FORMATTING INSTRUCTIONS FOR ICLR 2027 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Kernel density estimation is a fundamental nonparameteric method for estimating probabaility densities, but computational cost scales with the number of data points. Coresets compactly represent the data while approximately preserving the kernel density estimate. We improve coreset complexity bounds for a broad class of kernels by combining a discrepancy-theoretic approach with a chaining argument. For uniformly bounded data, we give randomized polynomial-time algorithms to construct coresets of size $O\left(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log\log\frac{1}{\varepsilon}}\right)$ for the Gaussian and Laplacian kernels. For the Laplacian kernel in constant dimension, we additionally obtain coresets of size $O_d\left(\frac{1}{\varepsilon}\sqrt{\log\log\frac{1}{\varepsilon}}\right)$. We further establish bounds of $O\left(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log(2\max\{1,\alpha\})}\right)$ for the exponential, Hellinger, and Jensen-Shannon kernels, where $(1/\alpha)$ denotes the kernel bandwidth.

1 INTRODUCTION

Kernel density estimators provide a non-parametric method for estimating probability distributions. In contrast to parametric machine learning models, in which a set of training data is used to determine an optimal parameter vector to predict future outcomes without the training data, for non-parametric methods the number of parameters has the capability to grow with the size of the dataset. Our paper will focus on *kernel density estimation*, in which one estimates a probability distribution using kernel functions of the form $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. Kernels are often defined with a *bandwidth parameter* $1/\alpha$ that controls the width of the kernel function.

Given a probability distribution ρ and a collection of points $X = \{x_1, ..., x_n\} \sim \rho$ sampled independently, it is well known that for certain well-behaved kernel functions K, the distribution ρ is approximated very well by the *kernel density estimator* $\mathrm{KDE}_X(y) = \frac{1}{n}\sum_{i\in[n]} K(x_i, y)$. In particular, it is known that under certain conditions on the kernel function, KDE_X approximates ρ at the minimax optimal rate as $|X| \to \infty$ (see Tsybakov (2008)).

Although this is an elegant theoretical result, in practice it is computationally inefficient to store and make computations with arbitrarily many data points. One solution to reduce the computational complexity is to use an ε-*coreset*.

Definition 1 (KDE ε-coreset)**.** *For fixed $\varepsilon > 0$, kernel function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and data set $X \subseteq \mathcal{D}$, an* ε-coreset *for K is a subset $Q \subseteq X$ so that $\|\mathrm{KDE}_X(y) - \mathrm{KDE}_Q(y)\|_\infty \leq \varepsilon$. The* coreset complexity *of a kernel function K is the minimum possible size of a ε-coreset Q for K.*

In general, coreset complexity bounds will depend on ε and the dimension d of the kernel domain, and they will be independent of the size of the set X. These bounds are also often independent of the choice of $X \subseteq \mathcal{D}$, although several previous results and several of our results give an explicit dependence on X that may allow improvement over existing bounds for sufficiently nice data sets (see Section 1.3). In particular, several of our bounds depend on the radius of the set X. For more details about non-parametric methods and kernel density estimation, see Tsybakov (2008).

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

1.1 THE DISCREPANCY APPROACH

One powerful method for proving bounds on the coreset complexity of kernel functions is the *discrepancy approach*, previously used by Phillips & Tai (2020); Tai (2022); Phillips (2013); Karnin & Liberty (2019). This method is based on a method for computing range counting coresets (see for example Chazelle & Matousek (1996); Phillips (2013); Bentley & Saxe (1980); Phillips (2008)). Following the notational conventions of Phillips & Tai (2020), we make the following definition.

Definition 2 (Kernel Discrepancy)**.** *Given a data set $X \subseteq \mathcal{D}$, a kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and a coloring $\beta \in \{\pm 1\}^X$, the kernel discrepancy at a point $y \in \mathcal{D}$ is defined as*

$$\mathrm{disc}_K(X, \beta, y) := \Big| \sum_{x \in X} \beta(x) K(x, y) \Big|.$$

The kernel discrepancy is defined as $\mathrm{disc}_K(n) = \max_{\substack{X \subseteq \mathcal{D}: \\ |X|=n}} \min_{\beta \in \{\pm 1\}^X} \max_{y \in \mathcal{D}} \mathrm{disc}_K(X, \beta, y)$.

We will also use the notation $\mathrm{disc}_K(X)$ to denote the kernel discrepancy with respect to a fixed data set X. Bounds on kernel discrepancy can be leveraged to obtain bounds on the coreset complexity for a given kernel K via the following strategy, often called the "halving trick", as in Chazelle & Matousek (1996); Bentley & Saxe (1980); Phillips (2008). We construct a coreset of X by iteratively removing half of X, and we select which half of the points are removed by creating colorings $\beta \in \{\pm 1\}^X$ minimizing the kernel discrepancy and then removing those points assigned -1 (there are standard techniques to assume that the points are evenly split as $+1, -1$, see for example Matousek (1999)). Given an optimal choice of signs $\beta \in \{\pm 1\}^X$ such that

$$\sup_{y \in \mathcal{D}} \Big| \sum_{x \in X} \beta(x) K(x, y) \Big| \leq f(n),$$

note that, letting X^+ be the set of points assigned $+1$ and X^- be the set of points assigned -1, then under the assumption that $|X^-| = |X|/2$, for any $y \in \mathcal{D}$,

$$\begin{aligned} \frac{1}{|X|} \sum_{x \in X} \beta(x) K(x, y) &= \frac{1}{|X|} \sum_{x \in X^+} K(x, y) - \frac{1}{|X|} \sum_{x \in X^-} K(x, y) \\ &= \frac{1}{|X|} \sum_{x \in X} K(x, y) - \frac{1}{|X|/2} \sum_{x \in X^-} K(x, y). \end{aligned} \tag{1}$$

Taking a supremum over $y \in \mathcal{D}$, the final line of equation 1 is exactly $\mathrm{KDE}_X(y) - \mathrm{KDE}_{X^-}(y)$. Iterating this procedure t times, and denoting the set at iteration s by X_s (with $X_0 := X$), we find that

$$\|\mathrm{KDE}_X - \mathrm{KDE}_{X_t}\|_\infty \leq \sum_{s \in [t]} \|\mathrm{KDE}_{X_{s-1}} - \mathrm{KDE}_{X_s}\|_\infty \leq \sum_{s \in [t]} \frac{2^{s-1}}{n} f\big(n/2^{s-1}\big).$$

Assuming that the function f grows sufficiently slowly[1], this sum will be dominated by the final term, which allows us to calculate the size of a coreset yielding error at most ε. Based on this connection, our proofs focus on bounding $\mathrm{disc}_K(n)$ and $\mathrm{disc}_K(X)$ for different kernels K and datasets X, and then the "halving trick" is used to determine the corresponding size of the coreset.

1.2 POSITIVE DEFINITE KERNELS AND REPRODUCING KERNEL HILBERT SPACES

A kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is *positive definite* if given any selection of points $x_1, ..., x_m \in \mathcal{D}$, the Gram matrix G given by $G_{ij} = K(x_i, x_j)$ is positive definite. The importance of our assumption that our kernels are positive definite rests on the following famous theorem.

Theorem 3 (Aronszajn (1950))**.** *Let T be a set and K a positive definite function on $T \times T$. Then there is a map $\phi : T \to \mathcal{H}_K$ to a unique Hilbert space $\mathcal{H}_K$ so that $\forall s, t \in T$, $K(s, t) = \langle \phi(s), \phi(t) \rangle_{\mathcal{H}_K}$.*

The Hilbert space $\mathcal{H}_K$ is called the *reproducing kernel Hilbert space* (RKHS) associated to K.

[1] It suffices if $f(n) \leq n^c$ for some fixed constant $0 < c < 1$.

Definition 4. *The kernel distance associated to a kernel function K is the function*

$$D_K(x, y) := \|\phi(x) - \phi(y)\|_{\mathcal{H}_K}.$$

Here $\|h\|_{\mathcal{H}_K} = \sqrt{\langle h, h\rangle_{\mathcal{H}_K}}$ is the Euclidean norm in $\mathcal{H}_K$. In general, it is only true that D_K is a pseudometric on the domain of K; however, this is all we will need for our proofs. Almost all of the kernels commonly used in machine learning applications are positive definite.

1.3 Summary of Known Results

The problem of coreset complexity for a wide variety of kernel functions has been studied for several decades. Early approaches by Silverman (2018); Scott (1992); Joshi et al. (2011b) focused on random samples, but more recently analytic approaches and new algorithms have been discovered, leading to much stronger bounds. We focus here on previous discrepancy-based approaches, which are most directly related to our techniques. For a more comprehensive overview of coreset bounds for kernel density estimation, see the appendix.

The following results apply to *Lipschitz kernels*, that is, kernels where we can bound the *Lipschitz factor* $C_K := \max_{x,y,z\in\mathcal{D}} \frac{\|K(z,x)-K(z,y)\|_2}{\|x-y\|_2}$ of K. In the case that C_K is a small constant, as is the case for most kernels of interest, it is easy to see that taking a $2\varepsilon/(C_K\sqrt{d})$-net G_ε over the domain of K and mapping each point $x \in X$ to the closest point in G_ε (with multiplicity) to obtain X_{G_ε}, we find that $\sup_{y\in\mathcal{D}} |\mathrm{KDE}_X(y) - \mathrm{KDE}_{X_{G_\varepsilon}}(y)| \le \varepsilon$. Phillips (2013) showed a bound of $O((\alpha/\varepsilon)^{2d/(d+2)} \log^{d/(d+2)}(\alpha/\varepsilon))$ for kernels with a Lipchitz factor $C_K = O(\alpha)$. Using a sorting argument, Phillips also showed in this paper that for $d = 1$ one can achieve a coreset of size $O(1/\varepsilon)$. In general the coreset complexity is bounded below by $O(1/\varepsilon)$ Phillips (2013). Phillips & Tai (2020) improved these results by combining discretization and discrepancy approaches to give a bound of $O((1/\varepsilon)\sqrt{d\log(1/\varepsilon)})$ for any positive definite, Lipschitz kernel with *bounded influence*, a restriction similar to the impact radius condition that we define. This result applies to a very wide class of kernels, including all of the kernels that we discuss in our paper, and also the sinc kernel, for which no earlier non-trivial bounds were known. They also provide a lower bound of $\Omega(\sqrt{d}/\varepsilon)$ for $d \in [2, 1/\varepsilon^2)$, and a tight lower bound of $O(1/\varepsilon^2)$ in the case that $d \ge 1/\varepsilon^2$, for all shift and rotation invariant kernels that are somewhere-steep. Tai (2022) proved an $O_d(1/\varepsilon)$ bound for the Gaussian kernel with d constant, matching the optimal lower bound in terms of ε. This result surpresses an exponential dependence on d but is still interesting for small-dimensional data sets.

A related but not directly comparable result due to Karnin & Liberty (2019) applies to kernels that are analytic functions of the dot product and satisfy the very strong condition that $\sup_{x,x'\in\mathcal{D}} \|x-x'\|_2 \le R_K$, where R_K is a fixed constant determined by the kernel K. In this setting, they show a bound of $O(\sqrt{d}/\varepsilon)$. However, for kernels such as the Gaussian or Laplacian kernel, one can only assume that $\sup_{x,x'\in\mathcal{D}} \|x - x'\|_2 \le n\log n$, where $n = |X|$. Their result does however give a bound of $O(\alpha\exp(\alpha)\sqrt{d}/\varepsilon)$ for the exponential kernel because it has a constant diameter. Finally, Dwivedi & Mackey (2024) also uses a repeated halving, discrepancy theory based approach to address a closely related problem called *kernel thinning*. Though similar in flavor, their L^∞ guarantees are weaker than Phillips & Tai (2020), and therefore our results as well, in our setting of interest.

1.4 Our Contributions

Our results require the following two definitions.

Definition 5 (Impact Radius)**.** *Given a kernel $K : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ for $\mathcal{D} \subseteq \mathbb{R}^d$, we define the impact radius of K as $r_K(n) := \inf\{r \ge 0 : \forall x, y \in \mathcal{D},\ \|x - y\|_2 \ge r \implies |K(x, y)| \le 1/n\}$.*

Definition 6 (Query Space)**.** *Given a kernel $K : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ and a data set $X \subseteq \mathcal{D}$, we define the query space of K with respect to X as*

$$Q := \Big\{y \in \mathcal{D} :\ \exists x \in X \text{ s.t. } \|x - y\|_2 \le r_K(|X|)\Big\} = \mathcal{D} \cap \Big(\bigcup_{x\in X} (x + r_K(|X|)B_2^d)\Big).$$

Note that in general both impact radius and query space may depend explicitly on the bandwidth parameter α, but this dependence often cancels out, making the bounds obtained independent of α.

One notable exception is when the domain $\mathcal{D}$ of the kernel is compact, for example for the exponential, Hellinger, and Jensen-Shannon kernels; we return to this idea in Section 4 and Appendices C and D. We also note that if the query space Q given by a particular data set is disconnected, it suffices to consider the largest connected component of Q. To see this, note that by the definition of impact radius, the query points and data points from distinct connected components do not interact, thus we can apply the same bound to each connected component independently. For the proof of our results, we will assume that Q is connected.

Our two main results are in terms of the size of the query space and the impact radius of the kernel, respectively. One of the key challenges in earlier applications of the discrepancy approach was that the domain $\mathcal{D}$ is often $\mathbb{R}^d$, the sphere S^d, the standard $(d-1)$-dimensional simplex Δ^d, or some other potentially unbounded and/or uncountably infinite space. These domains make bounding the discrepancy challenging, as the size of these spaces make the union bound ineffective. The following lemma shows how the query space and impact radius can simplify this problem to at least ensure that the domain is bounded for kernels with sufficiently nice decay properties; for a proof see Appendix B.

Lemma 7. *Let $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be a kernel, $X \subseteq \mathcal{D}$ a data set, and Q the query space associated to K and X. Then $\operatorname{disc}_K(n) \leq \operatorname{disc}_{K|_Q}(n) + O(1)$.*

Earlier approaches using the discrepancy approach, for example Phillips & Tai (2020); Tai (2022), needed an even stronger version of Lemma 7 for Lipschitz kernels that also ensured that the query space could be made finite up to an $O(1)$ error using the Lipschitz constant of the kernel and a sufficiently small ε-net. This lemma provides an extra factor of $\sqrt{\log n}$ in many results that we will avoid by using *chaining* (see Vershynin (2018)), a multi-step construction of ε-nets. Note that in general, the query space can have volume up to $n\operatorname{Vol}_d(r_K B_2^d)$, as it is possible that the impact radii of data points do not intersect. This observation is why the results of Karnin & Liberty (2019) cannot be applied to the general Gaussian kernel, for example. Our first result gives a discrepancy bound that depends explicitly on the choice of the data set $X \subseteq \mathbb{R}^d$.

Theorem 8. *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel with bandwidth parameter $1/\alpha$ and $X \subseteq \mathbb{R}^d$ be a dataset. Denote the query space of (K, X) by Q, and define $R = R(X)$ so that $Q \subseteq RB_2^d$. If K satisfies the following properties:*

(i) K is positive definite;
(ii) K satisfies $K(x, x) = 1$ for all $x \in Q$;
(iii) $K = \kappa(\alpha\|x - y\|_2)$, where $\kappa : \mathbb{R}_{\geq 0} \to [-1, 1]$ is strictly decreasing and continuous; and
(iv) The following bound holds, where $[0, b_K)$ is the domain of the integrand:

$$\int_0^{b_K} \sqrt{-\ln\left[\kappa^{-1}\left(\frac{2-r^2}{2}\right)\right]}\,\mathrm{d}r = O(1);$$

then $\operatorname{disc}_K(X) = O(\sqrt{d\log(2\max\{R\alpha, 1\})})$. Moreover, there is a randomized algorithm that on input of X and K, finds an according coloring $\beta \in \{\pm 1\}^X$ in polynomial time.

We will see in Section 4 that Theorem 8 gives strong improvements on the current best known bounds for the Gaussian and Laplacian kernels in the case that the data set X has bounded diameter.

Corollary 9. *Let K be a kernel satisfying the conditions of Theorem 8, and let $X \subseteq \mathbb{R}^d$ be any data set. Then $\operatorname{disc}_K(X, n) = O(\sqrt{d\log(\operatorname{rad}(X) + \kappa^{-1}(1/n))})$.*

Note that because of the discretization used in Phillips & Tai (2020), even in the case where the data set is bounded uniformly, the $\sqrt{\log n}$ factor in their discrepancy bound cannot be removed. In the case where we do not account for the geometry of the data set X, we present the following stronger bound in the case that the dimension d is taken to be constant.

Theorem 10. *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel with bandwidth parameter $1/\alpha$ and $X \subseteq \mathbb{R}^d$ a data set. If K satisfies the following properties:*

(i) K is positive definite;

(ii) K satisfies $K(x, x) = 1$ for all $x \in Q$;

(iii) $K = \kappa(\alpha\|x - y\|_2)$, where $\kappa : \mathbb{R}_{\geq 0} \to [-1, 1]$ is strictly decreasing and continuous; and

(iv) The following bound holds: $\int_0^{b_K} \sqrt{-\ln\left[\kappa^{-1}\left(\frac{2-r^2}{2}\right)\right]}\,\mathrm{d}r = O(1)$*;*

then $\mathrm{disc}_K(n) = 2^{O(d)}\sqrt{\log(\kappa^{-1}(1/n))}$*. Moreover, there is a randomized algorithm that on input of* X *and* K*, finds an according coloring* $\beta \in \{\pm 1\}^X$ *in polynomial time.*

We will see in Section 4 that Theorem 10 yields significantly stronger bounds than Phillips & Tai (2020) for several important kernels in machine learning for low-dimensional data.

Our last three results demonstrate that even if the kernel is not a function of the Euclidean distance, a similar technique still provides strong bounds. In particular, we give $O(\sqrt{d})$ bounds for the Jensen-Shannon, exponential, and Hellinger kernels, with a logarithmic dependence on the bandwidth parameter α. We define $\Delta_d := \{x \in \mathbb{R}^d_{\geq 0} \mid \sum_{i\in[d]} x_i = 1\}$ as the *standard* $(d-1)$*-dimensional simplex* in $\mathbb{R}^d$.

Theorem 11. *The Jensen-Shannon (JS) kernel* $K_{JS} : \Delta_d \times \Delta_d \to \mathbb{R}$*, given by*

$$K_{JS}(x,y) = \exp\left(-\alpha\big(H\big(\tfrac{x+y}{2}\big) - \tfrac{H(x)+H(y)}{2}\big)\right),$$

where $H(x) = -\sum_{i\in[d]} x_i \log x_i$ *is the Shannon entropy function, has discrepancy satisfying* $\mathrm{disc}_{K_{JS}}(n) = O(\sqrt{d\log(2\max\{\alpha,1\})})$*. Further, the JS Kernel has coreset complexity*

$$O(\sqrt{d\log(2\max\{\alpha,1\})}/\varepsilon),$$

and such a coreset can be constructed in randomized polynomial time.

This result greatly improves on the current best bound of $O(\sqrt{d\log n})$ on the discrepancy of the JS kernel, in particular by dropping all dependence on the size n of the data set. We also have the same bound for the exponential kernel.

Theorem 12. *The exponential kernel* $K_e : S^d \times S^d \to \mathbb{R}$*, given by* $K_e(x,y) = \exp(-\alpha(1-\langle x,y\rangle))$

$$\mathrm{disc}_{K_e}(n) = O(\sqrt{d\log(2\max\{\alpha,1\})}).$$

Further, the exponential kernel has coreset complexity $O(\sqrt{d\log(2\max\{\alpha,1\})}/\varepsilon)$*, and such a coreset can be constructed in randomized polynomial time.*

As mentioned in Section 1.3, Karnin & Liberty (2019) gives a bound of $O(\alpha\exp(\alpha)\sqrt{d})$ on this quantity, hence our analysis gives a doubly-exponential improvement in terms of the bandwidth parameter α. As a corollary of Theorem 12 we obtain the same bound for the Hellinger kernel, again giving the best known result for this kernel.

Theorem 13. *The Hellinger kernel* $K_H : \Delta^d \times \Delta^d \to \mathbb{R}$*, given by*

$$K_H(x,y) = \exp\left(-\alpha\sum_{i=1}^{d}\left(\sqrt{x}-\sqrt{y}\right)\right),$$

has discrepancy satisfying $\mathrm{disc}_{K_H}(n) = O(\sqrt{d\log(2\max\{\alpha,1\})})$*. Further, the Hellinger kernel has coreset complexity* $O(\sqrt{d\log(2\max\{\alpha,1\})}/\varepsilon)$*, and such a coreset can be constructed in randomized polynomial time.*

We note that the JS, exponential, and Hellinger kernels are not shift or rotation invariant, so no better lower bound than $O(1/\varepsilon)$ is known. See Appendices C, D for proofs of Theorems 11, 12, and 13.

2 Preliminaries

2.1 Banaszczyk's Theorem and the Gram-Schmidt Walk

As we always work with positive definite kernels, by Theorem 3 each kernel K has an associated RKHS $\mathcal{H}_K$ and a map $\phi_K : \mathcal{D} \to \mathcal{H}_K$. Having mapped the data set X into $\mathcal{H}_K$ via ϕ_K, a key step of our analysis will be to apply a result of Banaszczyk (1998), in its algorithmic form due to Bansal et al. (2018), to the vectors $\{\phi_K(x) : x \in X\}$, which we observed above have norm 1 for any kernel satisfying $K(x,x) = 1$ for $x \in \mathcal{D}$. The algorithmic variant of Banaszczyk's Theorem can be stated as follows.

Theorem 14 (Gram-Schmidt Walk)**.** *There is a polynomial-time randomized algorithm that takes as input vectors $v_1, \dots, v_n \in \mathbb{R}^m$ of ℓ_2 norm at most 1 and outputs random signs $\beta \in \{\pm 1\}^n$ such that the (mean zero) random variable $\sum_{i=1}^n \beta_i v_i$ is $O(1)$-subgaussian.*

We give one of several equivalent definitions of subgaussian random variables, as well as the definition of the *subgaussian norm*, which we will need in Section 2.2 (see Vershynin (2018) for more details).

Definition 15. *A random variable X is K-*subgaussian *if the tails of X satisfy*

$$\mathbb{P}[|X| \geq t] \leq 2\exp(-t^2/K^2) \ \ \forall t \geq 0. \tag{2}$$

The subgaussian norm of X is then $\|X\|_{\psi_2} := \inf\{s > 0 : equation\ 2\ holds\ for\ K = s\}$.

2.2 Chaining and Dudley's Theorem

As mentioned in Section 1.4, Phillips & Tai (2020) introduce a $1/n$-net on the query space, which necessitates at least $O(n^d)$ data points in the query space. As the final bound is proportional to $\sqrt{\log |Q|}$, this is which is precisely the factor yielding $\sqrt{d \log n}$. To avoid this extra factor of $\sqrt{\log(n)}$, we replace the union bound-type argument of Phillips & Tai (2020) by a *chaining* approach. The term chaining refers to the fact that rather than applying an ε-net to the entire space, we continue to construct finer and finer ε-nets. For more details see Vershynin (2018).

Definition 16. *Given a (pseudo)metric space (T, d) and $r > 0$:*

- $B_d(s, r) = \{t \in T : \ d(s, t) \leq r\}$*;*
- $\operatorname{diam}(d) := \sup_{t,s \in T} d(t, s)$*;*
- $\mathcal{N}(T, d, r) = \min\left\{|S| : \ S \subseteq T \subseteq \bigcup_{s \in S} B_d(s, r)\right\}$ *is the size of a minimal r-cover of T w.r.t. d.*

Theorem 17 (Dudley's Integral Inequality)**.** *Let $(X_t)_{t \in T}$ be a random process on a pseudometric space (T, d) satisfying $\|X_t - X_s\|_{\psi_2} \leq d(t, s) \ \ \forall t, s \in T$. Then for any $t_0 \in T$,*

$$\mathbb{E} \sup_{t \in T} |X_t| \lesssim \int_0^{\operatorname{diam}(d)} \sqrt{\log \mathcal{N}(T, d, r)}\, \mathrm{d}r + \|X_{t_0}\|_{\psi_2}.$$

There is also a concentration version of this inequality.

Theorem 18. *Let $(X_t)_{t \in T}$ be a random process on a pseudometric space (T, d) satisfying $\|X_t - X_s\|_{\psi_2} \leq d(t, s)\ \forall t, s \in T$. Then, for every $u \geq 0$, the event*

$$\sup_{t,s \in T} |X_t - X_s| \leq C\Big[\, u \cdot \operatorname{diam}(T) + \int_0^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T, d, r)}\, \mathrm{d}r\Big]$$

holds with probability at least $1 - 2\exp(-u^2)$, where $C > 0$ is a universal constant.

Condition (iv) in Theorems 8 and 10 is exactly the quantity that will appear in the integral of interest once we make a comparison between the pseudometric D_K and the Euclidean distance. For this comparison we will also need the following facts about covering numbers (see for example Artstein-Avidan et al. (2015)).

Lemma 19. *Let $N(A, B)$ denote the number of copies of body B needed to cover A. For any convex set $K \subseteq \mathbb{R}^d$ and $r > 0$, one has $N(K, rB_2^d) \leq 2^d \frac{\operatorname{Vol}_d(K + \frac{r}{2} B_2^d)}{\operatorname{Vol}_d(rB_2^d)}$.*

Lemma 20. *For any symmetric convex body $P \subseteq \mathbb{R}^d$ and $r > 0$, one has $N(P, rP) \leq \left(1 + \frac{2}{r}\right)^d$.*

3 Proofs of Theorems 8 and 10

We begin with the proof of Theorem 8.

Proof. We may assume that $R\alpha \geq 1$, otherwise replace R by $\frac{1}{\alpha}$. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $X \subseteq \mathbb{R}^d$ satisfy conditions (i)-(iv) outlined in Theorem 8. As K is positive definite, there exists an RKHS $\mathcal{H}_K$ and a map $\phi : \mathbb{R}^d \to \mathcal{H}_K$ so that $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_K}$ for all $x, y \in \mathbb{R}^d$. We apply the Gram-Schmidt walk from Theorem 14 to the vectors $\phi(x)$ for $x \in X$, noting that by condition (ii), $\|\phi(x)\|_{\mathcal{H}_K} = 1$ for each $x \in X$. The algorithm yields a distribution $\mathcal{P}$ over $\{\pm 1\}^X$ so that the random variable $\Sigma := \sum_{x \in X} \beta(x)\phi(x)$, with $\beta \sim \mathcal{P}$, is $O(1)$-subgaussian. In particular, as we know that $\|\phi(y)\|_{\mathcal{H}_K} = 1$ for any $y \in Q$, the (mean zero) random variable

$$\Sigma_y := \big\langle \Sigma, \phi(y) \big\rangle = \sum_{x \in X} \beta(x) K(x, y) = \operatorname{disc}_K(X, \beta, y)$$

is $O(1)$-subgaussian. We will apply Theorem 17 to the mean zero process $(\Sigma_y)_{y \in Q}$ with the pseudometric $D_K(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_K}$, the kernel distance defined in Section 1.2. To see why D_K satisfies the condition of Theorem 17, note that for any $y, q \in Q$,

$$\operatorname{Var}[\Sigma_y - \Sigma_q]^{1/2} = \mathbb{E}[\big\langle \Sigma, \phi(y) - \phi(q) \big\rangle^2]^{1/2} \lesssim \|\phi(y) - \phi(q)\|_{\mathcal{H}_K} = D_K(y, q).$$

Thus by Theorem 17,

$$\mathbb{E} \operatorname{disc}_{K|_Q}(X) = \mathbb{E} \sup_{y \in Q} |\Sigma_y| \lesssim \int_0^{\operatorname{diam}(D_K)} \sqrt{\log \mathcal{N}(Q, D_K, r)}\, \mathrm{d}r + \|\Sigma_{y_0}\|_{\psi_2} \tag{3}$$

for any fixed $y_0 \in Q$. Here $\|\Sigma_{y_0}\|_{\psi_2} \leq O(1)$, and so we may ignore this lower order term. To estimate $\mathcal{N}(Q, D_K, r)$, we use conditions (ii) and (iii) to note that

$$D_K(q, y) = \sqrt{2 - 2\kappa(\alpha \|q - y\|_2)} \leq r \quad \iff \quad \|q - y\|_2 \leq \frac{1}{\alpha} \kappa^{-1}\left(\frac{2 - r^2}{2}\right),$$

where we use that κ is strictly decreasing. From this calculation we conclude that

$$\mathcal{N}(Q, D_K, r) = \mathcal{N}\left(Q, \|\cdot\|_2, \frac{1}{\alpha} \kappa^{-1}\left(\frac{2 - r^2}{2}\right)\right).$$

Taking $c := \frac{1}{\alpha} \kappa^{-1}\left(\frac{2-r^2}{2}\right)$ for a moment, we can bound the quantity on the right using Lemma 19:

$$\mathcal{N}\left(Q, \|\cdot\|_2, c\right) = N(Q,\ cB_2^d) \leq N(RB_2^d, cB_2^d) \leq 2^d \frac{\operatorname{Vol}_d((R + \frac{c}{2})B_2^d)}{\operatorname{Vol}_d(cB_2^d)} \leq \left(\frac{4R}{c}\right)^d, \tag{4}$$

where we use Lemma 19 and $c \leq R$ (if $c > R$ then the covering number is trivially 1). Using this bound in equation 3, we can bound $\mathbb{E} \operatorname{disc}_{K|_Q}(X)$ by

$$\begin{aligned} &\int_0^{\operatorname{diam}(D_K)} \sqrt{\log\left[\left(\frac{4R\alpha}{\kappa^{-1}((2-r^2)/2)}\right)^d\right]}\, \mathrm{d}r \lesssim \sqrt{d} \int_0^{\operatorname{diam}(D_K)} \sqrt{\log\left[\frac{4R\alpha}{\kappa^{-1}\left(\frac{2-r^2}{2}\right)}\right]}\, \mathrm{d}r \\ &\lesssim \sqrt{d \log(2R\alpha)} + \sqrt{d} \int_0^{\operatorname{diam}(D_K)} \sqrt{\log\left[\frac{1}{\kappa^{-1}\left(\frac{2-r^2}{2}\right)}\right]}\, \mathrm{d}r \lesssim \sqrt{d \log(2R\alpha)} + O(\sqrt{d}). \end{aligned} \tag{5}$$

To justify these calculations, note that in the event that the quantity inside the radical is negative, it means that $4R\alpha \leq \kappa^{-1}\left(\frac{2-r^2}{2}\right)$, and hence by equation 4 we know that $\mathcal{N}(Q, D_K, r) = 2^{O(d)}$. As $\operatorname{diam}(D_K) \leq 2$, the domain of r values for which this occurs has length ≤ 2, so we can simply restrict the domain to this point while losing at most an additive $O(\sqrt{d})$. Note that as κ is assumed strictly decreasing and continuous, $\kappa^{-1}\left(\frac{2-r^2}{2}\right)$ is an increasing, continuous function of r. Thus in particular $\kappa^{-1}\left(\frac{2-r^2}{2}\right) \to 0$ as $r \to 0$, based on our assumption that $K(x, x) = 1$ for all $x \in \mathbb{R}^d$. Finally, by Lemma 7, $\mathbb{E} \operatorname{disc}_K(X) \leq \mathbb{E} \operatorname{disc}_{K|_Q}(X) + O(1) \lesssim \sqrt{d \log(2\alpha R)}$. □

Next we prove Theorem 10.

Proof. Fix a kernel K satisfying conditions (i)-(iv) with impact radius r_K, and let $X \subseteq \mathbb{R}^d$. Fix the associated query space Q as given by Definition 6. As we assume that K is positive definite and satisfies $K(x,x) = 1$ for all $x \in Q$, there exists a map $\phi : \mathbb{R}^d \to \mathcal{H}_K$, where $\mathcal{H}_K$ is a RKHS such that $K(x,y) = \langle \phi(x), \phi(y) \rangle$ for all $x, y \in \mathbb{R}^d$, and $\|\phi(x)\|_{\mathcal{H}_K} = 1$ for all $x \in \mathbb{R}^d$. We first find a maximal set of points $q_1, \dots, q_m \in Q$ such that for any $\|q_i - q_j\|_2 \geq r_K$ for any $i, j \in [m]$. We then partition the space Q into disjoint cells $Q := R_1 \dot\cup \cdots \dot\cup R_m$ so that $R_i \subseteq q_i + r_K B_2^d$ for each $i \in [m]$. These cells also partition the set X into smaller sets $X_i := X \cap R_i$. We begin with the set X_1; because the conditions in Theorems 8 and 10 are the same, by identical arguments to the proof of Theorem 8 we obtain a random variable $\Sigma^1 := \sum_{x \in X_1} \beta^{(1)}(x)\phi(x)$, where $\beta^{(1)} \sim \{\pm 1\}^{X_1}$ is drawn according to the Gram-Schmidt walk. By Theorem 14, Σ^1 is $O(1)$-subgaussian, so in particular, for any $y \in Q$, as $\|\phi(y)\|_{\mathcal{H}_K} = 1$, the random variable $\Sigma^1(y) := \Big\langle \phi(y), \sum_{x \in X_1} \beta^{(1)}(x) \cdot \phi(x) \Big\rangle$ is also $O(1)$-subgaussian.

For each $j \in [m]$, we now have the collection of (mean zero) random variables $\{\Sigma^j(y) : y \in R_j\}$. We fix $j \in [m]$. By construction, our query space R_j has volume at most $\mathrm{Vol}_d(r_K B_2^d)$. By a calculation analogous to that in equation 4, we have that for all $r > 0$,

$$\mathcal{N}(R_j, D_K, r) \leq \max\Big\{\Big(\frac{4\alpha r_K}{r}\Big)^d, 1\Big\}. \tag{6}$$

Here we note that because of our assumption that $K(x,y) = \kappa(\alpha\|x-y\|_2)$ and the definition of $r_K(n)$, we have for x, y with $\|x - y\|_2 = r_K(n)$,

$$K(x,y) = \kappa(\alpha r_K(n)) \leq 1/n \implies \alpha r_K(n) \leq \kappa^{-1}(1/n) \implies r_K(n) \leq \frac{1}{\alpha}\kappa^{-1}(1/n). \tag{7}$$

Thus the quantity αr_K appearing in equation 6 can be bounded by a term independent of α. Note that by assumption (iii), κ^{-1} is a decreasing function. By the same analysis in the proof of Theorem 10,

$$\mathrm{disc}_{K|_{R_j}}(X_1) \lesssim \int_0^{\mathrm{diam}(D_K|_{R_j})} \sqrt{\log \mathcal{N}(R_j, D_K, r)}\, \mathrm{d}r = O(\sqrt{d \log \kappa^{-1}(1/n)}).$$

Applying Theorem 18 with $u := C_0\sqrt{d}$ yields that

$$\mathbb{P}\Big\{ \sup_{y \in R_j} |\Sigma^1(y)| \geq C \cdot C_0\sqrt{d}\Big\} \leq 2e^{-dC_0^2}, \tag{8}$$

where C is the constant from Theorem 18 and $C_0 > 0$. Moreover, by standard packing arguments, this probability is 0 for all but at most $2^{O(d)}$ choices of $j \in [m]$, as any R_j not adjacent to (or the same as) R_1 will have distance $\Omega(r_K)$ from any point $x \in X_1$, hence by the definition of the impact radius r_K and the bound in equation 7, for any such R_j the above probability is 0. We apply the union bound over the $2^{O(d)}$ cells where equation 8 could fail and obtain that for C_0 large enough,

$$\begin{aligned}\mathbb{P}\Big\{\forall j \in [m] : \sup_{y \in R_j} |\Sigma^1(y)| \geq C \cdot C_0\sqrt{d \log r_K}\Big\} &= \mathbb{P}\Big\{ \sup_{y \in Q} |\Sigma^1(y)| \geq C_0\sqrt{d \log r_K}\Big\} \\ &\leq 2^{O(d)} \cdot 2(\kappa^{-1}(1/n))^{-C_0^2 d} < 1.\end{aligned} \tag{9}$$

Thus with positive probability, $\mathrm{disc}_K(X_1) \leq C \cdot C_0\sqrt{d \log \kappa^{-1}(1/n)}$. We then fix such an outcome and repeat this construction independently for $R_2, ..., R_m$, at each step repeating the Gram-Schmidt algorithm until we get the outcome as in equation 9. After repeating this construction m times, we have a choice of signs $\beta = (\beta^{(1)}, ..., \beta^{(m)}) \in \{\pm 1\}^X$ so that by the triangle inequality

$$\mathrm{disc}_K(X) \leq \sum_{i \in [m]} \mathrm{disc}_K(X_i). \tag{10}$$

However, for our purpose this bound is too weak. But because the function $K = \kappa(\|x-y\|_2)$ is symmetric in x and y, by the same packing argument we made above, we know that for each fixed point $y \in Q$, only $2^{O(d)}$ terms in the summand in equation 10 can contribute a discrepancy larger than $O(1)$. Thus we conclude that $\mathrm{disc}_K(X) \leq 2^{O(d)}\sqrt{d \log \kappa^{-1}(1/n)}$. □

4 APPLICATIONS TO SPECIFIC KERNELS OF INTEREST

The Gaussian and Laplacian kernels are two of the most important and most commonly used kernels in machine learning (see Scott (1992)). The Gaussian kernel $K_G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined by $K_G(x, y) = \exp(-\alpha^2\|x - y\|_2^2)$, and the Laplacian kernel $K_L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined by $K_L(x, y) = \exp(-\alpha\|x - y\|_2)$ where again $1/\alpha > 0$ is the bandwidth parameter of the kernels. Currently the best known bound for both the Gaussian and the Laplacian kernels in arbitrary dimension d is $O(\sqrt{d \log n})$, and because these bounds rely on taking $1/n$-nets of the query space, even if the given data set X satisfies nice properties such as boundedness, the $\sqrt{\log n}$ term remains. However, Theorem 8 allows us to give significant improvements on this $O(\sqrt{d \log n})$ bound given such a data set. Note that for all of the following results, the coreset complexity bound follows immediately by applying the halving trick described in Section 1.1.

Corollary 21. *For the Gaussian kernel K_G with any bandwidth parameter $\alpha > 0$ and any data set $X \subseteq RB_2^d$ for any fixed constant $R > 0$, we have $\operatorname{disc}_{K_G}(X) = O(\sqrt{d \log \log n})$. In particular, one can find (in randomized polynomial time) a coreset for X of size $O\big(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log \log \frac{1}{\varepsilon}}\big)$.*

Proof. We first check the conditions of Theorem 8 for the Gaussian kernel. It is well-known that K_G is positive definite, and clearly $K_G(x, x) = e^0 = 1$ for any $x \in \mathbb{R}^d$. Taking $\kappa_G(z) := e^{-z^2}$ gives us $K_G(x, y) = \kappa_G(\alpha\|x - y\|_2)$, and κ is strictly decreasing on $\mathbb{R}^+$; finally,

$$\sqrt{-\log\left[\kappa^{-1}\left(\tfrac{2-r^2}{2}\right)\right]} \lesssim \sqrt{\log\left[\frac{1}{\log\left(\frac{2}{2-r^2}\right)}\right]}$$

is dominated by $O(1/r^2)$ for $r \in \mathbb{R}^+$, hence is integrable on its domain. The query space Q is given by $Q := \bigcup_{x \in X} B_2^d(x, r_{K_G}) \subseteq O(r_{K_G})B_2^d$, as we assume that $\operatorname{radius}(X) = O(1)$. Thus Theorem 8 yields $\operatorname{disc}_{K_G}(X) \lesssim O(\sqrt{d \log R\alpha}) = O(\sqrt{d \log \log n})$ by our assumption that $R = O(r_K)$ and the same observation as in the proof of Theorem 10 that $\alpha r_{K_G} \leq \kappa_G^{-1}(1/n) = \sqrt{\log n}$. □

An identical argument for the discrepancy yields the proof of Corollary 9 stated in Section 1.4. However, if the diameter of the set X depends on the number of data points, one has to be more careful obtaining bounds on coreset complexity, as the dependence may change as we remove points.

The following theorem follows from almost identical arguments to those in the previous proof, noting that the Laplacian kernel can be written as $K_L(x, y) = \kappa_L(\alpha\|x - y\|_2)$, with $\kappa_L(z) = e^{-z}$.

Corollary 22. *For the Laplacian kernel K_L with any $\alpha > 0$ and any data set $X \subseteq RB_2^d$ for any fixed constant $R > 0$ we have $\operatorname{disc}_{K_L}(X) = O(\sqrt{d \log \log n})$. In particular, one can find (in randomized polynomial time) a coreset for X of size $O\big(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log \log \frac{1}{\varepsilon}}\big)$.*

Thus in the interest of applications, we can invoke useful properties of the data set X to obtain better bounds, something that was not possible due to the discretization technique used in Phillips & Tai (2020). Karnin & Liberty (2019) provide bounds assuming that the entire query space lies in a bounded region (for a particular constant depending on the kernel), but our assumption is much weaker: we are only interested in the geometry of the data set, not the query space.

For constant d, Theorem 10 yields the following improvement on the current best known bound for the Laplacian kernel, due to Phillips & Tai (2020). The same bound follows for the Gaussian kernel, though Tai (2022) gives a bound of $O(1)$ on the discrepancy of the Gaussian kernel for d constant. Tai's technique does not appear to generalize to other kernels, whereas our technique applies to a broader class of kernels.

Corollary 23. *For the Laplacian kernel K_L with any $\alpha > 0$, we have $\operatorname{disc}_{K_L}(n) = 2^{O(d)}\sqrt{\log \log n}$. In particular, given any data set $X \subseteq \mathbb{R}^d$, we can find (in randomized polynomial time) a coreset of size $\frac{2^{O(d)}}{\varepsilon}\sqrt{\log \log \frac{1}{\varepsilon}}$.*

Proof. The verification of conditions (i)-(iv) of Theorem 10 is essentially identical to showing that the Gaussian kernel satisfies these properties. The Laplacian kernel can be written as $K_L(x, y) = \kappa_L(\alpha\|x - y\|_2)$, where $\kappa_L(z) = e^{-z}$. Applying Theorem 10 then yields the bound. □

AI USE STATEMENT

No generative AI tools were used in the research or preparation of this work.

REFERENCES

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950. URL https://api.semanticscholar.org/CorpusID:54040858.

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic Geometric Analysis, Part I*. Mathematical surveys and monographs. American Mathematical Society, 2015. ISBN 9781470421939. doi: 10.1090/surv/202. Includes bibliographical references and indexes.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1355–1362, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Wojciech Banaszczyk. Balancing vectors and Gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The Gram-Schmidt walk: A Cure for the Banaszczyk Blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pp. 587–597, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188850. URL https://doi.org/10.1145/3188745.3188850.

Jon Louis Bentley and James B Saxe. Decomposable searching problems i. static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980. ISSN 0196-6774. doi: https://doi.org/10.1016/0196-6774(80)90015-2. URL https://www.sciencedirect.com/science/article/pii/0196677480900152.

Bernard Chazelle and Jiri Matousek. On linear-time deterministic algorithms for optimization problems in fixed dimension. *Journal of Algorithms*, 21(3):579–597, 1996. ISSN 0196-6774. doi: https://doi.org/10.1006/jagm.1996.0060. URL https://www.sciencedirect.com/science/article/pii/S0196677496900607.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 109–116, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Efren Cruz and Clayton Scott. Sparse approximation of a kernel mean. *IEEE Transactions on Signal Processing*, PP, 03 2015. doi: 10.1109/TSP.2016.2628353.

Raaz Dwivedi and Lester Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25 (152):1–77, 2024.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42 (6):2301 – 2339, 2014. doi: 10.1214/14-AOS1252. URL https://doi.org/10.1214/14-AOS1252.

Nick Harvey and Samira Samadi. Near-optimal herding. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 1165–1182, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/harvey14.html.

Sarang Joshi, Raj Varma Kommaraji, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, SoCG '11, pp. 47–56, New York, NY, USA, 2011a. Association for Computing Machinery. ISBN 9781450306829. doi: 10.1145/1998196.1998204. URL https://doi.org/10.1145/1998196.1998204.

Sarang Joshi, Raj Varma Kommaraji, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, SoCG '11, pp. 47–56, New York, NY, USA, 2011b. Association for Computing Machinery. ISBN 9781450306829. doi: 10.1145/1998196.1998204. URL https://doi.org/10.1145/1998196.1998204.

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1975–1993. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/karnin19a.html.

Simon Lacoste-Julien, Fredrik Lindsten, and Francis R. Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. *ArXiv*, abs/1501.02056, 2015. URL https://api.semanticscholar.org/CorpusID:215825724.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1452–1461. JMLR.org, 2015.

Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 1999.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. 2017.

Jeff Phillips and Wai Tai. Near-optimal coresets of kernel density estimates. *Discrete and Computational Geometry*, 63, 06 2020. doi: 10.1007/s00454-019-00134-6.

Jeff M. Phillips. Algorithms for ε-approximations of terrains. In *Proceedings of the 35th International Colloquium on Automata, Languages and Programming - Volume Part I*, ICALP '08, pp. 447–458, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3540705740. doi: 10.1007/978-3-540-70575-8_37. URL https://doi.org/10.1007/978-3-540-70575-8_37.

Jeff M. Phillips. ε-samples for Kernels. In *Proceedings of the Twenty-Fourth Annual Symposium on Discrete Algorithms*, pp. 1622–1632. SIAM, 2013.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992. ISBN 9780471547709. URL https://books.google.hu/books?id=7crCUS_F2ocC.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Le Song, Xinhua Zhang, Alex Smola, Arthur Gretton, and Bernhard Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 992–999, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390281. URL https://doi.org/10.1145/1390156.1390281.

Wai Ming Tai. Optimal Coreset for Gaussian Kernel Density Estimation. In Xavier Goaoc and Michael Kerber (eds.), *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 63:1–63:15, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-227-3. doi: 10.4230/LIPIcs.SoCG.2022.63. URL https://drops.dagstuhl.de/opus/volltexte/2022/16071.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527. URL https://books.google.hu/books?id=mwB8rUBsbqoC.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

A Appendix A: Summary of Non-discrepancy based KDE Coreset Results

As mentioned in Section 1.3, the problem of coreset complexity for various kernels has been studied for several decades. Early approaches by Silverman (2018); Scott (1992); Joshi et al. (2011b) focused on random samples, but more recently analytic approaches and new algorithms have been discovered, leading to much stronger bounds. In the main paper we focused on discrepancy-based approaches, as they are direct precursors to our work. Here, we describe in more depth other related results derived using different techniques.

In terms of sampling-based approaches, Joshi et al. (2011a) used a sampling technique to prove a bound of $O((1/\varepsilon^2)(d + \log(1/\delta)))$ on the coreset complexity of any centrally symmetric, non-increasing kernel, where δ is the probability of failure. Fasy et al. (2014) used sampling to prove a different bound of $O((d/\varepsilon^2)\log(d\Delta/\varepsilon\delta))$, where $\Delta := \alpha \sup_{x,x' \in X} \|x - x'\|_2$ is the diameter of the data set. This result may improve upon that of Joshi et al. in the case that $K(x, x) > 1$, and it also applies to the broader class of Lipschitz kernels (see below).

The following collection of results applies to the collection of *characteristic kernels*, a subset of positive definite kernels that satisfy the additional property that the mapping ϕ_K into the associated Reproducing Kernel Hilbert Space (RKHS) is injective, which implies that the induced kernel distance D_K is a metric. This class contains many positive definite kernels, with a notable exception being the exponential kernel (see Theorem 12). Again using random sampling, Lopez-Paz et al. (2015); Muandet et al. (2017) gave a simpler bound of $O((1/\varepsilon^2)\log(1/\delta))$.

Improved results were proved using an iterative technique called *kernel herding* introduced by Chen et al. (2010) to solve a closely related problem called *kernel mean approximation*, which is shown to upper bound kernel density approximation in the case of reproducing kernels in Chen et al. (2010); Song et al. (2008). Chen et al. proved a bound of $O(1/(\varepsilon r_X))$, where r_X is the largest radius of a ball centered at $\frac{1}{|X|}\sum_{x \in X} \phi_K(x) \in \mathcal{H}_K$ that is completely contained in $\mathrm{conv}\{\phi(x) : x \in X\}$. This paper claimed that r_X is always a constant greater than 0; however, Bach et al. (2012) gave a new interpretation of the algorithm and argued that although r_X is arbitrarily small for continuous distributions, the constant $1/r_X$ is finite when X is finite. Their interpretation provided a bound of $(1/r_X^2)\log(1/\varepsilon)$. Lacoste-Julien et al. (2015) also provided a bound of $O(1/\varepsilon^2)$ on the *weighted coreset complexity*, where points in X can be assigned a non-negative weight, and this result was later improved to the setting of unweighted coresets.

Harvey & Samadi (2014) applied the kernel herding technique to an even more general problem called *general mean approximation in* $\mathbb{R}^d$ to provide bounds on the coreset complexity of order $O((1/\varepsilon)\sqrt{n}\log^{2.5}(n))$, where $n = |X|$. The dependence on n is introduced by the worst case outcome of manipulating r_X using affine scaling. Lacoste-Julien et al. (2015) showed that actually one can take $n = O(1/\varepsilon^2)$ which improves the bound to $O((1/\varepsilon^2)\log^{2.5}(1/\varepsilon))$.

Finally, Cruz & Scott (2015)'s work on the sparse kernel mean problem combined with the discretization argument above implies a bound of $O((\Delta/\varepsilon)^d)$ on the coreset-complexity, in the case that Δ is bounded and K is a Lipschitz kernel.

B Appendix B: Proof of Lemma 7

In this section we prove Lemma 7, restated below.

Lemma 24. *Let $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be a kernel, $X \subseteq \mathcal{D}$ a data set, and Q the query space associated to K and X. Then*

$$\mathrm{disc}_K(n) \leq \mathrm{disc}_{K|_Q}(n) + O(1).$$

Proof. To prove the lemma, it suffices to show that for any $y \in \mathcal{D} \setminus Q$, $\mathrm{disc}_K(X, \beta, y) = O(1)$ for all $\beta \in \{\pm 1\}^X$. But this follows immediately, as we know by the definition of Q that for any $y \in \mathcal{D} \setminus Q$ and $x \in X$, $|K(x, y)| \leq 1/n$. Thus

$$\mathrm{disc}_K(X, \beta, y) = \Big| \sum_{x \in X} \beta(x) K(x, y) \Big| \leq \sum_{x \in X} |K(x, y)| \leq 1.$$

□

C THE EXPONENTIAL AND THE HELLINGER KERNEL

In this section we apply Theorem 8 to obtain improved bounds for the exponential kernel

$$K_e : S^d \times S^d \to \mathbb{R}^d, \quad K_e(x,y) = \exp(-\alpha(1 - \langle x, y\rangle))$$

and the Hellinger kernel

$$K_H : \Delta_d \times \Delta_d \to \mathbb{R}, \quad K_H(x,y) = \exp\Big(-\alpha \sum_{i\in[d]} (\sqrt{x_i} - \sqrt{y_i})^2\Big).$$

We restate Theorem 12 for clarity.

Theorem. *The exponential kernel K_e with bandwidth parameter $\alpha > 0$ has kernel discrepancy satisfying*

$$\operatorname{disc}_{K_e}(n) = O(\sqrt{d\log(2\max\{\alpha,1\})})$$

In particular, we can find in randomized polynomial time a coreset of size

$$O\left(\frac{\sqrt{d}}{\varepsilon}\sqrt{d\log(2\max\{\alpha,1\})}\right).$$

Proof of Theorem 12. First we note that $x, y \in S^d$, hence in particular $\|x\|_2 = \|y\|_2 = 1$, so we can re-write the exponential kernel as

$$K_e(x,y) = \exp\Big(-\frac{\alpha}{2}\|x-y\|_2^2\Big) = K_G|_{S^d},$$

from which we see that K_e satisfies the conditions of Theorem 8. As the query space is $Q = S^d \subset B_2^d$, we can apply Theorem 8 with $R = 1$. □

We now argue that the same bound applies to the Hellinger kernel.

Lemma 25. *The Hellinger kernel K_H has discrepancy at most that of the exponential kernel K_e, i.e.*

$$\operatorname{disc}_{K_H}(n) \le \operatorname{disc}_{K_e}(n).$$

Proof of Lemma 25. Given a data set $X \subseteq \Delta_d$ of size n, we make the transformation

$$f : \mathbb{R}^d \to \mathbb{R}^d, \quad (x_1, \dots, x_d) \to (\sqrt{x_1}, \dots, \sqrt{x_d}).$$

Here we make two observations: first, f maps Δ^d into S^d, and second, for $x, y \in S^d$, and

$$K_H(x,y) = \exp\Big(-\alpha \sum_{i\in[d]} (\sqrt{x_i} - \sqrt{y_i})^2\Big) = \exp(-\alpha\|f(x) - f(y)\|_2^2) = K_e(f(x), f(y)).$$

By the first observation, we can apply our bound on the exponential kernel discrepancy of the set $\{f(x) : x \in X\} \subseteq S^d$ to find signs $\beta \in \{\pm 1\}^X$ so that

$$\begin{aligned}\sup_{y\in S^d}\Big|\sum_{x\in X}\beta(x)K_e(f(x),y)\Big| &= \sup_{z\in\Delta^d}\Big|\sum_{x\in X}\beta(x)K_e(f(x),f(z))\Big| \\ &= \sup_{z\in\Delta^d}\Big|\sum_{x\in X}\beta(x)K_H(x,z)\Big| \\ &= \operatorname{disc}_{K_H}(X)\end{aligned}$$

satisfies the bound of Theorem 12. Thus we conclude that

$$\operatorname{disc}_{K_H}(X) \le \operatorname{disc}_{K_e}(X),$$

and as X was an arbitrary data set, the lemma follows.

□

Theorem 13 then follows immediately from Theorem 12 and Lemma 25.

D APPENDIX C: THE JENSEN-SHANNON KERNEL

In this section we prove the following bound for a specific kernel that does not directly satisfy the conditions of Theorems 8 and 10: the Jensen-Shannon (JS) Kernel, defined by

$$K_{JS} : \Delta_d \times \Delta_d \to \mathbb{R} \quad K_{JS}(x, y) = \exp\left[-\alpha\left(H\left(\frac{x+y}{2}\right) - \frac{H(x)+H(y)}{2}\right)\right],$$

where $H(z) = -\sum_{i\in[d]} z_i \ln z_i$ is the Shannon entropy function. Note that the quantity in the exponent of the JS Kernel, $H\left(\frac{x+y}{2}\right) - \frac{H(x)+H(y)}{2}$, is precisely the Jensen gap for the Shannon entropy function (hence the name Jensen-Shannon kernel).

The Jensen-Shannon kernel, as well as the Hellinger kernel introduced in Section 4, are commonly used in information theory, as they represent similarity between histograms (i.e. the data points are discrete probability distributions over d variables). One example from text analysis Phillips & Tai (2020) is encoding with each data point X, corresponding to a text, the probabilities of d certain words appearing in that text. Although the JS kernel does not satisfy many of the conditions of our theorems, a similar proof strategy yields Theorem 11.

To prove this result, we will need the following lemmas bounding the one-dimensional entropy function, which we will denote by $h(x) := x\ln(\frac{1}{x}) = -x\ln(x)$.

Lemma 26. *For $a, \delta \geq 0$ one has $|2h(a+\delta) - (h(a) + h(a+2\delta))| \leq 3\delta$.*

Proof. By continuity of h it suffices to prove the inequality for the case of $a > 0$. We have

$$\begin{aligned}
& |2h(a+\delta) - h(a) - h(a+2\delta)| \\
= \; & |2(a+\delta)\ln(a+\delta) - a\ln(a) - (a+2\delta)\ln(a+2\delta)| \\
= \; & \left|2(a+\delta)\cdot\left(\ln(a) + \ln\left(1+\frac{\delta}{a}\right)\right) - a\ln(a) - (a+2\delta)\cdot\left(\ln(a) + \ln\left(1+\frac{2\delta}{a}\right)\right)\right| \\
= \; & \left|2(a+\delta)\ln\left(1+\frac{\delta}{a}\right) - (a+2\delta)\ln\left(1+\frac{2\delta}{a}\right)\right| \\
\overset{\text{triangle ineq.}}{\leq} \; & a\cdot\underbrace{\left|2\ln\left(1+\frac{\delta}{a}\right) - \ln\left(1+\frac{2\delta}{a}\right)\right|}_{\leq \delta/a} + 2\delta\cdot\underbrace{\left|\ln\left(1+\frac{\delta}{a}\right) - \ln\left(1+\frac{2\delta}{a}\right)\right|}_{\leq 1} \leq 3\delta
\end{aligned}$$

Here one can verify that $|2\ln(1+z) - \ln(1+2z)| \leq z$ for all $z \geq 0$ which we use to bound the left term. To bound the right term we use that $|\ln(1+z) - \ln(1+2z)| \leq 1$ for all $z \geq 0$. □

Lemma 27. *For $a, b \geq 0$ one has $|h(\frac{a+b}{2}) - \frac{h(a)+h(b)}{2}| \leq \frac{3}{4}|a-b|$.*

Proof. Suppose $b \geq a$. Applying Lemma 26 with a and $\delta := \frac{b-a}{2}$ gives

$$\left|2h\Big(\frac{a+b}{2}\Big) - (h(a)+h(b))\right| \leq 3\cdot\frac{b-a}{2} = \tfrac{3}{2}|a-b|.$$

The claim follows after dividing both sides by 2. □

We restate Theorem 11 here for clarity.

Theorem. *The Jensen-Shannon (JS) kernel $K_{JS} : \Delta_d \times \Delta_d \to \mathbb{R}$, given by*

$$K_{JS}(x, y) = \exp\left(-\alpha\big(H\big(\tfrac{x+y}{2}\big) - \tfrac{H(x)+H(y)}{2}\big)\right),$$

where $H(x) = -\sum_{i\in[d]} x_i \log x_i$ is the Shannon entropy function, has discrepancy satisfying $\mathrm{disc}_{K_{JS}}(n) = O(\sqrt{d\log(2\max\{\alpha, 1\})})$. Further, the JS Kernel has coreset complexity

$$O(\sqrt{d\log(2\max\{\alpha, 1\})}/\varepsilon),$$

and such a coreset can be constructed in randomized polynomial time.

Proof of Theorem 11. As the JS Kernel is positive definite, there exists a map $\phi : \Delta^d \to \mathcal{H}_{K_{JS}}$, where $\mathcal{H}_{K_{JS}}$ is a RKHS for K_{JS}, i.e. for any choice of $x, y \in \Delta_d$,

$$K_{JS}(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_{K_{JS}}}.$$

Then as $\|\phi(x)\|_{\mathcal{H}_{K_{JS}}} = e^0 = 1$ for any choice of $x \in \Delta^d$, we can apply the Gram-Schmidt walk to the collection of vectors $\{\phi(x)\}_{x \in \Delta_d}$ exactly as in the proofs of Theorems 8 and 10 to obtain a distribution $\mathcal{D}$ over $\{\pm 1\}^X$ and a corresponding family of $O(1)$-subgaussian random variables

$$\Sigma_y := \operatorname{disc}_{K_{JS}}(X, \beta, y), \quad y \in \Delta_d, \ \beta \sim \mathcal{D}.$$

Then applying Theorem 17 to this collection of random variables with the pseudometric

$$D_{K_{JS}}(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_{K_{JS}}} = \sqrt{2 - 2K_{JS}(x, y)},$$

we find that

$$\mathbb{E} \operatorname{disc}_{K_{JS}}(X) \lesssim \int_0^{\operatorname{diam}(D_{K_{JS}})} \sqrt{\log \mathcal{N}(\Delta_d, D_{K_{JS}}, r)} \, \mathrm{d}r. \tag{11}$$

We focus on bounding the term $\mathcal{N}(\Delta_d, D_{K_{JS}}, r)$, for which we will use Lemma 27. First, note that for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, breaking the d-dimensional entropy function H down component-wise as $H(x) = \sum_{i \in [d]} h(x_i)$:

$$D_{K_{JS}}(x, y) \le r \iff \sum_{i \in [d]} \left(h\left(\frac{x_i + y_i}{2} \right) - \frac{h(x_i) + h(y_i)}{2} \right) \le \frac{1}{\alpha} \log\left(\tfrac{2}{2 - r^2} \right). \tag{12}$$

By Lemma 27, we know that for each $i \in [d]$,

$$h\left(\frac{x_i + y_i}{2} \right) - \frac{h(x_i) + h(y_i)}{2} \le |x_i - y_i|,$$

hence

$$\sum_{i \in [d]} \left(h\left(\frac{x_i + y_i}{2} \right) - \frac{h(x_i) + h(y_i)}{2} \right) \le \sum_{i \in [d]} |x_i - y_i| = \|x - y\|_1.$$

In particular, by equation 12, $\|x - y\|_1 \le \frac{1}{\alpha} \log\left(\frac{2}{2 - r^2} \right) =: c$ implies that $D_{K_{JS}}(x, y) \le r$, so we have the bound

$$\mathcal{N}(\Delta_d, D_{K_{JS}}, r) \le \mathcal{N}(\Delta_d, \|\cdot\|_1, c) \le N(B_1^d, cB_1^d) \le \left(1 + \frac{2}{c}\right)^d \le \left(\frac{3}{c}\right)^d$$

for $c \le 1$ using Lemma 20 (and if $c > 1$, the covering number is 1 anyway). Returning to equation 11, we conclude

$$\mathbb{E} \operatorname{disc}_{K_{JS}}(X) \lesssim \int_0^{\operatorname{diam}(D_K)} \sqrt{\log \left[\left(\frac{3\alpha}{\log \frac{2}{2 - r^2}} \right)^d \right]} \, \mathrm{d}r \lesssim \sqrt{d \log(2 \max\{1, \alpha\})},$$

by the same argument as in the proofs of Theorems 8 and 10. □