Coresets for Clustering Under Stochastic Noise

Lingxiao Huang* Nanjing University

Zhize Li Singapore Management University

Nisheeth K. Vishnoi Yale University

Runkai Yang Nanjing University

Haoyu Zhao Princeton University

Abstract

We study the problem of constructing coresets for (k, z)-clustering when the input dataset is corrupted by stochastic noise drawn from a known distribution. In this setting, evaluating the quality of a coreset is inherently challenging, as the true underlying dataset is unobserved. To address this, we investigate coreset construction using surrogate error metrics that are tractable and provably related to the true clustering cost. We analyze a traditional metric from prior work and introduce a new error metric that more closely aligns with the true cost. Although our metric is defined independently of the noise distribution, it enables approximation guarantees that scale with the noise level. We design a coreset construction algorithm based on this metric and show that, under mild assumptions on the data and noise, enforcing an ε-bound under our metric yields smaller coresets and tighter guarantees on the true clustering cost than those obtained via classical metrics. In particular, we prove that the coreset size can improve by a factor of up to $\mathrm{poly}(k)$, where n is the dataset size. Experiments on real-world datasets support our theoretical findings and demonstrate the practical advantages of our approach.

1 Introduction

Clustering is a foundational tool in machine learning, with applications ranging from image segmentation and customer behavior analysis to sensor data summarization [67, 80, 5, 20]. An important class of clustering problems is called (k, z)-CLUSTERING where, given a dataset $P \subset \mathbb{R}^d$ of n points and a $k \geq 1$, the goal is to find a set $C \subset \mathbb{R}^d$ of k points that minimizes the cost $\mathrm{cost}_z(P, C) := \sum_{x \in P} d^z(x, C)$. Here $d^z(x, C) := \min\{d^z(x, c) : c \in C\}$ is the distance of x to the center set C and d^z denotes the z-th power of the Euclidean distance. Examples of (k, z)-CLUSTERING include k-MEDIAN (when $z = 1$) and k-MEANS (when $z = 2$). In many applications, the dataset P is large, and it is desirable to have a small representative subset that requires less storage and computation while allowing us to solve the underlying clustering problem. Coresets have been proposed as a solution towards this [48] – a coreset is a subset $S \subseteq P$ that approximately preserves the clustering cost for all center sets. Coresets have found further applications in sublinear models, including streaming [48, 16], distributed [8, 56], and dynamic settings [51] due to the ability to merge and compose them (see, e.g., [83, Section 3.3]).

Yet, a critical limitation of existing coreset constructions is their reliance on exact, noise-free data—a condition rarely met in practice. In practical applications, data is frequently corrupted by measurement error, transmission artifacts, or deliberate noise insertion for privacy and robustness. One reason is that the measurement process may itself introduce noise into the data, or corruption may occur during

*Alphabetical order. Correspondence to: huanglingxiao@nju.edu.cn, zhizeli@smu.edu.sg, nisheeth.vishnoi@gmail.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

the recording or reporting processes [46, 3, 75, 58]. Further, noise can be introduced intentionally in data due to privacy concerns [44, 31, 43], or to ensure robustness [84, 66]. In these scenarios, instead of the true dataset P, one observes a noisy dataset $\widehat{P} \subset \mathbb{R}^d$. Various types of noise can emerge depending on the context: stochastic noise, adversarial noise, and noise due to missing data [7, 11, 58].

The degree of knowledge about the noise may range from complete uncertainty to full specification of its distributional parameters. Stochastic noise, in particular, has been studied in problems such as clustering [58] and regression [81], and is commonly encountered in fields like the social sciences [12, 72, 39], economics [34], and machine learning [31, 84, 66]. When the attributes of data interact weakly with each other, independent and additive stochastic noise is considered [86, 40, 65]. In such case, for each point $p \in P$ and every attribute $j \in [d]$, the observed point is $\widehat{p}_j = p_j + \xi_{p,j}$ where $\xi_{p,j}$ is drawn from a known distribution D_j. Various choices for D_j have been explored, including Gaussian distribution [77, 49], Laplace distribution [17], uniform distribution [3, 74], and Dirac delta distribution [87]. Such noise can reflect inherent individual variability—for example, fluctuations in STEM scores across repeated exams [12, 72], where the mean and variance can be estimated by multiple exams, or from employers making decisions based on statistical information about the groups individuals belong to [34]. In settings focused on privacy or robustness, noise with known parameters might be deliberately added to data, such as the use of i.i.d. Gaussian noise in the Gaussian mechanism [30], or the introduction of i.i.d. Gaussian noise to enhance robustness against adversarial attacks in deep learning [84, 66]. The noise level might also be well known or estimable when data is collected using sensors [73, 79]; see Sections 2 and B.2 for further discussions. Numerous studies have examined the effects of modeled noise, specifically Gaussian noise [85, 58], on clustering tasks [29, 42, 58]. They analyze the relationship between the level of noise and the performance of clustering algorithms, showing that a small amount of noise can actually benefit centroid-based clustering methods [58].

Given the widespread use of coresets in clustering problems, it is crucial to explore the possibility of constructing compact coresets that remain effective in the presence of noise. The effectiveness of a coreset S is usually measured via the approximation quality of S's optimal center set $\mathsf{C}(S)$ in P:

$$r_P(\mathsf{C}(S)) := \frac{\mathrm{cost}_z(P,\mathsf{C}(S))}{\min_{C\subset\mathbb{R}^d:|C|=k} \mathrm{cost}_z(P,C)} = \frac{\mathrm{cost}_z(P,\mathsf{C}(S))}{\mathsf{OPT}_P}. \tag{1}$$

In the noise-free setting, challenges for evaluating quality $r_P(\mathsf{C}(S))$ lie in the computation of the optimal clustering cost OPT_P, which is NP-hard. To address this, one traditionally considers a surrogate error metric that bounds the maximum (over all possible center sets) ratio [36, 24]:

$$\mathsf{Err}(S,P) := \sup_{C\subset\mathbb{R}^d:|C|=k} \frac{|\mathrm{cost}_z(S,C)-\mathrm{cost}_z(P,C)|}{\mathrm{cost}_z(S,C)} \tag{2}$$

This ratio, serving as an upper bound for $r_P(\mathsf{C}(S))$ (see Section 2), helps derive the minimum size necessary for coreset construction. There is a long and productive line of work focused on analyzing the optimal tradeoff between the coreset size $|S|$ and the associated estimation error [21, 22, 55] in the noise-free setting. However, in the noisy setting, there is an additional challenge for evaluating $r_P(\mathsf{C}(S))$: the true underlying dataset P is unobservable. This leads to a natural question: *Can the traditional surrogate error* Err *guide coreset construction when the observed data is noisy?*

Our contributions. We study the problem of coreset construction for clustering in the presence of noise. Motivated by the applications mentioned above, we consider a stochastic additive noise model where each D_j (as defined earlier) is parameterized by a known *noise-level* θ, constrained by a bounded-moment condition (see Definition 2.1). Our first result is adapting the use of the Err measure for coreset construction to this noise model, along with a bound for the coreset's quality $r_P(\mathsf{C}(S))$ (Theorem 3.1). To our knowledge, this is the first result to study how coreset performance degrades under noise. We show that Err can significantly overestimate coreset error under noise, resulting in overly pessimistic guarantees (Section 2).

To address this limitation of using Err, we introduce a new surrogate metric Err_α, termed *approximation-ratio* (Equation (6)). Using this new metric, we design a cluster-wise sampling algorithm (Algorithm 1) that partitions the given noisy dataset into k clusters and takes a uniform sample from each cluster. We show that, under natural and necessary assumptions on P, this new algorithm yields smaller coresets (by a factor of up to $\mathrm{poly}(k)$) and tighter quality guarantees for $r_P(\mathsf{C}(S))$ (Theorem 3.3). These improvements hinge on distinguishing the influence of noise on the clustering cost and the location of the optimal center set $\mathsf{C}(S)$, and may be of independent interest.

Empirical results, in Section 4, support our theoretical findings even for datasets that do not meet the assumptions required by our theoretical analysis (see, e.g., Table 1), and in scenarios involving non-i.i.d. noise across dimensions (see, e.g., Table 7). Overall, our algorithm effectively generates small coresets with theoretical quality bounds, which can be integrated into clustering frameworks, enhancing robustness in noisy environments.

2 Noise models and metrics for coreset quality

This section formalizes the noisy data models we consider, defines the ideal (but unobservable) metric for coreset quality, and introduces two surrogate metrics-Err (standard) and Err_α (ours). We compare their behavior under noise and motivate our new construction.

Noise distribution and models. Given a probability distribution D on $\mathbb{R}$ with mean μ and variance σ^2, D is said to satisfy *Bernstein condition* [15, 14, 33] if there exists some constant $b > 0$ such that for every integer $i \geq 3$, $\mathbb{E}_{X\sim D}\left[|X-\mu|^i\right] \leq \frac{1}{2}i!\sigma^2 b^{i-2}, i = 3, 4, \ldots$. This condition imposes an upper bound on each moment of D, allowing control over tail behaviors, and we consider such noise distributions in this paper. Several well-known distributions satisfy the Bernstein condition, including the Gaussian distribution, Laplace distribution, sub-Gaussian distributions, sub-exponential distributions, and so on [82]; see Section B.1 for a discussion. We begin with a probabilistic noise model that reflects real-world data corruption: with some probability, each point either remains untouched or receives independent additive noise on each coordinate.

Definition 2.1 (Noise model I). Let $\theta \in [0, 1]$ be a noise parameter and $D_1, \ldots, D_d$ be probability distributions on $\mathbb{R}$ with mean 0 and variance 1 that satisfy the Bernstein condition.[2] Every point $\widehat{p}$ ($p \in P$) is i.i.d. drawn from the following distribution: 1) with probability $1-\theta$, $\widehat{p} = p$; 2) with probability θ, for every $j \in [d]$, $\widehat{p}_j = p_j + \xi_{p,j}$ where $\xi_{p,j}$ is drawn from D_j.

When $\theta = 0$, $\widehat{P} = P$ and as $\theta \to 1$, $\widehat{P}$ becomes increasingly noisy. Note that this noise model roughly selects a fraction θ of underlying points and adds an independent noise to each feature $j \in [d]$ that is drawn from a certain distribution D_j.

We also consider the following model, called *noise model II*: For every $i \in [n]$ and $j \in [d]$, $\widehat{p}_{i,j} = p_{i,j} + \xi_{p,j}$, where $\xi_{p,j}$ is drawn from D_j, a probability distribution on $\mathbb{R}$ with mean 0, variance σ^2, and satisfying the Bernstein condition. The main difference from noise model I is that we add noise to every coordinate of each point with a changeable variance σ^2 instead of 1. Moreover, we consider more general noise models where the noise is non-independent across dimensions. For example, the covariance matrix of each noise vector ξ_p is $\Sigma \in \mathbb{R}^{d\times d}$, which extends $\Sigma = \sigma^2 \cdot I_d$ when each $D_j = N(0, \sigma^2)$ under the noise model II.

Several applications mentioned in Section 1 use these noise models. For example, setting $\theta = 1$ and each D_j as a Gaussian distribution corresponds to the Gaussian mechanism in differential privacy [30, 66]. The noise parameter θ is usually known in real-world scenarios. In domains like healthcare, location services, and financial analytics, adding Laplace or Gaussian noise to data points is a common strategy to protect privacy [73, 79], creating a noisy dataset, $\widehat{P}$. In such cases, θ is known in advance, removing the need to compute it during coreset construction. Other scenarios where θ is known include: 1) measurement errors in sensor data, and 2) noise in STEM exam scores (see Section B.2).

The ideal metric. Let $\mathcal{C}$ denote the collection of all subsets $C \subset \mathbb{R}^d$ of size k, termed *center sets*. For any dataset $X \subset \mathbb{R}^d$, let $\mathsf{C}(X)$ denote the optimal center set for (k, z)-CLUSTERING, i.e., $\mathsf{C}(X) := \arg\min_{C\in\mathcal{C}} \mathrm{cost}_z(X, C)$. Given a dataset $P \subset \mathbb{R}^d$ of size n, a coreset is a weighted set $S \subset \mathbb{R}^d$ with a function $w : S \to \mathbb{R}_{\geq 0}$ such that for all $C \in \mathcal{C}$,

$$\mathrm{cost}_z(S, C) := \textstyle\sum_{x\in S} w(x) \cdot d^z(x, C) \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(P, C). \tag{3}$$

Ideally, we would measure the effectiveness of a coreset S by how well the clustering solution it yields on S generalizes to the true dataset P. This leads to the *ideal quality measure*:

$$r_P(\mathsf{C}(S)) := \tfrac{\mathrm{cost}_z(P,\mathsf{C}(S))}{\mathsf{OPT}_P},$$

where $\mathsf{OPT}_P := \min_{C\in\mathcal{C}} \mathrm{cost}_z(P, C)$. Note that $r_P(\mathsf{C}(S)) \geq 1$, with equality if $\mathsf{C}(S) = \mathsf{C}(P)$. However, computing $\mathsf{C}(S)$ is generally NP-hard, and even estimating $r_P(\mathsf{C}(S))$ is infeasible in

[2] The variance of each D_j can be fixed to any $t > 0$ since we can scale each point in the dataset by $\frac{1}{t}$.

the noisy setting where P is unobservable. To account for approximate clustering, we define for any $\alpha \geq 1$ the set of α-approximate center sets for S: $\mathcal{C}_\alpha(S) := \{C \in \mathcal{C} : r_S(C) \leq \alpha\}$, where $r_S(C) := \mathrm{cost}_z(S, C)/\mathsf{OPT}_S$. We then define the *worst-case quality* over this set:

$$r_P(S, \alpha) := \max_{C \in \mathcal{C}_\alpha(S)} \frac{\mathrm{cost}_z(P,C)}{\mathsf{OPT}_P}. \tag{4}$$

This function is monotonically increasing in α, and $r_P(S, 1) = r_P(\mathsf{C}(S))$. We focus on settings where α is close to 1, such as when a PTAS is available for (k, z)-CLUSTERING. As discussed in Section 1, directly evaluating $r_P(S, \alpha)$ is computationally hard and, in noisy settings, fundamentally infeasible—motivating the need for surrogate metrics.

The metric Err**.** In the noise-free setting, a standard surrogate for $r_P(S, \alpha)$ is the relative error:

$$\mathsf{Err}(S, P) := \sup_{C \in \mathcal{C}} \frac{|\mathrm{cost}_z(S,C) - \mathrm{cost}_z(P,C)|}{\mathrm{cost}_z(S,C)}.$$

This quantity provides a bound on how much the clustering cost on S deviates from that on P, uniformly over all center sets. In particular, for the optimal center set $\mathsf{C}(S)$ of S, we have: $\frac{|\mathrm{cost}_z(S,\mathsf{C}(S)) - \mathrm{cost}_z(P,\mathsf{C}(S))|}{\mathrm{cost}_z(S,\mathsf{C}(S))} \leq \mathsf{Err}(S, P)$, which implies $\mathrm{cost}_z(P, \mathsf{C}(S)) \in \left[\frac{1}{1+\mathsf{Err}(S,P)}, (1 + \mathsf{Err}(S, P))\right] \cdot \mathrm{cost}_z(S, \mathsf{C}(S))$. Since $\mathsf{C}(S)$ is optimal for S, this yields a lower bound on OPT_P and extends to all α-approximate center sets:

$$\forall C \in \mathcal{C}_\alpha(S), \quad \mathrm{cost}_z(P, C) \leq (1+\mathsf{Err}(S, P)) \cdot \mathrm{cost}_z(S, C) \leq (1+\mathsf{Err}(S, P)) \cdot \alpha \cdot \mathrm{cost}_z(S, \mathsf{C}(S)).$$

Combining these gives:

$$r_P(S, \alpha) \leq (1 + \mathsf{Err}(S, P))^2 \cdot \alpha. \tag{5}$$

This justifies the use of $\mathsf{Err}(S, P)$ as a surrogate for $r_P(S, \alpha)$ in the noise-free setting.

In the noisy setting, however, P is unobservable. A natural alternative is to compute $\mathsf{Err}(S, \widehat{P})$ instead. This raises the question: how does $\mathsf{Err}(S, \widehat{P})$ relate to the true coreset quality $r_P(S, \alpha)$? We explore this relationship and show how $\mathsf{Err}(S, \widehat{P})$ can still guide coreset construction (see Theorem 3.1).

The new metric. While $\mathsf{Err}(S, P)$ is a valid surrogate in the noise-free setting, its adaptation to noisy data via $\mathsf{Err}(S, \widehat{P})$ is problematic: noise inflates clustering costs on $\widehat{P}$, weakening its connection to the true quality $r_P(S, \alpha)$. To mitigate this, we introduce a new surrogate metric that compares the *relative* quality of a center set on P versus S:

$$\mathsf{Err}_\alpha(S, P) := \sup_{C \in \mathcal{C}_\alpha(S)} \frac{r_P(C)}{r_S(C)} - 1. \tag{6}$$

Since $r_S(C) \leq \alpha$, we have $\mathsf{Err}_\alpha(S, P) \geq \sup_{C \in \mathcal{C}_\alpha(S)} \frac{r_P(C)}{\alpha} - 1$, and therefore $r_P(S, \alpha) \leq (1 + \mathsf{Err}_\alpha(S, P)) \cdot \alpha$. This bound justifies Err_α as a surrogate for $r_P(S, \alpha)$. The metric is monotonic in α, independent of the noise distribution, and aligns better with coreset quality under noise than Err, which measures absolute cost deviation.

In practice, we compute $\mathsf{Err}_\alpha(S, \widehat{P})$ as a proxy for $\mathsf{Err}_\alpha(S, P)$. We analyze this in Theorem 3.3, showing that it can guide coreset construction under noise. Notably, Err_α extends prior approximation-ratio-based coreset ideas [23, 53] to the noisy setting.

Comparing two metrics under noise. We illustrate the advantage of Err_α over Err through a simple 1-MEANS example in $\mathbb{R}$, with $k = d = 1$, $z = 2$, and $\alpha = 1$.

Let $P = P_- \cup P_+$, where P_- has $n/2$ points at -1 and P_+ has $n/2$ points at 1. The optimal center is $\mathsf{C}(P) = 0$ with $\mathsf{OPT}_P = n$. Now consider $\widehat{P}$, generated under noise model I with $\theta = 1$ and $D_j = \mathcal{N}(0, 1)$. This adds i.i.d. Gaussian noise to each point, inflating clustering cost by roughly $2n$ for any center c, i.e., $\mathrm{cost}_z(\widehat{P}, c) - \mathrm{cost}_z(P, c) \approx 2n$. Hence, $\mathsf{Err}(\widehat{P}, P) \approx \frac{2n}{3n} = \frac{2}{3}$, yielding a bound:

$$r_P(\widehat{P}, 1) \leq (1 + \mathsf{Err}(\widehat{P}, P))^2 \lesssim \tfrac{25}{9}.$$

In contrast, because the noise ξ_p averages out, the empirical center satisfies $\mathsf{C}(\widehat{P}) \in [-\sqrt{1/n}, \sqrt{1/n}]$ with high probability, and: $\mathrm{cost}_z(P, \mathsf{C}(\widehat{P})) \leq n + 1$. This implies $\mathsf{Err}_1(\widehat{P}, P) \leq \frac{1}{n}$, and therefore:

$$r_P(\widehat{P}, 1) \leq 1 + \tfrac{1}{n}.$$

Thus, Err_α provides a much tighter estimate of r_P under noise, by compensating for the uniform cost inflation that Err fails to account for. We elaborate on this example and provide additional comparisons in Section B.3.

Finally, we note that our setting differs fundamentally from "robust" coreset models [38, 52, 54], which assume direct access to the clean dataset P. In contrast, we construct coresets directly from noisy observations $\widehat{P}$; see Section A for a detailed comparison.

3 Theoretical results

This section gives theoretical guarantees for coreset construction under noise. We present two algorithms for k-MEANS ($z = 2$) using noise model I, based on the surrogate metrics Err (Theorem 3.1) and Err_α (Theorem 3.3). While both yield bounds on coreset quality, the Err_α-based method achieves smaller coresets and tighter guarantees under mild structural assumptions. We write cost for cost_2 throughout. Extensions to other noise models and (k, z)-CLUSTERING are in Section F.

We begin with Err-based coresets. Theorem 3.1 extends its use to noisy data and bounds $r_P(S, \alpha)$ in terms of θ, d, and n. See Section C for the proof.

Theorem 3.1 (Coreset using Err). *Let $\widehat{P}$ be drawn from P via noise model I with known $\theta \geq 0$. Let $\varepsilon \in (0, 1)$ and fix $\alpha \geq 1$. Let $\mathcal{A}$ be an algorithm that constructs a weighted subset $S \subset \widehat{P}$ for k-MEANS of size $\mathcal{A}(\varepsilon)$ and with guarantee $\mathsf{Err}(S, \widehat{P}) \leq \varepsilon$. Then with probability at least 0.9,*

$$\mathsf{Err}(S, P) \leq \varepsilon + O\left(\frac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\frac{\theta n d}{\mathsf{OPT}_P}}\right) \text{ and } r_P(S, \alpha) \leq \left(1 + \varepsilon + O\left(\frac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\frac{\theta n d}{\mathsf{OPT}_P}}\right)\right)^2 \cdot \alpha.$$

To ensure $\mathsf{Err}(S, \widehat{P}) \leq \varepsilon$, we may use an importance sampling algorithm [10] with coreset size

$$\mathcal{A}(\varepsilon) = \widetilde{O}(\min\{k^{1.5}\varepsilon^{-2}, k\varepsilon^{-4}\}), \tag{7}$$

which matches the state of the art in the noise-free setting. However, the resulting bounds for $\mathsf{Err}(S, P)$ and $r_P(S, \alpha)$ incur an additive term $O(\frac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\frac{\theta n d}{\mathsf{OPT}_P}})$, due to the gap $\mathsf{Err}(\widehat{P}, P)$ (see Lemma C.2). As discussed in Section 2, this gap can be overly conservative—especially when noise uniformly inflates clustering cost. In such cases, $\mathsf{Err}_\alpha(\widehat{P}, P) \ll \mathsf{Err}(\widehat{P}, P)$, as shown in our earlier example. While this inflation always occurs when $k = 1$, it may not persist for general k, where noise can change point-to-center assignments between P and $\widehat{P}$. To address this, we introduce structural assumptions that preserve assignments under noise.

Assumptions on data. To theoretically separate the performance of Err and Err_α, we impose mild structural assumptions on the dataset to ensure that point-to-center assignments remain stable under noise. A natural but strong assumption is to posit a generative model, such as a Gaussian mixture $\sum_{\ell=1}^{k} \frac{1}{k} N(\mu_\ell, 1)$, where the means $\mu_\ell \in \mathbb{R}^d$ are well separated, e.g., $\|\mu_\ell - \mu_{\ell'}\| \geq n$ [35, 57]. This would make the assignments between P and $\widehat{P}$ nearly identical. However, this is more than we require—we instead use structural assumptions that capture the relevant properties directly.

The first is *cost stability*, a widely studied notion in clustering and coreset literature [71, 6, 59, 2, 25, 10]. Let $\mathsf{OPT}_P(m)$ denotes the optimal cost of m-means on P. For $\gamma > 0$, a dataset P is γ-cost-stable if

$$\frac{\mathsf{OPT}_P(k-1)}{\mathsf{OPT}_P(k)} \geq 1 + \gamma.$$

As γ increases, the clusters are more well-separated, making assignments more robust to noise. In our setting, cost stability ensures that the assignment changes between P and $\widehat{P}$ remain limited, and we specify the required value of γ as a function of the noise level θ in Assumption 3.2. This assumption is also necessary to distinguish $\mathsf{Err}(\widehat{P}, P)$ from $\mathsf{Err}_\alpha(\widehat{P}, P)$. As shown in Appendix E.1, when cost stability is weak, the two metrics can behave similarly; e.g., in a 3-means instance with $\gamma = 1$, we find $\mathsf{Err}(\widehat{P}, P) \approx \mathsf{Err}_1(\widehat{P}, P)$.

We also assume that P does not contain strong outliers. Let $P_1, \ldots, P_k$ denote the partition of P induced by its optimal center set $\mathsf{C}(P)$. For each cluster P_i, define the average and maximum radius:

$$\overline{r}_i := \sqrt{\frac{1}{|P_i|} \sum_{p \in P_i} d^2(p, \mathsf{C}(P)_i)}, \quad r_i := \max_{p \in P_i} d(p, \mathsf{C}(P)_i).$$

We assume $r_i \le 8\overline{r}_i$ for all i, which rules out heavy-tailed clusters and helps distinguish noise from genuine outliers. This choice of 8 is made to simplify analysis and is satisfied by real-world datasets; see Table 2.

Assumption 3.2 (Cost stability and limited outliers). Given $\alpha \ge 1$ and $\theta \in [0,1]$, assume P is γ-cost-stable with

$$\gamma = O(\alpha) \cdot \left(1 + \frac{\theta n d \log^2(kd/\sqrt{\alpha-1})}{\mathsf{OPT}_P}\right),$$

and that $r_i \le 8\overline{r}_i$ for all $i \in [k]$.

Under these assumptions, the following theorem gives performance guarantees for a coreset algorithm based on the Err_α metric. The proof appears in Section D.

Theorem 3.3 (Coreset using the Err_α metric). *Let $\widehat{P}$ be an observed dataset drawn from P under the noise model I with known parameter $\theta \in [0, \frac{\mathsf{OPT}_P}{nd}]$. Let $\varepsilon \in (0,1)$ and fix $\alpha \in [1,2]$. Under Assumption 3.2, there exists a randomized algorithm that constructs a weighted $S \subset \widehat{P}$ for k-MEANS of size $O(\frac{k \log k}{\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}_P}} + \frac{(\alpha-1)k\log k}{(\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}_P})^2})$ and with guarantee $\mathsf{Err}_\alpha(S, \widehat{P}) \le \varepsilon$ with probability at least 0.99. Moreover,*

$$\mathsf{Err}_\alpha(S,P) \le \varepsilon + O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta kd \mathsf{OPT}_P} + \theta nd}{\mathsf{OPT}_P}\right) \text{ and}$$

$$r_P(S,\alpha) \le \left(1 + \varepsilon + O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta kd \mathsf{OPT}_P} + \theta nd}{\mathsf{OPT}_P}\right)\right) \cdot \alpha.$$

We consider the regime $\theta \le \frac{\mathsf{OPT}_P}{nd}$, ensuring that noise does not dominate the clustering cost, i.e., $\mathrm{cost}(\widehat{P}, C) = O(\mathrm{cost}(P, C))$. The coreset size depends on the knowledge of OPT_P, but we later show how to remove this dependence. In the special case $\varepsilon = 0$ and $\alpha = 1$, the bound becomes $\mathsf{Err}_\alpha(S,P) = \mathsf{Err}_1(\widehat{P}, P) = O\left(\frac{\theta kd}{\mathsf{OPT}_P}\right)$, which is a factor k/n smaller than the bound $\mathsf{Err}(\widehat{P}, P) = O\left(\frac{\theta nd}{\mathsf{OPT}_P} + \sqrt{\frac{\theta nd}{\mathsf{OPT}_P}}\right)$ from Theorem 3.1. This supports the use of Err_α under Assumption 3.2. Subsequently, we also provide an interpretation for different terms in the bounds of Err and Err_α.

Comparison of coreset performance using two metrics. We now compare the coreset size and error bounds for $r_P(S,\alpha)$ achieved by Theorem 3.1 (**CN**) and Theorem 3.3 ($\mathbf{CN}_\alpha$). Let $\varepsilon = 1/\mathrm{poly}(k)$ and $\alpha = 1 + c\varepsilon$ for a constant $0 < c < 0.5$, ensuring that an α-approximate center set can be efficiently computed [62]. Under this setting, the stability parameter in Assumption 3.2 becomes $\gamma = O(1 + \log^2(kd/\varepsilon))$, since $\theta \le \mathsf{OPT}_P/(nd)$.

Coreset size. When $\frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}_P} < c\varepsilon/2$, the coreset size from $\mathbf{CN}_\alpha$ is $\widetilde{O}(k/\varepsilon)$, which improves over the size $\widetilde{O}(\min\{k^{1.5}/\varepsilon^2,\, k/\varepsilon^4\})$ of **CN** by a factor of $\sqrt{k}/\varepsilon$.

Bound on r_P. The error bound in Theorem 3.3 includes a term $O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta kd \mathsf{OPT}_P} + \theta nd}{\mathsf{OPT}_P}\right)$, whose dominant component, when $n \gg \mathrm{poly}(k)$, is at most $O\left(\frac{1}{\mathrm{poly}(k)} \cdot \frac{\theta nd}{\mathsf{OPT}_P}\right)$. This is again tighter than the bound from Theorem 3.1, which scales as $O\left(\frac{\theta nd}{\mathsf{OPT}_P} + \sqrt{\frac{\theta nd}{\mathsf{OPT}_P}}\right)$, by a factor of at least $\mathrm{poly}(k)$.

For a general $\varepsilon \in (0,1)$, we provide another example that demonstrates improved coreset performance for $\mathbf{CN}_\alpha$. Let $\alpha = 1 + \varepsilon$ and $\theta = \frac{\mathrm{OPT}_P}{nd \cdot \mathrm{poly}(k)}$. Following the same analysis as above, we observe that the coreset size of $\mathbf{CN}_\alpha$ improves over that of **CN** by a factor of $\sqrt{k}/\varepsilon$, while the error bound improves by at least a $\mathrm{poly}(k)$ factor.

Overall, $\mathbf{CN}_\alpha$ yields smaller coresets and tighter theoretical guarantees for $r_P(S,\alpha)$, improving over **CN** by at least a factor of $\mathrm{poly}(k)$, owing to the more noise-aware nature of the Err_α metric.

Applying Theorems 3.1 and 3.3 in practice. In practical settings, we often aim to construct a coreset such that $r_P(S,\alpha) \le (1+\varepsilon)\cdot\alpha$. There are two ways to achieve this:

1. Use $\mathbf{CN}(\varepsilon')$ from Theorem 3.1, with $\varepsilon' = \varepsilon - O\left(\frac{\theta nd}{\mathsf{OPT}_P} + \sqrt{\frac{\theta nd}{\mathsf{OPT}_P}}\right)$.

2. Use $\mathbf{CN}_\alpha(\varepsilon_\alpha)$ from Theorem 3.3, with $\varepsilon_\alpha = \varepsilon - O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta kd\mathsf{OPT}_P}+\theta nd}{\mathsf{OPT}_P}\right)$.

Both approaches require an estimate of OPT_P, which can be obtained by computing an $O(1)$-approximate center set $\widehat{C}$ for $\widehat{P}$ and using $\mathrm{cost}(\widehat{P}, \widehat{C})$ as a proxy; see discussion in Section E.3. When $\theta \le \mathsf{OPT}_P/(nd)$, this yields a valid approximation. Notably, both $\mathbf{CN}$ and $\mathbf{CN}_\alpha$ already compute such a $\widehat{C}$ as a first step, so no additional overhead is incurred.

A practical question is when to prefer $\mathbf{CN}_\alpha(\varepsilon_\alpha)$ over $\mathbf{CN}(\varepsilon')$. This depends on whether P satisfies Assumption 3.2. Although P is unobserved, the observed dataset $\widehat{P}$ often satisfies an approximate version of the assumption. We discuss how to verify this in Section E.2, allowing practitioners to choose the tighter construction when applicable. In practice, we note that $\mathbf{CN}_\alpha$ performs well even when the assumptions are violated; see Section 4.

Key ideas in the proof of Theorem 3.1. The goal is to relate $\mathsf{Err}(S, \widehat{P})$ to $\mathsf{Err}(S, P)$, from which the bound on $r_P(S, \alpha)$ follows via Equation (5). The Err metric satisfies a standard composition bound: $\mathsf{Err}(S, P) \le \mathsf{Err}(S, \widehat{P}) + O(\mathsf{Err}(\widehat{P}, P))$ (Lemma C.3). Thus, it suffices to bound $\mathsf{Err}(\widehat{P}, P)$. We show $\mathsf{Err}(\widehat{P}, P) = O\left(\frac{\theta nd}{\mathsf{OPT}} + \sqrt{\frac{\theta nd}{\mathsf{OPT}}}\right)$ (Lemma C.2). This is obtained by controlling the difference:

$$\mathrm{cost}(\widehat{P}, C) - \mathrm{cost}(P, C) \le \textstyle\sum_{p\in P} (\|\xi_p\|_2 + \langle \xi_p, p - c\rangle)$$

for all $C \in \mathcal{C}$, and showing that the sum on the r.h.s. above normalized by $\mathrm{cost}(\widehat{P}, C)$ is $O\left(\frac{\theta nd}{\mathsf{OPT}} + \sqrt{\frac{\theta nd}{\mathsf{OPT}}}\right)$. The proof uses concentration from the independence of noise $\{\xi_p\}$, combined with bounded higher moments ensured by the Bernstein condition.

We note that the first term, $O\left(\frac{\theta nd}{\mathsf{OPT}_P}\right)$, is information-theoretically necessary and cannot be improved in the worst case; see Section C.2. The second term, $O\left(\sqrt{\frac{\theta nd}{\mathsf{OPT}_P}}\right)$, arises from a loose upper bound in our analysis and is not known to be tight. This term is introduced when bounding the cumulative error over all points and center sets C by applying Cauchy-Schwarz:

$$\sum_{p\in P} \langle \xi_p, p - c\rangle \le \sqrt{\left(\sum_p \|\xi_p\|_2^2\right) \cdot \left(\sum_p d^2(p, C)\right)} \le \sqrt{\theta nd \cdot \mathrm{cost}(P, C)}.$$

This worst-case analysis assumes full alignment of all error contributions, which may be overly pessimistic. A more refined analysis could potentially tighten or remove this term by accounting for error cancellation.

Key ideas in the proof of Theorem 3.3. The proof has two parts: (1) designing a coreset algorithm that guarantees $\mathsf{Err}_\alpha(S, \widehat{P}) \le \varepsilon$, and (2) bounding the gap between $\mathsf{Err}_\alpha(S, \widehat{P})$ and $\mathsf{Err}_\alpha(S, P)$.

Coreset design. We first construct a coreset S in the noise-free setting ($\widehat{P} = P$) such that $\mathsf{Err}_\alpha(S, \widehat{P}) \le \varepsilon$. Under Assumption 3.2, we partition $\widehat{P}$ into k well-separated clusters $\widehat{P}_1, \ldots, \widehat{P}_k$, each of diameter at most $2r_i$. From each $\widehat{P}_i$, the algorithm samples a uniform subset S_i of size $O\left(\frac{\log k}{\varepsilon-\Delta} + \frac{(\alpha-1)\log k}{(\varepsilon-\Delta)^2}\right)$, where $\Delta := \frac{\sqrt{\alpha-1}\theta nd}{\alpha\mathsf{OPT}_P}$. Let $S = \bigcup_i S_i$. Standard concentration bounds imply that $\mathsf{C}(S)$ is close to $\mathsf{C}(\widehat{P})$, and $\mathsf{OPT}_S \lesssim \mathsf{OPT}_{\widehat{P}}$, which yield $\mathsf{Err}_1(S, \widehat{P}) \le \varepsilon$ (Lemma D.4).

To extend this to $\mathsf{Err}_\alpha(S, \widehat{P}) \le \varepsilon$, we leverage a geometric property: for any $C \in \mathcal{C}_\alpha(S)$, each center c_i lies within distance $O\left(\sqrt{\frac{\alpha(\mathsf{OPT}+\theta nd\log^2(kd/\sqrt{\alpha-1}))}{n_i}}\right)$ of $\mathsf{C}(S)_i$ (Lemma D.3). This follows from cost stability, which ensures consistent cluster structure for all such C.

Directly applying this algorithm to the noisy dataset introduces several analytical challenges. First, noise can significantly increase the diameter of each cluster $\widehat{P}_i$, weakening the closeness guarantee between $\mathsf{C}(S)$ and $\mathsf{C}(\widehat{P})$. Second, highly noisy points may shift cluster assignments across partitions $\widehat{P}i$, breaking the geometric structure required for all $C \in \mathcal{C}_\alpha(S)$. To address both issues, we eliminate extremely noisy points from $\widehat{P}$ (see Line 3 of Algorithm 1), which is the key innovation of our

Algorithm 1 A coreset algorithm $\mathbf{CN}_\alpha$ using the Err_α metric under the noise model I

Input: a noisy dataset $\widehat{P}$ derived from P under noise model I with $\theta > 0$, $\varepsilon \in (0,1)$, $\alpha \in [1,2]$, and an $O(1)$-approximate center set $\widehat{C} = \{\widehat{c}_1, \ldots, \widehat{c}_k\} \in \mathcal{C}$.
Output: a weighted set $S \subseteq \widehat{P}$

1: Decompose $\widehat{P}$ into k clusters $\widehat{P}_i$ by $\widehat{C}$, where $\widehat{P}_i = \left\{p \in \widehat{P} : \arg\min_{c \in \widehat{C}} d(p,c) = \widehat{c}_i\right\}$.
2: For each $i \in [k]$, compute $\widehat{r}_i = \sqrt{\mathrm{cost}(\widehat{P}_i, \widehat{c}_i)/|\widehat{P}_i|}$.
3: Compute $P'_i = \widehat{P}_i \cap B_i$, where ball $B_i := B(\widehat{c}_i, R_i)$ with $R_i := 3\widehat{r}_i + O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})$.
4: For each $i \in [k]$, take a uniform sample S_i of size $O(\frac{\log k}{\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}}} + \frac{(\alpha-1)\log k}{(\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}})^2})$.
5: Return $S = \bigcup_{i\in[k]} S_i$ with $w(p) = \frac{|P'_i|}{|S_i|}$ for $p \in S_i$.

algorithm. In contrast to the existing approaches, our method explicitly excludes high-noise points and focuses the coreset on points whose assignments are reliable. This noise-aware sampling step is critical to preserving geometric stability. Moreover, we show that the effect of the removed points on the location of $\mathsf{C}(P')$ is negligible (Lemma D.5). Together, these steps ensure that Algorithm 1 achieves the desired bound $\mathsf{Err}_1(S, \widehat{P}) \le \varepsilon$.

Metric bounds. A natural approach is to compose $\mathsf{Err}_\alpha(S, \widehat{P})$ with $\mathsf{Err}_\alpha(\widehat{P}, P)$, as with the Err metric. However, this fails because a set $C \in \mathcal{C}_\alpha(S)$ may only satisfy $C \in \mathcal{C}_{\alpha(1+\varepsilon)}(\widehat{P})$, preventing direct use of $\mathsf{Err}_\alpha(\widehat{P}, P)$. To resolve this, we instead compose $\mathsf{Err}_\alpha(S, \widehat{P})$ with $\mathsf{Err}_{\alpha(1+\varepsilon)}(\widehat{P}, P)$, yielding:

$$\mathsf{Err}_\alpha(S, P) \le \varepsilon + O(\mathsf{Err}_{\alpha(1+\varepsilon)}(\widehat{P}, P)).$$

The remaining task is to bound $\mathsf{Err}_{\alpha(1+\varepsilon)}(\widehat{P}, P)$. Assumption 3.2 allows us to analyze this quantity cluster by cluster. For simplicity, we illustrate the argument on one cluster P_i and its noisy counterpart $\widehat{P}_i$. Note that $\mathsf{C}(\widehat{P}_i) = \mathsf{C}(P_i) + \frac{1}{|P_i|}\sum_{p\in P_i}\xi_p$ by the optimality condition of 1-MEANS. Then:

$$\mathrm{cost}(P_i, \mathsf{C}(\widehat{P}_i)) = \mathsf{OPT}_{P_i} + \tfrac{1}{|P_i|}\left\|\textstyle\sum_{p\in P_i}\xi_p\right\|_2^2 \text{ and hence,}$$

$$\mathsf{Err}_1(\widehat{P}_i, P_i) = \tfrac{\mathrm{cost}(P_i, \mathsf{C}(\widehat{P}_i))}{\mathsf{OPT}_{P_i}} - 1 = O\left(\tfrac{\theta d}{\mathsf{OPT}_{P_i}}\right),$$

using standard concentration for sums of independent noise vectors (Lemma D.12). This is significantly smaller than $\mathsf{Err}(\widehat{P}_i, P_i)$, which lacks cancellation.

Aggregating across clusters and plugging into the composition bound gives:

$$\mathsf{Err}_\alpha(S, P) \le \varepsilon + O\left(\tfrac{\theta kd}{\mathsf{OPT}_P} + \tfrac{\sqrt{\alpha(1+\varepsilon)-1}}{\alpha}\cdot\tfrac{\sqrt{\theta kd\mathsf{OPT}_P}+\theta nd}{\mathsf{OPT}_P}\right).$$

This bound matches Theorem 3.3, up to an extra $(1+\varepsilon)$ factor inside the square root. This artifact can be removed by composing through P' instead of $\widehat{P}$ (see Lemmas D.4 and D.5). Note that the term $O\left(\frac{\theta kd}{\mathsf{OPT}_P}\right)$ accounts for center drift due to noise. Intuitively, each optimal center in $\mathsf{C}(P)$ may shift by $O(\theta d)$ under noise, resulting in a total movement of $O(\theta kd)$. This implies that $\mathsf{C}(\widehat{P})$ forms a $(1 + \frac{\theta kd}{\mathsf{OPT}_P})$-approximate center set for P, yielding the corresponding error term. The final term,

$$O\left(\frac{\sqrt{\alpha-1}}{\alpha}\cdot\frac{\sqrt{\theta kd\cdot\mathsf{OPT}_P}+\theta nd}{\mathsf{OPT}_P}\right),$$

captures the additional approximation error from using α-approximate center sets rather than exact optima. It scales with the gap between α and 1, and vanishes as $\alpha \to 1$.

Thus, we complete the proof of Theorem 3.3. Note that other noise models primarily affect the concentration bounds of the term $\left\|\sum_{p\in P_i}\xi_p\right\|_2^2$ in the analysis; see details in Section F.

4 Empirical results

We now evaluate the empirical performance of our proposed coreset algorithms on real-world datasets under varying noise levels and tolerance thresholds. The goal is to test whether our theoretical guarantees translate into practical improvements in coreset size, accuracy, and robustness. We also assess how well the theoretical bounds track actual performance in both clean and noisy data regimes.

Setup. We consider the k-MEANS problem on the `Adult` [61] and `Census1990` [68] datasets from the UCI Repository. Both satisfy the limited outlier assumption but exhibit small cost-stability constants γ; see Table 2. We set $k = 10$. We perturb each dataset under noise model I, using Gaussian noise with $\theta \in \{0, 0.01, 0.05, 0.25\}$, where $\theta = 0$ denotes the noise-free case. For varying tolerance levels $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$, we construct a coreset S from $\widehat{P}$ using $\mathbf{CN}$ and $\mathbf{CN}_\alpha$. For the initialization of our algorithms, we run k-means++ with $\max\_\text{iter} = 5$ on $\widehat{P}$ to obtain a fast $O(1)$-approximate solution. Implementation details appear in Section G.1.

Metrics. For each coreset S, we report: (i) coreset size $|S|$, (ii) empirical approximation ratio $\widetilde{r}_S := \frac{\text{cost}(P, C_S)}{\text{cost}(P, C_P)}$, and (iii) tightness ratio $\kappa_S := \frac{\widetilde{r}_S}{u_S}$, where u_S is the theoretical bound for $r_P(C_S)$ from Theorems 3.1 and 3.3. The approximation ratio $\widetilde{r}_S$ measures how well the coreset solution approximates the true clustering cost on P. To calculate $\widetilde{r}_S$, we obtain C_S and C_P by running k-means++ 10 times (default settings, varied seeds) on S and P separately and selecting the best solution for each. Letting $\widetilde{\mathsf{OPT}} = \text{cost}(\widehat{P}, C_{\widehat{P}})$, we set:

$$u_S = \begin{cases} (1 + \varepsilon + \frac{\theta n d}{\widetilde{\mathsf{OPT}}} + \sqrt{\frac{\theta n d}{\widetilde{\mathsf{OPT}}}})^2 & (\mathbf{CN}) \\ 1 + \varepsilon + \frac{\theta k d}{\widetilde{\mathsf{OPT}}} + \frac{\theta n d}{\widetilde{\mathsf{OPT}}} & (\mathbf{CN}_\alpha). \end{cases}$$

A value $\kappa_S \leq 1$ implies the empirical ratio is below the theoretical bound; values closer to 1 indicate tighter guarantees. All experiments are repeated 10 times, and average metrics are reported.

Analysis. Table 1 reports results on the `Adult` dataset across noise levels. Results on `Census1990` appear in Section G.2 and follow similar trends. In all settings, $\mathbf{CN}_\alpha$ consistently produces smaller coresets and achieves κ_S values closer to 1 than $\mathbf{CN}$. For example, at $\varepsilon = 0.2, \theta = 0.01$, $\mathbf{CN}_\alpha$ yields a coreset of size 1940 (82% of $\mathbf{CN}$'s 2371), with $\kappa_S = 0.937$ vs. 0.596 for $\mathbf{CN}$ —indicating much tighter empirical bounds.

We also find that for higher tolerance levels ($\varepsilon \geq 0.2$), $\mathbf{CN}_\alpha$ often yields better empirical approximation: e.g., for $\varepsilon = 0.2, \theta = 0.01$, $\mathbf{CN}_\alpha$ attains $\widetilde{r}_S = 1.156$ vs. 1.193 for $\mathbf{CN}$. This suggests that $\mathbf{CN}_\alpha$ outperforms even in noise-free settings, indicating its potential value beyond noisy data applications. Moreover, in the noise-free case ($\theta = 0$), empirical ratios $\widetilde{r}_S$ consistently satisfy $\widetilde{r}_S \leq 1 + \varepsilon$ for $\mathbf{CN}_\alpha$ —further validating the theoretical guarantees.

Additional results under Laplace, uniform, non-i.i.d., and noise model II appear in Section F.1. These confirm the robustness and broader applicability of $\mathbf{CN}_\alpha$ across diverse noise regimes.

Summary. Overall, the empirical results demonstrate that $\mathbf{CN}_\alpha$ consistently achieves tighter approximation guarantees with smaller coresets across a range of noise levels. These findings validate the practical utility of the Err_α-based construction and suggest that it remains effective even when the theoretical assumptions (e.g., exact cost-stability) are only approximately satisfied. The robustness of $\mathbf{CN}_\alpha$ across multiple datasets and noise models underscores its suitability for integration into practical data preprocessing pipelines.

5 Conclusion, limitations, and future work

This paper studies the practically relevant problem of coreset construction for clustering in the presence of noise. The main contributions are two new algorithms that construct coresets with provable guarantees relative to the true (unobserved) dataset, one based on adapting the traditional surrogate metric Err, and the other introducing a new metric Err_α. We prove that the algorithm based on Err_α yields smaller coreset sizes and tighter performance bounds, assuming the dataset satisfies certain natural assumptions that are necessary in worst-case scenarios. Our analysis quantifies how noise impacts clustering costs and perturbs the optimal center sets, relying on properties such as noise cancellation and the concentration of the empirical center $\mathsf{C}(\widehat{P})$. The new metric Err_α may

Table 1: Results of `Adult` dataset under noise model I with Gaussian noise. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the tightness ratio of its empirical approximation ratio over the theoretical bound.

(a) $\theta = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1318	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.040	1.080	1.183	1.278	1.200
	$\mathbf{CN}_\alpha$	1.085	1.115	1.150	1.197	1.124
κ_S	$\mathbf{CN}$	0.859	0.817	0.821	0.818	0.710
	$\mathbf{CN}_\alpha$	0.986	0.969	0.959	0.958	0.865

(b) $\theta = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.037	1.081	1.193	1.187	1.244
	$\mathbf{CN}_\alpha$	1.114	1.069	1.156	1.154	1.145
κ_S	$\mathbf{CN}$	0.600	0.581	0.596	0.554	0.543
	$\mathbf{CN}_\alpha$	0.984	0.904	0.937	0.899	0.859

(c) $\theta = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.027	1.061	1.217	1.152	1.200
	$\mathbf{CN}_\alpha$	1.120	1.108	1.133	1.186	1.154
κ_S	$\mathbf{CN}$	0.369	0.359	0.389	0.348	0.343
	$\mathbf{CN}_\alpha$	0.886	0.844	0.830	0.839	0.788

(d) $\theta = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.044	1.049	1.067	1.234	1.341
	$\mathbf{CN}_\alpha$	1.123	1.163	1.173	1.158	1.222
κ_S	$\mathbf{CN}$	0.131	0.127	0.125	0.139	0.146
	$\mathbf{CN}_\alpha$	0.585	0.590	0.581	0.559	0.576

also have independent utility beyond the noisy setting, for instance, enabling further coreset size reduction in noise-free tasks where preserving near-optimal solutions suffices—such as in regression. Empirical evaluations strongly support the theoretical findings, demonstrating robust performance across a broad range of real-world datasets (which may violate the theoretical assumptions) and under diverse noise models. These results suggest that the proposed algorithms can be readily integrated into existing coreset-based clustering pipelines to improve robustness in noisy environments.

One limitation lies in extending the use of the Err_α metric to coreset construction in the streaming model. Although Err_α satisfies a composition property akin to Err, it may lack the *mergeability* property (the union of coresets is a coreset for the union of datasets), which is crucial for enabling coreset construction in streaming settings. As a result, adapting the Err_α metric to the streaming model remains an technically challenging problem. As an initial step toward this challenge, we present a relaxed version of mergeability for the Err_α metric in Section B.4, which may be applicable in certain scenarios.

Besides this, our work opens several promising avenues for future research. A natural next step is to explore weaker assumptions that are applicable beyond worst-case scenarios, thereby improving the generalizability of our results. Another key direction is to extend our analysis to more realistic noise models, including those with dependencies across data points, heavy-tailed noise, or adversarially structured noise. Furthermore, our work focuses on clustering problems in Euclidean spaces. Extending coreset construction under noise to general metric spaces presents an interesting direction for future research. However, without access to Euclidean coordinates, it becomes nontrivial to define additive noise on individual points. A natural alternative in such settings is to model noise additively on pairwise distances (edges) rather than on point coordinates. In addition to clustering, it would be valuable to investigate how noise affects coreset construction in other learning tasks such as regression and classification. Furthermore, studying connections between coreset constructions and other robustness notions in clustering could yield new insights. Finally, our empirical results suggest that coresets based on the Err_α metric may outperform those based on Err even in the absence of noise. Characterizing the conditions under which Err_α provides tighter guarantees than Err in the noise-free setting is an intriguing direction for future theoretical study.

We anticipate positive societal impact from this work by enabling more accurate and reliable data analysis pipelines. These improvements could benefit various sectors, including healthcare, finance, and technology, by enabling more robust data-driven decisions.

Acknowledgments

LH acknowledges support from the State Key Laboratory of Novel Software Technology, the New Cornerstone Science Foundation, and NSFC Grant No. 625707396. ZL was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant. NKV was supported in part by NSF Grant CCF-2112665.

References

[1] Margareta Ackerman, Shai Ben-David, David Loker, and Sivan Sabato. Clustering oligarchies. In *Artificial Intelligence and Statistics*, pages 66–74. PMLR, 2013.

[2] Manu Agarwal, Ragesh Jaiswal, and Arindam Pal. k-means++ under approximation stability. In *Theoretical Computer Science*, 2013.

[3] Parag Agrawal, Anish Das Sarma, Jeffrey D. Ullman, and Jennifer Widom. Foundations of uncertain-data integration. *Proc. VLDB Endow.*, 3(1):1080–1090, 2010.

[4] David Arthur, Bodo Manthey, and Heiko Röglin. Smoothed analysis of the k-means method. *Journal of the ACM (JACM)*, 58(5):1–31, 2011.

[5] 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, 2007.

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

[7] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680, 2008.

[8] Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general communication topologies. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1995–2003, 2013.

[9] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[10] Nikhil Bansal, Vincent Cohen-Addad, Milind Prabhu, David Saulpic, and Chris Schwiegelshohn. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1707–1723. IEEE, 2024.

[11] Gustavo E. A. P. A. Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.*, 17(5-6):519–533, 2003.

[12] Ariane Baye and Christian Monseur. Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education*, 4:1–16, 2016.

[13] Shai Ben-David and Nika Haghtalab. Clustering in the presence of background noise. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 280–288, 2014.

[14] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[15] Sergei Bernstein. The theory of probabilities. *Gostechizdat, Moscow*, page 348, 1946.

[16] Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

[17] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *J. Mach. Learn. Res.*, 20:94:1–94:34, 2019.

[18] Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *SODA*, volume 1, pages 642–651. Citeseer, 2001.

[19] Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 826–835, 2008.

[20] Adam Coates and Andrew Y. Ng. Learning feature representations with k-means. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 561–580. Springer, 2012.

[21] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 1038–1051. ACM, 2022.

[22] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for Euclidean k-means. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

[23] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in euclidean spaces. In *NeurIPS*, pages 21085–21098, 2021.

[24] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 169–182. ACM, 2021.

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

[26] Juan Antonio Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

[27] Matan Danos. *Coresets for clustering by uniform sampling and generalized rank aggregation*. PhD thesis, Master's thesis, Weizmann Institute of Science, 2021, 2021.

[28] Rajesh N Dave. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664, 1991.

[29] Rajesh N Dave. Robust fuzzy clustering algorithms. In *[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems*, pages 1281–1286. IEEE, 1993.

[30] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[31] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In David B. Shmoys, editor, *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 11–20. ACM, 2014.

[32] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit and differential variance. *Artif. Intell.*, 302:103609, 2022.

[33] Xiequan Fan, Ion Grama, and Quansheng Liu. Sharp large deviation results for sums of independent random variables. *Science China Mathematics*, 58:1939–1958, 2012.

[34] Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. *Microeconomics: Asymmetric & Private Information eJournal*, 2010.

[35] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2142–2150, 2011.

[36] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.

[37] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

[38] Dan Feldman and Leonard J. Schulman. Data reduction for weighted and outlier-resistant clustering. In Yuval Rabani, editor, *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 1343–1354. SIAM, 2012.

[39] Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K. Vishnoi. Fairness for AUC via feature augmentation. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, page 610. ACM, 2022.

[40] Alex Alves Freitas. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, 16:177–199, 2001.

[41] Zachary Friggstad, Kamyar Khodamoradi, Mohsen Rezapour, and Mohammad R Salavatipour. Approximation schemes for clustering with outliers. *ACM Transactions on Algorithms (TALG)*, 15(2):1–26, 2019.

[42] Luis A. García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustin Mayo-Iscar. A general trimming approach to robust cluster Analysis. *The Annals of Statistics*, 36(3):1324 – 1345, 2008.

[43] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[44] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. Fast data anonymization with low information loss. In Christoph Koch, Johannes Gehrke, Minos N. Garofalakis, Divesh Srivastava, Karl Aberer, Anand Deshpande, Daniela Florescu, Chee Yong Chan, Venkatesh Ganti, Carl-Christian Kanne, Wolfgang Klas, and Erich J. Neuhold, editors, *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007*, pages 758–769. ACM, 2007.

[45] Shalmoli Gupta, Ravi Kumar, Kefu Lu, Benjamin Moseley, and Sergei Vassilvitskii. Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment*, 10(7):757–768, 2017.

[46] Alon Y. Halevy, Anand Rajaraman, and Joann J. Ordille. Data integration: The teenage years. In Umeshwar Dayal, Kyu-Young Whang, David B. Lomet, Gustavo Alonso, Guy M. Lohman, Martin L. Kersten, Sang Kyun Cha, and Young-Kuk Kim, editors, *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 9–16. ACM, 2006.

[47] Frank R Hampel. A general qualitative definition of robustness. *The annals of mathematical statistics*, 42(6):1887–1896, 1971.

[48] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *36th Annual ACM Symposium on Theory of Computing,*, pages 291–300, 2004.

[49] Majed El Helou and Sabine Süsstrunk. Blind universal bayesian image denoising with gaussian noise level learning. *IEEE Transactions on Image Processing*, 29:4885–4897, 2019.

[50] Christian Hennig. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99:1154–1176, 2008.

[51] Monika Henzinger and Sagar Kale. Fully-dynamic coresets. In Fabrizio Grandoni, Grzegorz Herman, and Peter Sanders, editors, *28th Annual European Symposium on Algorithms, ESA 2020, September 7-9, 2020, Pisa, Italy (Virtual Conference)*, volume 173 of *LIPIcs*, pages 57:1–57:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[52] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. ε-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 814–825. IEEE Computer Society, 2018.

[53] Lingxiao Huang, Shaofeng H.-C. Jiang, and Jianing Lou. The power of uniform sampling for k-median. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 13933–13956. PMLR, 2023.

[54] Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou, and Xuan Wu. Near-optimal coresets for robust clustering. In *ICLR*. OpenReview.net, 2023.

[55] Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for Euclidean (k, z)-clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1594–1604, 2024.

[56] Lingxiao Huang, Zhize Li, Jialin Sun, and Haoyu Zhao. Coresets for vertical federated learning: Regularized linear regression and K-means clustering. In *Advances in Neural Information Processing Systems*, pages 29566–29581, 2022.

[57] Lingxiao Huang, K. Sudhir, and Nisheeth K. Vishnoi. Coresets for time series clustering. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22849–22862, 2021.

[58] Natthakan Iam-on. Clustering data with the presence of attribute noise: a study of noise completely at random and ensemble of multiple k-means clusterings. *Int. J. Mach. Learn. Cybern.*, 11(3):491–509, 2020.

[59] Ragesh Jaiswal and Nitin Garg. Analysis of k-means++ for separable data. In *International Workshop and International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 2012.

[60] Xiaodan Jin and Keigo Hirakawa. Approximations to camera sensor noise. In *Image Processing: Algorithms and Systems*, volume 8655 of *SPIE Proceedings*, page 86550H. SPIE, 2013.

[61] Ronny Kohavi and Barry Becker. UCI machine learning repository, 1996.

[62] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

[63] Shrinu Kushagra, Samira Samadi, and Shai Ben-David. Finding meaningful cluster structure amidst background noise. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 339–354. Springer, 2016.

[64] Shrinu Kushagra, Yaoliang Yu, and Shai Ben-David. Provably noise-robust, regularised k-means clustering. *arXiv preprint arXiv:1711.11247*, 2017.

[65] Pat Langley, Wayne Iba, and Kevin Thompson. An analysis of bayesian classifiers. In William R. Swartout, editor, *Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, USA, July 12-16, 1992*, pages 223–228. AAAI Press / The MIT Press, 1992.

[66] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9459–9469, 2019.

[67] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28(2):129–136, 1982.

[68] Chris Meek, Bo Thiesson, and David Heckerman. US Census Data (1990). UCI Machine Learning Repository, 2001. DOI: https://doi.org/10.24432/C5VP42.

[69] Steven J. Miller. *Chapter 16. The Chi-square Distribution*, pages 427–446. Princeton University Press, Princeton, 2017.

[70] Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *KI-Künstliche Intelligenz*, 32:37–53, 2018.

[71] Rafail M. Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 165–176, 2006.

[72] Rose E. O'Dea, Malgorzata Lagisz, Michael D. Jennions, and Shinichi Nakagawa. Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature Communications*, 9, 2018.

[73] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Differentially private location privacy in practice. *CoRR*, abs/1410.7744, 2014.

[74] Violeta Roizman, Matthieu Jonckheere, and Frédéric Pascal. A flexible EM-like clustering algorithm for noisy data. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2019.

[75] José A. Sáez, Mikel Galar, Julián Luengo, and Francisco Herrera. Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness. *Inf. Sci.*, 247:1–20, 2013.

[76] Benjamin Schelling and Claudia Plant. KMN - removing noise from k-means clustering results. In Carlos Ordonez and Ladjel Bellatreche, editors, *Big Data Analytics and Knowledge Discovery - 20th International Conference, DaWaK 2018, Regensburg, Germany, September 3-6, 2018, Proceedings*, volume 11031 of *Lecture Notes in Computer Science*, pages 137–151. Springer, 2018.

[77] Marco Secondini. Chapter 20 - information capacity of optical channels. In Alan E. Willner, editor, *Optical Fiber Telecommunications VII*, pages 867–920. Academic Press, 2020.

[78] Adiel Statman, Liat Rozenberg, and Dan Feldman. k-means+++: Outliers-resistant clustering. *Algorithms*, 13(12):311, 2020.

[79] Rishabh Subramanian. Differential privacy techniques for healthcare data. In *IDSTA*, pages 95–100. IEEE, 2022.

[80] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.

[81] Sergios Theodoridis. Parameter learning: a convex analytic path. *Machine Learning*, 2020.

[82] Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 2020.

[83] Zixiu Wang, Yiwen Guo, and Hu Ding. Robust and fully-dynamic coreset for continuous-and-bounded learning (with outliers) problems. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14319–14331, 2021.

[84] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168, 2019.

[85] Zhiwen Yu and Hau-San Wong. Class discovery from gene expression data based on perturbation and cluster ensemble. *IEEE Transactions on NanoBioscience*, 8:147–160, 2009.

[86] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev.*, 22(3):177–210, 2004.

[87] Manfred Zimmermann and Klaus M. Dostert. Analysis and modeling of impulsive noise in broad-band powerline communications. *IEEE Transactions on Electromagnetic Compatibility*, 44:249–258, 2002.

Contents

A Related work

Coresets for clustering (with noise). There is a substantial body of work on coreset construction for (k, z)-CLUSTERING across various metric spaces, including Euclidean spaces, doubling metrics, graph shortest-path metrics, and general discrete metrics [48, 36, 37, 16, 52, 24, 21, 55]. An alternative notion called *weak coresets* has also been studied [70, 24, 27], which only require preservation of $O(1)$-approximate solutions. However, weak coresets offer no significant improvement in size compared to standard coresets in practice.

In the context of noise, most prior work focuses on identifying and removing outliers [38, 52, 54]. These approaches assume that noise manifests as identifiable outliers, and the goal is to build a robust coreset by filtering them out. By contrast, our model assumes full observation of only noisy data, with no oracle access to the clean underlying distribution. This represents a conceptual shift: instead of excluding noise, our algorithms construct coresets that directly accommodate it. This setting reflects realistic scenarios in which noise and signal cannot be reliably disentangled—necessitating coreset constructions that remain robust without preprocessing or filtering steps.

Clustering under noise. Clustering with noisy data has been studied extensively, typically under two paradigms. The first assumes noise is generated stochastically from a known distribution [29, 42], while the second considers adversarial noise with bounded magnitude or cardinality [7, 13, 9, 63, 64]. Our work aligns more closely with the first setting.

Broadly, this literature branches into two directions. The first investigates the *robustness* of existing clustering algorithms to noise—quantifying their performance degradation in noisy environments [28, 47, 50, 1, 4]. The second direction designs new algorithms that can tolerate or adapt to noise [26, 29, 13, 64]. A related line of work in robust clustering allows the algorithm to discard a fraction of points as outliers [18, 19, 45, 76, 41, 74, 78]. While our work shares a similar structural noise model with some previous works, the goal is significantly different. Prior works mostly ask whether a standard algorithm can efficiently solve the problem on the noisy dataset $\widehat{P}$. For instance, [4] uses Gaussian perturbations for k-means and shows that k-means converges quickly on the noisy data. In contrast, we aim to construct a coreset from $\widehat{P}$ that approximates the clustering cost on the original, unobserved dataset P. Here, noise is the main challenge, not a tool for tractability.

B Additional discussion on assumptions and error metrics

This section provides additional context and justification for the modeling choices made in the main text. We first clarify the Bernstein condition and show it is satisfied by a wide class of noise distributions. We then discuss practical scenarios where the noise variance θ is known or can be estimated. Next, we compare the standard error metric Err with the proposed Err_α, using concrete examples to highlight the tighter guarantees enabled by our approach. Finally, we present a weak mergeability property for Err_α measure, which may be useful for streaming settings.

B.1 Distributions satisfying the Bernstein condition

The Bernstein condition plays a central role in our theoretical analysis, as it enables control over higher-order moments and supports concentration arguments under noise. We show that this condition is satisfied by several widely used distribution families.

Sub-Gaussian and sub-exponential distributions. A distribution D is called *sub-Gaussian* if there exists a constant $K > 0$ such that for all $t > 0$,

$$\Pr_{X\sim D}[|X| \geq t] \leq 2e^{-t^2/K^2}.$$

Similarly, D is called *sub-exponential* if there exists $K > 0$ such that

$$\Pr_{X\sim D}[|X| \geq t] \leq 2e^{-t/K}.$$

Gaussian distributions are sub-Gaussian; Laplace distributions are sub-exponential.

Lemma B.1 (Moment bounds [82, Sections 2.5 and 2.7]). *If D is sub-Gaussian, there exists a constant $K > 0$ such that for all integers $i \geq 1$,*

$$\mathbb{E}_{X\sim D}[|X|^i]^{1/i} \leq K\sqrt{i}.$$

If D is sub-exponential, then for some $K > 0$,

$$\mathbb{E}_{X\sim D}[|X|^i]^{1/i} \leq Ki.$$

Combining Lemma B.1 with Stirling's approximation, we conclude that both sub-Gaussian and sub-exponential distributions satisfy the Bernstein condition. On the other hand, heavy-tailed distributions such as the Pareto, log-normal, and Student's t-distributions generally do not.

B.2 Practical scenarios with known or estimable θ

Several real-world applications provide natural access to the noise parameter θ, either directly or through side-information.

Sensor networks. In settings such as environmental monitoring using sensor arrays (e.g., for air quality), each measurement $\widehat{p}_{i,j} = p_{i,j} + \xi_j$ is perturbed by Gaussian noise $\xi_j \sim \mathcal{N}(0, \theta)$ [60]. Although the exact θ may be unknown, sensor specifications often report an upper bound θ', which can be directly used in our algorithm without additional estimation.

Standardized assessments. In educational assessments (e.g., STEM exam scores), observed outcomes are noisy proxies for latent ability [32]. Here, $\widehat{p}_{i,j} = p_{i,j} + \xi_j$, with ξ_j representing unknown variation. Historical data or variance analyses over large student cohorts often enable institutions to estimate θ empirically, which can then be reused to construct robust coresets for future datasets.

B.3 Comparing Err and Err_α

We now illustrate how Err_α yields significantly tighter guarantees than Err, even in simple 1D settings.

Noisy setting. Consider 1-MEANS in $\mathbb{R}$, with $k = d = 1$. Let P consist of $n/2$ points at -1 and $n/2$ points at $+1$. Then $\mathsf{C}(P) = 0$ and $\mathsf{OPT} = n$. Both -1 and $+1$ are 2-approximate centers: $\mathrm{cost}(P, -1) = \mathrm{cost}(P, 1) = 2n$.

Let $\widehat{P}$ be the noisy version of P under model I, with $\theta = 1$ and $\xi_p \sim \mathcal{N}(0, 1)$. Using the decomposition:

$$\mathrm{cost}(\widehat{P}, c) - \mathrm{cost}(P, c) = \sum_{p\in P} \|\xi_p\|_2^2 + 2\sum_{p\in P} \langle \xi_p, p - c\rangle,$$

the first term concentrates around n. Bounding the second term naively:

$$2\sum_{p\in P} \langle \xi_p, p - c\rangle \leq \sum_{p\in P} \left(\|\xi_p\|_2^2 + \|p - c\|_2^2\right),$$

we get $\mathsf{Err}(\widehat{P}, P) \approx \frac{2n}{3n} = \frac{2}{3}$, yielding $r_P(\widehat{P}, 1) \leq (1 + \frac{2}{3})^2 = \frac{25}{9}$.

Now consider $\mathsf{Err}_\alpha(\widehat{P}, P)$. We have:

$$\mathsf{C}(\widehat{P}) = \frac{1}{n}\sum_{p\in P} \xi_p, \quad |\mathsf{C}(\widehat{P}) - \mathsf{C}(P)| = O(n^{-1/2}), \quad \Rightarrow \mathrm{cost}(P, \mathsf{C}(\widehat{P})) = n + O(1).$$

Thus, $\mathsf{Err}_1(\widehat{P}, P) \lesssim \frac{1}{n}$, giving $r_P(\widehat{P}, 1) \lesssim 1 + \frac{1}{n}$—a much sharper guarantee.

Noise-free setting. Consider the 1-Median problem in $\mathbb{R}$, with $n - 1$ points at 0 and one point at 1. Then $\mathsf{C}(P) = 0$ and $\mathsf{OPT} = 1$. Let S have $n - 1$ points at 0 and one at $1/n$. Then:

$$\mathsf{C}(S) = 0, \quad \mathsf{OPT}_S = 1/n, \quad r_P(\mathsf{C}(S)) = 1, \quad r_S(\mathsf{C}(S)) = 1 \Rightarrow \mathsf{Err}_1(S, P) = 0.$$

However,

$$\mathrm{Err}(S, P) = \frac{|\mathrm{cost}(P, \mathsf{C}(S)) - \mathrm{cost}(S, \mathsf{C}(S))|}{\mathrm{cost}(S, \mathsf{C}(S))} = n.$$

This example underscores that even in noise-free settings, Err_α can yield dramatically smaller errors than Err, especially when datasets differ slightly but preserve the same optimal solution.

B.4 Weak mergeability under Err_α

We discuss the mergeability under our proposed measure Err_α. In the ideal case, mergeability means that if two coresets S_1, S_2 for disjoint datasets P_1, P_2 each satisfy $\mathsf{Err}_\alpha(S_\ell, P_\ell) \le \varepsilon$, then their union also satisfies

$$\mathsf{Err}_\alpha(S_1 \cup S_2,\ P_1 \cup P_2) \le \varepsilon.$$

However, this guarantee can fail with our new metric Err_α. Consider the cases that S_1 and S_2 summarize datasets with significantly different cluster structures. In such cases, the optimal center set for $S_1 \cup S_2$ may differ substantially from the union of the individual optimal, and the combined coreset may not encode enough information to capture this emergent structure.

Mergeability remains plausible when the two coresets are structurally similar. If there exists $1 \le \alpha' \le \alpha$ such that

$$\mathcal{C}_{\alpha'}(S_1) \subseteq \mathcal{C}_\alpha(S_2) \quad \text{and} \quad \mathcal{C}_{\alpha'}(S_2) \subseteq \mathcal{C}_\alpha(S_1),$$

then S_1 and S_2 approximately share the same set of near-optimal solutions. This overlap enables a weaker form of mergeability under Err_α.

Claim B.2 (**Weak mergeability under** Err_α). Let $\varepsilon > 0$ and $\alpha, \alpha' \ge 1$ with $\alpha' \le \alpha$. Given datasets P_1 and P_2 and weighted subsets S_1 and S_2, suppose $\mathsf{Err}_\alpha(S_\ell, P_\ell) \le \varepsilon$ and $\mathcal{C}_{\alpha'}(S_\ell) \subseteq \mathcal{C}_\alpha(S_{3-\ell})$ for $\ell = 1, 2$. Define:

$$\kappa = \frac{\min\{\mathsf{OPT}_{S_1}, \mathsf{OPT}_{S_2}\}}{\mathsf{OPT}_{S_1 \cup S_2}}, \quad \tau = \max\left\{ \frac{\mathsf{OPT}_{S_1}/\mathsf{OPT}_{P_1}}{\mathsf{OPT}_{S_2}/\mathsf{OPT}_{P_2}}, \frac{\mathsf{OPT}_{S_2}/\mathsf{OPT}_{P_2}}{\mathsf{OPT}_{S_1}/\mathsf{OPT}_{P_1}} \right\}.$$

Then:

$$\mathsf{Err}_{1+(\alpha-1)\kappa}(S_1 \cup S_2,\ P_1 \cup P_2) < \alpha' \cdot \tau \cdot (1+\varepsilon) - 1.$$

In the limiting case where $\alpha', \alpha \to 1$ and $\tau \to 1$, this bound recovers the ideal mergeability condition:

$$\mathsf{Err}_\alpha(S_1 \cup S_2,\ P_1 \cup P_2) \le \varepsilon.$$

Proof of Claim B.2. We first prove that for any center set $C \in \mathcal{C}_{1+(\alpha-1)\kappa}(S_1 \cup S_2)$, $C \in \mathcal{C}_\alpha(S_1) \cap \mathcal{C}_\alpha(S_2)$. By contradiction if there exists some $C \in \mathcal{C}_{1+(\alpha-1)\kappa}(S_1 \cup S_2)$ such that $C \notin \mathcal{C}_\alpha(S_1) \cap \mathcal{C}_\alpha(S_2)$. Then

$$\begin{aligned}
& \mathrm{cost}(S_1 \cup S_2, C) \\
> \ & \mathsf{OPT}_{S_1 \cup S_2} + (\alpha-1) \cdot \min\{\mathsf{OPT}_{S_1}, \mathsf{OPT}_{S_2}\} && (C \notin \mathcal{C}_\alpha(S_1) \cap \mathcal{C}_\alpha(S_2)) \\
\ge \ & \mathsf{OPT}_{S_1 \cup S_2} + (\alpha-1)\kappa \cdot \mathsf{OPT}_{S_1 \cup S_2} && (\text{Defn. of } \kappa) \\
= \ & (1 + (\alpha-1)\kappa) \cdot \mathsf{OPT}_{S_1 \cup S_2}.
\end{aligned}$$

Thus, $C \notin \mathcal{C}_{1+(\alpha-1)\kappa}(S_1 \cup S_2)$, which is a contradiction.

Fix a center set $C \in \mathcal{C}_{1+(\alpha-1)\kappa}(S_1 \cup S_2)$. We now have $C \in \mathcal{C}_\alpha(S_1) \cap \mathcal{C}_\alpha(S_2)$. Since $\mathsf{Err}_\alpha(S_\ell, P_\ell) \le \varepsilon$, we have

$$\frac{r_{P_\ell}(C) - r_{S_\ell}(C)}{r_{S_\ell}(C)} \le \varepsilon,$$

implying that $r_{P_\ell}(C) \le (1+\varepsilon) r_{S_\ell}(C)$. Moreover, let $\mathsf{C}(P)$ denote the optimal center set of S_1 and we have $\mathsf{C}(P) \in \mathcal{C}_\alpha(S_1) \subseteq \mathcal{C}_{\alpha'}(S_2)$. Thus, we have

$$\mathsf{OPT}_{S_1 \cup S_2} \le \mathrm{cost}(S_1 \cup S_2, \mathsf{C}(P)) \le \mathsf{OPT}_{S_1} + \alpha' \cdot \mathsf{OPT}_{S_2} \le \alpha'(\mathsf{OPT}_{S_1} + \mathsf{OPT}_{S_2}). \quad (8)$$

Thus,

$$
\begin{aligned}
& \frac{r_{P_1\cup P_2}(C) - r_{S_1\cup S_2}(C)}{r_{S_1\cup S_2}(C)} \\
= \;& \frac{r_{P_1\cup P_2}(C)}{r_{S_1\cup S_2}(C)} - 1 \\
= \;& \frac{\mathrm{cost}(P_1, C) + \mathrm{cost}(P_2, C)}{\mathrm{cost}(S_1, C) + \mathrm{cost}(S_2, C)} \cdot \frac{\mathsf{OPT}_{S_1\cup S_2}}{\mathsf{OPT}_{P_1\cup P_2}} - 1 \\
= \;& \frac{\mathsf{OPT}_{P_1} \cdot r_{P_1}(C) + \mathsf{OPT}_{P_2} \cdot r_{P_2}(C)}{\mathsf{OPT}_{S_1} \cdot r_{S_1}(C) + \mathsf{OPT}_{S_2} \cdot r_{S_2}(C)} \cdot \frac{\mathsf{OPT}_{S_1\cup S_2}}{\mathsf{OPT}_{P_1\cup P_2}} - 1 \\
\leq \;& (1+\varepsilon)\frac{\mathsf{OPT}_{P_1} \cdot r_{S_1}(C) + \mathsf{OPT}_{P_2} \cdot r_{S_2}(C)}{\mathsf{OPT}_{S_1} \cdot r_{S_1}(C) + \mathsf{OPT}_{S_2} \cdot r_{S_2}(C)} \cdot \frac{\mathsf{OPT}_{S_1\cup S_2}}{\mathsf{OPT}_{P_1\cup P_2}} - 1 \\
& (r_{P_\ell}(C) \leq (1+\varepsilon) r_{S_\ell}(C)) \\
\leq \;& (1+\varepsilon)\max\left\{\frac{\mathsf{OPT}_{P_1}}{\mathsf{OPT}_{S_1}}, \frac{\mathsf{OPT}_{P_2}}{\mathsf{OPT}_{S_2}}\right\} \cdot \frac{\mathsf{OPT}_{S_1\cup S_2}}{\mathsf{OPT}_{P_1\cup P_2}} - 1 \\
\leq \;& (1+\varepsilon)\max\left\{\frac{\mathsf{OPT}_{P_1}}{\mathsf{OPT}_{S_1}}, \frac{\mathsf{OPT}_{P_2}}{\mathsf{OPT}_{S_2}}\right\} \cdot \frac{\alpha'(\mathsf{OPT}_{S_1} + \mathsf{OPT}_{S_2})}{\mathsf{OPT}_{P_1} + \mathsf{OPT}_{P_2}} - 1 && \text{(Ineq. (8))} \\
\leq \;& \alpha'(1+\varepsilon) \cdot \max\left\{\frac{\mathsf{OPT}_{P_1}}{\mathsf{OPT}_{S_1}}, \frac{\mathsf{OPT}_{P_2}}{\mathsf{OPT}_{S_2}}\right\} \cdot \max\left\{\frac{\mathsf{OPT}_{S_1}}{\mathsf{OPT}_{P_1}}, \frac{\mathsf{OPT}_{S_2}}{\mathsf{OPT}_{P_2}}\right\} - 1 \\
\leq \;& \alpha'\tau(1+\varepsilon) - 1. && \text{(Defn. of } \tau\text{)}
\end{aligned}
$$

□

C Proof of Theorem 3.1: Using Err

In this section, we prove Theorem 3.1. Our proof primarily relies on bounding $\mathsf{Err}(\widehat{P}, P)$, which we will show in Section C.1. Additionally, we provide a lower bound for $\mathsf{Err}(\widehat{P}, P)$ in Section C.2. This helps explain each additive term in Theorem 3.1.

Theorem C.1 (Restatement of Theorem 3.1). *Let $\widehat{P}$ be drawn from P via noise model I with known $\theta \geq 0$. Let $\varepsilon \in (0,1)$ and fix $\alpha \geq 1$. Let $\mathcal{A}$ be an algorithm that constructs a weighted subset $S \subset \widehat{P}$ for k-MEANS of size $\mathcal{A}(\varepsilon)$ and with guarantee $\mathsf{Err}(S, \widehat{P}) \leq \varepsilon$. Then with probability $p > 0.9$,*

$$\mathsf{Err}(S, P) \leq \varepsilon + O(\tfrac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\tfrac{\theta n d}{\mathsf{OPT}_P}}) \text{ and } r_P(S, \alpha) \leq (1 + \varepsilon + O(\tfrac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\tfrac{\theta n d}{\mathsf{OPT}_P}}))^2 \cdot \alpha.$$

For simplicity, we use OPT to denote OPT_P in the following discussion. For preparation, we provide the following lemmas.

Lemma C.2 (Bounding $\mathsf{Err}(\widehat{P}, P)$). *For any $\widehat{P}$ derived from an n-point dataset $P \in \mathbb{R}^d$ using noise model I, with parameter θ we have that with probability $p > 0.9$,*

$$\mathsf{Err}(\widehat{P}, P) \leq O(\tfrac{\theta n d}{\mathsf{OPT}} + \sqrt{\tfrac{\theta n d}{\mathsf{OPT}}}). \tag{9}$$

Lemma C.3 (Composition property). *Given $P, \widehat{P}, S \subset \mathbb{R}^d$, suppose $\mathsf{Err}(S, \widehat{P}) \in (0,1)$, then we have*

$$\mathsf{Err}(S, P) \leq \mathsf{Err}(S, \widehat{P}) + 2\mathsf{Err}(\widehat{P}, P)$$

Proof. Suppose $\mathsf{Err}(S, \widehat{P}) = \varepsilon$ and $\mathsf{Err}(\widehat{P}, P) = \varepsilon'$. For every center set C, we have

$$|\mathrm{cost}(\widehat{P}, C) - \mathrm{cost}(S, C)| \leq \varepsilon \cdot \mathrm{cost}(S, C),$$

and

$$|\mathrm{cost}(P, C) - \mathrm{cost}(\widehat{P}, C)| \leq \varepsilon' \cdot \mathrm{cost}(\widehat{P}, C).$$

Combine these inequalities above, we have

$$\begin{aligned}
&|\mathrm{cost}(P,C) - \mathrm{cost}(S,C)| \\
\le\ &|\mathrm{cost}(\widehat{P},C) - \mathrm{cost}(S,C)| + |\mathrm{cost}(P,C) - \mathrm{cost}(\widehat{P},C)| && \text{(Triangle Inequality)} \\
\le\ &\varepsilon \cdot \mathrm{cost}(S,C) + \varepsilon' \cdot \mathrm{cost}(\widehat{P},C) \\
\le\ &(\varepsilon + \varepsilon'(1+\varepsilon)) \cdot \mathrm{cost}(S,C) \\
\le\ &(\varepsilon + 2\varepsilon') \cdot \mathrm{cost}(S,C) && (\varepsilon < 1)
\end{aligned}$$

Thus $\mathsf{Err}(S,P) = \sup_{C\in\mathcal{C}} \frac{|\mathrm{cost}_z(S,C) - \mathrm{cost}_z(P,C)|}{\mathrm{cost}_z(S,C)} \le \mathsf{Err}(S,\widehat{P}) + 2\mathsf{Err}(\widehat{P},P)$. □

We are ready to prove Theorem 3.1.

Proof of Theorem 3.1. Suppose a weighted subset $S \subset \widehat{P}$ constructed by Algorithm $\mathcal{A}$ satisfies that $\mathsf{Err}(S,\widehat{P}) \le \varepsilon$. By Lemma C.2, with probability $p > 0.9$,

$$\mathsf{Err}(\widehat{P},P) \le O(\frac{\theta nd}{\mathsf{OPT}} + \sqrt{\frac{\theta nd}{\mathsf{OPT}}}).$$

By Lemma C.3,

$$\mathsf{Err}(S,P) \le \mathsf{Err}(S,\widehat{P}) + 2\mathsf{Err}(\widehat{P},P) = \varepsilon + O(\frac{\theta nd}{\mathsf{OPT}} + \sqrt{\frac{\theta nd}{\mathsf{OPT}}}).$$

Moreover, for the optimal center set $\mathsf{C}(S)$ of S, we have

$$\frac{|\mathrm{cost}(S,\mathsf{C}(S)) - \mathrm{cost}(P,\mathsf{C}(S))|}{\mathrm{cost}(S,\mathsf{C}(S))} \le \mathsf{Err}(S,P),$$

which implies

$$\mathrm{cost}_z(P,\mathsf{C}(S)) \in \left[\frac{1}{1+\mathsf{Err}(S,P)}, (1+\mathsf{Err}(S,P))\right] \cdot \mathrm{cost}_z(S,\mathsf{C}(S)).$$

For an α-approximate center set C of S, we have

$$\mathrm{cost}(P,C) \le (1+\mathsf{Err}(S,P)) \cdot \mathrm{cost}(S,C) \le (1+\mathsf{Err}(S,P)) \cdot \alpha \cdot \mathrm{cost}(S,\mathsf{C}(S)).$$

Combining these gives:

$$r_P(S,\alpha) \le (1+\mathsf{Err}(S,P))^2 \cdot \alpha. \tag{10}$$

This implies that $r_P(S,\alpha) \le (1+\varepsilon+O(\frac{\theta nd}{\mathsf{OPT}} + \sqrt{\frac{\theta nd}{\mathsf{OPT}}}))^2 \cdot \alpha$. □

C.1 Proof of Lemma C.2: Bounding $\mathsf{Err}(\widehat{P},P)$

For each $p \in P$, recall that $\xi_p = \widehat{p} - p$ is the noise vector. We first have the following claim that bounds norms of these noise vectors.

Claim C.4 (**Bounding $\sum_{p\in P} \|\xi_p\|_2^2$**). With probability at least 0.95, $\sum_{p\in P} \|\xi_p\|_2^2 \le 60\theta nd$.

Proof. Note that for every $p \in P$,

$$\begin{aligned}
\mathbb{E}_{\xi_p}\left[\|\xi_p\|_2^2\right] &= \theta \mathbb{E}_{\xi_{p,j}\sim D_j\sim}\left[\sum_j \xi_{p,j}^2\right] \\
&= \theta \sum_j \mathrm{Var}_{\xi_{p,j}\sim D_j\sim}[\xi_{p,j}] - \mathbb{E}_{\xi_{p,j}\sim D_j\sim}[\xi_{p,j}]^2 \\
&= \theta d. && \text{(Defn. of } D_j\text{)}
\end{aligned}$$

Thus, we have

$$\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right]=\theta nd. \tag{11}$$

Furthermore,

$$\begin{aligned}
&\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right]\\
=\;& n\cdot\mathrm{Var}_{\xi_p}\left[\|\xi_p\|_2^2\right]\\
=\;& n\cdot\left(\mathbb{E}_{\xi_p}\left[\|\xi_p\|_2^4\right]-\mathbb{E}_{\xi_p}\left[\|\xi_p\|_2^2\right]^2\right)\\
\le\;& \theta n\cdot(2d+d^2-\theta d^2)\\
\le\;& 3\theta nd^2.
\end{aligned} \tag{12}$$

where $\chi^2(d)$ represents the chi-square distribution with d degrees of freedom, whose variance is known to be $2d$ [69].

If $\theta\le\frac{1}{20n}$, we have that

$$\Pr_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2=0\right]=(1-\theta)^n\ge(1-\frac{1}{20n})^n\ge 0.95,$$

implying that $\Pr_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\le 60\theta nd\right]\ge 0.95$. Otherwise, if $\theta\le\frac{1}{n}$, we have that

$$\begin{aligned}
&\Pr_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2>60\theta nd\right]\\
\le\;& \Pr_{\widehat{P}}\left[\left|\sum_{p\in P}\|\xi_p\|_2^2-\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right]\right|>33\sqrt{\theta n}\sqrt{\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right]}\right] && \text{(Eq. (11) and Ineq. (12))}\\
\le\;& \frac{1}{1000\theta n} && \text{(Chebyshev's ineq.)}\\
\le\;& 0.05. && (\theta\ge\frac{1}{20n})
\end{aligned}$$

Thus, we complete the proof of the claim. □

Now we are ready to prove Lemma C.2.

Proof of Lemma C.2. By Claim C.4, with probability at least 0.95, $\sum_{p\in P}\|\xi_p\|_2^2\le 60\theta nd$, which we assume happens in the following. It suffices to prove that for any center set $C\in\mathcal{C}$,

$$|\mathrm{cost}(P,C)-\mathrm{cost}(\widehat{P},C)|\le(\frac{60\theta nd}{\mathsf{OPT}}+4\sqrt{\frac{15\theta nd}{\mathsf{OPT}}})\cdot\mathrm{cost}(\widehat{P},C). \tag{13}$$

By the triangle inequality, we know that for each $p\in P$, $|d(p,C)-d(\widehat{p},C)|\le\|\xi_p\|_2$, which implies that $|d^2(p,C)-d^2(\widehat{p},C)|\le\|\xi_p\|_2^2+2\|\xi_p\|_2\cdot d(p,C)$.

By a similar analysis of Lemma D.8, we have $\mathrm{cost}(\widehat{P}, C) = O(\mathrm{cost}(P, C))$ since we assume $\mathsf{OPT} > \theta nd$, thus we have

$$\begin{aligned}
& \frac{|\mathrm{cost}(P, C) - \mathrm{cost}(\widehat{P}, C)|}{\mathrm{cost}(\widehat{P}, C)} \\
\leq\ & \frac{\sum_{p \in P} \|\xi_p\|_2^2 + 2\|\xi_p\|_2 \cdot d(p, C)}{\mathrm{cost}(\widehat{P}, C)} \\
\leq\ & \frac{60\theta nd}{\mathsf{OPT}} + \frac{2\sum_{p \in P} \|\xi_p\|_2 \cdot d(p, C)}{\mathrm{cost}(\widehat{P}, C)} && \text{(by assumption)} \\
\leq\ & \frac{60\theta nd}{\mathsf{OPT}} + \frac{2\sqrt{(\sum_{p \in P} \|\xi_p\|_2^2) \cdot (\sum_{p \in P} d^2(p, C))}}{\mathrm{cost}(\widehat{P}, C)} && \text{(Cauchy-Schwarz)} \\
\leq\ & \frac{60\theta nd}{\mathsf{OPT}} + 4\sqrt{\frac{15\theta nd}{\mathrm{cost}(P, C)}} && \text{(by assumption)} \\
\leq\ & \frac{60\theta nd}{\mathsf{OPT}} + 4\sqrt{\frac{15\theta nd}{\mathsf{OPT}}}, && \text{(Defn. of } \mathsf{OPT}\text{)}
\end{aligned}$$

which completes the proof of Inequality (13). □

C.2 Lower bound of $\mathrm{Err}(\widehat{P}, P)$

We provide the following lower bound for $\mathrm{Err}(\widehat{P}, P)$.

Claim C.5 (**Lower Bound of** $\mathsf{Err}(\widehat{P}, P)$)**.** With probability $p > 0.8$, $\mathsf{Err}(\widehat{P}, P) = \Omega(\frac{\theta nd}{\mathsf{OPT}_P})$.

Proof of Claim C.5. We show this by a worst-case example. Let $\theta = 0.1$, $n = 10000$ and $k = n - 1$. Let $P = \{p_i = 100ne_i : i \in [n]\} \subset \mathbb{R}^n$ where e_i is the i-th unit basis in $\mathbb{R}^n$. An optimal solution $\mathsf{C}(P) = \left\{p_1, \ldots, p_{n-2}, \frac{p_{n-1}+p_n}{2}\right\}$, and hence, $\mathsf{OPT} = 10000n^2$. Note that with probability at least 0.8, the following events hold: $\sum_{p \in P} \|\xi_p\|_2^2 = \Theta(\theta nd)$ and for every $p \in P$, $\|\xi_p\|_2^2 \leq 10d \log n$. Conditioned on these events, the optimal solution $\mathsf{C}(\widehat{P})$ of $\widehat{P}$ must consist of $n - 2$ points $\widehat{p} \in \widehat{P}$ and the average of the remaining two points in $\widehat{P}$. Assume $\mathsf{C}(\widehat{P}) = \left\{\widehat{p}_1, \ldots, \widehat{p}_{n-2}, \frac{\widehat{p}_{n-1}+\widehat{p}_n}{2}\right\}$. By calculation, we obtain that

$$\mathrm{cost}(P, \mathsf{C}(\widehat{P})) \geq (1 + \Omega(\frac{\theta nd}{\mathsf{OPT}})),$$

implying that $\mathsf{Err}(\widehat{P}, P) = \Omega(\frac{\theta nd}{\mathsf{OPT}_P})$. □

D Proof of Theorem 3.3: Using Err_α

Theorem D.1 (**Restatement of Theorem 3.3**)**.** *Let $\widehat{P}$ be an observed dataset drawn from P under the noise model I with known parameter $\theta \in [0, \frac{\mathsf{OPT}_P}{nd}]$. Let $\varepsilon \in (0, 1)$ and fix $\alpha \in [1, 2]$. Under Assumption 3.2, there exists a randomized algorithm that constructs a weighted $S \subset \widehat{P}$ for k-MEANS of size $O(\frac{k \log k}{\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}_P}} + \frac{(\alpha-1)k \log k}{(\varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\alpha \mathsf{OPT}_P})^2})$ and with guarantee $\mathsf{Err}_\alpha(S, \widehat{P}) \leq \varepsilon$ with probability at least 0.99. Moreover,*

$$\mathsf{Err}_\alpha(S, P) \leq \varepsilon + O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha - 1}}{\alpha} \cdot \frac{\sqrt{\theta kd\mathsf{OPT}_P} + \theta nd}{\mathsf{OPT}_P}\right),$$

and

$$r_P(S, \alpha) \leq \left(1 + \varepsilon + O\left(\frac{\theta kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha - 1}}{\alpha} \cdot \frac{\sqrt{\theta kd\mathsf{OPT}_P} + \theta nd}{\mathsf{OPT}_P}\right)\right) \cdot \alpha.$$

Recall that we use Algorithm 1 to construct such a coreset S. For ease of analysis, we provide the following assumption for the center set $\widehat{C}$ obtained in Line 1 of Algorithm 1. We will justify this assumption in Section D.4.

Assumption D.2 (Locality). Assume that for each $\widehat{c}_i \in \widehat{C}$ $(i \in [k])$, $d(\widehat{c}_i, c_i^\star) \leq O(r_i + \sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})$. Here, $c_i^\star$ represents the i-th center in $\mathsf{C}(P)$.

In the following discussion, we denote $c_i^\star$ to be the i-th center in $\mathsf{C}(P)$, and use $\mu(X) = \frac{\sum_{p\in X} p}{|X|}$ to denote the mean point of any dataset X.

For preparation, we first show that given $C \in \mathcal{C}_\alpha(S)$, every center $c_i \in C$ lies within a local ball centered at an optimal center $c_i^\star \in \mathsf{C}(P)$.

Lemma D.3 (Structural properties of $\mathcal{C}_\alpha(S)$). *With high probability, for every $C \in \mathcal{C}_\alpha(S)$ and every $i \in [k]$, there exist a center $c \in C$ such that $c \in B(c_i^\star, O(\sqrt{\frac{\alpha(\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i}}))$. Moreover, for each point $p \in P_i, i, j \in [k], i \neq j$, $d(p, c_i) < d(p, c_j)$.*

Next, we provide some useful geometric properties of S.

Lemma D.4 (Properties of S). *With probability $p > 0.99$,*

- *For any $i \in [k]$, $\|\mu(P_i') - \mu(S_i)\|_2^2 \leq O(\frac{\varepsilon\mathsf{OPT}_i'}{n_i})$.*
- $\mathsf{OPT}_S \leq (1 + O(\frac{\varepsilon}{\sqrt{\alpha-1}}))\mathsf{OPT}'$.
- $\mathsf{Err}_\alpha(S, \widehat{P}) \leq \varepsilon$.

Finally, we provide an upper bound for the movement distance of centers from P' to P, which helps bound the error caused by noise.

Lemma D.5 (Movement of centers of P'). *With high probability, for any $i \in [k]$,*

$$\|\mu(P_i) - \mu(P_i')\|_2^2 \leq O(\frac{\theta d}{n_i} + \sqrt{\alpha-1}\theta d)$$

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Suppose we run Algorithm 1 to construct a weighted $S \subset \widehat{P}$ for k-MEANS of size m. By Lemma D.4, we directly have $\mathsf{Err}_\alpha(S, \widehat{P}) \leq \varepsilon$. It remains to prove $r_P(S, \alpha) \leq (1 + O(\varepsilon + \frac{\theta kd}{\mathsf{OPT}} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta kd\mathsf{OPT}}+\theta nd}{\mathsf{OPT}})) \cdot \alpha$.

Given an α-approximate center set $C_\alpha = \{c_1, \ldots, c_k\}$ of S , by Lemma D.3, we can assume $\min_{c\in C_\alpha} d(p, c) = c_i$ for any $i \in [k]$ and any point $p \in P_i$. This implies that any point $p \in P_i$ satisfies $d(p, c_i) < d(p, c_j)$ for any $j \neq i$.

Then we have

$$
\begin{aligned}
& r_P(C_\alpha) \\
= \; & \frac{\mathrm{cost}(P, C_\alpha)}{\mathsf{OPT}} \\
= \; & \frac{\sum_{i\in k} \mathrm{cost}(P_i, c_i) - \mathrm{cost}(P_i, \mu(P_i))}{\mathsf{OPT}} + 1 \\
= \; & \frac{\sum_{i\in k} n_i \|c_i - \mu(P_i)\|_2^2}{\mathsf{OPT}} + 1 \\
\le \; & \frac{\sum_{i\in k} n_i \cdot (||\mu(P_i) - \mu(P_i')|| + ||\mu(P_i') - \mu(S_i)|| + ||\mu(S_i) - c_i||)^2}{\mathsf{OPT}} + 1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Triangle inequality)} \\
\le \; & \frac{O(\varepsilon \mathsf{OPT}' + \theta k d + \sqrt{\alpha-1}\theta n d + \sqrt{(\alpha-1)\theta k d \mathsf{OPT}}) + (\alpha-1)\mathsf{OPT}_S}{\mathsf{OPT}} + 1 \\
\le \; & \frac{O(\varepsilon \mathsf{OPT}' + \theta k d + \sqrt{\alpha-1}\theta n d + \sqrt{(\alpha-1)\theta k d \mathsf{OPT}}) + (\alpha-1)(1 + O(\frac{\varepsilon}{\sqrt{\alpha-1}}))\mathsf{OPT}}{\mathsf{OPT}} + 1 \\
\le \; & \alpha \cdot \left(1 + O(\varepsilon + \frac{\theta k d}{\mathsf{OPT}} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\theta k d \mathsf{OPT}} + \theta n d}{\mathsf{OPT}})\right).
\end{aligned}
$$

This completes the proof.

□

D.1 Proof of Lemma D.3: structural properties of $\mathcal{C}_\alpha(S)$

Useful properties of $\widehat{P}$. For preparation, we first show some properties of $\widehat{P}$. In the following, Claim D.6 and Lemma D.7 show that the clustering cost for 1-MEANS remains stable under noise perturbation. It also helps bound the cost difference in each cluster in the general k-MEANS setting.

Claim D.6 (**Statistics of cost difference**). In 1-MEANS problem, for any center $c \in \mathbb{R}^d$, we have

$$\mathrm{cost}(\widehat{P}, c) - \mathrm{cost}(P, c) = \sum_{p\in P} \|\xi_p\|_2^2 + 2\sum_{p\in P} \langle \xi_p, p - c\rangle.$$

Moreover, we have

- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] = \theta n d$ and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] \le 3\theta n d^2$;
- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P} \langle \xi_p, p - c\rangle\right] = 0$ and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P} \langle \xi_p, p - c\rangle\right] = \theta \cdot \mathrm{cost}(P, c)$.

Proof. We have $\mathrm{cost}(\widehat{P}, c) - \mathrm{cost}(P, c) = \sum_{p\in P} d^2(\widehat{p}, c) - d^2(p, c) = \sum_{p\in P} \|\xi_p\|_2^2 + 2\langle \xi_p, p - c\rangle$. For the first error term $\sum_{p\in P} \|\xi_p\|_2^2$, we have $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] = \theta n d$, and

$$
\begin{aligned}
& \mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] \\
= \; & n \cdot \mathrm{Var}_{\xi_p}\left[\|\xi_p\|_2^2\right] \\
= \; & n \cdot \left(\mathbb{E}_{\xi_p}\left[\|\xi_p\|_2^4\right] - \mathbb{E}_{\xi_p}\left[\|\xi_p\|_2^2\right]^2\right) \\
\le \; & \theta n \cdot (2d + d^2 - \theta d^2) \\
\le \; & 3\theta n d^2.
\end{aligned}
$$

where $\chi^2(d)$ represents the chi-square distribution with d degrees of freedom, whose variance is known to be $2d$ [69].

For the second error term $\sum_{p\in P}\langle \xi_p, p-c\rangle$, its expectation is obvious 0 and we have

$$\begin{aligned}\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\langle \xi_p, p-c\rangle\right] &= \sum_{p\in P}\mathrm{Var}_{\xi_p}\left[\langle \xi_p, p-c\rangle\right] \\ &= \sum_{p\in P}\mathbb{E}_{\xi_p}\left[\langle \xi_p, p-c\rangle^2\right] \\ &= \theta\cdot\sum_{p\in P}\|p-c\|_2^2 \\ &= \theta\cdot \mathrm{cost}(P,c).\end{aligned}$$

□

By Claim D.6, it suffices to prove that

$$\sup_{c\in\mathbb{R}^d}\frac{\left|\sum_{p\in P}\|\xi_p\|_2^2+2\sum_{p\in P}\langle \xi_p, p-c\rangle\right|}{\mathrm{cost}(P,c)} \le O\left(\frac{\theta n d}{\mathsf{OPT}}+\sqrt{\frac{\theta d}{\mathsf{OPT}}}\right). \tag{14}$$

By Claim D.6 and Chebyshev's inequality, we directly have that $\sum_{p\in P}\|\xi_p\|_2^2 \le 2\theta n d$ happens with probability at least 0.95. For the second error term $2\sum_{p\in P}\langle \xi_p, p-c\rangle$, we provide an upper bound by the following lemma, which strengthens the second property of Claim D.6.

Lemma D.7 (Error bound for 1-MEANS). *In* 1-MEANS *problem, with probability at least* $1-0.05\delta$ *for* $0<\delta\le 1$*, the following holds*

$$\sup_{c\in\mathbb{R}^d}\frac{|\sum_{p\in P}\langle \xi_p, p-c\rangle|}{\mathrm{cost}(P,c)} = 10\cdot\sqrt{\frac{\theta d}{\delta\mathsf{OPT}}}.$$

Proof. Let $X = |\{p\in P:\xi_p\neq 0\}|$ be a random variable. When $\theta n\le 0.01\delta$, we have

$$\Pr[X=0] = (1-\theta)^n \ge 1-0.02\delta.$$

Conditioned on the event that $X=0$, we have

$$\sup_{c\in\mathbb{R}^d}\frac{|\sum_{p\in P}\langle \xi_p, p-c\rangle|}{\mathrm{cost}(P,c)} = 0,$$

which completes the proof.

In the following, we analyze the case that $\theta n > 0.01\delta$. Let E be the event that $\|\sum_{p\in P}\xi_p\|_2 \le O(\sqrt{\theta n d/\delta})$. We have the following claim:

$$\Pr[E]\ge 1-0.02\delta. \tag{15}$$

Note that $E[X]=\theta n$ and hence, $\Pr[X\le 100\theta n/\delta]\ge 1-0.01\delta$ by Markov's inequality. Hence, we only need to prove $\Pr[E\mid X\le 100\theta n/\delta]$ for Claim 15. Also note that $\sum_{p\in P}\xi_p$ has the same distribution as $N(0, X\cdot I_d)$. Then by Theorem 3.1.1 in [82],

$$\Pr\left[\|\sum_{p\in P}\xi_p\|_2 \le 10\sqrt{\theta n d/\delta}\mid X\le 100\theta n\right]\ge 1-0.01\delta.$$

Thus, we prove (15).

In the remaining proof, we condition on event E. Fix an arbitrary center $c\in\mathbb{R}^d$ and let $l = \|c-\mathsf{C}(P)\|_2$. By the optimality of $\mathsf{C}(P)$, it is well known that

$$\mathrm{cost}(P,c) = \mathrm{cost}(P,\mathsf{C}(P)) + n\cdot\|c-\mathsf{C}(P)\|_2^2 = \mathsf{OPT}+nl^2.$$

Note that $|\sum_{p\in P}\langle \xi_p, p-c\rangle| \le |\sum_{p\in P}\langle \xi_p, p-\mathsf{C}(P)\rangle| + |\sum_{p\in P}\langle \xi_p, c-\mathsf{C}(P)\rangle|$. By Chebyshev's inequality, we have

$$\Pr_{\widehat{P}}[|\sum_{p\in P}\langle \xi_p, p-\mathsf{C}(P)\rangle| \ge 10\sqrt{\theta\cdot\mathsf{OPT}/\delta}] \le \frac{\theta\cdot\mathrm{cost}(P,\mathsf{C}(P))}{(10\sqrt{\theta\cdot\mathsf{OPT}/\delta})^2} = 0.01\delta. \tag{16}$$

We also have

$$\begin{aligned} |\sum_{p\in P}\langle \xi_p, c-\mathsf{C}(P)\rangle| \leq & \quad \|\sum_{p\in P}\xi_p\|_2 \|c-\mathsf{C}(P)\| \quad \text{(Cauchy-schwarz)} \\ \leq & \quad 10l\cdot\sqrt{\theta nd/\delta} \end{aligned} \tag{17}$$

Combining with Inequalities 16 and 17, we conclude that

$$\frac{|\sum_{p\in P}\langle \xi_p, p-c\rangle|}{\mathrm{cost}(P,c)} \leq 20\cdot\frac{l\cdot\sqrt{\theta nd/\delta}}{\mathsf{OPT}+nl^2} \leq 10\cdot\sqrt{\frac{\theta d}{\delta\mathsf{OPT}}},$$

happens with probability at least 0.95, which completes the proof of Lemma D.7. □

Recall that we use P_i to denote the data clustered to c_i in P, where c_i is the i-th center in the optimal cluster of P. We use $\widetilde{P}_i$ to denote the points in P_i with the presence of the noise, and $\widetilde{c}_i$ to denote the mean of $\widetilde{P}_i$. Now we bound $\mathsf{OPT}_{\widehat{P}}$ in the general k-MEANS case.

Lemma D.8 (Bounding $\mathsf{OPT}_{\widehat{P}}$). *With probability at least 0.9, we have for all $i\in[k]$*

$$\mathrm{cost}(\widetilde{P}_i,\widetilde{c}_i) \leq \mathsf{OPT}_i + O\left(\theta n_i dk + \sqrt{\theta dk\cdot\mathsf{OPT}_i}\right) \leq 1.5\mathsf{OPT}_i + O(\theta n_i dk).$$

Besides, we also have with high probability

$$\mathsf{OPT}_{\widehat{P}} \leq \sum_i \mathrm{cost}(\widetilde{P}_i,\widetilde{c}_i) \leq \mathsf{OPT} + O(\theta nd + \sqrt{\theta d\cdot\mathsf{OPT}})$$

Proof. Similar to the 1-MEANS setting (Claim D.6), we have the following decomposition of the error

$$\mathrm{cost}(\widetilde{P}_i,\widetilde{c}_i) - \mathrm{cost}(P_i,c_i) \leq \mathrm{cost}(\widetilde{P}_i,c_i) - \mathrm{cost}(P_i,c_i) = \sum_{p\in P_i}\|\xi_p\|_2^2 + 2\sum_{p\in P_i}\langle\xi_p, p-c_i\rangle.$$

Besides, we have the following variance of the error terms

- $\mathbb{E}_{\widetilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right] = \theta n_i d$ and $\mathrm{Var}_{\widetilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right] = O(\theta n_i d^2)$;
- $\mathbb{E}_{\widetilde{P}_i}\left[\sum_{p\in P_i}\langle\xi_p, p-c_i\rangle\right] = 0$ and $\mathrm{Var}_{\widetilde{P}_i}\left[\sum_{p\in P_i}\langle\xi_p, p-c_i\rangle\right] = \theta\cdot\mathsf{OPT}_i$.

From Lemma D.7 by choosing $\delta = 1/k$, we know that with probability at least $1-\frac{0.05}{k}$, we have

$$\sup_{c\in\mathbb{R}^d}\frac{|\sum_{p\in P_i}\langle\xi_p, p-c_i\rangle|}{\mathrm{cost}(P_i,c_i)} = 10\cdot\sqrt{\frac{\theta dk}{\mathsf{OPT}_i}}.$$

Besides from Chybeshev's inequality, we also have $\sum_{p\in P_i}\|\xi_p\|^2 \leq 2\theta n_i dk$ happens with probability at least $1-\frac{0.05}{k}$. Then we conclude the proof by applying union bound on $i\in[k]$. □

As shown in [10], under the stability assumption of the dataset, for any sufficiently good approximate k-MEANS solution $\{c_i,\ldots,c_k\}$, centers c_i are pairwise well-separated.

Lemma D.9 (Restatement of Lemma C.1 in [10]). *Given a γ-stable dataset $P\subset\mathbb{R}^d$ and $\alpha\leq 1+\frac{\gamma}{2}$, let $C=\{c_1,\ldots,c_k\}$ be a α-approximation center set. Then for any $i,j\in[k]$ with $i\neq j$, it holds that*

$$d^2(c_i,c_j) \geq \frac{\gamma\mathsf{OPT}}{2\min(n_i,n_j)}$$

By $\gamma = \alpha\cdot O(1+\frac{\theta nd\log^2(\frac{kd}{\sqrt{\alpha}-1})}{\mathsf{OPT}})$, we have $d^2(c_i,c_j) \geq \alpha\cdot O(\frac{\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha}-1})}{\min(n_i,n_j)})$.

Now we are ready to prove Lemma D.3

Proof of Lemma D.3. For ease of analysis, we prove the structural property of P', and the property of S naturally holds, as S is obtained by uniform sampling from P'. We first show that with high probability, for all i, $|P'_i| > 0.5n_i$. Note that for any $\widehat{p} \in \widetilde{P}_i$, $\widehat{p} \notin P'_i$ implies that $\|\xi_p\| > R_i - d(\widehat{p}, \widehat{c}_i) > O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})$.

Note that since $\xi_{p,j}$ satisfies Bernstein condition for any $j \in [d]$ if $\xi_p \neq 0$, we have

$$\Pr\left[|\xi_{p,j}| \geq t | \xi_p \neq 0\right] \leq 2\exp\left(-\frac{t^2}{2(1+bt)}\right), \forall j \in [d], t > 0.$$

Thus we have for large enough t (larger than b),

$$\Pr\left[|\xi_{p,j}| \geq t | \xi_p \neq 0\right] \leq 2\exp\left(-\Omega(t)\right).$$

It follows that

$$\Pr\left[\|\xi_p\| \geq t | \xi_p \neq 0\right] \leq d \cdot \Pr\left[|\xi_{p,j}| \geq \frac{t}{\sqrt{d}} | \xi_p \neq 0\right] \leq 2d\exp\left(-\Omega(\frac{t}{\sqrt{d}})\right).$$

Thus for any single point $p \in P_i$, with probability at most $o(1)$, $\widehat{p} \notin B(c_i^\star, R_i)$. Then we have $|P'_i| \geq 0.5n_i$ with high probability by Chernoff bound.

Note that for any points $\widehat{p}$ in P'_i,

$$\begin{aligned}
d(\widehat{p}, c_i^\star) \leq\ & d(\widehat{p}, \widehat{c}_i) + d(c_i^\star, \widehat{c}_i) \\
\leq\ & R_i + O(r_i) + O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}) && \text{(Assumption D.2)} \\
\leq\ & O(\overline{r_i}) + O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}) && (r_i \leq 8\overline{r_i}) \\
\leq\ & O(\sqrt{\frac{\mathsf{OPT}}{n_i}}) + O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}) \\
\leq\ & O(\sqrt{\frac{\alpha(\mathsf{OPT} + \theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i}}).
\end{aligned}$$

Suppose there exists $i \in [k]$ such that $C \cap B(c_i^\star, O(\sqrt{\frac{\alpha(\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i}})) = \emptyset$. Then

$$\begin{aligned}
\mathrm{cost}(P', C) \geq\ & \mathrm{cost}(P'_i, C) \\
\geq\ & \sum_{p \in P'_i} d^2(C, c_i^\star) - d^2(p, c_i^\star) \\
\geq\ & \frac{n_i}{2} \cdot O(\frac{\alpha(\mathsf{OPT} + \theta nd)}{n_i}) \\
\geq\ & \alpha \cdot O(\mathsf{OPT} + \theta nd).
\end{aligned}$$

By Lemma D.8, we have with high probability

$$\mathsf{OPT}' \leq O(\mathsf{OPT} + \theta nd).$$

Then we have

$$\frac{\mathrm{cost}(P', C)}{\mathsf{OPT}'} \geq \frac{\mathrm{cost}(P'_i, C)}{\mathsf{OPT}'} \geq \frac{\alpha \cdot O(\mathsf{OPT} + \theta nd)}{\mathsf{OPT}'} \geq \alpha.$$

Thus we prove it by contradiction. This directly implies that for any $i \in [k]$, there exists exactly one center $c_i \in C$ satisfying $c \in B(c_i^\star, O(\sqrt{\frac{\alpha(\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i}}))$. Moreover, by Lemma D.9, for any $j \in [k], j \neq i$,

$$d^2(c_i, c_j) \geq \alpha \cdot O(\frac{\mathsf{OPT} + \theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{\min(n_i, n_j)}).$$

Then we have

$$d(p, c_i) \le r_i + d(c_i, c_i^\star) \le O(d(c_i, c_j)).$$

Thus for any point $p \in P_i$,

$$d(p, c_i) < d(c_i, c_j) - d(p, c_i) < d(p, c_j),$$

which completes the proof. □

D.2 Proof of Lemma D.4: properties of S

Lemma D.10 (Movement of coreset centers). *Let P be a set of n points and let S be a set of m points sampled uniformly at random with replacement from P, let $\mathsf{OPT} = \sum_{x\in P} \|x - \mu(P)\|_2^2$, then*

$$\mathbb{E}(\|\mu(P) - \mu(S)\|_2^2) = \frac{\mathsf{OPT}}{nm}.$$

Proof. For each $p \in P$, let $d_p := p - \mu(P)$. Obviously $\sum_{p\in P} d_p = 0$.

Thus

$$\begin{aligned}
\mathbb{E}\left(\|\mu(P) - \mu(S)\|_2^2\right) &= \mathbb{E}\left(\left\|\frac{\sum_{p\in S}(p - \mu(P))}{m}\right\|_2^2\right) \\
&= \frac{1}{m^2}\mathbb{E}\left(\|\sum_{p\in S} d_p\|_2^2\right) \\
&= \frac{1}{m^2}\mathbb{E}\left(\sum_{p\in S}||d_p||_2^2 + 2\sum_{p_i,p_j\in S, i<j}\langle d_{p_i}, d_{p_j}\rangle\right) \\
&= \frac{\sum_{p\in P}\|d_p\|_2^2}{mn} + \frac{1}{\mathcal{P}(m,n)}\sum_{p\in P}\sum_{q\in P} < d_p, d_q > \\
&= \frac{\mathsf{OPT}}{nm} + \frac{1}{\mathcal{P}(m,n)}\left\langle\sum_{p\in P} d_p, \sum_{q\in P} d_q\right\rangle \\
&= \frac{\mathsf{OPT}}{nm},
\end{aligned}$$

where $\mathcal{P}(m, n)$ is a polynomial in m and n. □

Now we prove Lemma D.4.

Proof of Lemma D.4. For any $i \in [k]$, by Lemma D.10, $\mathbb{E}(\|\mu(P_i') - \mu(S_i)\|^2) = \frac{\mathsf{OPT}_i'}{|P_i'|\cdot m_i}$.

Note that $|P_i'| > 0.5n_i$ and $m_i > O(\frac{1}{\varepsilon})$. Let $\varepsilon' = \varepsilon - \frac{\sqrt{\alpha-1}\theta nd}{\mathsf{OPT}}$. Then by Hoeffding bound, with $p \ge 0.99$,

$$\|\mu(P_i') - \mu(S_i)\|^2 \le O\left(\frac{\log k\mathsf{OPT}_i'}{|P_i'|\cdot m_i}\right) = O\left(\frac{\varepsilon'\mathsf{OPT}_i'}{n_i}\right).$$

Next we discuss the upper bound of OPT_S. Let $C' = c_1', \ldots, c_k'$ be the optimal center set for P' with cost OPT'. Given an α-approximate center set $C_\alpha = \{c_1, \ldots, c_k\}$, For each point $p \in S_i$, $w_p = \frac{|P_i'|}{|S_i|}$, thus

$$\mathbb{E}[\mathrm{cost}(S_i, C')] = \frac{m_i}{|P_i'|}\sum_{x\in P_i'} d^2(x, C')\cdot w_x = \mathsf{OPT}_i'.$$

Then by Chernoff bound, we have for any $t > 0$ and $i \in [k]$,

$$\Pr\left[\left|\mathrm{cost}(S_i, c_i') - \mathsf{OPT}_i'\right| \geq t\mathsf{OPT}_i'\right] \leq \exp\left(-\Omega(\frac{t^2}{m_i})\right).$$

Set $t = O\left(\sqrt{\frac{k\log k}{m}}\right)$ and by union bound, it follows that

$$\Pr\left[\left|\mathrm{cost}(S, C') - \mathsf{OPT}'\right| \geq t\mathsf{OPT}'\right] \leq 0.01$$

Thus, with probability > 0.99,

$$\mathsf{OPT}_S \leq \mathrm{cost}(S, C') \leq \left(1 + O\left(\sqrt{\frac{k\log k}{m}}\right)\right)\mathsf{OPT}' \leq (1 + O(\frac{\varepsilon}{\sqrt{\alpha - 1}}))\mathsf{OPT}'.$$

Having the above properties, we conclude that

$$\begin{aligned}
\mathsf{Err}_\alpha(S, \widehat{P}) &\leq \frac{\mathrm{cost}(\widehat{P}, C_\alpha)}{\alpha\mathsf{OPT}_{\widehat{P}}} - 1 \\
&\leq \frac{\sum_i \mathrm{cost}(\widetilde{P}_i, c_i) - \mathrm{cost}(\widetilde{P}_i, c_i')}{\alpha\mathsf{OPT}'} - (1 - \frac{1}{\alpha}) \\
&\leq \frac{\sum_i n_i\|\mu(\widetilde{P}_i) - c_i\|_2^2}{\alpha\mathsf{OPT}'} - (1 - \frac{1}{\alpha}) \\
&\leq \frac{\sum_i n_i(\|\mu(\widetilde{P}_i) - \mu(P_i')\| + \|\mu(P_i') - \mu(S_i)\| + \|\mu(S_i) - c_i\|)^2}{\alpha\mathsf{OPT}'} - (1 - \frac{1}{\alpha}) \\
&\qquad\qquad \text{(Triangle inequality)} \\
&\leq \frac{O(\varepsilon'\mathsf{OPT}' + \sqrt{\alpha - 1}\theta nd) + (\alpha - 1)\mathsf{OPT}_S}{\alpha\mathsf{OPT}'} - (1 - \frac{1}{\alpha}) \\
&\leq O\left(\frac{\varepsilon}{\alpha} + \frac{(\alpha - 1)(\mathsf{OPT}_S - \mathsf{OPT}')}{\alpha\mathsf{OPT}'}\right) \\
&\leq O\left(\frac{\varepsilon}{\alpha} + \frac{\sqrt{\alpha - 1}\varepsilon}{\alpha}\right) \qquad \text{(Lemma D.4)} \\
&\leq \varepsilon
\end{aligned}$$

The last inequality is ensured by fixing a sufficient large constant factor in the sample size. This completes the proof. □

D.3 Proof of Lemma D.5: movement of centers of P'

Recall that $\widetilde{P}_i := \{p + \xi_p | p \in P_i\} \subset \widehat{P}$ represent the cluster with noise corresponding to P_i. Let $O_i := \widetilde{P}_i \setminus B_i$, $O_{i\to j} := \widetilde{P}_i \cap B_j$ where $j \neq i$, $I_i := \cup_{j\neq i} O_{j\to i}$. By the above definition, we have $P_i' = (\widetilde{P}_i \setminus O_i) \cup I_i$.

For simplicity, we use n_i, n_i^{in}, n_i^{out} to denote $|\tilde{P}_i|, |O_i|, |I_i|$ respectively.

Lemma D.11 (Impact of removed points). *With probability $p > 0.99$, for any $i \in [k]$,*

- $\frac{n_i^{out}}{n_i^2}\sum_{p\in O_i}\|p - \mu(P_i')\|_2^2 \leq O(\sqrt{\alpha - 1}\cdot\theta d)$.
- $\frac{n_i^{in}}{n_i^2}\sum_{p\in I_i}\|p - \mu(P_i')\|_2^2 \leq O(\sqrt{\alpha - 1}\cdot\theta d)$.

Proof. Note that since $\xi_{p,j}$ satisfies Bernstein condition for any $j \in [d]$ if $\xi_p \neq 0$, we have

$$\Pr\left[|\xi_{p,j}| \geq t | \xi_p \neq 0\right] \leq 2\exp\left(-\frac{t^2}{2(1 + bt)}\right), \forall j \in [d], t > 0.$$

Thus we have for large enough t (larger than b),

$$\Pr\left[|\xi_{p,j}| \geq t | \xi_p \neq 0\right] \leq 2\exp\left(-\Omega(t)\right).$$

By union bound, with high probability,

$$\Pr\left[\|\xi_p\| \geq t | \xi_p \neq 0\right] \leq d \cdot \Pr\left[|\xi_{p,j}| \geq \frac{t}{\sqrt{d}} \middle| \xi_p \neq 0\right] \leq 2d\exp\left(-\Omega(\frac{t}{\sqrt{d}})\right).$$

Note that for any $p \in O_i$,

$$\|p - \mu(P_i')\|_2 \leq \|p - \xi_p - \mu(P_i')\|_2 + \|\xi_p\|_2 \leq 2(r_i + O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})) + \|\xi_p\|_2$$

Then we can decompose the cost as

$$\begin{aligned}
& \mathbb{E}\left[\sum_{p\in O_i}\|p-\mu(P_i')\|_2^2\right] \\
\leq\ & n_i\cdot\theta\cdot\int_{r_i+O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})}^{}\left(2(r_i+O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}))+t\right)^2\rho(\|\xi_p\|=t)dt \\
\leq\ & n_i\cdot\theta\cdot\int_{r_i+O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})}^{}9t^2\rho(\|\xi_p\|=t)dt.
\end{aligned}$$

Note that when $\Pr\left[\|\xi_p \geq t | \xi_p \neq 0\right] \leq 2d\exp\left(-\Omega(\frac{t}{\sqrt{d}})\right)$, for a constant c, $\rho(\|\xi_p\| = t)$ satisfies that

$$\rho(\|\xi_p\| = t) \leq 2c\sqrt{d}\exp\left(-\frac{ct}{\sqrt{d}}\right).$$

Thus

$$\begin{aligned}
\int_{r_i+O(\theta\sqrt{d}\log k)}^{} t^2\rho(\|\xi_p\|=t)dt \leq\ & d(r_i+O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}))^2\cdot\exp[\Omega(-\frac{r_i+O(\sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})}{\sqrt{d}})] \\
\leq\ & O(\sqrt{\alpha-1}d).
\end{aligned}$$

Then applying Markov's inequality, we have with probability $p > 0.99$,

$$\frac{n_i^{out}}{n_i}\sum_{p\in O_i}\|p-\mu(P_i')\|_2^2 \leq \frac{1}{n_i}\sum_{p\in O_i}\|p-\mu(P_i')\|_2^2 \leq O(\sqrt{\alpha-1}\cdot\theta d).$$

By the definition of I_i, $p \in B(\widetilde{c_i}, R_i)$ for any $p \in I_i$, thus

$$\max_{p\in I_i}\|p-\mu(\widetilde{P_i})\|_2 \leq 2R_i.$$

Note that for any point $p \in I_i \cap \widetilde{P_j}$,

$$\|\widehat{c_i} - \widehat{c_j}\|_2 \leq R_i + \|p - \widehat{c_j}\|_2 \leq R_i + R_j + \|\xi_p\|_2,$$

By the stability of dataset P, every point $p \in I_i \cap \widetilde{P_j}$ satisfies that

$$\|\xi_p\|_2^2 \geq \alpha\cdot O(\frac{\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{\min(n_i,n_j)}) - O(R_i+R_j) \geq O(\frac{\mathsf{OPT}+\theta nd\log^2(\frac{kd}{\sqrt{\alpha-1}})}{\min(n_i,n_j)})$$

Thus

$$\begin{aligned}
\mathbb{E}(\frac{n_i^{in}}{n_i^2}\sum_{p\in I_i}\|p-\mu(P_i')\|_2^2) \le\ & \mathbb{E}(\frac{n_i^{in}}{n_i^2}\cdot 4n_i^{in}R_i^2)\\
&\le \frac{4R_i^2}{n_i^2}\mathbb{E}\left((n_i^{in})^2\right)\\
&\le \frac{4R_i^2n^2\theta^2}{n_i^2}\exp\left[-\Omega(R_i)\right]\cdot\exp\left[-\Omega(\frac{\theta n\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i})\right]\\
&\le O((\frac{\theta n}{n_i})^2)\cdot\exp\left[-\Omega(\frac{\theta n\log^2(\frac{kd}{\sqrt{\alpha-1}})}{n_i})\right]\\
&\le O(\theta d\cdot\sqrt{\alpha-1})
\end{aligned}$$

Then we complete the proof by Markov inequality. □

For the center of P_i and $\tilde{P}_i$, we have the following properties:

Lemma D.12 (Movement of noisy centers). *With high probability, for any $i\in[k]$,*

$$\|\mu(P_i)-\mu(\tilde{P}_i)\|_2^2\le O(\frac{\theta d}{n_i}).$$

Proof. Note that

$$\mu(\tilde{P}_i)=\frac{\sum_{\hat{p}\in\tilde{P}_i}\hat{p}}{n_i}=\mu(P_i)+\frac{\sum_{p\in P_i}\xi_p}{n_i}.$$

Thus it remains to show

$$\|\frac{\sum_{p\in P_i}\xi_p}{n_i}\|_2^2\le O(\frac{\theta d}{n_i}).$$

As shown in Claim D.6, $\mathbb{E}_{\tilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right]=\theta n_i d$ and $\mathrm{Var}_{\tilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right]\le 3\theta n_i d^2$.

We also note that $\mathbb{E}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2\right]\le O(\sqrt{\theta n_i d})$ and

$$\mathrm{Var}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2\right]\le\mathbb{E}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2^2\right]=\mathbb{E}_{\tilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right]=\theta n_i d,$$

which implies that $\left\|\sum_{p\in P_i}\xi_p\right\|_2$ is $O(\sqrt{\theta n_i d})$ with high probability.

Therefore, with high probability, we have:

$$\left\|\mu(\tilde{P}_i)-\mu(P_i)\right\|_2^2\le O\left(\frac{\theta d}{n_i}\right).$$

This completes the proof. □

Now we prove Lemma D.5.

Proof of Lemma D.5. For any $i \in [k]$, we have

$$\begin{aligned}
\|\mu(\widetilde{P}_i) - \mu(P'_i)\|_2^2 = & \ \|\frac{\sum_{p\in\widetilde{P}_i} p}{n_i} - \mu(P'_i)\|_2^2 \\
= & \ \|\frac{\sum_{p\in P'_i} p + \sum_{p\in O_i} p - \sum_{p\in I_i} p}{n_i} - \mu(P'_i)\|_2^2 \\
= & \ \|\frac{\sum_{p\in O_i}(p-\mu(P'_i)) - \sum_{p\in I_i}(p-\mu(P'_i))}{n_i}\|_2^2 \\
\le & \ \frac{2n_i^{out}}{n_i^2}\cdot\sum_{p\in O_i}\|p-\mu(P'_i)\|_2^2 + \frac{2n_i^{in}}{n_i^2}\cdot\sum_{p\in I_i}\|p-\mu(P'_i)\|_2^2 \\
\le & \ O(\theta d\cdot\sqrt{\alpha-1}) \qquad \text{(Lemma D.11)}
\end{aligned}$$

By Lemma D.12, with high probability,

$$\|\mu(P_i) - \mu(\tilde{P}_i)\|_2^2 \le O(\frac{\theta d}{n_i}).$$

Thus

$$\|\mu(P_i) - \mu(P'_i)\|_2^2 \le 2\|\mu(P_i) - \mu(\tilde{P}_i)\|_2^2 + 2\|\mu(\widetilde{P}_i) - \mu(P'_i)\|_2^2 \le O(\theta d\cdot\sqrt{\alpha-1} + \frac{\theta d}{n_i}),$$

which completes the proof. □

D.4 Justifying Assumption D.2: Finding center sets within local balls

Lemmas D.8 and D.12 provide us with control of the clustering cost and the location of the optimal center set $\mathsf{C}(\widehat{P})$ of $\widehat{P}$. Together with Assumption 3.2, we know that $\widehat{P}$ is $O(\alpha)$-cost-stable. By [71, 59], we can efficiently compute an $O(1)$-approximate center set $C \in \mathcal{C}$ for such stable $\widehat{P}$. Partition $\widehat{P}$ into $\widehat{P}_1, \ldots, \widehat{P}_k$ of this C. Using a similar argument as for Lemma D.11, we can show that $\widehat{P} \cap B(c_i^\star, O(r_i + \sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}})) \subseteq \widehat{P}_i$. Furthermore, by a similar argument as for D.12, we can show that $\mu(\widehat{P}_i) \in B(c_i^\star, O(r_i + \sqrt{d}\log\frac{1+\theta kd}{\sqrt{\alpha-1}}))$. Thus, letting $\widehat{C}$ be the collection of $\mu(\widehat{P}_i)$'s satisfies Assumption D.2.

E Missing results and discussions from Section 3

This appendix provides additional results and clarifications supporting our theoretical analysis. We first examine the necessity of the cost-stability assumption in worst-case settings and then illustrate how to examine Assumption 3.2 in practice. Finally, we show the impact of using an approximate optimal solution of P in the algorithms.

E.1 Necessity of the stability assumption in the worst case

We present an example to demonstrate the necessity of the cost-stability assumption (Assumption 3.2) for ensuring a meaningful separation between the traditional error Err and the proposed Err_α.

Consider the 3-Means problem in $\mathbb{R}$ (i.e., $k=3$, $d=1$). Let $P \subset \mathbb{R}$ consist of $\frac{n}{4}$ points each at positions $-3.25\sqrt{2}$, $-1.25\sqrt{2}$, $1.25\sqrt{2}$, and $3.25\sqrt{2}$. A simple calculation shows that the optimal center set $\mathsf{C}(P)$ is either $\{-3.25\sqrt{2}, -1.25\sqrt{2}, 2.25\sqrt{2}\}$ or $\{3.25\sqrt{2}, 1.25\sqrt{2}, -2.25\sqrt{2}\}$, yielding $\mathsf{OPT}_P(k) = n$ in both cases. Moreover, for the 2-Means problem, the optimal centers are at $\{\pm 2.25\sqrt{2}\}$, giving $\mathsf{OPT}_P(k-1) = 2n = 2\cdot\mathsf{OPT}_P(k)$. Hence, P is γ-cost-stable with $\gamma = 1$.

Now let $\widehat{P}$ be a noisy version of P generated under noise model I with $\theta = 1$, where each perturbation is sampled from $N(0,1)$. Using the known expectation $\mathbb{E}_{x\sim N(0,1)}[|x| \mid x \ge 0] = \sqrt{2/\pi}$, we can approximate the means of the three clusters in $\widehat{P}$ as $-2.25\sqrt{2} - \sqrt{2/\pi}$, 0, and $2.25\sqrt{2} + \sqrt{2/\pi}$,

respectively. Thus, the likely optimal center set for $\widehat{P}$ is $\mathsf{C}(S) = \{-2.25\sqrt{2} - \sqrt{2/\pi}, 0, 2.25\sqrt{2} + \sqrt{2/\pi}\}$, leading to:

$$\mathsf{Err}_1(\widehat{P}, P) = \frac{\mathrm{cost}(P, \mathsf{C}(S))}{\mathsf{OPT}_P(k)} - 1 \approx 0.82.$$

By further computation, we approximate $\widehat{P}$ by assuming $\xi_p = \pm\mathbb{E}(|\xi_p|) = \pm\sqrt{2/\pi}$ for each point p, then we observe that $\mathsf{OPT}_{\widehat{P}} \approx 1.24n$, thus $\mathsf{Err}(\widehat{P}, P) \approx \frac{|\mathsf{OPT}_{\widehat{P}} - \mathrm{cost}(P, \mathsf{C}(S))|}{\mathsf{OPT}_{\widehat{P}}} \approx 0.47$. This implies $\mathsf{Err}(\widehat{P}, P) \lesssim \mathsf{Err}_1(\widehat{P}, P)$.

This example underscores the importance of Assumption 3.2: in worst-case scenarios without cost-stability, the two errors may become comparable, and the benefit of using Err_α may diminish.

Empirical comparison of metrics. To further compare the behavior of Err_α and Err, we conduct simulations on synthetic datasets with varying separation levels β.

Let P consist of $\frac{n}{4}$ points each at positions $p_1 = -2\sqrt{2} - \frac{\beta\sqrt{2}}{2}$, $p_2 = \frac{-\beta\sqrt{2}}{2}$, $p_3 = \frac{\beta\sqrt{2}}{2}$, and $p_4 = 2\sqrt{2} + \frac{\beta\sqrt{2}}{2}$. When $\beta = 2.5$, this setup matches the example discussed above. Note that the distances $|p_2 - p_1| = |p_4 - p_3| = 2\sqrt{2}$ remain fixed, while only the inter-cluster gap $|p_3 - p_2|$ varies. This ensures that $\mathsf{OPT}_P = \theta nd$ and $\gamma = 1$ hold for all $\beta \geq 2$.

We set $n = 10{,}000$, $k = 3$, and vary β from 2 to 3 in steps of 0.05. The dataset P is perturbed under noise model I with $\theta = 1$ to obtain $\widehat{P}$.

We compute near-optimal k-means++ centers for P and $\widehat{P}$, denoted $C^\star$ and $\widehat{C^\star}$, respectively. We approximate $\mathsf{Err}_1(\widehat{P}, P)$ by computing $\frac{\mathrm{cost}(P, \widehat{C^\star})}{\mathrm{cost}(P, C^\star)} - 1$.

To estimate $\mathsf{Err}(\widehat{P}, P)$, we randomly sample 500 candidate k-center sets $C_1, \ldots, C_{500}$, each constructed by uniformly sampling k points from the interval $[\min(\widehat{P}), \max(\widehat{P})]$. We define

$$\widehat{\mathsf{Err}}(\widehat{P}, P) := \max_{1 \leq i \leq 500} \frac{|\mathrm{cost}(\widehat{P}, C_i) - \mathrm{cost}(P, C_i)|}{\mathrm{cost}(\widehat{P}, C_i)}$$

as a proxy for the classical Err metric.

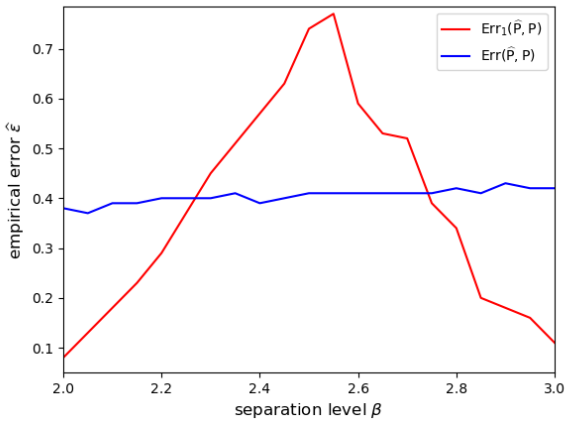

Figure 1: Empirical comparison of error metrics $\mathsf{Err}_1(\widehat{P}, P)$ and $\mathsf{Err}(\widehat{P}, P)$ on synthetic data as the separation level β varies. While Err_1 responds to changing cluster geometry, Err remains nearly constant, illustrating its insensitivity to structural properties in noisy settings.

Figure 1 presents the results. At $\beta = 2.5$, we observe that $\mathsf{Err}(\widehat{P}, P) \approx 0.41$ and $\mathsf{Err}_1(\widehat{P}, P) \approx 0.74$, consistent with our theoretical discussion. The overall trend confirms that Err and Err_α can diverge significantly in value.

Interpretation. This simulation supports our comparison between the classical error metric $\mathsf{Err}(\widehat{P}, P)$ and the proposed surrogate $\mathsf{Err}_\alpha(\widehat{P}, P)$. It shows that Err is relatively insensitive to changes in the structural properties of the dataset, potentially underrepresenting approximation error as data becomes less separable. In contrast, Err_α, approximated here via Err_1, decreases with β (when $\beta > 2.5$), reflecting the growing challenge of summarizing data as cluster separation decreases.

Even when the optimal clustering is unaffected by noise—i.e., OPT_P and cluster geometry remain stable—Err remains nearly unchanged, while Err_α adapts to the increased representational difficulty. This divergence (e.g., 0.41 vs. 0.74 at $\beta = 2.5$) underscores their differing sensitivities.

These findings reinforce our theoretical argument: Err_α is a structurally aware and more reliable metric for noisy settings. This justifies its use in guiding coreset construction and motivates algorithms such as $\mathbf{CN}_\alpha$ that explicitly target this measure.

Interestingly, we observe that Err_1 remains small even when the separation is minor (e.g., $\mathsf{Err}_1 < 0.1$ at $\beta = 2$), suggesting that Err_α may also better align with the ideal metric r_P in the presence of noise when clusters are close. This opens a promising direction for strengthening our theoretical guarantees in low-separation regimes.

E.2 Examining Assumption 3.2 in practice

We detail the process for examining $\widehat{P}$ in $\mathbf{CN}_\alpha$, whose goal is to estimate whether P meets the data assumptions.

Given $\widehat{P}$ and an $O(1)$-approximate center set $\widetilde{C}_k$ for k-MEANS problem, let $\widetilde{\mathsf{OPT}}(k) := \mathrm{cost}(\widehat{P}, \widetilde{C}_k)$. Correspondingly, compute an $O(1)$-approximate center set $\widetilde{C}_{k-1}$, let $\widetilde{\mathsf{OPT}}(k-1) := \mathrm{cost}(\widehat{P}, \widetilde{C}_{k-1})$. For *cost stability* assumption, we directly verify whether

$$\frac{\widetilde{\mathsf{OPT}}(k-1)}{\widetilde{\mathsf{OPT}}(k)} \geq 1 + O(\alpha) \cdot \left(1 + \frac{\theta n d \log^2(kd/\sqrt{\alpha-1})}{\widetilde{\mathsf{OPT}}(k)}\right).$$

Suppose P does not meet the stability assumption, for example, there exists a solution C_{k-1} for $(k-1)$-Means problem, such that $\frac{\mathrm{cost}(P, C_{k-1})}{\mathsf{OPT}_P} < 1+\gamma$. Then it's likely that a corresponding solution for $\widehat{P}$ will also violate the stability assumption.

For the assumption that $r_i \leq 8\overline{r}_i$, consider removing the top 1% of points in $\widehat{P}$ with the greatest distances to $\widetilde{C}$, resulting in a trimmed dataset $\widetilde{P}$. Recompute $r_i, \overline{r}_i$ for this adjusted dataset and $\widetilde{C}$, and check if $r_i \leq 8\overline{r}_i$ holds for every $i \in [k]$.

E.3 Impact of Cost Estimation of OPT_P on Algorithm Performance

We discuss how the estimation of the optimal cost OPT_P affects the performance of $\mathbf{CN}_\alpha$ and $\mathbf{CN}$. Generally speaking, a c-estimate ($c \geq 2$) of the optimal cost on P worsens the error guarantee for $\mathbf{CN}$ (using Err measure) and hardens the data assumption for $\mathbf{CN}_\alpha$ (using Err_α measure). Below we illustrate these points.

Suppose we are given a center set $\widehat{C}$ for coreset construction, whose clustering cost is $\mathrm{cost}(\widehat{P}, \widehat{C}) = c \cdot \mathrm{OPT}_P$ for some $c > 1$ (c-estimate), where P is the underlying dataset and $\widehat{P}$ is a given noisy dataset.

Impact on CN. Let S be a coreset derived from CN. Employing a c-approximate estimate $\widehat{C}$ can weaken the quality bound of S by up to a factor of c in the following sense: With an exact estimate of OPT_P, Theorem 3.1 provides a bound on the worst-case quality of an α-approximate center set of S as

$$r_P(S, \alpha) \leq \left(1 + \varepsilon + O\left(\frac{\theta n d}{\mathsf{OPT}_P} + \sqrt{\frac{\theta n d}{\mathsf{OPT}_P}}\right)\right)^2 \cdot \alpha,$$

In contrast, using CN with the approximate estimate $\widehat{C}$ yields the following bound for the derived coreset S:

$$r_P(S, \alpha) \leq \left(1 + c \cdot \varepsilon + O\left(\frac{\theta n d}{\mathrm{OPT}_P} + \sqrt{\frac{\theta n d}{\mathrm{OPT}_P}}\right)\right)^2 \cdot \alpha.$$

Compared to the exact estimation case, this bound increases the error term from ε to $c \cdot \varepsilon$. This arises because the importance sampling procedure from [10], used in $\mathbf{CN}$, introduces an additional error term $\varepsilon \cdot \frac{\mathrm{cost}(P,\widehat{C})}{\mathsf{OPT}_P} = c \cdot \varepsilon$ in its analysis, whereas the term $O\left(\frac{\theta nd}{\mathsf{OPT}_P} + \sqrt{\frac{\theta nd}{\mathsf{OPT}_P}}\right)$ results from $\mathsf{Err}(\widehat{P}, P)$ and remains unaffected by the quality of the estimation.

Impact on $\mathbf{CN}_\alpha$. Let S be a coreset derived from $\mathbf{CN}_\alpha$. Using a c-approximation estimate $\widehat{C}$ weakens Assumption 3.2 by a factor of c; specifically, the required cost-stability constant γ increases by this factor. Consequently, the range of datasets satisfying the theoretical bounds of $\mathbf{CN}_\alpha$ (as stated in Theorem 3.3) shrinks. This occurs because the performance guarantee of $\mathbf{CN}_\alpha$ relies on correctly identifying k separated clusters, a task complicated by the approximation error in $\widehat{C}$. Increasing γ by a factor of c ensures these clusters are accurately identified, restoring the desired guarantees.

Additionally, using a c-estimate may still result in a high-qualified coreset as with perfect estimate in practice.

F Extensions of Theorems 3.1 and 3.3

This section extends Theorems 3.1 and 3.3 to other noise models and general (k, z)-CLUSTERING.

F.1 Extension to other noise models

Noise model II. Recall that under noise model II, for every $i \in [n]$ and $j \in [d]$, $\widehat{p}_{i,j} = p_{i,j} + \xi_{p,j}$, where $\xi_{p,j}$ is drawn from D_j, a probability distribution on $\mathbb{R}$ with mean 0, variance σ^2, and satisfying the Bernstein condition. The following Theorem shows the performance of coreset generated using the Err metric under noise model II. The only difference from Theorem 3.1 is that we replace θ by σ^2.

Theorem F.1 (Coreset using Err under noise model II). *Let $\widehat{P}$ be drawn from P via noise model II with known $\sigma^2 \geq 0$. Let $\varepsilon \in (0,1)$ and fix $\alpha \geq 1$. Let $\mathcal{A}$ be an algorithm that constructs a weighted subset $S \subset \widehat{P}$ for k-MEANS of size $\mathcal{A}(\varepsilon)$ and with guarantee $\mathsf{Err}(S, \widehat{P}) \leq \varepsilon$. Then*

$$\mathsf{Err}(S,P) \leq \varepsilon + O\left(\frac{\sigma^2 nd}{\mathsf{OPT}_P} + \sqrt{\frac{\sigma^2 nd}{\mathsf{OPT}_P}}\right) \text{ and } r_P(S,\alpha) \leq \left(1 + \varepsilon + O\left(\frac{\sigma^2 nd}{\mathsf{OPT}_P} + \sqrt{\frac{\sigma^2 nd}{\mathsf{OPT}_P}}\right)\right)^2 \cdot \alpha.$$

Proof. By a similar argument as that of Theorem 3.1, we only need to prove the following property:

$$\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] = O(\sigma^2 nd), \text{ and } \mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P} \|\xi_p\|_2^2\right] = O(\sigma^2 nd^2),$$

which again holds by the Bernstein condition. □

For the Err_α metric, we first give the data assumption under noise model II, adapting from Assumption 3.2.

Assumption F.2 (Data assumption under noise model II). Given $\alpha \geq 1$ and $\sigma^2 \geq 0$, assume P is γ-cost-stable with

$$\gamma = O(\alpha) \cdot \left(1 + \frac{\sigma^2 nd \log^2\left(\frac{kd}{\sqrt{\alpha-1}}\right)}{\mathsf{OPT}_P}\right),$$

and that $r_i \leq 8\overline{r}_i$ for all $i \in [k]$.

Theorem F.3 (Coreset using the Err_α metric under noise model II). *Let $\widehat{P}$ be an observed dataset drawn from P under the noise model II with known parameter $\sigma^2 \in [0, \frac{\mathsf{OPT}_P}{nd}]$. Let $\varepsilon \in (0,1)$ and fix $\alpha \in [1,2]$. Under Assumption F.2, there exists a randomized algorithm that constructs a weighted $S \subset \widehat{P}$ for k-MEANS of size $O\left(\frac{k \log k}{\varepsilon - \frac{\sqrt{\alpha-1}\sigma^2 nd}{\alpha \mathsf{OPT}_P}} + \frac{(\alpha-1)k\log k}{(\varepsilon - \frac{\sqrt{\alpha-1}\sigma^2 nd}{\alpha \mathsf{OPT}_P})^2}\right)$ and with guarantee $\mathsf{Err}_\alpha(S, \widehat{P}) \leq \varepsilon$. Moreover,*

$$\mathsf{Err}_\alpha(S,P) \leq \varepsilon + O\left(\frac{\sigma^2 kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\sigma^2 kd\mathsf{OPT}_P} + \sigma^2 nd}{\mathsf{OPT}_P}\right) \text{ and}$$

$$r_P(S,\alpha) \leq \left(1 + \varepsilon + O\left(\frac{\sigma^2 kd}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha} \cdot \frac{\sqrt{\sigma^2 kd\mathsf{OPT}_P} + \sigma^2 nd}{\mathsf{OPT}_P}\right)\right) \cdot \alpha.$$

Proof. By a similar argument as that of Theorem 3.3, we have the following property:

- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(\sigma^2 nd)$, and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(\sigma^2 nd^2)$,
- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\langle \xi_p, p-c\rangle\right] = 0$ and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\langle \xi_p, p-c\rangle\right] = \sigma^2\cdot \mathrm{cost}(P,c)$.

which holds by the Bernstein condition.

Similar to the proof of Lemma D.12, for any $i\in[k]$, $\mathbb{E}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2\right] \le O(\sqrt{\sigma^2 n_i d})$ and

$$\mathrm{Var}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2\right] \le \mathbb{E}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2^2\right] = \mathbb{E}_{\tilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right] = O(\sigma^2 n_i d),$$

Then by Chebyshev's inequality, with high probability, $\left\|\sum_{p\in P_i}\xi_p\right\|_2 \le O(\sqrt{\sigma^2 n_i d})$, which implies

$$\left\|\mu(\tilde{P}_i)-\mu(P_i)\right\|_2^2 = \|\frac{\sum_{p\in P_i}\xi_p}{n_i}\|_2^2 \le O(\frac{\sigma^2 d}{n_i}).$$

The remaining steps remain the same as in Theorem 3.3. □

Non-independent noise across dimensions. For simplicity, we consider a specific setting where the covariance matrix of each noise vector ξ_p is $\Sigma\in\mathbb{R}^{d\times d}$. Note that under the noise model II, $\Sigma = \sigma^2\cdot I_d$ when each $D_j = N(0,\sigma^2)$. In contrast, this setting considers non-independent noise across dimensions.

Note that the above proofs rely on certain concentration properties of the terms $\sum_{p\in P}\|\xi_p\|_2^2$ and $\sum_{p\in P}\langle\xi_p, p-c\rangle$. In general, $\mathbb{E}[\|\xi_p\|_2^2] = \mathrm{trace}(\Sigma)$. Hence, by a similar argument as in the proof of Claim D.6, one can show that $\sum_{p\in P}\|\xi_P\|_2^2$ concentrates on $n\cdot\mathrm{trace}(\Sigma)$ and $\sum_{p\in P}\langle\xi_p, p-c\rangle \le O(\sqrt{\mathrm{trace}(\Sigma)\cdot\mathrm{cost}(P,c)})$. The only difference with Theorem 3.1 and 3.3 is that we replace the original variance term θd to $\mathrm{trace}(\Sigma)$.

Theorem F.4 (Coreset using Err under non-independent noise model). *Let $\widehat{P}$ be drawn from P under non-independent noise model where each ξ_p is drawn from $D_j = N(0,\Sigma)$. Let $\varepsilon\in(0,1)$ and fix $\alpha\ge 1$. Let $\mathcal{A}$ be an algorithm that constructs a weighted subset $S\subset\widehat{P}$ for k-MEANS of size $\mathcal{A}(\varepsilon)$ and with guarantee $\mathsf{Err}(S,\widehat{P})\le\varepsilon$. Then*

$$\mathsf{Err}(S,P) \le \varepsilon + O(\frac{n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P} + \sqrt{\frac{n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P}}) \text{ and}$$

$$r_P(S,\alpha) \le (1+\varepsilon+O(\frac{n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P} + \sqrt{\frac{n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P}}))^2\cdot\alpha.$$

Proof. By a similar argument as that of Theorem 3.1, we only need to prove the following property:

$$\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(n\cdot\mathrm{trace}(\Sigma)), \text{ and } \mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(n\cdot\mathrm{trace}(\Sigma)^2).$$

The remaining steps of the proof remain the same as in Theorem 3.1. □

For the Err_α metric, we also extend Theorem 3.3 to this noise model.

Assumption F.5 (Data assumption under non-independent noise model). Given $\alpha\ge 1$ and a covariance matrix $\Sigma\in\mathbb{R}^{d\times d}$, assume P is γ-cost-stable with

$$\gamma = O(\alpha)\cdot\left(1+\frac{n\cdot\mathrm{trace}(\Sigma)\cdot\log^2(\frac{kd}{\sqrt{\alpha}-1})}{\mathsf{OPT}_P}\right),$$

and that $r_i\le 8\overline{r}_i$ for all $i\in[k]$.

Theorem F.6 (Coreset using the Err_α metric under non-independent noise model). *Let $\widehat{P}$ be an observed dataset drawn from P under non-independent noise model where each ξ_p is drawn from $D_j = N(0,\Sigma)$ with $\mathrm{trace}(\Sigma) \in [0, \frac{\mathsf{OPT}_P}{nd}]$. Let $\varepsilon \in (0,1)$ and fix $\alpha \in [1,2]$. Under Assumption F.5, there exists a randomized algorithm that constructs a weighted $S \subset \widehat{P}$ for k-MEANS of size $O(\frac{k\log k}{\varepsilon - \frac{\sqrt{\alpha-1}n\cdot\mathrm{trace}(\Sigma)}{\alpha\mathsf{OPT}_P}} + \frac{(\alpha-1)k\log k}{(\varepsilon - \frac{\sqrt{\alpha-1}n\cdot\mathrm{trace}(\Sigma)}{\alpha\mathsf{OPT}_P})^2})$ and with guarantee $\mathsf{Err}_\alpha(S,\widehat{P}) \le \varepsilon$. Moreover,*

$$\mathsf{Err}_\alpha(S,P) \le \varepsilon + O(\frac{k\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha}\cdot\frac{\sqrt{k\cdot\mathrm{trace}(\Sigma)\cdot\mathsf{OPT}_P} + n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P}) \text{ and}$$

$$r_P(S,\alpha) \le (1+\varepsilon+O(\frac{k\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P} + \frac{\sqrt{\alpha-1}}{\alpha}\cdot\frac{\sqrt{k\cdot\mathrm{trace}(\Sigma)\cdot\mathsf{OPT}_P} + n\cdot\mathrm{trace}(\Sigma)}{\mathsf{OPT}_P}))\cdot\alpha.$$

Proof. By a similar argument as that of Theorem 3.3, we have the following property:

- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(n\cdot\mathrm{trace}(\Sigma))$, and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O(n\cdot\mathrm{trace}(\Sigma)^2)$,
- $\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\langle\xi_p, p-c\rangle\right] = 0$ and $\mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\langle\xi_p, p-c\rangle\right] = \mathrm{trace}(\Sigma)\cdot\mathrm{cost}(P,c)$.

which holds by the Bernstein condition.

Similar to the proof of Lemma D.12, for any $i\in[k]$,

$$\mathrm{Var}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2\right] \le \mathbb{E}_{\tilde{P}_i}\left[\|\sum_{p\in P_i}\xi_p\|_2^2\right] = \mathbb{E}_{\tilde{P}_i}\left[\sum_{p\in P_i}\|\xi_p\|_2^2\right] = O(n_i\cdot\mathrm{trace}(\Sigma)),$$

Then by Chebyshev's inequality, with high probability, $\left\|\sum_{p\in P_i}\xi_p\right\|_2 \le O(\sqrt{n_i\cdot\mathrm{trace}(\Sigma)})$, which implies

$$\left\|\mu(\tilde{P}_i) - \mu(P_i)\right\|_2^2 = \|\frac{\sum_{p\in P_i}\xi_p}{n_i}\|_2^2 \le O(\frac{\mathrm{trace}(\Sigma)}{n_i}).$$

The remaining steps remain the same as in Theorem 3.3. □

F.2 Extension to (k,z)-CLUSTERING

The following theorem extends Theorem 3.1 to (k,z)-CLUSTERING, under both noise models I and II.

Theorem F.7 ((k,z)-CLUSTERING coreset using the Err_α metric in the presence of noise). *Let $\widehat{P}$ be drawn from P via noise model I (or II) with known $\theta \ge 0$ (or $\sigma^2 \ge 0$). Let $\varepsilon\in(0,1)$ and fix $\alpha \ge 1$. Let $\mathcal{A}$ be an algorithm that constructs a weighted subset $S\subset\widehat{P}$ for k-MEANS of size $\mathcal{A}(\varepsilon)$ and with guarantee $\mathsf{Err}(S,\widehat{P}) \le \varepsilon$. Then for noise model I,*

$$\mathsf{Err}(S,P) \le \varepsilon + O(\frac{\theta n d^{z/2}}{\mathsf{OPT}_P} + \sqrt[z]{\frac{\theta n d^{z/2}}{\mathsf{OPT}_P}}) \text{ and } r_P(S,\alpha) \le (1+\varepsilon+O(\frac{\theta n d^{z/2}}{\mathsf{OPT}_P} + \sqrt[z]{\frac{\theta n d^{z/2}}{\mathsf{OPT}_P}}))^2\cdot\alpha.$$

For noise model II,

$$\mathsf{Err}(S,P) \le \varepsilon + O(\frac{\sigma^z n d^{z/2}}{\mathsf{OPT}_P} + \sqrt[z]{\frac{\sigma^z n d^{z/2}}{\mathsf{OPT}_P}}) \text{ and } r_P(S,\alpha) \le (1+\varepsilon+O(\frac{\sigma^z n d^{z/2}}{\mathsf{OPT}_P} + \sqrt[z]{\frac{\sigma^z n d^{z/2}}{\mathsf{OPT}_P}}))^2\cdot\alpha.$$

Proof. The case of k-MEANS has been proved in Theorem 3.1. Now we show how to extend to (k,z)-CLUSTERING. For the upper bound, the main difference is that we have

$$|d^z(p,C) - d^z(\widehat{p},C)| \le O_z\left(\|\xi_p\|_2^z + \|\xi_p\|_2\cdot d^{z-1}(p,C)\right),$$

Table 2: Datasets used in our experiments. γ represents the cost stability constant. r_i and $\overline{r}_i$ are as defined in Assumption 3.2. Note that the assumption $\max_i \frac{r_i}{\overline{r}_i} < 8$ holds for both datasets. The `Census1990` dataset consists of 2458285 data points and we subsample 100000 from them. We also drop the categorical features of the datasets and only keep the continuous features for clustering.

Dataset	**Size**	**Dim**	k	γ	$\max_i \frac{r_i}{\overline{r}_i}$
`Adult`	41188	10	10	0.07	7.52
`Census1990`	10^5	68	10	0.03	5.93

where $O_z(\cdot)$ hides constant factor $2^{O(z)}$. Similar to Claim D.6, we assume $\sum_{p\in P}\|\xi_p\|_2^z \le O_z(\theta n d^{z/2})$ by the Bernstein condition, which happens with probability at least 0.9. Then we have

$$
\begin{aligned}
& \frac{|\mathrm{cost}_z(P,C)-\mathrm{cost}_z(\widehat{P},C)|}{\mathrm{cost}_z(\widehat{P},C)} \\
\le\ & \frac{O_z\left(\sum_{p\in P}\|\xi_p\|_2^z + \|\xi_p\|_2\cdot d^{z-1}(p,C)\right)}{\mathrm{cost}_z(\widehat{P},C)} \\
\le\ & O(\frac{\theta n d}{\mathsf{OPT}}) + \frac{O_z\left(\sum_{p\in P}\|\xi_p\|_2\cdot d^{z-1}(p,C)\right)}{\mathsf{OPT}} && \text{(by assumption)} \\
\le\ & O(\frac{\theta n d^{\frac{z}{2}}}{\mathsf{OPT}}) + \frac{O_z\left(\sqrt[z]{(\sum_{p\in P}\|\xi_p\|_2^z)(\sum_{p\in P} d^z(p,C))^{z-1}}\right)}{\mathsf{OPT}} && \text{(Generalized Hölder inequality)} \\
\le\ & O(\frac{\theta n d}{\mathsf{OPT}}) + \sqrt[z]{\frac{O(\theta n d^{z/2})}{\mathsf{OPT}}} && \text{(by assumption)} \\
\le\ & O(\frac{\theta n d}{\mathsf{OPT}} + \sqrt[z]{\frac{\theta n d^{z/2}}{\mathsf{OPT}}}), && \text{(Defn. of OPT)}
\end{aligned}
$$

which completes the proof for (k,z)-CLUSTERING under noise model I.

Similarly, for (k,z)-CLUSTERING under noise model II, we only need to prove the following property:

$$
\mathbb{E}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O_z(\sigma^z n d^{z/2}),\ \text{and}\ \mathrm{Var}_{\widehat{P}}\left[\sum_{p\in P}\|\xi_p\|_2^2\right] = O_z(\sigma^{2z} n d^z),
$$

which again holds by the Bernstein condition. This completes the proof. □

The extension of Theorem 3.3 to general (k,z)-CLUSTERING introduces additional technical challenges, as the optimal center for $k=1$ is not the mean point, making it more difficult to control the location of centers. Adapting the use of the Err_α metric to general (k,z)-CLUSTERING remains an interesting direction for future work.

G Additional empirical results

In this section, we provide the implementation details of **CN** and $\mathbf{CN}_\alpha$, and give more empirical results to corroborate our findings in Section 4. All experiments are conducted using Python 3.11 on an Apple M3 Pro machine with an 11-core CPU, 14-core GPU, and 36 GB of memory.

G.1 Implementation details of CN and $\mathbf{CN}_\alpha$

Implementation of CN. Given the perturbed dataset $\widehat{P}$ and budget $\varepsilon > 0$, **CN** first computes an approximate optimal center set $\widetilde{C}$ of $\widehat{P}$ by k-means++ with max_iter $= 5$ and computes $\widetilde{OPT} = \mathrm{cost}_2(\widehat{P},\widetilde{C})$. Then it constructs a coreset by the importance sampling algorithm [10] with coreset size $|S| = 3k^{1.5}\varepsilon^{-2}$, which represents the size bound derived by Err . The runtime of **CN** is $O(ndk)$.

Table 3: Result of `census1990` dataset under noise model I with Gaussian noise. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the ratio of its empirical approximation ratio over the theoretical bound.

(a) $\theta = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\|S\|$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6835	3260	1922	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.028	1.026	1.043	1.052	1.056
	$\mathbf{CN}_\alpha$	1.057	1.067	1.079	1.071	1.087
κ_S	$\mathbf{CN}$	0.849	0.776	0.724	0.673	0.625
	$\mathbf{CN}_\alpha$	0.961	0.928	0.899	0.857	0.836

(b) $\theta = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\|S\|$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6779	3260	1940	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.038	1.030	1.043	1.052	1.060
	$\mathbf{CN}_\alpha$	1.058	1.059	1.063	1.071	1.071
κ_S	$\mathbf{CN}$	0.655	0.602	0.565	0.530	0.498
	$\mathbf{CN}_\alpha$	0.945	0.905	0.872	0.843	0.812

(c) $\theta = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\|S\|$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6834	3260	1928	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.024	1.036	1.057	1.053	1.059
	$\mathbf{CN}_\alpha$	1.069	1.061	1.065	1.097	1.091
κ_S	$\mathbf{CN}$	0.451	0.427	0.409	0.384	0.363
	$\mathbf{CN}_\alpha$	0.893	0.851	0.821	0.815	0.781

(d) $\theta = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\|S\|$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6708	3176	1922	1320	960
$\widetilde{r}_S$	$\mathbf{CN}$	1.041	1.041	1.054	1.086	1.088
	$\mathbf{CN}_\alpha$	1.079	1.067	1.104	1.122	1.106
κ_S	$\mathbf{CN}$	0.201	0.192	0.187	0.184	0.177
	$\mathbf{CN}_\alpha$	0.682	0.653	0.656	0.648	0.620

Implementation of $\mathbf{CN}_\alpha$. Given the perturbed dataset $\widehat{P}$ and budget $\varepsilon > 0$, $\mathbf{CN}_\alpha$ also first computes an approximate optimal center set $\widetilde{C}$ of $\widehat{P}$ by k-means++ with max_iter $= 5$ and computes $\widetilde{OPT} = \mathrm{cost}_2(\widehat{P}, \widetilde{C})$. Next, we decompose $\widehat{P}$ into k clusters $\widehat{P}_i$ by $\widetilde{C}$. For each $i \in [k]$, compute $\widehat{r}_i = \sqrt{\frac{\mathrm{cost}(\widehat{P}_i, \widehat{c}_i)}{|\widehat{P}_i|}}$. Then we compute $P_i' = \widehat{P}_i \cap B_i$, where ball $B_i := B(\widehat{c}_i, R_i)$ with $R_i := \widehat{r}_i + \sqrt{d}\log 10(1 + \theta k d)$. For each $i \in [k]$, take a uniform sample S_i of size $\min(|P_i'|, \frac{9}{\varepsilon} + \frac{6}{\varepsilon^2})$ as an approximation of theoretical results. Finally, it returns $S = \bigcup_{i\in[k]} S_i$ with $w(p) = \frac{|P_i'|}{S_i}$ for $p \in S_i$. Similarly, the runtime of $\mathbf{CN}_\alpha$ is $O(ndk)$.

G.2 Results on the `Census1990` dataset

The setup has been shown in Section 4. See Table 3 for results. All observations are consistent with that for `Adult`.

G.3 Results under other noise settings

We present additional results under various noise settings using the `Adult` dataset, with the same noise levels and tolerance parameters as in Section 4. Tables 4 and 5 evaluate coreset performance under noise model I, replacing Gaussian noise with Laplacian and Uniform noise, respectively. The results are consistent with Table 1, validating the robustness of our theoretical findings across different noise types. Table 6 assesses performance under noise model II, illustrating the practical relevance of Theorem F.3, an extension of Theorem 3.3. Table 7 considers non-independent noise, as analyzed in Theorem F.6, and again shows consistency with Table 1. Overall, these results further support our theoretical analysis, confirming that coreset performance under noise is primarily governed by the noise variance.

Table 4: Result of `Adult` dataset under noise model I with Laplacian noise $\mathrm{Lap}(0, \frac{1}{\sqrt{2}})$. This choice of Laplacian noise ensures a variance of 1 per coordinate, matching that of the Gaussian noise. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the tightness ratio of its empirical approximation ratio over the theoretical bound.

(a) $\theta = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6428	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.048	1.072	1.121	1.179	1.310
	$\mathbf{CN}_\alpha$	1.138	1.096	1.173	1.155	1.153
κ_S	$\mathbf{CN}$	0.866	0.810	0.778	0.755	0.775
	$\mathbf{CN}_\alpha$	1.035	0.953	0.978	0.924	0.887

(b) $\theta = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.036	1.031	1.093	1.118	1.243
	$\mathbf{CN}_\alpha$	1.083	1.111	1.120	1.188	1.173
κ_S	$\mathbf{CN}$	0.600	0.554	0.547	0.522	0.542
	$\mathbf{CN}_\alpha$	0.956	0.940	0.908	0.926	0.880

(c) $\theta = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.030	1.075	1.217	1.095	1.281
	$\mathbf{CN}_\alpha$	1.081	1.113	1.174	1.116	1.147
κ_S	$\mathbf{CN}$	0.370	0.364	0.389	0.331	0.367
	$\mathbf{CN}_\alpha$	0.855	0.847	0.860	0.789	0.783

(d) $\theta = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.036	1.073	1.132	1.167	1.384
	$\mathbf{CN}_\alpha$	1.158	1.125	1.183	1.204	1.221
κ_S	$\mathbf{CN}$	0.130	0.130	0.132	0.132	0.151
	$\mathbf{CN}_\alpha$	0.603	0.571	0.586	0.581	0.576

Table 5: Result of `Adult` dataset under noise model I with Uniform noise $U[-\sqrt{3}, \sqrt{3}]$. This choice of Uniform noise ensures a variance of 1 per coordinate, matching that of the Gaussian noise. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the tightness ratio of its empirical approximation ratio over the theoretical bound.

(a) $\theta = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.057	1.139	1.043	1.281	1.265
	$\mathbf{CN}_\alpha$	1.106	1.097	1.142	1.169	1.155
κ_S	$\mathbf{CN}$	0.873	0.861	0.724	0.820	0.748
	$\mathbf{CN}_\alpha$	1.006	0.954	0.951	0.935	0.889

(b) $\theta = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.026	1.050	1.128	1.224	1.158
	$\mathbf{CN}_\alpha$	1.089	1.125	1.098	1.208	1.130
κ_S	$\mathbf{CN}$	0.594	0.564	0.564	0.571	0.505
	$\mathbf{CN}_\alpha$	0.962	0.951	0.891	0.942	0.848

(c) $\theta = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3145	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.061	1.051	1.092	1.151	1.208
	$\mathbf{CN}_\alpha$	1.131	1.133	1.136	1.163	1.111
κ_S	$\mathbf{CN}$	0.381	0.356	0.349	0.348	0.346
	$\mathbf{CN}_\alpha$	0.894	0.862	0.832	0.822	0.759

(d) $\theta = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.064	1.061	1.099	1.228	1.295
	$\mathbf{CN}_\alpha$	1.117	1.136	1.217	1.220	1.214
κ_S	$\mathbf{CN}$	0.133	0.128	0.128	0.139	0.141
	$\mathbf{CN}_\alpha$	0.582	0.576	0.602	0.589	0.573

Table 6: Result of `Adult` dataset under noise model II with Gaussian noise $N(0, \sigma^2)$. Here, σ^2 controls the variance of noise, playing the same role as θ under noise model I. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the tightness ratio of its empirical approximation ratio over the theoretical bound.

(a) $\sigma^2 = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.023	1.071	1.133	1.219	1.220
	$\mathbf{CN}_\alpha$	1.085	1.126	1.170	1.156	1.140
κ_S	$\mathbf{CN}$	0.845	0.810	0.787	0.780	0.722
	$\mathbf{CN}_\alpha$	0.987	0.979	0.975	0.925	0.877

(b) $\sigma^2 = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.039	1.066	1.075	1.104	1.280
	$\mathbf{CN}_\alpha$	1.096	1.120	1.137	1.164	1.161
κ_S	$\mathbf{CN}$	0.602	0.573	0.538	0.515	0.558
	$\mathbf{CN}_\alpha$	0.967	0.947	0.922	0.908	0.871

(c) $\sigma^2 = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.045	1.074	1.218	1.135	1.234
	$\mathbf{CN}_\alpha$	1.098	1.115	1.129	1.123	1.139
κ_S	$\mathbf{CN}$	0.375	0.363	0.389	0.343	0.353
	$\mathbf{CN}_\alpha$	0.869	0.849	0.828	0.794	0.778

(d) $\sigma^2 = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.042	1.098	1.127	1.292	1.323
	$\mathbf{CN}_\alpha$	1.079	1.125	1.217	1.215	1.231
κ_S	$\mathbf{CN}$	0.130	0.133	0.132	0.146	0.145
	$\mathbf{CN}_\alpha$	0.562	0.571	0.602	0.587	0.581

Table 7: Result of `Adult` dataset under noise model II with non-independent Gaussian noise $N(0, \sigma^2\Sigma)$ with $\mathrm{trace}(\Sigma) = d$. We first generate the covariance matrix Σ randomly, and then apply it for different noise levels σ^2. This choice of $\mathrm{trace}(\Sigma)$ ensures a variance of $\sigma^2 d$ per point, matching that of noise model II. $|S|$ represents the coreset size, $\widetilde{r}_S$ represents its empirical approximation ratio, and κ_S denotes the tightness ratio of its empirical approximation ratio over the theoretical ratio.

(a) $\sigma^2 = 0$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.029	1.051	1.101	1.265	1.227
	$\mathbf{CN}_\alpha$	1.064	1.149	1.136	1.119	1.175
κ_S	$\mathbf{CN}$	0.851	0.795	0.764	0.810	0.726
	$\mathbf{CN}_\alpha$	0.967	0.999	0.947	0.895	0.904

(b) $\sigma^2 = 0.01$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.022	1.039	1.133	1.147	1.240
	$\mathbf{CN}_\alpha$	1.077	1.165	1.150	1.088	1.164
κ_S	$\mathbf{CN}$	0.592	0.559	0.566	0.535	0.541
	$\mathbf{CN}_\alpha$	0.951	0.985	0.933	0.848	0.873

(c) $\sigma^2 = 0.05$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6418	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.038	1.036	1.043	1.152	1.337
	$\mathbf{CN}_\alpha$	1.083	1.086	1.141	1.165	1.167
κ_S	$\mathbf{CN}$	0.373	0.350	0.333	0.348	0.383
	$\mathbf{CN}_\alpha$	0.857	0.826	0.836	0.824	0.797

(d) $\sigma^2 = 0.25$

ε		**0.1**	**0.15**	**0.2**	**0.25**	**0.3**
$\lvert S\rvert$	$\mathbf{CN}$	9486	4216	2371	1517	1054
	$\mathbf{CN}_\alpha$	6445	3178	1940	1320	960
$\widehat{r}_S$	$\mathbf{CN}$	1.090	1.145	1.221	1.219	1.304
	$\mathbf{CN}_\alpha$	1.129	1.199	1.202	1.193	1.239
κ_S	$\mathbf{CN}$	0.136	0.138	0.143	0.138	0.142
	$\mathbf{CN}_\alpha$	0.588	0.609	0.595	0.576	0.584

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Contributions are shown in Sections 2, 3, and 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Assumption 3.2, Section C and Section D.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)? s

   Answer: [Yes]

   Justification: See Sections 4 and G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See https://github.com/xiaohuangyang/Coresets-for-Clustering-Under-Stochastic-Noise.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4 and Table 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: As mentioned in Section 4, we repeat each setting 10 times and report the averages.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: All submissions are anonymous.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mention that datasets are from the UCI Machine Learning Repository in Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.