removing $\vec{a}$ from O will decrease $\mathrm{cost}'_{\mathrm{pen}}(O)$ by at most

$$\begin{aligned} &\iint_{D'} \max\left[\delta'(\vec{x}, b) - \delta'(\vec{x}, \vec{a}), 0\right] \\ &\overset{(i)}{\leq} \iint_{D'} (\alpha - 1)\,\delta(\vec{x}, a) + \alpha\varepsilon(n-1) d\mu \\ &\leq \frac{1}{\varepsilon^2}\Big((\alpha - 1)\,\frac{2\pi}{3}(\alpha + \varepsilon)^3 + (\alpha - 1)\varepsilon\pi(\alpha + \varepsilon)^2 \\ &\qquad + \alpha\varepsilon(n-1)\pi(\alpha + \varepsilon)^2\Big). \end{aligned} \tag{9}$$

$\mathrm{cost}'_{\mathrm{pen}}(O') - \mathrm{cost}'_{\mathrm{pen}}(O) \leq \frac{1}{\varepsilon^2}\left((\alpha^4 - \alpha^3)\frac{2\pi}{3} - \frac{\pi}{3\alpha^2} + O(n\varepsilon)\right)$ follows from (7). The latter is less than zero for $\alpha \in (1, 1.2)$ for large enough n. This contradicts O minimizing $\mathrm{cost}'_{\mathrm{pen}}$.

Consider $O' \in \mathcal{F}_{k^2}$ to be any optimum solution of $\mathcal{I}(S, n)$. We have established that O contains k^2 centres, one in each $C_{i,j}$ for $1 \leq i, j, \leq k$. Hence O' has the same property. Let $f : O' \to O$ be the unique function such that for any $o \in O'$, if $o \in C_{i,j}$ for some $1 \leq i, j, \leq k$, then also $f(o) \in C_{i,j}$. Therefore, $\sum_{o \in O_p} \delta(o, f_p(o)) \leq k^2\varepsilon(n-1) \leq \beta$. So $\delta^{bij}_{\mathcal{I}}(O, O') \leq \beta$. Thus, $\mathcal{I}(S, n)$ is (α, β)-stable. □

4.2 Stable k-Median in Doubling Metrics

Focusing on the setting without penalties, we prove that (α, β)-stable k-Median with doubling constant 8 is hard, for any $1 < \alpha < 1.2$, $0 < \beta$. The main proof idea for the following theorem is that while the point set is located in $\mathbb{R}^3$, introducing a metric δ^{pen} with doubling dimension 3 (different from standard norm two metric) allows us to replicate the effect of penalties.

Theorem 4 (Hardness of Almost Stable k-Median)**.** *For any $\alpha \in (1, 1.2)$, $\beta > 0$, there is no $f(k)n^{o(\sqrt{k})}$-time algorithm for (α, β)-stable k-Median in metrics of bounded doubling constant, unless ETH fails.*

5 Hardness of α-stable k-Median in Bounded-Dimensional Metrics

[Friggstad *et al.*, 2019a] set out to resolve the complexity of $(1 + \varepsilon)$-stable instances of k-Median and k-Means for constant values of ε, which is highly desirable since this is the

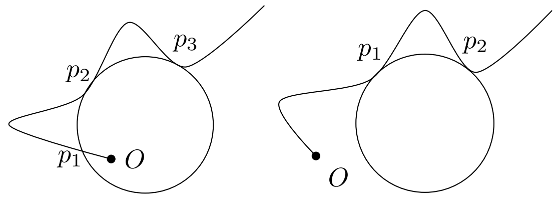

Figure 1: In $\mathbb{R}^4$ (left), the 3-sphere that goes through the points p_1, p_2, p_3 on the moment curve and is tangent to the moment curve on p_2 and p_3, has no other intersections with the moment curve after the origin. In $\mathbb{R}^3$ (right), the sphere that is tangent to the moment curve on p_1 and p_2 has no other intersections with the moment curve.

range where the ρ-swap local search can produce solutions in polynomial time. However, the computational complexity of the problem remains unknown when we go beyond constant error values. This question is of particular interest when discussing Fixed Parameter Tractable (FPT) algorithms that allow us slightly super-polynomial running times. Therefore, we examine the computational complexity of $(1+\varepsilon)$-stable instances of k-MEDIAN (and k-MEANS) when ε is an inverse polynomial in n. As the main result of this section, we show that for $\varepsilon = \frac{1}{poly(n)}$, the problem is still hard.

Inspired by [Cohen-Addad *et al.*, 2018], our general approach is to reduce PVC to k-MEDIAN using moment curves. We first establish some useful properties regarding moment curves and spheres. These properties resemble those shown in Section 2.1 of [Cohen-Addad *et al.*, 2018], but are modified to our purpose. The proofs, in particular, utilize Descartes' rule of signs (see [Curtiss, 1918]).

Lemma 6. *Let $0 < t_1 < t_2 < t_3$. Consider the 4-dimensional moment curve $\left(t, t^2, t^3, t^4\right)$ and the 3-sphere in $\mathbb{R}^4$ that goes through the moment curve at $t = t_1$ and is tangent to the curve at $t = t_2, t_3$. The segments on the moment curve for $t \in (t_1, t_2) \cup (t_2, t_3) \cup (t_3, \infty)$ all lie outside of the sphere.*

Lemma 7. *Let $0 < t_1 < t_2$. Consider the 3-dimensional moment curve $\left(t, t^2, t^3\right)$ and the unique 2-sphere in $\mathbb{R}^3$ that is tangent to the moment curve at $t = t_1, t_2$. The sphere only touches the curve at t_1 and t_2.*

5.1 $(1 + \frac{1}{poly(n)})$-Stable Euclidean k-MEDIAN with Penalties

This subsection establishes the hardness of $(1 + 1/poly(n))$-stable Euclidean k-MEDIAN with penalties in $\mathbb{R}^4$. Like [Cohen-Addad *et al.*, 2018], we reduce from PARTIAL VERTEX COVER (PVC) by creating a candidate centre on a moment curve for each vertex of the input graph, along with a data point for each edge of the input graph $G = (V, E)$. For each edge, we ensure that the point corresponding to that edge is closer to the two centres representing each vertex of the edge than to any other candidate centre. This allows us to reduce a PVC instance to a k-MEDIAN instance.

However, to ensure the stability of k-MEDIAN with penalties, we need to guarantee that the difference between the distances of the representation of an edge to its corresponding vertex and to the representation of other vertices is significant. Moreover, with the reduction introduced by [Cohen-Addad *et al.*, 2018], perturbing the metric function risks having the optimum solution switch from one maximal partial cover of the graph to another. To avoid this, we ensure that the distance between the representation of edges and the representation of their respective vertices remains constant.

Theorem 5 (Hardness of $(1 + \frac{1}{poly(n)})$-Stable Penalty k-MEDIAN)**.** *There is no $f(k)n^{o(k)}$-time algorithm for $(1 + \mathcal{O}\left(\frac{1}{n^{12}}\right))$-stable Euclidean k-MEDIAN with penalties in $\mathbb{R}^4$, unless ETH fails.*

5.2 $(1 + \frac{1}{poly(n)})$-Stable Euclidean k-MEDIAN

Finally, we focus on k-MEDIAN without penalties and show that $(1 + \frac{1}{poly(n)})$-stable Euclidean k-MEDIAN in $\mathbb{R}^6$ is hard. The proof is non-trivial and builds on similar ideas as the proof of Theorem 5.

Theorem 6 (Hardness of $(1 + \frac{1}{poly(n)})$-Stable k-MEDIAN)**.** *There is no $f(k)n^{o(k)}$-time algorithm for $(1+\mathcal{O}\left(\frac{1}{n^{16}}\right))$-stable Euclidean k-MEDIAN in $\mathbb{R}^6$, unless ETH fails.*

6 Conclusion

This paper addresses the complexity of stable instances of k-MEANS and k-MEDIAN under generalized stability notions, and in the generalized setting with penalties. For our most general stability notion, i.e., (α, β)-stability, the problem is proven highly intractable, even in small-dimensional Euclidean spaces. Some of our proofs build on proof ideas from the related literature; yet they require new and non-trivial technical insights in order to address our generalized settings. We fall short of answering the following question.

- **Open Problem 1:** Does a PTAS (a $(1+\varepsilon)$-approximation for any constant $\varepsilon > 0$) exist for almost stable instances of k-MEANS/k-MEDIAN?

We conjecture that the answer is yes, as there is evidence that the local search may be producing feasible solutions that are simultaneously near optimal and β-close to one optimal solution. For instance, one can show that the algorithm visits at least one nearly-good solution in a polynomial number of iterations. We also prove that $(1+\varepsilon)$-stable instances of the problems are solvable in polynomial time, while the $(1+\varepsilon)$-stable instances are intractable again for inverse polynomial values of ε. Our analysis leaves the following question open:

- **Open Problem 2:** What is the parameterized complexity of the $(1 + 1/poly(n))$-stable instances?

Acknowledgments

K. Khodamoradi was supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (application no. RGPIN-2024-06360). S. Zilles was supported by a Canada CIFAR AI Chair (held at Amii), an NSERC Canada Research Chair, an NSERC Discovery Grant (application no. RGPIN-2017-05336), and through the New Frontiers in Research Fund under grant no. NFRFE-2023-00109.

References

[Awasthi *et al.*, 2010] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a PTAS for k-Median and k-Means clustering. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 309–318. IEEE Computer Society, 2010.

[Balcan and Liang, 2016] Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM J. Comput.*, 45(1):102–155, 2016.

[Ben-David *et al.*, 2006] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In Gábor Lugosi and Hans Ulrich Simon, editors, *Learning Theory, 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006, Proceedings*, volume 4005 of *Lecture Notes in Computer Science*, pages 5–19. Springer, 2006.

[Bilu and Linial, 2012] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, 2012.

[Cohen-Addad *et al.*, 2018] Vincent Cohen-Addad, Arnaud De Mesmay, Eva Rotenberg, and Alan Roytman. The bane of low-dimensionality clustering. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 441–456. SIAM, 2018.

[Curtiss, 1918] DR Curtiss. Recent extentions of descartes' rule of signs. *Annals of Mathematics*, 19(4):251–278, 1918.

[Drineas *et al.*, 2004] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33, 2004.

[Friggstad *et al.*, 2019a] Zachary Friggstad, Kamyar Khodamoradi, and Mohammad R Salavatipour. Exact algorithms and lower bounds for stable instances of Euclidean k-Means. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2958–2972. SIAM, 2019.

[Friggstad *et al.*, 2019b] Zachary Friggstad, Mohsen Rezapour, and Mohammad R Salavatipour. Local search yields a PTAS for k-Means in doubling metrics. *SIAM Journal on Computing*, 48(2):452–480, 2019.

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

[Khodamoradi *et al.*, 2025] Kamyar Khodamoradi, Farnam Mansouri, and Sandra Zilles. The computational complexity of almost stable clustering with penalties. *ArXiv*, abs/2510.03143, 2025.

[Makarychev *et al.*, 2012] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In Howard J. Karloff and Toniann Pitassi, editors, *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 367–384. ACM, 2012.

[Vazirani, 2001] Vijay V. Vazirani. *Approximation algorithms*. Springer, 2001.