Approximate Cluster Recovery from Noisy Labels

Buddhima Gamlath BUDDHIMA@GOOGLE.COM
Google Zurich

Silvio Lattanzi SILVIOL@GOOGLE.COM
Google Research

Ashkan Norouzi-Fard ASHKANNOROUZI@GOOGLE.COM
Google Research

Ola Svensson OLA.SVENSSON@EPFL.CH [*]
EPFL

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Designing algorithms for machine learning problems targeting beyond worst-case analysis and, in particular, analyzing the effect of side-information on the complexity of such problems is a very important line of research with many practical applications. In this paper we study the classic k-means clustering problem in the presence of noisy labels: in addition to a set of points and parameter k, we receive cluster labels of each point generated by either an adversarial or a random perturbation of the optimal solution. Our main goal is to formally study the effect of this extra information on the complexity of the k-means problem. In particular, in the context of random perturbations, we give an efficient algorithm that finds a clustering of cost within a factor $1 + o(1)$ of the optimum even when the label of each point is perturbed with a large probability (think 99%). In contrast, we show that side-information with adversarial perturbations does not help, namely the problem remains as hard as the original k-means problem even if only a small ϵ fraction of the labels are perturbed. We complement this negative result by giving a simple algorithm in the case when the adversary is only allowed to perturb an ϵ fraction of the labels per *each cluster*.

Keywords: k-means, beyond worst-case, noisy labels.

1. Introduction

Clustering is a central problem in unsupervised learning with many real world applications. Perhaps the most widely studied clustering problem is the k-means problem. In this problem, we are given a finite set of points $P \subset \mathbb{R}^d$ in a d-dimensional Euclidean space. The goal is to find a set of k centers $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ such that the sum of squared distances between the points in P to their closest centers is minimized. I.e., the k-means problem asks to solve the following optimization problem:

$$(\mathbf{c}_1^*, \ldots, \mathbf{c}_k^*) = \operatorname*{arg\,min}_{\mathbf{c}_1,\ldots,\mathbf{c}_k \in \mathbb{R}^d} \sum_{\mathbf{p} \in P} \left(\min_{i \in k} \|\mathbf{p} - \mathbf{c}_i\| \right)^2.$$

The k-means problem has attracted a lot of attention in the last decades Arthur and Vassilvitskii (2007); Arya et al. (2004); Byrka et al. (2014); Charikar and Guha (2005); Kanungo et al. (2004); Jain et al. (2003); Li (2011); Lloyd (1982); Ahmadian et al. (2017). A wide range of techniques

[*] Research supported by the Swiss National Science Foundation project 200021-184656 "Randomness in Problem Instances and Randomized Algorithms."

have been introduced to achieve high quality solutions and hardness results; the best known approximation guarantee for the problem is 6.357 Ahmadian et al. (2017).

While algorithms with strong worst-case guarantees are certainly desirable, additional information regarding instances often allow for dramatically improved solution guarantees. Classic examples for k-means clustering are assumptions on the "clusterability" of instances Awasthi et al. (2010); Bilu and Linial (2012); Cohen-Addad and Schwiegelshohn (2017); Kumar and Kannan (2010); Ostrovsky et al. (2013). Another natural viewpoint is to analyze the impact of side-information such as noisy labels. This viewpoint is motivated by the fact that, in many real-world scenarios, in addition to the input points, we have access to extra information about the points that can be used to improve the quality of the solution. For example, consider a machine learning pipeline that receive as input a set of points and a cluster labeling computed by an earlier stage of the pipeline. Naturally, one cannot assume that the input labeling is perfect but it is reasonable to assume that it provides some noisy labels. Similarly, imagine a setting where humans provide initial clustering labels in a crowd-sourcing setting. Also in this case, the labels will not be perfect but they will contain useful information. Motivated by such examples, many machine learning problems including clustering problems in generative models have been studied with noisy labels Abbe et al. (2021); Saad and Nosratinia (2018); Esmaeili et al. (2019); Lesieur et al. (2016); Lelarge and Miolane (2019). However, perhaps surprisingly, there has been no such work on the classic k-means problem where there is no assumption on the underlying data distribution. We address this question with the goal to understand what kind and what amount of noise on the labels allow for the development of efficient algorithms with strong approximation guarantees.

Our results and overview of techniques. We study two natural noise models, the adversarial and the stochastic. In the adversarial noise model, starting from a correct labeling, an adversary changes a fraction of the labels. In the stochastic noise model, the label of a fraction of the points are altered at random. More formally, in the noisy k-means problem, in addition to the set of points P and number of centers k, we are provided with a clustering $O' = (O'_1, \ldots, O'_k)$ that is a noisy version of an optimum solution O[1]. For each point $\mathbf{p} \in P$, we refer to the id of the cluster that it belongs in O' as its label, i.e., the label of a point $\mathbf{p} \in O'_i$ is i. The goal of this work is to investigate the impact of this extra information on the complexity of the problem, its benefits and limitations. Therefore, we consider different models based on the source of the noise.

The first model that we consider is the adversarial noise model. In this model, an adversary is allowed to pick any optimum solution and change the label of ε-fraction of the points to any desired label, for some $\varepsilon > 0$. Therefore, in this model, at least $(1-\varepsilon)$-fraction of the labels are not changed. At first, this model might sound very strong as revealing the labels of most of the points can potentially be very useful in finding the optimal solution, but this is not the case. Our first result shows that this labeling does not provide any useful information and any hardness result for the k-means problem also applies to this setting. Intuitively, this is true because one can take any hard instance of the k-means problem on n points and add n/ε additional points extremely close to each other but far away from all the points in the hard instance. In this new instance, the adversary can assign any label to the points in the initial hard instance while ensuring that the labels for at least $(1-\varepsilon)$-fraction of the points are correct. An improved and formal argument for this is presented in Appendix F (Theorem F.1). Motivated by this, we define a *Balanced Adversarial Noise* model. In

1. We note that our noise model is strictly more challenging than a model where an $(1-\varepsilon)$-fraction of the clustering labels are revealed either adversarially or randomly.

this model, the adversary creates a set of perturbed labels $O' = \{O'_1, O'_2, \ldots, O'_k\}$ such that there exists an optimum solution $O = \{O_1, O_2, \ldots, O_k\}$ so that for all $i \in [k]$, the size of the symmetric difference $|O_i \Delta O'_i| \leq \varepsilon |O_i|$. This setting is basically the same as adversarial setting with the caveat that a fraction of points in each cluster is preserved instead of a fraction of the entire instance. Interestingly, we can show that in this setting it is possible to achieve a high quality solution.

Theorem 1.1 *(Informal version of Theorem B.1) There exists a $(1 + O(\varepsilon))$-approximation algorithm for the noisy k-means problem in the balanced adversarial noise model.* [2]

To obtain such a result, we consider each cluster O'_j separately. We start by observing that if $O'_j \subseteq O_j$ (i.e., no points were added to the cluster), then the centroid of O'_j is close to that of O_j, which gives a $(1 + O(\varepsilon))$ approximate solution. However, the points that are added to O_j can move the centroid of O'_j much farther from that of O_j. Let $A_j = O'_j \setminus O_j$ be the added points. If a point in A_j is close to the centroid of O_j, its contribution to the movement of centroid of O'_j is not significant. Thus, only the points in A_j that are far away from the centroid of O_j are problematic. Intuitively, if we somehow remove such "outliers" from consideration and compute the centroid considering the remaining points, it should result in a good solution. In fact, we show that the following outlier removal approach works: Consider a set of candidate centers that is guaranteed to contain a point close to the centroid of O_j and find the center that gives the minimum sum of squared distances to points in O'_j excluding the farthest $2 \cdot \varepsilon$ fraction of the points. Note that by excluding the farthest away points, we might also exclude some points in O_i, but since it is only a small fraction, we are still able to show that the resulting center is close to the centroid of O_j. A formal version of this argument is presented in Appendix B.

Furthermore, we show that any algorithm that considers only the points in O'_j to find its centroid, cannot achieve an asymptotically better guarantee.

Theorem 1.2 *In the balanced adversarial noise model, any (potentially randomized) algorithm has an approximation guarantee of $(1 + \Omega(\varepsilon))$ if it computes the center of each cluster only as a function of the input points with the label of that cluster.*

In many practical scenarios, the adversarial setting is too pessimistic and a stochastic noise model is more suitable. We consider two natural variants of such noise: the proportional stochastic model where the noisy labels are proportional to cluster sizes and the uniform stochastic model where noise is uniform across labels.

We first study the algorithmically easier *proportional stochastic model*. Starting from an arbitrary optimum solution $O = \{O_1, O_2, \ldots, O_k\}$, each point keeps its label with probability $1 - \varepsilon$ and its label is changed to label j with probability $\varepsilon \frac{|O_j|}{|P|}$. More precisely, the set of perturbed labels $O' = \{O'_1, O'_2, \ldots, O'_k\}$ is constructed as follows: for each cluster $i \in [k]$ and $\mathbf{p} \in O_i$, put $\mathbf{p}$ in O'_i with probability $1 - \varepsilon$ and put $\mathbf{p}$ in O'_j with probability $\varepsilon \frac{|O_j|}{|P|}$ for each $j \in [k]$.[3] This model, as in the adversarial model, ensures that a fraction of the points from each cluster is kept and the expected number of points added to each cluster i is at most $\varepsilon |O_i|$. However, in contrast to the adversarial model, we are here able to get an approximation guarantee of $1 + o(1)$ even if only a small (e.g., $1/\log |P|$) fraction of the labels are accurate.

2. By definition, the running time of approximation algorithm is polynomial with respect the size of input.
3. Thus the probability that a point $\mathbf{p} \in O_i$ retains its original label is $1 - \varepsilon + \varepsilon \frac{|O_i|}{|P|}$.

Theorem 1.3 *(Informal version of Theorem 3.1) For $\varepsilon \leq 1-1/\log|P|$, there exists a $(1+O(1/\log|P|))$-approximation algorithm with a success probability of $1 - 1/|P|$ for the proportional stochastic model, assuming that $|O_i| > \mathrm{poly}(\log|P|)$ for $1 \leq i \leq k$.*

Our approach has two main steps: In the first step, we find a set of points B that contains most of the points of an optimum cluster while having the diameter[4] bounded by the average cost of that cluster. In the second step, we focus on the points identified in B and use them to estimate the centroid of this cluster. The fact that the diameter of the points in B is bounded enables us to achieve a good estimate with high probability. As presented in Section 2, we do so by considering the centroid of $B \cap O'_j$ as a linear combination of the centroids of unperturbed and perturbed points in $B \cap O'_j$.

Perhaps the most natural stochastic model is the *uniform stochastic noise model*: each point keeps its label with probability $1 - \varepsilon$ and with the remaining probability its label is sampled uniformly. It turns out that this model is more challenging than the proportional model. Indeed a major challenge with uniform noise is that, for the clusters that have small (say n^{ε}) size, the amount of noisy labels added is almost n/k. Therefore, the number of noisy-labeled points can be a factor $n^{(1-\varepsilon)}/k$ more than the correctly labeled points. In other words, the signal is much weaker than the noise for small clusters. We overcome this with a more complex approach that iteratively finds the currently large clusters and discards them. This allows us to achieve similarly strong guarantees for the uniform stochastic model: we get close-to-optimal solutions even if only $1/\log|P|$ fraction of the points are correctly labeled.

Theorem 1.4 *(Informal version of Theorem 4.1) For $\varepsilon \leq 1-1/\log|P|$, there exists a $(1+O(1/\log|P|))$-approximation algorithm with a success probability of $1 - 1/|P|$ for the uniform stochastic model, assuming that $|O_i| > \mathrm{poly}(k, \log|P|)$ for $1 \leq i \leq k$.*

The algorithm for the uniform stochastic noise model is explained in Section 4.

Related work. Thanks to its natural motivation, many machine learning problems have been studied in presence of noisy labels. For example, in the context of learning halfspaces, both the adversarial Haussler (1992); Kearns et al. (1994), the stochastic Angluin and Laird (1988), and the Massart noise models Diakonikolas et al. (2019) have been studied.

When we restrict our attention to clustering, side information has been extensively studied to improve the results on the stochastic block model. For example, Mossel et al. (2014) use the output of another algorithm to show that belief propagation run on that output recovers the underlying clustering. Several papers also consider side information in the form of clustering labels Abbe et al. (2021); Saad and Nosratinia (2018); Esmaeili et al. (2019). In metric spaces, side information has been used to improve results on mixture of Gaussians Lesieur et al. (2016); Lelarge and Miolane (2019). In this context, our work extend this line of work in a setting where we have no assumption on the underlying structure of the clustering.

More generally, our problem is also related to the problem of estimating a vector of discrete variables, using a set of on noisy observations on the pairs has also been studied El Alaoui and Montanari (2021).

Finally, our paper is also related to the literature studying beyond worst-case analysis Roughgarden (2019). A closely related line of work in this are is the semi-supervised active clustering

4. The maximum distance between two points.

framework (SSAC) Ashtiani et al. (2016). In this model we are given a set X of n points and an oracle answering to same-cluster queries of the form "are these two points in the same cluster?". In this model both k-meas clustering and clustering with side information have been studied Ashtiani et al. (2016); Mazumdar and Saha (2017).

2. Stochastic Noise Model

In this section we consider the the k-means problems with stochastic noise, and we focus on the problem of designing an algorithm to estimate the centers for each cluster separately. Later, in Section 3, we show how to use our algorithm as a black box in a straightforward manner to solve the k-means problem in the proportional stochastic noise setting. Then, in Section 4, we present a more elaborate, iterative algorithm that again uses the algorithm of this section as a black box to solve the k-means problem in the uniform stochastic noise setting.

Formally, we consider the following problem, which we refer to as the *noisy center estimation* problem: Let $P \subset \mathbb{R}^d$ be a set of points, $O \subseteq P$ be a cluster of interest, and let $\varepsilon, \delta \in (0,1)$ be a pair of parameters denoting probabilities. An instance of the noisy center estimation problem consists of the quadruple $(P, O, \varepsilon, \delta)$. Let P_{noisy} be a random subset of P constructed by including each point of P independently with probability ε. Let $O_{\text{good}} = O \setminus P_{\text{noisy}}$ and let O_{bad} be a random subset of P_{noisy} which includes each point of P_{noisy} independently with probability δ. Let $O_{\text{noisy}} = O_{\text{good}} \cup O_{\text{bad}}$. We call such O_{noisy} an (ε, δ)-noisy version of O. Given P, an (ε, δ)-noisy version O_{noisy} of O, and parameters ε, δ, the noisy center estimation problem asks to find a center $\hat{\mathbf{o}}$ that is close to the centroid $\mathbf{o}$ of O.

Let $\text{cost}(X, \mathbf{c}) := \sum_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{c}\|^2$ denote the sum of squared distances from a points in a set $X \subset \mathbb{R}^d$ to a center $\mathbf{c}$, and we further use $\text{cost}(X, C) = \sum_{\mathbf{x} \in X} \min_{\mathbf{c} \in C} \|\mathbf{x} - \mathbf{c}\|^2$ for a subset C of centers. Let $r_{\text{avg}} = \sqrt{\text{cost}(O, \mathbf{o})/|O|}$ be the *average radius* of O.[5] Our aim is to find a center $\hat{\mathbf{o}}$ such that $\|\hat{\mathbf{o}} - \mathbf{o}\| = O\left(1/\log^{0.5} |P|\right) \cdot \sqrt{\text{cost}(O, \mathbf{o})/|O|}$. Since $\text{cost}(O, \hat{\mathbf{o}}) = \text{cost}(O, \mathbf{o}) + |O| \cdot \|\hat{\mathbf{o}} - \mathbf{o}\|^2$, such a center $\hat{\mathbf{o}}$ satisfies $\text{cost}(O, \hat{\mathbf{o}}) \leq (1 + O\left(1/\log |P|\right)) \cdot \text{cost}(O, \mathbf{o})$. We formally state this result below in Theorem 2.3.

The success of our algorithm depends on several natural assumptions on the input. First, we assume that the underlying cluster is not too small, and the noise parameter ε (i.e., the probability with which a point is selected for noisy labeling) is not too large. We also assume that the noise parameter δ (i.e., the probability with which the perturbed points are relabeled as belonging to O) is not too high so that the variance of the size of O_{noisy} is small compared to the size of the underlying cluster O. Formally, we define the following:

Definition 2.1 *We say an instance $(P, O, \varepsilon, \delta)$ of the noisy cluster estimation problem is* nice *if $|O| \geq \log^{200} |P|$, $\varepsilon \leq 1 - \frac{1}{\log |P|}$, and $\delta \leq \frac{|O|^{1.1}}{|P| \log^2 |P|}$.*

We remark that it is *not* necessary for the algorithm to know the parameters ε and δ exactly; instead it is sufficient to know good approximations ε' and δ', as formalized below:

Definition 2.2 *For an instance $(P, O, \varepsilon, \delta)$ of the noisy cluster estimation problem, we say that the probability parameters $\varepsilon', \delta' \in (0,1)$ are* close *approximations of ε and δ if $|\varepsilon - \varepsilon'| \leq |O|^{-0.4} \cdot \varepsilon$ and $|\delta - \delta'| \leq |O|^{-0.4} \cdot \delta$.*

5. Note that, despite the term '*average* radius', this is in fact the quadratic mean of point-to-center distances.

As the main result of this section, we prove the following theorem:

Theorem 2.3 *There exists an algorithm* ONECENTER *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ *be a* nice *instance of the noisy center estimation problem,* O_{noisy} *be an* (ε, δ)*-noisy version of* O*, and* ε', δ' *be* close *approximations of* ε *and* δ*. The algorithm* ONECENTER *takes* $(P, O_{\text{noisy}}, \varepsilon', \delta')$ *as input and outputs a center* $\hat{\mathbf{o}}$ *such that*

$$\text{cost}(O, \hat{\mathbf{o}}) \leq (1 + O(1/\log |P|)) \cdot \text{cost}(O, \mathbf{o})$$

with probability at least $1 - \exp\left(-|O|^{0.2}\right)$*, where the probability is over the randomness of the noisy labels and the algorithm's internal random bits.*

Overview of the techniques

Recall that O_{noisy} is the union of the following two disjoint sets:

1. O_{good}, i.e., the points that were originally in O but were not included in P_{noisy}. Observe that O_{good} is a random set of points constructed by including each point in O independently with probability $1 - \varepsilon$.

2. O_{bad}, i.e., the points that were included in O_{noisy} due to noise. Note that O_{bad} can be viewed as a random set of points constructed by including each point of P independently with probability $\varepsilon \cdot \delta$. However, it is *not* independent from O_{good}.

Observe that the centroid $\mathbf{o}_n$ of O_{noisy} is a linear combination of the centroids $\mathbf{o}_g$ and $\mathbf{o}_b$ of O_{good} and O_{bad} respectively. Namely, we have $\mathbf{o}_n = (|O_{\text{good}}|/|O_{\text{noisy}}|) \cdot \mathbf{o}_g + (|O_{\text{bad}}|/|O_{\text{noisy}}|) \cdot \mathbf{o}_b$, and substituting $|O_{\text{good}}| = |O_{\text{noisy}}| - |O_{\text{bad}}|$, we get

$$\mathbf{o}_g = \frac{\mathbf{o}_n - (|O_{\text{bad}}|/|O_{\text{noisy}}|) \cdot \mathbf{o}_b}{1 - |O_{\text{bad}}|/|O_{\text{noisy}}|}. \tag{1}$$

Intuitively, the centroid $\mathbf{s}$ of a set S of points can be approximated by taking the centroid of a random subset T obtained by including each point in S independently with some fixed probability (see Lemma A.3 for a statement of this kind). In particular, the latter is an unbiased estimator for the former, whose variance is $O\left(1/|T| \cdot \text{cost}(S,\mathbf{s})/|S|\right)$. Consequently, $\mathbf{o}_g$ is an unbiased estimator for the true centroid $\mathbf{o}$ of O. We may thus use Eq. (1) to estimate the centroid of O since we can find/estimate each term on the right hand side as follows: We can directly compute the centroid $\mathbf{o}_n$ from the input and closely approximate the ratio $|O_{\text{bad}}|/|O_{\text{noisy}}|$ using the standard concentration bounds. As for the centroid $\mathbf{o}_b$ of O_{bad}, we can again invoke Lemma A.3 since we can view O_{bad} as a random subset obtained by including each point of P independently with probability $\varepsilon \cdot \delta$.

However, there are two issues with the above approach:

1. The bound we have for the variance of $\mathbf{o}_b$ is in terms of $\text{cost}(P, \mathbf{p})$ instead of $\text{cost}(O, \mathbf{o})$. The former can be arbitrarily large compared to the latter.

2. Even if $\text{cost}(P, \mathbf{p})$ is comparable to $\text{cost}(O, \mathbf{o})$, the failure probability bound we get is only inversely proportional to $|O|$ as we are using Chebyshev-style inequality to bound $\|\mathbf{o}_g - \mathbf{o}\|$ and $\|\mathbf{o}_b - \mathbf{p}\|$. Ideally, we prefer Chernoff-style bounds that are exponentially small in terms of $|O|$, which would imply a wider applicability.

To circumvent the first issue, we employ a two stage algorithm: In the first step, we find a ball B that contains most (i.e., $1 - o(1)$ fraction) of the points of cluster O while having the diameter bounded by $O(\log^{0.5} |P|) \cdot \sqrt{\text{cost}(O, \mathbf{o})/|O|}$, and restrict our attention to only the points in side B. Since B contains most of the points of O, the centroid $\mathbf{o}$ of O is close to that of $O \cap B$ by Lemma B.2 as will be shown in Appendix B. To this end, we define the following:

Definition 2.4 *For a subset $Q \subseteq P$, we define a ball with center $\mathbf{c}$ and radius r, denoted by $\text{BALL}_Q(\mathbf{c}, r)$ as the set of all points in Q that are within distance r from $\mathbf{c}$. Namely*

$$\text{BALL}_Q(\mathbf{c}, r) = \{\mathbf{q} \in Q : \|\mathbf{c} - \mathbf{q}\| \leq r\}.$$

We denote by $\mathcal{B}^{\text{all}}$ the set of all balls for P whose center belongs to P and whose radius corresponds to the distance between the center and some point in P. Formally,

$$\mathcal{B}^{\text{all}} = \{\text{BALL}_P(\mathbf{c}, r) : \mathbf{c} \in P \text{ and } r = \|\mathbf{c} - \mathbf{c}'\| \text{ for some } \mathbf{c}' \in P\}.$$

By definition, $\mathcal{B}^{\text{all}}$ contains at most $|P|^2$ balls.

Definition 2.5 *We say a ball $B \in \mathcal{B}^{\text{all}}$ is good if the following two conditions hold:*

1. *B contains at least $1 - \frac{1}{\log |P|}$ fraction of the points in O, and*
2. *the diameter of B, i.e., $\max_{\mathbf{x},\mathbf{y} \in B} \|\mathbf{x} - \mathbf{y}\|$, is at most $32 \cdot (\log^{0.5} |P|) \cdot r_{\text{avg}}$.*

We denote set of all good balls by $\mathcal{B}^{\text{good}}$.

The way we find a good ball B is to guess a center and a radius for the ball and perform a statistical test to check if the guess is good. Namely, for each candidate center and radius, we compare the number of points in O_{noisy} that are outside the corresponding ball with the expected number of such points if the guess were to be good. This yields the following lemma which we prove later in Appendix C.1.

Lemma 2.6 *There exists an algorithm* GOODBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ be a nice *instance of the noisy center estimation problem, O_{noisy} be an (ε, δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. The algorithm* GOODBALL *takes $(P, O_{\text{noisy}}, \varepsilon', \delta')$ as input and outputs a ball $B \in \mathcal{B}^{\text{all}}$ such that $B \in \mathcal{B}^{\text{good}}$ with probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$, where the probability is over the randomness of O_{noisy}.*

In the second step, we use Eq. (1) considering only the points inside the identified ball to estimate the centroid of $O \cap B$. To boost the success probability and overcome the second issue we mentioned above, we use further randomization in the second step followed by a high-dimensional median trick (see Corollary A.5). Namely, instead of naïvely using our approach on the points of B to estimate a single centroid for $O \cap B$, we use it on many random partitions of B. Then we take the geometric median of the estimated centers as our final output. We describe the key ideas of this approach in Section 2.1, but we defer the technicalities of the high-dimensional median trick to Appendix C.2. This approach yields the following lemma. In Section 2.1, we prove a weaker version of it and present a proof-sketch for the stronger version. Due to the technicalities involved, we defer the complete proof of Lemma 2.7 to Appendix C.2.

Lemma 2.7 *There exists an algorithm* CENTERINBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ *be a* nice *instance of the noisy center estimation problem,* O_{noisy} *be an* (ε, δ)*-noisy version of* O*, and* ε', δ' *be* close *approximations of* ε *and* δ*. The algorithm* CENTERINBALL *takes* $(P, O_{\text{noisy}}, \varepsilon', \delta')$ *and* $B \subseteq P$ *as input and outputs a center* $\tilde{\mathbf{o}}_B$*. With probability at least* $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$ *over the randomness of* O_{noisy} *and the algorithm's internal random bits, it holds that*

$$\|\tilde{\mathbf{o}}_B - \text{centroid}(O \cap B)\| \leq \frac{6 \cdot r_{\text{avg}}}{\log^{0.5} |P|} \text{ for all } B \in \mathcal{B}^{\text{good}}.$$

Using the algorithm GOODBALL from Lemma 2.6 and the algorithm CENTERINBALL from Lemma 2.7, Theorem 2.3 now follows. Indeed, we first find a good ball using GOODBALL and then we estimate its center using CENTERINBALL. The high probability guarantee of Lemma 2.7 allows us to union-bound the failure probabilities over all possible good balls that the first stage (i.e., the algorithm GOODBALL) may output. For the formal argument how Theorem 2.3 is implied by the two lemmas, we refer the reader to Appendix C.4.

2.1. Estimating the centroids in good balls (proof of Lemma 2.7)

In this section, we first prove a weaker version of Lemma 2.7 to demonstrate our key techniques. Later we explain how it can be combined with random partitioning and a high dimensional median trick to prove the stronger result of Lemma 2.7. We defer the full proof of Lemma 2.7 to Appendix C.2. We remark that, to simplify the calculations, the weaker version stated below assumes that the algorithm knows the probability parameters ε and δ exactly.

Lemma 2.8 *There exists an algorithm that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ *be a* nice *instance of the noisy center estimation problem and* O_{noisy} *be an* (ε, δ)*-noisy version of* O*. Let* $B \in \mathcal{B}^{\text{good}}$ *be good ball for the considered instance and further suppose that* $\varepsilon\delta|B| \geq |O|^{0.95}$*. The algorithm takes* $(P, O_{\text{noisy}}, \varepsilon, \delta)$ *and* B *as input and outputs a center* $\hat{\mathbf{o}}$*. With probability at least* $\frac{2}{3}$ *over the randomness of* O_{noisy}*, it holds that*

$$\|\hat{\mathbf{o}} - \tilde{\mathbf{o}}\| \leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5} |P|},$$

where $\tilde{\mathbf{o}}$ *is the centroid of* $O \cap B$*.*

Proof The algorithm proceeds as follows: it first computes $\alpha = \frac{\varepsilon\delta|B|}{|O_{\text{noisy}} \cap B|}$, the centroid $\mathbf{o}_n$ of $O_{\text{noisy}} \cap B$ and the centroid $\mathbf{b}$ of B; it then outputs the center

$$\hat{\mathbf{o}} = \frac{\mathbf{o}_n - \alpha \cdot \mathbf{b}}{1 - \alpha}.$$

Recall that in the noisy cluster estimation problem, we can view O_{noisy} as the union of two disjoint sets, O_{good} and O_{bad}, which are constructed as follows: First a random set P_{noisy} is constructed by including each point of P independently with probability ε. Then the set O_{good} is created by removing all points in P_{noisy} from O, and the set O_{bad} is constructed by including each point in P_{noisy} independently with probability δ.

With this viewpoint, we can note that the definition of $\hat{\mathbf{o}}$ is analogous to the right hand side of Eq. (1) considering only the points of B. Here the ratio α is an estimator for the quantity $|O_{\text{bad}} \cap B|/|O_{\text{noisy}} \cap B|$.

We claim that $\hat{\mathbf{o}}$ computed as above is a good estimator for the centroid $\tilde{\mathbf{o}}$ of $O \cap B$. To this end, observe that

$$\|\hat{\mathbf{o}} - \tilde{\mathbf{o}}\| = \left\| \frac{\mathbf{o}_n - \alpha \cdot \mathbf{b}}{1-\alpha} - \tilde{\mathbf{o}} \right\| = \frac{1}{1-\alpha} \|\mathbf{o}_n - \alpha\mathbf{b} - (1-\alpha)\tilde{\mathbf{o}}\|$$

Now, letting $\mathbf{o}' = (1-\alpha)\mathbf{o}_g + \alpha\mathbf{o}_b$ where $\mathbf{o}_g$ and $\mathbf{o}_b$ are the centroids of $O_{\text{good}} \cap B$ and $O_{\text{bad}} \cap B$, respectively, we get

$$\begin{aligned}
\|\hat{\mathbf{o}} - \tilde{\mathbf{o}}\| &= \frac{1}{1-\alpha} \left\|\mathbf{o}_n - \mathbf{o}' + \mathbf{o}' - \alpha\mathbf{b} - (1-\alpha)\tilde{\mathbf{o}}\right\| \\
&= \frac{1}{1-\alpha} \left\|\mathbf{o}_n - \mathbf{o}' + ((1-\alpha)\mathbf{o}_g + \alpha\mathbf{o}_b) - \alpha\mathbf{b} - (1-\alpha)\tilde{\mathbf{o}}\right\| \\
&\leq \frac{1}{1-\alpha} \left(\|\mathbf{o}_n - \mathbf{o}'\| + (1-\alpha) \cdot \|\mathbf{o}_g - \tilde{\mathbf{o}}\| + \alpha \cdot \|\mathbf{o}_b - \mathbf{b}\|\right) \\
&\leq \frac{1}{1-\alpha}\|\mathbf{o}_n - \mathbf{o}'\| + \|\mathbf{o}_g - \tilde{\mathbf{o}}\| + \frac{\alpha}{1-\alpha} \cdot \|\mathbf{o}_b - \mathbf{b}\|. \qquad (2) \\
& \qquad (3)
\end{aligned}$$

Thus, to bounding $\|\hat{\mathbf{o}} - \tilde{\mathbf{o}}\|$ now boils down to bounding the the terms $1/(1-\alpha)$, $\|\mathbf{o}_n - \mathbf{o}'\|$, $\|\mathbf{o}_g - \tilde{\mathbf{o}}\|$ and $\|\mathbf{o}_b - \mathbf{b}\|$.

Bounding $1/(1-\alpha)$ With Chernoff bounds, one can verify that, with probability at least 0.99,

$$|O_{\text{noisy}} \cap B| \geq (1 - |O|^{-0.4})((1-\varepsilon)|O \cap B| + \varepsilon\delta|B|) \geq (1 + |O|^{-0.2})\varepsilon\delta|B|, \qquad (4)$$

where we also use that $|O \cap B| \geq |O|/2$ since B is good and $\varepsilon\delta|B| \leq \delta|P| \leq |O|^{1.1}$ since $(P, O, \varepsilon, \delta)$ is nice. Let $\mathcal{E}_1$ be the event that $|O_{\text{noisy}} \cap B| \geq (1 + |O|^{-0.2})\varepsilon\delta|B|$. Conditioned on $\mathcal{E}_1$, we have that $\alpha = \frac{\varepsilon\delta|B|}{|O_{\text{noisy}} \cap B|} \leq \frac{1}{1+|O|^{-0.2}}$, which implies that $1/(1-\alpha) \leq 2|O|^{0.2}$.

Bounding $\|\mathbf{o}_n - \mathbf{o}'\|$ Again using the Chernoff bounds, we can verify that, with probability at least 0.99,

$$\left||O_{\text{bad}} \cap B| - \varepsilon\delta|B|\right| \leq |O|^{-0.25}\varepsilon\delta|B|. \qquad (5)$$

Let $\mathcal{E}_2$ be the event that Eq. (5) holds.

We can write the centroid $\mathbf{o}_n$ as a linear combination of $\mathbf{o}_g$ and $\mathbf{o}_b$.

$$\mathbf{o}_n = \frac{|O_{\text{good}} \cap B|}{|O_{\text{noisy}} \cap B|} \cdot \mathbf{o}_g + \frac{|O_{\text{bad}} \cap B|}{|O_{\text{noisy}} \cap B|} \cdot \mathbf{o}_b.$$

Observe that both $\mathbf{o}_n$ and $\mathbf{o}'$ (defined earlier in the derivation of Eq. (2)) lie on the same line segment between $\mathbf{o}_g$ and $\mathbf{o}_b$. The point $\mathbf{o}_n$ is $\frac{|O_{\text{bad}} \cap B|}{|O_{\text{noisy}} \cap B|} \cdot \|\mathbf{o}_g - \mathbf{o}_b\|$ away from $\mathbf{o}_g$ while the point $\mathbf{o}'$ is $\alpha \cdot \|\mathbf{o}_g - \mathbf{o}_b\|$ distance away from $\mathbf{o}_g$. Thus, conditioned on both $\mathcal{E}_1$ and $\mathcal{E}_2$, it holds that

$$\begin{aligned}
\|\mathbf{o}_n - \mathbf{o}'\| = \left|\frac{|O_{\text{bad}} \cap B|}{|O_{\text{noisy}} \cap B|} - \alpha\right| \cdot \|\mathbf{o}_g - \mathbf{o}_b\| &\leq \frac{|O|^{-0.25}\varepsilon\delta|B|}{(1+|O|^{-0.2})\varepsilon\delta|B|} \cdot \|\mathbf{o}_g - \mathbf{o}_b\| \\
&\leq \frac{64(\log^{0.5}|P|) \cdot r_{\text{avg}}}{|O|^{0.25}}.
\end{aligned}$$

The first inequality follows from Eq. (5), and the second one uses the fact that both $\mathbf{o}_b$ and $\mathbf{o}_g$ are in B and that B's diameter is bounded by $32 \cdot (\log^{0.5}|P|) \cdot r_{\text{avg}}$ since $B \in \mathcal{B}^{\text{good}}$.

Bounding $\|\mathbf{o}_g - \tilde{\mathbf{o}}\|$ and $\|\mathbf{o}_b - \mathbf{b}\|$ By the Chernoff bounds, one can verify that, with probability at least 0.99, it holds that

$$|O_{\text{good}} \cap B| \geq \frac{1}{2}(1-\varepsilon)|O| \geq |O|^{0.5} \text{ and } |O_{\text{bad}} \cap B| \geq \frac{1}{2}\varepsilon\delta|B| \geq |O|^{0.5}.$$

This, combined with Lemma A.3, yields that, with probability at least $1 - 0.99 - \frac{1}{|O|^{0.2}} - \frac{1}{|O|^{0.2}} \geq 0.95$,

$$\|\mathbf{o}_g - \tilde{\mathbf{o}}\| \leq \frac{r_{\text{avg}}}{\log^{0.5}|P|} \text{ and } \|\mathbf{o}_b - \mathbf{b}\| \leq \frac{r_{\text{avg}}}{|O|^{0.25}}.$$

Let $\mathcal{E}_3$ denote this event.

Thus, using Eq. (2) and conditioned on the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$, we observe that

$$\|\hat{\mathbf{o}} - \tilde{\mathbf{o}}\| \leq 2|O|^{0.2} \cdot \frac{64 \cdot (\log^{0.5}|P|) \cdot r_{\text{avg}}}{|O|^{0.25}} + \frac{r_{\text{avg}}}{\log^{0.5}|P|} + |O|^{0.2} \cdot \frac{r_{\text{avg}}}{|O|^{0.25}} \leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5}|P|},$$

where the last inequality uses that $|O| \geq \log^{200}|P|$. To conclude the proof, observe that $\Pr[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3] \geq 1 - \overline{\mathcal{E}_1} - \overline{\mathcal{E}_2} - \overline{\mathcal{E}_3} \geq \frac{2}{3}$.

■

Note that Lemma 2.8 is a weaker version of Lemma 2.7 in a few aspects. First, it assumes that the algorithm knows the parameters ε, δ exactly, but this can be easily relaxed at the expense of a slightly more involved calculation.

Secondly, it is constrained on the assumption that $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$. This assumption can also be removed as follows: Suppose that $\varepsilon\delta|B| \leq \frac{3}{2}|O|^{0.95}$. In this case, we can show that, with high probability, $|O_{\text{bad}} \cap B|$ is in the order of $o(1) \cdot |O_{\text{good}} \cap B|$. Thus, we can simply use the algorithm for the adversarial setting to estimate a good center for $O_{\text{good}} \cap B$. We can use a simple probabilistic check (which succeeds with high probability) to determine which algorithm to run depending on the value of $\varepsilon\delta|B|$.

Finally, the most crucial issue is that the probability guarantee of Lemma 2.8 is not nearly sufficient, as we need the estimate to be good regardless of which good ball is selected, and we need this guarantee with high probability. Namely, the algorithm must succeed for a fixed good ball with high probability ($\geq 1 - \frac{1}{2\cdot|P|^2}\exp(-|O|^{0.2})$) so that we can union bound over all ($\leq |P|^2$) such balls. Next, we discuss how to achieve this.

Suppose that instead of estimating a center using all points in B, we first take a random sample of $Q \subseteq B$ and run our approach considering only the points in Q. Then, with a good probability, we can show that the centroid of $O \cap Q$ is close to that of $O \cap B$, and hence the center estimated using the approach of Lemma 2.8 considering only the points of Q is close to the the centroid of $O \cap B$ with a constant probability.

Now we can take many random samples of $Q_1, \ldots, Q_t \subset B$, estimate a center $\hat{\mathbf{o}}_{Q_i}$ using each sample, and then take the geometric median of the samples as our final estimate. *If the majority of the estimated centers are close to* $\mathbf{o}$, then using the result of Minsker Minsker (2015), we get that the geometric median of the estimated centers is also close to $\mathbf{o}$.

Now, how can we argue that the majority of these estimates are good? First, we make the different runs of the approach of Lemma 2.8 independent on different random samples by ensuring that Q_i's disjoint. However, this poses an additional challenge: Let $\tilde{\mathbf{o}}_{Q_i}$ be the centroid of $O \cap Q_i$. Then the events that $\tilde{\mathbf{o}}_{Q_i}$'s being close to the centroid of $O \cap B$ are no longer independent. To get

around this challenge, we take many disjoint random samples of B, but only run our algorithm on a $\frac{1}{\log^3 |P|}$ fraction of them. This allows us to view each considered sample Q_i as a random subset of a set B_i where $B_i \subseteq B$ contains at least $1 - O\left(\frac{1}{\log^3 |P|}\right)$ fraction of the points of $B \cap O$'s points. (Note that this only holds conditioned some high probability events such as all sample sizes being not too large compared to their expected sizes.) Hence, for each Q_i, we can show that the centroid of $Q_i \cap O$ is close to that of $B_i \cap O$ with constant probability, which in turn is close to the centroid of $B \cap O$ due to Lemma B.2. By using an augmented Chernoff-style bound, we then show that, with high probability, the above holds for majority of Q_i's.

We formalize all these ideas in Appendix C.2 and prove Lemma 2.7 in its full generality.

3. Proportional Stochastic Noise Model

In this section, we present an algorithm to estimate the centers in the proportional stochastic noise model. Unlike the $1 + O(\varepsilon)$ approximation guarantee for the balanced adversarial noise model, we now aim for a $1 + o(1)$ approximation guarantee.

Theorem 3.1 *There exists an algorithm such that the following holds:*

Let $(P, \{O_1, \ldots, O_k\}, \varepsilon)$ be an instance of the k-means problem in the proportional stochastic noise setting where $|O_i| \geq \log^{200} |P|$ for all $i \in [k]$ and $\varepsilon \leq 1 - \frac{1}{\log |P|}$. Let $O'_1, \ldots, O'_k$ be the ε-noise added versions of $O_1, \ldots, O_k$. The algorithm takes as input $(P, \{O'_1, \ldots, O'_k\}, \varepsilon)$ and outputs centers $\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k$ such that, with probability at least $1 - \frac{1}{|P|}$, we have

$$\operatorname{cost}(O_i, \hat{\mathbf{o}}_i) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \operatorname{cost}(O_i, \mathbf{o}_i)$$

where $\mathbf{o}_i$ denote the centroid of O_i. Consequently, the output of the algorithm satisfies

$$\operatorname{cost}(O_i, \{\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k\}) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \mathrm{OPT}\,.$$

with probability at least $1 - \frac{1}{|P|}$. The probability is over the randomness of O'_i's and the algorithm's internal random bits.

The proof follows from the observation that $(P, O_i, \varepsilon, \delta)$, where $\delta = \frac{|O_i|}{|P|}$, is a nice instance of the noisy center estimation problem. Moreover, we have with high probability that ϵ, δ' are close approximations of ϵ, δ, where $\delta' = \frac{|O'_i|}{|P|}$. Hence, using the algorithm ONECENTER whose existence is guaranteed by Theorem 2.3, we can with high probability find a very accurate center $\hat{\mathbf{o}}_i$ for each cluster O_i. This implies that the cost of each cluster is low and thus also the cost of the whole clustering. The formal proof is given in Appendix D.

4. Uniform IID Noise Model

In this section, we present an algorithm to estimate the centers in the *uniform* stochastic noise model. Similarly to the proportional stochastic noise model, we obtain a $(1 + o(1))$ approximation guarantee, as formally stated below.

Theorem 4.1 *There exists an algorithm such that the following holds:*

Let $(P, \{O_1, \ldots, O_k\}, \varepsilon)$ be an instance of the k-means problem in the uniform stochastic noise setting where $|O_i| \geq \max(\log^{200}|P|, k^{200}\log|P|)$ for all $i \in [k]$ and $\varepsilon \leq 1 - \frac{1}{\log|P|}$. Let $O'_1, \ldots, O'_k$ be the ε-noise added versions of $O_1, \ldots, O_k$. The algorithm takes $(P, \{O'_1, \ldots, O'_k\}, \varepsilon)$ as input and outputs centers $\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_{k'}$ where $k' \leq k$ such that

$$\mathrm{cost}(P, \{\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_{k'}\}) \leq \left(1 + O\left(\frac{1}{\log|P|}\right)\right) \cdot \mathrm{OPT}$$

with probability at least $1 - \frac{1}{|P|}$, where the probability is over the randomness of O'_i's and the algorithm's internal random bits.

Suppose that we attempt to recover one cluster a time as before. Observe that for a fixed cluster label $i \in [k]$, recovering a good center for O_i can be viewed as solving the instance $(P, O_i, \varepsilon, \frac{1}{k})$ of the noisy center estimation problem from Section 2 with $(P, O'_i, \varepsilon, \frac{1}{k})$ as input. However, we can no longer use the approach of Section 2 to solve this if $|O_i| = O(\sqrt{|P|/k})$ due to the following: In Section 2, we first identified a ball that contains most of the points of $|O_i|$ while having a diameter bounded in terms of the *average* radius of O_i, using a statistical test that checks whether a guessed center-radius pair $(\mathbf{c}, r)$ defines a good ball. Suppose that $B = \textsc{Ball}_P(\mathbf{c}, r)$ is good. Our test may fail to identify that B is good if the variance of the number of points that are outside B and *bad* (i.e., the variance of $|(P \setminus \textsc{Ball}_P(\mathbf{c}, r)) \cap O_{\mathrm{bad}}|$) is significant compared to the number of *good* points that are inside the ball (i.e., $|\textsc{Ball}_P(\mathbf{c}, r) \cap O_{\mathrm{good}}|$), which may happen with a non-negligible probability if $|O_i| = O(\sqrt{|P|/k})$.

Nevertheless, we can estimate good centers for all clusters $|O_i|$ that are sufficiently large using the techniques of Section 2. Observe that we can always find some clusters that are sufficiently large; for example, we can always estimate a good center for the largest cluster since $|O_i| \geq |P|/k$ for that cluster, and consequently, the instance $(P, O_i, \varepsilon, \frac{1}{k})$ is nice. Intuitively, if we assume that we also correctly identify most of the points that belong to these clusters, we can remove all such points from the instance, and recursively apply the same technique on the remaining points to estimate centers for the remaining clusters. Of course, we cannot correctly identify all such points, but we show that we can assign some fraction of the points in the instance that are closest to the already estimated centers, and then recursively estimate centers for the remaining portion of the instance after removing the already assigned points. Note that the points we assign to estimated centers may not overlap with the points that truly belong to the respective clusters. However, using the fact that we assign a fraction of the points that are *closest* to the estimated centers and that the estimated centers are close to the true centroids of the respective clusters (considering only the remaining points), we develop an elaborate charging scheme to bound the total assignment cost of this approach. We describe our algorithm in detail in Appendix E.

As a final remark, note that our approach for the stochastic uniform noise model requires the minimum cluster size to be larger compared to the number of clusters k. This is to ensure that the sizes of clusters remain not too small even in the recursively solved instances.

References

Emmanuel Abbe, Elisabetta Cornacchia, Yuzhou Gu, and Yury Polyanskiy. Stochastic block model entropy and broadcasting on trees with survey. In *Conference on Learning Theory*, pages 1–25. PMLR, 2021.

Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 61–72. Ieee, 2017.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing*, 33(3):544–562, 2004.

Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *Advances in Neural Information Processing Systems 29*, pages 3216–3224, 2016.

Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 309–318. IEEE, 2010.

Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(5):643–660, 2012.

Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 737–756. SIAM, 2014.

Moses Charikar and Sudipto Guha. Improved combinatorial algorithms for facility location problems. *SIAM Journal on Computing*, 34(4):803–824, 2005.

Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 49–60. IEEE, 2017.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems 32*, pages 4751–4762, 2019.

Ahmed El Alaoui and Andrea Montanari. On the computational tractability of statistical estimation on amenable graphs. *Probability Theory and Related Fields*, 181(4):815–864, 2021.

Mohammad Esmaeili, Hussein Saad, and Aria Nosratinia. Community detection with side information via semidefinite programming. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 420–424. IEEE, 2019.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *Journal of the ACM (JACM)*, 50(6):795–824, 2003.

Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

Marc Lelarge and Léo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643. IEEE, 2019.

Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608. IEEE, 2016.

Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. In *International Colloquium on Automata, Languages, and Programming*, pages 77–88. Springer, 2011.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Jirı Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24 (1):61–84, 2000.

Arya Mazumdar and Barna Saha. Query complexity of clustering with side information. In *Advances in Neural Information Processing Systems 30*, pages 4682–4693, 2017.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pages 356–370. PMLR, 2014.

Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.

Tim Roughgarden. Beyond worst-case analysis. *Communications of the ACM*, 62(3):88–96, 2019.

Hussein Saad and Aria Nosratinia. Community detection with side information: Exact recovery under the stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5): 944–958, 2018.

Appendix A. Basic Results

In this section, we present some basic results that are frequently used in the subsequent sections.

We extensively use the following Chernoff bounds. The first one is the standard Chernoff bound and the second one is a slightly modified version for the case when events are not completely independent.

Lemma A.1 (Chernoff) *Fix a positive integer n. For each $i \in [n]$, let X_i be an independent Bernoulli random variable. Let $X = \sum_{i\in[n]} X_i$ and $\mu = \mathbb{E}[X]$. Then for all $\delta \in (0,1)$,*

$$\Pr\left[|X - \mu| > \delta\mu\right] < 2\exp(-\delta^2\mu/4),$$

and for all $\delta \geq 1$,

$$\Pr\left[X \geq (1+\delta)\cdot\mu\right] \leq \exp(-\delta\mu/4).$$

Lemma A.2 (Chernoff when events are dependent) *Let $X_1, \ldots, X_t$ be binary random variables and $p_1, \ldots, p_t \in [0,1]$ be real numbers such that $\Pr[X_i = 1 | X_1, \ldots, X_{i-1}] \geq p_i$ for all outcomes of $X_1, \ldots, X_{i-1}$. Let $X = \sum_{i=1}^t X_i$ and $\mu = \sum_{i=1}^t p_i$. Then for $\delta \in (0,1)$, we have*

$$\Pr[X < (1-\delta)\mu] \leq \exp(-\delta^2\mu/4).$$

Proof For $i = 1, \ldots, t$, let Y_i be independent random variables such that

$$\Pr[Y_i = 1] = \min_{x_1,\ldots,x_{i-1}} \Pr[X_i = 1 | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}].$$

Note that $\Pr[Y_i] \geq p_i$ for all $i = 1, \ldots, t$. Let $Y = \sum_i^t Y_i$. By coupling, we get $\Pr[X > t] \geq \Pr[Y > t]$ for any t, and hence we get

$$\Pr[X < (1-\delta)\mu] \leq \Pr[Y < (1-\delta)\mu] \leq \exp(-\delta^2\mu/4)$$

where the second inequality follows from Lemma A.1. ■

Estimating the centroid of a point based on a small size sample is one of the procedures that we often use in our algorithms. Here we provide a utility lemma that bounds the error of this estimation.

Lemma A.3 (Centroid estimation) *Let $S \subset \mathbb{R}^d$ be a finite set of points, and let $\mathbf{s}$ be its centroid. Let $m \geq 1$ be a positive integer. Let $T \subseteq S$ be a uniformly random subset of m points, and let $\mathbf{t}$ be its centroid. With probability at least $1-\delta$, we have*

$$\|\mathbf{s} - \mathbf{t}\|^2 \leq \frac{1}{\delta m}\left(\frac{\mathrm{cost}(S,\mathbf{s})}{|S|}\right).$$

Proof

We compute the variance of the estimator $\mathbf{t}$ and apply Markov's inequality. For a point $\mathbf{x} \in S$, let $\mathbb{I}_{\mathbf{x}}$ denote the indicator variable for the event $x \in T$. Fix some $i \in [d]$ and consider the deviation of $\mathbf{t}$ from $\mathbf{s}$ in the i-th dimension. Note that for any point $\mathbf{x} \in S$, it belongs T with probability $|T|/|S|$, and for two points $\mathbf{x}, \mathbf{y} \in S$, both of them belong to T with probability $(|T|/|S|) \cdot (|T|-1/|S|-1)$.

Observe that

$$\begin{aligned}(s_i - t_i)^2 &= \left(\tfrac{1}{|S|}\sum_{\mathbf{x}\in S} x_i - \tfrac{1}{|T|}\sum_{\mathbf{x}\in T} x_i\right)^2 \\ &= \sum_{\mathbf{x},\mathbf{y}\in S} x_i \cdot y_i \cdot \left(\tfrac{1}{|S|} - \tfrac{\mathbb{I}_{\mathbf{x}}}{|T|}\right)\cdot\left(\tfrac{1}{|S|} - \tfrac{\mathbb{I}_{\mathbf{y}}}{|T|}\right).\end{aligned}$$

Using the linearity of expectation and the fact that $\Pr[\mathbf{x} \in T] = |T|/|S|$, we thus get

$$\begin{aligned}\mathbb{E}[(s_i - t_i)^2] &= \sum_{\mathbf{x},\mathbf{y}\in S} x_i \cdot y_i \cdot \mathbb{E}\left[\tfrac{1}{|S|^2} - \tfrac{\mathbb{I}_{\mathbf{x}}}{|S|\cdot|T|} - \tfrac{\mathbb{I}_{\mathbf{y}}}{|S|\cdot|T|} + \tfrac{\mathbb{I}_{\mathbf{x}}\cdot\mathbb{I}_{\mathbf{y}}}{|T|^2}\right] \\ &= \sum_{\mathbf{x},\mathbf{y}\in S} x_i \cdot y_i \cdot \left(\mathbb{E}\left[\tfrac{\mathbb{I}_{\mathbf{x}}\cdot\mathbb{I}_{\mathbf{y}}}{|T|^2}\right] - \tfrac{1}{|S|^2}\right).\end{aligned}$$

Note that $\mathbf{x}$ and $\mathbf{y}$ are the same point, we have

$$\mathbb{E}\left[\tfrac{\mathbb{I}_{\mathbf{x}}\cdot\mathbb{I}_{\mathbf{y}}}{|T|^2}\right] = \tfrac{|T|}{|S|}\cdot\tfrac{1}{|T|^2} = \tfrac{1}{|S|\cdot|T|},$$

and, when $\mathbf{x}$ and $\mathbf{y}$ are different, we have that

$$\mathbb{E}\left[\tfrac{\mathbb{I}_{\mathbf{x}}\cdot\mathbb{I}_{\mathbf{y}}}{|T|^2}\right] = \tfrac{|T|}{|S|}\cdot\tfrac{(|T|-1)}{(|S|-1)}\cdot\tfrac{1}{|T|^2} = \tfrac{|T|-1}{|S|\cdot|T|\cdot(|S|-1)}.$$

Thus we conclude that

$$\begin{aligned}\mathbb{E}[(s_i - t_i)^2] &= \sum_{\mathbf{x},\mathbf{y}\in S,\mathbf{x}\neq\mathbf{y}} x_i \cdot y_i \cdot \left(\tfrac{|T|-1}{|S|\cdot(|S|-1)\cdot|T|} - \tfrac{1}{|S|^2}\right) + \sum_{\mathbf{x}\in S} x_i^2 \cdot \left(\tfrac{1}{|S|\cdot|T|} - \tfrac{1}{|S|^2}\right) \\ &= \tfrac{1}{|S|(|S|-1)}\left(\sum_{\mathbf{x},\mathbf{y}\in S,\mathbf{x}\neq\mathbf{y}} x_i \cdot y_i \cdot \left(\tfrac{|T|-1}{|T|} - \tfrac{|S|-1}{|S|}\right) + \sum_{\mathbf{x}\in S} x_i^2 \cdot \left(\tfrac{|S|-1}{|T|} - \tfrac{|S|-1}{|S|}\right)\right) \\ &= \tfrac{1}{|S|(|S|-1)}\left(\sum_{\mathbf{x}\in S} x_i^2 \cdot \left(\tfrac{|S|-1}{|T|} - \tfrac{|S|-1}{|S|}\right) - \sum_{\mathbf{x},\mathbf{y}\in S,\mathbf{x}\neq\mathbf{y}} x_i \cdot y_i \cdot \left(\tfrac{1}{|T|} - \tfrac{1}{|S|}\right)\right) j \\ &= \tfrac{1}{|S|(|S|-1)}\left(\tfrac{1}{|T|} - \tfrac{1}{|S|}\right)\left(\sum_{\mathbf{x}\in S} x_i^2\,(|S|-1) - \sum_{\mathbf{x},\mathbf{y}\in S,\mathbf{x}\neq\mathbf{y}} x_i \cdot y_i\right) \\ &= \tfrac{|S|-|T|}{|S|^2(|S|-1)|T|}\left(|S|\sum_{\mathbf{x}\in S} x_i^2 - \sum_{\mathbf{x},\mathbf{y}\in S} x_i \cdot y_i\right) \\ &= \tfrac{|S|-|T|}{|S|(|S|-1)|T|}\left(\sum_{\mathbf{x}\in S} x_i^2 - \tfrac{1}{|S|}\cdot\left(\sum_{\mathbf{x}\in S} x_i\right)^2\right). \qquad (6)\end{aligned}$$

Now, taking the summation over all dimensions and noting that $|S| - |T| \leq |S| - 1$, we get

$$\mathbb{E}[\|\mathbf{s} - \mathbf{t}\|^2] \leq \frac{1}{m}\left(\frac{\text{cost}(S, \mathbf{s})}{|S|}\right), \tag{7}$$

where we use that $|T| = m$. The proof now follows from Markov's inequality. ■

Next, we present a useful lemma due to Minsker (2015) that helps to boost the success probability of centroid estimates by combining multiple estimates. Namely, if we take the geometric median of a set of points $U \subset \mathbb{R}^d$ and consider any point $\mathbf{u} \in \mathbb{R}^d$ that is far from the median, then there is a large subset of points in U that are far from $\mathbf{u}$.

Lemma A.4 (Lemma 2.1 of Minsker (2015)) *Let $\mathbf{u}$ be a point in $\mathbb{R}^d$ and let $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_t\} \subset \mathbb{R}^d$ be a set of points. Let $\widehat{\mathbf{u}}$ be the geometric median of U. If $\|\mathbf{u} - \widehat{\mathbf{u}}\| > ((1-\alpha)/\sqrt{1-2\alpha}) \cdot r$, then there exists $V \subseteq U$ such that $|V| \geq \alpha|U|$ and $\|\mathbf{u} - \mathbf{u}_i\| > r$ for all $\mathbf{u}_i \in V$.*

Lemma A.4 implies that, if sufficiently many points in U are good estimates for a point $\mathbf{u}$, then so is the geometric median of U. In particular, substituting α such that $(1-\alpha)/\sqrt{1-2\alpha} = 2$, i.e., $\alpha \simeq 0.46$ in , we get the following corollary, which can be viewed as a generalization of the so-called median-trick to higher dimensions.

Corollary A.5 *Let $\mathbf{u}$ be a point in $\mathbb{R}^d$ and let $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_t\} \subset \mathbb{R}^d$ be a set of points. Let $\widehat{\mathbf{u}}$ be the geometric median of U. If more than $0.6|U|$ points $\mathbf{u}_i \in U$ satisfies $\|\mathbf{u}_i - \mathbf{u}\| \leq r$ for some point $\mathbf{u}$, then $\|\widehat{\mathbf{u}} - \mathbf{u}\| \leq 2 \cdot r$.*

Finally, the following lemma bounds the cost increase due to adding a small number of points to an existing set of points.

Lemma A.6 *Let Q be a set of points in $\mathbb{R}^d$, and let $S \subset Q$ such that $|Q \setminus S| \leq \gamma^2|Q|$ for some $1/2 \geq \gamma > 0$. Let $\mathbf{c} \in \mathbb{R}^d$ be any point and let $\mathbf{q} \in \mathbb{R}^d$ be the centroid of Q. Then we have that*

$$\text{cost}(Q, \mathbf{c}) \leq (1 + 4 \cdot \gamma)\left(\text{cost}(S, \mathbf{c}) + \text{cost}(Q, \mathbf{q})\right).$$

Proof Observe that for real numbers $a, b \in \mathbb{R}$, and $\delta \in (0, \frac{1}{2}]$, we have

$$\begin{aligned}(a+b)^2 &= a^2 + b^2 + 2ab \leq a^2 + b^2 + \delta \cdot a^2 + (1/\delta) \cdot b^2 \\ &\leq (1+\delta) \cdot a^2 + (1 + 1/\delta) \cdot b^2 \leq (1+\delta) \cdot a^2 + \frac{2}{\delta} \cdot b^2,\end{aligned}$$

which we refer to as the squared triangle inequality.

Let $T = Q \setminus S$. Let $\mathbf{s}$ and $\mathbf{t}$ be the centroids of S and T. Note that we also have $\|\mathbf{s} - \mathbf{q}\| = (|T|/|S|)\|\mathbf{t} - \mathbf{q}\|$. We thus have the following:

$$\begin{aligned}\text{cost}(Q, \mathbf{c}) &= \text{cost}(Q, \mathbf{q}) + |Q| \cdot \|\mathbf{c} - \mathbf{q}\|^2 \\ &\leq \text{cost}(Q, \mathbf{q}) + \frac{|S|}{1-\gamma^2} \cdot \left((1+\gamma)\|\mathbf{c} - \mathbf{s}\|^2 + \frac{2}{\gamma}\|\mathbf{s} - \mathbf{q}\|^2\right) \\ &= \text{cost}(Q, \mathbf{q}) + \frac{|S|}{1-\gamma^2} \cdot \left((1+\gamma)\|\mathbf{c} - \mathbf{s}\|^2 + \frac{2}{\gamma}\frac{|T|^2}{|S|^2}\|\mathbf{t} - \mathbf{q}\|^2\right) \\ &\leq \text{cost}(Q, \mathbf{q}) + (1 + 4 \cdot \gamma) \cdot |S| \cdot \|\mathbf{c} - \mathbf{s}\|^2 + 4 \cdot \gamma \cdot |T| \cdot \|\mathbf{t} - \mathbf{q}\|^2 \\ &\leq \text{cost}(Q, \mathbf{q}) + (1 + 4 \cdot \gamma) \cdot \text{cost}(S, \mathbf{c}) + 4 \cdot \gamma \cdot \text{cost}(Q, \mathbf{q}) \\ &= (1 + 4 \cdot \gamma)\left(\text{cost}(Q, \mathbf{q}) + \text{cost}(S, \mathbf{c})\right),\end{aligned}$$

where the first inequality follows from the squared triangle inequality and the second inequality follows since $\gamma \leq 1/2$. In the third inequality, we use that for a non-empty set of points $A \subset \mathbb{R}^d$ with centroid $\mathbf{a}$ and for any point $\mathbf{b} \in \mathbb{R}^d$, it holds that $\mathrm{cost}(A, \mathbf{b}) = \mathrm{cost}(A, \mathbf{a}) + |A| \cdot \|\mathbf{a} - \mathbf{b}\|^2 \geq |A| \cdot \|\mathbf{a} - \mathbf{b}\|^2$. ■

Appendix B. Balanced Adversarial Noise Model

In this section, we present our algorithm for the k-means problem in the balanced adversarial noise model and prove the following threorem which is the formal version of Theorem 1.1.

Theorem B.1 *There exists a deterministic algorithm that, given an instance P of the k-means problem in the balanced adversarial noise setting with the underlying optimal clusters $O_1, \ldots, O_k$ and the noise parameter $\varepsilon \in [0, 1]$, outputs centers $\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k$ such that*

$$\mathrm{cost}(P, \{\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k\}) \leq (1 + O(\varepsilon)) \cdot \mathrm{cost}(P, \{\mathbf{o}_1, \ldots, \mathbf{o}_k\}),$$

where $\mathbf{o}_i$ denote the centroid of O_i for each $i \in [k]$.

Note that when $\varepsilon \geq 1/4$, any constant factor approximation algorithm for k-means (without any side information) will satisfy the claimed statement. Thus, without loss of generality, we assume that $\varepsilon < 1/4$.

Our algorithm recovers one cluster at a time and is based on the following observations: In the adversarial noise model, for each $i \in [k]$, O'_i is obtained by removing some points from O_i and adding some other points to O_i. In the *balanced* adversarial model, we have that $|O_i \Delta O'_i| \leq \varepsilon |O_i|$, for each $i \in [k]$, so the number of points added to O_i and the number of points removed from O_i are both upper-bounded by $\varepsilon |O_i|$.

Now, fix some $i \in [k]$ and consider the set O_i of the points with true label i. If we remove a small number of points from O_i, the optimum center cannot move too far, so the new centroid is close to the original centroid. On the other hand, suppose that we add small number of points to O_i. The new points can be arbitrary far away, but we can remove such outliers by disregarding the distant points. Note that when we disregard distant points, we could end up ignoring some points that are in O_i, but by the previous argument on removing a small fraction of the points, ignoring a small fraction of points that were originally in O_i does not move the centroid by much.

We start by formalizing the robustness of the centroid with respect to removing points.

Lemma B.2 *Let S be a set of points whose centroid is $\mathbf{s}$. Let $T \subseteq S$ be any subset obtained by removing $\varepsilon |S|$ points for some $0 \leq \varepsilon \leq \frac{2}{3}$, and let $\mathbf{t}$ be the centroid of T. Then $\|\mathbf{s} - \mathbf{t}\| \leq \sqrt{2\varepsilon \cdot \mathrm{cost}(S, \mathbf{s})/|S|}$.*

Proof Let $T' = S \setminus T$ and let $\mathbf{t}'$ be the centroid of T'. Observe that $|T'| = \varepsilon |S|$. Since we work with Euclidean distances, we have

$$\begin{aligned}
\mathrm{cost}(S, \mathbf{s}) &= \mathrm{cost}(T, \mathbf{s}) + \mathrm{cost}(T', \mathbf{s}) \\
&= \mathrm{cost}(T, \mathbf{t}) + |T| \cdot \|\mathbf{s} - \mathbf{t}\|^2 + |T'| \cdot \|\mathbf{s} - \mathbf{t}'\|^2 + \mathrm{cost}(T', \mathbf{t}') \\
&\geq (1 - \varepsilon) \cdot |S| \cdot \|\mathbf{s} - \mathbf{t}\|^2 + \varepsilon \cdot |S| \cdot \|\mathbf{s} - \mathbf{t}'\|^2.
\end{aligned}$$

In the last inequality, we use that the costs are non-negative.

Note that $\mathbf{s}$ lies on the line segment between $\mathbf{t}$ and $\mathbf{t}'$ such that $\|\mathbf{s}-\mathbf{t}\| = \varepsilon \cdot \|\mathbf{t}-\mathbf{t}'\|$. Hence we have

$$\begin{aligned}\operatorname{cost}(S,\mathbf{s}) &\geq (1-\varepsilon)\cdot\varepsilon^2\cdot|S|\cdot\|\mathbf{t}-\mathbf{t}'\|^2 + \varepsilon\cdot(1-\varepsilon)^2\cdot|S|\cdot\|\mathbf{t}-\mathbf{t}'\|^2 \\ &= \varepsilon\cdot(1-\varepsilon)\cdot|S|\cdot\|\mathbf{t}-\mathbf{t}'\|^2,\end{aligned}$$

which yields that

$$\|\mathbf{s}-\mathbf{t}\| = \varepsilon\cdot\|\mathbf{t}-\mathbf{t}'\| \leq \varepsilon\cdot\sqrt{\frac{\operatorname{cost}(S,\mathbf{s})}{\varepsilon\cdot(1-\varepsilon)\cdot|S|}} = \sqrt{\frac{\varepsilon\cdot\operatorname{cost}(S,\mathbf{s})}{(1-\varepsilon)\cdot|S|}} \leq \sqrt{\frac{2\varepsilon\cdot\operatorname{cost}(S,\mathbf{s})}{|S|}},$$

where the last inequality uses that $\varepsilon \leq \frac{2}{3}$. ■

We now present the algorithm, which we refer to as OUTLIERREMOVAL, for approximating the center of some fixed cluster O which has been perturbed by adding and/or removing points. The algorithm uses a set of candidate centers that includes a center $\mathbf{o}^*$ such that $\operatorname{cost}(O,\mathbf{o}^*) \leq (1+\varepsilon)\cdot\operatorname{cost}(O,\mathbf{o})$ where $\mathbf{o}$ is the centroid of O. Such a set can be constructed as explained in Matoušek (2000) with polynomial size with respect to the input size.

The algorithm OUTLIERREMOVAL takes as input a set O' and ε, goes over each candidate center $\mathbf{c}$, and compute the cost of O' with respect to center $\mathbf{c}$ ignoring the farthest away $2\varepsilon|O'|$ points from $\mathbf{c}$. Out of all candidate centers, it outputs the one that gives the minimum cost. Algorithm 1 outlines these steps, and Lemma B.3 shows that it always outputs a good center for the considered cluster O, provided that $|O'\Delta O| \leq \varepsilon|O|$.

Algorithm 1: The outline of OUTLIERREMOVAL.

Input: Set of points O' and $\varepsilon \in [0, 1/4]$.
$\hat{\mathbf{o}} \leftarrow \infty$
cost $\leftarrow$ nil
for *each center* $\mathbf{c}$ *in the set of candidate centers* **do**
  $O'_{\mathbf{c}} \leftarrow O' \setminus \{2\varepsilon|O'|$ farthest points from $\mathbf{c}$ in $O'\}$
  if $\operatorname{cost}(O'_{\mathbf{c}},\mathbf{c}) <$ cost **then**
    cost $\leftarrow \operatorname{cost}(O'_c,\mathbf{c})$
    $\hat{\mathbf{o}} \leftarrow \mathbf{c}$
return $\hat{\mathbf{o}}$.

Let $\hat{\mathbf{o}}$ be the center returned by OUTLIERREMOVAL. Recall that we use $\mathbf{o}$ to denote the centroid of O, and that the set of candidate centers includes a center $\mathbf{o}^*$ such that $\operatorname{cost}(O,\mathbf{o}^*) \leq (1+\varepsilon)\cdot\operatorname{cost}(O,\mathbf{o})$. Moreover, we can also show that $|O'\setminus O| \leq 2\varepsilon|O'|$, implying that the center $\hat{\mathbf{o}}$ incurs a cost of at most $(1+\varepsilon)\cdot\operatorname{cost}(O,\mathbf{o})$ for its nearest $(1-2\varepsilon)\cdot|O'|$ points. On the other hand, this set of nearest $(1-2\varepsilon)\cdot|O'|$ points to $\hat{\mathbf{o}}$ must have a considerable overlap with O, and hence the fraction of the points outside the overlap must be small. As a result, by Lemma B.2, the centroid of these overlapping points must be close to both

1. the center returned by the algorithm $\hat{\mathbf{o}}$, and

2. the centroid $\mathbf{o}$ of O.

Hence $\hat{\mathbf{o}}$ and $\mathbf{o}$ must be close to each other. We formalize this intuition in Lemma B.3.

Lemma B.3 *Given O' and $\varepsilon \in [0, 1/4]$ such that $|O \Delta O'| \leq \varepsilon |O|$,* OUTLIERREMOVAL *returns a center $\hat{\mathbf{o}}$ such that $\|\hat{\mathbf{o}} - \mathbf{o}\| \leq 12\sqrt{\varepsilon \frac{\mathrm{cost}(O,\mathbf{o})}{|O|}}$. Consequently, we have $\mathrm{cost}(O, \hat{\mathbf{o}}) \leq (1 + O(\varepsilon)) \cdot \mathrm{cost}(O, \mathbf{o})$.*

Proof

In this proof, we use the following inequalities, which can be easily verified using that $\varepsilon \in [0, 1/4]$ and that $|O \Delta O'| \leq \varepsilon |O|$:

$$(1 - \varepsilon) \cdot |O| \leq |O'| \leq (1 + \varepsilon) \cdot |O|, \text{ and} \tag{8}$$

$$(1 - 2\varepsilon) \cdot |O'| \leq |O| \leq (1 + 2\varepsilon) \cdot |O'|. \tag{9}$$

Similarly, we also observe that

$$|O \cap O'| = |O'| - |O' \setminus O| \geq |O'| - \varepsilon |O| \geq (1 - 2\varepsilon) \cdot |O'|. \tag{10}$$

Let $\widehat{O}$ be the closest $(1 - 2\varepsilon) \cdot |O'|$ points in O' to $\hat{\mathbf{o}}$. We first prove that

$$\mathrm{cost}(\widehat{O}, \hat{\mathbf{o}}) \leq (1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o}). \tag{11}$$

To see this, let $\mathbf{o}^*$ be the candidate center such that $\mathrm{cost}(O, \mathbf{o}^*) \leq (1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o})$. Then we have that

$$\mathrm{cost}(O \cap O', \mathbf{o}^*) \leq \mathrm{cost}(O, \mathbf{o}^*) \leq (1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o}).$$

Moreover, by (10), O' contains at least $(1 - 2\varepsilon) \cdot |O'|$ points of O. Thus OUTLIERREMOVAL could have picked $\mathbf{o}^*$ together with some $(1 - 2\varepsilon) \cdot |O'|$ points such that the cost of the selected points with respect to $\mathbf{o}^*$ is at most $(1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o})$. Since the algorithm output the center with the minimum cost, we have $\mathrm{cost}(\widehat{O}, \hat{\mathbf{o}}) \leq (1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o})$.

We now bound $\mathrm{cost}(O, \hat{\mathbf{o}})$ by showing that $\|\mathbf{o} - \hat{\mathbf{o}}\|$ is small. Let $\tilde{O} = O \cap \widehat{O}$ and let $\tilde{\mathbf{o}}$ be the centroid of the points in $\tilde{O}$. Observe that $|\tilde{O}| \geq (1 - 4\varepsilon) \cdot |O'|$. This is because $\widehat{O}$ contains $(1 - 2\varepsilon) \cdot |O'|$ points and O' can have at most $\varepsilon |O| \leq 2\varepsilon |O'|$ points that does not belong to O. Thus, $|\widehat{O} \setminus \tilde{O}| / |\widehat{O}| \leq \frac{2\varepsilon}{1 - 2\varepsilon} \leq 4\varepsilon$ when $\varepsilon \leq 1/4$. Now, invoking Lemma B.2 with $S = \widehat{O}$ and $T = \tilde{O}$ yields

$$\begin{aligned}
\|\tilde{\mathbf{o}} - \hat{\mathbf{o}}\| &\leq \sqrt{8\varepsilon \frac{\mathrm{cost}(\widehat{O}, \hat{\mathbf{o}})}{|\widehat{O}|}} \\
&\leq \sqrt{8\varepsilon \frac{(1 + \varepsilon) \cdot \mathrm{cost}(O, \mathbf{o})}{(1 - 2\varepsilon) \cdot |O'|}} \\
&\leq \sqrt{8\varepsilon \frac{(1 + \varepsilon) \cdot \mathrm{cost}(O, o)}{(1 - 2\varepsilon) \cdot (1 - \varepsilon) \cdot |O|}} \\
&\leq 8\sqrt{\varepsilon \frac{\mathrm{cost}(O, o)}{|O|}}.
\end{aligned}$$

For the second inequality, we used (11) and that $\widehat{O}$ is a set of $(1-2\varepsilon)\cdot|O'|$ points. The third inequality follows from (8) and the last inequality is due to ε being at most $1/4$.

On the other hand, we have $|\widehat{O}| \geq (1-2\varepsilon)\cdot|O'| \geq (1-2\varepsilon)\cdot(1-\varepsilon)\cdot|O|$ where the last inequality follows from (8). Moreover, $\widehat{O}$ is a subset of $O \cup O'$ and $|O' \setminus O| \leq \varepsilon|O|$. Thus $\tilde{O} = O \cap \widehat{O}$ must contain at least $(1-2\varepsilon)\cdot(1-\varepsilon)\cdot|O| - \varepsilon|O|$ points. Using $\varepsilon \leq 1/4$, we therefore get that $\tilde{O}$ contains at least $(1-4\varepsilon)\cdot|O|$ points, which yields that $|O \setminus \tilde{O}|/|O| \leq 4\varepsilon$. Now invoking Lemma B.2 with $S = O$ and $T = \tilde{O}$, we get

$$\|\mathbf{o} - \tilde{\mathbf{o}}\| \leq \sqrt{8\varepsilon \frac{\mathrm{cost}(O,\mathbf{o})}{|O|}} \leq 4\sqrt{\varepsilon \frac{\mathrm{cost}(O,\mathbf{o})}{|O|}}.$$

By the triangle inequality, we thus have that

$$\|\mathbf{o} - \hat{\mathbf{o}}\| \leq \|\mathbf{o} - \tilde{\mathbf{o}}\| + \|\tilde{\mathbf{o}} - \hat{\mathbf{o}}\| \leq 12\sqrt{\varepsilon \frac{\mathrm{cost}(O,\mathbf{o})}{|O|}}.$$

Thus, we get that

$$\begin{aligned}
\mathrm{cost}(O,\hat{\mathbf{o}}) &= \mathrm{cost}(O,\mathbf{o}) + |O|\cdot\|\mathbf{o} - \hat{\mathbf{o}}\|^2 \\
&\leq \mathrm{cost}(O,\mathbf{o}) + |O|\cdot 144\cdot\varepsilon\frac{\mathrm{cost}(O,\mathbf{o})}{|O|} \\
&= (1+144\varepsilon)\cdot\mathrm{cost}(O,\mathbf{o}).
\end{aligned}$$

■

The proof of Theorem B.1 now follows from Lemma B.3 as we can separately invoke OUTLIERREMOVAL O'_i for each $i \in [k]$.

Appendix C. Details of Stochastic Noise Model

In this section, we present our algorithm for solving the stochastic noisy center estimation problem introduced in Section 2. Recall that our overall approach consists of two stages, which we describe in Appendix C.1 and Appendix C.3.

C.1. Find good balls for clusters (proof of Lemma 2.6)

In this section, we prove Lemma 2.6, which we restate below. As mentioned in Section 2, we show that we can identify good balls for a cluster using a statistical test to determine if a guessed (center, radius)-pair is good. For balls defined by all possible (center, radius) combinations, we essentially check whether the numbers of points in O' that fall outside the considered ball significantly deviates from the expected number of such points. Our analysis extensively uses Chernoff bounds to show that all comparisons work as expected with very high probability.

Lemma 2.6 *There exists an algorithm* GOODBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ be a nice *instance of the noisy center estimation problem, O_{noisy} be an (ε,δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. The algorithm* GOODBALL *takes $(P, O_{\mathrm{noisy}}, \varepsilon', \delta')$ as input and outputs a ball $B \in \mathcal{B}^{\mathrm{all}}$ such that $B \in \mathcal{B}^{\mathrm{good}}$ with probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$, where the probability is over the randomness of O_{noisy}.*

Algorithm 2: Outline of IsRadiusTooSmall.

$\hat{n}_O \leftarrow (|O_{\text{noisy}}| - \varepsilon' \cdot \delta' \cdot |P|)/(1-\varepsilon')$
$n_P^{\text{out}} \leftarrow |P \setminus \text{Ball}_P(c, r)|$
$m_O^{\text{out}} \leftarrow |O_{\text{noisy}} \setminus \text{Ball}_{O_{\text{noisy}}}(c, r)|$
$n^{\text{thresh}} \leftarrow (1-\varepsilon')\frac{\hat{n}_O}{4\log|P|} + \varepsilon' \cdot \delta' \cdot n_P^{\text{out}}$
if $m_O^{\text{out}} \geq n^{\text{thresh}}$ **then**
  return YES
else
  return NO

As stated earlier, the idea behind GoodBall is to use a statistical test to determine whether a considered candidate center $\mathbf{c}$ and a radius r defines a *good* ball. In this regard, we first define and analyze an auxiliary algorithm called IsRadiusTooSmall. Its goal is to decide whether a significant fraction of the points of O are outside a given ball $\text{Ball}_P(\mathbf{c}, r)$.

To do so, given $(P, O_{\text{noisy}}, \varepsilon', \delta')$, a candidate center $\mathbf{c}$, and a radius r, IsRadiusTooSmall checks if the number of points in $O_{\text{noisy}} \setminus \text{Ball}_P(\mathbf{c}, r)$ is significantly more than the number of such points we would expect if the number of points in O that are outside the considered ball is close to $\frac{1}{\log|P|}$ fraction of the size of O. Namely, we check whether $|O \setminus \text{Ball}_P(\mathbf{c}, r)| \simeq |O|/\log|P|$.

Let $n_P^{\text{out}} = |P \setminus \text{Ball}_P(\mathbf{c}, r)|$, $n_O^{\text{out}} = |O \setminus \text{Ball}_P(\mathbf{c}, r)|$, $m_O^{\text{out}} = |O_{\text{noisy}} \setminus \text{Ball}_P(\mathbf{c}, r)|$ be the numbers points in P, O, and O_{noisy} respectively that are outside the considered ball. Then we have $\mathbb{E}[m_O^{\text{out}}] = (1-\varepsilon)\cdot n_O^{\text{out}} + \varepsilon\cdot\delta\cdot n_P^{\text{out}}$. On the other hand, we have that $\mathbb{E}[|O_{\text{noisy}}|] = (1-\varepsilon)\cdot|O| + \varepsilon\cdot\delta\cdot|P|$, and hence we can approximate $|O|$ with $\hat{n}_O = (|O_{\text{noisy}}| - \varepsilon\cdot\delta\cdot|P|)/(1-\varepsilon)$. Thus, assuming $|O \setminus \text{Ball}_P(\mathbf{c}, r)| \simeq |O|/\log|P|$, we expect about $(1-\varepsilon)\hat{n}_O/\log|P| + \varepsilon\cdot\delta\cdot n_P^{\text{out}}$ points to be in $O_{\text{noisy}} \setminus \text{Ball}_P(\mathbf{c}, r)$. Using this observation, we outline IsRadiusTooSmall in Algorithm 2.

We have the following lemma regarding the performance of IsRadiusTooSmall.

Lemma C.1 *Let $(P, O, \varepsilon, \delta)$ be a* nice *instance of the noisy center estimation problem, O_{noisy} be an (ε, δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. Given $(P, O_{\text{noisy}}, \varepsilon', \delta')$, a candidate center $\mathbf{c}$, and a radius r, the algorithm* IsRadiusTooSmall *outputs* YES *or* NO *such that the following holds with probability at least $1 - \frac{1}{2|P|^2}\exp\left(-|O|^{0.2}\right)$ over the randomness of O_{noisy}:*

1. *If* IsRadiusTooSmall *outputs* NO*, then $|\text{Ball}_O(c, r)| \geq \left(1 - \frac{1}{\log|P|}\right)|O|$, and*
2. *If* IsRadiusTooSmall *outputs* YES*, then $|\text{Ball}_O(c, r)| \leq \left(1 - \frac{1}{16\log|P|}\right)|O|$.*

Proof

Recall that $\mathbb{E}[|O_{\text{noisy}}|] = (1-\varepsilon)|O| + \varepsilon\delta|P|$. Since the instance is nice, we have $1 - \varepsilon \geq \frac{1}{\log|P|}$ and $\delta \leq \frac{|O|^{1.1}}{|P|}$. Thus we have $(1-\varepsilon)|O| \geq \frac{|O|}{\log|P|}$ and $\delta|P| \leq |O|^{1.1}$. Furthermore, trivially $\varepsilon\delta|P| \geq 0$ and $(1-\varepsilon)|O| \leq |O| \leq |O|^{1.1}$. Thus we have $\frac{|O|}{\log|P|} \leq \mathbb{E}[|O_{\text{noisy}}|] \leq 2\cdot|O|^{1.1}$. Now

by Lemma A.1,

$$\Pr\left[\left||O_{\text{noisy}}| - \mathbb{E}[|O_{\text{noisy}}|]\right| \geq |O|^{0.95}\right] \leq 2 \cdot \exp\left(-\frac{|O|^{1.9}}{4 \cdot \mathbb{E}[|O_{\text{noisy}}|]}\right)$$
$$\leq \frac{1}{4|P|^2} \exp\left(-|O|^{0.2}\right). \tag{12}$$

Now let $\mathcal{E}_1$ be the event that $\left||O_{\text{noisy}}| - \mathbb{E}[|O_{\text{noisy}}|]\right| < |O|^{0.95}$. Recall that $n^{\text{thresh}} = \frac{1}{4\log|P|}(|O_{\text{noisy}}| - \varepsilon'\delta'|P|) + \varepsilon'\delta' n_P^{\text{out}}$. Thus, conditioned on $\mathcal{E}_1$, we have

$$\begin{aligned}
n^{\text{thresh}} &\leq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \left(1 - |O|^{-0.4}\right)^2 \varepsilon\delta|P|\right) + \left(1 + |O|^{-0.5}\right)^2 \varepsilon\delta n_P^{\text{out}} \\
&\leq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \left(1 - 2|O|^{-0.4}\right)\varepsilon\delta|P|\right) + \left(1 + 3|O|^{-0.5}\right)\varepsilon\delta n_P^{\text{out}} \\
&\leq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} + 5|O|^{-0.4}\varepsilon\delta|P| \\
&\leq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} + 5|O|^{-0.4}|O|^{1.1} \\
&< \frac{1}{4\log|P|}\left(\mathbb{E}[|O_{\text{noisy}}|] + |O|^{0.95} - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} + 5|O|^{0.7} \\
&\leq \frac{1}{4\log|P|}\left((1-\varepsilon)|O| + |O|^{0.95}\right) + \varepsilon\delta n_P^{\text{out}} + 5|O|^{0.7} \\
&\leq \frac{(1-\varepsilon)|O|}{2\log|P|} + \varepsilon\delta n_P^{\text{out}}.
\end{aligned}$$

Similarly, conditioned on $\mathcal{E}_1$, we also have,

$$\begin{aligned}
n^{\text{thresh}} &\geq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \left(1 + |O|^{-0.4}\right)^2 \varepsilon\delta|P|\right) + \left(1 - |O|^{-0.4}\right)^2 \varepsilon\delta n_P^{\text{out}} \\
&\geq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \left(1 + 3|O|^{-0.4}\right)\varepsilon\delta|P|\right) + \left(1 - 2|O|^{-0.4}\right)\varepsilon\delta n_P^{\text{out}} \\
&\geq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} - 5|O|^{-0.4}\varepsilon\delta|P| \\
&\geq \frac{1}{4\log|P|}\left(|O_{\text{noisy}}| - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} - 5|O|^{-0.4}|O|^{1.1} \\
&> \frac{1}{4\log|P|}\left(E[|O_{\text{noisy}}|] - |O|^{0.95} - \varepsilon\delta|P|\right) + \varepsilon\delta n_P^{\text{out}} - 5|O|^{0.7} \\
&\geq \frac{1}{4\log|P|}\left((1-\varepsilon)|O| - |O|^{0.95}\right) + \varepsilon\delta n_P^{\text{out}} - 5|O|^{0.7} \\
&\geq \frac{(1-\varepsilon)|O|}{8\log|P|} + \varepsilon\delta n_P^{\text{out}}.
\end{aligned}$$

Note that $\mathbb{E}[m_O^{\text{out}}] = (1-\varepsilon)n_O^{\text{out}} + \varepsilon\delta n_P^{\text{out}}$, and observe that $\mathbb{E}[m_O^{\text{out}}] \leq \mathbb{E}[O_{\text{noisy}}] \leq 2 \cdot |O|^{1.1}$. Consider the following two cases:

Case 1. Suppose that $\mathbb{E}[m_O^{\text{out}}] \leq |O|^{0.95}$. This implies, by Lemma A.1, that

$$\Pr\left[m_O^{\text{out}} - \mathbb{E}[m_O^{\text{out}}] \geq |O|^{0.95}\right] \leq \exp\left(-\frac{1}{4}|O|^{0.95}\right) \leq \frac{1}{4|P|^2}\exp\left(-|O|^{0.2}\right).$$

Case 2. Suppose instead that $\mathbb{E}[m_O^{\text{out}}] \geq |O|^{0.95}$. This implies, by Lemma A.1, that

$$\Pr\left[|m_O^{\text{out}} - \mathbb{E}[m_O^{\text{out}}]| \geq |O|^{0.95}\right] \leq \exp\left(-\frac{|O|^{1.9}}{4\mathbb{E}[m_O^{\text{out}}]}\right) \leq \frac{1}{4|P|^2}\exp\left(-|O|^{0.2}\right).$$

In either case, with probability at least $1 - \frac{1}{4|P|^2}\exp\left(-|O|^{0.2}\right)$, it holds that $|m_O^{\text{out}} - \mathbb{E}[m_O^{\text{out}}]| \leq |O|^{0.95}$. Let $\mathcal{E}_2$ denote this event.

By the union bound, both $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with probability at least $1 - \frac{1}{2|P|^2}\exp\left(-|O|^{0.2}\right)$, and conditioned on $\mathcal{E}_1$ and $\mathcal{E}_2$, we have the following:

1. If $|\textsc{Ball}_O(\mathbf{c}, r)| < \left(1 - \frac{1}{\log|P|}\right)|O|$, then

$$\begin{aligned} m_O^{\text{out}} &\geq \mathbb{E}[m_O^{\text{out}}] - |O|^{0.95} \\ &\geq \frac{(1-\varepsilon)|O|}{\log|P|} + \varepsilon\delta n_P^{\text{out}} - |O|^{0.95} \\ &\geq \frac{(1-\varepsilon)|O|}{\log|P|} + \varepsilon\delta n_P^{\text{out}} - |O|^{0.95} \\ &\geq \frac{(1-\varepsilon)|O|}{2\log|P|} + \varepsilon\delta n_P^{\text{out}} \\ &> n^{\text{thresh}}. \end{aligned}$$

   Thus IsRadiusTooSmall outputs YES.

2. If $|\textsc{Ball}_O(c, r)| > \left(1 - \frac{1}{16\log|P|}\right)|O|$, then

$$\begin{aligned} m_O^{\text{out}} &\leq \mathbb{E}[m_O^{\text{out}}] + |O|^{0.95} \\ &\leq (1-\varepsilon)n_O^{\text{out}} + \varepsilon\delta n_P^{\text{out}} + |O|^{0.95} \\ &\leq \frac{(1-\varepsilon)|O|}{16\log|P|} + \varepsilon\delta n_P^{\text{out}} + |O|^{0.95} \\ &\leq \frac{(1-\varepsilon)|O|}{8\log|P|} + \varepsilon\delta n_P^{\text{out}} \\ &< n^{\text{thresh}}. \end{aligned}$$

   Thus IsRadiusTooSmall outputs NO.

Finally, observe that the output of the algorithm violates the two claims of the lemma if it outputs NO when $|\textsc{Ball}_O(c, r)| < \left(1 - \frac{1}{\log|P|}\right)|O|$ or YES when $|\textsc{Ball}_O(c, r)| > \left(1 - \frac{1}{16\log|P|}\right)|O|$, which never happens if $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. The claim of the lemma now follows because both $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with probability at least $1 - \frac{1}{2|P|^2}\exp\left(-|O|^{0.2}\right)$. ■

Having access to ISRADIUSTOOSMALL, we now introduce an algorithm to find a good ball for the desired cluster. Namely, the algorithm, which refer to as GOODBALL, tries all possible (center, radius)-combinations and finds the combination with the smallest radius for which ISRADIUSTOOSMALL returns *NO*. The outline of GOODBALL is given in Algorithm 3. We now analyze its performance and show that it satisfies the requirements of Lemma 2.6.

Algorithm 3: Outline of GOODBALL.

```
r ← ∞
c ← Null
for c′ ∈ P do
    for r′ ∈ {‖p − c′‖ : p ∈ P} do
        if not IsRadiusTooSmall((P, O_noisy, ε′, δ′), c′, r′) then
            if r′ < r then
                c ← c′, r ← r′
            break
return c, r
```

Proof [Proof of Lemma 2.6]

Let $(\mathbf{c}, r)$ be the center and the radius returned by GOODBALL when it is called with P and O' as input. To prove Lemma 2.6, we show that $\text{BALL}_P(\mathbf{c}, r)$ is good with probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$. In other words, we need to show that the following two requirements are met with that probability:

1. $|\text{BALL}_O(c, r)| \geq \left(1 - \frac{1}{\log|O|}\right)|O|$, and
2. $r \leq 16 \cdot (\log^{0.5}|P|) \cdot \sqrt{\frac{\text{cost}(O,\mathbf{o})}{|O|}}$ where $\mathbf{o}$ is the centroid of O.

Note that GOODBALL makes at most $|P|^2$ call to ISRADIUSTOOSMALL, and each call fails with probability at most $\frac{1}{2|P|^2}\exp\left(-|O|^{0.2}\right)$. Therefore, by the union bound, with probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$, all calls to ISRADIUSTOOSMALL succeeds.

Suppose that all the calls to ISRADIUSTOOSMALL are successful. Observe that, for any candidate center, the ISRADIUSTOOSMALL call with minimum radius always returns YES and the ISRADIUSTOOSMALL call with maximum radius always returns NO. Thus, GOODBALL always returns a (center, radius) pair where the radius is finite. Since all calls to ISRADIUSTOOSMALL succeeded, we already have Item 1 above.

To prove Item 2, let $\mathbf{c}, r$ be the center and the radius returned by GOODBALL and let $\mathbf{c}'$ be the point in P that is closest to the centroid $\mathbf{o}$ of O. This implies that

$$2 \cdot \text{cost}(O, \mathbf{o}) \geq \text{cost}(O, \mathbf{o}) + \sum_{\mathbf{x} \in O} \|\mathbf{x} - \mathbf{o}\|^2 \geq \text{cost}(O, \mathbf{o}) + |O| \cdot \|\mathbf{o} - \mathbf{c}'\|^2 = \text{cost}(O, \mathbf{c}').$$

On the other hand, since r is the smallest radius for which ISRADIUSTOOSMALL returned NO, it must be the case that $|\text{BALL}_O(\mathbf{c}', r')| < \left(1 - \frac{1}{16\log|O|}\right)|O|$, for any $r' < r$. In particular, there

are at least $\frac{|O|}{16 \log |O|}$ points outside $\textsc{Ball}_O(\mathbf{c}', r/2)$, which implies that

$$\mathrm{cost}(O, \mathbf{c}') \geq \frac{r^2 |O|}{2^2 \cdot 16 \log |O|}.$$

Combining this together with the previous bound for $\mathrm{cost}(O, \mathbf{c}')$, we get that

$$r \leq 16 \cdot (\log^{0.5} |O|) \cdot r_{\mathrm{avg}},$$

which implies the requirement in Item 2.

■

C.2. Estimating center with high dimensional median trick (proof of Lemma 2.7)

In this section, we prove Lemma 2.7 which we restate below. Previously, in Section 2.1, we proved a weaker version of this result where the success probability was only constant. To prove the stronger version, we use the same techniques from Section 2.1 on random partitions of the input to obtain multiple centroid estimates and then combine it with a high dimensional median trick (Corollary A.5).

Lemma 2.7 *There exists an algorithm* CENTERINBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ *be a* nice *instance of the noisy center estimation problem,* O_{noisy} *be an* (ε, δ)*-noisy version of* O*, and* ε', δ' *be* close *approximations of* ε *and* δ*. The algorithm* CENTERINBALL *takes* $(P, O_{\mathrm{noisy}}, \varepsilon', \delta')$ *and* $B \subseteq P$ *as input and outputs a center* $\tilde{\mathbf{o}}_B$*. With probability at least* $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$ *over the randomness of* O_{noisy} *and the algorithm's internal random bits, it holds that*

$$\|\tilde{\mathbf{o}}_B - \mathrm{centroid}(O \cap B)\| \leq \frac{6 \cdot r_{\mathrm{avg}}}{\log^{0.5} |P|} \text{ for all } B \in \mathcal{B}^{\mathrm{good}}.$$

Before we proceed, recall that in the noisy cluster estimation problem, O_{noisy} is obtained as follows: First a random set P_{noisy} is constructed by including each point of P independently with probability ε. Then a set O_{good} is created by removing all points in P_{noisy} from O, and a set O_{bad} is constructed by including each point in P_{noisy} independently with probability δ. Finally O_{noisy} is defined as the union of O_{good} and O_{bad}. Also recall that, since B is a good ball, it has bounded diameter in terms of r_{avg}.

As mentioned before, our goal is to estimate a good center for $O \cap B$ using only a subset $Q \subseteq B$ of points. We now introduce the properties we require subsets Q to satisfy. Note that depending on the how much *bad* points we expect to see in the considered ball B, (i.e., the size of $O_{\mathrm{bad}} \cap B$, we have different requirements for the partitions. This is because we plan to use different algorithms depending on the situation. In any case, we first need the centroid of *good* points in the partition be be close to the centroid of $O \cap B$. Next, if we expect many *bad* points in B, then we plan to use the approach of Lemma 2.8, and hence we have the following requirements: Namely, the number of bad points in the partition should be close to the expected number of such points, the number of good points in the partition must not be too small, and the centroid of bad points in the partition must be close the centroid of the partition. On the other hand, if expect the number of bad points in B to be small, then we plan to use the algorithm for the adversarial setting, so we require the number bad points in the partition to be very small compared to the number of good points in the partition. We formalize these requirements in the following definition.

Definition C.2 *Let $B \in \mathcal{B}^{\text{good}}$ be a good ball for the considered instance, and let $Q \subseteq P$. We say that Q is good for a ball B and a fixed realization of O_{good} and O_{bad} if the following implications hold:*

- $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95} \Rightarrow$ *Properties 1-4 below hold, and*
- $\varepsilon\delta|B| \leq \frac{3}{2}|O|^{0.95} \Rightarrow$ *Properties 1 and 5 below hold.*

The properties are:

1. $\|\operatorname{centroid}(O_{\text{good}} \cap Q) - \operatorname{centroid}(O \cap B)\| \leq \frac{r_{\text{avg}}}{\log^{0.5}|P|}$.
2. $|O_{\text{bad}} \cap Q| \in \left[\left(1 - |O|^{-0.25}\right) \cdot \varepsilon\delta|Q|, \left(1 + |O|^{-0.25}\right) \cdot \varepsilon\delta|Q|\right]$.
3. $|O_{\text{good}} \cap Q| \geq |O|^{-0.2} \cdot \varepsilon\delta|Q|$.
4. $\|\operatorname{centroid}(O_{\text{bad}} \cap Q) - \operatorname{centroid}(Q)\| \leq \frac{r_{\text{avg}}}{|O|^{0.25}}$.
5. $|O_{\text{bad}} \cap Q| \leq \frac{|O_{\text{good}} \cap Q|}{\log^2|P|}$.

Now consider a set $Q \subseteq B$ that is good for some fixed ball $B \in \mathcal{B}^{\text{good}}$ satisfying $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$. As we see later, Properties 2-4 of the goodness definition ensures that the center $\tilde{\mathbf{o}}$ we estimated for $O_{\text{good}} \cap Q$ using Eq. (1) is close to $\operatorname{centroid}(O_{\text{good}} \cap Q)$, and the first property guarantees that $\tilde{\mathbf{o}}$, in turn, is close to $\operatorname{centroid}(O \cap B)$. This is formalized in Lemma C.3.

Lemma C.3 *Consider an instance $(P, O, \varepsilon, \delta)$ of the noisy center estimation problem. Let O_{noisy} be an (ε, δ)-noise added version of O and ε', δ' be parameters such that $(P, O_{\text{noisy}}, \varepsilon', \delta')$ is nice. Let $\mathcal{B}^{\text{good}}$ be the set of good balls for the considered instance, let $B \in \mathcal{B}^{\text{good}}$, and suppose that $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$. Let $Q \subseteq B$, $C = O \cap Q$, $C_{\text{noisy}} = O_{\text{noisy}} \cap Q$, and define $\alpha = \frac{\varepsilon'\delta'|Q|}{|C_{\text{noisy}}|}$. Let $\mathbf{c}$, $\mathbf{c}_n$, and $\mathbf{q}$, respectively, denote the centroids of C, C_{noisy}, and Q, and define $\hat{\mathbf{c}} = \frac{\mathbf{c}_n - \alpha \cdot \mathbf{q}}{1-\alpha}$ (or $\mathbf{0}$ if the fraction is undefined). If Q is good for B and O_{noisy}, Then we have*

$$\|\hat{\mathbf{c}} - \operatorname{centroid}(B \cap O)\| \leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.$$

Proof Let $C_{\text{bad}} = O_{\text{bad}} \cap Q$ and let $C_{\text{good}} = C_{\text{noisy}} \setminus C_{\text{bad}} = O_{\text{good}} \cap Q$. Note that since $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$ and Q is good for B, Properties 1-4 of Definition 2.5 hold. From Property 2 and Property 3, it follows that,

$$|C_{\text{noisy}}| = |C_{\text{good}}| + |C_{\text{bad}}| \geq (1 + |O|^{-0.2} - |O|^{-0.25}) \cdot \varepsilon\delta|Q| \geq (1 + \tfrac{1}{2}|O|^{-0.2}) \cdot \varepsilon\delta|Q|. \quad (13)$$

Since the instance is nice (Definition 2.1), we have that ε', δ' are $1 \pm |O|^{-0.4}$ approximations to ε, δ. Hence we have

$$|C_{\text{noisy}}| \geq \frac{1 + \frac{1}{2}|O|^{-0.2}}{(1 + |O|^{-0.4})^2} \cdot \varepsilon'\delta'|Q| \geq (1 + \tfrac{1}{4}|O|^{-0.2}) \cdot \varepsilon'\delta'|Q|. \quad (14)$$

This yields that $\alpha = \frac{\varepsilon'\delta'|Q|}{|C_{\text{noisy}}|} \leq \frac{1}{1+\frac{1}{4}|O|^{0.2}}$, implying

$$\frac{1}{1-\alpha} \leq \frac{1+\frac{1}{4}|O|^{-0.2}}{\frac{1}{4}|O|^{-0.2}} \leq 8|O|^{0.2}. \tag{15}$$

Also observe that

$$\begin{aligned}\big||C_{\text{bad}}| - \varepsilon'\delta'|Q|\big| &\leq \big||C_{\text{bad}}| - \varepsilon\delta|Q|\big| + \big|\varepsilon\delta|Q| - \varepsilon'\delta'|Q|\big| \\ &\leq |O|^{-0.25}\varepsilon\delta|Q| + |O|^{-0.5}\varepsilon\delta|Q| \\ &\leq 2|O|^{-0.25}\varepsilon\delta|Q|, \end{aligned} \tag{16}$$

where the first line is due to the triangle inequality and the second line uses Property 2 of Definition C.2 and that ε' and δ' are $\left(1 \pm |O|^{-0.4}\right)$ approximations for ε and δ respectively.

Now, observe that the sets C_{good} and C_{bad} are analogous to sets O_{good} and O_{bad} in Eq. (1). We first write the centroid $\mathbf{c}_n$ of C_{noisy} as a linear combination of centroids of C_{good} and C_{bad}.

$$\begin{aligned}\mathbf{c}_n &= \frac{1}{|C_{\text{noisy}}|}\sum_{\mathbf{x}\in C_{\text{noisy}}} \mathbf{x} \\ &= \frac{1}{|C_{\text{noisy}}|}\left(\sum_{\mathbf{x}\in C_{\text{good}}} \mathbf{x} + \sum_{\mathbf{x}\in C_{\text{bad}}} \mathbf{x}\right) \\ &= \frac{|C_{\text{good}}|}{|C_{\text{noisy}}|}\cdot\left(\frac{1}{|C_{\text{good}}|}\cdot\sum_{\mathbf{x}\in C_{\text{good}}} \mathbf{x}\right) + \frac{|C_{\text{bad}}|}{|C_{\text{noisy}}|}\cdot\left(\frac{1}{|C_{\text{bad}}|}\cdot\sum_{\mathbf{x}\in C_{\text{bad}}} \mathbf{x}\right).\end{aligned}$$

Denoting $\mathbf{c}_g = \frac{1}{|C_{\text{good}}|}\cdot\sum_{\mathbf{x}\in C_{\text{good}}} \mathbf{x}$ and $\mathbf{c}_b = \frac{1}{|C_{\text{bad}}|}\cdot\sum_{\mathbf{x}\in C_{\text{bad}}} \mathbf{x}$, we thus get

$$\mathbf{c}_n = \frac{|C_{\text{good}}|}{C_{\text{noisy}}}\mathbf{c}_g + \frac{|C_{\text{bad}}|}{C_{\text{noisy}}}\mathbf{c}_b.$$

Define $\mathbf{c}' = (1-\alpha)\mathbf{c}_g + \alpha\mathbf{c}_b$.

Since Q is good for B, by Definition C.2, we have that $\|\mathbf{c}_g - \mathbf{c}\| \leq \frac{r_{\text{avg}}}{\log^{0.5}|P|}$ and that $\|\mathbf{c}^b - \mathbf{q}\| \leq \frac{r_{\text{avg}}}{|O|^{0.25}}$. Thus we can estimate $\mathbf{c}$ with $(\mathbf{c}' - \alpha\mathbf{q})/(1-\alpha)$ if we know $\mathbf{c}'$.

We now show that $\mathbf{c}_n$ is very close to $\mathbf{c}'$. Observe that $\mathbf{c}_n$ and $\mathbf{c}'$ both lie on the same line segment between $\mathbf{c}_g$ and $\mathbf{c}_b$. The point $\mathbf{c}_n$ is $\frac{|C_{\text{bad}}|}{|C_{\text{noisy}}|}\cdot\|\mathbf{c}_g - \mathbf{c}_b\|$ away from $\mathbf{c}_g$ while the point $\mathbf{c}'$ is $\alpha\cdot\|\mathbf{c}_g - \mathbf{c}_b\|$ is distance away from $\mathbf{c}_g$. Consequently, it holds that

$$\begin{aligned}\|\mathbf{c}_n - \mathbf{c}'\| &= \left|\frac{|C_{\text{bad}}|}{|C_{\text{noisy}}|} - \alpha\right|\cdot\|\mathbf{c}_g - \mathbf{c}_b\| \\ &= \left|\frac{|C_{\text{bad}}|}{|C_{\text{noisy}}|} - \frac{\varepsilon'\delta'|Q|}{|C_{\text{noisy}}|}\right|\cdot\|\mathbf{c}_g - \mathbf{c}_b\| \\ &\leq \frac{2\varepsilon\delta|Q|}{|C_{\text{noisy}}|\cdot|O|^{0.25}}\cdot\|\mathbf{c}^g - \mathbf{c}^b\| \\ &\leq \frac{64(\log^{0.5}|P|)\cdot r_{\text{avg}}}{|O|^{0.25}}.\end{aligned}$$

The first inequality follows from Eq. (16) and the second one uses Eq. (14) together with the fact that both $\mathbf{c}_b$ and $\mathbf{c}_g$ are in B. Note that B's diameter is bounded by $32 \cdot (\log^{0.5}|P|) \cdot r_{\text{avg}}$ since $B \in \mathcal{B}^{\text{good}}$.

We thus conclude that

$$\begin{aligned}
\|\hat{\mathbf{c}} - \mathbf{c}\| &= \left\| \frac{\mathbf{c}_n - \alpha\mathbf{q}}{1-\alpha} - \mathbf{c} \right\| \\
&= \frac{1}{1-\alpha} \|\mathbf{c}_n - \alpha\mathbf{q} - (1-\alpha)\mathbf{c}\| \\
&= \frac{1}{1-\alpha} \left\|\mathbf{c}_n - \mathbf{c}' + \mathbf{c}' - \alpha\mathbf{q} - (1-\alpha)\mathbf{c}\right\| \\
&= \frac{1}{1-\alpha} \left\|\mathbf{c}_n - \mathbf{c}' + ((1-\alpha)\mathbf{c}_g + \alpha\mathbf{c}_b) - \alpha\mathbf{q} - (1-\alpha)\mathbf{c}\right\| \\
&\leq \frac{1}{1-\alpha} \left(\|\mathbf{c}_n - \mathbf{c}'\| + (1-\alpha) \cdot \|\mathbf{c}_g - \mathbf{c}\| + \alpha \cdot \|\mathbf{c}_b - \mathbf{q}\|\right) \\
&\leq \frac{1}{1-\alpha} \|\mathbf{c}_n - \mathbf{c}'\| + \|\mathbf{c}_g - \mathbf{c}\| + \frac{\alpha}{1-\alpha} \cdot \|\mathbf{c}_b - \mathbf{q}\|. \\
&\leq 8|O|^{0.2} \cdot \frac{64 \cdot (\log^{0.5}|P|) \cdot r_{\text{avg}}}{|O|^{0.25}} + \frac{r_{\text{avg}}}{\log^{0.5}|P|} + |O|^{0.2} \cdot \frac{r_{\text{avg}}}{|O|^{0.25}} \\
&\leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.
\end{aligned}$$

The second to last inequality uses that $|P|$ sufficiently large and that $|O| \geq \log^{100}|P|$. ■

The key to prove Lemma 2.7 is to find a large collection of sets Q such that, for any good ball B, we have many good sets for B in the collection. To show that we can find such a collection with high probability, we proceed by first defining a collection of sets that is nice for a fixed good ball B.

Definition C.4 *Let $\mathcal{Q} = \{Q_1, \ldots, Q_t\}$ be a collection of at least $|O|^{0.3}$ sets $Q_j \subseteq P$. We say that $\mathcal{Q}$ is* nice *for a ball $B \in \mathcal{B}^{\text{good}}$, if there are at least $\frac{3}{4}t$ sets $Q \in \mathcal{Q}$ that are* good *for B.*

The next lemma guarantees that we can find a *nice* collection of sets with high probability for any ball $B \in \mathcal{B}^{\text{good}}$. The candidate algorithm that we use to prove Lemma C.5 simply partitions a given ball into $\Theta(t \log|P|)$ sets uniformly at random, and then return the first t subsets of the partition, where $t = \Theta(|O|^{0.4})$. However, due to the dependencies involved, the analysis is rather technical, and we defer its proof to Appendix C.3.

Lemma C.5 *There exists an algorithm* RANDOMPARTITION *such that the following holds: Let $(P, O, \varepsilon, \delta)$ be an instance of the noisy center estimation problem, let O_{noisy} be an (ε, δ)-noise added version of O, and let ε', δ' be parameters such that $(P, O_{\text{noisy}}, \varepsilon', \delta')$ is nice.*

RANDOMPARTITION *takes $(P, O_{\text{noisy}}, \varepsilon', \delta')$ and a ball $B \in \mathcal{B}^{\text{all}}$ as input and outputs a collection $\mathcal{Q} = \{Q_1, \ldots, Q_t\}$ of subsets of B where $t \geq |O|^{0.4}$. If $B \in \mathcal{B}^{\text{good}}$ then $\mathcal{Q}$ is* nice *for B with probability at least $1 - \frac{1}{2|P|^2}\exp\left(-|O|^{0.2}\right)$.*

We now use Lemma C.5, Lemma C.3, and Corollary A.5 to prove Lemma 2.7.

Proof (of Lemma 2.7)

Algorithm 4: Outline of CENTERINBALL.

1 $\mathcal{Q} \leftarrow \text{RANDOMPARTITION}(P, O_{\text{noisy}}, \varepsilon', \delta', B)$
2 $C \leftarrow \emptyset$
3 **for** *each* $Q \in \mathcal{Q}$ **do**
4 **if** $\varepsilon'\delta'|B| \leq \left(\frac{|O_{\text{noisy}}| - \varepsilon'\delta'|P|}{1-\varepsilon'}\right)^{0.95}$ **then**
5 $\mathbf{c} \leftarrow \text{OUTLIERREMOVAL}\left(O_{\text{noisy}} \cap Q, \frac{1}{\log|P|}\right)$
6 **else**
7 $\mathbf{c} \leftarrow$ the estimate from Lemma C.3 for Q and B.
8 $C \leftarrow C \cap \{\mathbf{c}\}$
9 **return** geometric median of C

Let CENTERINBALL be the algorithm outlined in Algorithm 4. We show that CENTERINBALL satisfies the claim of Lemma 2.7 with the required success probability probability.

Let $\mathcal{B}^{\text{good}}_{\text{low}} \subseteq \mathcal{B}^{\text{good}}$ be the set of good balls B such that $\varepsilon\delta|B| \leq \frac{3}{2}|O|^{0.95}$, and let $\mathcal{B}^{\text{good}}_{\text{high}} \subseteq \mathcal{B}^{\text{good}}$ be the set of good balls B such that $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$. Let $\mathcal{Q}_B$ be the collection of subsets of B chosen in Line 1. Note that $|\mathcal{Q}_B| = |O|^{0.3}$. Let C_B be the state of set C at the end of the algorithm. Let $\mathcal{E}_B$ be the event that $\mathcal{Q}_B$ is *nice* for B, i.e., the event that at least $\frac{3}{4}|\mathcal{Q}_B|$ sets in $\mathcal{Q}_B$ are *good* for B.

First, suppose that $B \in \mathcal{B}^{\text{good}}_{\text{low}}$ and consider a set $Q \in \mathcal{Q}_B$ that is good for B. By Property 1 of Definition C.2, we have that

$$\|\operatorname{centroid}(O_{\text{good}} \cap Q) - \operatorname{centroid}(O \cap B)\| \leq \frac{r_{\text{avg}}}{\log^{0.5}|P|}.$$

And by Property 5 of Definition C.2, we have that $|O_{\text{bad}} \cap Q| \leq \frac{|O_{\text{good}} \cap Q|}{\log^2|P|}$. Moreover, note that

$$\operatorname{cost}(O_{\text{good}} \cap Q, \operatorname{centroid}(O_{\text{good}} \cap Q)) \leq 32^2 \cdot |O_{\text{good}} \cap Q| \cdot (\log|P|) \cdot r^2_{\text{avg}}$$

as B's diameter is bounded. Let $\hat{\mathbf{q}}$ be the center returned by $\text{OUTLIERREMOVAL}\left(O_{\text{noisy}} \cap Q, \frac{1}{\log^2|P|}\right)$. Thus by Lemma B.3, we get that

$$\begin{aligned}
\|\hat{\mathbf{q}} - \operatorname{centroid}(O_{\text{good}} \cap Q)\| &\leq 12\left(\frac{2}{\log^2|P|} \cdot \frac{\operatorname{cost}(O_{\text{good}} \cap Q, \operatorname{centroid}(O_{\text{good}} \cap Q))}{|O_{\text{good}} \cap Q|}\right)^{0.5} \\
&\leq \frac{12 \cdot 32\sqrt{2} \cdot r_{\text{avg}}}{\log|P|} \\
&\leq \frac{r_{\text{avg}}}{\log^{0.5}|P|},
\end{aligned}$$

and by the triangle inequality, we conclude that

$$\begin{aligned}
\|\hat{\mathbf{q}} - \operatorname{centroid}(O \cap B)\| &\leq \|\hat{\mathbf{q}} - \operatorname{centroid}(O_{\text{good}} \cap B)\| + \|\operatorname{centroid}(O_{\text{good}} \cap B) - \operatorname{centroid}(O \cap B)\| \\
&\leq \frac{2 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.
\end{aligned}$$

Now suppose that $B \in \mathcal{B}^{\text{good}}_{\text{high}}$ and again consider a set $Q \in \mathcal{Q}_B$ that is good for B. Note that in this case, Properties 1-4 of Definition C.2 hold for Q (because $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$), and the center $\hat{\mathbf{q}}$ estimated as in Lemma C.3 satisfies

$$\|\hat{\mathbf{q}} - \text{centroid}(O \cap B)\| \leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.$$

To conclude the proof, observe the following: From Eq. (12), we have that with probability at least $1 - \frac{1}{4|P|^2}\exp(-|O|^{0.2})$, $\Pr\left[||O_{\text{noisy}}| - \mathbb{E}[|O_{\text{noisy}}|]| \geq |O|^{0.95}\right]$. Moreover, we have that ε' and δ' are $1 \pm |O|^{-0.4}$ approximations to ε and δ respectively. With this, it is easy to verify that $\varepsilon'\delta'|B| \leq |O|^{0.95} \Rightarrow \varepsilon\delta|P| \geq \frac{1}{2}|O|^{0.95}$ and $\varepsilon'\delta'|B| < |O|^{0.95} \Rightarrow \varepsilon\delta|P| \leq \frac{3}{2}|O|^{0.95}$.

Therefore, in any case, if $Q \in \mathcal{Q}_B$ is good for B, the centroid estimate $\hat{\mathbf{q}}$ found by the algorithm CENTERINBALL at Line 5 or Line 7 satisfies

$$\|\hat{\mathbf{q}} - \text{centroid}(O \cap B)\| \leq \frac{3 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.$$

When $\mathcal{E}_B$ happens, there are $\frac{3}{4}|\mathcal{Q}_B|$ sets in $\mathcal{Q}_B$ that are good for B, and hence, by Corollary A.5, the median $\tilde{\mathbf{o}}_B$ of C_B satisfies:

$$\|\tilde{\mathbf{o}}_B - \text{centroid}(O \cap B)\| \leq \frac{6 \cdot r_{\text{avg}}}{\log^{0.5}|P|}.$$

Note that for a fixed ball $B \in \mathcal{B}^{\text{good}}$, $\mathcal{E}_B$ holds with probability at least $1 - \frac{1}{2|P|^2}\exp\left(|O|^{-0.2}\right)$. Thus, by the union bound over all good balls, we have that

$$\|\tilde{\mathbf{o}}_B - O \cap B\| \leq \frac{6 \cdot r_{\text{avg}}}{\log^{0.5}|P|}$$

for all balls in $\mathcal{B}^{\text{good}}$ with probability at least $1 - \frac{1}{2}\exp\left(|O|^{-0.2}\right)$ as $|\mathcal{B}^{\text{good}}| \leq |\mathcal{B}^{\text{all}}| \leq |P|^2$. ■

C.3. The proof of Lemma C.5

We now prove Lemma C.5 by showing that a random partition yields a nice collection with high probability.

Let $t = 2 \cdot \left(\frac{|O_{\text{noisy}}| - \varepsilon'\delta'|P|}{1-\varepsilon'}\right)^{0.4}$ and $s = \log^3|P|$. From Eq. (12), we have that with probability at least $1 - \frac{1}{4|P|^2}\exp(-|O|^{0.2})$, $\Pr\left[||O_{\text{noisy}}| - \mathbb{E}[|O_{\text{noisy}}|]| \geq |O|^{0.95}\right]$. Let $\mathcal{E}_{\text{O}}$ be this event. Note that we have that ε' and δ' are $1 \pm |O|^{-0.4}$ approximations to ε and δ respectively. Thus, conditioned on $\mathcal{E}_{\text{O}}$, it is easy to verify that $|O|^{0.4} \leq t \leq 4 \cdot |O|^{0.4}$.

We partition B into disjoint sets $\mathcal{P} = \{P_1, \ldots, P_{2 \cdot t \cdot s}\}$ as follows: For each point $\mathbf{p} \in P$, put $\mathbf{p}$ in a uniformly random set in $\mathcal{P}$. Then, let $\mathcal{Q} = \{P_1, \ldots, P_t\}$ be the first t sets in $\mathcal{P}$.

In the following analysis, we assume that $B \in \mathcal{B}^{\text{good}}$ and that $|P|$ is sufficiently large. Denote by $\mathcal{Q}^{\text{good}} \subseteq \mathcal{Q}$ the collection of sets that are good for B. To prove Lemma C.5, we show that, with probability at least $1 - \exp\left(-|O|^{0.2}\right)$, $|\mathcal{Q}^{\text{good}}| \geq \frac{3}{4}t$.

Let $O_B = O \cap B$. Since B is good, we have $|O_B| \geq \left(1 - \frac{1}{\log|P|}\right)|O| \geq \frac{|O|}{2}$. We consider a set $Q \in \mathcal{P}$ to be good with respect to size if both $|Q|$ and $|Q \cap O_B|$ are close to their expectations.

Definition C.6 *We say that a set $Q \in \mathcal{P}$ is* size-wise good *for B if*

$$|Q| \in \left[\frac{1}{4}\cdot\frac{|B|}{s\cdot t},\ \frac{3}{4}\cdot\frac{|B|}{s\cdot t}\right] \text{ and } |Q\cap O_B| \in \left[\frac{1}{4}\cdot\frac{|O_B|}{s\cdot t},\ \frac{3}{4}\cdot\frac{|O_B|}{s\cdot t}\right].$$

Using standard concentration bounds, it follows that, with high probability, all sets in $\mathcal{P}$ are size-wise good.

Lemma C.7 *With probability at least $1-\exp\left(-|O|^{0.5}\right)$, all $Q\in\mathcal{P}$ are size-wise good for B.*

Proof Fix some $Q\in\mathcal{P}$, and let $\mathcal{E}_Q$ be the event that Q is not size-wise good for B. Let $X=|Q|$. We have $\mathbb{E}[X]=\frac{|B|}{2\cdot s\cdot t}\geq\frac{|O_B|}{2\cdot s\cdot t}\geq\frac{|O|}{8\cdot s\cdot t}\geq\frac{|O|^{0.6}}{4\log^3}$, and by Chernoff bounds, we have

$$\Pr\left[|X-\mathbb{E}[X]|\geq\frac{1}{2}\mathbb{E}[X]\right]\leq 2\cdot\exp\left(-\frac{1}{16}\mathbb{E}[X]\right)=2\cdot\exp\left(-\frac{|O|^{0.6}}{64\log^3|P|}\right).$$

Similarly, setting $Y=|Q\cap O_B|$, we get $\mathbb{E}[Y]=\frac{O_B}{2\cdot s\cdot t}\geq\frac{|O|^{0.6}}{4\log^3|P|}$ as before. And by the same Chernoff bound above, we conclude that

$$\Pr\left[|Y-\mathbb{E}[Y]|>\frac{1}{2}\mathbb{E}[Y]\right]\leq 2\cdot\exp\left(-\frac{|O|^{0.6}}{64\log^3|P|}\right).$$

Thus by the union bound, we have $\Pr[\mathcal{E}_Q]\leq 4\exp(-|O|^{0.55})$, and

$$\Pr[\cup_{Q\in\mathcal{P}}\mathcal{E}_Q]\leq 8\cdot|\mathcal{P}|\cdot\exp\left(-|O|^{0.55}\right)=16\cdot s\cdot t\cdot\exp\left(-|O|^{0.55}\right)\leq\exp\left(-|O|^{0.5}\right).$$

■

As we see later, Lemma C.7 allows us to prove that the centroid of $O\cap P_j$ is close to the centroid of $B\cap O$ with good probability for each $j\leq t$. This is because, conditioned on $P_1,\ldots,P_{j-1}$ and that they are size-wise good, we can view P_j as a uniformly random sample from the remaining points. Due to $P_1,\ldots,P_{j-1}$ being size-wise good, there are many remaining points in both B and O_B. Consequently, the centroids of the remaining points in B and O_B are close the centroids of B and O_B respectively.

Let $\mathcal{E}^{size}$ be the event that P_j is size-wise good for B for all $P_j\in\mathcal{P}$. For each $j\leq t$, we now show that, conditioned on the choices of $P_1,\ldots,P_{j-1}$ and $\mathcal{E}^{\text{size}}$, P_j is good for B as defined in Definition C.2 with a good probability.

Lemma C.8 *Let $\mathcal{E}_j$ denote the event that P_j is good for B. For each $j=1,\ldots,t$, we have that $\Pr[\mathcal{E}_j|P_1,\ldots,P_{j-1},\mathcal{E}^{\text{size}}]\geq 0.9$.*

Proof

Let $\mathcal{E}_j^1,\mathcal{E}_j^2,\mathcal{E}_j^3,\mathcal{E}_j^4$ and $\mathcal{E}_j^5$ be the respective events that the Properties 1-5 of Definition C.2 does not hold when Q is set to P_j. Namely, they are events that the following properties hold:

1. $\|\operatorname{centroid}(O_{\text{good}}\cap P_j)-\operatorname{centroid}(O_B)\|\leq\frac{r_{\text{avg}}}{\log^{0.5}|P|}$.
2. $|O_{\text{bad}}\cap P_j|\in\left[\left(1-|O|^{-0.25}\right)\cdot\varepsilon\delta|P_j|,\left(1+|O|^{-0.25}\right)\cdot\varepsilon\delta|P_j|\right]$.

3. $|O_{\text{noisy}} \cap P_j| \geq (1 + |O|^{-0.2}) \cdot \varepsilon\delta|P_j|$.
4. $\|\operatorname{centroid}(O_{\text{bad}} \cap P_j) - \operatorname{centroid}(P_j)\| \leq \frac{r_{\text{avg}}}{|O|^{0.25}}$.
5. $|O_{\text{bad}} \cap P_j| \leq \frac{|O_{\text{good}} \cap P_j|}{\log^2 |P|}$.

Conditioned on $\mathcal{E}^{\text{size}}$ and any outcome of $P_1, \ldots, P_{j-1}$ that is compatible with $\mathcal{E}^{\text{size}}$, we show the following:

1. If $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$, then $\Pr[\mathcal{E}_j^i | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{1}{\log |P|}$ for $i = 1, 2, 3,$ and, $4,$ and hence
$$\Pr[\mathcal{E}_j^1 \cup \mathcal{E}_j^2 \cup \mathcal{E}_j^3 \cup \mathcal{E}_j^4 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{4}{\log |P|} \leq 0.1.$$
2. If $\varepsilon\delta|B| \leq \frac{3}{2}|O|^{0.95}$, then $\Pr[\mathcal{E}_j^i | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{1}{\log |P|}$ for $i = 1$ and 5, and hence
$$\Pr[\mathcal{E}_j^1 \cup \mathcal{E}_j^5 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{2}{\log |P|} \leq 0.1.$$

One important observation we use in the proof is that for any subset Q of B with centroid $\mathbf{q}$, the average radius $\sqrt{\operatorname{cost}(Q, \mathbf{q})/|Q|}$ is at most $32 \cdot (\log^{0.5} |P|) \cdot r_{\text{avg}}$ as the diameter of B is bounded by that quantity.

Suppose that $\mathcal{E}^{\text{size}}$ holds, and fix any choices of $P_1, \ldots, P_{j-1}$ that is compatible with $\mathcal{E}^{\text{size}}$. Regardless of the value of $\varepsilon\delta|B|$, we have the following:

Bounding $\Pr[\mathcal{E}_j^1 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}]$

Let $O_B^{\text{rem}} = O_B \setminus (\cup_{\ell=1}^{j-1} P_\ell)$. Since $j \leq t$ and we are conditioning on $\mathcal{E}^{size}$, we have that $|O_B^{\text{rem}}| \geq |O_B| - 3 \cdot (j-1) \cdot \frac{|O_B|}{4 \cdot s \cdot t} \geq \left(1 - \frac{1}{s}\right) |O_B|$. Noting that $s = \log^3 |P|$, from Lemma B.2, we get that

$$\|\operatorname{centroid}(O_B^{\text{rem}}) - \operatorname{centroid}(O_B)\| \leq \sqrt{\frac{2}{\log^3 |P|}} \cdot 32 \cdot (\log^{0.5} |P|) \cdot r_{\text{avg}} \leq \frac{r_{\text{avg}}}{3 \log^{0.5} |P|}.$$

Furthermore, since we are conditioning on $\mathcal{E}^{size}$, we know that $O_B \cap P_j$ is a uniformly random subset of O_B^{rem} of size at least $\frac{|O_B|}{4 \cdot s \cdot r} \geq \frac{|O|}{8 \cdot |O|^{0.4} \log^3 |P|} \geq |O|^{0.6} \geq \log^6 |P|$. From Lemma A.3, we thus have that, with probability at least $1 - \frac{1}{2 \log |P|}$,

$$\|\operatorname{centroid}(O_B \cap P_j) - \operatorname{centroid}(O_B^{\text{rem}})\| \leq \frac{32 \cdot (\log^{0.5} |P|) \cdot r_{\text{avg}}}{\log^3 |P| / (\sqrt{2} \cdot \log^{0.5} |P|)} \leq \frac{r_{\text{avg}}}{3 \log^{0.5} |P|}.$$

Moreover, observe that $O_{\text{good}} \cap P_j$ is constructed by including each element of $O_B \cap P_j$ independently with probability $1 - \varepsilon \geq \frac{1}{\log |P|}$. We have

$$\mathbb{E}[|O_{\text{good}} \cap P_j|] \geq (1 - \varepsilon) \cdot |O_B \cap P_j| \geq \frac{|O_B|}{4 \cdot s \cdot t \cdot \log |P|} \geq \frac{|O|^{0.6}}{8 \cdot \log^4 |P|}$$

and by Lemma A.1, we get that $|O_{\text{good}} \cap P_j| \geq \frac{1}{2}\frac{|O|^{0.6}}{8\cdot\log^4|P|} \geq \log^6|P|$ with probability at least $1 - \frac{1}{4\log|P|}$. Conditioned on $|O_{\text{good}} \cap P_j| \geq \log^6|P|$, with probability at least $1 - \frac{1}{4\log|P|}$, we get

$$\|\operatorname{centroid}(O_{\text{good}} \cap P_j) - \operatorname{centroid}(O_B \cap P_j)\| \leq \frac{32 \cdot (\log^{0.5}|P|) \cdot r_{\text{avg}}}{\log^3|P|/(2 \cdot \log^{0.5}|P|)} \leq \frac{r_{\text{avg}}}{3\log^{0.5}|P|},$$

due to Lemma A.3. Thus from the union bound and the triangle inequality, we conclude that, $\Pr[\mathcal{E}_j^1|P_1,\ldots,P_{j-1},\mathcal{E}^{\text{size}}] \leq \frac{1}{2\log|P|} + \frac{1}{4\log|P|} + \frac{1}{4\log|P|} = \frac{1}{\log|P|}$.

Now suppose that $\varepsilon\delta|B| \geq \frac{1}{2}|O|^{0.95}$. We bound the probabilities of $\mathcal{E}_j^2, \mathcal{E}_j^3$ and $\mathcal{E}_j^4$ conditioned on $P_1,\ldots,P_{j-1}$ and $\mathcal{E}^{\text{size}}$ as follows:

Bounding $\Pr[\mathcal{E}_j^2|P_1,\ldots,P_{j-1},\mathcal{E}^{\text{size}}]$

We have $\mathbb{E}[|O_{\text{bad}} \cap P_j|] = \varepsilon\delta|P_j| \geq \varepsilon\delta\frac{|B|}{4\cdot s\cdot t} \geq \frac{|O|^{0.95}}{8\cdot s\cdot t\cdot\log|P|} = \frac{|O|^{0.55}}{8\log^4|P|}$. By Lemma A.1, we have that

$$\big||O_{\text{bad}} \cap P_j| - \varepsilon\delta|P_j|\big| \leq \frac{\varepsilon\delta|P_j|}{|O|^{0.25}}$$

with probability at least $1 - \exp\left(-\frac{|O|^{0.55}}{32|O|^{0.5}\log^4|P|}\right)$. Thus we have $\Pr[\mathcal{E}_j^2|P_1,\ldots,P_{j-1},\mathcal{E}^{\text{size}}] \leq \frac{1}{\log|P|}$.

Bounding $\Pr[\mathcal{E}_j^3|P_1,\ldots,P_{j-1},\mathcal{E}^{\text{size}}]$

We have $\mathbb{E}[|O_{\text{noisy}} \cap P_j|] = (1-\varepsilon)|O_B \cap P_j| + \varepsilon\delta|P_j|$. Using the second term, by the same reasoning as in the previous bound, we get

$$\mathbb{E}[|O_{\text{noisy}} \cap P_j|] \geq \frac{|O|^{0.55}}{8\log^5|P|}.$$

On the other hand, we also have that

$$\mathbb{E}[|O_{\text{noisy}} \cap P_j|] \leq (1-\varepsilon)\frac{3|O_B|}{4\cdot s\cdot t} + \varepsilon\delta\frac{3|B|}{4\cdot s\cdot t} \leq \frac{\delta|B|}{s\cdot t} \leq \frac{|O|^{1.1}}{s\cdot t} \leq \frac{|O|^{0.7}}{\log^4|P|}.$$

By Chernoff bounds, we have

$$\Pr\left[||O_{\text{noisy}} \cap P_j| - \mathbb{E}[|O_{\text{noisy}} \cap P_j|| \geq \frac{|O|^{0.55}}{\log^6|P|}\right] \leq 2\cdot\exp\left(-\frac{|O|^{1.1}}{4\cdot(\log^{12}|P|)\cdot\mathbb{E}[|O_{\text{noisy}} \cap P_j|]}\right) \leq \frac{1}{\log|P|}.$$

This yields that probability at least $1 - \frac{1}{\log |P|}$,

$$\begin{aligned}
|O_{\text{noisy}} \cap P_j| &\geq (1-\varepsilon)|O_B| + \varepsilon\delta|P_j| - \frac{|O|^{0.55}}{\log^6 |P|} \\
&\geq \varepsilon\delta|P_j| + \frac{|O_B|}{4 \cdot s \cdot t \cdot (\log |P|)} - \frac{|O|^{0.55}}{\log^6 |P|} \\
&\geq \varepsilon\delta|P_j| + \frac{|O|^{0.95}}{8|O|^{0.4} \log^5 |P|} - \frac{|O|^{0.55}}{\log^6 |P|} \\
&\geq \varepsilon\delta|P_j| + \frac{|O|^{0.55}}{8 \log^5 |P|} - \frac{|O|^{0.55}}{\log^6 |P|} \\
&\geq \varepsilon\delta|P_j| + \frac{|O|^{0.55}}{\log^6 |P|}.
\end{aligned}$$

Since $\varepsilon\delta|P_j| \leq \frac{|O|^{0.7}}{\log^4 |P|}$, we get that $|O_{\text{noisy}} \cap P_j| \geq (1 + |O|^{-0.2}) \cdot \varepsilon\delta|P_j|$. Thus we have

$$\Pr[\mathcal{E}_j^3 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{1}{\log |P|}.$$

Bounding $\Pr[\mathcal{E}_j^4 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}]$

From the same reasoning we did in the bound for the conditional probability of $\mathcal{E}_2$, we have $\mathbb{E}[|O_{\text{bad}} \cap P_j|] \geq \frac{|O|^{0.55}}{8 \cdot \log^4 |P|}$, and by Lemma A.1, we get that with probability at least $1 - \frac{1}{2 \log |P|}$,

$$|O_{\text{bad}} \cap P_j| \geq \frac{1}{2}\mathbb{E}[|O_{\text{bad}} \cap P_j|] \geq \frac{|O|^{0.55}}{16 \cdot \log^4 |P|}.$$

Conditioned on $|O_{\text{bad}} \cap P_j| \geq \frac{|O|^{0.55}}{16 \cdot \log^4 |P|}$, by Lemma A.3, we get that, with probability at least $1 - \frac{1}{2 \log |P|}$,

$$\| \operatorname{centroid}(O_{\text{bad}} \cap B \cap P_j) - \operatorname{centroid}(B \cap P_j) \| \leq \frac{32 \cdot (\log^{0.5} |P|) r_{\text{avg}}}{|O|^{0.275} / (4\sqrt{2} \cdot \log^{2.5} |P|)} \leq \frac{r_{\text{avg}}}{|O|^{0.25}}.$$

Thus we have that $\Pr[\mathcal{E}_j^2 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}] \leq \frac{1}{2 \cdot \log |P|} + \frac{1}{2 \cdot \log |P|} \leq \frac{1}{\log |P|}$.

Finally, to conclude the proof, suppose that $\varepsilon\delta|B| \leq |O|^{0.95}$. In this case, using the conditioning on $\mathcal{E}^{\text{size}}$, it follows that $\mathcal{E}_j^5$ always holds.

Bounding $\Pr[\mathcal{E}_j^5 | P_1, \ldots, P_{j-1}, \mathcal{E}^{\text{size}}]$

Since P_j is size-wise good, we have

$$|O_{\text{bad}} \cap P_j| \leq \frac{3|B|}{4 \cdot s \cdot t} \leq \frac{|O|^{0.95}}{s \cdot t}$$

and

$$|O_{\text{good}} \cap P_j| \geq \frac{|O_B|}{4 \cdot s \cdot t} \geq \frac{|O|}{8 \cdot s \cdot t}.$$

Thus we conclude that

$$\frac{|O_{\text{bad}} \cap P_j|}{|O_{\text{good}} \cap P_j|} \leq \frac{8}{|O|^{0.05}} \leq \frac{1}{\log^2 |P|},$$

and hence, conditioned on $P_1, \ldots, P_{j-1}$ and $\mathcal{E}^{\text{size}}$, $\mathcal{E}_j^5$ always holds. ■

Now to conclude the proof of Lemma C.5, recall that $\mathcal{Q}^{\text{good}}$ denote the collection subsets in $\mathcal{Q}$ that are good for ball B. By Lemma A.2, we have that,

$$\begin{aligned}
\Pr\left[|\mathcal{Q}_B| < 3t/4 \mid \mathcal{E}^{\text{size}}\right] &\leq \Pr\left[\left||\mathcal{Q}^{\text{good}}| - 0.9t\right| < \frac{1}{6} \cdot 0.9t \,\middle|\, \mathcal{E}^{\text{size}}\right] \\
&\leq \exp\left(-\frac{0.9|O|^{0.4}}{6^2}\right) \\
&\leq \exp\left(-|O|^{0.25}\right).
\end{aligned}$$

Using Lemma C.8 together with Lemma A.2, we have the following:

$$\begin{aligned}
\Pr\left[|\mathcal{Q}^{\text{good}}| \geq \frac{3}{4}t\right] &\geq 1 - \Pr[\overline{\mathcal{E}^{\text{size}}}] - \Pr\left[|\mathcal{Q}_B| < 3t/4 \mid \mathcal{E}^{\text{size}}\right] \\
&\geq 1 - \exp(-|O|^{0.25}) - \exp(-|O|^{0.25}) \\
&\geq 1 - \frac{1}{2|P|^2}\exp(-|O|^{0.2}).
\end{aligned}$$

C.4. Putting things together (proof of Theorem 2.3)

In this section we prove Theorem 2.3 assuming Lemma 2.6 and Lemma 2.7. For convenience, we start be restating these lemmas and the theorem.

Lemma 2.6 *There exists an algorithm* GOODBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ be a nice *instance of the noisy center estimation problem, O_{noisy} be an (ε, δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. The algorithm* GOODBALL *takes $(P, O_{\text{noisy}}, \varepsilon', \delta')$ as input and outputs a ball $B \in \mathcal{B}^{\text{all}}$ such that $B \in \mathcal{B}^{\text{good}}$ with probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$, where the probability is over the randomness of O_{noisy}.*

Lemma 2.7 *There exists an algorithm* CENTERINBALL *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ be a nice *instance of the noisy center estimation problem, O_{noisy} be an (ε, δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. The algorithm* CENTERINBALL *takes $(P, O_{\text{noisy}}, \varepsilon', \delta')$ and $B \subseteq P$ as input and outputs a center $\tilde{\mathbf{o}}_B$. With probability at least $1 - \frac{1}{2}\exp\left(-|O|^{0.2}\right)$ over the randomness of O_{noisy} and the algorithm's internal random bits, it holds that*

$$\|\tilde{\mathbf{o}}_B - \text{centroid}(O \cap B)\| \leq \frac{6 \cdot r_{\text{avg}}}{\log^{0.5} |P|} \text{ for all } B \in \mathcal{B}^{\text{good}}.$$

Theorem 2.3 *There exists an algorithm* ONECENTER *that satisfies the following:*

Let $(P, O, \varepsilon, \delta)$ be a nice *instance of the noisy center estimation problem, O_{noisy} be an (ε, δ)-noisy version of O, and ε', δ' be* close *approximations of ε and δ. The algorithm* ONECENTER *takes $(P, O_{\text{noisy}}, \varepsilon', \delta')$ as input and outputs a center $\hat{\mathbf{o}}$ such that*

$$\text{cost}(O, \hat{\mathbf{o}}) \leq (1 + O(1/\log |P|)) \cdot \text{cost}(O, \mathbf{o})$$

with probability at least $1 - \exp(-|O|^{0.2})$, where the probability is over the randomness of the noisy labels and the algorithm's internal random bits.

Proof of Theorem 2.3. The algorithm first finds a ball B using the algorithm GOODBALL with $(P, O_{\text{noisy}}, \varepsilon', \delta')$ as input, and then uses the algorithm CENTERINBALL with $(P, O_{\text{noisy}}, \varepsilon', \delta')$ and B as input to find a center $\hat{\mathbf{o}}$.

Let $\mathcal{E}_1$ be the event that ball B returned by GOODBALL is not *good*, i.e., the event that $B \notin \mathcal{B}^{\text{good}}$. Let $\mathcal{E}_2$ be the event that CENTERINBALL fails to find a good approximate center for some good ball, i.e., the event that

$$\|\tilde{\mathbf{o}}_B - \text{centroid}(O \cap B')\| > \frac{6 \cdot r_{\text{avg}}}{\log^{0.5} |P|} \text{ for some } B' \in \mathcal{B}^{\text{good}}.$$

Suppose that neither of the events $\mathcal{E}_1$ and $\mathcal{E}_2$ happen. Then, since $B \in \mathcal{B}^{\text{good}}$, we have that $B \cap O$ is a subset of O obtained by removing at most $\frac{|O|}{\log |P|}$ fraction of the points from O. Therefore, by Lemma B.2, we have that $\|\text{centroid}(B \cap O) - \mathbf{o}\| \leq \frac{\sqrt{2} \cdot r_{\text{avg}}}{\log^{0.5} |P|}$. Moreover, we also have that CENTERINBALL succeeds on B and hence $\|\hat{\mathbf{o}} - \text{centroid}(O \cap B)\| \leq \frac{6 \cdot r_{\text{avg}}}{\log^{0.5} |P|}$. Hence, from the triangle inequality, it follows that

$$\|\hat{\mathbf{o}} - \mathbf{o}\| \leq \|\hat{\mathbf{o}} - \text{centroid}(B \cap O)\| + \|\text{centroid}(B \cap O) - \mathbf{o}\| \leq \frac{8 \cdot r_{\text{avg}}}{\log^{0.5} |P|}.$$

Thus we conclude that

$$\begin{aligned}
\text{cost}(O, \hat{\mathbf{o}}) &= \text{cost}(O, \mathbf{o}) + |O| \cdot \|\hat{\mathbf{o}} - \mathbf{o}\|^2 \\
&\leq \text{cost}(O, \mathbf{o}) + \frac{64}{\log |P|} \cdot |O| \cdot r_{\text{avg}}^2 \\
&= \left(1 + \frac{64}{\log |P|}\right) \cdot \text{cost}(O, \mathbf{o}).
\end{aligned}$$

By the union bound, the failure probability of this algorithm is upper-bounded by

$$\Pr[\mathcal{E}_1 \cup \mathcal{E}_2] \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] \leq \exp(-|O|^{0.2}),$$

where in the last inequality we use that both $\Pr[\mathcal{E}_1]$ and $\Pr[\mathcal{E}_2]$ are at most $\frac{1}{2} \exp(-|O|^{0.2})$ due to Lemma 2.6 and Lemma 2.7. ■

Appendix D. Proportional Stochastic Noise Model: Proof of Theorem 3.1

In this section, we formally prove Theorem 3.1 which we restate below. The algorithm for this setting simply uses our earlier algorithm of Section 2 as a black box.

Theorem 3.1 *There exists an algorithm such that the following holds:*

Let $(P, \{O_1, \ldots, O_k\}, \varepsilon)$ be an instance of the k-means problem in the proportional stochastic noise setting where $|O_i| \geq \log^{200} |P|$ for all $i \in [k]$ and $\varepsilon \leq 1 - \frac{1}{\log |P|}$. Let $O'_1, \ldots, O'_k$ be the ε-noise added versions of $O_1, \ldots, O_k$. The algorithm takes as input $(P, \{O'_1, \ldots, O'_k\}, \varepsilon)$ and outputs centers $\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k$ such that, with probability at least $1 - \frac{1}{|P|}$, we have

$$\operatorname{cost}(O_i, \hat{\mathbf{o}}_i) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \operatorname{cost}(O_i, \mathbf{o}_i)$$

where $\mathbf{o}_i$ denote the centroid of O_i. Consequently, the output of the algorithm satisfies

$$\operatorname{cost}(O_i, \{\hat{\mathbf{o}}_1, \ldots, \hat{\mathbf{o}}_k\}) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \mathrm{OPT}\,.$$

with probability at least $1 - \frac{1}{|P|}$. The probability is over the randomness of O'_i's and the algorithm's internal random bits.

Proof As in the balanced adversarial model, the high-level idea of the candidate algorithm is to recover the centers for each label separately.

Fix any $i \in [k]$ and let $X = |O'_i|$. We have $\mathbb{E}[X] = (1-\varepsilon)|O_i| + \varepsilon \cdot \frac{|O_i|}{|P|} \cdot |P| = |O_i|$, and by Chernoff bounds

$$\Pr[|X - \mathbb{E}[X]| \geq |O_i|^{0.6}] \leq 2 \exp\left(-\frac{|O_i|^{1.2}}{4\mathbb{E}[X]}\right) \leq \exp(-|O_i|^{0.1}) \leq |P|^{-3}.$$

Thus, with probability at least $1 - |P|^{-3}$, we have that $\frac{|O'_i|}{|P|}$ is a $(1 \pm |O_i|^{-0.4})$-approximation to $\frac{|O_i|}{|P|}$. When this happens, we have that $\left(P, O_i, \varepsilon, \frac{|O_i|}{|P|}\right)$ is a *nice* instance of the noisy center estimation problem. Hence, using the algorithm ONECENTER whose existence is guaranteed by Theorem 2.3, we can find a center $\hat{\mathbf{o}}_i$ such that, with probability at least $1 - \exp(-|O_i|^{0.2}) \geq 1 - \exp(-\log^{40} |P|) \geq 1 - |P|^{-3}$,

$$\operatorname{cost}(O_i, \hat{\mathbf{o}}_i) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \operatorname{cost}(O_i, \mathbf{o}_i).$$

Let $\mathcal{E}_1$ be the event that $\frac{|O'_i|}{|P|}$ is a $(1 \pm |O_i|^{-0.4})$-approximation to $\frac{|O_i|}{|P|}$, and let $\mathcal{E}_2$ be the event ONECENTER from Theorem 2.3 succeeds. Then, the failure probability of our algorithm is at most $\Pr[\overline{\mathcal{E}_1}] + \Pr[\overline{\mathcal{E}_2}] \leq |P|^{-3} + |P|^{-3} \leq |P|^{-2}$.

We apply the procedure mentioned above on each O_i separately, and the proof now follows by the union bound. ■

Appendix E. Uniform Stochastic Noise Model

In this section, we present our algorithm for the k-means problem in the uniform stochastic noise model and prove Theorem 4.1, which we restate below:

Theorem 4.1 *There exists an algorithm such that the following holds:*

Let $(P, \{O_1, \dots, O_k\}, \varepsilon)$ be an instance of the k-means problem in the uniform stochastic noise setting where $|O_i| \geq \max(\log^{200} |P|, k^{200} \log |P|)$ for all $i \in [k]$ and $\varepsilon \leq 1 - \frac{1}{\log |P|}$. Let $O'_1, \dots, O'_k$ be the ε-noise added versions of $O_1, \dots, O_k$. The algorithm takes $(P, \{O'_1, \dots, O'_k\}, \varepsilon)$ as input and outputs centers $\hat{\mathbf{o}}_1, \dots, \hat{\mathbf{o}}_{k'}$ where $k' \leq k$ such that

$$\operatorname{cost}(P, \{\hat{\mathbf{o}}_1, \dots, \hat{\mathbf{o}}_{k'}\}) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \mathrm{OPT}$$

with probability at least $1 - \frac{1}{|P|}$, where the probability is over the randomness of O'_i's and the algorithm's internal random bits.

Recall that, as outlined in Section 4, the idea is to estimate cluster centers in several phases, starting from the largest clusters. In each phase, we estimate the centers of clusters that are large compared to the current number of remaining points. Then we reduce the overall size of the instance by removing points that are closer to the recovered centers.

The iterative approach above leads to a more complicated analysis compared to that of the proportional stochastic noise model. Recall that in proportional stochastic noise model, our analysis was on a per-cluster basis. Namely, for each optimal cluster O_i, we found a center $\hat{\mathbf{o}}_i$ that is close to the centroid of O_i. For the iterative approach described above, this kind of analysis no longer works due to the following reason: Suppose that O_1 is the largest cluster, and for simplicity, assume that we estimate only one center $\hat{\mathbf{o}}_1$ (for O_1) in the first phase. Note that although $\hat{\mathbf{o}}_1$ is close to the true centroid of O_1, when we remove the points that are close to $\hat{\mathbf{o}}_1$, we might end up removing many points from $P \setminus O_1$ and *assign* them to $\hat{\mathbf{o}}_1$, while still keeping a lot of points of O_1 in the instance. (This may happen because points may not be symmetrically distributed around their centroid.) This poses two questions: First, how do we bound the cost of these assigned points, especially if we assign wrongly? Second, since we may end up in a situation where we do not have enough points remaining for the algorithm to proceed, how can we bound the cost of the remaining points at this stage?

In the first case, there are two kinds of points: For assigned points that belong to O_1 (i.e., correctly assigned points), the cost increase is small since $\hat{\mathbf{o}}_1$ is close to the centroid of O_1. (In the recursively solved instances, this still holds for a given cluster considering only the remaining points of that cluster.) However, for assigned points that does *not* belong to O_1 (i.e., wrongly assigned points), the cost increase can be large compared to their true assignment cost in the optimal clustering. Thus we make sure that for each such point, there always exists sufficiently many points (e.g., at least $\log |P|$ points) in O_1 that are so-far unassigned. Since we assign the *closest* points, the cost of assigned points that does not belong to O_1 can be *charged* (fractionally) to those unassigned points of O_1.

In the second case, let $\mathbf{p}$ be a remaining unassigned point when our algorithm terminates, and suppose that $\mathbf{p} \in O_i$. If we have already estimated a center for i in a previous stage, then we can easily bound the cost of assigning $\mathbf{p}$ to that center since the estimated center is close to the centroid of remaining points of O_i at the time we estimated that center. On the other hand, if we have not

estimated a center for i yet, this means that we have already assigned many points of O_i to other centers before (since the number of remaining points is very small). We show that we can assign such points $\mathbf{p}$ to those centers such that the total assignment cost is not too large compared to their total assignment cost in the optimal clustering.

E.1. Our algorithm

As discussed before, the algorithm reduces the instance size in each phase and continues until we have at most $\max(\log^{195}|P|, k^{200})$ points remaining. Let $\alpha = 0.99$ and let Φ be the smallest integer such that $|P|^{\alpha^\Phi} \leq \max(\log^{195}|P|, k^{200})$. Our algorithm continues for Φ iterations. For $\phi = 1, \dots, \Phi$, in the ϕ-th iteration, the algorithm operates on a subset $P_{\phi-1} \subseteq P$ where $P_0 = P$ and $|P_\phi| = |P|^{\alpha^{\phi-1}}$. In each phase ϕ, it estimates centers for sufficiently large clusters in $P_{\phi-1}$, and then produces a new instance P_ϕ by removing the points that are closest to the estimated centers until we arrive at the new required size.

At the beginning of each phase, the algorithm computes a set of cluster-labels L_ϕ for which it is going to estimate centers. It does so in such a way that the following holds:

1. L_ϕ does not contain any cluster-label considered in the previous iterations. I.e., $L_\phi \cap L_{\phi'} = \emptyset$ for all $\phi' < \phi$.
2. L_ϕ contains all cluster labels whose sizes in the current instance are at least $|P|^{\alpha^\phi}$ (except those cluster labels that are already considered in previous iterations). Namely, $j \in L_\phi$ for all $j \in [k] \setminus (L_1 \cup \dots \cup L_{\phi-1})$ such that $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}$.
3. L_ϕ does not contain any cluster label whose size in the current instance is smaller than $|P|^{\alpha^\phi}/2$. Formally, for all $j \in L_\phi$, it holds that $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}/2$.

If a set L_ϕ of cluster labels satisfies the above three conditions, we say that it is *nice*.

Note that the first condition above ensures that we never estimate two centers for the same cluster, so that we never output more than k centers. The second condition ensures that we recover centers for all sufficiently large clusters. The third condition ensures that we do not attempt to estimate centers for too small clusters. We show how to find *nice* sets of cluster labels in Appendix E.3 where we extensively rely on Chernoff bounds.

Once the set of labels is computed for a phase, the next step is to estimate a center for each label in the set. For a phase ϕ and a label $j \in L_\phi$, let $\mathbf{o}_j^\phi$ denote the centroid of $O_j \cap P_{\phi-1}$, i.e., the centroid of the points of O_j that remain in phase ϕ. We say that an estimate $\hat{\mathbf{o}}_j$ is *good* at phase ϕ if

$$\|\hat{\mathbf{o}}_j - \mathbf{o}_j^\phi\| \leq O(1/\log^{0.5}|P|) \cdot \sqrt{\mathrm{cost}(O_j \cap P_{\phi-1}, \mathbf{o}_j^\phi)/|O_j \cap P_{\phi-1}|}.$$

When this happens, we have that $\mathrm{cost}(O_j \cap P_{\phi-1}, \hat{\mathbf{o}}_j) = (1 + O(1/\log|P|)) \cdot \mathrm{cost}(O_j \cap P_{\phi-1}, \mathbf{o}_j^\phi)$. Our algorithm, at phase ϕ, estimates *good* centers $\hat{\mathbf{o}}_j$ for each $j \in L_\phi$. In Appendix E.4, we show how to estimate such centers using our algorithm from Section 2 as a black box. We denote the set of estimated centers at the end of phase ϕ by C_ϕ.

Finally, the algorithm reduces the current instance size by removing points that are close to the already estimated centers. For a point $\mathbf{c} \in \mathbb{R}^d$, let $\mathbf{c}^* \in P$ be the point in P that is closest to $\mathbf{c}$. We call $\mathbf{c}^*$ the representative of $\mathbf{c}$. At the end of each phase, the algorithm first construct a set C_ϕ^* which consists of the representatives of centers in C_ϕ. It then constructs P_ϕ by removing the points

closest to the points in C^*_ϕ until we end up with $|P|^{\alpha^{\phi-1}}$ points. As we see later, this additional complication (i.e., removing points closest to C^*_ϕ instead of those closest to C_ϕ) allows us to limit the number of intermediate states the algorithm may arrive at, which, in turn, allows us to union bound the failure probability of the algorithm over all such intermediate states.

We outline our algorithm in Algorithm 5.

Algorithm 5: Algorithm for estimating centers in the uniform stochastic noise model.

1 **Input:** An instance P of the k-means problem in the uniform stochastic noise model.
2 Let $\alpha = 0.99$ and let Φ be such that $|P|^{\alpha^\Phi} \leq \max(\log^{195} |P|, k^{200})$.
3 $P_0 \leftarrow P$, $C_0 \leftarrow \emptyset$
4 **for** *each* $\phi = 1, \ldots, \Phi$ **do**
5 Compute a *nice* set of cluster labels L_ϕ.
6 **for** *each* $j \in L_\phi$ **do**
7 find a *good* center $\hat{\mathbf{o}}_j$.
8 $C_\phi \leftarrow C_{\phi-1} \cup \{\hat{\mathbf{o}}_j : j \in L_\phi\}$.
9 Let C^*_ϕ be the representatives of C_ϕ.
10 $P_\phi \leftarrow \{\text{farthest } |P|^{\alpha^{\phi-1}} \text{ points in } P_{\phi-1} \text{ from } C^*_\phi\}$
11 **return** C_Φ

E.2. Analysis of Algorithm 5

In later sections, we show how the tasks in Line 5 and Line 7 are performed with high success probability. In this section, we prove Theorem 4.1, assuming the algorithm successfully executes Line 5 and Line 7 in all phases. Namely, we prove the following lemma:

Lemma E.1 *Let $C = C_\Phi$ be the output of the procedure outlined in Algorithm 5. We have that*

$$\text{cost}(P, C) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \text{cost}(P, \{\mathbf{o}_1, \ldots, \mathbf{o}_k\}).$$

Proof We use $\phi(j)$ to denote the phase in which a center $\hat{\mathbf{o}}_j$ is estimated. I.e., $j \in L_{\phi(j)}$ and $\hat{\mathbf{o}}_j \in C_{\phi(j)}$. For $j \in [k]$ for which no center was estimated, define $\phi(j) = \Phi + 1$. Let $L = \cup_{\phi \in [\Phi]} L_\phi$ be the set of all cluster labels for which the algorithm estimates centers.

To bound the cost of the output clustering (defined in terms of the centers C_Φ), we consider three types of points.

1. A^{good}: the set of points that belong to cluster of L that were present in the instance when the respective center was estimated. Namely, $A^{\text{good}} = \cup_{j \in L} O_j \cap P_{\phi(j)-1}$.
2. A^{bad}: the set of points that belong to some cluster O_j for $j \in [k]$ but were removed from consideration in some phase ϕ where $\phi < \phi(j)$. Namely, $A^{\text{bad}} = \cup_{j \in [k]} \{\mathbf{p} \in O_j : \mathbf{p} \notin P_{\phi(j)}\}$.
3. A^{ugly}: all the remaining points. Namely, $\cup_{j \in [k] \setminus L} O_j \cap P_\Phi$.

Bounding the cost of A^{good} Conditioned on Line 7 succeeds in finding *good* centers for all $j \in L$, we have the following bound for the cost of A^{good}:

$$\begin{aligned}\text{cost}(A^{\text{good}}, C) &\leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \sum_{j \in L} \text{cost}\left(O_j \cap P_{\phi(j)-1}, \mathbf{o}_j^{\phi(j)}\right) \\ &\leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \sum_{j \in L} \text{cost}(O_j, \mathbf{o}_j). \end{aligned} \tag{17}$$

The first inequality holds because all centers in C are good and the second one holds because $O_j \cap P_{\phi(j)-1} \subseteq O_j$ and $\mathbf{o}_j^{\phi(j)}$ and $\mathbf{o}_j$ are, respectively, the centroids of $O_j \cap P_{\phi(j)-1}$ and O_j. (Observe that $\text{cost}(O_j \cap P_{\phi(j)-1}, \mathbf{o}_j^{\phi(j)}) \leq \text{cost}(O_j \cap P_{\phi(j)-1}, \mathbf{o}_j) \leq \text{cost}(O_j, \mathbf{o}_j)$.)

We also establish a cost bound for A^{good} with respect to the representative centers. This will be useful in our charging argument for bounding the cost of A^{bad} later. For this, define $\text{cost}^*(A^{\text{good}}) = \sum_{j \in L} \text{cost}(O_j \cap P_{\phi(j)}, \hat{\mathbf{o}}_j^*)$, where $\hat{\mathbf{o}}_j^*$ is the representative of $\hat{\mathbf{o}}_j^*$. Using the fact that the representative is the closest point and that $\hat{\mathbf{o}}_j^*$ is good, we can show that for each point $\mathbf{p} \in O_j \cap P_{\phi(j)-1}$,

$$\|\mathbf{p} - \hat{\mathbf{o}}_j^*\|^2 \leq O(1) \cdot \left(\frac{\text{cost}(O_j \cap P_{\phi(j)-1}, \mathbf{o}_j^{\phi(j)})}{|O_j \cap P_{\phi(j)-1}|} + \|\mathbf{p} - \mathbf{o}_j^{\phi(j)}\|^2\right).$$

Summing over points in each $O_j \cap P_{\phi(j)-1}$ for $j \in L$, we thus get that

$$\begin{aligned}\text{cost}^*(A^{\text{good}}) &\leq O(1) \cdot \sum_{j \in L} \text{cost}\left(O_j \cap P_{\phi(j)-1}, \mathbf{o}_j^{\phi(j)}\right) \\ &\leq O(1) \cdot \sum_{j \in L} \text{cost}(O_j, \mathbf{o}_j). \end{aligned} \tag{18}$$

Bounding the cost of A^{bad} To bound $\text{cost}(A^{\text{bad}}, C)$ we show that for each point in A^{bad}, there are many (at least $\log |P|$) unique points in A^{good} that have a greater cost.

For a phase $\phi \in [\Phi]$, let $L_{\leq \phi} = \cup_{\phi'=1}^{\phi} L_\phi$ and r_ϕ be the minimum distance from a point in P_ϕ to any center in C_ϕ^*. Let $P_\phi^{\text{pend}} = P_\phi \cap (\cup_{j \in [k] \setminus L_{\leq \phi}} O_j)$ be the points that both (1) remain in the instance at the end of phase ϕ and (2) belong to so far un-recovered clusters. Since each cluster in $[k] \setminus L_{\leq \phi}$ has size at most $|P|^{\alpha^\phi}$ (due to Line 5), we have

$$|P_\phi^{\text{pend}}| \leq k \cdot |P|^{\alpha^\phi} \leq |P|^{\alpha^\phi/200} \cdot |P|^{\alpha^\phi} \leq |P|^{1.006\alpha^\phi}.$$

In the second inequality, we use that $|P|^{\alpha^\phi} \geq |P|^{\alpha^\Phi} = |P|^{\alpha^{\Phi-1}\alpha} \geq k^{200\alpha}$.

On the other hand, after phase ϕ, we have $|P_\phi| = |P|^{\alpha^{\phi-1}} \geq |P|^{1.01\alpha^\phi}$. Let $P_\phi^{\text{good}} = P_\phi \cap (\cup_{j \in L_{\leq \phi}} O_j)$ be the set of points in P_ϕ that belong to already recovered clusters. We thus have that $|P_\phi^{\text{good}}| \geq |P_\phi| - |P_\phi^{\text{pend}}| \geq |P|^{1.01\alpha^\phi} - |P|^{1.006\alpha^\phi} \geq |P|^{1.006\alpha^\phi} \log^2 |P|$. We now argue as follows: In P_Φ^{good}, we can select $|P|^{1.006\alpha^\Phi} \log^2 |P|$ points (denoted by A_Φ^{good}) that are distance at least r_Φ away from their respective representative center. In $P_{\Phi-1}^{\text{good}}$, we can select $|P|^{1.006\alpha^{\Phi-1}} \log^2 |P|$ points (denoted by $A_{\Phi-1}^{\text{good}}$) that are distance at least $r_{\Phi-1}$ away from their respective representative center, and out of these points, we can select $(|P|^{1.006\alpha^{\Phi-1}} - |P|^{1.006\alpha^\Phi}) \cdot \log^2 |P|$ points

that does not overlap with A_{Φ}^{good}. Namely, for each $\phi = \Phi - 1, \ldots, 1$, in P_{ϕ}^{good}, we can select $|P|^{1.006\alpha^{\phi}} \log^2 |P|$ points (denoted by A_{ϕ}^{good} that are distance at least r_{ϕ} away from the representative center estimated for them, and out of these, we can select $(|P|^{1.006\alpha^{\phi}} - |P|^{1.006\alpha^{\phi+1}}) \cdot \log^2 |P|$ many points that does not overlap with $A_{\phi+1}^{\text{good}}$. This yields that

$$\begin{aligned}\text{cost}^*(A^{\text{good}}) &\geq r_1 \cdot (|P|^{1.006\alpha^1} \log^2 |P| - |P|^{1.006\alpha^2} \log^2 |P|) \\ &\quad + r_2 \cdot (|P|^{1.006\alpha^2} \log^2 |P| - |P|^{1.006\alpha^3} \log^2 |P|) + \ldots \\ &\geq r_1 \cdot |P|^{1.006\alpha^1} \log |P| + r_2 \cdot |P|^{1.006\alpha^2} \log |P| + \ldots. \end{aligned} \tag{19}$$

Now let $P_{\phi}^{\text{bad}} = P_{\phi-1} \cap \left(\cup_{j \in [k] \setminus L_{\leq \phi}} O_j \setminus P_{\phi} \right)$ be the points of so far un-recovered clusters that were removed from consideration at the end of phase ϕ. As with P_{ϕ}^{pend} before, we have that $|P_{\phi}^{\text{bad}}| \leq |\cup_{j \in [k] \setminus L_{\leq \phi}} O_j| \leq k \cdot |P|^{\alpha^{\phi}} \leq |P|^{1.006\alpha^{\phi}}$. Observe that $A^{\text{bad}} = \cup_{\phi} P_{\phi}^{\text{bad}}$, and thus

$$\text{cost}(A^{\text{bad}}, C) = \sum_{\phi} \text{cost}(P_{\phi}^{\text{bad}}, C_{\phi}) \leq \sum_{\phi} r_{\phi} \cdot |P|^{1.006\alpha^{\phi}}. \tag{20}$$

Combining Eq. (19) with Eq. (20), we conclude that

$$\text{cost}(A^{\text{bad}}, C) \leq \frac{1}{\log |P|} \cdot \text{cost}^*(A^{\text{good}}). \tag{21}$$

Bounding the cost increase due to A^{ugly} Finally, observe that for any $j \in [k] \setminus L$, we have $|P_{\Phi} \cap O_j| \leq \max(\log^{195} |P|, k^{200})$ whereas we have $|O_j| \geq \max(\log^{200} |P|, k^{200} \log^2 |P|)$. Thus, by applying Lemma A.6 to each cluster in $[k] \setminus L$ separately, we get that

$$\text{cost}(A^{\text{bad}} \cup A^{\text{ugly}}, C) \leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \left(\text{cost}(A^{\text{bad}}, C) + \sum_{j \in [k] \setminus L} \text{cost}(O_j, \mathbf{o}_j)\right). \tag{22}$$

Combining the bounds Eq. (17), Eq. (21), and Eq. (22), we thus get

$$\begin{aligned}\text{cost}(P, C) &\leq \text{cost}(A^{\text{good}} \cup A^{\text{bad}} \cup A^{\text{ugly}}, C) \\ &= \text{cost}(A^{\text{good}}, C) + \text{cost}(A^{\text{bad}} \cup A^{\text{ugly}}, C)\end{aligned}$$

$$
\begin{aligned}
&\leq \mathrm{cost}(A^{\mathrm{good}}) + \left(1 + O\left(\frac{1}{\log |P|}\right)\right)\left(\mathrm{cost}(A^{\mathrm{bad}}, C) + \sum_{j \in [k] \setminus L} \mathrm{cost}(O_j, \mathbf{o}_j)\right) \\
&\leq \mathrm{cost}(A^{\mathrm{good}}) + \left(1 + O\left(\frac{1}{\log |P|}\right)\right)\left(\frac{O(1) \cdot \mathrm{cost}^*(A^{\mathrm{good}})}{\log |P|} + \sum_{j \in [k] \setminus L} \mathrm{cost}(O_j, \mathbf{o}_j)\right) \\
&\leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \left(\mathrm{cost}(A^{\mathrm{good}}) + \sum_{j \in [k] \setminus L} \mathrm{cost}(O_j, \mathbf{o}_j)\right) \\
&\leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \left(\left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \sum_{j \in L} \mathrm{cost}(O_j, \mathbf{o}_j) + \sum_{j \in [k] \setminus L} \mathrm{cost}(O_j, \mathbf{o}_j)\right) \\
&\leq \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \left(\sum_{i \in [k]} \mathrm{cost}(O_i, \mathbf{o}_i)\right) \\
&= \left(1 + O\left(\frac{1}{\log |P|}\right)\right) \cdot \mathrm{cost}(P, \{\mathbf{o}_1, \ldots, \mathbf{o}_k\}).
\end{aligned}
$$

The first inequality uses Eq. (22), the second one uses Eq. (21), and the fourth inequality follows from Eq. (17). ■

In the next two sections, we show how to implement Line 5 and Line 7. Before we proceed, we introduce the concept of an intermediate state of Algorithm 5. We union bound the failure probabilities of Line 5 and Line 7 over all possible intermediate states the algorithm may have.

Definition E.2 (Intermediate state) *We denote an intermediate state of Algorithm 5 by a tuple $(\phi, P_{\phi-1})$ where ϕ denote the current phase and P_ϕ is the set of remaining points.*

Note that when $\phi = 1$, we have only one possible intermediate state $(1, P_0)$. For each intermediate state $(\phi, P_{\phi-1})$, observe that there is at most $O(|P|^k)$ possible intermediate states $(\phi + 1, P_\phi)$ for phase $\phi + 1$ that can be arrived from $(\phi, P_{\phi-1})$. This is because we choose at most k representative centers in each phase out of $|P|$ possible points, and P_ϕ is completely determined by the current state $(\phi, P_{\phi-1})$ and the chosen representative centers. Since the algorithm may have at most $\Phi = O(\log |P|)$ phases, we now have the following observation:

Observation E.3 *The total number of possible intermediate states is $O(|P|^{k \log |P|})$.*

E.3. Implementing Line 5 of Algorithm 5

The following lemma guarantees that we can successfully implement Line 5 of Algorithm 5.

Lemma E.4 *There exists an algorithm $\mathcal{A}_{labels}$ that satisfies the following with probability at least $1 - \frac{1}{2}|P|^{-1}$: Given any intermediate state $(\phi, P_{\phi-1})$, $\mathcal{A}_{labels}$ identifies a set of labels $L \subseteq [k]$ such that $j \in L$ if $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}$ and $j \in L$ only if $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}/2$.*

Proof Given $(\phi, P_{\phi-1})$, the candidate algorithm simply computes $n_j = |O'_j \cap P_{\phi-1}|$ for all $j \in [k]$, and includes $j \in L$ if and only if $n_j \geq \frac{3}{4} \cdot (1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}}$. We have that $\mathbb{E}[n_j] = (1-\varepsilon)|O_j \cap P_{\phi-1}| + (\varepsilon/k)|P_{\phi-1}|$, and

$$\begin{aligned}\Pr\left[|n_j - \mathbb{E}[n_j]| \geq 100(\log|P|)\sqrt{\mathbb{E}[n_j]}\right] &\leq \exp\left(-\log|P|\sqrt{\mathbb{E}[n_j]}\right)\\ &\leq \exp\left(-10k^{10}\log^3|P|\right)\\ &\leq |P|^{-10k^{10}\log^2|P|}.\end{aligned}$$

In the first inequality, we used that $\mathbb{E}[n_j] \geq (\varepsilon/k)|P_{\phi-1}| \geq k^{100}\log^4|P|$. Thus we have that, if $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}$, with probability at least $1 - |P|^{-10k^{10}\log^2|P|}$,

$$\begin{aligned}n_j &\geq E[n_j] - 100(\log|P|)\sqrt{\mathbb{E}[n_j]}\\ &\geq (1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}} + 100(\log|P|)|P|^{0.5\alpha^{\phi-2}}\\ &\geq (1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}} + 100(\log|P|)|P|^{0.7\alpha^{\phi}}\\ &\geq \frac{3}{4}(1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}}.\end{aligned}$$

The last inequality follows because $1 - \varepsilon \geq \frac{1}{\log|P|}$. On the other hand, if $|O_j \cap P_{\phi-1}| \leq \frac{1}{2}|P|^{\alpha^\phi}$, with probability at least $1 - |P|^{-10k^2}$,

$$\begin{aligned}n_j &\leq E[n_j] + 100(\log|P|)\sqrt{\mathbb{E}[n_j]}\\ &\leq \frac{1}{2}(1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}} + 100(\log|P|)|P|^{0.5\alpha^{\phi-2}}\\ &\leq \frac{1}{2}(1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}} + 100(\log|P|)|P|^{0.7\alpha^{\phi}}\\ &\leq \frac{3}{4}(1-\varepsilon)|P|^{\alpha^\phi} + (\varepsilon/k)|P|^{\alpha^{\phi-2}}.\end{aligned}$$

As before, the last inequality follows because $1 - \varepsilon \geq \frac{1}{\log|P|}$. Now, by the union bound over all $j \in [k]$, the candidate algorithm succeeds for a given $(\phi, P_{\phi-1})$ with probability at least $1 - k|P|^{-10k^{10}\log^2|P|}$, and it succeeds for all intermediate states with probability at least

$$1 - k \cdot O(|P|^{k\log|P|}) \cdot |P|^{-10k^{10}\log^2|P|} \geq 1 - \frac{1}{2}|P|^{-1}.$$

■

E.4. Implementing Line 7 of Algorithm 5

The following lemma ensures that we can successfully implement Line 7 of Algorithm 5.

Lemma E.5 *There exists an algorithm* CENTERINUNIFORMNOISE *that satisfies the following: The algorithm* CENTERINUNIFORMNOISE *takes an instance P of the k-means problem in the uniform stochastic noise model, an intermediate state $(\phi, P_{\phi-1})$, and a label $j \in [k]$ as input and*

returns a center $\hat{\mathbf{o}}$*. We say that the center is good if* $\|\hat{\mathbf{o}} - \mathbf{o}_j^\phi\| \leq \frac{100 r_{\text{avg}}^{\phi,j}}{\log^{0.5}|P|}$*, where* $\mathbf{o}_j^\phi$ *denotes the centroid of* $O_j \cap P_{\phi-1}$*. With probability at least* $1 - \frac{1}{2}|P|^{-1}$*, for all intermediate states* $(\phi, P_{\phi-1})$*, and labels* $j \in [k]$ *such that* $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}$*, the returned center is good.*

Proof Fix some intermediate state $(\phi, P_{\phi-1})$ and a label $j \in [k]$ such that $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi}$. Note that $(P_{\phi-1}, O_j \cap P_{\phi-1}, \varepsilon, \frac{1}{k}, O'_j \cap P_{\phi-1}, \varepsilon, \frac{1}{k})$ is an instance of the noisy center estimation problem. Moreover, we have $|O_j \cap P_{\phi-1}| \geq |P|^{\alpha^\phi} \geq |P|^{\alpha^{\phi-2}\alpha^2} = |P_{\phi-1}|^{\alpha^2}$ which implies that $\delta = \frac{1}{k} \leq 1 \leq |O|^{1/\alpha^2}/|P| \leq |O|^{1.1}/|P|$. Thus the instance is in fact *nice*, and hence ONECENTER succeeds in finding a good center with probability at least $1 - \exp(-|O_j|^{0.2}) \geq 1 - \exp(-10 \cdot k^1 0 \log^2 |P|)$. Taking the union bound over all possible intermediate states and j, we have that with probability at least

$$1 - k \cdot O(|P|^{k \log |P|}) \cdot |P|^{-10k^{10} \log^2 |P|} \geq 1 - \frac{1}{2}|P|^{-1},$$

the algorithm succeeds on all intermediate instances. ■

Finally, due to Lemma E.4 and Lemma E.4, it follows that both Line 5 and Line 7 succeeds with probability at least $1 - 1/|P|$.

Appendix F. Hardness and Impossibility Results

In this section, we start by presenting an example that intuitively shows that the labels does not provide any additional information for the adversarial noise model.

Theorem F.1 *For constant* $\alpha > 1$*, any* α*-approximation hardness result for the k-means problem also extends to k-means problem in the adversarial noise setting, even in the case that adversary can alter only the labels of* k^ϵ *clusters or* $|P|^\varepsilon$ *points, for any constant* $\epsilon > 0$*.*

Proof Consider an instance of a k-means problem with points P without any labels. First we add a new cluster far away from the points P as follows. We add a set of points P_0 of size $m >> |P|^c$ for some constant c in the same place with distance more than the cost of the optimum solution. We denote the new instance by P' and use $k+1$ as the number of center that we want to open. Observe that the cost of the optimum solution for these two instances are the same. If there is no bound on the number of elements that the adversary changes from each cluster, adversary can change the labels of the all the point in P to an arbitrary or random color. Therefore the labels in the input does not contain any information regarding the point in P, so any such α-approximate hardness results for the k-kmeans problem applies to this case as well. Furthermore, instead of adding one cluster one can add k' clusters, where k' is significantly higher or independent from k. Therefore even if the we bound the number of the clusters that the adversary is allowed to alter their labels of k^ϵ, the input labels do not have any information and the hardness results still apply. ■

Lets us now show that our result is tight among all the algorithm that looks at the colors independently. Formally we show that:

Theorem 1.2 *In the balanced adversarial noise model, any (potentially randomized) algorithm has an approximation guarantee of* $(1+\Omega(\varepsilon))$ *if it computes the center of each cluster only as a function of the input points with the label of that cluster.*

Proof We start by showing the result for $k = 2$ and then we extend it to general k values. We define the following set of points in one dimensional space:

- Let Q_{-1}, be a set of ϵn points at coordinate -1.
- Let Q_0, be a set of n points at coordinate 0.
- Let Q_1, be a set of ϵn points at coordinate $+1$.

In both instances these three sets of the points exist with label 1. In P_1, P_2 there is additional set of n points with label 2 in coordinates -1 and 1, referred to as W_{-1} and W_1 respectively. Therefore,

$$P_1 = Q_{-1} \cup Q_0 \cup Q_1 \cup W_{-1} \qquad \text{and} \qquad P_2 = Q_{-1} \cup Q_0 \cup Q_1 \cup W_1,$$

which concludes the construction of the instances. One can observe that $O_1 = \{Q_{-1} \cup W_{-1}, Q_0 \cup Q_1\}$ is an optimal solution with centroids at positions -1 and ϵ for $\mathcal{A}_1$ and satisfies the constraint of the label of the adversarial setting. Similarly $O_1 = \{Q_{-1} \cup Q_0, W_1 \cup Q_1\}$ is an optimal solution with centroids at positions $-\epsilon$ and 1 for $\mathcal{A}_2$ which also satisfies the constraint of the label of the adversarial setting. Moreover the cost of both these solutions is

$$\begin{aligned} \text{cost}(O_1) = \text{cost}(O_2) \leq & \epsilon^2 n + (1-\epsilon)^2 \cdot \epsilon n \\ \leq & \epsilon^2 n + \epsilon n - 2\epsilon^2 n + \epsilon^3 n \\ \leq & \epsilon n(1 - \epsilon + \epsilon^2) \end{aligned} \tag{23}$$

Any algorithm that computes the center of the clusters only based on the input points with the label of that cluster; provides same distribution D_1 for the center of the points of labels 1 for $\mathcal{A}_1$ and $\mathcal{A}_2$. Notice that the best distribution D for these two instances is the one that opens a center in position 0 with high probability. The points with label 2 are in the same coordinate in both of these two instances and opening a center there has cost zero. Therefore in the best case the cost of the solution is at least ϵn. Combining with Eq. (23) the approximation ratio is

$$\frac{\epsilon n}{\epsilon n(1 - \epsilon + \epsilon^2)} \in 1 + \Omega(\varepsilon).$$

■