Beyond Worst-Case Coreset Bounds for k-Clustering via Determinantal Sampling

Diptarka Chakraborty
School of Computing
National University of Singapore

Satyaki Mukherjee
Department of Mathematics
National University of Singapore

Gaurav Vallabhdas Revankar
Department of Mathematics
Indian Institute of Technology Bombay

Hoang-Son Tran
Department of Mathematics
National University of Singapore

Abstract

Massive datasets in modern machine learning have made data reduction a central challenge, particularly for clustering tasks where memory and computational constraints demand compact yet faithful summaries. A standard approach is to construct an ε-*coreset*: a small weighted subset that approximately preserves the clustering cost for every plausible choice of centers. For the (k, z)-*clustering problem*, existing worst-case bounds on coreset size are essentially tight, ruling out substantially smaller coresets in general. However, such worst-case instances are often unrepresentative of real-world data.

In this work, we show that significantly smaller coresets are possible under mild and natural assumptions on the underlying data distribution. We introduce a new correlated sampling framework, called *determinantal sampling*, based on a novel application of determinantal point processes. Using this framework, we obtain an efficiently constructible ε-coreset for (k, z)-clustering in $\mathbb{R}^d$ whose dependence on $1/\varepsilon$ has exponent strictly smaller than 2 when d is fixed. This improves over the worst-case ε^{-2} barrier under our beyond-worst-case assumptions. To the best of our knowledge, this is the first result that provably surpasses these lower bounds through beyond-worst-case assumptions. Finally, we validate our approach on synthetic and real-world benchmark datasets, where it consistently achieves smaller coresets than existing state-of-the-art methods, even without explicitly enforcing the assumptions used in the analysis.

1 Introduction

The proliferation of massive datasets in modern machine learning applications necessitates the development of robust data reduction and summarization techniques. These methods are essential for minimizing memory usage and enabling distributed computations across systems with limited local storage. A prevalent strategy for managing such massive data involves sparsifying the input while meticulously preserving its underlying structural properties. This ensures that subsequent machine learning tasks, whether supervised or unsupervised, can be executed (nearly) accurately using only the summarized subset. Among the most prominent data summarization paradigms is the *coreset*. Since its introduction, coresets have garnered substantial attention in recent years for their ability to facilitate highly efficient algorithms in terms of memory, time, and parallel processing, with applications in streaming [HPM04, FS05, BFL+16, BFLR19], distributed [BEL13, HLSZ22], and dynamic settings [HK20, WGD21]. Their utility spans diverse domains, including computational geometry [AHPV04], graph sparsification [SS08, BSS09], linear

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and logistic regression [DDH+09, MSSW18], clustering [HPM04, FL11, CAGLS+22, HHHW23], classification [MMR21], subspace approximation [WY23], and learning mixture models [LFKF18].

Clustering remains a fundamental data processing task with pervasive applications across fields such as image segmentation, document categorization, anomaly detection, and healthcare analytics; e.g., see [Llo82, TSK+06, AV07, CN12]. One of the well-studied variants is the (k, z)-*clustering* that, given a set P of points from a metric space $\mathcal{X}$ and a non-negative integer k, asks to find a set C of at most k centers from the underlying metric space $\mathcal{X}$ that minimizes the cost, defined as $\texttt{cost}_z(P, C) := \sum_{x \in P} \partial^z(x, C)$, where $\partial(x, C)$ denotes the minimum distance from x to a center point in C. Notable special cases of this variant include the classical k-*median* clustering when $z = 1$, and the ubiquitous k-*means* clustering in Euclidean space when $z = 2$.

Given the massive scale of modern datasets, it is highly advantageous to extract a (weighted) representative subset that faithfully approximates the clustering cost for any arbitrary set of centers. Specifically, this subset must preserve the cost within a multiplicative approximation factor of $(1+\varepsilon)$ for any $\varepsilon > 0$. Such a representative subset is formally known as an ε-*coreset* (see Section 2 for a formal definition). Extensive research in this area has established a tight theoretical understanding of coreset sizes for the (k, z)-clustering problem [Che09, LS10, BBCA+19, FSS20, HV20]. The best-known upper bound on the size of such a coreset is $\tilde{O}_z(k\varepsilon^{-z-2})$ [CASS21, CAGLS+22], where $\tilde{O}_z(\cdot)$ hides logarithmic factors and dependencies only on z. Later, a similar bound is also achieved using *sensitivity sampling* for Euclidean k-means (i.e., for $z = 2$) [BCAP+24]. The above bound is known to be essentially tight for Euclidean spaces. More specifically, a lower bound of $\Omega(k\varepsilon^{-2})$ for the k-means [CALSS22] and for arbitrary z, a lower bound of $\Omega(k\varepsilon^{-z-2})$ for $\varepsilon \geq \Omega\left(k^{-1/(z+2)}\right)$ [HLW24], are already established.

These bounds definitely preclude the existence of smaller coresets in the worst-case scenario. However, empirical data encountered in real-world applications rarely exhibit worst-case behavior. That motivates the researchers to explore various beyond-worst-case assumptions to exploit when designing efficient clustering algorithms [KK10, ORSS13, BBG13, AMM17, CAS17]. However, to the best of our knowledge, only recently, [BCAP+24] looked into coreset construction from the beyond-worst-case perspective – more specifically, with *cost-stable clusterability criteria* from [ORSS13]. Unfortunately, even with that assumption, it is not possible to surpass the $\Omega(k\varepsilon^{-2})$ lower bound [BCAP+24].

It naturally motivates a compelling research question: Is it possible to circumvent these worst-case lower bounds by introducing realistic or natural assumptions about the distribution or structure of the input points? In this work, we provide an affirmative answer to this question, demonstrating that beyond-worst-case analysis can yield significantly more compact coresets.

Our Contribution. The primary contribution of this paper is an efficient coreset construction for the (k, z)-clustering problem that, under mild assumptions on the underlying data distribution, improves upon the known worst-case lower bound. Our approach is based on a correlated sampling method, which we call *determinantal sampling*, arising from a novel application of determinantal point processes. More precisely, we establish the following informal theorem.

Theorem 1.1 (Informal)**.** *Under certain mild assumptions on the underlying data distributions, as specified in Section 3, the (k, z)-clustering problem on $\mathbb{R}^d$ admits an efficiently constructible ε-coreset of size $\tilde{O}_{d,z}(k^{d/(d+1)}\varepsilon^{-2d/(d+1)})$.*

Here, $\tilde{O}_{d,z}(\cdot)$ hides polylogarithmic factors as well as factors depending only on d and z. In particular, for constant dimension d, this bound improves upon the worst-case lower bound of $\Omega(\varepsilon^{-2})$ established in [CALSS22]. This is especially relevant in practice, since many real-world datasets are effectively low-dimensional. Even when represented in a high-dimensional ambient space, by applying standard dimensionality-reduction techniques such as the Johnson–Lindenstrauss lemma, PCA, and related tools, one can often work in a reduced-dimensional space without significantly distorting the clustering structure. Consequently, our coreset construction is well-suited for practical scenarios in which the effective dimension is small, allowing one to beat the known worst-case bound. We also note that to derive Theorem 1.1, we state and prove a more general result in Theorem 3.1 for any Polish space. As such, similar results hold in more generality as long as one can find DPPs that satisfy the required variance bounds.

Table 1: Comparison between the tools imported from [BGSOT24] and our contributions.

Imported from [BGSOT24]	**Established in this paper**
Bernstein-type concentration for DPP linear statistics ([BGSOT24, Theorem 1] = our Theorem A.2)	Uniform guarantee for parametric losses under determinantal sampling via root-Lipschitz uniformization and Hausdorff covering (Theorem 3.1)
Discretized multivariate OPE kernel and its variance rate $O(m^{-1-1/d})$ [BGL21, BGSOT24]	Euclidean (k, z)-clustering with explicit size and exponent of $1/\varepsilon$ below 2 for fixed d (Corollary 3.2)
General DPP-coreset viewpoint	Verification of **(B.1)** under weak conditions (Appendix B.2); multi-kernel benchmark

Our work is motivated by the observation that real-world datasets arising in machine learning are unlikely to exhibit worst-case behavior. This has already led to recent attempts to obtain smaller coresets under structural or distributional assumptions, initiated in [BCAP+24]. However, the cost-stable clusterability conditions introduced there are not sufficient to overcome the worst-case lower bound. Ours is the first to provably surpass that lower bound by leveraging natural assumptions, like subgaussianity, on the data distribution. Although our analysis relies on certain assumptions, they are mild, well-motivated, and commonly satisfied in practice (see Section 3).

On the technical side, we introduce a new correlated sampling framework, called *determinantal sampling*, based on a novel use of determinantal point processes. The key idea is to exploit correlations in the sampling process to obtain stronger coreset guarantees than those achievable with standard independent sampling methods. **Relation to [BGSOT24]:** We use two tools from [BGSOT24]: the concentration inequality for DPP linear statistics and the discretized OPE kernel with its variance rate. Uniform control over the (k, z)-clustering loss class, which their generic coreset theorems do not cover, is new here. Table 1 summarizes the differences.

An additional practical benefit is that once such a kernel has been constructed, it can be reused to efficiently generate multiple coresets. This is particularly appealing in applications where repeated sampling is desirable, such as improving robustness, obtaining stronger performance guarantees for downstream clustering tasks, attaining stability over noise in ensemble-style clustering, and applications requiring reliability, uncertainty estimation, or hyperparameter selection (e.g., [YCDF+19, DF03, ZQL+25]).

To make the framework concrete and evaluate its practical performance, we implement several DPP kernel constructions tailored to the clustering setting, and test them on both synthetic and real-world benchmark datasets. The synthetic experiments support the improved ε-dependence predicted by our analysis, while the real-world benchmark experiments test the method without explicitly enforcing the assumptions required by the theory. Even in this setting, our coreset construction outperforms state-of-the-art methods in terms of coreset size for comparable clustering error. The results also highlight the modular role of the kernel: our theorem applies whenever the induced DPP sampler satisfies the required variance conditions, while the experiments include additional kernels for which we do not yet prove these guarantees. Their strong empirical performance suggests that determinantal sampling may be effective beyond the regime currently covered by our analysis.

2 Preliminaries

Notations. Throughout this paper, we consider *Polish spaces*, which are separable, complete metric spaces, as defined in [Coh80]. From now on, we assume the metric spaces we consider are separable and complete. It is worth remarking that most general metric spaces satisfy the separability and completeness property; for more details, see a discussion in Appendix E.

For a metric space $(\mathcal{X}, \partial)$, a subset $N \subseteq \mathcal{X}$ is a *u-net* if every point in $\mathcal{X}$ is within distance u of some point in N. The *covering number* $\mathcal{N}(\mathcal{X}, u, \partial)$ denotes the minimum size of a u-net for $(\mathcal{X}, \partial)$.

(k, z)-Clustering and Coresets. The *(k, z)-clustering* problem is defined as follows: Let $(\mathcal{X}, \partial)$ be a metric space, where $\mathcal{X}$ could be either continuous or discrete. Given a dataset $\mathcal{P} \subseteq \mathcal{X}$ of n points and a parameter $z \geq 1$, the goal is to choose a set $C \subseteq \mathcal{X}$ of (at most) k centers minimizing

$$\mathrm{cost}_z(\mathcal{P}, C) := \sum_{x \in \mathcal{P}} \partial(x, C)^z, \tag{1}$$

where $\partial(x, C) := \min_{y \in C} \partial(x, y)$ denotes the distance from x to its nearest center in C.

A coreset is a data-reduction technique that approximates an optimization objective using a small (weighted) subset of the data. In the context of (k, z)-Clustering, let $\mathcal{X}^k$ denote the set of all the k-sized subsets (potential center sets) of $\mathcal{X}$. A coreset is defined as follows.

Definition 2.1 (Coreset [LS10, FL11])**.** Consider a metric space $(\mathcal{X}, \partial)$. For a dataset $\mathcal{P} \subseteq \mathcal{X}$ of n points, an *ε-coreset* for the (k, z)-Clustering problem is a weighted subset $\mathcal{S} \subseteq \mathcal{P}$ with weight $w : \mathcal{S} \to \mathbb{R}_{\geq 0}$ such that

$$\forall C \in \mathcal{X}^k, \qquad \sum_{x \in \mathcal{S}} w(x) \cdot \partial(x, C)^z \in (1 \pm \varepsilon) \cdot \mathrm{cost}_z(\mathcal{P}, C).$$

The goal is to give good upper bounds for the size of the smallest possible ε-coreset. Our guarantees are for centers in a fixed compact set $K \supseteq \mathcal{P}$, i.e. C ranges over k-point subsets of K, as in our experiments (Appendix 6).

Samplers via determinantal point processes. Let $\mathcal{P}$ be a Polish space equipped with a reference measure ν. We recall that a point process $\mathcal{S}$ on $\mathcal{P}$ is a random locally finite subset of $\mathcal{P}$. The standard definition of determinantal point processes ([Mac75, HKPV06]) is as follows:

Definition 2.2. A point process $\mathcal{S}$ on $\mathcal{P}$ is called a *determinantal point process* (DPP) with respect to ν if there exists a measurable kernel $\mathbf{K} : \mathcal{P} \times \mathcal{P} \to \mathbb{C}$ such that, for all $k \in \mathbb{N}$ and all bounded measurable functions $f : \mathcal{P}^k \to \mathbb{R}$, we have

$$\mathbb{E}\Big[\sum_{\substack{x_1, \ldots, x_k \in \mathcal{S} \\ \forall i \neq j, x_i \neq x_j}} f(x_1, \ldots, x_k) \Big] = \int_{\mathcal{P}^k} f(x_1, \ldots, x_k) \, \det\big[K(x_i, x_j)\big]_{i,j=1}^k \, \mathrm{d}\nu^{\otimes k}(x_1, \ldots, x_k). \quad (2)$$

The function $\mathbf{K}$ is called a *kernel* of the DPP, and ν the *background measure*. If $\mathbf{K}(x, y) = \overline{\mathbf{K}(y, x)}$, we call $\mathbf{K}$ a Hermitian kernel.

An important setting in machine learning is when the background space $\mathcal{P}$ is a finite set, which could be interpreted as a set of data points. The standard reference measure on $\mathcal{P}$ is the counting measure. In this discrete setting, DPPs have the following simple characterization:

Definition 2.3 (Discrete DPPs)**.** Let $\mathcal{P}$ be a finite set of n points. A random subset $\mathcal{S} \subset \mathcal{P}$ is a DPP on $\mathcal{P}$ if there exists an $n \times n$-matrix $\mathbf{K}$ (indexed by elements of $\mathcal{P}$) such that

$$\mathbb{P}(T \subseteq \mathcal{S}) = \det(\mathbf{K}_T), \qquad \forall\, T \subseteq \mathcal{P},$$

where $\mathbf{K}_T$ denotes the submatrix of $\mathbf{K}$ indexed by the subset $T \subset \mathcal{P}$.

Determinantal point processes (DPPs) have emerged as an attractive framework for repulsive sampling in machine learning, and several advanced algorithms have been developed for efficient DPP sampling and inference (e.g. [RRS+22]). For further discussion on DPPs, particularly from a machine learning perspective, we refer to the survey [KT12] and the recent survey [TPBG25].

DPPs as coresets. Given a dataset $\mathcal{P} = \{X_1, \ldots, X_n\} \subseteq \mathcal{X}$ consisting of n iid samples from a distribution μ and a function $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$, define

$$L(f) := \sum_{i=1}^n f(X_i). \quad (3)$$

Given a coreset $\mathcal{S} \subseteq \mathcal{P}$ with associated weights $w : \mathcal{S} \to \mathbb{R}_{\geq 0}$, define the *weighted coreset estimator* as $L_{\mathcal{S}}(f) := \sum_{x \in \mathcal{S}} w(x) \, f(x)$. We can now define a weighted coreset estimator using a Determinantal Point Process as follows.

Definition 2.4 (Determinantal Coreset Estimator)**.** Sample $\mathcal{S} \subset \mathcal{P}$ by a DPP with Kernel $\mathbf{K}$ and reference measure ν as the counting measure on $\mathcal{P}$. We choose the weight $\omega(x) = \mathbf{K}(x, x)^{-1}$. Thus $L_{\mathcal{S}}(f) = \sum_{x \in \mathcal{S}} \mathbf{K}(x, x)^{-1} f(x)$. By (2), this choice of ω makes $L_{\mathcal{S}}(f)$ an unbiased estimator for $L(f)$. In this case, we say that $\mathcal{S}$ is sampled by a *determinantal sampler* .

3 Coreset Construction beyond Worst-Case

3.1 General determinantal sampler guarantees

We begin with a uniform coreset bound for a general parametric loss class under determinantal sampling, which we later specialize to clustering objectives. Let $\mathcal{F} = \{f_C : C \in \mathcal{C}\}$ be a class of nonnegative functions indexed by a metric space $(\mathcal{C}, \partial_{\mathcal{C}})$. (Note that the space $(\mathcal{C}, \partial_{\mathcal{C}})$ is different from the metric space $(\mathcal{X}, \partial)$ where the points in datasets belong.) We assume:

(A.1) The parameter space $(\mathcal{C}, \partial_{\mathcal{C}})$ is bounded and has parametric covering dimension ℓ: there exists $c_{\mathcal{C}} > 0$ such that

$$\mathcal{N}(\mathcal{C}, u, \partial_{\mathcal{C}}) \le \left(\frac{c_{\mathcal{C}}}{u}\right)^{\ell}, \qquad 0 < u \le 1.$$

(A.2) The map $C \mapsto f_C(x)^{1/z}$ is 1-Lipschitz uniformly in x: for all $C, C' \in \mathcal{C}$ and all $x \in \mathcal{X}$,

$$\left| f_C(x)^{1/z} - f_{C'}(x)^{1/z} \right| \le \partial_{\mathcal{C}}(C, C').$$

For the (k, z)-clustering loss, these assumptions are natural. Taking $f_C(x) = \partial(x, C)^z$, where C denotes a k-tuple of centers, we verify in the next section that $\mathcal{F}$ satisfies Assumptions **(A.1)** and **(A.2)** with $\ell = dk$. Thus ℓ can be viewed as the effective dimension of the parameter space; in the standard Euclidean setting, it coincides with the usual parameter dimension.

For technical reasons, we will also need some assumption which is essentially equivalent to the requirement that the dataset $\{x_1, ..., x_n\}$ is non-degenerate. We will later show (in Theorem B.2) that this ensuing assumption will hold with high probability when the data is, for instance, drawn independently from a subgaussian measure (e.g. Appendix B.2 shows that much weaker conditions suffice).

(B.1) $\frac{1}{n}|L(f)| \ge c$, for some $c > 0$, uniformly on $\mathcal{F}$.

Theorem 3.1 (DPP Coreset Bound for General (k, z)-Clustering)**.** *Let $\mathcal{S}$ be drawn from a determinantal sampler with a Hermitian kernel $\mathbf{K}$ and with respect to the counting measure of a finite set $\mathcal{P} = \{x_1, \dots, x_n\} \subset \mathcal{X}$. Let $m = \mathbb{E}[|\mathcal{S}|]$. Assume that for all $i \in \{1, \dots, n\}$, $\mathbf{K}(x_i, x_i) \ge \rho \cdot m/n$ for some $\rho > 0$ not depending on m, n. Suppose $V \ge \sup_{C \in \mathcal{C}} \mathrm{Var}\left[\frac{1}{n} L_{\mathcal{S}}(f_C)\right]$ and $|\mathcal{S}| \le B \cdot m$ almost surely for some $B > 0$. Then under **(A.1)**, **(A.2)**,**(B.1)** and for some constants c_1, c_2, c_3 depending on $\mathcal{C}, B, \rho, z, c$ we have, for any $0 \le \varepsilon \le \frac{c_3 m V}{\sup_{C \in \mathcal{C}} \|f_C\|_\infty}$,*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \ge \varepsilon\right) \le 2 \exp\left(c_1 \ell - \ell \log \varepsilon - \frac{c_2 \varepsilon^2}{V}\right).$$

Proof. Define

$$\mathcal{E}(C) := \left(\frac{L_{\mathcal{S}}(f_C)}{n}\right)^{1/z} - \left(\frac{L(f_C)}{n}\right)^{1/z}, \qquad \tilde{\varepsilon} := \varepsilon c^{1/z}/(2z).$$

Then a simple algebraic inequality (detailed in Proposition A.1 in appendix) implies that under assumption **(B.1)** we have,

$$\mathbb{P}\left(\exists\, f \in \mathcal{F} : \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \ge \varepsilon\right) \le \mathbb{P}\left(\sup_{C \in \mathcal{C}} |\mathcal{E}(C)| \ge \tilde{\varepsilon}\right).$$

Let $C, C' \in \mathcal{C}$. By Assumption **(A.2)**, $|f_C(x)^{1/z} - f_{C'}(x)^{1/z}| \le \partial_{\mathcal{C}}(C, C')$ for all $x \in \mathcal{P}$. Thus by Minkowski's inequality applied to $L_{\mathcal{S}}(f)/n = \sum_{x \in \mathcal{S}} \mathbf{K}(x, x)^{-1} n^{-1} f(x)$,

$$\left|(L_{\mathcal{S}}(f_C)/n)^{1/z} - (L_{\mathcal{S}}(f_{C'})/n)^{1/z}\right| \le \partial_{\Theta}(C, C') \Big(\sum_{x \in \mathcal{S}} \tfrac{1}{nK(x,x)}\Big)^{1/z}$$

Since $\mathbf{K}(x, x) \ge \rho m/n$ by (i) and $|\mathcal{S}| \le Bm$, we have a.s.

$$\left|(L_{\mathcal{S}}(f_C)/n)^{1/z} - (L_{\mathcal{S}}(f_{C'})/n)^{1/z}\right| \le (B/\rho)^{1/z}\, \partial_{\mathcal{C}}(C, C').$$

The same argument with unit weights gives $\left|(L(f_C)/n)^{1/z} - (L(f_{C'})/n)^{1/z}\right| \leq \partial_{\mathcal{C}}(C, C')$, and therefore

$$|\mathcal{E}(C) - \mathcal{E}(C')| \leq c_0\, \partial_{\mathcal{C}}(C, C') \quad \text{almost surely with,} \qquad c_0 := 1 + (B/\rho)^{1/z}.$$

Let Γ be a $\tilde{\varepsilon}/(2c_0)$-net for $\mathcal{C}$. Then $\sup_{C\in\mathcal{C}} |\mathcal{E}(C)| \leq \sup_{C'\in\Gamma} |\mathcal{E}(C')| + \frac{\tilde{\varepsilon}}{2}$, so $\mathbb{P}(\sup_{C\in\mathcal{C}} |\mathcal{E}(C)| \geq \tilde{\varepsilon}) \leq \sum_{C'\in\Gamma} \mathbb{P}\big(|\mathcal{E}(C')| \geq \frac{\tilde{\varepsilon}}{2}\big)$.

Fix $C' \in \Gamma$. Again some simple algebra detailed in Proposition A.1 (in appendix) implies that under assumption **(B.1)** we have,

$$\mathbb{P}\left(|\mathcal{E}(C')| \geq \frac{\tilde{\varepsilon}}{2}\right) \leq \mathbb{P}\left(\frac{1}{n}\big|L_{\mathcal{S}}(f_{C'}) - L(f_{C'})\big| \geq \frac{\tilde{\varepsilon}}{2}\, c^{(z-1)/z}\right). \tag{4}$$

Using equation (2) it is easy to verify that $\mathbb{E}\left[L_{\mathcal{S}}(f_{C'})\right] = L(f_{C'})$. We can thus conclude the results by using standard concentration results on Determinantal Samplers given by, e.g., [BGSOT24],[BD14]. In particular using Theorem A.2 (Theorem 1 of [BGSOT24]) stated in appendix A.2 with $\phi(x) = f(x)/K(x,x)$ to the R.H.S. of equation (4), we obtain

$$\mathbb{P}\left(|\mathcal{E}(C')| \geq \frac{\tilde{\varepsilon}}{2}\right) \leq 2\exp\left(-\frac{c^{2(z-1)/z}\tilde{\varepsilon}^2}{16AV}\right)$$

Using $\mathbf{K}(x,x) \geq \rho m/n$ the above inequality holds for all $0 \leq \tilde{\varepsilon} \leq 4A\rho mV/(3c^{(z-1)/z}\|f_C\|_\infty)$.

Substituting $\tilde{\varepsilon} = \varepsilon c^{1/z}/(2z)$ and writing $c_2 = \frac{c^2}{64Az^2}$, and $c_3 = \frac{8zA\rho}{3c}$, we obtain

$$\mathbb{P}\left(|\mathcal{E}(C')| \geq \frac{\tilde{\varepsilon}}{2}\right) \leq 2\exp\left(-\frac{c_2\varepsilon^2}{V}\right),$$

and the constraint on $\tilde{\varepsilon}$ translates to $0 \leq \varepsilon \leq \frac{c_3 mV}{\|f_C\|_\infty}$.

Recall, Γ is a $\tilde{\varepsilon}/(2c_0)$-net for $\mathcal{C}$. By Assumption **(A.1)**, $|\Gamma| \leq (2c_0c_{\mathcal{C}}/\tilde{\varepsilon})^{\ell}$. After enlarging c_1 if necessary, $\log|\Gamma| \leq c_1\ell - \ell\log\varepsilon$. Combining all estimates,

$$\mathbb{P}\left(\exists\, f \in \mathcal{F} : \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq 2\exp\left(c_1\ell - \ell\log\varepsilon - \frac{c_2\varepsilon^2}{V}\right)$$

for all $0 \leq \varepsilon \leq \min\{1,\, c_3 mV/\|f_C\|_\infty\}$. □

3.2 Euclidean (k,z)-Clustering

In the ensuing section, we now apply Theorem 3.1 to the (k,z) problem with $\mathcal{X} = \mathbb{R}^d$ to derive the following Corollary. We will then show that Theorem 1.1 holds.

Corollary 3.2 (Coreset Size for Euclidean (k,z)-Clustering)**.** *Given $\varepsilon, \delta \geq 0$, let $\mathcal{S}$ be drawn from a determinantal sampler with a Hermitian kernel $\mathbf{K}$ and with respect to the counting measure of a finite set $\mathcal{P} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$. Let $K \subset \mathbb{R}^d$ be a bounded set containing $\mathcal{P}$, and let $\mathcal{C}$ be the family of k-point subsets of K. Let $m = \mathbb{E}[|\mathcal{S}|]$. Assume that for all $i \in \{1, \ldots, n\}$, $\mathbf{K}(x_i, x_i) \geq \rho \cdot m/n$ for some $\rho > 0$ not depending on n. Suppose $\sup_{C\in\mathcal{C}} \mathrm{Var}\left[\frac{1}{n}L_{\mathcal{S}}(f_C)\right] = \mathcal{O}(m^{-1-1/d})$ and $|\mathcal{S}| \leq B \cdot m$ almost surely for some constant $B > 0$. Further suppose for some constants, c and c', m satisfies*

$$c' \cdot \frac{1}{\varepsilon^d} \geq m \geq c\left(\frac{kd + kd\log(1/\varepsilon) + \log(2/\delta)}{\varepsilon^2}\right)^{\frac{d}{d+1}},$$

*Then under assumption **(B.1)** on the data, $\mathcal{S}$ is an ε-coreset with probability at least $1 - \delta$ (over the randomness of the coreset construction algorithm).*

Before proceeding with the proof argument, observe that Corollary 3.2 along with Remarks 1 and 2 immediately imply Theorem 1.1.

Proof Sketch. We verify that the Euclidean (k, z)-clustering loss class satisfies the hypotheses of Theorem 3.1. Equip the space $\mathcal{C}$ of k-center sets with the Hausdorff metric ∂_H. The nearest-center map is Lipschitz under this metric: if $y^* \in C$ is a closest center to x, then some $y' \in C'$ lies within $\partial_H(C, C')$ of y^*, so the triangle inequality gives $\partial(x, C') - \partial(x, C) \leq \partial_H(C, C')$; reversing the roles of C and C' yields $|\partial(x, C) - \partial(x, C')| \leq \partial_H(C, C')$.

Since $f_C(x)^{1/z} = \partial(x, C)$, Assumption **(A.2)** follows. Moreover, because the candidate centers lie in a bounded subset $K \subset \mathbb{R}^d$, a Euclidean u-net of K of size $(C_K/u)^d$ induces a u-net of $\mathcal{C}$ of size $(C_K/u)^{kd}$. Thus Assumption **(A.1)** holds with $\ell = kd$. The boundedness of K also gives $\sup_C \|f_C\|_\infty < \infty$, and the admissible-range condition of Theorem 3.1, namely $\varepsilon \leq c_3 m V / \sup_C \|f_C\|_\infty$, together with the variance scaling $V = \mathcal{O}(m^{-(1+1/d)})$, gives $\varepsilon = \mathcal{O}(m^{-1/d})$, or equivalently $m = \mathcal{O}(\varepsilon^{-d})$. This is precisely the upper bound $m \leq c' \cdot \varepsilon^{-d}$ stated in the corollary, ensuring the theorem is applied within its valid regime. Applying Theorem 3.1 with $\ell = kd$ gives

$$\mathbb{P}(\text{failure}) \leq 2 \exp\left(c_1 kd - kd \log \varepsilon - \frac{c_2 \varepsilon^2}{V} \right).$$

Requiring this probability to be at most δ, substituting $V = \mathcal{O}(m^{-(1+1/d)})$, and solving for m gives

$$m \gtrsim \left(\frac{kd + kd \log(1/\varepsilon) + \log(2/\delta)}{\varepsilon^2} \right)^{d/(d+1)},$$

which is the lower bound in the corollary. Combining both bounds shows that any m in the stated range guarantees failure probability at most δ, so $\mathcal{S}$ is an ε-coreset with probability at least $1 - \delta$. □

Discussions on Assumptions

Remark 1 (On the assumption **(B.1)**)**.** Recall that Assumption **(B.1)** requires $L(f)/n \geq c$ uniformly over $f \in \mathcal{F}$. However, since the dataset $\mathcal{P}$ is randomly generated, $L(f) = \sum_{i=1}^n f(X_i)$ depends on the particular realization of the data, making the verification of Assumption **(B.1)** potentially delicate. We address this issue through a technical result in the appendix; see Theorem B.2. This result verifies Assumption **(B.1)** with high probability in the dataset $\mathcal{P}$, under mild assumptions on the data distribution μ. Appendix 5 discusses how to check the assumptions in practice.

Remark 2 (Variance reduction and discretized multivariate OPE)**.** A key ingredient in the above corollary is the variance bound $V = \mathcal{O}(m^{-1-1/d})$, which is achieved by the discretized multivariate OPE kernel constructed in [BGL21, BGSOT24]. More precisely, this DPP satisfies $\mathrm{Var}[n^{-1} L_{\mathcal{S}}(f)] = \mathcal{O}(m^{-1-1/d})$ for any sufficiently regular test function f, with high probability over the dataset $\mathcal{P} = \{X_1, \ldots, X_n\}$, where the X_i are i.i.d. samples from a distribution supported on a d-dimensional hypercube. For a more detailed discussion of the construction of these discrete OPE kernels, we refer to [BGL21, BGSOT24] and the references therein. An analogous variance-reduction property has recently been established for wavelet-based DPPs [TGBG26].

We remark that the property $V = \mathcal{O}(m^{-1-\alpha})$ for some $\alpha > 0$, often referred to as *variance reduction*, is one of the key features of DPP samplers. It has played an important role in proving theoretical guarantees that DPP-based samplers can perform better than i.i.d. sampling; see, for example, [BH20] for DPP-based quadratures, [BGL21] for DPP-based minibatches, and [BGSOT24] for DPP-based coresets. An ongoing line of work investigates this phenomenon for substantially broader classes of continuous kernels, together with mechanisms for transferring the corresponding variance-reduction guarantees to discrete kernels.

Remark 3. Corollary 3.2 gives a trade-off between the size m of a DPP-based coreset and the accuracy rate ε. In particular, for an OPE-based coreset of size m with $V = \mathcal{O}(m^{-1-1/d})$, one obtains the high-probability accuracy rate $\varepsilon = \mathcal{O}(m^{-1/2-1/(2d)})$, whereas independent-sampling coresets achieve $\varepsilon = \mathcal{O}(m^{-1/2})$.

4 Experiments

We evaluate our DPP-based coreset samplers against standard baselines, including state-of-the-art near-optimal sensitivity construction by [HLW24] (abbreviated HLW). Two questions drive the evaluation: whether DPP samplers exhibit the faster error decay predicted by theory, and whether this advantage persists on real-world clustering benchmarks. All algorithms are implemented in Python 3

Algorithm 1 Determinantal coreset construction (expected size m)

Require: data $\mathcal{P} = \{x_1, \ldots, x_n\} \subset [-1, 1]^d$, target expected size m, kernel family (OPE, Haar or Vandermonde)

1: **Kernel construction.** Build a marginal kernel $\mathbf{K}$ on $\mathcal{P}$ (Appendix C):
DPP-OPE: Jacobi reference measure, discretized multivariate OPE features, rank-m projection;
DPP-Haar: 2^{jd} dyadic cells, $j \approx \log_2(m)/d$, cell-projection kernel;
DPP-Vandermonde: graded monomial Vandermonde matrix orthonormalized by QR.

2: **Sampling.** Draw $\mathcal{S} \sim \mathrm{DPP}(\mathbf{K})$ (for DPP-Haar, the fixed-size variant of Appendix C).

3: **Weighting.** Give each $x_i \in \mathcal{S}$ the weight $w_i = \mathbf{K}_{ii}^{-1}$.

4: **return** $\{(x_i, w_i) : x_i \in \mathcal{S}\}$

and run on an Apple MacBook with a 10-core M5 chip and 8GB of RAM, using `NumPy`, `SciPy`, and `scikit-learn` [PVG+11]. The OPE feature construction uses the `MultivariateJacobiOPE` class from `DPPy` [GPBV19]. For each method and coreset size m we draw $N_{\mathrm{REP}} = 150$ independent coresets and evaluate them on a fixed family of $N_{\mathrm{QUERY}} = 150$ random k-tuples of data points (as potential centers), drawn once per dataset k pair. We report the average-case relative error over this query family; supplementary supremum-style metrics appear in Appendix G.

Methods. We compare seven coreset samplers; each takes a dataset $\mathcal{P} \subset \mathbb{R}^d$ and a target size m and returns a weighted subset $\mathcal{S} \subset \mathcal{P}$. For a query center set C, the full data loss is $L(C) := \frac{1}{n}\sum_{x\in\mathcal{P}} \partial^2(x, C)$, and the coreset estimate is $L_{\mathcal{S}}(C) := \sum_{x\in\mathcal{S}} w(x)\partial^2(x, C)$, with method-specific weights w chosen to estimate the average loss $L(C)$. For independent importance sampling with probabilities p_x, we use the unbiased weight $w(x) = 1/(mnp_x)$ per sampled copy; for a discrete DPP with marginal kernel K, we use $w(x) = 1/(nK(x, x))$.

Classical baselines. Uniform draws m points uniformly with replacement in $O(m)$ time. Sensitivity [LS10] is the importance-sampling coreset using the practical sensitivity upper bounds of Bachem et al. [BLK17] from a D^2-weighted initialisation [AV07], in time $O(nk + nm)$. HLW [HLW24] is the near-optimal construction by Huang, Li, and Wu, based on a bicriteria D^2-approximation, ring decomposition, and per-ring sensitivity bounds. Uniform has the usual Monte Carlo $m^{-1/2}$ error rate for a fixed query family, but is not a worst-case clustering coreset. Sensitivity sampling gives the standard sensitivity-based $\widetilde{O}(S\varepsilon^{-2})$ coreset size, where S is the total sensitivity upper bound. For k-means in Euclidean space, HLW gives the near-optimal worst-case bound $\widetilde{O}(k^{3/2}\varepsilon^{-2})$.

Negative-dependence (DPP) samplers. Stratified [BGSOT24] is a simple projection DPP for $\mathcal{P} \subset [-1, 1]^d$: it partitions $[-1, 1]^d$ into a grid of m equal bins and draws one point uniformly from $\mathcal{P}$ in each occupied bin, running in $O(nm)$. Its limitation is immediate: it relies on the data being sufficiently well-spread so that many bins are occupied, which is unlikely for non-uniform or high-dimensional datasets. DPP-Haar is a dyadic-cell Haar-wavelet projection DPP: it partitions $[-1, 1]^d$ into 2^{jd} cells with $j \approx \log_2(m)/d$, samples one representative per selected non-empty cell, and reweights by cell and in-cell selection probabilities. DPP-OPE [BGSOT24] (Example 5 therein) is the discretised orthogonal polynomial ensemble, implemented via `MultivariateJacobiOPE` from `DPPy` [GPBV19]; its Jacobi reference-measure parameters are fitted coordinatewise from marginal moments with a 10% uniform fallback; it runs in $O(nm^2)$, and it is restricted to $d \leq 6$. DPP-Vandermonde [TBA19] (Algorithm 2) is a projection DPP built from the monomial Vandermonde matrix with graded total-degree multi-indices and an economy QR factorization in $O(nm^2)$; it requires no density estimation and applies up to $d \leq 16$. We call DPP-OPE and DPP-Vandermonde the polynomial DPPs. Full constructions and a comparison of the kernels are in Appendix C (Table 2).

Datasets. All datasets are normalized coordinatewise to $[-1, 1]^d$. The synthetic collection comprises six datasets: Uniform ($n = 982$, $d = 2$), Trimodal ($n = 981$, $d = 2$), Gaussian Blobs 2D ($n = 4,800$,$d = 2$), Gaussian Blobs 3D ($n = 4,702$, $d = 3$), Anisotropic Gaussians ($n = 2,880$, $d = 2$), and BIRCH Benchmark ($n = 9,600$, $d = 2$). Uniform 2D is drawn uniformly from $[-1, 1]^2$. Trimodal 2D comprises three isotropic Gaussian clusters with standard deviations 0.04, 0.15, and 0.40. Gaussian Blobs 2D and Gaussian Blobs 3D are five well-separated isotropic Gaussian blobs (std 0.3) generated with `make_blobs`. Anisotropic Gaussians 2D contains three anisotropic Gaus-

sian clusters, two of which are elongated along rotated axes. BIRCH Benchmark is a subsampled instance of 100 Gaussian blobs, used as a low-dimensional many-cluster benchmark [ZRL97]. All these synthetic datasets are sub-Gaussian and thus satisfy our required assumptions with high probability.

For real-world evaluation, we use two natively low-dimensional datasets and two higher-dimensional datasets reduced by PCA to two dimensions. We also consider additional datasets in the Appendix **(B.1)**. 3D Road Network ($n = 9,424$, $d = 3$) is the *3D Road Network (North Jutland, Denmark)* dataset from the UCI Machine Learning Repository [Kau13]. Twitter GPS ($n = 9,644$, $d = 2$) [CGS18] consists of geotagged longitude-latitude coordinates; its strong spatial clustering structure makes it particularly well-suited to spatial DPP methods.

High-dimensional datasets. UCI Adult (PCA-2) ($n = 9,338$, $d = 2$) is the UCI Adult dataset [BK96], represented by its six numeric attributes (age, fnlwgt, education-num, capital-gain/loss, hours-per-week) and then projected onto the top two principal components via PCA [PVG+11]; the $d = 2$ setting enables all DPP samplers and gives a clean benchmark for the polynomial constructions on real economic data. DIGIT (PCA-2) ($n = 1,797$, $d = 2$) is the UCI handwritten digits dataset [AK98] (64 dimensions, from `scikit-learn`), projected the same way.

Performance metric. We report the *mean relative error* $\overline{\mathrm{RelErr}}(m)$, defined as the average over $N_{\mathrm{REP}} = 150$ independent coreset draws of the per-draw average relative error:

$$\overline{\mathrm{RelErr}}(m) := \frac{1}{N_{\mathrm{REP}}} \sum_{r=1}^{N_{\mathrm{REP}}} \frac{1}{N_{\mathrm{QUERY}}} \sum_{j=1}^{N_{\mathrm{QUERY}}} \frac{|L_{\mathcal{S}^{(r)}}(C_j) - L(C_j)|}{L(C_j)},$$

where $\mathcal{S}^{(r)}$ denotes the r-th independently drawn coreset and the query centers $C_1, \ldots, C_{N_{\mathrm{QUERY}}}$ are fixed random k-tuples drawn once per dataset-k pair. Coreset sizes range over $m \in \{1^2, 2^2, \ldots, 29^2\}$ (29 values), clipped to $m \leq 300$ for datasets with $n < 5{,}000$ and to $m \leq n/3$ in all cases. We fix $k = 5$ throughout; supplementary variance and supremum-style plots appear in Appendix G.

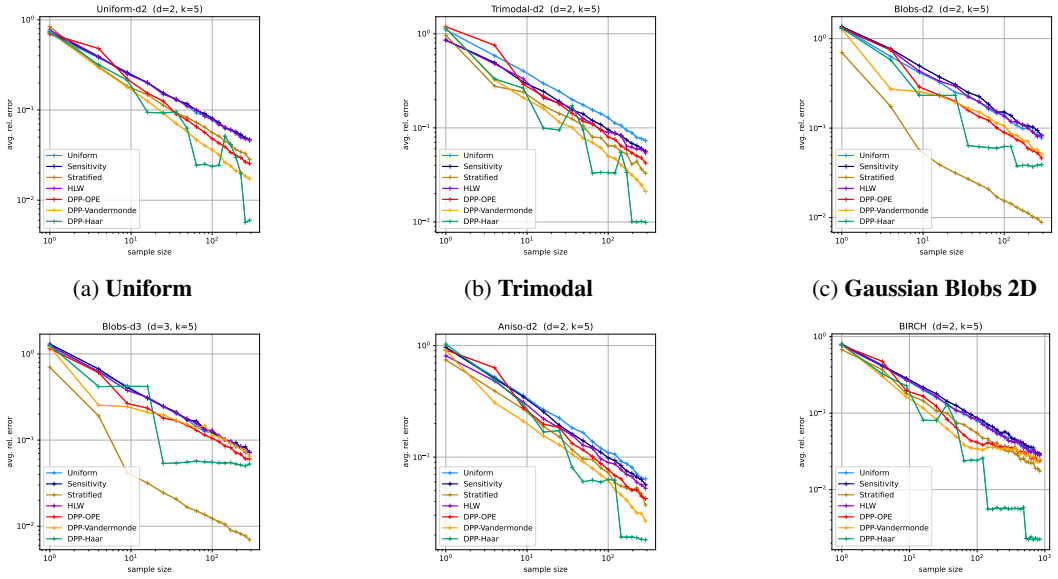
Research questions.

1. **(RQ1) Faster error decay.** Do DPP-based samplers achieve a steeper log–log decay of $\overline{\mathrm{RelErr}}(m)$, improving upon the $m^{-1/2}$ rate of i.i.d. methods, as predicted by Theorem 3.1?

2. **(RQ2) Smaller coreset on real data.** Do DPP methods attain a given target error with a smaller coreset than i.i.d. baselines on real-world benchmarks?

Results: RQ1 Across all six synthetic datasets (Figure 1), DPP-OPE and DPP-Vandermonde show a visibly steeper log–log decay than the i.i.d. baselines Uniform, Sensitivity, and HLW [HLW24]. On the $d = 2$ benchmarks, the empirical slope is consistent with the predicted $m^{-(1/2+1/4)}$ rate from Theorem 3.1, whereas on Gaussian Blobs 3D the gap narrows to the expected $m^{-(1/2+1/6)}$ regime. DPP-Haar matches or exceeds DPP-OPE and DPP-Vandermonde on the more spatially regular datasets. Stratified improves over Uniform when the data are well spread, but degrades on clustered or strongly non-uniform datasets, exactly as expected from its grid-based construction [BGSOT24]. The DPP-OPE rate is backed by prior results (Remark 2). Since the original submission of this paper, new results have been released justifying DPP-Haar ([TGBG26]). DPP-Haar oscillates in m because its variance reduction is cleanest when $m = 2^{jd}$, $j \in \mathbb{N}$, matches the dyadic partition, which does not affect DPP-OPE and DPP-Vandermonde. The other DPP results are empirical

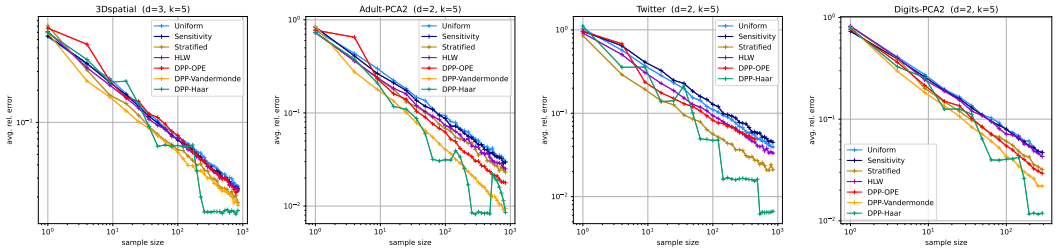
Results: RQ2 Figure 2 shows the same qualitative pattern on real data: for moderate and large m, at least one DPP sampler reaches a given error with fewer points than the independent baselines. The gains are smaller on 3D Road Network, consistent with the weaker $d = 3$ rate. On UCI Adult (PCA-2), DPP-OPE and DPP-Vandermonde are strongest, while on Twitter GPS, DPP-Haar performs best, suggesting that its multiscale spatial partition is well matched to geographic data. Stratified is less reliable on these non-uniform datasets, in line with its occupancy-based limitations [BGSOT24].



(a) **Uniform** (b) **Trimodal** (c) **Gaussian Blobs 2D**

(d) **Gaussian Blobs 3D** (e) **Anisotropic Gaussian** (f) **BIRCH**

Figure 1: $\overline{\mathrm{RelErr}}(m)$ vs. m (log–log, $k = 5$), synthetic datasets.



(a) **3D Road Network** (b) **UCI Adult (PCA-2)** (c) **Twitter GPS** (d) **DIGIT (PCA-2)**

Figure 2: $\overline{\mathrm{RelErr}}(m)$ vs. m (log–log, $k = 5$), real-world datasets.

5 Conclusion and Future Work

In this paper, we study the (k, z)-clustering problem from a beyond-worst-case perspective and show that substantially smaller coresets are possible under natural assumptions on the input distribution. To achieve this, we introduce determinantal sampling and use it to obtain stronger guarantees than those available through standard independent sampling techniques. Beyond its theoretical significance, we show the efficacy of our method by performing experiments on various standard datasets. Overall, our results provide further evidence that realistic structural assumptions can meaningfully improve the landscape of coreset construction for clustering.

It would be interesting to relax or further generalize the distributional assumptions used in our analysis, and to better understand the weakest conditions under which one can still beat the worst-case lower bounds. Another important open question is whether similar improvements can be achieved for broader classes of clustering objectives, more general metric spaces, or in high-dimensional settings without relying on dimensionality reduction. Finally, it would be valuable to explore whether the determinantal-sampling framework can be extended beyond clustering to other coreset problems in regression, subspace approximation, and related summarization tasks.

Acknowledgments and Disclosure of Funding

Funding. Diptarka Chakraborty was supported in part by an MoE AcRF Tier 1 grant (T1 251RES2303) and a Google South & South-East Asia Research Award. Satyaki Mukherjee was funded by grant A-8003802-00-00 at the National University of Singapore, under the supervision of Subhroshekhar Ghosh. Hoang Son Tran was supported by the Singapore MOE grants A-8003576-00-00 and A-8002488-00-00.

Competing interests. The authors declare no competing interests.

References

[AHPV04] Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Approximating extent measures of points. *Journal of the ACM (JACM)*, 51(4):606–635, 2004.

[AK98] E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. DOI: 10.24432/C50P49.

[AMM17] Haris Angelidakis, Konstantin Makarychev, and Yury Makarychev. Algorithms for stable and perturbation-resilient problems. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 438–451, 2017.

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM, 2007.

[BBCA+19] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pages 1039–1050, 2019.

[BBG13] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *Journal of the ACM (JACM)*, 60(2):1–34, 2013.

[BCAP+24] Nikhil Bansal, Vincent Cohen-Addad, Milind Prabhu, David Saulpic, and Chris Schwiegelshohn. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1707–1723. IEEE, 2024.

[BD14] Jonathan Breuer and Maurice Duits. The nevai condition and a local law of large numbers for orthogonal polynomial ensembles. *Advances in Mathematics*, 265:441–484, 2014.

[BEL13] Maria-Florina F Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general topologies. *Advances in neural information processing systems*, 26, 2013.

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

[BFLR19] Vladimir Braverman, Dan Feldman, Harry Lang, and Daniela Rus. Streaming coreset constructions for m-estimators. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 2019.

[BGL21] Rémi Bardenet, Subhroshekhar Ghosh, and Meixia Lin. Determinantal point processes based on orthogonal polynomials for sampling minibatches in sgd. In *Advances in Neural Information Processing Systems*, volume 34, pages 16226–16237, 2021.

[BGSOT24] Rémi Bardenet, Subhroshekhar Ghosh, Hugo Simon-Onfroy, and Hoang Son Tran. Small coresets via negative dependence: Dpps, linear statistics, and concentration. *Advances in Neural Information Processing Systems*, 37:84329–84349, 2024.

[BH20] Rémi Bardenet and Adrien Hardy. Monte carlo with determinantal point processes. *The Annals of Applied Probability*, 2020.

[BK96] Barry Becker and Ron Kohavi. Adult [Dataset]. UCI Machine Learning Repository, 1996.

[BLK17] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

[BSS09] Joshua D Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 255–262, 2009.

[CAGLS+22] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for euclidean k-means. *Advances in Neural Information Processing Systems*, 35:2679–2694, 2022.

[CALSS22] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1038–1051, 2022.

[CAS17] Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 49–60. IEEE, 2017.

[CASS21] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 169–182, 2021.

[CGS18] T-H. Hubert Chan, Arnaud Guerquin, and Mauro Sozio. Twitter data set. https://github.com/fe6Bc5R4JvLkFkSeExHM/k-center, 2018.

[Che09] Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[CN12] Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 561–580. Springer, 2012.

[Coh80] Donald L. Cohn. *Polish Spaces and Analytic Sets*, pages 251–296. Birkhäuser Boston, Boston, MA, 1980.

[DDH+09] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

[DF03] Sandrine Dudoit and Jane Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[FL11] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of Computing*, pages 569–578, 2011.

[FS05] Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of Computing*, pages 209–217, 2005.

[FSS20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.

[Gar17] D. J. H. Garling. *Analysis on Polish Spaces and an Introduction to Optimal Transportation*, volume 89 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2017.

[GKMV19] Jennifer Gillenwater, Alex Kulesza, Zelda Mariet, and Sergei Vassilvitskii. A tree-based method for fast repeated sampling of determinantal point processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2260–2268, 2019.

[GPBV19] Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. DPPy: DPP sampling with Python. *Journal of Machine Learning Research*, 20(180):1–7, 2019. MLOSS.

[HHHW23] Lingxiao Huang, Ruiyuan Huang, Zengfeng Huang, and Xuan Wu. On coresets for clustering in small dimensional euclidean spaces. In *International Conference on Machine Learning*, pages 13891–13915. PMLR, 2023.

[HK20] Monika Henzinger and Sagar Kale. Fully-Dynamic Coresets. In Fabrizio Grandoni, Grzegorz Herman, and Peter Sanders, editors, *28th Annual European Symposium on Algorithms (ESA 2020)*, volume 173 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 57:1–57:21, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[HKPV06] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability surveys*, 2006.

[HLSZ22] Lingxiao Huang, Zhize Li, Jialin Sun, and Haoyu Zhao. Coresets for vertical federated learning: Regularized linear regression and k-means clustering. *Advances in Neural Information Processing Systems*, 35:29566–29581, 2022.

[HLW24] Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for euclidean (k, z)-clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1594–1604, 2024.

[HPM04] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of Computing*, pages 291–300, 2004.

[HV20] Lingxiao Huang and Nisheeth K Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1416–1429, 2020.

[Kau13] Michael Kaul. 3D Road Network (North Jutland, Denmark) [Dataset]. UCI Machine Learning Repository, 2013.

[KK10] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[KT12] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

[LFKF18] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets. *Journal of Machine Learning Research*, 18(160):1–25, 2018.

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LS10] Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[Mac75] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

[MMR21] Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification–simplified and strengthened. *Advances in Neural Information Processing Systems*, 34:11643–11654, 2021.

[MRS20] Konstantin Makarychev, Aravind Reddy, and Liren Shan. Improved guarantees for k-means++ and k-means++ parallel. In *Advances in Neural Information Processing Systems*, volume 33, pages 16142–16152, 2020.

[MSSW18] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. *Advances in Neural Information Processing Systems*, 31, 2018.

[ORSS13] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RRS+22] Aravind Reddy, Ryan A. Rossi, Zhao Song, Anup Rao, Tung Mai, Nedim Lipka, Gang Wu, Eunyee Koh, and Nesreen Ahmed. One-pass algorithms for MAP inference of nonsymmetric determinantal point processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18463–18482, 2022.

[SS08] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 563–568, 2008.

[TBA19] Nicolas Tremblay, Simon Barthélmé, and Pierre-Olivier Amblard. Determinantal point processes for coresets. *Journal of Machine Learning Research*, 20(168):1–70, 2019.

[TGBG26] Hoang-Son Tran, Pranav Gupta, Rémi Bardenet, and Subhroshekhar Ghosh. State-of-art minibatches via novel DPP kernels: discretization, wavelets, and rough objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

[TPBG25] Hoang-Son Tran, Vladimir Petrovic, Remi Bardenet, and Subhroshekhar Ghosh. Negative dependence as a toolbox for machine learning: review and new developments. *arXiv preprint arXiv:2502.07285*, 2025.

[TSK+06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.

[WGD21] Zixiu Wang, Yiwen Guo, and Hu Ding. Robust and fully-dynamic coreset for continuous-and-bounded learning (with outliers) problems. *Advances in Neural Information Processing Systems*, 34:14319–14331, 2021.

[WY23] David Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In *International Conference on Machine Learning*, pages 37238–37272. PMLR, 2023.

[YCDF+19] Han Yu, Brian Chapman, Arianna Di Florio, Ellen Eischen, David Gotz, Mathews Jacob, and Rachael Hageman Blair. Bootstrapping estimates of stability for clusters, observations and model selection: H. yu et al. *Computational Statistics*, 34(1):349–372, 2019.

[ZQL+25] Xu Zhang, Haoye Qiu, Weixuan Liang, Hui Liu, Junhui Hou, and Yuheng Jia. Generalization performance of ensemble clustering: From theory to algorithm. In *International Conference on Machine Learning*, pages 75901–75927. PMLR, 2025.

[ZRL97] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.

A Technical results related to Theorem 3.1

A.1 A simple algebraic identity

Proposition A.1. *For $C \in \mathcal{C}$, define*

$$Y(C) := \frac{L_{\mathcal{S}}(f_C)}{n}, \qquad Z(C) := \frac{L(f_C)}{n}, \qquad \mathcal{E}(C) := Y(C)^{1/z} - Z(C)^{1/z},$$

and let $\tilde{\varepsilon} := \varepsilon c^{1/z}/(2z)$, where $z \geq 1$ and $0 < \varepsilon \leq 1$. Under Assumption (B.1),

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{C \in \mathcal{C}} |\mathcal{E}(C)| \geq \tilde{\varepsilon}\right).$$

For each fixed $C \in \mathcal{C}$, separately,

$$\mathbb{P}\left(|\mathcal{E}(C)| \geq \frac{\tilde{\varepsilon}}{2}\right) \leq \mathbb{P}\left(\frac{1}{n}|L_{\mathcal{S}}(f_C) - L(f_C)| \geq \frac{\tilde{\varepsilon}}{2} c^{(z-1)/z}\right).$$

Proof. First suppose z is a positive integer. If $\sup_{C \in \mathcal{C}} |\mathcal{E}(C)| \leq \tilde{\varepsilon}$, then

$$\frac{Y(C)}{Z(C)} - 1 = \mathcal{E}(C)\, Z(C)^{-1/z} \sum_{i=0}^{z-1} \left(\frac{Y(C)}{Z(C)}\right)^{i/z}.$$

By simplifying the above, since $Z(C) \geq c$ by assumption (B.1),

$$\frac{Y(C)^{1/z}}{Z(C)^{1/z}} = 1 + \frac{\mathcal{E}(C)}{Z(C)^{1/z}} \leq 1 + \frac{\tilde{\varepsilon}}{c^{1/z}} = 1 + \frac{\varepsilon}{2z} \leq e^{\varepsilon/(2z)}.$$

Hence, for every $0 \leq i \leq z-1$,

$$\left(\frac{Y(C)}{Z(C)}\right)^{i/z} = \left(\frac{Y(C)^{1/z}}{Z(C)^{1/z}}\right)^{i} \leq e^{i\varepsilon/(2z)} \leq e^{\varepsilon/2} < 2,$$

provided $0 \leq \varepsilon \leq 1$. Therefore, the sum is bounded by $2z$, and thus

$$\left|\frac{Y(C)}{Z(C)} - 1\right| < \tilde{\varepsilon}\, c^{-1/z} \cdot 2z = \varepsilon.$$

On the other hand, since

$$Y(C) - Z(C) = \mathcal{E}(C) \sum_{i=0}^{z-1} Y(C)^{i/z} Z(C)^{(z-1-i)/z},$$

and all terms are nonnegative; keeping only the $i = 0$ term gives

$$|Y(C) - Z(C)| \geq |\mathcal{E}(C)|\, Z(C)^{(z-1)/z} \geq |\mathcal{E}(C)|\, c^{(z-1)/z}.$$

Therefore

$$\mathbb{P}\left(|\mathcal{E}(C)| \geq \frac{\tilde{\varepsilon}}{2}\right) \leq \mathbb{P}\left(\frac{1}{n}\left|L_{\mathcal{S}}(f_C) - L(f_C)\right| \geq \frac{\tilde{\varepsilon}}{2}\, c^{(z-1)/z}\right). \tag{5}$$

For real $z \geq 1$, the same conclusions follow by setting $a = (Y(C)/Z(C))^{1/z}$: if $|a - 1| \leq \varepsilon/(2z)$, Bernoulli's inequality and the exponential bound give

$$1 - \varepsilon/2 \leq (1 - \varepsilon/(2z))^z \leq a^z \leq (1 + \varepsilon/(2z))^z \leq e^{\varepsilon/2} < 1 + \varepsilon.$$

Also, $|a^z - 1| \geq |a - 1|$ for $a \geq 0$, which gives the second inequality. □

A.2 Concentration bounds of DPP

The following is a restatement of Theorem 1 in [BGSOT24] stated here for the reader's convenience:

Theorem A.2 (Theorem 1 in [BGSOT24])**.** *Let $\mathcal{S}$ be a DPP on a Polish space $\mathcal{X}$ with reference measure μ and Hermitian kernel K. Then there is a universal constant A such that for any bounded test function $\phi : \mathcal{X} \to \mathbb{R}$, and $\forall\, 0 \le \varepsilon \le 2A\frac{\mathrm{Var}\left[\sum_{x\in\mathcal{S}}\phi(x)\right]}{3\|\phi\|_\infty}$ we have*

$$\mathbb{P}\left(\left|\sum_{x\in\mathcal{S}}\phi(x) - \mathbb{E}\left[\sum_{x\in\mathcal{S}}\phi(x)\right]\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{\varepsilon^2}{4A\mathrm{Var}\left[\sum_{x\in\mathcal{S}}\phi(x)\right]}\right).$$

By the moment-generating-function estimate in [BGSOT24, Lemma 3], the result also holds with any positive variance upper bound V in both the exponent and the admissible range.

B Technical results related to Corollary 3.2

B.1 Proof of Corollary 3.2

Proof. Let $K \subset \mathbb{R}^d$ be a bounded set containing the data $\mathcal{P}$ and all candidate centers, let $\mathcal{C}$ be the family of nonempty subsets of K of size at most k, equipped with the Hausdorff metric ∂_H, and set $f_C(x) = \partial(x, C)^z$.

$$\partial_H(C, C') := \max\left\{\max_{y\in C}\min_{y'\in C'}\partial(y, y'),\ \max_{y'\in C'}\min_{y\in C}\partial(y, y')\right\}.$$

Then, for every $x \in \mathcal{X}$ and every $C, C' \in \mathcal{C}$,

$$\left|\partial(x, C) - \partial(x, C')\right| \le \partial_H(C, C').$$

Fix $x \in \mathcal{X}$ and $C, C' \in \mathcal{C}$. Let $y^* \in C$ satisfy $\partial(x, C) = \partial(x, y^*)$. By the definition of the Hausdorff metric, there exists $y' \in C'$ such that $\partial(y^*, y') \le \partial_H(C, C')$. Therefore

$$\begin{aligned}\partial(x, C') - \partial(x, C) &\le \partial(x, y') - \partial(x, y^*)\\ &\le \partial(y^*, y')\\ &\le \partial_H(C, C'),\end{aligned}$$

where we used the reverse triangle inequality in the second step. Interchanging the roles of C and C' gives $\partial(x, C) - \partial(x, C') \le \partial_H(C, C')$. Hence $\left|\partial(x, C) - \partial(x, C')\right| \le \partial_H(C, C')$. Consequently, $f_C(x) := \partial(x, C)^z$, hence

$$\left|f_C(x)^{1/z} - f_{C'}(x)^{1/z}\right| = \left|\partial(x, C) - \partial(x, C')\right| \le \partial_H(C, C').$$

Assumption **(A.2)** holds with $\partial_{\mathcal{C}} = \partial_H$. Since K is bounded in $\mathbb{R}^d$, there exists $C_K > 0$ such that $\mathcal{N}(K, u, \|\cdot\|) \le (C_K/u)^d$ for $0 < u \le 1$. Replacing each center of $C \in \mathcal{C}$ by its nearest point in a u-net $N_u \subseteq K$ yields $\hat{C}$ with $\partial_H(C, \hat{C}) \le u$, so

$$\mathcal{N}(\mathcal{C}, u, \partial_H) \le |N_u|^k \le \left(\frac{C_K}{u}\right)^{kd}.$$

Hence Assumption **(A.1)** holds with $\ell = kd$. We next verify the upper bound on ε required by Theorem 3.1. Taking the sup norms over the finite data domain, we have

$$M := \sup_{C\in\mathcal{C}}\|f_C\|_\infty = \sup_{C\in\mathcal{C},\, x\in\mathcal{P}}\partial(x, C)^z \le \mathrm{diam}(K)^z < \infty.$$

Choose the variance upper bound $V := c_V m^{-1-1/d} > 0$. For $0 < \varepsilon \le 1$, the theorem applies when $\varepsilon \le c_3 mV/M = (c_3 c_V/M)m^{-1/d}$, equivalently $m \le (c_3 c_V/M)^d \varepsilon^{-d}$, as required by the corollary's upper size restriction. Applying Theorem 3.1,

$$\mathbb{P}(\text{failure}) \le 2\exp\left(c_1 kd - kd\log\varepsilon - \frac{c_2\varepsilon^2}{V}\right).$$

For this probability to be at most δ, it suffices to have

$$c_1 kd + kd \log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{2}{\delta}\right) \le \frac{c_2 \varepsilon^2}{V}.$$

Using $V \le c_V m^{-(1+1/d)}$ and rearranging,

$$m^{\frac{d+1}{d}} \ge \frac{c_V}{c_2 \varepsilon^2}\left[c_1 kd + kd \log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{2}{\delta}\right)\right],$$

and therefore

$$m \ge \left(\frac{c_V}{c_2 \varepsilon^2}\left[c_1 kd + kd \log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{2}{\delta}\right)\right]\right)^{\frac{d}{d+1}}.$$

The $\tilde{\mathcal{O}}$ form follows by absorbing all constants and logarithmic factors. □

B.2 Discussion on Assumption (B.1)

Definition B.1 (Norm-tail subgaussian random vector)**.** Let $X \in \mathbb{R}^d$ have finite mean. We say that X is σ-subgaussian if

$$\mathbb{P}\left(\|X - \mathbb{E}[X]\|_2 > t\right) \le 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad \text{for all } t \ge 0.$$

Theorem B.2 (Uniform empirical lower bound for k-point z-cost)**.** *Let $X_1, \dots, X_n$ be i.i.d. samples from a probability measure μ on $\mathbb{R}^d$ with $\mathbb{E}_{X\sim\mu}\|X\|_2^z < \infty$. Let $k \ge 1$ and $z \ge 1$, we define the population k-point z-cost of μ by*

$$q_{k,z}(\mu) := \inf_{1 \le |\mathcal{C}| \le k} \mathbb{E}_{X\sim\mu}\left[\partial(X, \mathcal{C})^z\right].$$

Assume that $q := q_{k,z}(\mu) > 0$, then there exist constants

$$c = c(z, q, R) > 0 \qquad \text{and} \qquad n_0 = n_0(k, d, z, q, \mu) \ge 1,$$

such that for every $n \ge n_0$, we have

$$\mathbb{P}\left(\inf_{1 \le |\mathcal{C}| \le k} \frac{1}{n} \sum_{i=1}^{n} \partial(X_i, \mathcal{C})^z \ge \frac{q}{2}\right) \ge 1 - e^{-cn}.$$

Proof. Let $X \sim \mu$ and $V := \|X - \mathbb{E}[X]\|_2$.

Since $z \ge 1$ and $\mathbb{E}\|X\|_2^z < \infty$, the mean $\mathbb{E}[X]$ exists and $\mathbb{E}[V^z] < \infty$. By dominated convergence, $\mathbb{E}[V^z \mathbf{1}_{\{V>R\}}] \to 0$ as $R \to \infty$, so we can fix R with

$$R \ge q^{1/z} \qquad \text{and} \qquad 3^z\, \mathbb{E}\left[V^z \mathbf{1}_{\{V>R\}}\right] \le \frac{q}{4}. \tag{6}$$

Note that $q < \infty$ because $q \le \mathbb{E}\|X - \mathbb{E}X\|_2^z < \infty$.

For the choice of R satisfying (6), define the compact set K_R to be the ball of radius R centered at $\mathbb{E}[X]$. And define the compact set K_R^+ to be the ball of radius $R + q^{1/z}$ centered at $\mathbb{E}[X]$. Note that if $x \in K_R$ and $C \notin K_R^+$, then

$$\|x - C\|_2 \ge \|C - \mathbb{E}[X]\|_2 - \|x - \mathbb{E}[X]\|_2 > R + q^{1/z} - R = q^{1/z}.$$

Thus any in-ball sample $x \in K_R$ whose closest center lies outside K_R^+ already contributes more than q to the z-cost.

Let $\mathbf{C}$ be a k-tuple $\mathbf{C} = (C_1, \dots, C_k) \in (K_R^+)^k$, We note that repeated centers are allowed, so k-tuples parametrize all center sets of size at most k. Then we define the truncated loss

$$g_{\mathbf{C}}(x) := \partial(x, \mathbf{C})^z \mathbf{1}_{\{x \in K_R\}}.$$

We claim that

$$\mathbb{E}[g_{\mathbf{C}}(X)] \ge \frac{3q}{4} \qquad \text{for every } \mathbf{C} \in (K_R^+)^k. \tag{7}$$

Indeed, for $\mathbf{C} \in (K_R^+)^k$,

$$\mathbb{E}[g_{\mathbf{C}}(X)] = \mathbb{E}[\partial(X, \mathcal{C}_\Theta)^z] - \mathbb{E}\left[\partial(X, \mathbf{C})^z \mathbf{1}_{\{X \notin K_R\}}\right] \geq q - \mathbb{E}\left[\partial(X, \mathbf{C})^z \mathbf{1}_{\{X \notin K_R\}}\right],$$

where we used $\mathbb{E}[\partial(X, \mathbf{C})^z] \geq q$ which follows by the definition of q. Now we observe that on the event $\{X \notin K_R\}$, we have $V = \|X - \mathbb{E}[X]\| > R$. For every center $C_j \in K_R^+$, we have $\|C_j - \mathbb{E}[X]\|_2 \leq R + q^{1/z} \leq 2R$. Hence, on $\{X \notin K_R\}$, for $C_j \in K_R^+$

$$\|X - C_j\|_2 \leq \|X - \mathbb{E}[X]\|_2 + \|C_j - \mathbb{E}[X]\|_2 \leq V + 2R \leq 3V.$$

Therefore,

$$\partial(X, \mathbf{C})^z \mathbf{1}_{\{X \notin K_R\}} \leq 3^z V^z \mathbf{1}_{\{V > R\}}.$$

Taking expectation gives

$$\mathbb{E}\left[\partial(X, \mathbf{C})^z \mathbf{1}_{\{X \notin K_R\}}\right] \leq \mathbb{E}[3^z V^z \mathbf{1}_{\{V > R\}}] \leq \frac{q}{4}.$$

Therefore

$$\mathbb{E}[g_{\mathbf{C}}(X)] \geq q - \frac{q}{4} = \frac{3q}{4} \qquad \text{for every } \mathbf{C} \in (K_R^+)^k.$$

For $x \in K_R$ and $\mathbf{C} \in (K_R^+)^k$,

$$\partial(x, \mathbf{C}) \leq R + R + q^{1/z} \leq 3R.$$

Thus

$$0 \leq g_{\mathbf{C}}(x) \leq (3R)^z \qquad \text{for all } x \in \mathbb{R}^d.$$

As in the proof of Corollary 3.2, for every $x \in K_R$,

$$\left|\partial(x, \mathbf{C}) - \partial(x, \mathbf{C}')\right| \leq \partial_H(\mathbf{C}, \mathbf{C}')$$

Consider the function $\varphi : t \mapsto t^z$ on $[0, 3R]$, we have $|\varphi'(t)| \leq z(3R)^{z-1} := L_R$. Thus, we have

$$|g_{\mathbf{C}}(x) - g_{\mathbf{C}'}(x)| \leq L_R \partial_H(\mathbf{C}, \mathbf{C}') \qquad \text{for all } x \in \mathbb{R}^d.$$

Set $\delta := q/(16L_R)$. Let $\mathcal{N}_\delta$ be a δ-net of K_R^+. We may take $|\mathcal{N}_\delta| \leq \left(1 + \frac{4R}{\delta}\right)^d$. Therefore $\mathcal{N}_\delta^k$ is a δ-net of $(K_R^+)^k$, and $|\mathcal{N}_\delta^k| \leq \left(1 + \frac{4R}{\delta}\right)^{kd}$. For each fixed $\mathbf{C} \in \mathcal{N}_\delta^k$, Hoeffding's inequality gives

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n g_{\mathbf{C}}(X_i) < \mathbb{E}[g_{\mathbf{C}}(X)] - \frac{q}{8}\right) \leq \exp\left(-\frac{nq^2}{32(3R)^{2z}}\right).$$

Taking a union bound over $\mathcal{N}_\delta^k$, there is an event $\mathcal{E}$ with probability at least

$$1 - \left(1 + \frac{4R}{\delta}\right)^{kd} \exp\left(-\frac{nq^2}{32(3R)^{2z}}\right),$$

such that on $\mathcal{E}$ we have

$$\frac{1}{n}\sum_{i=1}^n g_{\mathbf{C}}(X_i) \geq \mathbb{E}[g_{\mathbf{C}}(X)] - \frac{q}{8}, \quad \forall \mathbf{C} \in \mathcal{N}_\delta^k.$$

On this event, fix an arbitrary $\mathbf{C} \in (K_R^+)^k$ and choose $\mathbf{C}' \in \mathcal{N}_\delta^k$ such that $\partial_H(\mathbf{C}, \mathbf{C}') \leq \delta$. By the Lipschitz bound, we have

$$\frac{1}{n}\sum_{i=1}^n g_{\mathbf{C}}(X_i) \geq \frac{1}{n}\sum_{i=1}^n g_{\mathbf{C}'}(X_i) - L_R\delta,$$

and $\mathbb{E}[g_{\mathbf{C}'}(X)] \geq \mathbb{E}[g_{\mathbf{C}}(X)] - L_R\delta$. Therefore,

$$\frac{1}{n}\sum_{i=1}^n g_{\mathbf{C}}(X_i) \geq \mathbb{E}[g_{\mathbf{C}}(X)] - \frac{q}{8} - 2L_R\delta.$$

Since $\delta = q/(16L_R)$, $2L_R\delta = q/8$. Thus

$$\frac{1}{n}\sum_{i=1}^{n} g_{\mathbf{C}}(X_i) \geq \mathbb{E}[g_{\mathbf{C}}(X)] - \frac{q}{4}.$$

By (7), we have $\mathbb{E}[g_{\mathbf{C}}(X)] \geq 3q/4$. Thus, on the event $\mathcal{E}$:

$$\frac{1}{n}\sum_{i=1}^{n} g_{\mathbf{C}}(X_i) \geq \frac{q}{2} \qquad \text{for every } \mathbf{C} \in (K_R^+)^k. \tag{8}$$

It remains to pass from centers in K_R^+ to arbitrary centers in $\mathbb{R}^d$. Let $\Pi : \mathbb{R}^d \to K_R^+$ be the Euclidean projection onto the closed convex set K_R^+. Since $K_R \subseteq K_R^+$, projection can only decrease distance from points $x \in K_R$, i.e.

$$\|x - \Pi(C)\|_2 \leq \|x - C\|_2 \qquad \text{for all } x \in K_R,\ C \in \mathbb{R}^d.$$

Take arbitrary $\mathbf{C} = (C_1, \ldots, C_k) \in (\mathbb{R}^d)^k$, and define $\Pi\mathbf{C} := (\Pi(C_1), \ldots, \Pi(C_k)) \in (K_R^+)^k$. For $x \in K_R$, we then have $\partial(x, \Pi\mathbf{C}) \leq \partial(x, \mathbf{C})$. For $x \notin K_R$, note that $g_{\Pi\mathbf{C}}(x) = 0$. Therefore, for every $x \in \mathbb{R}^d$,

$$\partial(x, \mathbf{C})^z \geq g_{\Pi\mathbf{C}}(x).$$

Consequently,

$$\frac{1}{n}\sum_{i=1}^{n} \partial(X_i, \mathbf{C})^z \geq \frac{1}{n}\sum_{i=1}^{n} g_{\Pi\mathbf{C}}(X_i).$$

Since $\Pi\mathbf{C} \in (K_R^+)^k$, we deduce by equation (8), that on the event $\mathcal{E}$:

$$\frac{1}{n}\sum_{i=1}^{n} \partial(X_i, \mathbf{C})^z \geq \frac{q}{2}.$$

Since $\mathbf{C} \in (\mathbb{R}^d)^k$ is arbitrary, we deduce that on the event $\mathcal{E}$:

$$\inf_{1 \leq |\mathbf{C}| \leq k} \frac{1}{n}\sum_{i=1}^{n} \partial(X_i, \mathbf{C})^z \geq \frac{q}{2}.$$

Thus, for every n,

$$\mathbb{P}\left(\inf_{1 \leq |\mathbf{C}| \leq k} \frac{1}{n}\sum_{i=1}^{n} \partial(X_i, \mathbf{C})^z \geq \frac{q}{2}\right) \geq 1 - \left(1 + \frac{4R}{\delta}\right)^{kd} \exp\left(-\frac{nq^2}{32(3R)^{2z}}\right).$$

Define

$$n_0 := \left\lceil \frac{64(3R)^{2z}}{q^2} kd \log\left(1 + \frac{4R}{\delta}\right) \right\rceil.$$

If $n \geq n_0$, then

$$kd \log\left(1 + \frac{4R}{\delta}\right) \leq \frac{nq^2}{64(3R)^{2z}}.$$

Hence

$$\left(1 + \frac{4R}{\delta}\right)^{kd} \exp\left(-\frac{nq^2}{32(3R)^{2z}}\right) \leq \exp\left(-\frac{nq^2}{64(3R)^{2z}}\right).$$

Therefore, the theorem holds with

$$c := \frac{q^2}{64(3R)^{2z}}.$$

□

Remark 4. When μ is σ-subgaussian, (6) holds for an explicit R as follows (and hence we can explicitly write c and n_0). In the subgaussian case,

$$\mathbb{P}(V > t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad \text{for all } t \ge 0.$$

Using the tail-integral formula, $\mathbb{E}[V^{2z}] = \int_0^\infty 2z\, t^{2z-1}\mathbb{P}(V > t)\, dt$, we get

$$\mathbb{E}[V^{2z}] \le 2\int_0^\infty 2z\, t^{2z-1}\exp\left(-\frac{t^2}{2\sigma^2}\right) dt = 2(2\sigma^2)^z\Gamma(z+1) =: A_{z,\sigma}.$$

Let $R := \max\{q^{1/z}, (\frac{4\cdot 3^z A_{z,\sigma}}{q})^{1/z}\}$. Then $R \ge q^{1/z}$, and by Markov's inequality,

$$\mathbb{E}\left[V^z \mathbf{1}_{\{V>R\}}\right] \le \frac{\mathbb{E}[V^{2z}]}{R^z} \le \frac{A_{z,\sigma}}{R^z}.$$

Therefore, for this choice of R we have,

$$3^z\mathbb{E}\left[V^z \mathbf{1}_{\{V>R\}}\right] \le \frac{q}{4},$$

We now show that no moment condition is needed for Assumption **(B.1)** itself.

Proposition B.3 (Nondegeneracy without moment conditions)**.** *Let μ be a probability measure on $\mathbb{R}^d$, $k \ge 1$, $z \ge 1$. Suppose there are $k+1$ closed balls $B_1, \dots, B_{k+1}$ with $\partial(B_i, B_j) \ge \Delta > 0$ for $i \ne j$ and $\mu(B_i) \ge p > 0$ for all i. If $X_1, \dots, X_n$ are i.i.d. from μ, then for every $n \ge 1$,*

$$\mathbb{P}\left(\inf_{1\le|\mathcal{C}|\le k} \frac{1}{n}\sum_{i=1}^n \partial(X_i, \mathcal{C})^z \;\ge\; \left(\frac{\Delta}{2}\right)^z \frac{p}{2}\right) \;\ge\; 1 - (k+1)e^{-np/8}.$$

Proof. For sets A, B and a point c, $\partial(A, B) \le \partial(A, c) + \partial(c, B)$. Hence, a point c cannot be at distance less than $\Delta/2$ from two different balls, since otherwise those two balls would be at distance less than Δ. Let $\mathcal{C}$ be a set of at most k centers. Each center is at distance strictly less than $\Delta/2$ from at most one ball, and there are $k+1$ balls, so some ball B_{i^*} satisfies $\partial(c, B_{i^*}) \ge \Delta/2$ for every $c \in \mathcal{C}$. Every $x \in B_{i^*}$ then has $\partial(x, \mathcal{C}) \ge \Delta/2$. Writing $N_i := |\{t \le n : X_t \in B_i\}|$,

$$\frac{1}{n}\sum_{t=1}^n \partial(X_t, \mathcal{C})^z \;\ge\; \left(\frac{\Delta}{2}\right)^z \frac{N_{i^*}}{n} \;\ge\; \left(\frac{\Delta}{2}\right)^z \min_{1\le i\le k+1} \frac{N_i}{n},$$

and the right-hand side does not depend on $\mathcal{C}$. Each N_i is binomial with mean $n\mu(B_i) \ge np$, so the multiplicative Chernoff bound $\mathbb{P}(N_i \le \frac{1}{2}\mathbb{E}N_i) \le e^{-\mathbb{E}N_i/8}$ gives $\mathbb{P}(N_i < np/2) \le e^{-np/8}$. A union bound over the $k+1$ balls concludes. □

This is essentially minimal: if μ is supported on at most k points, placing the centers on these points gives zero cost. Examples covered by these results include any law with a finite z-th moment (for k-means, any law with finite variance) and laws with no finite moments, such as the multivariate Cauchy distribution, mixtures, and discrete laws, as long as the support is not concentrated on at most k points.

Remark 5 (**On Assumption (B.1)**)**.** For k-means with positive optimal cost, we can obtain a positive lower bound on OPT_k/n before sampling using k-means++ [AV07]. Its cost A satisfies $\mathbb{E}[A] \le \beta\,\mathrm{OPT}_k$, where $\beta = 5(\ln k + 2)$ [MRS20]. By Markov's inequality, $A/(2\beta n) \le \mathrm{OPT}_k/n$ with probability at least $1/2$. Thus, setting $c = A_{\min}/(2\beta n)$, where $A_{\min}$ is the lowest cost from r independent runs, satisfies Assumption **(B.1)** with probability at least $1 - 2^{-r}$. The variance condition depends jointly on the kernel, data distribution, and loss class; it holds for the discretized OPE kernel under the hypotheses of Remark 2. If the data lie in a lower-dimensional affine subspace, the kernel can be constructed in PCA coordinates spanning that subspace. Our experiments also use PCA for dimensionality reduction.

Remark 6 (**Bounded centers and scope**)**.** Restricting centers to a bounded region does not weaken the known lower bound. The $\Omega(k\varepsilon^{-2})$ bound for k-means in [CALSS22, Theorem 5] uses the dataset $\{e_1, \dots, e_{d'}\} \subset \mathbb{R}^{2d'}$, where $d' = k/(36\varepsilon^2)$, and center sets consisting of unit vectors.

It therefore applies to centers in the unit ball and, in particular, to the weighted-subset coresets considered here. However, this hard instance has dimension $\Theta(k\varepsilon^{-2})$, rather than fixed.

Extending our guarantee to unrestricted centers is natural, since distant centers induce nearly constant costs on the data and the DPP approximately preserves the total weight. However, this requires uniform control of the variance and an admissible concentration range for the rescaled losses, which we leave to future work.

Finally, our improvement is primarily relevant in fixed or intrinsically low dimension. Standard JohnsonLindenstrauss guarantees for preserving clustering costs give target dimensions of order ε^{-2} up to logarithmic factors, in which regime the exponent $2d/(d+1)$ approaches 2. The explicit dimension dependence is given in Corollary 3.2.

C Kernel constructions

DPP-OPE. Jacobi weights fitted coordinatewise by matching marginal moments define a reference density q on $[-1,1]^d$. With $p_1,\dots,p_m$ the first m multivariate orthonormal polynomials for q and $\tilde{\gamma}$ a kernel density estimate of the data, the features $P_{ij} = \sqrt{q(x_i)/\tilde{\gamma}(x_i)}\, p_j(x_i)$ span the rank-m projection kernel of [BGSOT24, Example 5], computed in $O(nm^2)$.

DPP-Vandermonde. The first m graded monomials evaluated on $\mathcal{P}$ form a matrix F; with an economy QR factorization $F = QR$, the kernel is $\mathbf{K} = QQ^\top$ [TBA19, Algorithm 2]. No density estimate is needed.

DPP-Haar. The cube $[-1,1]^d$ is split into 2^{jd} dyadic cells with $j = \mathrm{round}(\log_2(m)/d)$. The cell-projection kernel selects one uniformly chosen point in each non-empty cell Q, with weight $|Q \cap \mathcal{P}|/n$; if more than m cells are non-empty, m of them are selected uniformly and the weights are divided by the inclusion probability.

Table 2: Kernel selection criteria for the three DPP samplers. Only the variance rate of the pure OPE kernel is proven; the statements for the other two kernels are empirical.

	DPP-OPE	**DPP-Vandermonde**	**DPP-Haar**
Construction	discretized multivariate OPE, density-corrected	graded monomials, orthonormalized by QR	dyadic cells, one point per non-empty cell
Parameters	Jacobi weights (a_j, b_j), KDE bandwidth, 10% uniform fallback	none	resolution j
Dimension	$d \le 6$	$d \le 16$	any d
Cost	$O(nm^2)$ to build and per exact draw	$O(nm^2)$ to build and per exact draw	$O(nd)$ to build, $O(m)$ per draw
Theoretical status	rate $O(m^{-1-1/d})$ proved for the pure kernel [BGSOT24]	empirical	empirical; a related property is proved for wavelet DPPs [TGBG26]
Intended role	smooth, low-dimensional data	smooth data, no density estimate	spatially clustered data, large n

D Running times

Per coreset, Uniform costs $O(m)$, Stratified $O(nm)$, Sensitivity $O(nk+nm)$, DPP-Haar is near-linear in n, and DPP-OPE and DPP-Vandermonde cost $O(nm^2)$ with the exact sampler of `DPPy`. The kernel is built once and reused across draws, and faster samplers for repeated draws [GKMV19] can be used as drop-in replacements. Table 3 reports wall-clock times on three point sets of $n \approx 10^4$; each entry is one call producing $R = 150$ coresets at $k = 5$.

Table 3: Wall-clock seconds for one call of each sampler, producing $R = 150$ coresets of size m ($k = 5$; divide by 150 for one coreset), measured in-process with `time.perf_counter` on point sets of $n \approx 10^4$. Σ: sum over the six sizes shown. MNIST-PCA4 (MNIST projected onto its top four principal components) is used for timing only.

Sampler	$m = 25$	$m = 100$	$m = 225$	$m = 400$	$m = 576$	$m = 729$	Σ
BIRCH ($d = 2$)							
DPP-Haar	0.081	0.099	0.124	0.124	0.152	0.153	0.733
DPP-OPE	1.1	7.9	48.5	166	354	560	1,137
DPP-Vandermonde	1.3	12.3	74.1	250	512	841	1,692
Uniform	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sensitivity	0.007	0.008	0.009	0.011	0.013	0.014	0.062
Stratified	0.011	0.012	0.012	0.012	0.013	0.013	0.072
HLW	0.030	0.031	0.032	0.034	0.035	0.037	0.198
3D Road Network ($d = 3$)							
DPP-Haar	0.018	0.090	0.156	0.156	0.157	0.157	0.734
DPP-OPE	1.1	7.9	46.9	168	364	568	1,156
DPP-Vandermonde	1.3	12.6	73.4	253	508	914	1,762
Uniform	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sensitivity	0.007	0.008	0.009	0.011	0.013	0.014	0.062
Stratified	0.012	0.012	0.012	0.012	0.013	0.013	0.074
HLW	0.030	0.031	0.032	0.034	0.036	0.037	0.201
MNIST-PCA4 ($d = 4$)							
DPP-Haar	0.085	0.055	0.148	0.152	0.149	0.149	0.738
DPP-OPE	1.1	8.0	49.1	170	364	567	1,159
DPP-Vandermonde	1.3	12.4	77.1	254	496	826	1,667
Uniform	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Sensitivity	0.007	0.008	0.010	0.011	0.013	0.014	0.064
Stratified	0.013	0.012	0.013	0.013	0.014	0.014	0.079
HLW	0.032	0.032	0.034	0.036	0.037	0.039	0.211

E On Polish spaces.

A Polish space is a topological space whose topology can be induced by a complete separable metric. This class is broad enough for many standard spaces used in analysis, probability, statistics, and data science. In particular, every compact metrizable space is Polish [Gar17, Corollary 3.3.2]. Therefore, whenever the data domain is modeled as a compact metric space, the assumption that the underlying space is Polish adds no extra restriction.

This includes many common finite-dimensional data domains. For example, the cube $[0, 1]^d$ is compact and metric, hence Polish. An $H \times W$ normalized RGB image space can be identified with $[0, 1]^{3HW}$ so it is also a compact metric space and hence Polish.

In fact, any finite label space with the discrete metric is a compact metric space and hence Polish.

Similarly, the probability simplex

$$\Delta_{k-1} = \left\{ p \in [0, 1]^k : \sum_{i=1}^{k} p_i = 1 \right\}$$

is a closed subset of the compact cube $[0, 1]^k$, so it is compact metric and hence Polish. Thus, the Polish-space framework is general enough to cover many compact metric spaces that arise naturally in data science applications.

F The Average Squared Distance Objective

Consider an alternative but closely related objective to the standard k-means clustering. Instead of the minimum distance to the nearest center, define the *average squared distance* to a k-center set

$C = \{c_1, \dots, c_k\} \subset \mathbb{R}^d$ as

$$f_C(x) = \frac{1}{k} \sum_{c \in C} \|x - c\|^2. \tag{9}$$

Proposition F.1. *Let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{F} = \{f_C : C \subset \mathbb{R}^d, |C| = k\}$ be the class of functions defined in (9). The linear span of $\mathcal{F}$ has finite dimension $D = d + 2$, independently of k, thereby satisfying Assumption* (A.1) *of [BGSOT24].*

Proof. Fix any center set $C = \{c_1, \dots, c_k\}$. Expanding and averaging the squared distances gives

$$f_C(x) = \|x\|^2 - 2\bar{c}^\top x + R^2,$$

where $\bar{c} = \frac{1}{k}\sum_{i=1}^k c_i$ is the centroid and $R^2 = \frac{1}{k}\sum_{i=1}^k \|c_i\|^2$ is the mean squared norm of the centers. Hence every $f_C \in \mathcal{F}$ is a linear combination of the $d + 2$ linearly independent functions $h_0(x) = \|x\|^2$, $h_j(x) = x_j$ for $j = 1, \dots, d$, and $h_{d+1}(x) = 1$. Translating a fixed k-center configuration with centroid zero by $\pm t$ shows that constants and all coordinate functions belong to the span; the expansion then gives $\|x\|^2$ as well. Therefore $\dim(\mathrm{span}_{\mathbb{R}}(\mathcal{F})) = d + 2 < \infty$. □

Under the remaining regularity and nondegeneracy hypotheses of the cited result, since $\mathcal{F}$ lies in a fixed $(d + 2)$-dimensional function space, the first part of Theorem 6 from [BGSOT24] applies. With high probability over the dataset $P \sim \mu^n$, the DPP-sampled weighted coreset $\mathcal{S}$ of expected size m satisfies

$$\mathbb{P}_{\mathcal{S}}\left(\exists\, f \in \mathcal{F} : \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \geq \varepsilon\right) \leq 2\exp\!\big(6(d+2) - C'\varepsilon^2 m^{1+1/d}\big)$$

for $\varepsilon = O(m^{-1/d})$, where $C' > 0$ does not depend on m, n. For $d > 1$ Remark 6.1 of [BGSOT24] then gives, with probability tending to 1,

$$\forall\, f \in \mathcal{F}, \qquad \left|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\right| \leq m^{-(1/2+1/(2d))+o(1)}.$$

Solving the displayed concentration bound at fixed confidence gives the sufficient coreset size

$$m = \tilde{\mathcal{O}}\Big(\varepsilon^{-\frac{2d}{d+1}}\Big),$$

hiding constants and lower-order logarithmic factors.

G Supplementary Benchmark Plots

This appendix collects benchmark diagnostics for nine of the datasets evaluated in Section 4: the six synthetic datasets (Uniform 2D, Trimodal 2D, Gaussian Blobs 2D, Gaussian Blobs 3D, Anisotropic Gaussians 2D, BIRCH Benchmark) and the three real-world datasets (3D Road Network, UCI Adult PCA-2, Twitter GPS). For each dataset, we show three complementary metrics alongside the average relative error reported in the main text. Together, these figures allow the reader to distinguish variance effects from slope effects, and to compare average-case against worst-case behavior across methods and coreset sizes.

Metric definitions. Throughout, we fix $N_{\mathrm{REP}} = 150$ independent coreset draws $\mathcal{S}^{(1)}, \dots, \mathcal{S}^{(N_{\mathrm{REP}})}$ and a fixed family of $N_{\mathrm{QUERY}} = 150$ random query center sets $C_1, \dots, C_{N_{\mathrm{QUERY}}}$ drawn once per dataset-k pair.

- **Supremum variance** ("Sup. variance"). For each query C_j, let $\widehat{\sigma}_j^2(m) := \frac{1}{N_{\mathrm{REP}}-1}\sum_{r=1}^{N_{\mathrm{REP}}}\big(L_{\mathcal{S}^{(r)}}(C_j) - \overline{L}(C_j)\big)^2$ be the empirical variance of the coreset estimator over repetitions, where $\overline{L}(C_j)$ is its sample mean. The supremum variance is then

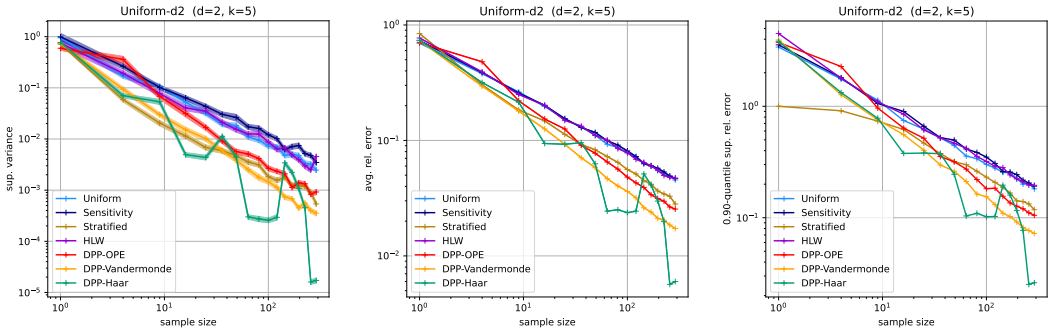
  $$\sup_{j \in [N_{\mathrm{QUERY}}]} \widehat{\sigma}_j^2(m).$$

  With the experimental mean-loss convention, this finite-query variance maximum is the empirical analog of $V = \sup_C \mathrm{Var}[n^{-1}L_{\mathcal{S}}(f_C)]$ from Theorem 3.1.

- **Average relative error** ("Avg. rel. error"). This is $\overline{\mathrm{RelErr}}(m)$ as defined in Section 4, reproduced here for ease of comparison.
- 0.9**-quantile supremum relative error** ($Q_{\mathcal{S}}(0.9)$). For each repetition r, let $E^{(r)}(m) := \sup_{j \in [N_{\mathrm{QUERY}}]} \frac{|L_{\mathcal{S}^{(r)}}(C_j) - L(C_j)|}{L(C_j)}$ be the worst-case (supremum) relative error over all queries. The metric $Q_{\mathcal{S}}(0.9)$ is the 0.9-quantile of $E^{(1)}(m), \ldots, E^{(N_{\mathrm{REP}})}(m)$:

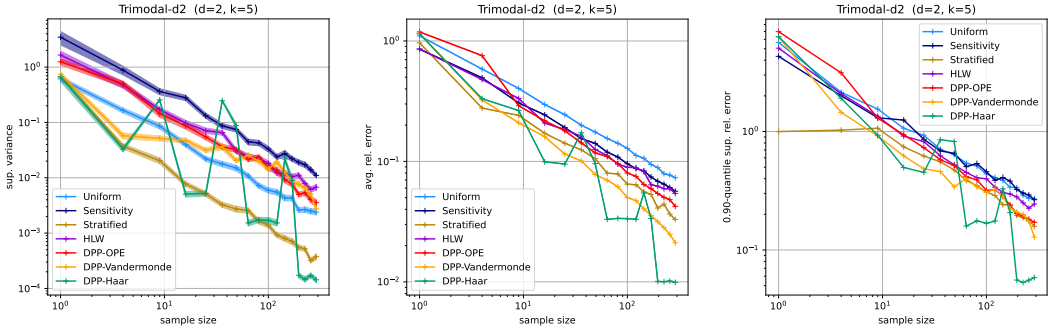
  $$Q_{\mathcal{S}}(0.9) := \mathrm{quantile}_{0.9}\big(E^{(1)}(m), \ldots, E^{(N_{\mathrm{REP}})}(m)\big).$$

  This robustly captures worst-query performance: it reports the supremum error that is exceeded in only 10% of independent coreset draws, and is thus more stable than the hard maximum over repetitions.



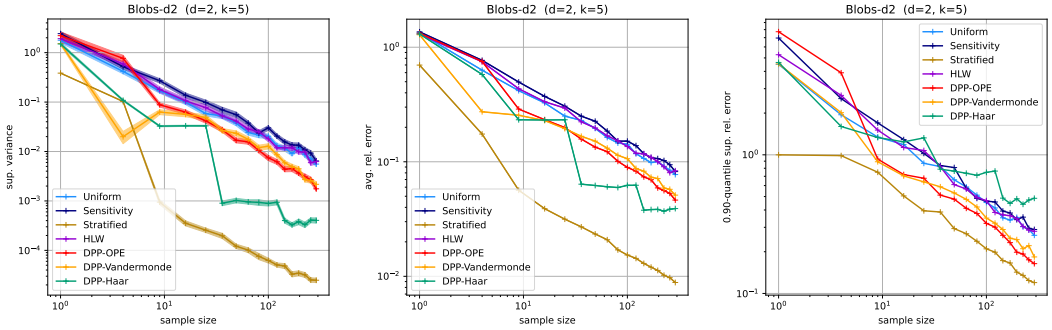
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 3: Supplementary metrics for **Uniform 2D** ($d = 2$, $k = 5$).



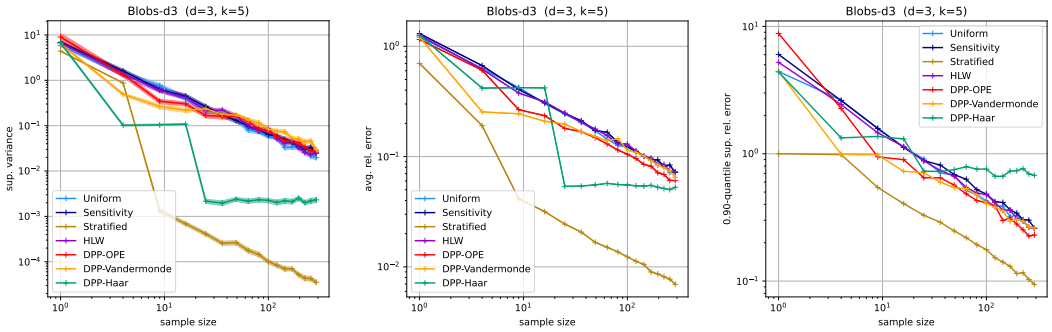
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 4: Supplementary metrics for **Trimodal 2D** ($d = 2$, $k = 5$).



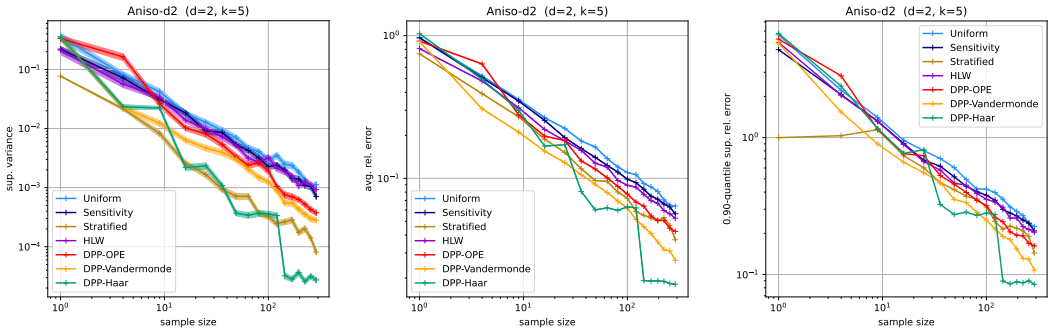
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 5: Supplementary metrics for **Gaussian Blobs 2D** ($d = 2$, $k = 5$).



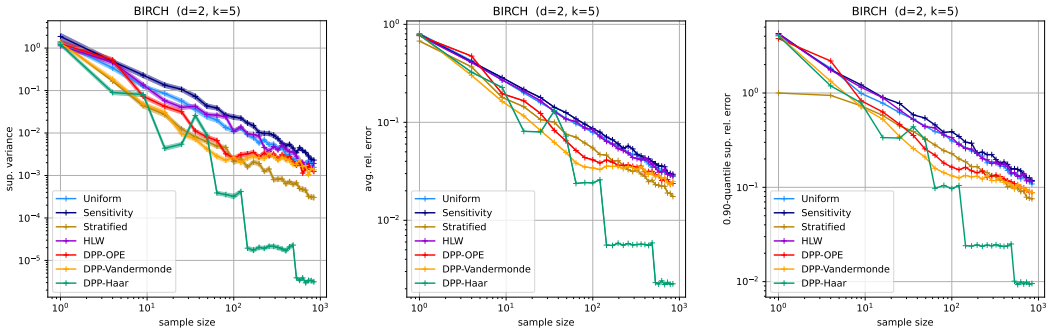
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 6: Supplementary metrics for **Gaussian Blobs 3D** ($d = 3$, $k = 5$).



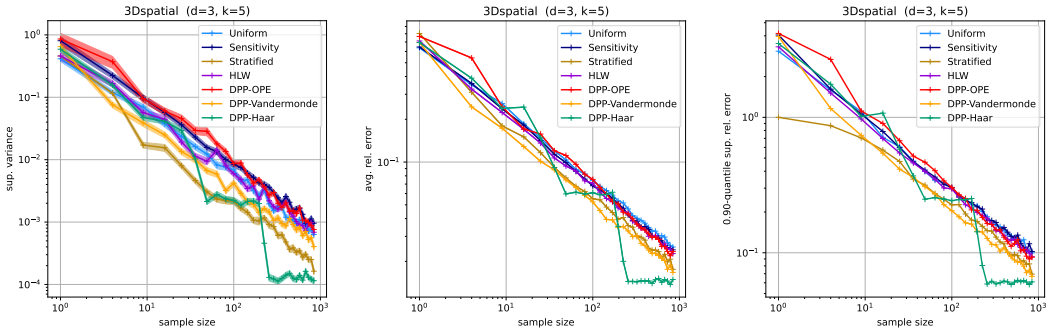
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 7: Supplementary metrics for **Anisotropic Gaussians 2D** ($d = 2$, $k = 5$).



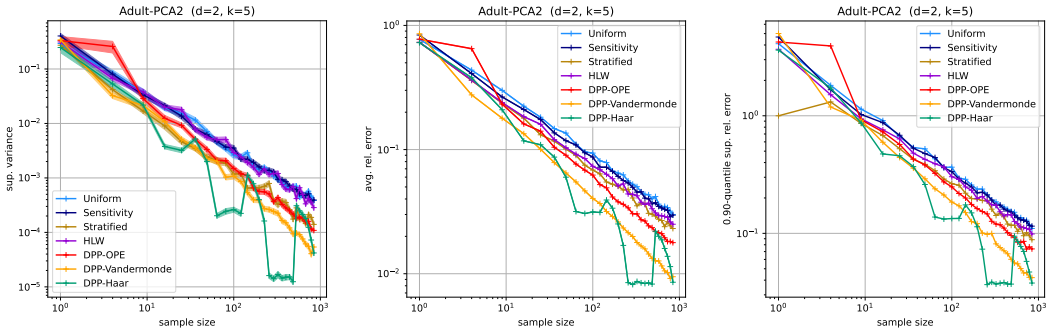
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 8: Supplementary metrics for **BIRCH Benchmark** ($d = 2$, $k = 5$).



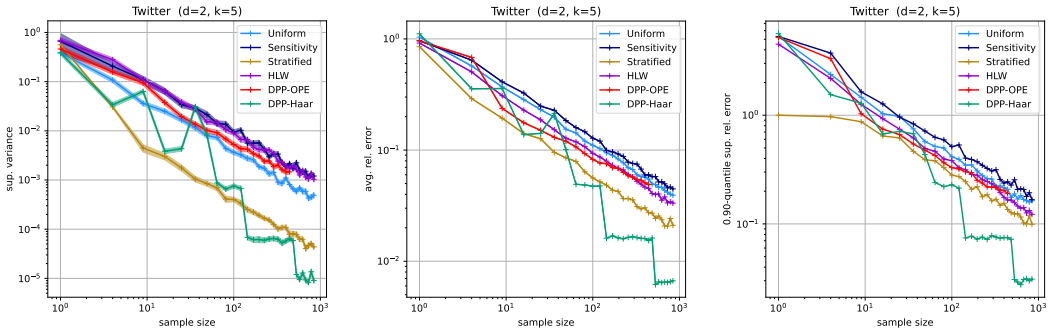
Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 9: Supplementary metrics for **3D Road Network** ($d = 3$, $k = 5$).



Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 10: Supplementary metrics for **UCI Adult (PCA-2)** ($d = 2$, $k = 5$).



Sup. variance Avg. rel. error $Q_{\mathcal{S}}(0.9)$

Figure 11: Supplementary metrics for **Twitter GPS** ($d = 2$, $k = 5$).

H License

- **Twitter GPS dataset** [CGS18]. Collected and publicly released by Chan, Guerquin, and Sozio alongside their WWW 2018 paper "Fully Dynamic k-Center Clustering" (ACM DOI: 10.1145/3178876.3186124; HAL preprint: hal-02315471), available at `https://github.com/fe6Bc5R4JvLkFkSeExHM/k-center`. The dataset contains only GPS coordinates (longitude, latitude) and timestamps; no tweet text is included, in accordance with Twitter's privacy policy. The paper and accompanying dataset are published under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license, as stated on the paper's title page.
- **3D Road Network dataset** [Kau13]. Obtained from the UCI Machine Learning Repository under **CC BY 4.0**.
- **UCI Adult dataset** [BK96]. Obtained from the UCI Machine Learning Repository (DOI: 10.24432/C5XW20) under **CC BY 4.0**.
- **DIGIT dataset** [AK98]. Optical Recognition of Handwritten Digits, obtained through `scikit-learn` (a copy of the test set of the UCI dataset, DOI: 10.24432/C50P49), under **CC BY 4.0**.
- **BIRCH Benchmark** [ZRL97]. Synthetic benchmark generated by our own code following the procedure described in [ZRL97]; no third-party data are used.
- **DPPy library** [GPBV19]. Used for `MultivariateJacobiOPE`; released under the **MIT License** (`https://github.com/guilgautier/DPPy`).
- **scikit-learn** [PVG+11]. Used for PCA and baseline clustering utilities; released under the **BSD 3-Clause License** (`https://github.com/scikit-learn/scikit-learn`).
- All remaining synthetic datasets (Uniform, Trimodal, Gaussian Blobs, Anisotropic Gaussians) are generated entirely by our own code and involve no third-party data or assets.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We have provided all the relevant proofs and experimental results in Sections 3 and 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: As such, there are no particular limitations of our work, but we have made some beyond-worst-case assumptions which we explain in detail in Section 3 and also highlighted in all the theorem statements.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The proof sketches are in section 3 with the missing details of the proof deferred to the Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The details of the experimental setup are provided in Section 4. Additionally, all the code and references to the dataset (publicly available) used are submitted as supplementary materials (as a zip file).

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: All the code and references to the dataset (publicly available) used are submitted as supplementary materials. We have also provided a README.md containing a detail instruction in the submitted zip file.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The details are provided in Section 4.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Yes we provide suitable error bars in the Appendix G

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The details of the experimental setup, including the machine, are provided in Section 4; wall-clock times of all samplers are reported in Appendix D.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our results do not violate any of the code of ethics provided by NeurIPS.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our work is primarily theoretical in nature and, as such, does not have any significant social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets, software libraries, and algorithmic baselines used in this paper are in Appendix H.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: As such, there are no new assets, but we have provided all the codes used for our empirical evaluation along with proper documentation as supplementary material.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: We have used LLM only for editing and basic code assistance, and as per the Handbook, the use of spell checkers and grammar suggestions, aid for editing purposes, and basic code assistance does not need to be documented.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.