Near-Optimal Coresets of Kernel Density Estimates

Jeff M. Phillips · Wai Ming Tai

Received: 22 June 2018 / Revised: 14 August 2019 / Accepted: 27 August 2019 / Published online: 25 September 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
We construct near-optimal coresets for kernel density estimates for points in \( \mathbb{R}^d \) when the kernel is positive definite. Specifically we provide a polynomial time construction for a coreset of size \( O\left(\sqrt{d/\varepsilon} \cdot \sqrt{\log 1/\varepsilon}\right) \), and we show a near-matching lower bound of size \( \Omega\left(\min\{\sqrt{d/\varepsilon}, 1/\varepsilon^2\}\right) \). When \( d \geq 1/\varepsilon^2 \), it is known that the size of coreset can be \( O\left(1/\varepsilon^2\right) \). The upper bound is a polynomial-in-\( (1/\varepsilon) \) improvement when \( d \in [3, 1/\varepsilon^2) \) and the lower bound is the first known lower bound to depend on \( d \) for this problem. Moreover, the upper bound restriction that the kernel is positive definite is significant in that it applies to a wide variety of kernels, specifically those most important for machine learning. This includes kernels for information distances and the sinc kernel which can be negative.

Keywords Coreset · Kernel density estimate · Discrepancy theory

1 Introduction

Kernel density estimates are pervasive objects in data analysis. They are the classic way to estimate a continuous distribution from a finite sample of points [46,47]. With some points negatively weighted, they are the prediction function in kernel SVM classifiers [44]. They are the core of many robust topological reconstruction approaches [7,17,41]; and they arise in many other applications including mode estimation [1], outlier detection [45], regression [16], and clustering [42].
Consider a dataset $P \subset \mathbb{R}^d$ of size $n$, and a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, for instance the Gaussian kernel $K(x, p) = \exp(-\alpha^2 \|x - p\|^2)$ with $1/\alpha$ as a bandwidth parameter. Then a kernel density estimate is defined at any point $x \in \mathbb{R}^d$ as $\text{KDE}_P(x) = \frac{1}{n} \sum_{p \in P} K(x, p)$.

Given that it takes $O(n)$ time to evaluate $\text{KDE}_P$, and that data sets are growing to massive sizes, in order to continue to use these powerful modeling objects, a common approach is to replace $P$ with a much smaller data set $Q$ so that $\text{KDE}_Q$ approximates $\text{KDE}_P$. While statisticians have classically studied various sorts of average deviations ($L_2$ [46,47] or $L_1$ error [13]), for most modern data modeling purposes, a worst-case $L_\infty$ is more relevant (e.g., for preserving classification margins [44], density estimates [51], topology [41], and hypothesis testing on distributions [22]). Specifically this error guarantee preserves

$$\|\text{KDE}_P - \text{KDE}_Q\|_\infty = \sup_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon.$$  

We call such a set $Q$ an $\varepsilon$-KDE coreset of $P$. In this paper we study how small can $Q$ be as a function of error $\varepsilon$, dimension $d$, and properties of the kernels.\footnote{This combines results published in SOCG 2018 [39] and SODA 2018 [38].}

1.1 Background on Kernels and Related Coresets

Traditionally the approximate set $Q$ has been considered to be constructed as a random sample of $P$ [28,46,47], sometimes known as a Nyström approximation [14]. However, in the last decade, a slew of data-aware approaches have been developed that can obtain a set $Q$ with the same $L_\infty$ error guarantee, but with considerably smaller size.

To describe the random sample results and the data-aware approaches, we first need to be more specific about the properties of the kernel functions. We start with positive definite kernels, the central class required for most machine learning approaches to work [25].

**Positive definite kernels** Consider a kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ defined over some domain $\mathcal{D}$ (often $\mathbb{R}^d$). It is called a positive definite kernel if any $m$ points $x_1, x_2, \ldots, x_m \in \mathcal{D}$ are used to define an $m \times m$ Gram matrix $G$ so that each $i$, $j$ entry is $G_{i,j} = K(x_i, x_j)$, and the matrix $G$ is positive definite. Recall, a symmetric matrix $G$ is positive definite if any vector $z \in \mathbb{R}^n$ that is not all zeros satisfies $z^T G z > 0$. Moreover, a positive definite matrix $G$ can always be decomposed as a product $H^T H$ with real-valued matrix $H$.

Also if $K$ is positive definite, it is said to have the reproducing property [2,50]. This implies that $K(p, x)$ is an inner product in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$. Specifically, there exists a lifting map $\phi_K : \mathbb{R}^d \to \mathcal{H}_K$ where $\phi_K(x) = K(x, \cdot)$ and such that $K(p, x) = \langle \phi_K(p), \phi_K(x) \rangle_{\mathcal{H}_K}$. Moreover the entire set $P$ can be represented as $\Phi_K(P) = \sum_{p \in P} \phi_K(p)$, which is a single element of $\mathcal{H}_K$ and has the norm $\|\Phi_K(P)\|_{\mathcal{H}_K} = \sqrt{\sum_{p \in P} \sum_{p' \in P} K(p, p')}$. A single point $x \in \mathbb{R}^d$ also has the
norm $\|\phi_K(x)\|_{\mathcal{H}_K} = \sqrt{K(x, x)} = 1$ in this space. A kernel mean of a point set $P$ and a reproducing kernel $K$ is defined as $\hat{\mu}_P = \frac{1}{|P|} \sum_{p \in P} \phi_K(p) = \Phi_K(P)/|P| \in \mathcal{H}_K$.

### Example positive definite kernels

| Kernel | Formula | Domain |
|--------|---------|--------|
| Gaussian | $\exp(-\alpha^2\|x - p\|^2)$ | $\mathbb{R}^d$ |
| Laplacian | $\exp(-\alpha\|x - p\|)$ | $\mathbb{R}^d$ |
| Exponential | $\exp(-\alpha(1 - \langle x, p \rangle))$ | $\mathbb{S}^d$ |
| JS | $\exp(-\alpha(H(\frac{x+p}{2}) - H(x) + H(p)))$ | $\Delta^d$ |
| Helinger | $\exp(-\alpha \sum_{i=1}^d (\sqrt{x_i} - \sqrt{y_i})^2)$ | $\Delta^d$ |
| Sinc | $\frac{\sin(\alpha\|x-p\|)}{\alpha\|x-p\|}$ | $\mathbb{R}^{d \leq 3}$ |

There are many positive definite kernels, and we will next highlight a few. We normalize all kernels so that $K(x, x) = 1$ for all $x \in \mathcal{D}$ and therefore $|K(x, y)| \leq 1$ for all $x, y \in \mathcal{D}$. We will use $\alpha > 0$ as a parameter, where $1/\alpha$ represents the bandwidth, or smoothness of the kernel. For $\mathcal{D} = \mathbb{R}^d$ the most common positive definite kernels [49] are the Gaussian (described earlier) and the Laplacian, defined as $\exp(-\alpha\|x - y\|)$ for $x, y \in \mathbb{R}^d$. Another common domain is $\Delta^d = \{ x \in \mathbb{R}^{d+1} | \sum_{i=1}^d x_i = 1, \ x_i \geq 0 \}$, for instance in representing discrete distributions such as normalized counts of words in a text corpus or fractions of tweets per geographic region. Common positive definite kernels for $x, y \in \Delta^d$ include the Hellinger kernel $\exp(-\alpha \sum_{i=1}^d (\sqrt{x_i} - \sqrt{y_i})^2)$ and the Jensen–Shannon (JS) divergence kernel $\exp(-\alpha(H(\frac{x+y}{2}) - H(x) + H(y)))$, where $H(x) = \sum_{i=1}^d x_i \log x_i$ is entropy [24]. In other settings it is more common to normalize data points $x$ to lie on the sphere $\mathbb{S}^d = \{ x \in \mathbb{R}^{d+1} | \|x\| = 1 \}$. Then with $x, y \in \mathbb{S}^d$, the exponential kernel $\exp(-\alpha(1 - \langle x, y \rangle))$ is positive definite [25]. Perhaps surprisingly, positive definite kernels do not need to satisfy $K(x, y) \geq 0$. For $x, y \in \mathbb{R}^d$, the sinc kernel is defined as $\frac{\sin(\alpha\|x-y\|)}{\alpha\|x-y\|}$ and is positive definite for $d = \{1, 2, 3\}$ [43].

**Other classes of kernels** There are other ways to characterize kernels, which provide sufficient conditions for various other coreset bounds. For clarity, we describe these for kernels with domain $\mathbb{R}^d$, but they can apply more generally.

We say a kernel $K$ is $C_K$-Lipschitz if, for any $x, y, z \in \mathbb{R}^d$, $|K(x, z) - K(y, z)| \leq C_K \cdot \|x - y\|$. This ensures that the kernels do not fluctuate too widely, a necessity for robustness, but also prohibits “binary” kernels; for instance the binary ball kernel is defined as $K(x, y) = 1$ if $\|x - y\| \leq r$; and 0 otherwise. Such binary kernels are basically range counting queries (for instance the ball kernel corresponds with a range defined by a ball), and as we will see, this distinction allows the bounds for $\epsilon$-KDE coresets to surpass lower bounds for coresets for range counting queries. Aside from the ball kernel, all kernels we discuss in this paper will be $C_K$-Lipschitz.

Another way to characterize a kernel is with its shape. We can measure this by considering binary ranges defined by super-level sets of kernels. For instance, given a
fixed \( K \) and \( x \in \mathbb{R}^d \), and a threshold \( \tau \in \mathbb{R} \), the super-level set is \( \{ p \in \mathbb{R}^d \mid K(x, p) \geq \tau \} \). For a fixed \( K \), the family of such sets over all choices of \( x \) and \( \tau \) describes a range space with ground set \( \mathbb{R}^d \). For many kernels the VC-dimension of this range space is bounded; in particular, for common kernels, this range is equivalent to those defined by balls in \( \mathbb{R}^d \). Notably, the sinc kernel, which is positive definite for \( \mathbb{R}^d \) with \( d \leq 3 \) does not correspond to a range space with bounded VC-dimension.

Finally, we mention that kernels being characteristic is an important property for many bounds that rely on \( \phi_K \). It includes most, but not all positive definite kernels including Gaussian and Laplace kernels; the notable exceptions are the Euclidean dot product \( \langle x, y \rangle \), and anything derivative of it such as the exponential kernel. A characteristic kernel requires that the kernel \( K \) is positive definite, and the mapping \( \phi_K(x) \) is injective and ultimately this implies its induced distance

\[
D_K(p, x) = \sqrt{\|\phi_K(x)\|_{H_K}^2 + \|\phi_K(p)\|_{H_K}^2 - 2\langle\phi_K(p), \phi_K(x)\rangle_{H_K}}
\]

is a metric [35,49].

**Kernel distance** This \( D_K \) is known as the kernel distance [20,24,28,40] (or current distance or maximum mean discrepancy). If we define the similarity between the two point sets as

\[
k(P, Q) = \frac{1}{|P|} \frac{1}{|Q|} \sum_{p \in P} \sum_{q \in Q} K(p, q) = \langle \hat{\mu}_P, \hat{\mu}_Q \rangle_{H_K},
\]

then the kernel distance can be defined more generally between point sets (implicitly endowed with uniform probability measures) as

\[
D_K(P, Q) = \sqrt{k(P, P) + k(Q, Q) - 2k(P, Q)} = \|\hat{\mu}_P - \hat{\mu}_Q\|_{H_K}.
\]

When \( Q \) is a single point \( x \), then \( k(P, x) = \langle \hat{\mu}_P, \phi_K(x)\rangle_{H_K} = \text{KDE}_P(x) \).

**Relationship between kernel mean and \( \varepsilon \)-KDE coresets** It is possible to convert between bounds on the subset size required to approximate the kernel mean and an \( \varepsilon \)-kernel coreset of an associated kernel range space. But they are not symmetric.

The Koksma–Hlawka inequality (in the context of reproducing kernels [10, 48] when \( K(x, x) = 1 \)) states that

\[
\|\text{KDE}_P - \text{KDE}_Q\|_\infty \leq \|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{H}_K}.
\]

Since \( \text{KDE}_P(x) = k(P, x) = \langle \hat{\mu}_P, \phi(x)\rangle_{\mathcal{H}_K} \) and via Cauchy–Schwartz, for any \( x \in \mathbb{R}^d \),

\[
|\text{KDE}_P(x) - \text{KDE}_Q(x)| = |\langle \hat{\mu}_P, \phi(x)\rangle_{\mathcal{H}_K} - \langle \hat{\mu}_Q, \phi(x)\rangle_{\mathcal{H}_K}| \\
= |\langle \hat{\mu}_P - \hat{\mu}_Q, \phi(x)\rangle_{\mathcal{H}_K}| \leq \|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{H}_K}.
\]
Thus to bound \( \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon \) it is sufficient to bound
\[
\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{C}_K} \leq \varepsilon.
\]

On the other hand, if we have a bound \( \max_{x \in \mathbb{R}^d} |\text{KDE}_P(x) - \text{KDE}_Q(x)| \leq \varepsilon \), then we can only argue that
\[
\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{C}_K} \leq \sqrt{2\varepsilon}.
\]
We observe that
\[
\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{C}_K}^2 = D_K(P, Q)^2 = \kappa(P, P) + \kappa(Q, Q) - 2\kappa(P, Q)
\]
\[
= \frac{1}{|P|} \sum_{p \in P} \text{KDE}_P(p) + \frac{1}{|Q|} \sum_{q \in Q} \text{KDE}_Q(q)
- \frac{1}{|P|} \sum_{p \in P} \text{KDE}_Q(p) - \frac{1}{|Q|} \sum_{q \in Q} \text{KDE}_P(p)
= \frac{1}{|P|} \sum_{p \in P} (\text{KDE}_P(p) - \text{KDE}_Q(p)) + \frac{1}{|Q|} \sum_{q \in Q} (\text{KDE}_Q(q) - \text{KDE}_P(q))
\]
\[
\leq \frac{1}{|P|} \sum_{p \in P} \varepsilon + \frac{1}{|Q|} \sum_{q \in Q} \varepsilon = 2\varepsilon.
\]

Taking square root of both sides leads to the implication.

Unfortunately, the second reduction does not map the other way; a bound on \( \|\hat{\mu}_P - \hat{\mu}_Q\|_{\text{H}}^2 \) only ensures an average \((L_2 \text{ error})\) for \( \text{KDE}_P \) holds, not the desired stronger \( L_\infty \) error. Indeed, the above bound is tight. Consider \( P \) and \( Q \) such that \(|P| = |Q| = 1/\varepsilon\), and all pairs of points \( x, y \in P \cup Q \) are sufficiently far away from each other so that \( K(x, y) \leq \varepsilon^2 \), and hence it must be that \( \|\text{KDE}_P - \text{KDE}_Q\|_\infty \leq \varepsilon^2 \). However, then we can also bound
\[
\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{C}_K}^2 = \frac{1}{|P| |Q|} \left( \sum_{p \in P} \sum_{p' \in P} K(p, p') - 2 \sum_{p \in P} \sum_{q \in Q} K(p, q) + \sum_{q \in Q} \sum_{q' \in Q} K(q, q') \right)
\]
\[
\geq \frac{1}{|P| |Q|} \left( |P| + (|P| - 1)^2 0 - 2|P| \|Q\| \varepsilon^2 + |Q| + (|Q| - 1)^2 0 \right)
\]
\[
= \varepsilon^2 (|P| - 2|P| \|Q\| \varepsilon^2 + |Q|)
\]
\[
= 2\varepsilon - 2\varepsilon^2 = \Omega(\varepsilon).
\]

**Discrepancy-based approaches** Our approach for creating an \( \varepsilon \)-KDE coreset will follow a technique for creating range counting coresets \([6, 9, 36]\). It focuses on assigning a coloring \( \chi: P \to \{-1, +1\} \) to \( P \). Then retains either all \( P_+ = \{ p \in P \mid \chi(p) = +1 \} \) or the remainder \( P_- \), and recursively applies this halving until a small enough coreset \( Q \) has been retained.

Classically, when the goal is to compute a range counting coreset for a range space \((P, \mathcal{R})\), the specific goal of the coloring is to minimize the discrepancy...
\[ \text{disc}_R(P, \chi) = \left| \sum_{p \in P \cap R} \chi(p) \right| \]

over all choices of ranges \( R \in \mathbb{R} \). In the KDE-setting we consider a kernel range space \((P, \mathcal{K})\) where \( \mathcal{K} = \{K(x, \cdot) \mid x \in \mathcal{D}\} \) is defined by a kernel \( K : \mathcal{D} \times \mathcal{D} \to \mathbb{R} \) and a fixed domain \( \mathcal{D} \) which is typically assumed to be \( \mathcal{D} = \mathbb{R}^d \). We instead want to minimize the kernel discrepancy

\[ \text{disc}(P, \chi, x) = \left| \sum_{p \in P} \chi(p)K(x, p) \right|. \]

Now in contrast to the case with the binary range space \((P, \mathcal{R})\), each point \( p \in P \) is partially inside the “range” where the amount inside is controlled by the kernel \( K \).

Understanding the quantity

\[ \text{disc}(n, \mathcal{K}) = \max_{P:|P|=n} \min_{\chi \in \mathcal{D}} \max_{x \in \mathcal{D}} \text{disc}(P, \chi, x) \]

is key. If for a particular \( \mathcal{K} \) we have \( \text{disc}(n, \mathcal{K}) = n^\tau \) or \( \text{disc}(n, \mathcal{K}) = \log^\eta n \), then applying the recursive halving algorithm one obtains an \( \varepsilon \)-KDE coreset of size \( O(1/\varepsilon^{1/(1-\tau)}) \) and \( O((1/\varepsilon) \log^{\eta}(1/\varepsilon)) \), respectively [37].

1.2 Known Results on KDE Coresets

In this section we survey known bounds on the size \(|Q|\) required for \( Q \) to be an \( \varepsilon \)-KDE coreset. We assume \( P \subset \mathbb{R}^d \), it is of size \( n \), and the diameter of \( P \) is \( \Delta = \alpha \max_{p, p' \in P} \|p - p'\| \), where \( 1/\alpha \) is the bandwidth parameter of the kernel. We sometimes allow a \( \delta \) probability that the algorithm does not succeed. Results are summarized in Table 1.

Halving approaches  Phillips [37] showed that kernels with a bounded Lipschitz factor (i.e., \(|K(x, p) - K(x, q)| \leq C\|p - q\| \) for some constant \( C \), including Gaussian, Laplace, and triangle kernels which have \( C = O(\alpha) \), admit coresets of size \( O((\alpha/\varepsilon)^{1/\log(\alpha/\varepsilon)}) \) in \( \mathbb{R}^2 \). For points in \( \mathbb{R}^d \) (for \( d > 1 \)) this generalizes to a bound of \( O((\alpha/\varepsilon)^{2d/(d+2)} \log^{d/(d+2)}(\alpha/\varepsilon)) \). That paper also observed that for \( d = 1 \), selecting evenly spaced points in the sorted order achieves a coreset of size \( O(1/\varepsilon) \).

Sampling bounds  Denote \( \delta \) to be the failure probability. Joshi et al. [28] showed that a random sample of size \( O((1/\varepsilon^2)(d + \log(1/\delta))) \) results in an \( \varepsilon \)-kernel coreset for any centrally symmetric, non-increasing kernel. This works by reducing to a VC-dimensional [30] argument with ranges defined by balls.

Fasy et al. [17] provide an alternative bound on how random sampling preserves the \( L_\infty \) error in the context of statistical topological data analysis. Their bound can be converted to require size \( O((d/\varepsilon^2) \log(d \Delta / \varepsilon \delta)) \), which can improve upon the bound of Joshi [28] if \( K(x, x) > 1 \) (otherwise, herein we only consider the case \( K(x, x) = 1 \)).
| Paper                      | Coreset size                                 | Restrictions             | Algorithm         |
|---------------------------|----------------------------------------------|--------------------------|-------------------|
| Joshi et al. [28]         | $d/\varepsilon^2$                            | Bounded VC               | Random sample     |
| Fasy et al. [17]          | $(d/\varepsilon^2) \log(d/\varepsilon)$    | Lipschitz                | Random sample     |
| Lopaz-Paz et al. [31]     | $1/\varepsilon^2$                            | Characteristic kernels   | Random sample     |
| Chen et al. [10]          | $1/(\varepsilon r_P)$                        | Characteristic kernels   | Iterative         |
| Bach et al. [3]           | $(1/r_P^2) \log(1/\varepsilon)$             | Characteristic kernels   | Iterative         |
| Bach et al. [3]           | $1/\varepsilon^2$                            | Characteristic kernels, weighted | Iterative         |
| Lacoste-Julien et al. [29]| $1/\varepsilon^2$                            | Characteristic kernels   | Iterative         |
| Harvey and Samadi [23]    | $(1/\varepsilon)\sqrt{n} \log^{2.5}(n)$    | Characteristic kernels   | Iterative         |
| Cortez and Scott [12]     | $k_0 \leq (\Delta/\varepsilon d)$           | Lipschitz; $d$ is constant | Iterative         |
| Phillips [37]             | $(1/\varepsilon)^{2d/(d+2)} \log^{d/(d+2)}(1/\varepsilon)$ | Lipschitz; $d$ is constant | Discrepancy-based |
| Phillips [37]             | $\Theta(1/\varepsilon)$                     | $d = 1$                 | Sorting           |

Table 1  Asymptotic $\varepsilon$-KDE coreset sizes in terms of error $\varepsilon$ and dimension $d$
Examining characteristic kernels which induce an RKHS in that function space leads to a simpler bound of $O((1/\varepsilon^2) \log(1/\delta))$ [34]; see Lopaz-Paz et al. [31] for a simple and complete proof.

**Iterative approaches** Motivated by the task of constructing samples from Markov random fields, Chen et al. [10] introduced a technique called *kernel herding* suitable for characteristic kernels. They showed that iteratively and greedily choosing the point $p \in P$ which when added to $Q$ most decreases the quantity $\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{H}_K}$, will decrease that term at rate $O(r_p/t)$ for $t = |Q|$. Here $r_p$ is the largest radius of a ball centered at $\hat{\mu}_P \in \mathcal{H}_K$ which is completely contained in the convex hull of the set $\{\phi(p) | p \in P\}$. They did not specify the quantity $r_p$ but claimed that it is a constant greater than 0.

Bach et al. [3] showed that this algorithm can be interpreted under the Frank–Wolfe framework [11,19]. Moreover, they argue that $r_p$ is not always a constant; in particular when $P$ is infinite (e.g., it represents a continuous distribution) then $r_p$ is arbitrarily small. However, when $P$ is finite, they prove that $1/r_p$ is finite without giving an explicit bound. They also make explicit that after $t$ steps, they achieve $\|\hat{\mu}_P - \hat{\mu}_Q, w\|_{\mathcal{H}_K} \leq 4/(r_p \cdot t)$. They also describe a method which includes “line search” to create a weighted coreset $(Q, w)$, so that each point $q \in Q$ is associated with a weight $w(q) \in [0,1]$ such that $\sum_{q \in Q} w(q) = 1$; then $\hat{\mu}_{Q,w} = \sum_{q \in Q} w(q)\phi(q)$.

For this method they achieve $\|\hat{\mu}_P - \hat{\mu}_{Q,w}\|_{\mathcal{H}_K} \leq \sqrt{\exp(-r_p^2 t)}$. Similarly, other recent progress in Frank–Wolfe analysis focuses on settings which achieve a “linear” rate of roughly $O(c^{-t})$ [18,27]. However, such faster linear convergence, unless some specific properties of the data exist, would violate our lower bound, and thus is not possible in general.

Bach et al. [3] also mentions a bound $\|\hat{\mu}_P - \hat{\mu}_{Q,w}\|_{\mathcal{H}_K} \leq \sqrt{8/t}$, that is independent of $r_p$. It relies on very general bound of Dunn [15] which uses line search, or one of Jaggi [26] which uses a fixed but non-uniform set of weights. These show this convergence rate for any smooth function, including $\|\hat{\mu}_P - \hat{\mu}_{Q,w}\|_{\mathcal{H}_K}^2$; taking the square root provides a bound for $\|\hat{\mu}_P - \hat{\mu}_{Q,w}\|_{\mathcal{H}_K} \leq \varepsilon$ after $t = O(1/\varepsilon^2)$ steps. This result is a weighted coreset, and it has been further improved to be unweighted [29].

Harvey and Samadi [23] further revisited kernel herding in the context of a general mean approximation problem in $\mathbb{R}^d$. That is, consider a set $P'$ of $n$ points in $\mathbb{R}^d$, they constructed a sequence $x_1, x_2, \ldots$ of points from $P'$ such that $\|\bar{P'} - \bar{Q'}\| \leq O(\sqrt{d^3 \log^{2.5}(n)/n})$ for all $t$, where $Q'_t = \{x_1, x_2, \ldots, x_t\}$ and $\bar{P'}$ and $\bar{Q'}_t$ are the Euclidean averages of $P'$ and $Q'_t$, respectively. This maps to the kernel mean problem with $P' = \{\phi_K(p) | p \in P\}$, and with the only bound of $d'$ as $n$ (for any positive definite kernel, the span of the $n$ points in the RKHS has dimension $d' \leq n$). They show that the $r_p$ term can be manipulated by affine scaling, but that in the worst case (after such transformations via John’s theorem) it is $O(\sqrt{d'} \log^{2.5}(n))$, and hence show one can always set $\varepsilon = O(\sqrt{d'} \log^{2.5}(n)/n) = O((1/t)\sqrt{n} \log^{2.5}(n))$, so after $t$ steps as desired $\|\bar{P'} - \bar{Q'}_t\| \leq \varepsilon$. Lacoste-Julien et al. [29] showed that one can always compress $P'$ to another set $P''$ of size $n = O(1/\varepsilon^2)$ (or for instance use the random sampling bound of Lopaz-Paz et al. [31], ignoring the $\log(1/\delta)$ factor); then solving for $t$ yields $t = O((1/\varepsilon^2) \log^{2.5}(1/\varepsilon))$, which is worse than the $O(1/\varepsilon^2)$ bound.
Harvey and Samadi also provide a lower bound to show that after \( t \) steps, at some step the kernel mean error may be as large as \( \Omega(\sqrt{d}/t) \) when \( t = \Theta(n) \). One might think this could imply (using the \( d' = \Omega(n) \) and \( P' \) of size \( \Theta(1/\varepsilon^2) \)) that we need \( t = \Omega(1/\varepsilon^2) \) steps to achieve \( \varepsilon \)-error for kernel density estimates. But this would contradict the bound of Phillips [37], which for instance shows a coreset of size \( O((1/\varepsilon)^2 \log(1/\varepsilon)) \) in \( \mathbb{R}^2 \). More specifically, it uses \( t = \Theta(d') \) steps to achieve this case, so if \( d' = n = \Theta(1/\varepsilon^2) \) then this requires asymptotically as many steps as there are points. Moreover, a careful analysis of their construction shows that the corresponding points in \( \mathbb{R}^d \) (using an inverse projection \( \phi^{-1}_K : \mathbb{K}_K \rightarrow \mathbb{R}^d \) to a set \( P \subseteq \mathbb{R}^d \)) would have them spread out that \( \text{KDE}_P(x) < c/\sqrt{n} \) (for constant \( c \), so \( = O(\varepsilon) \) for \( n = 1/\varepsilon^2 \) for all \( x \in \mathbb{R}^d \)); hence it is easy to construct a \( 2/\varepsilon \) size \( \varepsilon \)-kernel coreset for this point set. This distinction between bounds is indeed related to the difference between kernel mean approximations and \( \varepsilon \)-KDE coreset approximations.

**Discretization bounds** Another series of bounds comes from the Lipschitz factor of the kernels: \( C = \max_{x,y,z \in \mathbb{R}^d} \frac{K(z,x) - K(z,y)}{\|x-y\|} \). For most kernels, \( C \) is a small constant. Thus, we can for instance, lay down an infinite grid \( G_\varepsilon \subset \mathbb{R}^d \) of points so that for all \( x \in \mathbb{R}^d \) there exists some \( g \in G_\varepsilon \) such that \( \|g-x\| \leq \varepsilon/C \), and that means the side length of the grid is \( 2\varepsilon/(C\sqrt{d}) \).

Then we can map each \( p \in P \) to \( p_g \) the closest point \( g \in G_\varepsilon \) (with multiplicity), resulting in \( P_G \). By the additive property of KDE, we know that \( \|\text{KDE}_P - \text{KDE}_{P_G}\|_\infty \leq \varepsilon \).

Cortes and Scott [12] provide another approach to the sparse kernel mean problem. They run Gonzalez’s algorithms [21] for \( k \)-center on the points \( P \subseteq \mathbb{R}^d \) (iteratively add points to \( Q \), always choosing the furthest point from any in \( Q \) and terminate when the furthest distance to the nearest point in \( Q \) is \( \Theta(\varepsilon) \). Then they assign weights to \( Q \) based on how many points are nearby, similarly to the grid argument above. They make an “incoherence” based argument, specifically showing that \( \|\hat{\mu}_P - \hat{\mu}_Q\| \leq \sqrt{1 - v_Q} \) where \( v_Q = \min_{p \in P} \max_{q \in Q} K(p,q) \). This does not translate meaningfully in any direct way to any of the parameters we study. However, we can use the above discretization bound to argue that if \( \Delta \) is bounded, then this algorithm must terminate in \( O((\Delta/\varepsilon)^d) \) steps.

**Lower bounds** Finally, there is a simple lower bound of size \( [1/\varepsilon] - 1 \) for an \( \varepsilon \)-coreset \( Q \) for kernel density estimates [37]. Consider a point set \( P \) of size \( 1/\varepsilon - 1 \) where each point is very far away from every other point, then we cannot remove any point otherwise it would create too much error at that location.

**1.3 Our Results**

We show a new upper bound on the size of an \( \varepsilon \)-KDE coreset of \( O((1/\varepsilon)\sqrt{d} \log(1/\varepsilon)) \) in Sect. 2. The main restriction on the kernel \( K \) is that it is positive definite, a weaker bound than the similar characteristic assumption. There are also fairly benign restrictions (in Euclidean-like domains) that \( K \) is Lipschitz and only has a value greater than \( 1/|P| \) (or \( \geq \varepsilon^2 \)) for pairs of points both within a bounded region; these are due to the specifics of some geometric preprocessing. Notably, this upper bound applies

\[ \text{Springer} \]
Table 2  
Size bounds for $\varepsilon$-KDE coresets for Gaussian and Laplace kernels; also holds under more general assumption, see the text

| $d$     | Upper            | Lower            |
|---------|------------------|------------------|
| 1       | $1/\varepsilon$  | $1/\varepsilon$  | [37]          |
| [2, $1/\varepsilon^2$) | $\sqrt{d}/\varepsilon \cdot \sqrt{\log(1/\varepsilon)}$ | $\sqrt{d}/\varepsilon$ | New$^a$ |
| $\geq 1/\varepsilon^2$ | $1/\varepsilon^2$ | $1/\varepsilon^2$ | [29], new$^b$ |

$^a$ For $d = 2$, [37] matches upper bound  
$^b$ For the lower bound result

to a very wide range of kernels including the sinc kernel, whose super-level sets do not have bounded VC-dimension, and is not characteristic, so no non-trivial $\varepsilon$-KDE coreset bound was previously known. Moreover, unlike previous discrepancy-based approaches, we do not need to assume the dimension $d$ is constant.

We then show a nearly-matching lower bound on the size of an $\varepsilon$-KDE coreset of $\Omega(\sqrt{d}/\varepsilon)$, in Sect. 3. This construction requires a standard restriction that it is shift- and rotation-invariant, and a benign one that it is somewhere steep (see Sect. 3), satisfied by all common kernels. This closes the problem for many kernels (e.g., Gaussians, Laplace), except for a $\sqrt{\log(1/\varepsilon)}$ factor when $1 < d < 1/\varepsilon^2$. The gap filled by the new bounds is shown in Table 2.

Our approach and context  
Bounding the size $\varepsilon$-KDE coresets can be reduced to bounding kernel discrepancy. The range space discrepancy problem, for a range space $(P, \mathcal{R})$, has been widely studied in multiple areas [8,32]. For instance, Tusnady’s problem restricts $\mathcal{R}$ to represent axis-aligned rectangles in $\mathbb{R}^d$, has received much recent focus [33]. To achieve their result, Matoušek et al. [33] use a balancing technique of Banaszcyk [4] on a matrix version of discrepancy, by studying the so-called $\gamma_2$-norm.

Roughly speaking, we are able to show how to reduce directly the kernel discrepancy problem to the $\gamma_2$-norm, and the bound derived from Banaszcyk’s theorem [4]. In particular, the positive definiteness of a kernel allows us to define a specific Gram matrix $G$ which has a real-valued decomposition, which matches the structure studied with the $\gamma_2$-norm. Hence, while our positive definite restriction is similar to the characteristic restriction studied for $\varepsilon$-KDE coresets in many other settings [3,22] it uses a very different aspect of this property: the decomposability, not the embedding.

Finally, we show a lower bound, that there exist point sets $P$ in dimension $d$, such that any $\varepsilon$-kernel coreset requires $\Omega(\sqrt{d}/\varepsilon)$ points. Specifying this to the $d = 1/\varepsilon^2$ case proves a lower bound of $\Omega(1/\varepsilon^2)$ for any case with $d \geq 1/\varepsilon^2$. This applies to every shift-invariant kernel we considered, with a slightly weakened condition for the ball kernel.
2 Upper Bound for KDE Coreset

Recall that our result focuses on the case of \( d < 1/\varepsilon^2 \). We assume that \( P \) is finite and of size \( n \); however, as mentioned in the related work, for many settings, we can reduce this to a point set of size independent of \( n \) (size \( 1/\varepsilon^2 \) or \( d/\varepsilon^2 \), depending on the kernel). Indeed these techniques may start with inputs as continuous distributions as long as we can draw random samples or run iterative algorithm.

Consider a point set \( P \subset \mathbb{R}^d \) as input, but as Sect. 4 describes, it is possible to apply these arguments to other domains.

To prove our \( \varepsilon \)-kernel coreset upper bound we introduce two properties that the kernel must have.

- We say a kernel \( K \) has \( c_K \)-bounded influence if, for any \( x \in \mathbb{R}^d \) and \( \delta > 0 \), \(|K(x, y)| < \delta \) for all \( y \notin x + [-\sqrt{\delta} c_K, \sqrt{\delta} c_K]^d \) for some constant \( c_K \). By default we set \( \delta = 1/n \). If \( c_K \) is an absolute constant we simply say \( K \) has bounded influence.

- We say a kernel \( K \) is \( c_K \)-Lipschitz if, for any \( x, y, z \in \mathbb{R}^d \), \(|K(x, z) - K(y, z)| < c_K \|x - y\| \) for some \( C_K \). If \( C_K \) is an absolute constant within the context of the problem, we often just say the kernel is Lipschitz.

Next define a lattice \( R = \left\{ \left( \frac{i_1}{\sqrt{dn}}, \frac{i_2}{\sqrt{dn}}, \ldots, \frac{i_d}{\sqrt{dn}} \right) \mid i_j \text{ are integers} \right\} \). Also, denote, for each \( p \in P \), \( S_p = p + R \cap [-n c_K, n c_K]^d \) and \( S = \bigcup_{p \in P} S_p \).

The following lemma explains that we only need to consider the evaluation at a finite set (specifically \( S \)) rather than the entire space while preserving the discrepancy asymptotically. The advantage of doing this is we can then use a matrix representation of the discrepancy formula.

**Lemma 2.1** \( \max_{x \in \mathbb{R}^d} \text{disc}(P, \chi, x) \leq \max_{x \in S} \text{disc}(P, \chi, x) + O(1) \).

**Proof.** For any \( x \in \mathbb{R}^d \), if \( x \notin \bigcup_{p \in P} \left( p + [-n c_K, n c_K]^d \right) \), that is \( x \) is not within \( n c_K \) in all coordinates of some \( p \in P \), then \( K(p, x) \leq 1/n \) for all \( p \in P \). Hence we have

\[
\text{disc}(P, \chi, x) = \left| \sum_{p \in P} \chi(p) K(p, x) \right| \leq O(1).
\]

Otherwise, pick \( x_0 \in S \) to be the closest point to \( x \). We have

\[
\text{disc}(P, \chi, x) = \left| \sum_{p \in P} \chi(p) K(p, x) \right| = \left| \sum_{p \in P} \chi(p)(K(p, x) - K(p, x_0)) \right| \leq \sum_{p \in P} \chi(p) K(p, x_0) - K(p, x_0) \leq \text{disc}(P, \chi, x_0) + \sum_{p \in P} C_K \|x - x_0\|
\]
\[ \leq \text{disc}(P, \chi, x_0) + n \cdot C_K \cdot \sqrt{d \left( \frac{1}{\sqrt{dn}} \right)^2} \]

\[ = \text{disc}(P, \chi, x_0) + O(1). \]

Now we discuss the matrix view of discrepancy, known results, and then how to map the discretized kernel discrepancy problem into this setting. Consider any \( s \times t \) matrix \( A \), and define

\[ \text{disc}(A) = \min_{x \in [-1,1]^t} \|Ax\|_\infty. \]

Following Matoušek et al. [33] we define \( \gamma_2(A) = \min_{BC=A} r(B) \cdot c(C) \) where \( r(B) \) is largest Euclidean norm of row vectors of \( B \) and \( c(C) \) is largest Euclidean norm of column vectors of \( C \). There is an equivalent [33] geometric interpretation of \( \gamma_2 \). Let \( E_A \) be the set of ellipsoids in \( \mathbb{R}^s \) that contain all vector columns of \( A \). Then,

\[ \gamma_2(A) = \min_{E \in E_A} \max_{x \in E} \|x\|_\infty. \]

It is easy to see that \( \gamma_2 \) is a norm and \( \gamma_2(A) \leq \gamma_2(A') \) when the columns of \( A \) are a subset of the columns of \( A' \). We will apply these properties shortly.

A recent result by Matoušek et al. [33] shows the following property about connecting discrepancy to \( \gamma_2 \), which was recently made constructive in polynomial time [5].

**Lemma 2.2** (Matoušek et al. [33]) For an \( s \times t \) matrix \( A \), \( \text{disc}(A) \leq O(\sqrt{\log s}) \cdot \gamma_2(A) \).

Let the size of \( S \) be \( m = O(n^{O(d)}) \), and define an \( m \times n \) matrix \( G \) so that its rows are indexed by \( x \in S \) and columns indexed by \( p \in P \), and \( G_{x,p} = K(p, x) \). By examination, \( \text{disc}(G) = \min_{\chi} \max_{x \in S} \text{disc}(P, \chi, x) \).

**Lemma 2.3** \( \gamma_2(G) = 1 \).

**Proof** Denote \( G' \) to be an \( m \times m \) matrix with both rows and columns indexed by \( x, y \in S \) so that \( G'_{x,y} = K(x, y) \). Note that columns of \( G \) are a subset of columns of \( G' \) since \( P \subset S \). Since \( K \) is a positive definite kernel, it means that \( G' \) can be expressed as \( H^T H \) for some matrix \( H \). Now denote \( v_x \) as the \( x \)th column of \( H \) for all \( x \in S \). We have \( v_x^T v_x = G'_{x,x} = 1 \) which means the norm \( \|v_x\| = \sqrt{v_x^T v_x} = 1 \) for each column \( v_x \in H \). Hence the same holds for rows in \( H^T \), and this bounds \( \gamma_2(G') \leq 1 \). Then since \( \gamma_2(G) \leq \gamma_2(G') \), we have \( \gamma_2(G) \leq 1 \).

On the other hand, one of the coordinates in a column of \( G \) is 1. By the geometric definition, any ellipsoid containing columns of \( G \) has a point inside of it such that one of its coordinates is 1. Hence \( \gamma_2(G) \geq 1 \).

Combining all above lemmas, for any \( P \subset \mathbb{R}^d \) of size \( n \), we have

\[ \text{disc}(n, \mathcal{K}) \leq \max_{P:|P|=n} \min_{\chi} \max_{x \in S} \text{disc}(P, \chi, x) + O(1) \]

\[ = \max_{P:|P|=n} \text{disc}(G) + O(1) \]

\[ \leq O(\sqrt{d \log n} \cdot \gamma_2(G)) \]

\[ = O(\sqrt{d \log n}). \]
Theorem 2.4 Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded influence, Lipschitz, positive definite kernel. For any integer $n$, $\text{disc}(n, K_d) = O(\sqrt{d \log n})$.

Corollary 2.5 Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded influence, Lipschitz, positive definite kernel. For any set $P \subset \mathbb{R}^d$, there is a subset $Q \subset P$ of size $O\left(\frac{1}{\varepsilon} \sqrt{d \log \left(\frac{1}{\varepsilon}\right)}\right)$ such that

$$\max_{x \in \mathbb{R}^d} \left| \text{KDE}_P(x) - \text{KDE}_Q(x) \right| < \varepsilon.$$ 

Proof. In order to apply the standard halving technique [9,36], we need to make sure the coloring has the property that half of points are assigned $+1$ and the other half of them are assigned $-1$. We adapt a standard idea from combinatorial discrepancy [32].

This can be done by adding an all-one row to the discrepancy matrix $G$. It guarantees that the difference of the number of $+1$ and $-1$ is $O(\sqrt{d \log n})$ since $\gamma_2$ is a norm, and therefore we can apply the triangle inequality. Namely,

$$\gamma_2\left(\left[\mathbb{1}_{1 \times n} G\right]\right) \leq \gamma_2\left(\left[O_{1 \times n} G\right]\right) + \gamma_2\left(\left[\mathbb{1}_{1 \times n} O_{m \times n}\right]\right),$$

where $\mathbb{1}$ is all-one matrix and $O$ is zero matrix. Let $P_+ = \{p \in P \mid \chi(p) = +1\}$ and $P_- = \{p \in P \mid \chi(p) = -1\}$. Suppose there are more $+1$s than $-1$s. Choose $O\left(\frac{1}{\varepsilon} \sqrt{d \log \left(\frac{1}{\varepsilon}\right)}\right)$ points assigned $+1$ arbitrarily and flip them to $-1$ so that it makes the difference zero. $P'_+$ and $P'_-$ are defined in the same way as $P_+$ and $P_-$, after flipping some values. For any $x \in \mathbb{R}^d$,

$$\left| \sum_{p \in P'_+} K(x, p) - \sum_{p \in P'_-} K(x, p) \right| \leq \left| \sum_{p \in P_+} K(x, p) - \sum_{p \in P_-} K(x, p) \right| + \left| \sum_{p \in P'_+ \setminus P_+} K(x, p) \right| + \left| \sum_{p \in P_- \setminus P'_-} K(x, p) \right| = O(\sqrt{d \log n}).$$

Now, we can apply the standard halving technique to achieve

$$\max_{x \in \mathbb{R}^d} \left| \text{KDE}_P(x) - \text{KDE}_Q(x) \right| < \varepsilon.$$ 

Implementation Note that we do not need to decompose the entire matrix $G$. Instead, we just need a set of vectors $V = \{v_p \mid p \in P\}$ such that the inner product $\langle v_{p_1}, v_{p_2} \rangle = K(p_1, p_2)$ is input to the algorithm in [5]. This set $V$ can be computed in $\text{poly}(n, d) = \text{poly}(n)$ time assuming $d < n$. Using the standard Merge-Reduce framework [36], the coreset with desired size can be constructed in $O(n \text{poly}(1/\varepsilon))$ time.
3 Lower Bound for KDE Coreset

In this section, we add two new conditions on our kernel; both of these are common properties of kernels.

- A kernel $K$ is rotation- and shift-invariant if there exists a function $f$ such that $K(x, y) = f(\|x - y\|^2)$.
- A rotation- and shift-invariant kernel is somewhere $C_f$-steep if there exist a constant $C_f > 0$, and values $z_f > r_f > 0$ such that $f(z_1) - f(z_2) > C_f \cdot (z_2 - z_1)$ for all $z_1 \in (z_f - r_f, z_f)$ and $z_2 \in (z_f, z_f + r_f)$. When $C_f$ is an absolute constant, we often just say the kernel is somewhere steep.

Phillips [37] constructed an example of $P$ of size $1/\varepsilon$ where each point in $P$ is far away from all others. Therefore, if one of them is not picked for a KDE coreset, the evaluation of KDE at that point has large error. We divide $n = \sqrt{d}/\varepsilon$ points into $n/d$ groups where each group has $d$ points that form a simplex, and each group is far away from all other groups. It means that there is a group producing $\Omega(1/\sqrt{d})$ error when considered alone, and then, since we have $n/d$ groups, the final error would be $\Omega\left(\frac{1/\sqrt{d}}{n/d}\right) = \Omega(\varepsilon)$.

**Theorem 3.1** Suppose $\varepsilon > 0$. Consider a rotation- and shift-invariant, somewhere steep, bounded influence kernel $K$. Assume $1/\varepsilon^2 \geq d \geq 9z_f^2/r_f^2$, where $z_f$ and $r_f$ are absolute constants that depend on $K$ and are defined as they pertain to the somewhere steep criteria. There is a set $P \in \mathbb{R}^d$ such that, for any subset $Q$ of size $k \leq \sqrt{d}/(2\varepsilon)$, there is a point $x \in \mathbb{R}^d$ such that $|\text{KDE}_P(x) - \text{KDE}_Q(x)| > \varepsilon$.

**Proof.** Let $n = \sqrt{d}/\varepsilon$. We allow weighted coresets of $Q$; that is, for each $q \in Q$, there is a real number $\beta_q$ such that $\text{KDE}_Q(x) = \sum_{q \in Q} \beta_q K(x, q)$.

Let $k \leq n/2$ be the size of the potential coreset we consider. Construct $P$ of size $n$ in $\mathbb{R}^d$ as follows. Let $\{e_i\}_{i=1}^d$ be the standard basis and $L$ be a very large number. Set $P_j = \{p_{i,j} = \sqrt{\frac{z_f}{L}} e_i + j L e_1 \mid i = 1, 2, \ldots, d\}$ for all $j = 1, 2, \ldots, n/d$. Define $P = \bigcup_{j=1}^{n/d} P_j$. Namely, we divide $n$ points into $n/d$ groups so that each group has $d$ points which form a $d$-simplex. Also, the groups are sufficiently far away from each other. Suppose $Q = \bigcup_{j=1}^{n/d} \{p_{i,a,j} \mid a = 1, 2, \ldots, k_j\}$ where $k_j$ is the number of points in $Q$ at group $j$. Denote $Q_j = \{p_{i,a,j} \mid a = 1, 2, \ldots, k_j\}$. That is, $Q = \bigcup_{j=1}^{n/d} Q_j$ and $|Q_j| = k_j \leq d$ with $\sum_{j=1}^{n/d} k_j = |Q| = k$.

Since $\sum_{j=1}^{n/d} |Q_j| = k \leq n/2$, at least one $j$ must satisfy $k_j \leq d/2$. Denote $j'$ to be that $j$. We can assume $k_{j'} = d/2$, otherwise, pick enough points arbitrarily from $P_j \setminus Q_j$ and place them in $Q_{j'}$ to make $|Q_{j'}| = k_{j'} = d/2$, but set the corresponding weight to be 0. Denote by $\bar{\rho} = \frac{1}{d} \sum_{p \in P_{j'}} p$ the mean of $P_{j'}$; by $\bar{q}_+ = \frac{3}{d} \sum_{q \in Q_{j'}} q$ the mean of $Q_{j'}$; and by $\bar{q}_- = \frac{3}{d} \sum_{q \in P_{j'} \setminus Q_{j'}} q$ the mean of points in $P_{j'}$ not selected into $Q_{j'}$; see Fig. 1. Also, denote by $p^*_+ = \bar{q}_+ + \sqrt{\frac{z_f}{2}} \frac{\bar{q}_- - \bar{\rho}}{\|\bar{q}_- - \bar{\rho}\|}$ and $p^*_- = \bar{q}_- + \sqrt{\frac{z_f}{2}} \frac{\bar{q}_+ - \bar{\rho}}{\|\bar{q}_+ - \bar{\rho}\|}$ translates of these points away from the mean $\bar{\rho}$ by a specific vector. Note that $\|p^*_+ - q\|$ is the same for all $q \in Q_{j'}$, denoted by $l_1$, and $\|p^*_+ - q\|$ is same for all $q \in P_{j'} \setminus Q_{j'}$. 

\(\square\) Springer
denoted by $l_2$. By symmetry, we also have that $l_1 = \|p_-^* - q\|$ for all $q \in P_{j'} \setminus Q_{j'}$ and $l_2 = \|p_+^* - q\|$ for all $q \in Q_{j'}$.

If $\sum_{q \in Q_{j'}} \beta_q \geq d/n$, we evaluate the error at $p_+^*$.

$$
(KDE_Q - KDE_P)(p_+^*) = \sum_{q \in Q_{j'}} \left( \beta_q - \frac{1}{n} \right) f(\|p_+^* - q\|^2) + \sum_{q \in P_{j'} \setminus Q_{j'}} \left( -\frac{1}{n} \right) f(\|p_+^* - q\|^2) + s
$$
$$
\geq \frac{d}{2n} (f(l_1^2) - f(l_2^2)) + s
$$

where $|s|$ is arbitrarily small due to the choice of arbitrarily large number $L$ and the fact that $K$ is bounded influence. If $\sum_{q \in Q_{j'}} \beta_q \leq d/n$, we evaluate the error at $p_-^*$.

$$
(KDE_P - KDE_Q)(p_-^*) = \sum_{q \in P_{j'} \setminus Q_{j'}} \frac{1}{n} f(\|p_-^* - q\|^2) + \sum_{q \in Q_{j'}} \left( \frac{1}{n} - \beta_q \right) f(\|p_-^* - q\|^2) + s
$$
$$
\geq \frac{d}{2n} (f(l_1^2) - f(l_2^2)) + s.
$$

Therefore, in either case, we need to bound $f(l_1^2) - f(l_2^2)$ from below.

By direct computation, we have $l_1^2 = z_f - \frac{z_f}{d}$ and $l_2^2 = z_f + \frac{z_f}{d} + \frac{2z_f}{\sqrt{d}}$. By enforcing that

$$
z_f - r_f < z_f - \frac{z_f}{d} = l_1^2 < z_f
$$
and

\[ z_f < z_f + \frac{z_f}{d} + \frac{2z_f}{\sqrt{d}} = l_2^2 < z_f + \frac{3z_f}{\sqrt{d}} < z_f + r_f. \]

we can invoke the somewhere \( C_f \)-steep property that there exists an \( x \) in \( \mathbb{R}^d \) for which the inequality holds. Therefore,

\[ f(l_1^2) - f(l_2^2) > C_f \cdot (l_2^2 - l_1^2) > C_f \cdot z_f \cdot \frac{2}{\sqrt{d}}. \]

Hence, the error is at least

\[
\frac{d}{2n} \left( f(l_1^2) - f(l_2^2) \right) + s > \frac{d}{2n} \left( C_f \cdot z_f \cdot \frac{2}{\sqrt{d}} \right) + s
\]

\[
> \frac{\sqrt{d}}{n} \cdot C_f \cdot z_f + s = \Omega(\sqrt{d}/n) = \Omega(\epsilon). \]

\[ \square \]

Note that when \( d > 1/\epsilon^2 \) the above argument is still valid by considering \( d = 1/\epsilon^2 \). Hence, we have the following conclusion.

**Corollary 3.2** Suppose \( \epsilon > 0 \). Consider a rotation- and shift-invariant, somewhere steep, bounded influence kernel \( K \). Assume \( d \geq 9z_f^2/r_f^2 \), where \( z_f \) and \( r_f \) are absolute constants that depend on \( K \) and are defined as they pertain to the somewhere steep criteria. There is a set of \( P \subset \mathbb{R}^d \) such that, for any subset \( Q \) of size \( k \leq \min\{\sqrt{d}/\epsilon, 1/\epsilon^2\} \), there is a point \( x \in \mathbb{R}^d \) such that \( |\text{KDE}_P(x) - \text{KDE}_Q(x)| > \epsilon \).

#### 4 Applications to Specific Positive Definite Kernels

In this section, we work through the straightforward application of these bounds to some specific kernels and settings.

**Gaussian and Laplace kernels** These kernels are defined over \( \mathbb{R}^d \). They have bounded influence, so \( |K(x, p)| \leq 1/n \) for all \( p \notin [-n^{c_K}, n^{c_K}]^d \) for \( c_K = 1 \). They are also \( C_K \)-Lipschitz with constant \( C_K = \alpha \), so \( |K(x, z) - K(p, z)| \leq C_K \|x - p\| \) for any \( x, p \in \mathbb{R}^d \). These properties imply we can invoke the discrepancy upper bound in Theorem 2.4.

These kernels are also rotation- and shift-invariant, and somewhere steep with constant \( C_f = (\alpha/2) \exp(-\alpha^2) \). Hence we can invoke the lower bound in Theorem 3.1.

**Corollary 4.1** *For Gaussian or Laplacian kernels, for any set \( P \in \mathbb{R}^d \), there is an \( \epsilon \)-KDE coreset of size \( O((\sqrt{d}/\epsilon)\sqrt{\log 1/\epsilon}) \), and it cannot have an \( \epsilon \)-KDE coreset of size \( o(\sqrt{d}/\epsilon) \).*

\[ \square \] Springer
The Gaussian kernel has an amazing decomposition property that in \( \mathbb{R}^d \) if we fix any \( d' \) coordinates in any way, then conditioned on those, the remaining \( d - d' \) coordinates still follow a Gaussian distribution. Among other things, this means it is useful to construct kernels for complex scenarios. For instance, consider a large set \( T \) of \( n \) trajectories, each with \( k \) waypoints; e.g., backpacking or road trips or military excursions with \( k \) nights, and let the waypoints be the \( (x, y) \)-coordinates for the location of each night stay. We can measure the similarity between two trajectories \( t = (p_1, p_2, \ldots, p_k) \) and \( t' = (p'_1, p'_2, \ldots, p'_k) \) as the average similarity between the corresponding waypoints, and we can measure the similarity of any two corresponding waypoints \( p_j \) and \( p'_j \) with a two-dimensional Gaussian. Then, by the decomposition property, the full similarity between the trajectories is precisely a \( (2k) \)-dimensional Gaussian. We can thus define a kernel density estimate over these trajectories \( \text{KDE}_T \) using this \( (2k) \)-dimensional Gaussian kernel. Now, given Corollary 4.1 we know that to approximate \( \text{KDE}_T \) with a much smaller data set \( S \subset T \) so \( \| \text{KDE}_T - \text{KDE}_S \|_\infty \leq \varepsilon \), we can construct \( S \) so \( |S| = O(\sqrt{k}/\varepsilon \cdot \sqrt{\log 1/\varepsilon}) \) but cannot in general achieve \(|S| = o(\sqrt{k}/\varepsilon) \).

**Jensen–Shannon and Hellinger kernels** In order to apply our technique on \( \Delta^d \), observe that \( \Delta^d \) is a subset of a \( (d - 1) \)-dimensional Euclidean subspace of \( \mathbb{R}^d \); so we can simply create the grid needed for Lemma 2.1 within this subspace. Recall that these two kernels have the form of \( \exp(-\alpha d(x, y)) \) where \( d(x, y) = \delta_{\text{JS}}(x, y) = H((x + y)/2) - (H(x) + H(y))/2 \) for the Jensen–Shannon kernel and \( d(x, y) = \delta_{\text{H}}(x, y) = \sum_{i=1}^d (\sqrt{x_i} - \sqrt{y_i})^2 \) for Hellinger and note that \( |K(x, z) - K(y, z)| \leq \alpha |d(x, z) - d(y, z)| \) for any \( x, y, z \in \Delta^d \). It is easy to estimate that when \( x, y \) are sufficiently close, for JS kernel, \( |d(x, z) - d(y, z)| \leq 2d \max_i |x_i - y_i||\log |x_i - y_i|| \leq 2d \max_i \sqrt{|x_i - y_i|} \) and for the Hellinger kernel, \( |d(x, z) - d(y, z)| \leq 4d \max_i \sqrt{|x_i - y_i|} \). So even though these kernels are not Lipschitz, we can still modify the construction of the grid in Lemma 2.1 with width \( 1/n^4 \) (assuming \( d \leq n \)) instead of \( 1/\sqrt{dn} \) so that if \( x, y \) lie in the same cell then \( |K(x, z) - K(y, z)| = O(1/n) \) for any \( x, y, z \in \Delta^d \). Since all relevant points are in a bounded domain, both kernels have \( c_K \)-bounded influence; setting \( c_K = 1 \) is sufficient.

**Corollary 4.2** For Jensen–Shannon and Hellinger kernels, for any set \( P \in \Delta^d \), there is an \( \varepsilon \)-KDE coreset of size \( O((\sqrt{d}/\varepsilon)\sqrt{\log 1/\varepsilon}) \).

Note that these kernels are not rotation- and shift-invariant and therefore our lower bound result does not apply.

These kernels are based on widely-used information distances: the Jensen–Shannon distance \( \delta_{\text{JS}}(x, p) \) and the Hellinger distance \( \delta_{\text{H}}(x, p) \). These make sense when the input data \( x, p \in \Delta^d \) represent a “histogram,” a discrete probability distribution over a \( d \)-variate domain. These are widely studied objects in information theory, and more commonly text analysis. For instance, a common text modeling approach is to represent each document \( v \) in a large corpus of documents \( V \) (e.g., a collection of tweets, or news articles, or Wikipedia pages) as a set of word counts. That is, each coordinate \( v_j \) of \( v \) represents the number of times that word (indexed by) \( j \) occurs in that document. To remove length information from the documents (retaining only the topics), it is...
common to normalize each vector as \( v \mapsto v/\|v\| \) so the \( j \)th coordinate represents the probability that a random word on the page is \( j \). The most common modeling choice to measure distance between these distribution representations of documents are the Hellinger and Jensen–Shannon distances, and hence the most natural choice of similarity are the corresponding kernels we examine. In particular, with a very large corpus \( V \) of size \( n \), Corollary 4.2 shows that we can approximate \( \text{KDE}_V \), a kernel density estimate of \( V \), with one described by a much smaller set \( S \subseteq V \) such that \( \| \text{KDE}_V - \text{KDE}_S \| \leq \varepsilon \) and \( |S| = O(\sqrt{d/\varepsilon} \cdot \sqrt{\log 1/\varepsilon}) \). Noteably, when one has a fairly large \( d \), and desires high accuracy (small \( \varepsilon \)), then our new result will provide the best possible \( \varepsilon \)-KDE coreset.

**Exponential kernels** In order to apply our technique on \( \mathbb{S}^d \), we can rewrite the kernel to be \( K'(x, y) = K(x/\|x\|, y/\|y\|) \) for all \( x, y \in \mathbb{R}^d \setminus \{0\} \). We construct the grid in Lemma 2.1 on \( \mathbb{R}^d \) for \( K' \) and then only retain grid points which lie in the annulus \( A^d = \{ x \in \mathbb{R}^d \mid 1/2 \leq \|x\| \leq 3/2 \} \). This annulus contains all grid points which could be the closest point of some point on \( \mathbb{S}^d \), as required in Lemma 2.1. Moreover \( K' \) is \( C_K \)-Lipschitz on the annulus: it satisfies for any \( x, y, z \in A^d \) that \( |K'(x, z) - K'(y, z)| \leq C_K \|x - y\| \), with \( C_K = 4\alpha \). Since the domain is restricted to \( \mathbb{S}^d \), similarly to the domain \( \Delta^d \), any kernel has \( c_K \)-bounded influence and setting \( c_K = 1 \) is sufficient.

**Corollary 4.3** For the exponential kernel, for any set \( P \in \mathbb{S}^d \), there is an \( \varepsilon \)-KDE coreset of size \( O((\sqrt{d/\varepsilon})\sqrt{\log 1/\varepsilon}) \).

The exponential kernel is not rotation- and shift-invariant and therefore our lower bound result does not apply.

**Sinc kernel** Note that the sinc kernel is not everywhere positive, and as a result of its structure the VC-dimension is unbounded, so the approaches requiring those properties [28,37] cannot be applied. It is also not characteristic, so the embedding-based results [3,22] do not apply either. As a result, there is no non-trivial \( \varepsilon \)-KDE coreset for the sinc kernel. However, in our approach, the positivity of one single entry in the discrepancy matrix does not matter as long as the entire matrix is positive definite—which is the case for the sinc. Therefore, our result could be applied to sinc kernel, with \( c_K = 1 \) (it has \( 1 \)-bounded influence), \( C_K = \alpha/\pi \) (it is \( (\alpha/\pi) \)-Lipschitz) and \( C_f = \alpha^2/2\pi^2 \) (it is somewhere \( (\alpha^2/2\pi^2) \)-steep).

**Corollary 4.4** For sinc kernels, for any set \( P \in \mathbb{R}^d \), there is an \( \varepsilon \)-KDE coreset of size \( O((1/\varepsilon)\sqrt{\log 1/\varepsilon}) \) (for \( d = \{1, 2, 3\} \)), and it cannot have an \( \varepsilon \)-KDE coreset of size \( \Omega(1/\varepsilon) \).

5 Conclusion

We proved that the Gaussian kernel has an \( \varepsilon \)-KDE coreset of size \( O((1/\varepsilon)\sqrt{d \log(1/\varepsilon)}) \) and the size must satisfy \( \Omega(\min\{1/\varepsilon^2, \sqrt{d/\varepsilon}\}) \); both upper and lower bound results can be extended to a broad class of kernels. In particular the upper bounds only require
that the kernel be characteristic or in some cases only positive definite (typically the same restriction needed for most machine learning techniques) and that it has a domain which can be discretized over a bounded region without inducing too much error. This family of applicable kernels includes new options like the sinc kernel, which while positive definite in \( \mathbb{R}^d \) for \( d = \{1, 2, 3\} \), is not characteristic, is not always positive, and its super-level sets do not have bounded VC-dimension. This is the first non-trivial \( \varepsilon \)-KDE coreset result for these kernels.

By inspecting the new constructive algorithm for obtaining small discrepancy in the \( \gamma_2 \)-norm [5], the extra \( \sqrt{\log} \) factor comes from the union bound over the randomness in the algorithm. Indeed, if \( d = 1/\varepsilon^2 \) then the upper bound is \( O(1/\varepsilon^2) \), which is tight. This bound is deterministic and does not have an extra \( \sqrt{\log} \) factor. Therefore, a natural conjecture is that the upper bound result can be further improved to \( O(\sqrt{d/\varepsilon}) \), at least in a well-behaved setting like for the Gaussian kernel.

There are many other even more diverse kernels which are positive definite, which operate on domains as diverse as graphs, time series, strings, and trees [25]. The heart of the upper bound construction which uses the decomposition of the associated positive definite matrix will work even for these kernels. However, it is less clear how to generate a finite Gram or discrepancy matrix \( G \), whose size depends polynomially on the data set size for these discrete objects. Such constructions would further expand the pervasiveness of the \( \varepsilon \)-KDE coreset technique we present.

References

1. Arias-Castro, E., Mason, D., Pelletier, B.: On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. J. Mach. Learn. Res. 17, 43 (2016)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. 68(3), 337–404 (1950)
3. Bach, F., Lacoste-Julien, S., Obozinski, G.: On the equivalence between herding and conditional gradient algorithms. In: Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML’12), pp. 1355–1362. Omnipress (2012)
4. Banaszczuk, W.: Balancing vectors and Gaussian measures of \( n \)-dimensional convex bodies. Random Struct. Algorithms 12(4), 351–360 (1998)
5. Bansal, N., Dadush, D., Garg, S., Lovett, S.: The Gram–Schmidt walk: a cure for the Banaszczyk blues (STOC’18). In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 587–597. ACM, New York (2018)
6. Bentley, J.L., Saxe, J.B.: Decomposable searching problems I: static-to-dynamic transformations. J. Algorithms 1, 4 (1980)
7. Bobrowski, O., Mukherjee, S., Taylor, J.E.: Topological consistency via kernel estimation. Bernoulli 23(1), 288–328 (2017)
8. Chazelle, B.: The Discrepancy Method. Cambridge University Press, Cambridge (2000)
9. Chazelle, B., Matoušek, J.: On linear-time deterministic algorithms for optimization problems in fixed dimensions. J. Algorithms 21(3), 579–597 (1996)
10. Chen, Y., Welling, M., Smola, A.: Super-samples from kernel hearding. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI’10), pp. 109–116. AUAI Press, Arlington (2010)
11. Clarkson, K.: Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm. ACM Trans. Algorithms 4(6), 63 (2010)
12. Cortés, E.C., Scott, C.: Sparse approximation of a kernel mean. IEEE Trans. Signal Process. 65(5), 1310–1323 (2016)
13. Devroye, L., Györfi, L.: Nonparametric Density Estimation: The \( L_1 \) View. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. Wiley, New York (1985)
14. Drineas, P., Mahoney, M.W.: On the Nyström method for approximating a Gram matrix for improved kernel-based learning. J. Mach. Learn. Res. 6, 2153–2175 (2005)
15. Dunn, J.C.: Convergence rates for conditional gradient sequences generated by implicit step length rules. SIAM J. Control Optim. 18(5), 473–489 (1980)
16. Fan, J., Gijbels, I.: Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability, vol. 66. Chapman & Hall, London (1996)
17. Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A.: Confidence sets for persistence diagrams. Ann. Stat. 42(6), 2301–2339 (2014)
18. Freund, R.M., Grigas, P.: New analysis and results for the Frank–Wolfe method. Math. Program. 155(1–2), 199–230 (2016)
19. Gärtner, B., Jaggi, M.: Coresets for polytope distance. In: Proceedings of the 25th Annual Symposium on Computational Geometry (SCG’09), pp. 33–42. ACM, New York (2009)
20. Glaunès, J.: Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l’anatomie numérique. PhD thesis, Université Paris 13 (2005)
21. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoret. Comput. Sci. 38(2–3), 293–306 (1985)
22. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. 13, 723–773 (2012)
23. Harvey, N., Samadi, S.: Near-optimal herding. In: Proceedings of the 27th Conference on Learning Theory vol. 35, pp. 1165–1183 (2014)
24. Hein, M., Bousquet, O.: Hilbertian metrics and positive definite kernels on probability measures. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 136–143 (2005)
25. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. Ann. Stat. 36(3), 1171–1220 (2008)
26. Jaggi, M.: Revisiting Frank–Wolfe: projection-free sparse convex optimization. In: Proceedings of the 30th International Conference on Machine Learning, vol 28(1), pp. 427–435 (2013)
27. Jaggi, M., Lacoste-Julien, S.: On the global linear convergence of Frank–Wolfe optimization variants. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
28. Joshi, S., Kommaraju, R.V., Phillips, J.M., Venkatasubramanian, S.: Comparing distributions and shapes using the kernel distance. In: Proceedings of the 27th Annual Symposium on Computational Geometry (SoCG’11), pp. 47–56. ACM, New York (2011)
29. Lacoste-Julien, S., Lindsten, F., Bach, F.: Sequential kernel herding: Frank–Wolfe optimization for particle filtering. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, pp. 544–552 (2015)
30. Li, Y., Long, P.M., Srinivasan, A.: Improved bounds on the samples complexity of learning. J. Comput. Syst. Sci. 62(3), 516–527 (2001)
31. Lopaz-Paz, D., Muandet, K., Schölkopf, B., Tolstikhin, I.: Towards a learning theory of cause-effect inference. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 1452–1461 (2015)
32. Matoušek, J.: Geometric Discrepancy: An Illustrated Guide. Algorithms and Combinatorics, vol. 18, 2nd edn. Springer, Berlin (2010)
33. Matoušek, J., Nikolov, A., Talwar, K.: Factorization norms and hereditary discrepancy. Int. Math. Res. Not. https://doi.org/10.1093/imrn/rny033
34. Muandet, K., Fukumizu, K., Sriperumbudur, B.K., Schölkopf, B.: Kernel mean embedding of distributions: a review and beyond. Found. Trends Mach. Learn. 10, 1–141 (2017)
35. Müller, A.: Integral probability metrics and their generating classes of functions. Adv. Appl. Probab. 29(2), 429–443 (1997)
36. Phillips, J.M.: Algorithms for ε-approximations of terrains. In: ICALP (2008)
37. Phillips, J.M.: ε-Samples for kernels. In: Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA’13), pp. 1622–1632. SIAM, Philadelphia (2013)
38. Phillips, J.M., Tai, W.M.: Improved coresets for kernel density estimates. In: Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA’18), pp. 2718–2727. SIAM, Philadelphia (2018)
39. Phillips, J.M., Tai, W.M.: Near-optimal coresets for kernel density estimates. In: Proceedings 34th International Symposium on Computational Geometry (SoCG’18), pp. 66:1–66:13. Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2018)
40. Phillips, J.M., Venkatasubramanian, S.: A gentle introduction to the kernel distance. arXiv:1103.1625 (2011)
41. Phillips, J.M., Wang, B., Zheng, Y.: Geometric inference on kernel density estimates. In: Proceedings 31th International Symposium on Computational Geometry (SoCG’15). Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2015)
42. Rinaldo, A., Wasserman, L.: Generalized density clustering. Ann. Stat. 38(5), 2678–2722 (2010)
43. Schoenberg, I.J.: Metric spaces and completely monotone functions. Ann. Math. 39(4), 811–841 (1938)
44. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
45. Schubert, E., Zimek, A., Kriegel, H.P.: Generalized outlier detection with flexible kernel density estimates. In: Proceedings of the SIAM International Conference on Data Mining, pp. 542–550 (2014)
46. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York (1992)
47. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, London (1986)
48. Song, L., Zhang, X., Smola, A., Gretton, A., Schölkopf, B.: Tailoring density estimation via reproducing kernel moment matching. In: Proceedings of the 25th International Conference on Machine Learning (ICML’08), pp. 992–999. ACM, New York (2008)
49. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.G.: Hilbert space embeddings and metrics on probability measures. J. Mach. Learn. Res. 11, 1517–1561 (2010)
50. Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: Advances in Kernel Methods—Support Vector Learning, pp. 69–88. MIT Press, Cambridge (1999)
51. Zheng, Y., Phillips, J.M.: $L_\infty$ error and bandwidth selection for kernel density estimates of large data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’15), pp. 1533–1542. ACM, New York (2015)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.