CORESETS FOR k-MEANS: WHEN DOES UNIFORM SAMPLING SUFFICE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Data summarization techniques, such as coresets, are essential for scaling clustering algorithms to massive datasets. We study uniform sampling, which is an extremely simple and ultra-efficient method to construct coresets, in the context of weak, stable, and strong coresets for the k-means problem in Hilbert space. We begin with the fundamental case $k = 1$, and establish matching upper and lower bounds on the sample complexity of stable and weak coresets, finding that the two notions require different sample complexities. As a consequence, we obtain tight bounds for stable 1-median coresets in ℓ_1 and ℓ_2 (and thus also Kendall–tau and Jaccard metrics), removing the dimension dependence of Carmel and Krauthgamer (ICLR 2026). Our analysis is substantially simpler than prior work and reduces to bounding the deviation of the sample mean. For the general k-means problem on ρ-separated instances, a well-studied structural assumption on the input, we show that uniform sampling yields a stable coreset for all $O(1)$-approximate solutions, and a strong coreset if the instance additionally has bounded kurtosis. In all these results, the sample complexity is independent of the dimension and the instance size.

1 INTRODUCTION

Center-based clustering seeks to partition a set of points into groups of similar points and further assign a single representative center to each partition. Formally, in the (k, z)-clustering problem, the input is a set P of n points in a metric space $(\mathcal{X}, \mathrm{dist})$ and an integer $k \geq 1$, and the goal is to find a set of k centers $C \subseteq \mathcal{X}$ minimizing $\mathrm{cost}_z(C, P) := \frac{1}{n}\sum_{p\in P} \min_{c\in C} \mathrm{dist}(p, c)^z$. We focus on $z \in [1, 2]$ where the cases $z = 1$ and $z = 2$ are the classical k-median and k-means problems, respectively. The special case $k = 1$, namely the 1-median and 1-mean problems, is interesting in its own right and serves as a basic building block for the general problem.

Modern datasets are massive, and this motivates data summarization techniques that let downstream algorithms run on a smaller input rather than the original one. A coreset is a small, reweighted subset of the input that provably preserves the relevant properties of the clustering objective, so that solving the problem (optimally or near optimally) on the coreset gives a near-optimal solution for the original input. Coresets have been the subject of an extensive line of work, offering tunable tradeoffs between size and accuracy, and are the central tool we study here. (Since our results focus on coresets constructed by uniform sampling, we present the coreset definitions without weights.)

The prevailing coreset notion is the *strong ϵ-coreset*, a subset $Q \subseteq P$ such that for every candidate center set $C \subseteq \mathcal{X}$ of size k,

$$\mathrm{cost}_z(C, Q) \in (1 \pm \epsilon)\,\mathrm{cost}_z(C, P). \tag{1}$$

It is well known that constructing a strong coreset requires reading the entire input P, as a single far point (outlier) can dominate the cost, and not including it in Q would violate the guarantee. At the scale of modern datasets, reading the entire input is prohibitive and thus other notions were suggested. A *weak* (ϵ, η)*-coreset* is such a notion, that in contrast, only requires that near-optimal solutions on Q remain near-optimal on P. Writing $\mathrm{OPT}_z(Q) := \min_{C'\subseteq\mathcal{X},|C'|=k} \mathrm{cost}_z(C', Q)$, and similarly for P, it formally requires that for all $C \subseteq \mathcal{X}$ of size k,

$$\text{if } \mathrm{cost}_z(C, Q) \leq (1 + \epsilon)\ \mathrm{OPT}_z(Q), \quad \text{then } \mathrm{cost}_z(C, P) \leq (1 + \eta)\ \mathrm{OPT}_z(P). \tag{2}$$

While this guarantees the most basic property one would expect from a coreset, it says nothing about candidates C that are not near-optimal on Q. This is a serious limitation whenever the centers are constrained to lie in some subset $X \subset \mathcal{X}$, since an optimal solution over X need not be near-optimal over all of $\mathcal{X}$ (we give concrete examples of such constraints below).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recently, Carmel & Krauthgamer (2026) introduced an intermediate notion, the *stable* (ϵ, η)*-coreset*, which strengthens the weak coreset guarantee by approximately preserving the relative order of costs between *every* pair of candidate solutions. Formally it requires that, for all $C_1, C_2 \subseteq \mathcal{X}$ of size k,

$$\text{if } \operatorname{cost}_z(C_1, Q) \leq (1+\epsilon)\operatorname{cost}_z(C_2, Q), \quad \text{then } \operatorname{cost}_z(C_1, P) \leq (1+\eta)\operatorname{cost}_z(C_2, P). \tag{3}$$

This guarantee transfers to every submetric of $\mathcal{X}$, and hence to every variant of the problem in which the centers are restricted to a subset $X \subset \mathcal{X}$, without even needing to know X in advance, yet it is permissive enough to be achieved by uniform sampling without reading the entire input. Concretely, it is shown in (Carmel & Krauthgamer, 2026) that uniform sampling produces a stable coreset for 1-median in ℓ_1, and since tree metrics, Hamming, Kendall-tau, and Jaccard distance are all submetrics of ℓ_1, this transfer-to-submetric property immediately unlocks coreset results for all of these, which were not previously studied. This property also captures constraints on the solution itself, such as the discrete centers (medoids) variant, where centers must be input points, or kernel-based clustering, where the data points are images of a feature map and also the centers are constrained to lie in that map's image (Feldman et al., 2007; Jiang et al., 2024; Carmel et al., 2026).

Even though (k, z)-clustering is NP-hard in general (in fact APX-hard (Cohen-Addad et al., 2022a)), it is solved routinely in practice, and this apparent tension has motivated a large body of work isolating structural assumptions under which the problem becomes tractable (see (Ben-David, 2015; Roughgarden, 2020) for surveys). One recurring assumption is ρ-separation, introduced by Ostrovsky et al. (2013), which requires that the optimal cost with k centers is substantially smaller than with $k-1$ centers, formalizing the intuition behind the elbow method used in practice to choose k. Another common assumption is that each cluster has bounded dispersion, in the sense of not having a heavy tail, for instance because it arises from Gaussian or sub-exponential distributions; we formalize this via a bound on the kurtosis, which is the fourth moment of each cluster relative to the square of its variance.

All our coreset results are obtained by sampling points from P uniformly. Uniform sampling is a cheap, data-oblivious approach that requires no preprocessing of the data, and is therefore particularly suitable for streaming and distributed settings. This distinguishes it from sensitivity sampling and other strong coreset constructions, which inherently require processing the entire input. Our goal is to understand which guarantees can be obtained without processing the entire input. This stands in contrast to much of the literature, which has generally dismissed uniform sampling as a viable construction for (strong) coresets, in both theory and practice (Schwiegelshohn & Sheikh-Omar, 2022, Appendix B). Thus, the structural assumptions we use below for general k should be viewed as investigating the conditions under which uniform sampling succeeds. As discussed in (Meyerson et al., 2004; Huang et al., 2023), one structural requirement for such success is some form of balancedness of the optimal clusters, since a cluster of small size is likely to be completely missed by the sample. This is in line with practice, where a very small fraction of points is routinely treated as outliers rather than as an extra cluster. We adopt a standard formalization called β*-balancedness*, which requires that every cluster in the optimal solution contains at least a β/k-fraction of the points. Our objective is to find what uniform sampling can really achieve despite these apparent limitations.

1.1 Our contribution

Our results are obtained in Hilbert space, which from now on we denote by $\mathcal{H} = \mathcal{X}$. Our first technical result shows that uniform sampling yields a stable coreset for the 1-mean problem in $\mathcal{H}$, with a sample size that is independent of the dimension. Previously, no bound was known for this problem.

Theorem 1.1 (Stable coreset for 1-mean)**.** *Let $\mathcal{H}$ be a Hilbert space and let $P \subseteq \mathcal{H}$ be finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-2}\delta^{-1})$ from P is, with probability at least $1-\delta$, a stable $(\delta\epsilon,\, 8\epsilon)$-coreset for the 1-mean problem.*

Since for every $p \in [1, 2]$ the space ℓ_p^d ($\mathbb{R}^d$ endowed with the ℓ_p norm) embeds isometrically into ℓ_2-squared (Schoenberg, 1938; Bretagnolle et al., 1967), Theorem 1.1 immediately yields the same guarantee also for 1-median in ℓ_p^d.

Corollary 1.2 (Stable coreset for 1-median)**.** *Fix $p \in [1, 2]$ and let $P \subset \mathbb{R}^d$ be finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-2}\delta^{-1})$ from P is, with probability at least $1-\delta$, a stable $(\delta\epsilon,\, 8\epsilon)$-coreset for 1-median in ℓ_p^d.*

This corollary is perhaps our most important contribution for $k = 1$, as it improves the two known bounds on stable coresets for 1-median and has several applications (see also Appendix F for additional implications of Theorem 1.1). The endpoints $p = 1$ and $p = 2$ recover 1-median in ℓ_1^d and ℓ_2^d, and the same conclusion holds for every metric

Metric	Coreset type	Coreset size	Reference
ℓ_2^d	weak	$O(\epsilon^{-4})$	Cohen-Addad et al. (2021a)
		$O(\epsilon^{-3})$	Huang et al. (2023)†
		$O(\epsilon^{-2}\log\epsilon^{-1})$	Woodruff & Yasuda (2024)
ℓ_2^d	stable	$O(\epsilon^{-2}\log(d/\epsilon))$	Carmel & Krauthgamer (2026)
		$O(\epsilon^{-2}\log\epsilon^{-1})$	Carmel et al. (2026)
ℓ_1^d	stable	$O(\epsilon^{-2}\log d)$	Carmel & Krauthgamer (2026)
$\ell_p^d,\ p\in[1,2]$	stable	$O(\epsilon^{-2})$	Corollary 1.2

Table 1: Uniform-sampling coreset sizes for stable and weak $(\epsilon, O(\epsilon))$-coresets for the 1-median problem. †The cited work is primarily concerned with the more general k-median problem.

that embeds isometrically into ℓ_1, including Hamming, Kendall-tau, Spearman footrule, Jaccard, and tree metrics. These two prior results are as follows. The first (Carmel & Krauthgamer, 2026) gives a bound for ℓ_1^d that depends logarithmically on the dimension d, proved using VC-dimension machinery. The second (Carmel et al., 2026) employs an iterative size reduction while exploiting the construction of (Carmel & Krauthgamer, 2026), and its bound for ℓ_2^d carries a $\log(1/\epsilon)$ gap that we close. Table 1 summarizes these and related works on 1-median coresets.

Our key idea is to aim for a stronger result, namely, a coreset for 1-mean in Hilbert space, which implies a coreset for 1-median in ℓ_1^d or ℓ_2^d, by embedding these spaces isometrically into ℓ_2-squared and applying the submetric property of stable coresets. Surprisingly, our proof is much simpler as it avoids the VC-dimension machinery entirely, and it furthermore identifies the actual source of difficulty in analyzing stable and weak coresets constructed via uniform sampling. Since every stable coreset is also a weak coreset, our results immediately improve upon all previous weak coreset constructions summarized in Table 1, with substantially simpler analysis.

To place our stable coreset results in context, the next theorem states the corresponding bound for weak coresets. It was essentially proved in (Inaba et al., 1994; Maalouf et al., 2021), albeit only for weak $(0, \epsilon)$-coresets, and for completeness we provide a short proof of the extended bound in Appendix C. We remark that weak $(\epsilon, O(\epsilon))$-coresets were established in (Cohen-Addad et al., 2021a) also for a more general problem called Euclidean power mean, and were later improved in (Woodruff & Yasuda, 2024).

Theorem 1.3 (Weak coreset for 1-mean)**.** *Let $\mathcal{H}$ be a Hilbert space and let $P \subseteq \mathcal{H}$ be finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-1}\delta^{-1})$ from P is, with probability at least $1-\delta$, a weak $(\delta\epsilon,\ 8\epsilon)$-coreset for the 1-mean problem.*

We complement Theorem 1.1 and Theorem 1.3 with matching lower bounds. For weak coresets, there is already a lower bound for every power $z \geq 2$ in (Cohen-Addad et al., 2021a), and thus item (i) below was already known.

Theorem 1.4. *For every $\epsilon, \delta \in (0, 1/20)$, there exists a finite set $P \subset \mathbb{R}$ such that if a uniform sample $Q \subseteq P$ of size s is, with probability at least $1-\delta$,*

(i) a weak $(0,\epsilon)$-coreset for the 1-mean problem, then $s = \Omega(\epsilon^{-1}\delta^{-1})$;
(ii) a stable $(0,\epsilon)$-coreset for the 1-mean problem, then $s = \Omega(\epsilon^{-2}\delta^{-1})$.

Together, Theorems 1.1, 1.3 and 1.4 provide a direct comparison between the two notions for the 1-mean problem in Hilbert space, revealing a clear separation: stable coresets require $\Theta(\epsilon^{-2}\delta^{-1})$ samples, whereas weak coresets require only $\Theta(\epsilon^{-1}\delta^{-1})$. In contrast, there is no such separation for the 1-median problem, as both notions are known to require, up to polylogarithmic factors, $\Theta(\epsilon^{-2})$ samples for constant success probability (Carmel et al., 2026).

We then consider the case of general $k \geq 2$, a regime where uniform-sampling constructions are scarce. Prior stable-coreset results are restricted to $k = 1$ (Carmel & Krauthgamer, 2026; Carmel et al., 2026), and the closest prior work for $k \geq 2$ is a weak k-median coreset of (Huang et al., 2023), and we compare to it further below. For $\gamma \geq 1$, we consider the candidate center sets $\mathcal{S}_\gamma(P) := \{C :\ \mathrm{cost}(C,P) \leq \gamma \cdot \mathrm{OPT}_2(P)\}$ and similarly $\mathcal{S}_\gamma(Q) := \{C :\ \mathrm{cost}(C,Q) \leq \gamma \cdot \mathrm{OPT}_2(Q)\}$. We think of γ as a small constant, since constant-factor approximations for (k,z)-clustering are readily available. We first show that under β-balancedness and ρ-separation assumptions, a uniform sample already preserves the cost order (in the stable sense) of all solutions whose cost on P or on Q is within factor γ of optimal.

Theorem 1.5. *Let $\mathcal{H}$ be a Hilbert space and let $k \geq 2$, $\gamma \geq 1$, $\epsilon \in (0, 1/75)$, $\delta \in (0, 1)$. Then for every finite $P \subset \mathcal{H}$ that is β-balanced for $\beta \in (0, 1]$ and ρ-separated with $\rho^2 \leq \beta\delta/(576\gamma k)$, a uniform sample Q of (sufficiently large) size $O\big(k^2\gamma/(\beta\epsilon^2\delta^2)\big)$ is, with probability at least $1 - \delta$, a stable $(\delta\epsilon, O(\epsilon))$-coreset of P over solutions in $\mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$ for the k-means problem.*

Observe that by setting $\gamma = 1 + \epsilon$, we immediately obtain a weak $(\epsilon, O(\epsilon))$-coreset guarantee. The closest prior result, obtained by Huang et al. (2023), is a weak coreset of size $O\left(\frac{k^2}{\beta\epsilon^3}\log^2\frac{k}{\beta\epsilon}\log^2\frac{1}{\epsilon}\right)$ for k-median in ℓ_2^d, with constant success probability. They too require that P is β-balanced, but their weak coreset guarantee is further restricted to candidate solutions that are balanced on the sample Q. We require instead that the input P is ρ-separated, and for constant success probability our sample complexity matches their bound in k and improves the dependence on ϵ from ϵ^{-3} to ϵ^{-2}, while also giving the stronger stable guarantee over all $O(1)$-approximate solutions. Neither their result nor their technique extends to the k-means problem, whereas ours, stated above for squared-distance objective in Hilbert space, implies the Euclidean k-median setting as a special case. Another advantage of our result is that if the input P has in addition bounded kurtosis, then the stable guarantee strengthens to a strong one, i.e., it preserves the cost of every candidate center set in $\mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$. This improves over the strong coreset result of (Carmel & Krauthgamer, 2026), which is restricted to $k = 1$ and assumes bounded dispersion (each cluster diameter relative to its average distance). Bounded kurtosis is clearly a more relaxed requirement and arguably more typical in practice.

Theorem 1.6. *Let $\mathcal{H}$ be a Hilbert space and let $k \geq 2$, $\gamma \geq 1$, $\epsilon \in (0, 1/75)$, $\delta \in (0, 1)$. Then for every finite $P \subset \mathcal{H}$ that has kurtosis bounded by $M > 0$, is β-balanced for $\beta \in (0, 1]$, and is ρ-separated with $\rho^2 \leq \beta/(120\gamma k)$, a uniform sample Q of (sufficiently large) size $O\big(k^2 M/(\beta\epsilon^2\delta)\big)$ satisfies, with probability at least $1 - \delta$,*

$$\forall C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P), \qquad \mathrm{cost}(C, Q) \in (1 \pm \epsilon)\ \mathrm{cost}(C, P).$$

We complement our theoretical results with an experimental evaluation (Section 4) on real-world datasets, comparing uniform sampling against importance-sampling constructions of coreset for both 1-median and k-means. We find that uniform sampling matches the empirical accuracy of these more expensive methods, despite requiring several orders of magnitude less construction time and being oblivious to the input.

1.2 Technical overview

Upper bound for $k = 1$ (Theorem 1.1). The key idea is to analyze 1-mean in Hilbert space, rather than 1-median in ℓ_1^d or ℓ_2^d. We use the sufficient condition of (Carmel & Krauthgamer, 2026) for a stable coreset, called *relative cost-difference approximation* (RCDA) property: for every candidate center set C, the difference between its cost and the optimal cost is approximately the same whether measured on P or on Q (see Definition 2.2).

Now for the 1-mean problem in Hilbert space $\mathcal{H}$, we prove that the RCDA error at a center c factors into a center-dependent term and a sample-dependent term Δ, which is the deviation of the sample mean from the true mean. Since this Δ does not depend on c, a single concentration bound controls it, and hence also the RCDA error, simultaneously for all centers. Once we establish that a uniform sample Q has the RCDA property (with high probability), we just invoke the framework of (Carmel & Krauthgamer, 2026) (see Theorem 2.3) to conclude that Q is a stable coreset.

Lower bounds for $k = 1$ (Theorem 1.4). Both bounds use the same one-dimensional instance, where $\frac{1}{n}$-fraction of the mass is at a point $D > 0$ and the rest is at the origin, so the sample sees the rare point only through the number of times it is drawn. The separation between the two notions comes from how accurately the sample must reproduce the mean. A weak coreset only needs the sample mean to be within $O\big(\sqrt{\epsilon\,\mathrm{OPT}_2(P)}\big)$ of the true mean. A stable coreset must preserve the cost order of *every* pair of centers, and we detect the pair that is hardest to preserve; namely, two centers placed symmetrically on different sides of the sample mean $\hat{\mu}$. They have identical cost on Q, yet on P their cost ratio grows linearly with the deviation of $\hat{\mu}$ from the true mean. Any nonzero deviation therefore turns a tie on Q into a strict order on P, and forcing this deviation below ϵ demands the sample mean to be a $\sqrt{\epsilon}$ factor closer than in the weak coreset case. This quadratic gap in the tolerated deviation is exactly the ϵ^{-1} versus ϵ^{-2} separation, pinpointing an inherent difference between the two notions, weak and stable coresets. The proof appears in Section D.

Upper bounds for $k > 1$ (Theorems 1.5 and 1.6). The difficulty in moving from $k = 1$ to general k is that the clean factorization of the RCDA error no longer holds. For a single cluster, every candidate center sees the same sample mean, so a single concentration bound controls all candidates simultaneously. For general k, however, each candidate

solution (set of k centers) induces its own Voronoi partition. When the candidate centers vary ("move"), points near cluster boundaries may switch their assigned center, producing an additional error term that depends on the assignment itself. Consequently, unlike the 1-mean case, there is no single empirical quantity whose concentration controls the entire family of candidate solutions. The standard approach to handling such infinite families relies on VC-dimension or covering-number arguments (Huang et al., 2023). We take a different route.

Fix an optimal solution C^* and parametrize every candidate solution by its *displacement* $v = (v_1, \ldots, v_k)$ from the optimal centers, i.e. the candidate is $\{\mu_1 + v_1, \ldots, \mu_k + v_k\}$. In this parametrization, applying the König–Huygens identity separately within each optimal cluster of P and of the sample Q expresses both costs in closed form, and their difference splits into two terms. The first is the *matched-cost error*, in which every point is assigned to the center displaced from its own optimal center, regardless of which candidate center is actually nearest. This term depends only on the empirical weights and centroids of the optimal clusters, and is controlled by the same per-cluster concentration as in the 1-mean case. The second is the *boundary-gain error*, the correction for points whose nearest candidate center is displaced from a different optimal center. Bounding this term is the core technical challenge.

Before turning to it, we exploit the geometry of well-separated instances to show that every candidate of constant-factor approximation is *localized* around C^*: each of its centers lies within a small radius Γ of a distinct optimal center, so the displacements lie in a fixed set $\mathcal{V} \subset \mathcal{H}^k$. Crucially, the set $\mathcal{V}$ is determined by the input alone and does not depend on the sample.

The boundary-gain error depends on the minimum squared distance from each point to the candidate centers. This minimum is taken over all k centers at once, making the error non-linear and non-separable across clusters. To handle it, we first symmetrize the deviation, replacing the comparison between the sample and the population by a comparison between two independent samples, which yields a Rademacher process indexed by the displacements in $\mathcal{V}$. The boundary gain at a point is a maximum of functions of v that are affine except for a quadratic term, and we remove the two obstructions in turn. Maurer's vector contraction inequality (Maurer, 2016) removes the maximum; it bounds a Rademacher process indexed by vector-valued Lipschitz functions, up to a constant factor, by a Rademacher process indexed by the coordinates of those functions. The quadratic term is then removed by *lifting* each displacement v_r to $(v_r, \|v_r\|^2 / 2\Gamma)$ in one extra dimension, which makes every coordinate function linear in the lifted variable. The entire process thereby collapses to a single inner product in a Hilbert space between a bounded lifted displacement and one Rademacher vector, and Cauchy–Schwarz reduces the problem to the second moment of that vector, which we bound directly, without any tail assumption on the dataset P. Overall, we obtain a uniform bound on the *expected* boundary-gain error.

The final step is to convert this expectation bound into a high-probability guarantee, and this is exactly where our two results (Theorems 1.5 and 1.6) diverge. For the strong coreset guarantee, we bound the variance of the boundary-gain error directly. Since changing one sampled point may have a large effect when that point lies far from its optimal cluster center, the resulting variance bound depends on fourth moments, which we control by the bounded-kurtosis assumption. For the stable coreset guarantee, rather than approximating the cost itself, we first establish the ϵ-RCDA guarantee. The RCDA decomposition eliminates the term responsible for the fourth-moment dependence before concentration is applied, leaving only second-moment quantities to control. The trade-off is that this requires a bound on $\mathrm{cost}(C^*, Q)$, which follows easily from Markov's inequality, leading to the weaker dependence on the failure probability. The full proof appears in Section E due to space constraints.

2 Preliminaries

Let $\mathcal{H}$ be a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$. Throughout the remainder of the paper, we focus on the k-means problem, corresponding to the special case $z = 2$ of (k, z)-clustering. Accordingly, we omit the subscript z from the notation. In the *k-means problem* the input is a set $P \subseteq \mathcal{H}$ of size $|P| = n$ and an integer $k \geq 1$, and the goal is to find a k-center set $C \subseteq \mathcal{H}$ minimizing

$$\mathrm{cost}(C, P) := \frac{1}{n} \sum_{p \in P} \min_{c \in C} \|p - c\|^2 .$$

We denote by $\mathrm{OPT} := \min_{C \subseteq \mathcal{H}, |C| = k} \mathrm{cost}(C, P)$ the cost of an optimal solution. Each center set C induces a (Voronoi) partition of P into clusters $\{X_1, \ldots, X_k\}$, where $p \in X_i$ if c_i is a nearest center to p in C (ties broken by smallest label). We fix an optimal solution $C^* = \{\mu_1, \ldots, \mu_k\}$ for the k-means problem with induced partition

$\{X_1^*, \ldots, X_k^*\}$, for $k = 1$ we simply denote $C^* = \mu \in \mathcal{H}$. Define the per-cluster quantities

$$w_i := \frac{|X_i^*|}{n}, \quad V_i := \frac{1}{|X_i^*|} \sum_{p \in X_i^*} \|p - \mu_i\|^2, \quad \text{then } \operatorname{cost}(C^*, P) = \sum_{i=1}^{k} w_i V_i.$$

The following identity decomposes the cost of a single cluster into a bias term and a variance term (also known as the König–Huygens identity).

Claim 2.1. *Let $X \subseteq P$ of size m and $\mu = \frac{1}{m}\sum_{p \in X} p$. Then,*

$$\forall c \in \mathcal{H}, \qquad \operatorname{cost}(c, X) = \|c - \mu\|^2 + \operatorname{cost}(\mu, X). \tag{4}$$

It follows immediately that the centroid of X_i^*, i.e., $\mu_i = \frac{1}{|X_i^*|}\sum_{p \in X_i^*} p$, is the point minimizing the cost for the cluster X_i^*.

Let Q be a uniform sample of size s drawn i.i.d. from P. Define the per-cluster sample quantities

$$Q_i := Q \cap X_i^*, \quad s_i := |Q_i|, \quad \hat{w}_i := \frac{s_i}{s}, \quad \hat{\mu}_i := \frac{1}{s_i} \sum_{q \in Q_i} q, \quad \Delta_i := \hat{\mu}_i - \mu_i, \quad \hat{V}_i := \tfrac{1}{s_i} \sum_{q \in Q_i} \|q - \hat{\mu}_i\|^2 .$$

Now denote $\widehat{\mathrm{OPT}} := \sum_i \hat{w}_i \hat{V}_i$, and notice that this cost and its underlying partition are not necessarily optimal for Q, that is, $\min_{C \subseteq \mathcal{H}, |C|=k} \operatorname{cost}(C, Q) \le \widehat{\mathrm{OPT}}$.

A key tool for establishing stable coresets, introduced in (Carmel & Krauthgamer, 2026), is the following notion, which requires that the cost gap between any candidate center set C and the fixed optimal solution C^* be nearly the same whether measured on P or on the sample Q.

Definition 2.2 (ϵ-RCDA)**.** A subset $Q \subseteq P$ is an *ϵ-relative cost-difference approximation (ϵ-RCDA)* of P in $\mathcal{H}$ if

$$\forall\, C \subseteq \mathcal{H},\ |C| = k, \qquad \Big|\big[\operatorname{cost}(C, P) - \operatorname{cost}(C^*, P)\big] - \big[\operatorname{cost}(C, Q) - \operatorname{cost}(C^*, Q)\big]\Big| \ \le\ \epsilon \operatorname{cost}(C, P). \tag{5}$$

The ϵ-RCDA property implies a stable coreset, provided that the sample does not excessively overestimate the cost of the fixed optimum C^*. The proof given in (Carmel & Krauthgamer, 2026) for 1-median extends verbatim to k-means, for completeness we reproduce the proof in Appendix B.

Theorem 2.3 (Carmel & Krauthgamer (2026) (Theorem 3.1))**.** *Let $\epsilon \in (0, 1/5)$ and suppose that*

(I1) Q is an ϵ-RCDA of P in $\mathcal{H}$, and
(I2) $\operatorname{cost}(C^*, Q) \le \alpha \cdot \operatorname{cost}(C^*, P)$ for some $\alpha \ge 1$.

Then Q is a stable $(\epsilon/\alpha,\ 4\epsilon)$-coreset of P.

3 Upper bound for $k = 1$

3.1 RCDA in Hilbert space

In the $k = 1$ case, we denote the centers by lower-case letters and simply write $\operatorname{cost}(x, P) = \frac{1}{n}\sum \|x - p\|^2$. Also denote $V = \operatorname{cost}(\mu, P)$, $V_Q = \operatorname{cost}(\hat{\mu}, Q)$, where $\hat{\mu} := \frac{1}{s}\sum_{q \in Q} q$ is the sample mean of Q and $\Delta := \hat{\mu} - \mu$.

Applying Claim 2.1 to both P and Q, for all $c \in \mathcal{H}$:

$$\operatorname{cost}(c, P) = \|c - \mu\|^2 + V, \qquad \operatorname{cost}(\mu, P) = V, \tag{6}$$

$$\operatorname{cost}(c, Q) = \|c - \hat{\mu}\|^2 + V_Q, \qquad \operatorname{cost}(\mu, Q) = \|\Delta\|^2 + V_Q. \tag{7}$$

First, we note that the RCDA error can be written as a single inner product.

Claim 3.1. *For every $c \in \mathcal{H}$,*

$$\big[\operatorname{cost}(c, P) - \operatorname{cost}(\mu, P)\big] - \big[\operatorname{cost}(c, Q) - \operatorname{cost}(\mu, Q)\big] \ =\ 2\,\langle c - \mu,\ \Delta\rangle\,.$$

Proof. From (6) and (7):

$$\operatorname{cost}(c,P) - \operatorname{cost}(\mu,P) = \|c-\mu\|^2 + V - V = \|c-\mu\|^2,$$
$$\operatorname{cost}(c,Q) - \operatorname{cost}(\mu,Q) = \left(\|c-\hat{\mu}\|^2 + V_Q\right) - \left(\|\Delta\|^2 + V_Q\right) = \|c-\hat{\mu}\|^2 - \|\Delta\|^2.$$

Expanding $\|c-\hat{\mu}\|^2 = \|(c-\mu)-\Delta\|^2 = \|c-\mu\|^2 - 2\langle c-\mu,\, \Delta\rangle + \|\Delta\|^2$, so

$$\operatorname{cost}(c,Q) - \operatorname{cost}(\mu,Q) = \|c-\mu\|^2 - 2\langle c-\mu,\, \Delta\rangle.$$

Consequently,

$$\big[\operatorname{cost}(c,P) - \operatorname{cost}(\mu,P)\big] - \big[\operatorname{cost}(c,Q) - \operatorname{cost}(\mu,Q)\big] = \|c-\mu\|^2 - (\|c-\mu\|^2 - 2\langle c-\mu,\, \Delta\rangle) = 2\langle c-\mu,\, \Delta\rangle.$$

□

Second, we bound $\|\Delta\|^2$ via concentration (see also (Inaba et al., 1994) for similar concentration analysis).

Claim 3.2. *Let $\epsilon > 0$, then $\Pr\big[\|\Delta\|^2 > \epsilon^2 V\big] \le \frac{1}{s\epsilon^2}$.*

Proof. We can write $\Delta = \frac{1}{s}\sum_{q\in Q}(q-\mu)$, where the summands are i.i.d. with mean zero. By independence,

$$\mathbb{E}\big[\|\Delta\|^2\big] = \frac{1}{s^2}\sum_{q\in Q}\mathbb{E}\big[\|q-\mu\|^2\big] = \frac{1}{s^2}\cdot s\cdot V = \frac{V}{s}.$$

Applying Markov's inequality

$$\Pr\big[\|\Delta\|^2 > \epsilon^2 V\big] \;\le\; \frac{\mathbb{E}[\|\Delta\|^2]}{\epsilon^2 V} = \frac{1}{s\epsilon^2}.$$ □

We combine Claim 3.1 and Claim 3.2 to show that a uniformly sampled Q is an ϵ-RCDA of P.

Claim 3.3. *If $\|\Delta\|^2 \le \epsilon^2 V$, then Q is an ϵ-RCDA of P.*

Proof. By Claim 3.1, the RCDA error at center c equals $|2\langle c-\mu,\, \Delta\rangle|$. By Cauchy–Schwarz and Young's inequality ($2ab \le \epsilon a^2 + b^2/\epsilon$):

$$|2\langle c-\mu,\, \Delta\rangle| \;\le\; 2\|c-\mu\|\cdot\|\Delta\| \;\le\; \epsilon\|c-\mu\|^2 + \frac{1}{\epsilon}\|\Delta\|^2 \;=\; \epsilon(\operatorname{cost}(c,P) - V) + \frac{1}{\epsilon}\|\Delta\|^2.$$

By assumption, $\frac{1}{\epsilon}\|\Delta\|^2 \le \epsilon V$. Hence, $|2\langle c-\mu,\, \Delta\rangle| \le \epsilon\cdot\operatorname{cost}(c,P)$, for all $c\in\mathcal{H}$. □

3.2 Proof of Theorem 1.1

Proof of Theorem 1.1. Let $\epsilon \le 1/5$, $\delta > 0$ and set $s := \lceil 2/(\delta\epsilon^2)\rceil$. By Claim 3.2, $\Pr[\|\Delta\|^2 > \epsilon^2 V] \le \delta/2$. Since $\mathbb{E}[\operatorname{cost}(\mu,Q)] = V$, Markov's inequality gives $\Pr[\operatorname{cost}(\mu,Q) > 2V/\delta] \le \delta/2$. By the union bound, both $\|\Delta\|^2 \le \epsilon^2 V$ and $\operatorname{cost}(\mu,Q) \le 2V/\delta$ hold simultaneously with probability at least $1-\delta$. Conditioned on these events, Claim 3.3 gives that Q is an ϵ-RCDA of P, and $\operatorname{cost}(\mu,Q) \le 2V/\delta = 2\operatorname{cost}(\mu,P)/\delta$. Applying Theorem 2.3, Q is a stable $(\delta\epsilon/2,\, 4\epsilon)$-coreset of P, and the desired result is obtained by rescaling ϵ. □

3.3 Proof of Theorem 1.3

In contrast, we show that for weak coresets it suffices to have $\|\Delta\|^2 \le \epsilon V$, a single power of ϵ rather than ϵ^2 as in Claim 3.3, thus improving upon the required sample size. While a slightly weaker result, yielding a weak $(0,\epsilon)$-coreset via a similar technique, appears in (Inaba et al., 1994; Jubran et al., 2020), we include in Appendix C the proof of the more general statement for completeness.

Theorem 1.3 (Weak coreset for 1-mean)**.** *Let $\mathcal{H}$ be a Hilbert space and let $P \subseteq \mathcal{H}$ be finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-1}\delta^{-1})$ from P is, with probability at least $1-\delta$, a weak $(\delta\epsilon,\, 8\epsilon)$-coreset for the 1-mean problem.*

Table 2: Specifications of datasets used in Section 4.

Dataset	Size n	Dim. d	Description
Yellow Taxi NYC (YT) (NYC Taxi and Limousine Commission, 2024)	2.82M	16	New York City taxi trips in Jan. 2024
Covertype (Blackard, 1998)	581,012	54	Cartographic variables for forest plots in the Roosevelt National Forest of Northern Colorado
Census 1990 (Meek et al., 2001)	2.45M	68	Demographic attributes of individuals from the US Census Public Use Microdata Samples

4 EXPERIMENTS

We evaluate uniform sampling against importance-sampling coreset constructions on widely-used real-world datasets (summarized in Table 2). Generally speaking, the accuracy of uniform sampling in our experiments is consistent with experiments in prior work (Marom & Feldman, 2019; Baker et al., 2020; Jubran et al., 2020; Braverman et al., 2021a; Huang et al., 2023). These standard datasets were not chosen to satisfy the requirements of our analysis, and indeed they do not fully satisfy them. At every value of k, the measured separation is $\hat{\rho}^2 \geq 0.866$, while the measured kurtosis is less than 5.

Our results concern the three coreset guarantees: weak, stable and strong. For each, we define an empirical error that instantiates the corresponding definition, and we report these errors as percentages. The empirical weak-coreset error instantiates (2) with $\epsilon = 0$, as

$$\widehat{E}_{\text{weak}} := \frac{\text{cost}(C^Q, P) - \text{cost}(C^P, P)}{\text{cost}(C^P, P)}, \tag{8}$$

where C^Q and C^P are optimal solutions for Q and for P. The empirical stable-coreset error instantiates (3), as

$$\widehat{E}_{\text{stable}} := \max\Big\{ \tfrac{\text{cost}(C_1,P)}{\text{cost}(C_2,P)} - 1 \ : \ C_1, C_2 \in \mathcal{C},\ \text{cost}(C_1, Q) \leq \text{cost}(C_2, Q) \Big\}. \tag{9}$$

The empirical strong-coreset error instantiates (1), as

$$\widehat{E}_{\text{strong}} := \max_{C \in \mathcal{C}} \Big| \tfrac{\text{cost}(C,Q)}{\text{cost}(C,P)} - 1 \Big|. \tag{10}$$

Whereas Definitions (3) and (1) quantify over every center set in $\mathcal{H}$, in (9) and (10) the maximum is taken over a finite family $\mathcal{C}$ of candidate solutions, that we generated once per dataset and reused across every method and coreset size; both are therefore lower bounds on the true error. In Experiment 2, $\mathcal{C}$ consists of $5{,}000$ sets of k points drawn uniformly from P. In Experiment 1, $\mathcal{C}$ consists of centers along the direction that the proof of Theorem 1.4(ii) identifies as hardest.

Experimental setup. All experiments were conducted on a PC with Intel Core i7-13700 CPU and 32GB RAM running Python 3.12.10 on Windows 10. For each experiment, we report the average results over 20 independent runs with shaded regions representing one standard deviation across the runs. The source code used to run the experiments is available at `https://anonymous.4open.science/r/UniformSamplingCoresets-5C86/`.

Experiment 1: Comparison with importance sampling. We compare uniform sampling against sensitivity sampling (Feldman & Langberg, 2011) on the 1-median problem in the ℓ_1 metric over the Yellow Taxi NYC dataset. We vary the coreset size over $\{200, 250, \ldots, 1000\}$ and report both the empirical weak-coreset error (8) and the empirical stable-coreset error (9). We find that the stable-coreset error exceeds the weak-coreset error roughly by an order of magnitude at every coreset size, reflecting that preserving the order of costs is a substantially stronger requirement than preserving near-optimality alone.

Figure 1 shows no appreciable difference between the two sampling methods, under either notion. The difference in construction time is, in contrast, fundamental: sensitivity sampling examines the entire dataset before it can draw a single point, whereas uniform sampling requires constant time per sample and is oblivious to the dataset. And indeed, producing a coreset of 500 points took $2{,}160$ milliseconds for sensitivity sampling against 0.03 milliseconds for uniform sampling — almost five orders of magnitude faster.

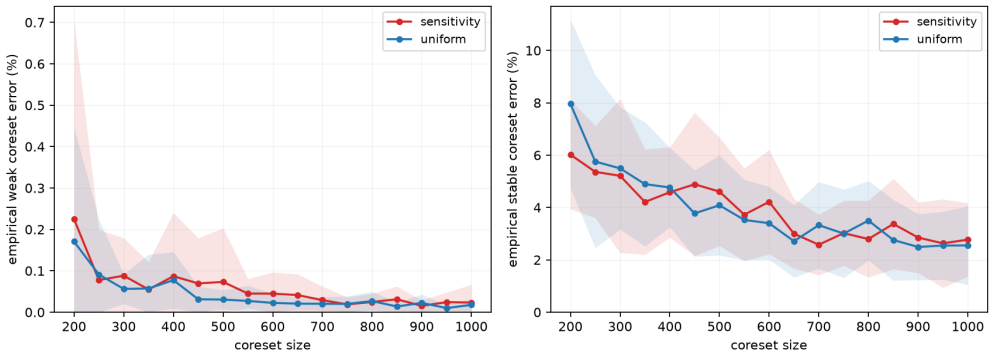


Figure 1: Comparison of uniform and sensitivity sampling on the Yellow Taxi NYC dataset: empirical weak-coreset error (left) and empirical stable-coreset error (right) as a function of coreset size, for 1-median in ℓ_1.

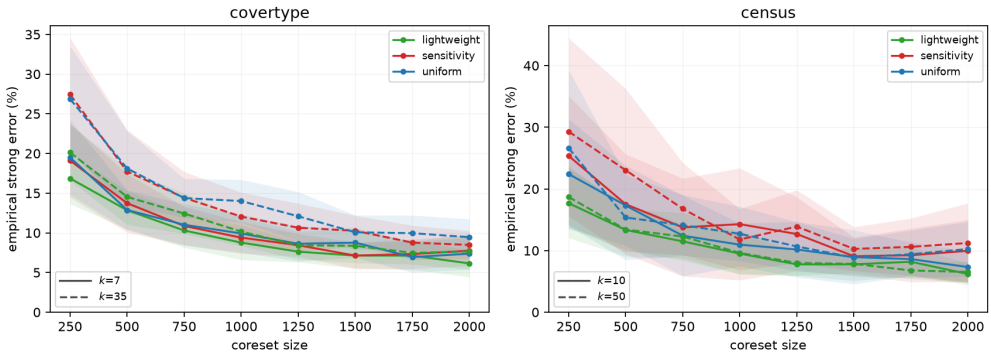


Figure 2: Empirical strong coreset error as a function of coreset size, on Covertype (left) and Census 1990 (right). Color denotes the sampling method and line style the value of k.

Experiment 2: Strong coresets for k-means. We compare uniform sampling against two importance-sampling constructions, Lightweight Coresets (Bachem et al., 2018) and Sensitivity Sampling (Feldman & Langberg, 2011), on the k-means problem over the Census 1990 and Covertype datasets, and report the empirical strong-coreset error (10). We vary the coreset size over $\{200, 250, \ldots, 1000\}$. Each dataset is run at two values of k: Covertype at its seven cover types and at 35, and Census, which carries no label, at 10 and 50.

Figure 2 shows that uniform sampling is comparable to the two importance-sampling methods, even though the assumptions of Theorem 1.6 are not satisfied. As in Experiment 1, there is a dramatic difference in construction time. On the larger value of k, producing a coreset of 1,000 points took 20,000 and 127,000 milliseconds for sensitivity sampling on Covertype and Census respectively, and 90 and 900 milliseconds for lightweight coresets, against 0.05 milliseconds for uniform sampling — at least three orders of magnitude faster and sometimes much more.

AI Disclosure

In this work, we used generative AI tools (Claude, Anthropic; ChatGPT, OpenAI) to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, expanding author-provided ideas and proof sketches into full write-ups, and to implement methods. The AI's mathematical contribution was confined to Appendix E: the main result and overall proof paradigm there are author-derived, but we subsequently used AI to tighten the bound, principally by developing the lifting step in the proof of Lemma E.9. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean and reformat a dataset, or interpret results. Generating synthetic data sets, assisting with translation, and supporting qualitative and thematic data analysis are not

applicable to this work. Additionally, we used generative AI tools to create and edit software code, summarize and analyze existing literature, edit author-written text to improve readability, identify relevant literature and propose a title and keywords for the paper. We have reviewed all AI-assisted work: AI-suggested proof steps were independently re-derived and verified by the authors, AI-implemented code was tested and checked for correctness by the authors, and AI-retrieved literature was checked against original sources. We take responsibility for the final content of this work, including text, claims, code, and mathematical arguments produced with the aid of generative AI.

REFERENCES

Marcel R. Ackermann, Johannes Blömer, and Christian Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6(4):59:1–59:26, 2010. doi:10.1145/1824777.1824779.

Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry*, volume 52 of *MSRI Publications*, chapter 1, pp. 1–30. Cambridge University Press, 2005. URL https://library2.msri.org/books/Book52/contents.html.

Haris Angelidakis, Konstantin Makarychev, and Yury Makarychev. Algorithms for stable and perturbation-resilient problems. In *STOC*, pp. 438–451. ACM, 2017. doi:10.1145/3055399.3055487.

Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a PTAS for k-median and k-means clustering. In *FOCS*, pp. 309–318. IEEE Computer Society, 2010. doi:10.1109/FOCS.2010.36.

Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, 2012. doi:10.1016/j.ipl.2011.10.006.

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1119–1127. Association for Computing Machinery, 2018. doi:10.1145/3219819.3219973.

Daniel N. Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 569–579. PMLR, 2020. URL http://proceedings.mlr.press/v119/baker20a.html.

Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM J. Comput.*, 45(1):102–155, 2016. doi:10.1137/140981575.

Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *J. ACM*, 60(2): 8:1–8:34, 2013. doi:10.1145/2450142.2450144.

Maria-Florina Balcan, Nika Haghtalab, and Colin White. k-center clustering under perturbation resilience. In *ICALP*, volume 55 of *LIPIcs*, pp. 68:1–68:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPIcs.ICALP.2016.68.

Shai Ben-David. A framework for statistical clustering with constant time approximation algorithms for *K*-median and *K*-means clustering. *Mach. Learn.*, 66(2-3):243–257, 2007. doi:10.1007/S10994-006-0587-3.

Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437, 2015. URL http://arxiv.org/abs/1501.00437.

Yonatan Bilu and Nathan Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, 2012. doi:10.1017/S0963548312000193.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering with missing values. In *NeurIPS*, pp. 17360–17372, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/90fd4f88f588ae64038134f1eeaa023f-Abstract.html.

Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In *SODA*, pp. 2679–2696. SIAM, 2021b. doi:10.1137/1.9781611976465.159.

Vladimir Braverman, Vincent Cohen-Addad, Shaofeng H.-C. Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Toftrup, and Xuan Wu. The power of uniform sampling for coresets. In *FOCS*, pp. 462–473. IEEE, 2022. doi:10.1109/FOCS54457.2022.00051.

J. Bretagnolle, D. Dacunha-Castelle, and J. L. Krivine. Lois stables et espaces l_p. In *Symposium on Probability Methods in Analysis*, pp. 48–54. Springer Berlin Heidelberg, 1967.

Mihai Bădoiu and Kenneth L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008. doi:10.1016/j.comgeo.2007.04.002.

Mihai Bădoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *STOC*, pp. 250–257. ACM, 2002. doi:10.1145/509907.509947.

Amir Carmel and Robert Krauthgamer. Stable coresets: Unleashing the power of uniform sampling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sOpAa8iR0A.

Amir Carmel, Chengzhi Guo, Shaofeng H.-C. Jiang, and Robert Krauthgamer. Coresets for 1-center in ℓ_1 metrics. In *ITCS*, volume 325 of *LIPIcs*, pp. 28:1–28:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPICS.ITCS.2025.28.

Amir Carmel, Robert Krauthgamer, and Nir Petruschka. Optimal Stable Coresets for Geometric Median via Uniform Sampling. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2026)*, volume 392 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 30:1–30:14, 2026. doi:10.4230/LIPIcs.APPROX/RANDOM.2026.30.

Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009. doi:10.1137/070699007.

Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *FOCS*, pp. 49–60. IEEE Computer Society, 2017. doi:10.1109/FOCS.2017.14.

Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in Euclidean spaces. In *NeurIPS*, pp. 21085–21098. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/b035d6563a2adac9f822940c145263ce-Abstract.html.

Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *STOC*, pp. 169–182. ACM, 2021b. doi:10.1145/3406325.3451022.

Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. Johnson coverage hypothesis: Inapproximability of k-means and k-median in ℓ_p-metrics. In *SODA*, pp. 1493–1530. SIAM, 2022a. doi:10.1137/1.9781611977073.63.

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *STOC*, pp. 1038–1051. ACM, 2022b. doi:10.1145/3519935.3519946.

Artur Czumaj and Christian Sohler. Sublinear-time approximation algorithms for clustering via random sampling. *Random Struct. Algorithms*, 30(1-2):226–256, 2007. doi:10.1002/RSA.20157.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *STOC*, pp. 569–578. ACM, 2011. doi:10.1145/1993636.1993712.

Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *SoCG*, pp. 11–18. ACM, 2007.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *STOC*, pp. 291–300. ACM, 2004. doi:10.1145/1007352.1007400.

Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In *STOC*, pp. 1416–1429. ACM, 2020. doi:10.1145/3357713.3384296.

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*, pp. 13933–13956. PMLR, 2023. URL https://proceedings.mlr.press/v202/huang23j.html.

Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for Euclidean (k, z)-clustering. In *STOC*, pp. 1594–1604. ACM, 2024. doi:10.1145/3618260.3649707.

Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering (extended abstract). In *SCG*, pp. 332–339. ACM, 1994. doi:10.1145/177424.178042.

Shaofeng H.-C. Jiang, Robert Krauthgamer, Jianing Lou, and Yubo Zhang. Coresets for kernel clustering. *Mach. Learn.*, 113(8):5891–5906, 2024. doi:10.1007/S10994-024-06540-Z.

Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4994–5005. PMLR, 2020. URL http://proceedings.mlr.press/v119/jubran20a.html.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *FOCS*, pp. 299–308. IEEE Computer Society, 2010. doi:10.1109/FOCS.2010.35.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *FOCS*, pp. 454–462. IEEE Computer Society, 2004. doi:10.1109/FOCS.2004.7.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Introduction to coresets: Approximated mean. *CoRR*, abs/2111.03046, 2021. URL https://arxiv.org/abs/2111.03046.

Konstantin Makarychev and Yury Makarychev. Certified algorithms: Worst-case analysis and beyond. In *ITCS*, volume 151 of *LIPIcs*, pp. 49:1–49:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICS.ITCS.2020.49.

Yair Marom and Dan Feldman. k-means clustering of lines for big data. In *NeurIPS*, pp. 12797–12806, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6084e82a08cb979cf75ae28aed37ecd4-Abstract.html.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *ALT*, volume 9925 of *Lecture Notes in Computer Science*, pp. 3–17, 2016. doi:10.1007/978-3-319-46379-7_1.

Chris Meek, Bo Thiesson, and David Heckerman. US Census Data (1990). UCI Machine Learning Repository, 2001. URL https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990). DOI: https://doi.org/10.24432/C5VP42.

Adam Meyerson, Liadan O'Callaghan, and Serge Plotkin. A k-median algorithm with running time independent of data size. *Mach. Learn.*, 56(1–3):61–87, 2004. doi:10.1023/B:MACH.0000033115.78247.f0.

Nina Mishra, Daniel Oblinger, and Leonard Pitt. Sublinear time approximate clustering. In *SODA*, pp. 439–447. ACM/SIAM, 2001. URL http://dl.acm.org/citation.cfm?id=365411.365499.

Dustin G. Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures with k-means. In *2016 IEEE Information Theory Workshop (ITW)*, pp. 211–215, 2016. doi:10.1109/ITW.2016.7606826.

NYC Taxi and Limousine Commission. TLC trip record data. NYC.gov Official Website, 2024. URL https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: March 2024.

Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6), January 2013. ISSN 0004-5411. doi:10.1145/2395116.2395117. URL https://doi.org/10.1145/2395116.2395117.

Milind Prabhu, Chris Schwiegelshohn, and Sudarshan Shyam. Approximation preserving coresets, 2026. URL https://arxiv.org/abs/2606.17316.

Tim Roughgarden (ed.). *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, 2020. doi:10.1017/9781108637435.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4):811–841, 1938. URL http://www.jstor.org/stable/1968466.

Chris Schwiegelshohn and Omar Ali Sheikh-Omar. An empirical evaluation of k-means coresets. In *ESA*, LIPIcs, pp. 84:1–84:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPICS.ESA.2022.84.

Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *FOCS*, pp. 802–813. IEEE Computer Society, 2018. doi:10.1109/FOCS.2018.00081.

J. H. Wells and L. R. Williams. *Embeddings and Extensions in Analysis*, volume 84 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1975. ISBN 978-3-642-66039-9. doi:10.1007/978-3-642-66037-5. URL https://doi.org/10.1007/978-3-642-66037-5.

David P. Woodruff and Taisuke Yasuda. Coresets for multiple ℓ_p regression. In *Forty-first International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research, pp. 53202–53233. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/woodruff24a.html.

A Related work

A large literature studies algorithms for Euclidean (k, z)-clustering under structural assumptions on the input, including perturbation resilience (Bilu & Linial, 2012; Balcan & Liang, 2016; Angelidakis et al., 2017; Balcan et al., 2016; Makarychev & Makarychev, 2020), various notions of separation and stability (Awasthi et al., 2010; 2012; Ostrovsky et al., 2013; Cohen-Addad & Schwiegelshohn, 2017) and more (Kumar & Kannan, 2010; Balcan et al., 2013; Mixon et al., 2016).

Another line of work studies sublinear-time clustering algorithms based on uniform random sampling (Mishra et al., 2001; Meyerson et al., 2004; Czumaj & Sohler, 2007; Ben-David, 2007). These works show that, under various settings and assumptions, approximate solutions for k-median and k-means can be obtained from samples whose size is independent of the input size.

Coresets have proven to be a valuable algorithmic tool for the (k, z)-clustering problem and have been studied extensively for over two decades (Har-Peled & Mazumdar, 2004; Chen, 2009; Feldman & Langberg, 2011; Sohler & Woodruff, 2018; Huang & Vishnoi, 2020; Braverman et al., 2021b; Cohen-Addad et al., 2021b). The dominant notion throughout this literature is that of a strong coreset, and the current state-of-the-art construction is essentially optimal, matching the lower bound up to lower-order terms (Cohen-Addad et al., 2022b; Huang et al., 2024). Strong coreset construction based on uniform sampling was originally proposed by Chen (2009), who suggested partitioning the metric space into rings and sample uniformly from each ring. The analysis was later improved in (Cohen-Addad et al., 2021a; Braverman et al., 2022). Yet, the algorithm still requires a linear scan of the dataset and hence the overall sampling distribution is non-uniform. For kernel clustering Jiang et al. (2024) give strong coresets of size $\mathrm{poly}(k/\epsilon)$, which as a special case yields strong coresets in ℓ_1^d (and every submetric of it).

Another related problem is Euclidean 1-center (the minimum enclosing ball), for which strong coresets require size exponential in the dimension (Agarwal et al., 2005), whereas weak coresets are dimension-free (Bădoiu et al., 2002; Bădoiu & Clarkson, 2008). In the ℓ_1 metric, however, *both* strong and weak coresets for 1-center must depend on the dimension (Carmel et al., 2025).

A related but conceptually different coreset notion, which is also intermediate between weak and strong coresets, is the *approximation-preserving coreset* of (Prabhu et al., 2026), which guarantees that an α-approximate solution on Q can be post-processed to an $\alpha\beta$-approximate solution on P, for every $\alpha, \beta \geq 1$, via an explicit algorithm.

B Proof of Theorem 2.3

Theorem 2.3 (Carmel & Krauthgamer (2026) (Theorem 3.1))**.** *Let $\epsilon \in (0, 1/5)$ and suppose that*

(I1) Q is an ϵ-RCDA of P in $\mathcal{H}$, and
(I2) $\mathrm{cost}(C^*, Q) \leq \alpha \cdot \mathrm{cost}(C^*, P)$ for some $\alpha \geq 1$.

Then Q is a stable $(\epsilon/\alpha,\ 4\epsilon)$-coreset of P.

Proof. For every $C \subseteq \mathcal{H}$ with $|C| = k$, using (I1) and (I2),

$$\operatorname{cost}(C, Q) \leq (\alpha - 1)\operatorname{cost}(C^*, P) + (1 + \epsilon)\operatorname{cost}(C, P) \leq (\alpha + \epsilon)\operatorname{cost}(C, P). \tag{11}$$

For every $C_1, C_2 \subseteq \mathcal{H}$ of size k, we have by the triangle inequality and property (I1),

$$\Big|\big[\operatorname{cost}(C_1, P) - \operatorname{cost}(C_2, P)\big] - \big[\operatorname{cost}(C_1, Q) - \operatorname{cost}(C_2, Q)\big]\Big| \;\leq\; \epsilon(\operatorname{cost}(C_1, P) + \operatorname{cost}(C_2, P)).$$

Now if $\operatorname{cost}(C_1, Q) \leq (1 + \epsilon/\alpha)\operatorname{cost}(C_2, Q)$, then

$$\begin{aligned}
\operatorname{cost}(C_1, P) &\leq \operatorname{cost}(C_2, P) + (\operatorname{cost}(C_1, Q) - \operatorname{cost}(C_2, Q)) + \epsilon(\operatorname{cost}(C_1, P) + \operatorname{cost}(C_2, P)) \\
&\leq \operatorname{cost}(C_2, P) + \frac{\epsilon}{\alpha}\operatorname{cost}(C_2, Q) + \epsilon(\operatorname{cost}(C_1, P) + \operatorname{cost}(C_2, P)) \\
&\leq \operatorname{cost}(C_2, P) + \frac{\epsilon}{\alpha}(\alpha + \epsilon)\operatorname{cost}(C_2, P) + \epsilon(\operatorname{cost}(C_1, P) + \operatorname{cost}(C_2, P)) \\
&= (1 + 2\epsilon + \tfrac{\epsilon^2}{\alpha})\operatorname{cost}(C_2, P) + \epsilon\operatorname{cost}(C_1, P),
\end{aligned}$$

where the third line applies (11) to C_2. By rearranging, $\operatorname{cost}(C_1, P) \leq \frac{1}{1-\epsilon}(1 + 2\epsilon + \frac{\epsilon^2}{\alpha})\operatorname{cost}(C_2, P) \leq (1 + 4\epsilon)\operatorname{cost}(C_2, P)$, where the last inequality uses $\epsilon \leq 1/5$. □

C Proof of Theorem 1.3

Theorem 1.3 (Weak coreset for 1-mean)**.** *Let $\mathcal{H}$ be a Hilbert space and let $P \subseteq \mathcal{H}$ be finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-1}\delta^{-1})$ from P is, with probability at least $1 - \delta$, a weak $(\delta\epsilon, 8\epsilon)$-coreset for the 1-mean problem.*

Proof. Let $c \in \mathcal{H}$ and let Q be a uniform sample of size $s \geq 2/(\epsilon\delta)$. If $\operatorname{cost}(c, Q) \leq (1 + \delta\epsilon/2)V_Q$, then by (7), $\|c - \hat{\mu}\|^2 \leq \delta\epsilon V_Q/2$, and we need to show that $\|c - \mu\|^2 \leq 4\epsilon V$. Observe

$$\|c - \mu\|^2 = \|c - \hat{\mu} + \hat{\mu} - \mu\|^2 \leq 2\|c - \hat{\mu}\|^2 + 2\|\hat{\mu} - \mu\|^2 \leq \delta\epsilon V_Q + 2\|\Delta\|^2.$$

It remains to bound V_Q and $\|\Delta\|^2$. As in Claim 3.2, $\Pr\big[\|\Delta\|^2 > \epsilon V\big] \leq \frac{1}{s\epsilon} \leq \frac{\delta}{2}$. Using Claim 2.1, $\operatorname{cost}(\mu, Q) = \|\Delta\|^2 + V_Q$, and therefore $V_Q \leq \operatorname{cost}(\mu, Q)$. Since $\mathbb{E}[\operatorname{cost}(\mu, Q)] = V$, Markov's inequality gives $\Pr\big[V_Q > \frac{2}{\delta}V\big] \leq \Pr\big[\operatorname{cost}(\mu, Q) > \frac{2}{\delta}V\big] \leq \frac{\delta}{2}$. Conditioned on both events, $\|c - \mu\|^2 \leq \delta\epsilon \cdot \frac{2}{\delta}V + 2 \cdot \epsilon V = 4\epsilon V$, with probability $1 - \delta$. The result is obtained by rescaling ϵ. □

D Lower bounds for $k = 1$

The proofs of the lower bounds use a two-point instance, mass $1/n$ at $D \neq 0$ and the rest at 0, with n tuned to s differently in each proof. We prove the weak bound first, as the stable construction reuses the same instance; its proof also appears in (Cohen-Addad et al., 2021a), and we include it to contrast with the stable guarantee.

Proof of Theorem 1.4 (i). Let $n = \frac{10}{3}s(\epsilon s + 1)$, Let P be such that $1/n$ points are at $D \neq 0$ and the rest are at 0 (w.l.o.g. n divides the size of P). Let X be a random variable denoting the number of sampled points at D in Q, then, $X \sim \operatorname{Bin}(s, p)$ with $p = \frac{1}{n}$. We have,

$$\mu = pD, \quad V = p(1 - p)D^2, \quad \Delta = D(\frac{X}{s} - p). \tag{12}$$

If Q is a weak $(0, \epsilon)$-coreset then $\operatorname{cost}(\hat{\mu}, P) \leq (1 + \epsilon)\operatorname{OPT}$, that is, $V + \|\Delta\|^2 \leq (1 + \epsilon)V$. By (12) it follows that $\big|\frac{X}{s} - p\big| \leq \sqrt{\epsilon p(1 - p)}$. Conditioned on the event $X = 1$, since $n \geq 10s/3$ we have $\big|\frac{X}{s} - p\big| \geq \frac{7}{10s}$, while $\sqrt{\epsilon p(1 - p)} \leq \sqrt{\epsilon p} \leq \sqrt{\frac{3\epsilon}{10s(\epsilon s + 1)}} \leq \frac{\sqrt{3}}{\sqrt{10}s} < \frac{3}{5s}$. Consequently, conditioned on the event $X = 1$, Q is not a weak $(0, \epsilon)$-coreset. As,

$$\delta \geq \Pr\big[X = 1\big] = sp(1 - p)^{s-1} \geq sp(1 - p)^n \geq \frac{1}{3}sp \geq \frac{1}{10\epsilon s + 10}.$$

We get $10\epsilon s\delta + 10\delta \geq 1$, since $\delta \leq 1/20$, it follows that $s \geq \frac{1}{20\epsilon\delta}$. □

The stable lower bound instead controls the cost ratio between two centers symmetric about $\hat{\mu}$, rather than the displacement Δ directly.

Proof of Theorem 1.4 (ii). Let P and X be as in the proof of Item (i) and further denote $t = \sqrt{V + \Delta^2}$.

Let $c_1 = \hat{\mu} + \operatorname{sign}(\Delta)\, t$ and $c_2 = \hat{\mu} - \operatorname{sign}(\Delta)\, t$. Since c_1, c_2 are symmetric around $\hat{\mu}$, Claim 2.1 gives $\operatorname{cost}(c_1, Q) = \operatorname{cost}(c_2, Q)$. Using $\operatorname{cost}(c, P) = V + (c - \mu)^2$, we get

$$\frac{\operatorname{cost}(c_1, P)}{\operatorname{cost}(c_2, P)} = \frac{V + (\Delta + t)^2}{V + (\Delta - t)^2} = \frac{2V + 2\Delta^2 + 2\Delta\sqrt{V + \Delta^2}}{2V + 2\Delta^2 - 2\Delta\sqrt{V + \Delta^2}} = \frac{\sqrt{V + \Delta^2} + |\Delta|}{\sqrt{V + \Delta^2} - |\Delta|}.$$

By setting $x := \frac{|\Delta|}{\sqrt{V}}$, this ratio equals $\frac{\sqrt{1+x^2}+x}{\sqrt{1+x^2}-x}$, which implies $\frac{\operatorname{cost}(c_1,P)}{\operatorname{cost}(c_2,P)} - 1 \geq 2x$.

Case 1: $s < 4/\epsilon^2$**.** Set $n = s$. Conditioned on $X = 0$,

$$x = \frac{|\Delta|}{\sqrt{V}} = \frac{\left|(\frac{X}{s} - p)D\right|}{\sqrt{p(1-p)}\,|D|} \geq \frac{\frac{1}{n}}{\sqrt{\frac{1}{n}}} = \frac{1}{\sqrt{n}} > \frac{\epsilon}{2}.$$

It follows that $\frac{\operatorname{cost}(c_1,P)}{\operatorname{cost}(c_2,P)} > 1 + \epsilon$, and therefore, conditioned on $X = 0$, Q is not a stable $(0, \epsilon)$-coreset.

As,

$$\delta \geq \Pr\big[X = 0\big] = (1-p)^s = (1 - \frac{1}{n})^n \geq \frac{1}{5}$$

Therefore the sample cannot succeed with probability at least $1 - \delta > 4/5$.

Case 2: $s \geq 4/\epsilon^2$**.** Set $n = s^2\epsilon^2$. Conditioned on $X = 1$,

$$x = \frac{|\Delta|}{\sqrt{V}} = \frac{\left|(\frac{X}{s} - p)D\right|}{\sqrt{p(1-p)}\,|D|} \geq \frac{\frac{1}{s} - \frac{1}{n}}{\sqrt{\frac{1}{n}}} = \frac{\sqrt{n}}{s} - \frac{1}{\sqrt{n}} = \epsilon - \frac{1}{s\epsilon} > \frac{\epsilon}{2}.$$

It follows that $\frac{\operatorname{cost}(c_1,P)}{\operatorname{cost}(c_2,P)} > 1 + \epsilon$, and therefore, conditioned on $X = 1$, Q is not a stable $(0, \epsilon)$-coreset.

As, $\delta \geq \Pr\big[X = 1\big] = sp(1-p)^{s-1} \geq \frac{1}{3s\epsilon^2}$. We get, $s \geq \frac{1}{3\epsilon^2\delta}$. □

E UPPER BOUNDS FOR $k > 1$

In this section we establish our coreset guarantees for k-means. We begin in Section E.1 by introducing the structural assumptions of balancedness and separation, and derive several geometric consequences that are used throughout the analysis.

In Section E.2, we parametrize candidate solutions by their *displacement* from the optimal centers. This yields a master identity that splits the difference between the costs on P and on Q into a matched-cost error and a boundary-gain error, reducing the analysis to bounding these two terms separately.

In Section E.3 we define a high-probability event $\mathcal{E}_Q$ on which the sampled cluster weights and sample-mean deviations concentrate, and show that on this event every candidate solution of interest is *localized*: its displacement lies in a fixed set $\mathcal{V}$, determined by the input alone. Localization is used only through this deterministic inclusion; all subsequent probabilistic statements are made uniformly over $\mathcal{V}$ and require no conditioning.

We then analyze the two error terms in Sections E.4 and E.5. We first bound the matched-cost error, and then establish a uniform bound on the boundary-gain error, which forms the core technical component of the analysis.

Finally, the analysis branches according to the desired guarantee. In Section E.6, we combine the matched-cost and boundary-gain bounds with a one-sided estimate on the sampled optimal cost to obtain an RCDA guarantee, which is then converted into a stable coreset. In Section E.7, we instead assume bounded kurtosis to strengthen the boundary-gain bound, leading to the strong coreset guarantee.

E.1 Structural assumptions

We start by defining the balancedness property, which is inherent to the uniform sampling approach.

Definition E.1. The dataset P is β-*balanced* for the k-means problem if there exists an optimal k-center set $C^* = \{\mu_1, \dots, \mu_k\}$ whose induced partition $\{X_1^*, \dots, X_k^*\}$ satisfies $|X_i^*| \geq \beta n/k$ for every $i \in [k]$.

Fix an optimal $C^* = \{\mu_1, \dots, \mu_k\}$ with induced partition $\{X_1^*, \dots, X_k^*\}$ and set $w_i := |X_i^*|/n \geq \beta/k$, $V_i := \frac{1}{|X_i^*|} \sum_{p \in X_i^*} \|p - \mu_i\|^2$, and $\mathrm{OPT} = \mathrm{OPT}(P) = \sum_i w_i V_i$. Since μ_i is the centroid of X_i^*, Claim 2.1 gives, for any c,

$$\frac{1}{|X_i^*|} \sum_{p \in X_i^*} \|p - c\|^2 = \|c - \mu_i\|^2 + V_i. \tag{13}$$

For $p \in P$, let $i(p)$ denote the index with $p \in X_{i(p)}^*$ and $a(p) = \|p - \mu_{i(p)}\|$. For convenience $\mathrm{dist}(p, C) = \min_{1 \leq j \leq k} \|p - c_j\|$.

The separation condition, which originated in (Ostrovsky et al., 2013), requires that the optimal cost drop substantially when passing from $k-1$ to k centers.

Definition E.2 (ρ-separated, (Ostrovsky et al., 2013))**.** $P \subseteq \mathcal{H}$ is ρ-separated for k-means if $\mathrm{OPT}_k \leq \rho^2 \, \mathrm{OPT}_{k-1}$, where $\mathrm{OPT}_k = \mathrm{OPT}$ and OPT_{k-1} is the optimal $(k-1)$-means cost.

The following lemma allows us to bound the distance between optimal centers via the ρ-separated guarantee (similar ideas appeared in (Ostrovsky et al., 2013)).

Lemma E.3. *If $\rho^2 \leq 1/2$ then $D := \min_{i \neq j} \|\mu_i - \mu_j\|$ satisfies $D^2 \geq \mathrm{OPT}/\rho^2$.*

Proof. Merging X_i^*, X_j^* and keeping the other clusters changes the normalized cost by $\frac{w_i w_j}{w_i + w_j} \|\mu_i - \mu_j\|^2$, so by Claim 2.1 applied to the two optimal centers, $\mathrm{OPT}_{k-1} \leq \mathrm{OPT} + \frac{w_i w_j}{w_i + w_j} \|\mu_i - \mu_j\|^2$. Since $\mathrm{OPT}_{k-1} \geq \mathrm{OPT}/\rho^2$, $\mathrm{OPT}(\rho^{-2} - 1) \leq \frac{w_i w_j}{w_i + w_j} \|\mu_i - \mu_j\|^2 \leq \frac{1}{2} \|\mu_i - \mu_j\|^2$. With $\rho^2 \leq 1/2$, $2(\rho^{-2} - 1) \geq \rho^{-2}$, giving $\|\mu_i - \mu_j\|^2 \geq \mathrm{OPT}/\rho^2$. □

Let $F = \{p \in P : a(p) \geq D/4\}$ denote the set of *far* points and $T = \frac{1}{n} \sum_{p \in P} a(p)^2 \mathbf{1}_{\{p \in F\}}$ their cost. Let Q be a uniform i.i.d. sample of size s, drawn with replacement and regarded as a multiset; set $Q_i := Q \cap X_i^*$, $s_i := |Q_i|$, $\hat{w}_i := s_i/s$, $\hat{\mu}_i := \frac{1}{s_i} \sum_{q \in Q_i} q$, $\Delta_i := \hat{\mu}_i - \mu_i$, $\hat{V}_i := \frac{1}{s_i} \sum_{q \in Q_i} \|q - \hat{\mu}_i\|^2$, and $\widehat{\mathrm{OPT}} := \sum_i \hat{w}_i \hat{V}_i$. Applying (13) within each Q_i,

$$\mathrm{cost}(C^*, Q) = \tfrac{1}{s} \textstyle\sum_{q \in Q} a(q)^2 = \widehat{\mathrm{OPT}} + \sum_i \hat{w}_i \|\Delta_i\|^2 \;\geq\; \widehat{\mathrm{OPT}}\,. \tag{14}$$

Note that $\widehat{\mathrm{OPT}}$ is the cost on Q of the fixed partition $\{Q_i\}$ when each Q_i is represented by its empirical centroid $\hat{\mu}_i$; hence it upper-bounds the optimal sample cost, $\mathrm{OPT}(Q) = \min_{|C| = k} \mathrm{cost}(C, Q) \leq \widehat{\mathrm{OPT}}$.

E.2 Displacements and the master identity

Our analysis parametrizes candidate solutions by their offset from the optimal centers. For a *displacement* $v = (v_1, \dots, v_k) \in \mathcal{H}^k$ write $C(v) := \{\mu_1 + v_1, \dots, \mu_k + v_k\}$, and define the pointwise *boundary gain*

$$b_v(p) := \|p - \mu_{i(p)} - v_{i(p)}\|^2 - \min_{1 \leq j \leq k} \|p - \mu_j - v_j\|^2 \;\geq 0, \tag{15}$$

which measures how much a point gains by being served by its nearest center in $C(v)$ rather than by the center displaced from its own optimal center. Set $B_v(P) := \frac{1}{n} \sum_{p \in P} b_v(p)$ and $B_v(Q) := \frac{1}{s} \sum_{q \in Q} b_v(q)$. The *matched costs* are obtained by assigning every point to the center displaced from its optimal center,

$$\mathrm{cost}^{\mathrm{m}}(C(v), P) := \tfrac{1}{n} \textstyle\sum_p \|p - \mu_{i(p)} - v_{i(p)}\|^2, \qquad \mathrm{cost}^{\mathrm{m}}(C(v), Q) := \tfrac{1}{s} \sum_q \|q - \mu_{i(q)} - v_{i(q)}\|^2,$$

so that $B_v(P) = \mathrm{cost}^{\mathrm{m}}(C(v), P) - \mathrm{cost}(C(v), P)$ and $B_v(Q) = \mathrm{cost}^{\mathrm{m}}(C(v), Q) - \mathrm{cost}(C(v), Q)$.

Applying (13) within each X_i^*, and likewise within each Q_i (whose centroid is $\hat{\mu}_i = \mu_i + \Delta_i$), the matched costs evaluate in closed form, and we obtain

$$\begin{aligned} \operatorname{cost}(C(v), P) &= \mathrm{OPT} + \textstyle\sum_i w_i \left\|v_i\right\|^2 - B_v(P), \\ \operatorname{cost}(C(v), Q) &= \widehat{\mathrm{OPT}} + \textstyle\sum_i \hat{w}_i \left\|v_i - \Delta_i\right\|^2 - B_v(Q). \end{aligned} \tag{16}$$

Subtracting the two lines of (16) and expanding $\left\|v_i - \Delta_i\right\|^2$ gives the master identity: for every $v \in \mathcal{H}^k$,

$$\operatorname{cost}(C(v), P) - \operatorname{cost}(C(v), Q) = \big[\operatorname{cost}^{\mathrm{m}}(C(v), P) - \operatorname{cost}^{\mathrm{m}}(C(v), Q)\big] + \big[B_v(Q) - B_v(P)\big], \tag{17}$$

$$\begin{aligned} &\operatorname{cost}^{\mathrm{m}}(C(v), P) - \operatorname{cost}^{\mathrm{m}}(C(v), Q) = \\ &(\mathrm{OPT} - \widehat{\mathrm{OPT}}) + \sum_i (w_i - \hat{w}_i) \left\|v_i\right\|^2 + 2 \sum_i \hat{w}_i \left\langle v_i,\, \Delta_i \right\rangle - \sum_i \hat{w}_i \left\|\Delta_i\right\|^2 . \end{aligned} \tag{18}$$

The first bracket in (17) is the *matched-cost error* and the second is the *boundary-gain error*, arising from points whose nearest center changes under sampling.

Using (14) to write $\operatorname{cost}(C^*, Q) - \widehat{\mathrm{OPT}} = \sum_i \hat{w}_i \left\|\Delta_i\right\|^2$, the last term of (18) cancels against the $\widehat{\mathrm{OPT}}$ term, and the RCDA error of $C(v)$ becomes

$$\begin{aligned} E(v) &:= \big(\operatorname{cost}(C(v), P) - \mathrm{OPT}\big) - \big(\operatorname{cost}(C(v), Q) - \operatorname{cost}(C^*, Q)\big) \\ &= \sum_i (w_i - \hat{w}_i) \left\|v_i\right\|^2 + 2 \sum_i \hat{w}_i \left\langle v_i,\, \Delta_i \right\rangle + \big[B_v(Q) - B_v(P)\big]. \end{aligned} \tag{19}$$

E.3 A GOOD EVENT AND LOCALIZATION

Throughout, $\epsilon \in (0, 1/5)$ and $\delta \in (0, 1)$.

Claim E.4. *Let $\mathcal{E}_Q$ denote the event that*

(E1) $|\hat{w}_i - w_i| \le \epsilon w_i$ *for all* $i \in [k]$,
(E2) $\left\|\Delta_i\right\|^2 \le \epsilon^2 V_i$ *for all* $i \in [k]$.

For a uniform sample Q of (sufficiently large) size $s = O\big(\frac{k^2}{\beta \epsilon^2 \delta}\big)$ it holds that $\Pr[\mathcal{E}_Q] \ge 1 - \delta/3$.

Proof. We bound $\Pr[\neg(\text{E1})] \le \delta/6$ and $\Pr[(\text{E1}) \wedge \neg(\text{E2})] \le \delta/6$. For the former, $s_i \sim \mathrm{Bin}(s, w_i)$ with $w_i \ge \beta/k$, so multiplicative Chernoff and a union bound give $\le 2k \exp(-\epsilon^2 s\beta/(3k)) \le \delta/6$ once $s \ge \frac{3k}{\beta\epsilon^2} \ln(12k/\delta)$. For the latter, on (E1) we have $s_i \ge s(1-\epsilon)\beta/k$; conditionally on $s_1, \dots, s_k$ the set Q_i is an i.i.d. uniform sample of size s_i from X_i^*, so $\mathbb{E}[\left\|\Delta_i\right\|^2 \mid s_i] = V_i/s_i$ and Markov gives per-cluster failure $\le 1/(s_i\epsilon^2)$. A union bound gives $\le k/(s_i\epsilon^2) \le k^2/(s(1-\epsilon)\beta\epsilon^2) \le \delta/6$. □

Fix a constant $\gamma \ge 1$ and set $\mathcal{S}_\gamma(Q) = \{C : \operatorname{cost}(C, Q) \le \gamma\, \widehat{\mathrm{OPT}}\}$, $\mathcal{S}_\gamma(P) = \{C : \operatorname{cost}(C, P) \le \gamma\, \mathrm{OPT}\}$. Since $\mathrm{OPT}(Q) \le \widehat{\mathrm{OPT}}$, we have $\{C : \operatorname{cost}(C, Q) \le \gamma\, \mathrm{OPT}(Q)\} \subseteq \mathcal{S}_\gamma(Q)$, so proving a guarantee uniformly over $\mathcal{S}_\gamma(Q)$ yields it over the $\mathrm{OPT}(Q)$-class as well.

The next lemma shows that every candidate of interest is pinned near the optimal centers. It is stated for a generic a-priori bound $\operatorname{cost}(C^*, Q) \le \lambda\, \mathrm{OPT}$, where $\lambda \ge 1$ will be determined later as it depends on the exact application; note that it implies $\widehat{\mathrm{OPT}} \le \lambda\, \mathrm{OPT}$ by (14). Throughout we write

$$\Gamma_\gamma^2 := \frac{6\gamma\lambda k}{\beta}\, \mathrm{OPT}, \qquad K := 11\gamma\lambda\, \mathrm{OPT}, \tag{20}$$

$$\mathcal{V} := \Big\{ v = (v_1, \dots, v_k) \in \mathcal{H}^k : \forall i \in [k]\ \ \left\|v_i\right\| \le \Gamma_\gamma, \text{ and } \textstyle\sum_i w_i \left\|v_i\right\|^2 \le K \Big\}. \tag{21}$$

We stress that Γ_γ, K, and hence the set $\mathcal{V}$, are determined by P, k, β, γ and λ alone, and in particular do not depend on the sample Q.

We start with a geometric claim bounding the boundary gain.

Claim E.5. *Assume $\Gamma_\gamma \le D/4$. Then for every v with $\|v_i\| \le \Gamma_\gamma$ for all i, and every $p \in P$, if $b_v(p) > 0$ then $p \in F$, and $0 \le b_v(p) \le 4\Gamma_\gamma\, a(p) \le 4a(p)^2$.*

Proof. Let $i = i(p)$ and let j attain $\min_l \|p - \mu_l - v_l\|$. If $b_v(p) > 0$ then $j \neq i$, and $\|p-\mu_j\| - \Gamma_\gamma \le \|p - \mu_j - v_j\| \le \|p - \mu_i - v_i\| \le a(p) + \Gamma_\gamma$, so $\|p-\mu_j\| \le a(p) + 2\Gamma_\gamma$ and $D \le \|\mu_i - \mu_j\| \le 2a(p) + 2\Gamma_\gamma$. Hence $a(p) \ge D/2 - \Gamma_\gamma \ge D/4$ (using $\Gamma_\gamma \le D/4$), i.e. $p \in F$ and $a(p) \ge \Gamma_\gamma$. Writing $a = a(p)$ and $a_2 = \min_{l \neq i} \|p - \mu_l\| \ge a$ (as $p \in X_i^*$),

$$b_v(p) \le \|p - \mu_i - v_i\|^2 - \|p - \mu_j - v_j\|^2 \le (a + \Gamma_\gamma)^2 - (a_2 - \Gamma_\gamma)^2 \le (a+\Gamma_\gamma)^2 - (a - \Gamma_\gamma)^2 = 4\Gamma_\gamma a \le 4a^2,$$

as $a_2 \ge a \ge \Gamma_\gamma$. □

In particular, averaging Claim E.5 over P and over Q gives

$$B_v(P) \le 4\,\mathrm{OPT}, \qquad B_v(Q) \le 4\,\mathrm{cost}(C^*, Q) \le 4\lambda\,\mathrm{OPT}\,. \tag{22}$$

We next prove the localization lemma.

Lemma E.6. *Suppose $\mathrm{cost}(C^*, Q) \le \lambda\,\mathrm{OPT}$ and $\Gamma_\gamma \le D/4$. Then, conditioned on $\mathcal{E}_Q$, every $C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$ satisfies $C = C(v)$ for some $v \in \mathcal{V}$.*

Proof. Fix $i \in [k]$ and let $R_i = \mathrm{dist}(\mu_i, C)$ and $c(p) \in C$ be a nearest center to p. First we show that for every $C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$, for all $i \in [k]$, $R_i \le \Gamma_\gamma$. For every $p \in X_i^*$ we have,

$$R_i^2 \le \|\mu_i - c(p)\|^2 \le 2\,\|p - \mu_i\|^2 + 2\,\|p - c(p)\|^2 = 2a(p)^2 + 2\,\mathrm{dist}^2(p, C).$$

Averaging over $p \in X_i^*$ and using $\frac{1}{|X_i^*|}\sum_{p \in X_i^*} \mathrm{dist}^2(p,C) \le \frac{1}{w_i}\,\mathrm{cost}(C,P)$ gives

$$R_i^2 \;\le\; 2V_i + \frac{2}{w_i}\,\mathrm{cost}(C, P). \tag{23}$$

If $C \in \mathcal{S}_\gamma(P)$, using $V_i \le \mathrm{OPT}\,/w_i$, $\mathrm{cost}(C,P) \le \gamma\,\mathrm{OPT}$, $\gamma \ge 1$ and $w_i \ge \beta/k$,

$$R_i^2 \le \frac{2(1+\gamma)\,\mathrm{OPT}}{w_i} \le \frac{4\gamma k}{\beta}\,\mathrm{OPT} \le \Gamma_\gamma^2.$$

Else, if $C \in \mathcal{S}_\gamma(Q)$ the same averaging over Q_i, with $\hat{w}_i, \hat{V}_i, \hat{\mu}_i$ in place of w_i, V_i, μ_i, gives $\widehat{R}_i^2 \le 2\hat{V}_i + \frac{2}{\hat{w}_i}\,\mathrm{cost}(C,Q)$ for $\widehat{R}_i = \mathrm{dist}(\hat{\mu}_i, C)$. Using $\hat{V}_i \le \widehat{\mathrm{OPT}}\,/\hat{w}_i$, $\mathrm{cost}(C,Q) \le \gamma\,\widehat{\mathrm{OPT}}$, $\widehat{\mathrm{OPT}} \le \lambda\,\mathrm{OPT}$, and $\hat{w}_i \ge (1-\epsilon)\beta/k \ge \frac{4\beta}{5k}$ by (E1),

$$\widehat{R}_i^2 \le \frac{2(1+\gamma)\lambda\,\mathrm{OPT}}{\hat{w}_i} \le \frac{5\gamma\lambda k}{\beta}\,\mathrm{OPT}\,.$$

By (E2), $\|\Delta_i\| \le \epsilon\sqrt{V_i} \le \frac{1}{5}\sqrt{k\,\mathrm{OPT}\,/\beta}$, so, by the triangle inequality

$$R_i^2 \le (\widehat{R}_i + \|\Delta_i\|)^2 \le (\sqrt{5\gamma\lambda} + \tfrac{1}{5})^2 \cdot \frac{k\,\mathrm{OPT}}{\beta} \le \frac{6\gamma\lambda k}{\beta}\,\mathrm{OPT} = \Gamma_\gamma^2,$$

where the last inequality follows from $\gamma\lambda \ge 1$.

Second, we show that for every $C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$, $\sum_i w_i\,\|v_i\|^2 \le K$. As $\Gamma_\gamma \le D/4$, the balls $\{c : \|c - \mu_i\| \le \Gamma_\gamma\}$, $i \in [k]$, are pairwise disjoint; each contains at least one center of C by the above, and $|C| = k$, so each contains exactly one. Letting $\sigma(i)$ be the index of the unique center of C inside the i-th ball and $v_i := c_{\sigma(i)} - \mu_i$ yields $C = C(v)$ with $\|v_i\| \le \Gamma_\gamma$ for all i.

For $C \in \mathcal{S}_\gamma(P)$, $\sum_i w_i\,\|v_i\|^2 = \mathrm{cost}(C(v), P) - \mathrm{OPT} + B_v(P) \le (\gamma+3)\,\mathrm{OPT} \le K$, using (16) for the first equality and (22) for the second inequality.

For $C \in \mathcal{S}_\gamma(Q)$, the second line of (16) together with $\mathrm{cost}(C,Q) \le \gamma\,\widehat{\mathrm{OPT}}$ and (22) gives

$$\sum_i \hat{w}_i\,\|v_i - \Delta_i\|^2 \le (\gamma - 1)\,\widehat{\mathrm{OPT}} + 4\lambda\,\mathrm{OPT} \le (\gamma+3)\lambda\,\mathrm{OPT}\,.$$

Hence, by $\|v_i\|^2 \le 2\|v_i - \Delta_i\|^2 + 2\|\Delta_i\|^2$, (E2) and $\epsilon \le 1/5$,

$$\textstyle\sum_i \hat{w}_i \|v_i\|^2 \le 2(\gamma+3)\lambda \,\mathrm{OPT} + 2(1+\epsilon)\epsilon^2 \,\mathrm{OPT} \le 2(\gamma+3)\lambda \,\mathrm{OPT} + \frac{1}{10}\,\mathrm{OPT},$$

and by (E1), $w_i \le \hat{w}_i/(1-\epsilon) \le \frac{5}{4}\hat{w}_i$, so $\sum_i w_i \|v_i\|^2 \le \frac{5}{2}(\gamma+3)\lambda \,\mathrm{OPT} + \frac{1}{8}\,\mathrm{OPT} \le 11\gamma\lambda \,\mathrm{OPT} = K$, using $\gamma + 3 \le 4\gamma$. □

E.4 Matched-cost error

We next bound the terms of (18), uniformly over $\mathcal{V}$.

Lemma E.7. *Assume $\Gamma_\gamma \le D/4$ and condition on $\mathcal{E}_Q$. Then for every $v \in \mathcal{V}$:*

(i) $\sum_i w_i \|v_i\|^2 \le 4\,\mathrm{cost}(C(v), P)$, *and,* $\left|\sum_i (\hat{w}_i - w_i)\|v_i\|^2\right| \le 4\epsilon\,\mathrm{cost}(C(v), P)$,
(ii) $\left|2\sum_i \hat{w}_i \langle v_i, \Delta_i\rangle\right| \le 5(1+\epsilon)\epsilon\,\mathrm{cost}(C(v), P)$,
(iii) $\left|\sum_i \hat{w}_i \|\Delta_i\|^2\right| \le (1+\epsilon)\epsilon^2\,\mathrm{OPT}$.

Proof. (i): by (16) and (22), $\sum_i w_i\|v_i\|^2 = \mathrm{cost}(C(v),P) + B_v(P) - \mathrm{OPT} \le \mathrm{cost}(C(v),P) + 3\,\mathrm{OPT} \le 4\,\mathrm{cost}(C(v),P)$, using $\mathrm{OPT} \le \mathrm{cost}(C(v),P)$. Then, by (E1), $\left|\sum_i(\hat{w}_i - w_i)\|v_i\|^2\right| \le \epsilon\sum_i w_i\|v_i\|^2 \le 4\epsilon\,\mathrm{cost}(C(v),P)$. (ii): by Cauchy–Schwarz and Young inequalities, $\left|2\sum_i \hat{w}_i\langle v_i, \Delta_i\rangle\right| \le \sum_i \hat{w}_i(\epsilon\|v_i\|^2 + \epsilon^{-1}\|\Delta_i\|^2)$. By (E1), the first part is bounded by $\epsilon(1+\epsilon)\sum_i w_i\|v_i\|^2 \le 4\epsilon(1+\epsilon)\,\mathrm{cost}(C(v),P)$. By (E1),(E2), the second part is at most $\epsilon^{-1}(1+\epsilon)\sum_i w_i\epsilon^2 V_i = (1+\epsilon)\epsilon\,\mathrm{OPT} \le (1+\epsilon)\epsilon\,\mathrm{cost}(C(v),P)$, so their sum is at most $5(1+\epsilon)\epsilon\,\mathrm{cost}(C(v),P)$. (iii): by (E1),(E2), $\left|\sum_i \hat{w}_i\|\Delta_i\|^2\right| \le (1+\epsilon)\sum_i w_i\epsilon^2 V_i = (1+\epsilon)\epsilon^2\,\mathrm{OPT}$. □

E.5 Boundary-gain error

We now bound $|B_v(Q) - B_v(P)|$ uniformly over $v \in \mathcal{V}$, which is the main technical difficulty of the proof. The obstacle is that $b_v(p)$ is nonlinear in the displacements v: it is a maximum of functions that are affine in v except for a quadratic term $\|v_i\|^2 - \|v_j\|^2$. The proof handles the two features separately. Maurer's vector contraction inequality removes the maximum, and a one-coordinate lifting turns the quadratic term into a linear one, after which the whole process is a single inner product in a Hilbert space and Cauchy–Schwarz finishes the argument.

Theorem E.8 (Maurer's vector contraction inequality (Maurer, 2016, Theorem 2))**.** *Let $\mathcal{S}$ be a countable set and let $\psi_t : \mathcal{S} \to \mathbb{R}$, $\phi_t : \mathcal{S} \to \ell_2$ $(1 \le t \le m)$ satisfy $\psi_t(\sigma) - \psi_t(\sigma') \le \|\phi_t(\sigma) - \phi_t(\sigma')\|$ for all $\sigma, \sigma' \in \mathcal{S}$. Let ξ_t and $\eta_{t,j}$ be independent Rademacher variables. Then*

$$\mathbb{E}_\xi \sup_{\sigma\in\mathcal{S}} \sum_t \xi_t\,\psi_t(\sigma) \;\le\; \sqrt{2}\,\mathbb{E}_\eta \sup_{\sigma\in\mathcal{S}} \sum_t \sum_j \eta_{t,j}\,\phi_t(\sigma)_j,$$

where $\phi_t(\sigma)_j$ denotes the j-th coordinate of $\phi_t(\sigma)$.

Lemma E.9. *Assume $\Gamma_\gamma \le D/4$ and let $Z := \sup_{v\in\mathcal{V}}\left|B_v(Q) - B_v(P)\right|$. Then*

$$\mathbb{E}_Q\, Z \;\le\; 63\sqrt{k}\,\Gamma_\gamma\sqrt{T/s}.$$

Proof. Fix $p \in X_i^*$. By Claim E.5, $b_v(p) > 0$ only when $p \in F$, so assume $p \in F$. With $A(p) = \{j : \|p - \mu_j\| \le a(p) + 2\Gamma_\gamma\}$ we may write

$$b_v(p) = \max_{j\in A(p)} \big(r_j(p) + h_j(p, v)\big), \tag{24}$$

where $r_j(p) = \|p-\mu_i\|^2 - \|p-\mu_j\|^2 \le 0$ and $h_j(p,v) = -2\langle p - \mu_i, v_i\rangle + 2\langle p - \mu_j, v_j\rangle + \|v_i\|^2 - \|v_j\|^2$. The restriction to $A(p)$ is valid since for $j \notin A(p)$ the argument of Claim E.5 gives $r_j + h_j < 0$, while $j = i \in A(p)$ attains 0.

Symmetrization. Let $Q' = (q'_1, \ldots, q'_s)$ be an independent i.i.d. copy and ξ_t Rademacher. Since $B_v(P) = \mathbb{E}_{Q'}[\frac{1}{s}\sum_t b_v(q'_t)]$ for every fixed v, and the index set $\mathcal{V}$ depends on neither Q nor Q', Jensen applied to $\sup_{v\in\mathcal{V}}$

and exchangeability of (q_t, q'_t) give

$$\mathbb{E}_Q Z = \frac{1}{s}\mathbb{E}_Q \sup_{v\in\mathcal{V}}\Big|\mathbb{E}_{Q'}\big[\textstyle\sum_t \big(b_v(q_t) - b_v(q'_t)\big)\big]\Big| \;\le\; \tfrac{1}{s}\mathbb{E}_{Q,Q'} \sup_{v\in\mathcal{V}}\Big|\textstyle\sum_t \big(b_v(q_t) - b_v(q'_t)\big)\Big|$$
$$= \frac{1}{s}\mathbb{E}_{Q,Q',\xi} \sup_{v\in\mathcal{V}}\Big|\sum_t \xi_t\big(b_v(q_t) - b_v(q'_t)\big)\Big| \;\le\; \frac{2}{s}\,\mathbb{E}_{Q,\xi} \sup_{v\in\mathcal{V}}\Big|\sum_t \xi_t\, b_v(q_t)\Big|,$$

the sum running over far points $q_t \in Q \cap F$ only, as b_v vanishes off F.

Contraction. Fix Q with far points $\{q_1, \dots, q_m\} = Q \cap F$, and write $a_t = a(q_t) \ge D/4$, $i_t = i(q_t)$. By (24), $v \mapsto b_v(q_t)$ is 1-Lipschitz with respect to the ℓ_∞, hence the ℓ_2, norm of the vector $\phi_t(v) = (h_j(q_t, v))_{j\in A(q_t)}$. Since $0 \in \mathcal{V}$ and $b_0 \equiv 0$ we have $\sup_{v\in\mathcal{V}} \sum_t \zeta_t b_v(q_t) \ge 0$ for every sign pattern ζ, so writing $|x| = \max(x, -x)$ and using the symmetry of ξ, Theorem E.8 gives

$$\mathbb{E}_\xi \sup_{v\in\mathcal{V}}\Big|\sum_t \xi_t b_v(q_t)\Big| \le 2\sqrt{2}\,\mathbb{E}_\eta \sup_{v\in\mathcal{V}} \sum_t \sum_{j\in A(q_t)} \eta_{t,j}\, h_j(q_t, v). \tag{25}$$

Lifting. The functions $h_j(p, \cdot)$ are not linear in v, but they all become linear after appending one coordinate to each displacement. Let $\mathcal{G} := \bigoplus_{r=1}^k (\mathcal{H} \oplus \mathbb{R})$ with the direct-sum inner product, and define the *lifted displacement*

$$u(v) := \big(u_1(v), \dots, u_k(v)\big) \in \mathcal{G}, \qquad u_r(v) := \Big(v_r,\ \frac{\|v_r\|^2}{2\Gamma_\gamma}\Big).$$

For $p \in X_i^*$ and $j \in [k]$, let $g_{p,j} \in \mathcal{G}$ be the vector whose only nonzero blocks are

$$(g_{p,j})_i = \big(-2(p - \mu_i),\ 2\Gamma_\gamma\big), \qquad (g_{p,j})_j = \big(2(p - \mu_j),\ -2\Gamma_\gamma\big);$$

for $j = i$ these two expressions cancel and $g_{p,i} = 0$, consistently with $h_i(p, \cdot) \equiv 0$. Expanding the inner products gives the identity

$$h_j(p, v) = \langle g_{p,j},\, u(v)\rangle_{\mathcal{G}} \qquad \text{for all } j \in [k],\ v \in \mathcal{H}^k. \tag{26}$$

Consequently the Rademacher process in (25) is a single inner product.

Setting $W := \sum_t \sum_{j\in A(q_t)} \eta_{t,j}\, g_{q_t,j} \in \mathcal{G}$,

$$\sup_{v\in\mathcal{V}} \sum_t \sum_{j\in A(q_t)} \eta_{t,j} h_j(q_t, v) = \sup_{v\in\mathcal{V}} \langle W,\, u(v)\rangle \;\le\; \|W\| \cdot U, \qquad U := \sup_{v\in\mathcal{V}} \|u(v)\|. \tag{27}$$

Bounding U. Since $\|v_r\| \le \Gamma_\gamma$ we have $\|u_r(v)\|^2 = \|v_r\|^2 + \frac{\|v_r\|^4}{4\Gamma_\gamma^2} \le \frac{5}{4}\|v_r\|^2$, and by the aggregate constraint in (21) together with $w_r \ge \beta/k$,

$$\sum_r \|v_r\|^2 \le \frac{k}{\beta}\sum_r w_r \|v_r\|^2 \le \frac{kK}{\beta} = \frac{11\gamma\lambda k}{\beta}\,\mathrm{OPT} = \frac{11}{6}\,\Gamma_\gamma^2.$$

Hence $U^2 \le \frac{5}{4}\cdot\frac{11}{6}\Gamma_\gamma^2 = \frac{55}{24}\Gamma_\gamma^2$.

Bounding $\|W\|$. For $j \in A(q_t)$ we have $\|q_t - \mu_j\| \le a_t + 2\Gamma_\gamma \le 3a_t$ and $\Gamma_\gamma \le D/4 \le a_t$, so

$$\|g_{q_t,j}\|^2 = 4\,\|q_t - \mu_{i_t}\|^2 + 4\Gamma_\gamma^2 + 4\,\|q_t - \mu_j\|^2 + 4\Gamma_\gamma^2 \le 4a_t^2 + 4a_t^2 + 36a_t^2 + 4a_t^2 = 48\,a_t^2.$$

The $\eta_{t,j}$ are independent with zero mean, so $\mathbb{E}_\eta\,\|W\|^2 = \sum_t \sum_{j\in A(q_t)} \|g_{q_t,j}\|^2 \le 48\sum_t |A(q_t)|\, a_t^2 \le 48k\sum_t a_t^2$, and by Jensen $\mathbb{E}_\eta\,\|W\| \le (48k\sum_t a_t^2)^{1/2}$. Combining with (27) and $\sqrt{55/24}\cdot\sqrt{48} = \sqrt{110} < 11$,

$$\mathbb{E}_\eta \sup_{v\in\mathcal{V}} \sum_t \sum_{j\in A(q_t)} \eta_{t,j} h_j(q_t, v) \;\le\; \sqrt{110\,k}\,\Gamma_\gamma\Big(\textstyle\sum_t a_t^2\Big)^{1/2} \;\le\; 11\sqrt{k}\,\Gamma_\gamma\Big(\sum_{q\in Q\cap F} a(q)^2\Big)^{1/2}.$$

Finally, combining with (25) and using Jensen (concavity of $\sqrt{\cdot}$) together with $\mathbb{E}_Q \sum_{q\in Q\cap F} a(q)^2 = sT$,

$$\mathbb{E}_Q Z \le \frac{2}{s}\cdot 2\sqrt{2}\cdot 11\sqrt{k}\,\Gamma_\gamma (sT)^{1/2} \le 63\,\sqrt{k}\,\Gamma_\gamma\sqrt{T/s}. \qquad \square$$

E.6 Stable coreset

We further augment $\mathcal{E}_Q$ with a one-sided control on the sampled optimal cost, i.e. the event (E3): $\operatorname{cost}(C^*, Q) \le (6/\delta)\operatorname{OPT}$. Since $\mathbb{E}[\operatorname{cost}(C^*, Q)] = \operatorname{OPT}$, Markov gives $\Pr[\neg\text{(E3)}] \le \delta/6$. In this section we write $\mathcal{E}_Q$ for (E1) $\wedge$ (E2) $\wedge$ (E3), so that Claim E.4 gives $\Pr[\mathcal{E}_Q] \ge 1 - \delta/2$.

We instantiate (20) with $\lambda = 6/\delta$, so $\Gamma_\gamma^2 = 36\gamma k \operatorname{OPT}/(\beta\delta)$. Assume $\rho^2 \le \beta\delta/(576\gamma k)$; then $\Gamma_\gamma^2 \le \operatorname{OPT}/(16\rho^2) \le D^2/16$ by Lemma E.3, so $\Gamma_\gamma \le D/4$ and Claim E.5 and Lemmas E.6, E.7 and E.9 all apply.

Lemma E.10. *Let $\rho^2 \le \beta\delta/(576\gamma k)$. Then for a uniform sample of (sufficiently large) size $O\big(k^2\gamma/(\beta\epsilon^2\delta^2)\big)$, with probability $\ge 1 - \delta/2$, $\sup_{v\in\mathcal{V}} |B_v(Q) - B_v(P)| \le \epsilon \operatorname{OPT}$.*

Proof. By Lemma E.9 and the trivial bound $T \le \frac{1}{n}\sum_p a(p)^2 = \operatorname{OPT}$,

$$\mathbb{E}_Q Z \le 63\sqrt{k}\,\Gamma_\gamma \sqrt{\operatorname{OPT}/s}, \qquad (\mathbb{E}_Q Z)^2 \le 3969\, k\Gamma_\gamma^2 \operatorname{OPT}/s.$$

For the variance we apply Efron–Stein to Z, which is a function of the i.i.d. coordinates $q_1, \dots, q_s$ only, the class $\mathcal{V}$ being fixed. By Claim E.5, $0 \le b_v(p) \le 4\Gamma_\gamma\, a(p)$ for every $v \in \mathcal{V}$, so replacing q_t by an independent copy q_t' changes $B_v(Q)$ by at most $\frac{4\Gamma_\gamma}{s}(a(q_t) + a(q_t'))$ *uniformly in* $v \in \mathcal{V}$, and hence changes the supremum Z by at most the same amount. By Efron–Stein and $\mathbb{E}[a^2] = \operatorname{OPT}$,

$$\operatorname{Var}(Z) \le \frac{1}{2}\sum_t \mathbb{E}\big[(Z - Z^{(t)})^2\big] \le \frac{8\Gamma_\gamma^2}{s^2}\sum_t \mathbb{E}[(a(q_t) + a(q_t'))^2] \le \frac{32\Gamma_\gamma^2 \operatorname{OPT}}{s}.$$

Hence $\mathbb{E}[Z^2] \le 4001\, k\Gamma_\gamma^2 \operatorname{OPT}/s$, and by Markov $\Pr[Z > \epsilon \operatorname{OPT}] \le \frac{4001 k\Gamma_\gamma^2}{s\epsilon^2 \operatorname{OPT}} \le \delta/2$ once $s \ge 8002\, k\Gamma_\gamma^2/(\epsilon^2 \operatorname{OPT}\delta) = \Theta\big(k^2\gamma/(\beta\epsilon^2\delta^2)\big)$. □

Lemma E.11. *On $\mathcal{E}_Q$ and the event of Lemma E.10, for every $C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$,*

$$\Big|(\operatorname{cost}(C, P) - \operatorname{OPT}) - (\operatorname{cost}(C, Q) - \operatorname{cost}(C^*, Q))\Big| \le 11\epsilon \operatorname{cost}(C, P).$$

That is, Q is an 11ϵ-RCDA of P over $\mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$.

Proof. On $\mathcal{E}_Q$, Lemma E.6 gives $C = C(v)$ for some $v \in \mathcal{V}$, so it suffices to bound $|E(v)|$ from (19) uniformly over $v \in \mathcal{V}$. Bounding the first two terms of (19) by items (i) and (ii) of Lemma E.7, and the third by Lemma E.10 together with $\operatorname{OPT} \le \operatorname{cost}(C(v), P)$, we get, using $\epsilon \le 1/5$,

$$|E(v)| \le 4\epsilon \operatorname{cost}(C(v), P) + 5(1+\epsilon)\epsilon \operatorname{cost}(C(v), P) + \epsilon \operatorname{cost}(C(v), P) \le 11\epsilon \operatorname{cost}(C(v), P). \qquad \square$$

Proof of Theorem 1.5. By a union bound, $\mathcal{E}_Q$ and the event of Lemma E.10 occur simultaneously with probability at least $1 - \delta$. On $\mathcal{E}_Q$ we have $\operatorname{cost}(C^*, Q) \le 6\operatorname{cost}(C^*, P)/\delta$, so combining Lemma E.11 with Theorem 2.3 gives the claim; we use the restricted-class version of Theorem 2.3, whose proof is identical after restricting every candidate center set appearing in it to $\mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$. Conclude by rescaling ϵ. □

E.7 Strong coreset

Fix $\epsilon \in (0, 1/5)$. To obtain a strong coreset guarantee over $\mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$ we require the additional assumption of bounded kurtosis.

Definition E.12. $P \subseteq \mathcal{H}$ has M-bounded kurtosis if for every i, $\frac{1}{|X_i^*|}\sum_{p\in X_i^*} \|p - \mu_i\|^4 \le M V_i^2$.

Claim E.13. *Under M-bounded kurtosis, $\frac{1}{n}\sum_{p\in P} a(p)^4 \le \frac{Mk}{\beta}\operatorname{OPT}^2$.*

Proof. $\frac{1}{n}\sum_p a(p)^4 = \sum_i w_i \frac{1}{|X_i^*|}\sum_{p\in X_i^*} \|p - \mu_i\|^4 \le M\sum_i w_i V_i^2$, and $V_i \le \operatorname{OPT}/w_i \le k\operatorname{OPT}/\beta$ gives $\sum_i w_i V_i^2 \le \frac{k}{\beta}\operatorname{OPT}^2$. □

We then bound the cost of the far points.

Claim E.14. $T \le 16\rho^2 \frac{Mk}{\beta}\,\mathrm{OPT}$.

Proof. $\mathbf{1}_{\{p\in F\}} \le (4a(p)/D)^2$ gives $T \le \frac{16}{D^2}\cdot\frac{1}{n}\sum_p a(p)^4$. Proof is concluded by applying Claim E.13 and $D^2 \ge \mathrm{OPT}/\rho^2$ (Lemma E.3). □

Consider the event ($\bar{\mathrm{E}}3$): $\left|\mathrm{cost}(C^*,Q) - \mathrm{OPT}\right| \le \epsilon\,\mathrm{OPT}$.

Claim E.15. *Under M-bounded kurtosis, if $s \ge 6Mk/(\beta\epsilon^2\delta)$ then $\Pr[\neg(\bar{E}3)] \le \delta/6$.*

Proof. $\mathbb{E}[\mathrm{cost}(C^*,Q)] = \mathrm{OPT}$ and, by Claim E.13, $\mathrm{Var}(\mathrm{cost}(C^*,Q)) \le \mathbb{E}[a^4]/s \le Mk\,\mathrm{OPT}^2/(\beta s)$. Using Chebyshev's inequality the probability of failure is at most $\frac{Mk}{s\beta\epsilon^2} \le \delta/6$. □

Augment $\mathcal{E}_Q$ with ($\bar{\mathrm{E}}3$), then $\Pr[\mathcal{E}_Q] \ge 1-\delta/2$.

Corollary E.16. *On $\mathcal{E}_Q$, $\left|\widehat{\mathrm{OPT}} - \mathrm{OPT}\right| \le 2\epsilon\,\mathrm{OPT}$.*

Proof. Using (14), $\mathrm{cost}(C^*,Q) = \widehat{\mathrm{OPT}} + \sum_i \hat{w}_i\,\|\Delta_i\|^2$. For the first direction, $\widehat{\mathrm{OPT}} \le \mathrm{cost}(C^*,Q) \le (1+\epsilon)\,\mathrm{OPT}$. For the second direction,

$$\widehat{\mathrm{OPT}} = \mathrm{cost}(C^*,Q) - \sum_i \hat{w}_i\,\|\Delta_i\|^2 \ge (1-\epsilon)\,\mathrm{OPT} - \sum_i (1+\epsilon) w_i \epsilon^2 V_i \ge (1-2\epsilon)\,\mathrm{OPT}\,. \qquad \square$$

On $\mathcal{E}_Q$ we have $\mathrm{cost}(C^*,Q) \le (1+\epsilon)\,\mathrm{OPT}$, so we instantiate (20) with $\lambda = 1.2$, giving $\Gamma_\gamma^2 = 7.2\gamma k\,\mathrm{OPT}/\beta$. Assume $\rho^2 \le \beta/(120\gamma k)$; then $\Gamma_\gamma^2 \le \mathrm{OPT}/(16\rho^2) \le D^2/16$, so $\Gamma_\gamma \le D/4$ and Claim E.5 and Lemmas E.6, E.7 and E.9 apply. We first bound the matched cost.

Lemma E.17. *On $\mathcal{E}_Q$, for every $v \in \mathcal{V}$, $\left|\mathrm{cost}^{\mathrm{m}}(C(v),P) - \mathrm{cost}^{\mathrm{m}}(C(v),Q)\right| \le 14\epsilon\,\mathrm{cost}(C(v),P)$.*

Proof. Bound the four terms of (18): the first by Corollary E.16, the remaining three by items (i), (ii), (iii) of Lemma E.7. Together with $\mathrm{OPT} \le \mathrm{cost}(C(v),P)$ this gives

$$\begin{aligned}\left|\mathrm{cost}^{\mathrm{m}}(C(v),P) - \mathrm{cost}^{\mathrm{m}}(C(v),Q)\right| &\le 2\epsilon\,\mathrm{OPT} + 4\epsilon\,\mathrm{cost}(C(v),P) + 5(1+\epsilon)\epsilon\,\mathrm{cost}(C(v),P) + (1+\epsilon)\epsilon^2\,\mathrm{OPT}\\ &\le 14\epsilon\,\mathrm{cost}(C(v),P),\end{aligned}$$

the last step using $\epsilon \le 1/5$. □

We next bound the boundary-gain error.

Lemma E.18. *Under M-bounded kurtosis and $\rho^2 \le \beta/(120\gamma k)$, for a sample of (sufficiently large) size $O(k^2M/(\beta\epsilon^2\delta))$, with probability $\ge 1-\delta/2$,*

$$\sup_{v\in\mathcal{V}} |B_v(Q) - B_v(P)| \le \epsilon\,\mathrm{OPT}\,.$$

Proof. By Lemma E.9, $\mathbb{E}_Q Z \le 63\sqrt{k}\,\Gamma_\gamma\sqrt{T/s}$. With $\Gamma_\gamma^2 = 7.2\gamma k\,\mathrm{OPT}/\beta$, Claim E.14 and $\rho^2 \le \beta/(120\gamma k)$ give $T \le \frac{2M}{15\gamma}\,\mathrm{OPT}$ and hence $\Gamma_\gamma^2 T \le \frac{7.2\gamma k\,\mathrm{OPT}}{\beta}\cdot\frac{2M\,\mathrm{OPT}}{15\gamma} \le \frac{kM}{\beta}\,\mathrm{OPT}^2$, so

$$\mathbb{E}_Q Z \le 63\sqrt{k\,\Gamma_\gamma^2 T/s} \le 63\;\mathrm{OPT}\sqrt{\frac{k^2M}{\beta s}}, \qquad (\mathbb{E}_Q Z)^2 \le 3969\,\frac{k^2M}{\beta s}\,\mathrm{OPT}^2\,.$$

For the variance, by Claim E.5, $0 \le b_v(p) \le 4a(p)^2$ for every $v \in \mathcal{V}$, so replacing q_t changes $B_v(Q)$ by at most $\frac{4}{s}(a(q_t)^2 + a(q_t')^2)$ uniformly in $v \in \mathcal{V}$, and hence changes Z by at most the same amount. By Efron–Stein and Claim E.13,

$$\mathrm{Var}(Z) \le \frac{1}{2}\sum_t \mathbb{E}[(Z - Z^{(t)})^2] \le \frac{32}{s}\mathbb{E}[a^4] \le \frac{32Mk}{\beta s}\,\mathrm{OPT}^2\,.$$

Hence $\mathbb{E}[Z^2] = (\mathbb{E}[Z])^2 + \mathrm{Var}(Z) \le \frac{4001k^2M}{\beta s}\,\mathrm{OPT}^2$, and Markov gives $\Pr[Z > \epsilon\,\mathrm{OPT}] \le \frac{4001k^2M}{\beta s\epsilon^2} \le \delta/2$. □

Proof of Theorem 1.6. By a union bound, $\mathcal{E}_Q$ and the event of Lemma E.18 occur simultaneously with probability at least $1-\delta$. By Lemma E.6, every $C \in \mathcal{S}_\gamma(Q) \cup \mathcal{S}_\gamma(P)$ equals $C(v)$ for some $v \in \mathcal{V}$, so applying the identity (17) together with Lemma E.17 and Lemma E.18,

$$|\mathrm{cost}(C,P) - \mathrm{cost}(C,Q)| \le 14\epsilon\,\mathrm{cost}(C,P) + \epsilon\,\mathrm{OPT} \le 15\epsilon\,\mathrm{cost}(C,P).$$

The theorem statement is obtained by rescaling ϵ. □

F Immediate Applications of Theorem 1.1

We derive immediate applications of Theorem 1.1 through the transfer-to-submetric property of stable coresets, we collect them here for potential future use.

Kernel Clustering. A kernel function $K\colon X \times X \to \mathbb{R}$ is characterized by the existence of a Hilbert space $\mathcal{H}$ and a feature map $\varphi\colon X \to \mathcal{H}$ such that $K(x,y) = \langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}}$ for all $x, y \in X$. In *kernel* 1*-mean*, the goal is to minimize $\mathrm{cost}_\varphi(c,P) := \frac{1}{n}\sum_{x\in P} \|\varphi(x) - \varphi(c)\|_{\mathcal{H}}^2$ over all $c \in X$.

Corollary F.1. *Let $P \subseteq X$ be a finite set with kernel $K\colon X \times X \to \mathbb{R}$. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-2}\delta^{-1})$ from P is, with probability at least $1-\delta$, a stable $(\delta\epsilon, 8\epsilon)$-coreset for kernel 1-mean with respect to K.*

Clustering in ℓ_p for $0 < p \le 2$. For every $0 < p \le 2$ and every $0 < \alpha \le p/2$, the metric space $(\ell_p^d, \|\cdot\|_p^\alpha)$ embeds isometrically into a Hilbert space, this follows from a classical theorem of Schoenberg (1938) (see also Bretagnolle et al. (1967) and Wells & Williams (1975, Theorem 4.10)). In addition to Corollary 1.2, taking $\alpha = p/2$ yields the following.

Corollary F.2. *Let $p \in (0,2]$ and $P \subset \mathbb{R}^d$ finite. Then for every $\epsilon \in (0, 1/10)$ and $\delta > 0$, a uniform sample of (sufficiently large) size $O(\epsilon^{-2}\delta^{-1})$ from P is, with probability at least $1-\delta$, a stable $(\delta\epsilon, 8\epsilon)$-coreset for the cost*

$$\mathrm{cost}(c,P) \;:=\; \frac{1}{n}\sum_{x\in P} \|x-c\|_p^p\,.$$

Approximation algorithms for k-median. A known framework reduces the design of approximation algorithms for k-median to obtaining weak coreset guarantees for the 1-median problem through uniform sampling (Kumar et al., 2004; Ackermann et al., 2010). This framework was recently instantiated by (Carmel & Krauthgamer, 2026) using stable coresets, yielding approximation algorithms for k-median in ℓ_1 and every submetric of ℓ_1, including Hamming, Kendall–tau, Jaccard, and tree metrics (see Theorem 5.4 in (Carmel & Krauthgamer, 2026)). Since the running time of the framework depends exponentially on the coreset size, our improved stable coreset bounds translate directly into faster approximation algorithms.