Dimension-Independent Kernel ε-Covers

Jeff M. Phillips
Kahlert School of Computing, University of Utah, USA

Hasan Pourmahmood-Aghababa
Kahlert School of Computing, University of Utah, USA

Abstract

We introduce the notion of an ε-cover for a kernel range space. A kernel range space concerns a set of points $X \subset \mathbb{R}^d$ and the space of all queries by a fixed kernel (e.g., a Gaussian kernel $K(p, \cdot) = \exp(-\|p - \cdot\|^2)$, where $p \in \mathbb{R}^d$). For a point set X of size n, a query returns a vector of values $R_p \in \mathbb{R}^n$, where the ith coordinate $(R_p)_i = K(p, x_i)$ for $x_i \in X$. An ε-cover is a subset of points $Q \subset \mathbb{R}^d$ so for any $p \in \mathbb{R}^d$ that $\frac{1}{n}\|R_p - R_q\|_1 \leq \varepsilon$ for some $q \in Q$. This is a smooth analog of Haussler's notion of ε-covers for combinatorial range spaces (e.g., defined by subsets of points within a ball query) where the resulting vectors R_p are in $\{0,1\}^n$ instead of $[0,1]^n$. The kernel versions of these range spaces show up in data analysis tasks where the coordinates may be uncertain or imprecise, and hence one wishes to add some flexibility in the notion of inside and outside of a query range.

Our main result is that, unlike combinatorial range spaces, the size of kernel ε-covers is independent of the input size n and dimension d. We obtain a bound of $2^{\tilde{O}(1/\varepsilon^2)}$, where $\tilde{O}(f(1/\varepsilon))$ hides log factors in $(1/\varepsilon)$ that can depend on the kernel. This implies that by relaxing the notion of boundaries in range queries, eventually the curse of dimensionality disappears, and may help explain the success of machine learning in very high-dimensions. We also complement this result with a lower bound of almost $(1/\varepsilon)^{\Omega(1/\varepsilon)}$, showing the exponential dependence on $1/\varepsilon$ is necessary.

Keywords and phrases ε-Cover, ε-Sample, Kernel range space, Dimensionality reduction.

Digital Object Identifier 10.57717/cgt.v4i1.48

Acknowledgements Jeff Phillips thanks his support from NSF CCF-2115677, CCF-1350888, and IIS-1816149. We thank an anonymous reviewer of an earlier version of the paper for suggesting the argument based on Rademacher complexity for reproducing kernels. We also thank Nico Spronk for providing the proof of Lemma 31 through email communication.

1 Introduction

Given a data set X a *range space* $(X, \mathcal{R})$ is the collection of possible ways that set X can be queried; $\mathcal{R}$ is a set of subsets of X, often defined by intersection with a type of geometric shape. For a data structure, ranges specify the shape of any range query [1]. For machine learning, ranges categorize the function class of possible classifiers [42]. For spatial scan statistics, ranges restrict the family of regions which might form an anomalous hotspot [24].

In each of these cases, it is common to allow $\varepsilon|X|$ additive error when considering the results of these queries. In that context, an *ε-cover*, which is an instance of a cover in a metric space, and Haussler [17] among others studied it for a particular class of metric spaces, is an important concept; it is a subset $\mathcal{Q}$ of all possible subsets in the collection $(X, \mathcal{R})$ so that for any range $R \in (X, \mathcal{R})$ there exists some set $Q \in \mathcal{Q}$ so that the symmetric difference $|Q \triangle R| \leq \varepsilon|X|$. In particular, if one allows $\varepsilon|X|$ error, then one only needs to consider each of the above listed data analysis challenges with respect to the ε-cover $\mathcal{Q}$, not the full collection of possible subsets.

Haussler introduced and bounded the size of ε-covers for range spaces with bounded VC-dimension [17]. In particular, if the VC-dimension ν is bounded, then there exist ε-covers of size $O(1/\varepsilon^\nu)$ and $\Omega(1/\varepsilon^\nu)$ may be needed. Consider the common range spaces for $X \subset \mathbb{R}^d$

like half-space, ball, and fixed radius ball range spaces, each has VC-dimension $\nu = d+1$. However, Haussler's lower bound does not apply to these geometric range spaces specifically. Nevertheless, we supply a lower bound of $\Omega((1/\varepsilon)^{d^{1-o(1)}})$ for them in Section 6.

In this paper, we consider how this changes when we consider kernelized versions of these objects; that is where ranges are defined by kernels, like Gaussian kernels $K(x,q) = \exp(-\|x-q\|^2)$. Indeed, kernel SVM is a common way to build non-linear classifiers [38], and kernelized versions of data structures queries [8, 7, 23] and scan statistics [15, 14, 18] are also common. Partially motivated by these cases, the complexity of kernel range spaces have also been studied, and in particular samples for density approximation. These *ε-KDE-samples* are subsets $S \subset X$ so for every query $p \in \mathbb{R}^d$ that

$$\left| \frac{1}{|X|} \sum_{x \in X} K(x,p) - \frac{1}{|S|} \sum_{s \in S} K(s,p) \right| = |\text{KDE}_X(p) - \text{KDE}_S(p)| \leq \varepsilon.$$

While for positive and symmetric kernels, a bound of $O(d/\varepsilon^2)$ for such an ε-KDE-sample can be derived using bounds for ball range spaces [21], more remarkably, for reproducing kernels, only size $O(1/\varepsilon^2)$ is needed [27, 25, 5], that is with no dependence on d.

We tackle whether a similar result, with no dependence on n or d is possible for an ε-cover of a kernel range space. In particular, a kernel range space (X, K) is defined by a set of input points $X \subset \mathbb{R}^d$, and a fixed kernel K, e.g., Gaussians of the form $K(p,\cdot) = \exp(-\|p - \cdot\|^2)$. In this setting, any *range* in the kernel range space is defined by a point $p \in \mathbb{R}^d$, and reports a signature vector $R_p^X = (K(p,x_1), K(p,x_2), \ldots, K(p,x_n)) \in \mathbb{R}^n$, which has a scalar value $K(p,x_i)$ for each $x_i \in X$; we consider when $K(p,x_i) \in [0,1]$. This generalizes the notion of a set, where these signatures are bit-vectors from $\{0,1\}^n$ instead of $[0,1]^n$. An *ε-cover of a kernel range space* (X,K) is then a set of kernel ranges $K(q,\cdot)$, defined by a set of points $Q \subset \mathbb{R}^d$, so for *any* query point $p \in \mathbb{R}^d$ there exists a $q \in Q$ so that

$$\frac{1}{|X|} \sum_{x \in X} |K(p,x) - K(q,x)| = \frac{1}{|X|} \|R_p^X - R_q^X\|_1 \leq \varepsilon.$$

This generalizes the notion of ε-cover of Haussler to function values. Notice that if we instead placed the absolute values outside the sum, this becomes trivial since one can simply choose $1/\varepsilon$ points Q where $\text{KDE}_X(q_i) = (i-1/2)/\varepsilon$ for $i = 1, \ldots, 1/\varepsilon$.

Our results. Our main result is that ε-covers for kernel range spaces have size complexity independent of n and d. Thus for constant error (e.g., $\varepsilon = 0.01$ for 1% error), the size of the ε-cover is constant; that is, to evaluate these functions up to a fixed error, one only needs to pre-compute or consider evaluating a fixed number of kernel range queries. In particular, we show that the size of ε-covers are at most $2^{\tilde{O}(1/\varepsilon^2)}$; where $\tilde{O}(f(1/\varepsilon))$ hides polylogarithmic factors in $1/\varepsilon$. This bound works for a large class of kernels we call "standard" and includes Gaussian, Laplace, truncated Gaussian, triangle, Epanechnikov, quartic, and triweight. Moreover, we show that this $(1/\varepsilon)^{\mathsf{poly}(1/\varepsilon)}$ is necessary. In particular, for Gaussian kernels we provide a construction that requires an ε-cover of size $(\frac{1}{\varepsilon})^{\Omega(1/\varepsilon^\lambda)}$ for any $\lambda \in (0,1)$ in $\mathbb{R}^{d'}$ with $d' = \Omega(1/\varepsilon^\lambda)$.

When viewed in comparison to the ε-cover size bound for traditional range spaces, e.g. for half-spaces or balls, where the size grows exponentially in d (see discussion in Section 6), we believe this result is quite surprising. Almost all learning or data structure bounds, even approximate ones, have exponential dependence on d in the number of queries considered. However, this result shows that if one relaxes the boundary of the query, that is there is not a hard or combinatorial cut-off separating "in" the query or "not in" the query, then this exponential dependence and curse of dimensionality (eventually) disappears.

Overview of techniques. One may think (as we initially hoped) that this ε-cover result is a not-too-hard consequence of the dimension-independent bounds for ε-KDE-samples. However, these results seem to provide the wrong sorts of guarantees; they would work if the definition of the ε-cover had the absolute values outside the sum. Moreover, both constructions for ε-KDE-samples rely on properties of reproducing kernels, namely that the kernel density estimate $\textsc{kde}_X$ can be viewed as a mean in a reproducing kernel Hilbert space. This quantity turns out to be easy to approximate with sub-gradient descent [25, 5] or sampling [27]. However, the ε-cover is a richer and more structured summary of a point set, and does not admit such simple analysis.

Our approach at its core uses the simple idea that for a kernel with a bounded support (of value above ε), one can place a grid around each data point with a gap of ε between grid points. The union of all grid points is the ε-cover. Naively, this provides a bound of roughly $n(1/\varepsilon)^d$ for $n = |X|$ points in $\mathbb{R}^d$. This paper shows how to preserve the correctness of this construction while reducing both n and d to only depend on ε.

Being able to remove the dependence on n is perhaps not that surprising given the existence of ε-KDE-samples and similar data reduction results. However, this required some new adaptations on existing ideas as the direct invocation of ε-KDE-samples does not work. We connect to ε-samples of (traditional) range spaces, which we call *semi-linked* to kernels, and their VC-dimension. These semi-linked ranges are defined as the *super-level sets of the difference of two kernel functions.* A key insight is that these semi-linked range spaces allow us to calculate an intermediate object called an ε-cover-sample, via a simple random sample, and this ε-cover-sample can be converted into an ε-cover. We show for Gaussian, triangle, Epanechnikov, quartic, and triweight kernels that this VC-dimension bound is $O(d^2)$. So this reduction eliminates the dependence on size n, but adds dependence on dimension d.

More surprising to the authors is that the dependence on the dimension can be eliminated. The argument works by showing the existence of an embedding into dimension $m = O((1/\varepsilon^2)\log n)$ where the measurement of all kernels on the input point set X is preserved up to ε error. This embedding is a result of invoking terminal JL [33]. It preserves the R_p^X signatures for each point p, and means it is sufficient to create an ε-cover in that m-dimensional space. However, since the terminal JL embedding map is not invertible for points not in X, we require a new combinatorial covering argument to show that an $\varepsilon/8$-cover in $\mathbb{R}^m$ is still an ε-cover in the original $\mathbb{R}^d$. While this process eliminates the dependence on d, it increases it with respect to the number of points n.

Iterating between these two approaches would reduce both n and d, but would naively require $\log^*(nd)$ iterations, potentially inflating the error by that factor, and so not be independent of one of those two terms. Luckily, however, we can adapt a new inductive framework [11] for analyzing such iterative reduction processes, and we show a complete elimination of the dependence on n and d. For positive definite kernels we apply a Rademacher complexity bound to calculate ε-cover-samples independent of d; this does not immediately remove ε-cover dependence on d, but does sidestep some of this iterative analysis.

For the lower bound, in low dimensions, the construction works like one may expect for fixed-radius balls, which when their radius is sufficiently large, act like half-spaces. The size is trivially $\Omega(1/\varepsilon)$ in $\mathbb{R}^1$, and as we add each dimension we add a point "orthogonal" to the existing dimensions. The ranges we must cover is the cross-product of these distance intervals from points in each dimension, leading to a $(1/\varepsilon)^d$ lower bound for fixed-radius balls. However, interestingly, this construction stops working for kernels as we approach $1/\varepsilon$ dimensions. This is the result of both the curvature of the level sets and the decaying contribution with distance: properties implicit in data with full-dimensional noise.

Implications. This model and result is relevant in data analysis applications where a complete trust in data coordinates is rare (e.g., due to sensing noise), and it is common to have high dimensional data, and this sort of additive ε-error is tolerated if not expected. We hope this sheds a bit of light onto why learning in such high-dimensional spaces is not as challenging as traditional curse-of-dimensionality bounds may suggest. That is, if one assumes sensing noise, adding more and more features (dimensions) does not always generate more implicit query complexity.

For instance, our new definition of kernel ε-cover is also the notion required to enumerate all possible ranges that could lead to a distinct solution in the case of approximate range searching applications or noise-aware statistical modeling (e.g., for evaluating a kriging model [34] or Nadaraya-Watson kernel regression [44]) or in enumeration for approximate spatial scan statistics [29]. Spatial scan statistics search (or "scan") *all* combinatorial distinct ranges (up to ε-error) to find one that maximizes some statistic on the data – a candidate for an anomaly. Recently, Han *et.al.* [15] defined a kernel spatial scan statistic where the ranges are replaced with kernel queries, and the goal is still to approximately maximize a function of data over all such queries; they provide a solution in $d = 2$. Our results show that eventually, the size of the space required to search stops growing exponentially with dimension.

Another line of work [3, 6, 13] attempts to bound the number of local maximums of a Gaussian KDE_X for $X \subset \mathbb{R}^d$ and $|X| = n$. A lower bound is $\binom{n}{d} + n = \Omega(n^d)$ for $n, d \geq 2$ [3], but the upper bound is not known to be finite. If one assumes finiteness, the best bound is $2^{d+\binom{n}{2}}(5 + 3d)^n$. If one only counts local maximum p, q that have $\frac{1}{|X|}\|R_p^X - R_q^X\|_1 > \varepsilon$ (i.e., are sufficiently distinct), then our result induces a bound of $O(2^{\tilde{O}(1/\varepsilon^2)})$.

Finally, this result induces the first dimension-independent bounds for ε-KDE-samples [36] for non-reproducing kernels including triangle, Epanechnikov, and truncated Gaussians.

1.1 Connection to uniform Glivenko-Cantelli classes

Starting with Vapnik and Chervonenkis [43], numerous learning theorists, probabilists, and combinatorists have studied a strong and general notion of convergence of function approximation under sampling known as the *uniform Glivenko-Cantelli* class. It concerns a class of functions $\mathcal{F}$ from a set $\mathcal{X}$ to $[0, 1]$. Then let $\mathcal{P}$ be a probability measure over $\mathcal{X}$ so that any $f \in \mathcal{F}$ is $\mathcal{P}$-measurable over $\mathcal{X}$. Next we use $\mathcal{P}(f) = \int_{x\in\mathcal{X}} f(x)\mathsf{d}\mathcal{P}$ to denote the mean of $f(x)$ for $x \sim \mathcal{P}$. Now for an independent random sample $x_1, x_2, \ldots, x_m \sim \mathcal{P}$, let $\mathcal{P}_m(f) = \frac{1}{m}\sum_{i=1}^m f(x_i)$ be its approximation by the sample of size m.

Dudley *et.al.* [12] defines that the family $\mathcal{F}$ is *ε-uniform Glivenko-Cantelli class* if

$$\lim_{m\to\infty} \sup_{\mathcal{P}} Pr[\sup_{k\geq m} \sup_{f\in\mathcal{F}} |\mathcal{P}_k(f) - \mathcal{P}(f)| > \varepsilon] = 0.$$

While bounded VC-dimension [43] implies that $\mathcal{F}$ is ε-uniform Glivenko-Cantelli, it does not completely characterize this process.

Alon *et.al.* [2] showed that a variant of VC-dimension, called *V_γ-dimension*, for any $\gamma > 0$, did characterize this form of convergence. We say $\mathcal{F}$ *V_γ-shatters* a set $A \subset \mathcal{X}$ if there exists a value $\alpha \in [0, 1]$ such that for each $E \subseteq A$, there is another function $f_E \in \mathcal{F}$ so for all $x \in A \setminus E$ then $f_E(x) \leq \alpha - \gamma$ and for all $x \in E$ then $f_E(x) \geq \alpha + \gamma$. The V_γ-dimension of $\mathcal{F}$ is the maximum cardinality $A \subseteq \mathcal{X}$ that is V_γ-shattered by $\mathcal{F}$. They showed that if the V_γ-dimension is finite, then $\mathcal{F}$ is $(b\gamma)$-uniform Glivenko-Cantelli for some constant $b \leq 48$. Moreover, if the V_γ-dimension is not finite, then $\mathcal{F}$ is not $(2\gamma - \tau)$-uniform Glivenko-Cantelli for any $\tau > 0$.

The Glivenko-Cantelli criteria is intimately tied to L_s ε-covers via a result by Dudley *et.al.* [12]. An L_s *ε-cover* is a set $F \subset \mathcal{F}$ so for any $f' \in \mathcal{F}$ it holds that some $f \in F$ satisfies $\|f - f'\|_s \leq \varepsilon$, where $\|f - f'\|_s = (\int_{x\in\mathcal{X}} |f(x) - f'(x)|^s \mathsf{d}\mathcal{P})^{1/s}$ for $0 < s < \infty$, and $\|f - f'\|_\infty = \max_{x\in\mathcal{X}} |f(x) - f'(x)|$. Let $N_s(\varepsilon, \mathcal{F}, X)$ be the size of the smallest L_s ε-cover of $(X, \mathcal{F})$. Now let

$$H_m(\varepsilon, \mathcal{F}) = \sup_{\substack{X\subset\mathcal{X}\\|X|=m}} \log_2(N_s(\varepsilon, \mathcal{F}, X)).$$

They showed that $\mathcal{F}$ is Glivenko-Cantelli if and only if $\lim_{m\to\infty} H_m(\varepsilon, \mathcal{F})/m = 0$. Moreover, if $\varepsilon > 0$ and $\lim_{m\to\infty} H_m(\varepsilon, \mathcal{F})/m = 0$, then $\mathcal{F}$ is (8ε)-uniform Glivenko-Cantelli. These results hold for any $s \in (0, \infty]$ in the definition of ε-cover.

Application to kernel range spaces. This paper studies a restrictive setting within this framework.

First we consider $\mathcal{P}$ as the uniform probability measure on a fixed size-n set X with $\mathcal{X} = \mathbb{R}^d$. Our results do allow n to become arbitrarily large, and the sample complexity results do not depend on n, so it seems these approaches may extend naturally to continuous distributions $\mathcal{P}$. However, we do not formalize this limiting case. Moreover, we describe our results algorithmically, where the algorithms take a finite size n, and have run times which depend on n.

Second, we consider a specific class of functions $\mathcal{F}_K$ where each $f_p(\cdot) = K(p, \cdot)$ takes the form of a kernel. That is each $f_p \in \mathcal{F}_K$ is parameterized by a point $p \in \mathbb{R}^d$. Then $\mathcal{P}(f_p) = \mathcal{P}(K(p, \cdot)) = \text{KDE}_X(p)$.

Moreover, one can apply a Chernoff-Hoeffding bound on any one covering element (parameterized by p) with $O((1/\varepsilon^2)\log(1/\delta))$ samples to, with probability $1 - \delta$, get ε error for any one evaluation point p. Then one can take a union bound over $N_1(\varepsilon/2, \mathcal{F}_K, X)$ covering elements, and using triangle inequality, ensure that with $k_\varepsilon = O((1/\varepsilon^2)\log(\frac{N_1(\varepsilon/2,\mathcal{F}_K,X)}{\delta})$ samples S, we have with probability at least $1 - \delta$ that

$$\sup_{f_p\in\mathcal{F}_K} |\mathcal{P}(f_p) - \mathcal{P}_m(f_p)| = \sup_{p\in\mathbb{R}^d} |\text{KDE}_X(p) - \text{KDE}_S(p)| \leq \varepsilon.$$

Classically, for binary functions families $f \in \mathcal{F}$ with $f(x) \in \{0, 1\}$, this argument also works using $s = \infty$ [43]; and in this binary setting, the L_1 and L_∞ distances are equivalent. However, for real-valued $f(x) \in [0, 1]$, the triangle inequality over $\mathcal{P}(f)$ requires an L_1 bound.

Hence for our setting, ε-uniform Glivenko-Cantelli convergence implies that a random sample $S \subset X$ of some size k_ε satisfies that $\sup_{p\in\mathbb{R}^d} |\text{KDE}_X(p) - \text{KDE}_S(p)| \leq \varepsilon$; hence a random sample S of size k_ε is an ε-KDE coreset. Thus our bound of $N_1(\varepsilon, \mathcal{F}_K, X) = 2^{\tilde{O}(1/\varepsilon^2)}$ implies that we can bound $k_\varepsilon = \tilde{O}(1/\varepsilon^4)$.

It was already known [27, 37] that $k_\varepsilon = O((1/\varepsilon^2)\log(1/\delta))$, with no dependence on n or d when the kernel K is reproducing (e.g., for Gaussian or Laplace kernels). So this reduction does not imply new ε-KDE coreset results for this class of kernels. But our main result applies to “simply computable” kernels (defined below), includes the Epanechnikov, Triangle, Quartic, Triweight, and the truncated Gaussian, which are not reproducing.

Moreover, by leveraging an intermediate result on semi-linked range spaces, we can also obtain bounds of size $k_\varepsilon = \tilde{O}(1/\varepsilon^6)$ (see Appendix A.3).

Next we discuss what the existing results [27, 37] which show k_ε being independent of n and d imply about the size of the ε-covering $N_\infty(\varepsilon, \mathcal{F}_K, X)$. Uniform Glivenko-Cantelli ensures that $\lim_{m\to\infty} H_m(\varepsilon, \mathcal{F}_K)/m = 0$. This means $H_m(\varepsilon, \mathcal{F}_K) = \sup_{\substack{X\subset\mathcal{X}\\|X|=m}} \log_2(N_1(\varepsilon, \mathcal{F}_K, X)) =$

$o(m)$, and so for $|X| = m$ then $N_1(\varepsilon, \mathcal{F}_K, X) = 2^{o(m)}$. Note that this bound can allow $N_1(\varepsilon, \mathcal{F}_K, X)$ to depend on the dimension d as long as it is fixed. Our result of $N_1(\varepsilon, \mathcal{F}_K, X) = 2^{\tilde{O}(1/\varepsilon^2)}$ is much stronger, as it implies no dependence on the $m = |X|$ or the dimension d.

Finally, we back up to the notion of V_γ-shattering, and consider the implications of setting $\gamma = \varepsilon$. In our setting, a V_ε-shattering dimension of D would imply (using the contrapositive of the above definition) that for each $x \in X$ that for any set of $D + 1$ (or more) points $A \subset X$, then no value $\alpha \in [0, 1]$ can have all subsets $E \subseteq A$ ε-separated. That is, there must be some subset $E \subseteq A$ so there is not a function $f_E \in \mathcal{F}_K$ so for $x \in E$ that $f_E(x) \leq \alpha - \varepsilon$, and for $x \in A \setminus E$ then $f_E(x) \geq \alpha + \varepsilon$. So a V_ε-shattering of $\mathcal{F}_K$ dimension of $D = \tilde{O}(1/\varepsilon^2)$ would imply that for $D + 1$ or more points, that there would always be some subset A that cannot be ε-separated by a kernel $K(p, \cdot)$. This is not what our main technical lemma shows. Instead, it uses a version of an L_1 distance, that shows that $K(p, \cdot)$ and $K(p', \cdot)$ differ *on average* over $x \in X$ by at most ε for p in an ε-cover, and p' as any point in $\mathbb{R}^d$. Whereas the Alon *et.al.* [2] paper uses a less discriminative L_∞ distance that requires $K(p, \cdot)$ and $K(p', \cdot)$ to differ *on all* points $x \in X$ by ε.

In summary, the famous Glivenko-Cantelli analysis of Alon *et.al.* [2] provides a complete analysis of uniform convergence rates, but does not provide a finite sample size bound k_ε necessary to achieve an ε error. In order to completely characterize this rate bound, they rely on a stronger L_∞ type of distance between functional ranges; however, to provide a finite sample bound, we show we only need an L_1 variant of the distance between these functional ranges. By restricting ourselves to this L_1 distance, we are able to show finite sample bounds for k_ε that are independent of n and d for a broad class of kernel range spaces defined over n points in $\mathbb{R}^d$.

2 Preliminaries: kernels, range Spaces, and covers

There can be many definitions of a kernel, we start by specifying what properties are needed in this work. A symmetric bi-variate function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is *centrally symmetric* if $K(p, x) = g(\|x - p\|)$, where $g : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ is a continuous function. We say that K is L-Lipschitz if g is. Given $\varepsilon > 0$, the ε-*critical radius* $r = r(\varepsilon)$ of K is the smallest positive real number such that $g(r') < \varepsilon$ for any $r' > r$. The ball $B_r(x) = \{p \in \mathbb{R}^d \mid \|p - x\| \leq r\}$ with $r = r(\varepsilon)$ is then called the ε-*critical ball* around x. We call K a (L, r)-*standard kernel* if it is a non-negative, L-Lipschitz, centrally symmetric function with critical radius r such that $r(\varepsilon/2) = O(r(\varepsilon))$ and for $\varepsilon < 1/2$, $r(\varepsilon) > C$ for some absolute constant $C > 0$. Standard kernels are continuous and bounded, which implies without loss of generality, by normalizing, we may assume that they take a maximum value of 1. The most common standard kernel is the Gaussian kernel $K(x, y) = e^{-\|x-y\|^2/\sigma^2}$; see others in Table 1, where the parameter $\sigma > 0$ is elsewhere assumed $\sigma = 1$.

A kernel K is called k-*simply computable*, for some constant positive integer k, if for any $p, q, x \in \mathbb{R}^d$ and $\tau \in \mathbb{R}^+$, the inequality $|K(p, x) - K(q, x)| \geq \tau$ can be verified in $O(d^{k-1})$ steps using the "simple" arithmetic operations $+, -, \times$, and $/$, jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisons, and $O(1)$ evaluations of the exponential function $z \mapsto e^z$ on reals. K is called *simply computable* if it is k-simply computable for a constant k. We will mostly work with simply computable kernels. It is easy to see Gaussian, Epanechnikov, quartic and triweight kernels are 2-simply computable; Theorem 32 in Appendix A.6 shows triangle kernels are also 2-simply computable. It is not clear if Laplace kernels are simply computable, but they are positive definite, which we handle separately in Theorems 13 and 19.

Kernel	Rule: $K(x,y)$	L	r	k	ε-cover sizes	Theorem
Gaussian	$e^{-\|x-y\|^2/\sigma^2}$	$\frac{\sqrt{2/e}}{\sigma}$	$\sigma\sqrt{\ln(1/\varepsilon)}$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log^2(\frac{1}{\varepsilon}))}$	17, 19
Laplace	$e^{-\|x-y\|/\sigma}$	$1/\sigma$	$\sigma\ln(1/\varepsilon)$	3	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log^3(\frac{1}{\varepsilon}))}$	19
Epanechnikov	$\max\{0, 1-\frac{\|x-y\|^2}{\sigma^2}\}$	$2/\sigma$	$\sigma\sqrt{1-\varepsilon}$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))}$	17
Triangle	$\max\{0, 1-\frac{\|x-y\|}{\sigma}\}$	$1/\sigma$	$\sigma(1-\varepsilon)$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))}$	17
Quartic	$\max\{0, 1-\frac{\|x-y\|^2}{\sigma^2}\}^2$	$\frac{8}{3\sqrt{3}\sigma}$	$\sigma\sqrt{1-\sqrt{\varepsilon}}$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))}$	17
Triweight	$\max\{0, (1-\frac{\|x-y\|^2}{\sigma^2})^3\}$	$\frac{96}{25\sqrt{5}\sigma}$	$\sigma\sqrt{1-\sqrt[3]{\varepsilon}}$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))}$	17
Trun-Gaussian	$\frac{1}{1-\tau}(e^{-\|x-y\|^2/\sigma^2}-\tau)$	$\frac{\sqrt{2/e}}{\sigma}$	$\sigma\sqrt{\ln(\frac{1}{\tau+(1-\tau)\varepsilon})}$	2	$(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log^2(\frac{1}{\varepsilon}))}$	17

Table 1 Examples of standard kernels with L, r, k values and their d, n-free ε-cover sizes.

2.1 Generalized range spaces and ε-covers

Recall that the symmetric difference of two sets A and B is $A \triangle B = (A \cup B) \setminus (A \cap B)$. Let $(X, \mathcal{A})$ be a range space, where $X \subset \mathbb{R}^d$ is a finite set, and let $\varepsilon > 0$. A range space $(X, \mathcal{A}_\triangle)$ is called an *ε-covering* of $(X, \mathcal{A})$ if $\mathcal{A}_\triangle \subset \mathcal{A}$, and for any $A \in \mathcal{A}$ there is an $A' \in \mathcal{A}_\triangle$ such that $|(X \cap A) \triangle (X \cap A')| \leq \varepsilon|X|$ [28]; that is the difference in elements that A and A' cover is $\leq \varepsilon|X|$.

Consider the n-dimensional hypercube

$$\mathbb{W}^n = [0,1]^n = \{w = (w_1, \ldots, w_n) \in \mathbb{R}^n : 0 \leq w_i \leq 1 \ (1 \leq i \leq n)\},$$

equipped with the normalized L_1-distance

$$d_\triangle(w, w') = \tfrac{1}{n}\|w - w'\|_1 = \tfrac{1}{n}\textstyle\sum_{i=1}^n |w_i - w'_i|,$$

where $w = (w_1, \ldots, w_n), w' = (w'_1, \ldots, w'_n) \in \mathbb{W}^n$.

Generalized ε-covers. For a point set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$, a query point $p \in \mathbb{R}^d$ defines a signature vector

$$R_p^{X,K} = (K(p, x_1), \ldots, K(p, x_n)) \in \mathbb{W}^n.$$

The *generalized symmetric difference distance* (with respect to K and X) on $\mathbb{R}^d$ is defined by

$$d_\triangle^{X,K}(p,q) = d_\triangle(R_p^{X,K}, R_q^{X,K}) = \tfrac{1}{|X|}\textstyle\sum_{x \in X} |K(p,x) - K(q,x)|.$$

Since the kernel K will be fixed, for simplicity, we remove it from the superscripts of $R_p^{X,K}$ and $d_\triangle^{X,K}(p,q)$. We may also remove the superscript X if there is no ambiguity. For $S \subset X$ if we set $w_i = 1$ when $x_i \in S$ and 0 otherwise (i.e., w is a corner of $\mathbb{W}^n$), and similarly for $S' \subset X$, then $d_\triangle(w, w') = |S \triangle S'|/n$.

For a kernel range space (X, K), an *ε-cover* is a set $Q \subset \mathbb{R}^d$ such that for every $p \in \mathbb{R}^d$ there exists a $q \in Q$ such that $d_\triangle^X(p,q) \leq \varepsilon$. If K is (L, r)-standard, then

$$d_\triangle^X(p,q) = \frac{1}{n}\sum_{i=1}^n |K(p,x_i) - K(q,x_i)| \leq \frac{L}{n}\sum_{i=1}^n \|p-q\| = L\|p-q\|;$$

hence a sufficient condition for ε-cover Q is for all $p \in \mathbb{R}^d$ to have some $q \in Q$ so $\|p-q\| \leq \varepsilon/L$.

Next we can restrict this condition further to just query points p in the critical ball defined by K around each $x_i \in X$, i.e., $p \in \tilde{B} = \bigcup_{i=1}^n B_r(x_i)$, where r is the ε-critical radius of K. Otherwise, i.e. if $p, q \notin \tilde{B}$, both $K(p, x_i)$ and $K(q, x_i)$ are less than ε and so the inequality $d^X_\triangle(p, q) \le \varepsilon$ holds trivially. In fact, one point $q \notin \tilde{B}$ can cover all points in $\mathbb{R}^d \setminus \tilde{B}$ as an ε-cover. We denote this point by q_∞. That is, if we get a τ-cover for X in the sense of metric spaces (τ depends on ε and K), then we have an ε-cover for (X, K).

▶ **Theorem 1.** *Consider a point set X of size n in $\mathbb{R}^d$ and a (L, r)-standard kernel $K(x, y)$. One can construct an ε-cover of size $n\left(\frac{3Lr}{\varepsilon}\right)^d$ for the kernel range space (X, K).*

Proof. A metric space $\frac{\varepsilon}{L}$-cover of $\tilde{B} = \cup_{i=1}^n B_r(x_i)$ will provide an ε-cover of (X, K). The covering number, $N(V, \varepsilon)$, of $V \subset \mathbb{R}^d$ is bounded by $\frac{\mathrm{Vol}(V)}{\mathrm{Vol}(B)}(\frac{3}{\varepsilon})^d$, where B shows the unit ball in $\mathbb{R}^d$. The volume of a ball of radius r in $\mathbb{R}^d$ is $\frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d$, where Γ is the Gamma function. So each ball of radius r can be covered by $\left(\frac{3r}{\varepsilon/L}\right)^d$ points. Therefore, $n\left(\frac{3Lr}{\varepsilon}\right)^d$ points are sufficient for an ε-cover of (X, K). ◀

▶ **Corollary 2.** *For point set $X \subset \mathbb{R}^d$ of size n, applying Theorem 1 and Table 1, these kernels admit ε-covers for (X, K):*

Gaussian	*size:* $n(3/\varepsilon)^d \ln^{d/2}(1/\varepsilon)$	*Laplace*	*size:* $n(3/\varepsilon)^d \ln^d(1/\varepsilon)$
Triangle	*size:* $n(3/\varepsilon)^d$	*Epanechnikov*	*size:* $n(3/\varepsilon)^d$
Quartic	*size:* $n(3/\varepsilon)^d$	*Triweight*	*size:* $n(3/\varepsilon)^d$

3 ε-cover-samples and reducing the number of points

We next build to the new definition of an ε-cover-sample. This is a subset of points that preserves the above gridding-based construction for an ε-cover. We start with the more well-known definition of an ε-sample, and its existing generalization to kernels. We observe this does not quite capture the right definition of a subset, so we evolve it to the ε-cover-sample.

An ε-*sample* [16] for a range space $(X, \mathcal{A})$ is a set $S \subset X$ such that

$$\max_{A \in \mathcal{A}} \left| \frac{|X \cap A|}{|X|} - \frac{|S \cap A|}{|S|} \right| \le \varepsilon.$$

If $K \le 1$ is a kernel defined on X, then, following [21], an ε-*KDE-sample* (also called ε-*sample for* (X, K)) is a set $S \subset X$ so

$$\max_{q \in \mathbb{R}^d} \left| \frac{1}{|X|} \sum_{x \in X} K(q, x) - \frac{1}{|S|} \sum_{s \in S} K(q, s) \right| \le \varepsilon.$$

A kernel is said to be *linked* [21] to a range space $(X, \mathcal{A})$ if for any possible input point $q \in \mathbb{R}^d$ and any value $v \in \mathbb{R}^+$ the super-level set of $K(\cdot, q)$ defined by v is equal to some $H \in \mathcal{A}$, i.e. $\{x \in \mathbb{R}^d : K(x, q) \ge v\} = H$. If S is an ε-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is linked to K, then S is also an ε-KDE-sample [21, Theorem 5.1].

An ε-*cover-sample* for X is a set $S \subset X$ such that for any $p, q \in \mathbb{R}^d$,

$$\left| d^X_\triangle(p, q) - d^S_\triangle(p, q) \right| \le \varepsilon.$$

Appendix A.2 shows that an ε-cover-sample is an ε-KDE-sample, so an ε-cover-sample is a generalization of an ε-KDE-sample. However, we can only show an ε-KDE-sample implies the following (note the absolute values outside the sum):

$$\left| \left| \frac{1}{|X|} \sum_{x \in X} [K(p, x) - K(q, x)] \right| - \left| \frac{1}{|S|} \sum_{s \in S} [K(p, s) - K(q, s)] \right| \right| \le 2\varepsilon.$$

Moreover, the example below shows that an ε-cover-sample may not be an ε-cover.

▶ **Example 3.** Fix $z \in \mathbb{R}$ and let $X = \{x_1, \dots, x_n\} \subset \mathbb{R}$, where $x_1 = \cdots = x_n = z$ (or small perturbations of z), and consider the Gaussian kernel $K(x, y) = e^{-\|x-y\|^2}$. Then $S = \{z\}$ will be an ε-cover-sample for any $\varepsilon > 0$ since $K(p, x_i) = K(p, z)$ for all i and all $p \in \mathbb{R}$. So, for any $p, q \in \mathbb{R}$,

$$\left|d_\triangle^X(p,q) - d_\triangle^S(p,q)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}|K(p,x_i) - K(q,x_i)| - |K(p,z) - K(q,z)|\right| = 0 < \varepsilon.$$

Now let $p \in \mathbb{R}$ be arbitrary. If S is an ε-cover for (X, K), then we should have $d_\triangle^X(p,z) < \varepsilon$ (remember that S is a singleton and the only choice from S is z) and so considering the fact that $R_z = (1, \dots, 1)$ (since $x_1 = \cdots = x_n = z$), we will need to have

$$|K(p,z) - 1| = \frac{1}{n}\sum_{i=1}^{n}|K(p,x_i) - 1| = \frac{1}{n}\sum_{i=1}^{n}|K(p,x_i) - K(z,x_i)| < \varepsilon.$$

However, this is impossible for an arbitrary $p \in \mathbb{R}$; and any ε-cover Q needs at least one point in each annulus $S_i = \{p : (i-1)\varepsilon < K(p,z) \le i\varepsilon\}$ for $1 \le i \le 1/\varepsilon$.

3.1 Semi-linked range spaces

The *semi-super-level set* of a kernel K with respect to the points $p, q \in \mathbb{R}^d$ and $\tau \in \mathbb{R}^+$ is

$$R_{p,q,\tau} = \{x \in \mathbb{R}^d : |K(p,x) - K(q,x)| \ge \tau\}.$$

Moreover, K is said to be *semi-linked* to a range space $(\mathbb{R}^d, \mathcal{A})$ if $R_{p,q,\tau} \in \mathcal{A}$ for any possible $p, q \in \mathbb{R}^d$, $\tau \in \mathbb{R}^+$. We also say $\mathcal{A}$ is semi-linked to K. This is extending the idea of super-level sets and linking kernels to range spaces from [21]. There are also ε-KDE-samples of size $O(1/\varepsilon^2)$ for characteristic kernels, either using a uniform random sample [27] or via an iterative greedy algorithm [5, 25]; see discussion in [36]. The size can be improved to $O(1/\varepsilon)$ when d is constant [40] for Gaussian kernels or for a bounded domain [22], or $O(\sqrt{d}/\varepsilon \cdot \sqrt{\log(1/\varepsilon)})$ for positive definite kernels [36].

In the following theorem we extend the linking-based result to ε-cover-samples that are now semi-linked to an appropriate range space. The proof mostly follows the strategy of Joshi *et al.* [21], and is deferred to Appendix A.4. The main idea is for any query based on values $p, q \in \mathbb{R}^d$, all points in X can be sorted in descending order by their value in $|K(p,x) - K(q,x)|$. For each subset in this order, it is guaranteed to have the appropriate error levels by the associated semi-linked range space. Through some case analysis, as in [21], one can show that the error cannot accumulate too much across different levels.

▶ **Theorem 4.** *Let S be an ε-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to a kernel K, where $K \le 1$. Then S is an ε-cover-sample for X.*

Consider a range space $(X, \mathcal{A})$, and a kernel K such that its critical radius is finite for any ε and $\mathcal{A}$ is semi-linked to K. Then $\mathcal{A}$ is linked to K, so this is a generalization of the linked range space result.

▶ **Theorem 5.** *Consider a (L, r)-standard kernel K in $\mathbb{R}^d$. One can construct an ε-cover with $s\left(\frac{Lr}{\varepsilon}\right)^d$ points for the kernel range space (X, K), where s is the size of an $\varepsilon/2$-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to K.*

Proof. Let S be an $\varepsilon/2$-sample of size s for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to K. Then

$$\forall p, q \in \mathbb{R}^d, \ \left|d_\triangle^X(p,q) - d_\triangle^S(p,q)\right| \le \varepsilon/2. \tag{1}$$

Applying Theorem 1 for S we can get an $\varepsilon/2$-cover Q for (S, K) of size $s\left(\frac{6Lr}{\varepsilon}\right)^d$. Let $p \in \mathbb{R}^d$ be arbitrary. Choose $q \in Q$ such that $d_{\triangle}^{S}(p,q) \leq \varepsilon/2$. Then utilizing (1) we get

$$d_{\triangle}^{X}(p,q) \leq \left|d_{\triangle}^{X}(p,q) - d_{\triangle}^{S}(p,q)\right| + d_{\triangle}^{S}(p,q) \leq \varepsilon. \qquad ◀$$

3.2 Bounding the VC-dimension of semi-linked range spaces

Given a range space $(X, \mathcal{A})$, a subset $Y \subset X$ is said to be *shattered* by $\mathcal{A}$ if all subsets $Z \subset Y$ can be realized as $Z = Y \cap R$ for some $R \in \mathcal{A}$. Then the *VC-dimension* of a range space $(X, \mathcal{A})$ is the cardinality of the largest subset $Y \subset X$ that can be shattered by $\mathcal{A}$.

Let $\mathcal{A}_d = \{R_{p,q,\tau} : p, q \in \mathbb{R}^d, \tau > 0\}$, where K is a standard kernel and $R_{p,q,\tau}$ is its semi-super-level set. Now consider the class of functions $H = \{h_a : \mathbb{R}^d \to \{0,1\} | a \in \mathbb{R}^N\}$, where $h_a(x) = h(a,x)$ and $h : \mathbb{R}^N \times \mathbb{R}^d \to \{0,1\}$ is a function. This each $h_a(\cdot)$ defines a subset of $\mathbb{R}^d$ – the points x which evaluate to 1. Suppose h is "simply computable", that is, computing $h(a,x)$ for any $a \in \mathbb{R}^N$ and $x \in \mathbb{R}^d$ requires no more than t of the arithmetic operations, jumps conditions (described in Preliminaries) and comparisons, and requires u times evaluation of the exponential function $z \mapsto e^z$. Then Theorem 8.14 of [4] implies

$$\dim_{\mathrm{VC}}((\mathbb{R}^d, H)) \leq d^2(u+1)^2 + 11d(u+1)(t + \log(9d(u+1))).$$

In the appendix we have shown that $\dim_{\mathrm{VC}}((\mathbb{R}^d, \mathcal{A}_1)) = 4$, $\dim_{\mathrm{VC}}((\mathbb{R}^d, \mathcal{A}_2)) \geq 6$ and $\dim_{\mathrm{VC}}((\mathbb{R}^d, \mathcal{A}_d)) \geq d+1$, where K is any of the kernels listed in Table 1. In order to get an upper bound on the VC-dimension of $\mathcal{A}_d$ we employ the above simple operations theorem.

▶ **Theorem 6.** *Let $\mathcal{A}_d = \{R_{p,q,\tau} : p, q \in \mathbb{R}^d, \tau > 0\}$, where K is a k-simply computable standard kernel and $R_{p,q,\tau} = \{x \in \mathbb{R}^d : |K(p,x) - K(q,x)| \geq \tau\}$. Then $\dim_{\mathrm{VC}}(\mathcal{A}_d) = O(d^k)$.*

Proof. For $p, q \in \mathbb{R}^d$, $\tau \in \mathbb{R}^+$ let $\chi^+_{p,q,\tau}$ and $\chi^-_{p,q,\tau}$ be the characteristic functions of the sets

$$R^+_{p,q,\tau} = \{x \in \mathbb{R}^d : K(p,x) - K(q,x) \geq \tau\} \quad \text{and} \quad R^-_{p,q,\tau} = \{x \in \mathbb{R}^d : K(p,x) - K(q,x) \leq -\tau\},$$

respectively. Then $h_{p,q,\tau} = \chi^+_{p,q,\tau} + \chi^-_{p,q,\tau}$ is the characteristic function of $R_{p,q,\tau}$, i.e. $h_{p,q,\tau}(x) = 1$ if $x \in R_{p,q,\tau}$ and 0 otherwise. Therefore, we have the class of functions

$$H = \{h_{p,q,\tau} : \mathbb{R}^d \to \{0,1\} | \, p, q \in \mathbb{R}^d, \, \tau \in \mathbb{R}^+\},$$

(we set $N = 2d+1$ here, for the coordinates to describe p, q, τ).

Because determining $|K(p,x) - K(q,x)| \geq \tau$ needs $O(d^{k-1})$ simple operations, one can easily observe that $h_{p,q,\tau}(x)$ can be computed by $t = O(d^{k-1})$ steps using above-mentioned simple operations and $u = O(1)$ evaluations of the exponential function. Therefore, by [4, Theorem 8.14], the VC-dimension of H is upper bounded by $\dim_{\mathrm{VC}}(H) = O(d^k)$. Hence, $\dim_{\mathrm{VC}}(\mathcal{A}_d) = O(d^k)$. ◀

For a range space $(X, \mathcal{A})$ with VC-dimension ν, a random sample from X of size $O((1/\varepsilon^2)(\nu + \log 1/\delta))$ is an ε-sample with probability at least $1 - \delta$ [41, 26]. Using this fact and applying Theorem 5 we obtain the following corollaries.

▶ **Corollary 7.** *Let K be a k-simply computable standard kernel on $\mathbb{R}^d$. A random sample from X of size $O((1/\varepsilon^2)(d^k + \log 1/\delta))$ is an ε-cover-sample for X with probability $\geq 1 - \delta$.*

Now applying Theorem 5 gives the following result.

▶ **Corollary 8.** *Let K be a k-simply computable (L, r)-standard kernel on $\mathbb{R}^d$. Then $O\big((d^k + \log(1/\delta))(6Lr)^d/\varepsilon^{d+2}\big)$ points suffice to construct an ε-cover for (X, K) with probability $\geq 1 - \delta$.*

4 An upper bound on ε-cover size for high dimensions

This section builds a dimension-free size upper bound for a kernel ε-cover. We use the following theorem, termed *ε-terminal dimensionality reduction* or *ε-terminal JL*, from [33] (see [10] for an algorithmic version of terminal JL. Appendix B also gives the algorithm for computing terminal JL):

▶ **Theorem 9.** *([33, Theorem 1.2]) Let $\varepsilon \in (0, 1)$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be arbitrary with $n > 1$. Then there exists a function $f : \mathbb{R}^d \to \mathbb{R}^m$ with $m = O(\log(n)/\varepsilon^2)$ such that for all $x_i \in X$ and $\underline{\text{all } p \in \mathbb{R}^d}$, $\|p - x_i\| \le \|f(p) - f(x_i)\| \le (1 + \varepsilon)\|p - x_i\|$.*

▶ **Lemma 10.** *Let $\varepsilon > 0$, K be a (L, r')-standard kernel, $r = r'(\varepsilon/2)$, and $S \subset \mathbb{R}^d$ be a finite subset of $\mathbb{R}^d$. Let also $f : \mathbb{R}^d \to \mathbb{R}^m$ be the $\varepsilon/(2Lr)$-terminal dimensionality reduction transform for S, where $m = O(L^2 r^2 \log(|S|)/\varepsilon^2)$. Then for any $p \in \mathbb{R}^d$ the following holds:*

$$\sum_{s\in S} |K(f(p), f(s)) - K(p, s)| < \frac{\varepsilon}{2}|S|.$$

Proof. Since K is L-Lipschitz, using terminal dimensionality reduction property, for any $p \in \mathbb{R}^d$ and $s \in S$, we infer that

$$|K(f(p), f(s)) - K(p, s)| \le L\big|\|f(p) - f(s)\| - \|p - s\|\big| \le \frac{\varepsilon}{2r}\|p - s\|. \tag{2}$$

Therefore, applying (2), for any $p, q \in \mathbb{R}^d$ and $s \in S$ (note $|f(S)| = |S|$ as f is invertible on S), we get

$$\begin{aligned}
&\sum_{s\in S} |K(f(p),f(s)) - K(p, s)| \\
&\qquad = \sum_{\|p-s\|>r} |K(f(p), f(s)) - K(p, s)| + \sum_{\|p-s\|\le r} |K(f(p), f(s)) - K(p, s)| \\
&\qquad \le \sum_{\|p-s\|>r} \frac{\varepsilon}{2} + \sum_{\|p-s\|\le r} \frac{\varepsilon}{2r}\|p - s\| \le \sum_{s\in S} \frac{\varepsilon}{2} = \frac{\varepsilon}{2}|S|.
\end{aligned}$$

◀

▶ **Lemma 11.** *Let $\varepsilon > 0$, K be a (L, r')-standard kernel, $r = r'(\varepsilon/2)$, and $S \subset \mathbb{R}^d$ be a finite subset of $\mathbb{R}^d$. Let also $f : \mathbb{R}^d \to \mathbb{R}^m$ be the $\varepsilon/(2Lr)$-terminal dimensionality reduction transform for S, where $m = O(L^2 r^2 \log(|S|)/\varepsilon^2)$. Then for any $p, q \in \mathbb{R}^d$ the following holds:*

$$|d_\Delta^{f(S)}(f(p), f(q)) - d_\Delta^S(p, q)| < \varepsilon.$$

Proof. Applying Lemma 10, for any $p, q \in \mathbb{R}^d$ and any $s \in S$ (note $|f(S)| = |S|$ as f is invertible on X), we get

$$\begin{aligned}
&d_\Delta^{f(S)}(f(p), f(q)) = \frac{1}{|S|}\sum_{s\in S} |K(f(p), f(s)) - K(f(q), f(s))| \\
&\le \frac{1}{|S|}\sum_{s\in S} \big(|K(f(p), f(s)) - K(p, s)| + |K(p, s) - K(q, s)| + |K(f(q), f(s)) - K(q, s)|\big) \\
&\le \frac{1}{|S|}\left[\frac{\varepsilon}{2}|S| + \sum_{s\in S} |K(p, s) - K(q, s)| + \frac{\varepsilon}{2}|S|\right] = d_\Delta^S(p, q) + \varepsilon.
\end{aligned}$$

Similarly, we can show $d_\Delta^S(p, q) \le d_\Delta^{f(S)}(f(p), f(q)) + \varepsilon$. ◀

The following corollary will help us in proving our main result (Theorem 17), where we need to construct an ε-cover for a set S from an ε-cover of its image under terminal dimensionality reduction transform.

▶ **Corollary 12.** *Let $\varepsilon > 0$, K be a (L, r')-standard kernel, $r = r'(\varepsilon/16)$, $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and S be an $\frac{\varepsilon}{4}$-cover-sample for X. Let also $f : \mathbb{R}^d \to \mathbb{R}^m$ be the $\varepsilon/(16Lr)$-terminal dimensionality reduction transform on S, where $m = O(L^2r^2\log(|S|)/\varepsilon^2)$. If Q is an $\frac{\varepsilon}{8}$-cover for $(f(S), K)$, then we can compute an ε-cover for (X, K) of size at most $|Q|$.*

Proof. Compute a naive $\frac{\varepsilon}{8}$-cover $Q_S = \{p_1, \dots, p_M\}$ for S of size $M = |S|\left(\frac{24Lr'}{\varepsilon}\right)^d$ by Theorem 1. We say a point p_i is covered by a point $q \in Q$ if $f(p_i) \in B_{\varepsilon/8}(q)$, the $\varepsilon/8$-radius ball in the metric space $(\mathbb{W}^{|f(S)|}, d_\Delta^{f(S)})$. Process points $p_i \in Q_S$ one-by-one, putting some in a new set Q'. To process a p_i, put it in Q', find a point $q \in Q$ that covers p_i, and remove q from Q. Remove all $p_j \in Q_S$ from Q_S that are also covered by q. The process concludes when the whole Q_S is processed. We claim that Q' is an ε-cover for (X, K); it is of size at most $|Q|$.

Let $p \in \mathbb{R}^d$. Since Q is an $\frac{\varepsilon}{8}$-cover for $(f(S), K)$, there is $q \in Q$ such that $d_\Delta^{f(S)}(f(p), q) < \frac{\varepsilon}{8}$. Let $p_i \in Q_S$ be such that $d_\Delta^S(p, p_i) < \frac{\varepsilon}{8}$. If $p_i \in Q'$, we are done. Otherwise, there must be a $p_j \in Q_S$ with $j < i$ such that $f(p_i), f(p_j) \in B_{\varepsilon/8}(q')$ for some $q' \in Q$, and p_j is included in Q' by the construction of Q'. Thus $d_\Delta^{f(S)}(f(p_i), f(p_j)) < \frac{\varepsilon}{4}$. The proof will be complete if we show $d_\Delta^X(p, p_j) < \varepsilon$. Employing Lemma 11 with $\varepsilon/(16Lr)$ and S we obtain $|d_\Delta^{f(S)}(f(p), f(p_i)) - d_\Delta^S(p, p_i)| \le \frac{\varepsilon}{8}$, and so $d_\Delta^{f(S)}(f(p), f(p_i)) < \frac{\varepsilon}{4}$. Therefore,

$$d_\Delta^{f(S)}(f(p), f(p_j)) \le d_\Delta^{f(S)}(f(p), f(p_i)) + d_\Delta^{f(S)}(f(p_i), f(p_j)) < \varepsilon/2.$$

Applying Lemma 11 in a similar fashion we get $|d_\Delta^{f(S)}(f(p), f(p_j)) - d_\Delta^S(p, p_j)| \le \frac{\varepsilon}{8}$. Combining last two inequalities we conclude that $d_\Delta^S(p, p_j) < \frac{5\varepsilon}{8}$. On the other hand, because S is an $\frac{\varepsilon}{4}$-cover-sample for X, we have $|d_\Delta^X(p, p_j) - d_\Delta^S(p, p_j)| \le \frac{\varepsilon}{4}$, which means that $d_\Delta^X(p, p_j) < \frac{7\varepsilon}{8} < \varepsilon$. ◀

4.1 Input-size and dimension-free bound for ε-cover size

The bound we proved in Corollary 8 on ε-cover size is input size ($n = |X|$)-free but depends on the dimension d. In Theorem 13 we give an input-size- and dimension-free upper bound on the size of ε-cover-sample for positive definite kernels using Rademacher complexity. Then in Theorem 15 we obtain a similar bound for k-simply computable standard kernels. These two theorems will result in input-size- and dimension-free upper bound for ε-cover size.

▶ **Theorem 13.** *Let $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, K be a positive definite bounded kernel on $\mathbb{R}^d$, and let $\delta \in (0, 1)$ be the probability of failure. Then a random sample of size $m > \frac{1}{49\varepsilon^2}\log(\frac{1}{\delta})$ is an ε-cover-sample for X with probability $\ge 1 - \delta$.*

Proof. For any $p, q \in \mathbb{R}^d$, clearly $\mathbb{E}_{x\sim X}[f_{p,q}(X)] = \frac{1}{n}\sum_{i=1}^n |K(p, x_i) - K(q, x_i)|$, where $\mathbb{E}$ denotes the expectation. Consider an i.i.d random sample $S = \{s_1, \dots, s_m\}$ of X of size m. Define the family of functions

$$G = \{f_{p,q} : X \to [0, 1] \mid p, q \in \mathbb{R}^d,\ f_{p,q}(x) = |K(p, x) - K(q, x)| \text{ for } x \in X\}.$$

Applying the two-sided Rademacher complexity bound theorem (see Theorem 3.3 of [31] for a one-sided bound) on X, S and G with probability of at least $1 - \delta$ we get

$$\left| \frac{1}{n}\sum_{i=1}^n |K(p, x_i) - K(q, x_i)| - \frac{1}{m}\sum_{j=1}^m |K(p, s_j) - K(q, s_j)| \right| \le 2\hat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log(4/\delta)}{2m}}, \quad (3)$$

where $\hat{\mathfrak{R}}_S(G)$ denotes the empirical Rademacher complexity of G with respect to the sample S. On the other hand, by Theorem 33 we know that $\hat{\mathfrak{R}}_S(B_1^{\mathcal{H}}(0)) \le \frac{1}{\sqrt{m}}$, where $B_1^{\mathcal{H}}(0)$ denotes the unit ball around the origin in RKHS. Notice $K(p,\cdot)$ belongs to $B_1^{\mathcal{H}}(0)$ for any p in $\mathbb{R}^d$. Now, by inspection of the definition of the Rademacher complexity one can see that the Rademacher complexity of $B_1^{\mathcal{H}}(0) - B_1^{\mathcal{H}}(0)$ will be at most $\frac{2}{\sqrt{m}}$. Employing Talagrand's contraction principle (see Lemma 5 in [30]) for $B_1^{\mathcal{H}}(0) - B_1^{\mathcal{H}}(0)$ and absolute value function (which is 1-Lipschitz), we observe that the Rademacher complexity of G is at most $\frac{2}{\sqrt{m}}$. Therefore, applying our notation in the paper, by (3) we obtain

$$|d_\triangle^X(p,q) - d_\triangle^S(p,q)| \le 4/\sqrt{m} + 3\sqrt{\log(4/\delta)/(2m)} \le 7\sqrt{\log(1/\delta)/m}.$$

Setting $7\sqrt{\log(1/\delta)/m} < \varepsilon$ gives $m > \frac{1}{49\varepsilon^2}\log(\frac{1}{\delta})$. ◀

The similar bound for ε-cover-sample size of k-simply computable standard kernels relies on the following observation, which is an easy application of triangle inequality.

▶ **Lemma 14.** *Let S be an $\varepsilon/2$-cover-sample of X and S' an $\varepsilon/2$-cover-sample of S. Then S' is an ε-cover-sample of X.*

The intuition behind the proof of the following theorem is recursively applying Lemma 14, by creating ε-cover-samples of ε-cover-samples, each of smaller size. At the start of each step i we have a size n_i and dimension d_i. We can apply terminal JL to reduce the dimension to $d_i' = O((1/\varepsilon^2)\log n_i)$, and then Corollary 7 to create an ε-cover-sample of size (roughly) $n_{i+1} = O((d_i'/\varepsilon)^2) = O((1/\varepsilon)^6 \log^2 n_i)$. Combining these steps does not immediately remove the dependence on n (or the initial d), but it does push the dependence on n into the log term. Applying this recursively the dependence on n can eventually be eliminated, but at the cost of a $\log^*(n)$ error factor (since we accumulate ε-error at each recursive step), which ultimately needs to be folded back into the size bound, adjusting $\varepsilon' = \varepsilon/\log^*(n)$. Instead we apply an inductive argument (inspired by the proof of Theorem 12.3 of [32]), so we only need to argue about one step. We show that applying the reductions with sufficiently small error parameter ε it can be independent of n and d. It again uses Lemma 14 but only once. However, this argument is complicated by the two-stage approach because the dependence on n and d are linked, and reducing one relies on the other. Like the recursive method sketched above, by combining them we can reduce the dependence on both terms.

▶ **Theorem 15.** *Let $\varepsilon, \delta \in (0,1)$, consider a finite point set $X \subset \mathbb{R}^d$ and let K be a k-simply computable (L,r)-standard kernel. Then with probability at least $1-\delta$, a random sample of size $O\big(\frac{1}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k(\frac{Lr}{\varepsilon\delta})\big)$ from X is an ε-cover-sample for X.*

Proof. Let $T(\varepsilon,\delta,X)$ denote the least positive integer such that a uniform sample of X of size $T(\varepsilon,\delta,X)$ is an ε-cover-sample of X with probability at least $1-\delta$. We prove the theorem by induction on ε. If $n \le \frac{1}{\varepsilon^{2+2k}}$, then $T(\varepsilon,\delta,X) \le n \le \frac{1}{\varepsilon^{2+2k}}$, and so X would be an ε-cover-sample of X satisfying the claim of the theorem. Thus assume that $n > \frac{1}{\varepsilon^{2+2k}}$. Let S be an $\varepsilon/2$-cover-sample of X with probability at least $1-\delta/2$. Then employing Lemma 14, any $\varepsilon/2$-cover-sample S' of S, with probability at least $1-\delta/2$, would be an ε-cover-sample of X with probability $\ge 1-\delta$. This means that for any $\varepsilon/2$-cover-sample S' of S we have

$$T(\varepsilon,\delta,X) \le T(\varepsilon/2,\delta/2,S) \le |S'|. \tag{4}$$

Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be the ε'-terminal dimensionality reduction transform for the set S and $\varepsilon' = \varepsilon/(6Lr(\varepsilon/6))$, where $m = O(L^2 r^2 \log(|S|)/\varepsilon^2)$. Now consider the range space

$(f(S), \mathcal{A}_m)$. By Theorem 6, $\dim_{\mathrm{VC}}(\mathcal{A}_m) = O(m^k)$. Hence, a random sample S'' of size $n'' = (C_1/\varepsilon^2)(m^k + \log 1/\delta)$ from $f(S)$ is an $\varepsilon/6$-sample for $f(S)$ with probability at least $1 - \delta/2$ [41, 26], where C_1 is a sufficiently large constant. Now Theorem 4 shows that S'' is an $\varepsilon/6$-cover-sample for $f(S)$ with probability at least $1 - \delta/2$. However, since f is invertible on S, we have $S'' = f(S')$, where $S' = \{s' \in S : f(s') = s'' \text{ for some } s'' \in S''\}$. Let us show that S' is an $\varepsilon/2$-cover-sample for S. Let $p, q \in \mathbb{R}^d$ be arbitrary. Then with probability $\geq 1 - \delta/2$,

$$\begin{aligned}
|d^S_\triangle(p,q) - d^{S'}_\triangle(p,q)| &\leq |d^S_\triangle(p,q) - d^{f(S)}_\triangle(f(p), f(q))| \\
&\quad + |d^{f(S')}_\triangle(f(p), f(q)) - d^{S'}_\triangle(p,q)| + |d^{f(S)}_\triangle(f(p), f(q)) - d^{f(S')}_\triangle(f(p), f(q))| \\
&\leq \varepsilon/6 + \varepsilon/6 + \varepsilon/6 = \varepsilon/2,
\end{aligned}$$

where we applied Lemma 11 two times and utilized the fact that $f(S') = S''$ is an $\varepsilon/6$-cover-sample for $f(S)$. Obviously, $|S'| = |S''| = n''$. By plugging in $m = C_2 L^2 r^2 \log(|S|)/\varepsilon^2$ for some constant $C_2 > 0$, we obtain

$$|S'| = \frac{C_1}{\varepsilon^2}\Big[\Big(\frac{C_2}{\varepsilon^2} L^2 r^2 \log(|S|)\Big)^k + \log\frac{1}{\delta}\Big] \leq \frac{C_1(C_2^k + 1)}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k\big(\frac{|S|}{\delta}\big).$$

Since the above inequality holds for any $\varepsilon/2$-cover-sample S of X, we can infer that

$$|S'| \leq \frac{C_1(C_2^k + 1)}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k\big(\frac{T(\varepsilon/2, \delta/2, X)}{\delta}\big).$$

Therefore, applying inductive hypothesis and the fact that k is constant and for any constant $a > 1$ and $x \geq a^{1/(a-1)}$, $\log^k(ax) \leq a^k \log^k(x)$, which can be easily observed, we get

$$\begin{aligned}
\log^k\Big(\frac{\frac{2^{2+2k}C}{\varepsilon^{2+2k}} L^{2k} r(\varepsilon/2)^{2k} \log^k(\frac{4Lr(\varepsilon/2)}{\varepsilon\delta})}{\delta}\Big) &\leq \log^k\Big(\Big(\frac{2C^{1/(2+2k)} Lr(\varepsilon/2)\log(\frac{4Lr(\varepsilon/2)}{\varepsilon\delta})}{\varepsilon\delta}\Big)^{2+2k}\Big) \\
&\leq (2+2k)^k C_3^k C^{k/(2+2k)} \log^k\Big(\frac{Lr\log(\frac{4Lr(\varepsilon/2)}{\varepsilon\delta})}{\varepsilon\delta}\Big) \\
&\leq (2+2k)^k C_3^k \sqrt{C}\Big(\log\big(\frac{Lr}{\varepsilon\delta}\big) + \log\log\big(\frac{C_3 Lr}{\varepsilon\delta}\big)\Big)^k \\
&\leq (2+2k)^k C_3^k \sqrt{C}\Big(1.4\log\big(\frac{C_3 Lr}{\varepsilon\delta}\big)\Big)^k \\
&\leq 1.4^k (2+2k)^k C_3^{2k} \sqrt{C} \log^k\big(\frac{Lr}{\varepsilon\delta}\big),
\end{aligned}$$

and thus

$$|S'| \leq \frac{1.4^k (2+2k)^k C_1 (C_2^k + 1) C_3^{2k} \sqrt{C}}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k\big(\frac{Lr}{\varepsilon\delta}\big),$$

where C_3 is a constant making sure that $4r(\varepsilon/2) \leq C_3 r(\varepsilon)$; recall that here $r = r(\varepsilon)$ is the ε-critical radius. Hence, if we put C such that $\sqrt{C} \geq 1.4^k (2+2k)^k C_1 (C_2^k + 1) C_3^{2k}$, then by (4) we conclude $T(\varepsilon, \delta, X) \leq \frac{C}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k(\frac{Lr}{\varepsilon\delta})$. ◀

With the same proof technique as in Theorem 15 one can prove the following corollary, which gives an input-size- and dimension-free upper bound on the ε-KDE-samples. For characteristic kernels it is known that smaller such upper bounds of size $O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ exist [27, 35, 9, 5, 25]. Our bound, however, applies for non-characteristic kernels as well. The only modification we need is applying super-level sets rather than semi-super-level sets. Notice, in this setting, more kernels can be considered as k-simply computable such as Laplacian kernel, where one can argue that it is 3-simply computable.

▶ **Corollary 16.** *Let $\varepsilon, \delta \in (0,1)$, consider a finite point set $X \subset \mathbb{R}^d$ and let K be a k-simply computable (L,r)-standard kernel in the sense of super-level sets. Then with probability $\geq 1-\delta$, a random sample of size $O\big(\frac{1}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k(\frac{Lr}{\varepsilon\delta})\big)$ from X is an ε-KDE-sample for X.*

Finally, we reach our goal for this section which was providing an input-size- and dimension-free upper bound on ε-cover size. Recall that for most kernels the Lipschitz factor L is a constant, and the critical radius r is a constant or $\text{polylog}(1/\varepsilon)$.

▶ **Theorem 17.** *Let $\varepsilon > 0$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, and K be a k-simply computable (L,r)-standard kernel. Then, with constant probability, we can compute an ε-cover for (X,K) of size $\left(\frac{Lr}{\varepsilon}\right)^{O\left(\frac{L^2r^2}{\varepsilon^2}\log(\frac{Lr}{\varepsilon})\right)}$.*

Proof. Let S be an $\varepsilon/4$-cover-sample of X of size $|S| = O(\frac{1}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k(\frac{Lr}{\varepsilon}))$; which is guaranteed to exist by Theorem 15, and fix the probability of failure $\delta = 0.1$, or any constant in $(0,1)$. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be the ε'-terminal dimensionality reduction transform for the set S and $\varepsilon' = \varepsilon/(4Lr)$, where $m = O(L^2r^2\log(|S|)/\varepsilon^2) = O(L^2r^2\log(Lr/\varepsilon)/\varepsilon^2)$.

Then by Corollary 12, an $\varepsilon/8$-cover Q of $(f(S),K)$ will provide us with an ε-cover of (X,K) of size at most $|Q|$. Finally, by Theorem 1, $f(S)$ admits an $\varepsilon/8$-cover of size $O((\frac{24}{\varepsilon})^{m+2+2k} L^{m+2k} r^{m+2k} \log^k(\frac{Lr}{\varepsilon})) = (\frac{Lr}{\varepsilon})^{O(L^2r^2\log(\frac{Lr}{\varepsilon})/\varepsilon^2)}$ for (X,K). ◀

By the discussion in Section 1.1 we conclude the following corollary, which interestingly is a k-free improvement upon Corollary 16.

▶ **Corollary 18.** *Let $\varepsilon, \delta \in (0,1)$, consider a finite point set $X \subset \mathbb{R}^d$ and let K be a k-simply computable (L,r)-standard kernel in the sense of super-level sets. Then with probability $\geq 1-\delta$, a random sample of size $\tilde{O}\big(\frac{1}{\varepsilon^4}\big)$ from X is an ε-KDE-sample for X.*

We include a similar upper bound for positive definite kernels too.

▶ **Theorem 19.** *Let $\varepsilon > 0$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, and K be a positive definite L-Lipschitz kernel with critical radius r. Then, with constant probability, we can compute an ε-cover for (X,K) of size $\left(\frac{Lr}{\varepsilon}\right)^{O(L^2r^2\log(\frac{1}{\varepsilon})/\varepsilon^2)}$.*

Proof. Let S be an $\varepsilon/4$-cover-sample of X of size $|S| = O(\frac{1}{\varepsilon^2})$; which is guaranteed to exist by Theorem 13 assuming constant probability of failure δ. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be the ε'-terminal dimensionality reduction transform for the set S and $\varepsilon' = \varepsilon/(4Lr)$, where $m = O(L^2r^2\log(|S|)/\varepsilon^2) = O(L^2r^2\log(1/\varepsilon)/\varepsilon^2)$. Then by Corollary 12, an $\varepsilon/8$-cover Q of $(f(S),K)$ will provide us with an ε-cover of (X,K) of size at most $|Q|$. Finally, by Theorem 1, $f(S)$ admits an $\varepsilon/8$-cover of size $O((\frac{24}{\varepsilon})^{m+2} L^m r^m) = (Lr/\varepsilon)^{O(L^2r^2\log(\frac{1}{\varepsilon})/\varepsilon^2)}$ for (X,K). ◀

Considering the fact that the Gaussian kernel is $(1, \sqrt{\ln(1/\varepsilon)})$-standard and positive definite, triangle kernel is $(1,1)$-standard, Epanechnikov, triangle, quartic and triweight kernels are $(2,1)$-standard, and the Laplace kernel is 1-Lipschitz and positive definite with critical radius $\ln(1/\varepsilon)$, the following corollary is an immediate consequence of Theorems 17 and 19. Notice we assumed $\sigma = 1$.

▶ **Corollary 20.** *Let $\varepsilon > 0$ and $X \subset \mathbb{R}^d$ be of size n. There exist a set of size $(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log^2(\frac{1}{\varepsilon}))}$ for Gaussian/truncated Gaussian kernel, a set of size $(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log^3(\frac{1}{\varepsilon}))}$ for Laplace kernel, and a set of size $(\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))}$ for Epanechnikov, triangular, quartic and triweight kernels that are ε-covers of (X,K).*

Algorithm and run time. The result in Theorem 17 is constructive with a runtime described in the next theorem. Recall, despite the involved analysis, the algorithm to find the ε-cover Q' of X is quite simple: (1) create a random sample $S \sim X$; (2) create a terminal JL map $f : \mathbb{R}^d \to \mathbb{R}^m$ for S; (3) map $S' \leftarrow f(S)$; (4) apply Theorem 1 on S' in $\mathbb{R}^m$ to get $Q \subset \mathbb{R}^m$. (5) find $Q' \subset \mathbb{R}^d$ from Q. Analyzing the runtime of most steps is straightforward. The most intricate part is step (5) since f is only invertible on S. This is handled via the procedure described in the proof of Corollary 12, whereby we create a naive ε-cover $Q_S \subset \mathbb{R}^d$ (via Theorem 1), and make sure to only place one point p_i from Q_S into the final ε-cover $Q' \subset \mathbb{R}^d$ for each $q \in Q \subset \mathbb{R}^m$.

▶ **Theorem 21.** *For a size n point set $X \subset \mathbb{R}^d$ and k-simply computable, (L, r)-standard kernel K we can compute an ε-cover for (X, K) of size $N = (Lr/\varepsilon)^{O(L^2r^2 \log(Lr/\varepsilon)/\varepsilon^2)}$ in time $(Lr/\varepsilon)^{d+O(\frac{1}{\varepsilon^2} \log \frac{Lr}{\varepsilon})}$, where we assume k is a constant.*

Proof. Assuming we can draw a random sample in $O(1)$ time, step (1) takes $O(|S|) = O(d^k/\varepsilon^2)$ time. Creating a terminal JL map $f : \mathbb{R}^d \to \mathbb{R}^m$ for S in steps (2) and (3) needs $O(d|S| \log(|S|)/\varepsilon^2)$ time (see Appendix B). In step (4) applying Theorem 1 on $f(S)$ requires $O(|S|(Lr/\varepsilon)^m)$ time, where $m = O(\log(|S|)/\varepsilon^2)$. Similarly, in step (5) creating an ε-cover $Q_S \subset \mathbb{R}^d$ for S (in the proof of Corollary 12) requires $O(|S|(3Lr/\varepsilon)^d)$ time. Then we need to calculate $f(p_i)$ for $p_i \in Q_S$ in Corollary 12, which needs $O(|Q_S|\sqrt{d}(|S|^3 + d^3) \log(1/\varepsilon))$ time (see Appendix B). In addition, computing each distance $\|f(p_i) - q\|$ for any $p_i \in Q_S$ and $q \in Q$ needs $O(m)$ time. Thus, noting that the process stops when Q is empty, these distance calculations need $O(m|Q||Q_S|)$ time. Therefore, putting all together, the run time of obtaining an ε-cover of size N needs $O(d|S| \log(|S|)/\varepsilon^2 + |Q_S|\sqrt{d}(|S|^3 + d^3) \log(1/\varepsilon) + |S|^2 + m|Q||Q_S|)$ time. Substituting $|S| = O\big(\frac{1}{\varepsilon^{2+2k}} L^{2k} r^{2k} \log^k(\frac{Lr}{\varepsilon})\big)$ (by Theorem 15 assuming a constant probability), $|Q| = |S|(3Lr/\varepsilon)^m = |S|(Lr/\varepsilon)^m$, $|Q_S| = |S|(3Lr/\varepsilon)^d$ and by simplifying we end up with $(Lr/\varepsilon)^{d+O(\log(Lr/\varepsilon)/\varepsilon^2)}$ time. ◀

We leave as an open question if we can remove the $1/\varepsilon^d$, so the runtime is $(\frac{1}{\varepsilon})^{\text{poly}(\frac{1}{\varepsilon})}$. This would seem to require a way to invert terminal JL for any point in $\mathbb{R}^m$.

5 A lower bound on ε-cover size for Gaussian and Laplace kernel

We now provide a lower bound, nearly matching the upper bound on ε-cover size shown in Corollary 20 for the Gaussian kernel; it also applies to the Laplace kernel. First we define criteria on a set of d spheres in $\mathbb{R}^d$ that can generate exactly two points in their intersections. Then we use this to design a point set that provides the desired lower bound.

The following lemma, as we will see in Lemma 23, will help us to make sure that every d-sphere in a collection of $1/\varepsilon$ same-centered spheres will intersect $1/\varepsilon^{d-1}$ spheres from another collection of spheres, each in 2 points. This will inductively provide us with $2/\varepsilon^d$ grid-like cells, where each cell is obtained by intersecting d annuluses formed by pairs of same-centered but different radius spheres with other annuluses. For instance, in $\mathbb{R}^2$ each grid-like cell is created as an intersection of 2 strips formed by pairs of same-centered circles. This is illustrated in Figure 1. So, the lemma makes it easy to count the number of grids generated in this way as we need to consider them for the lower bound on ε-cover. The proof is deferred to the appendix (see Lemma 31).

▶ **Lemma 22.** *Let $S_1, \ldots, S_d$ be d-spheres in $\mathbb{R}^d$, where S_k is centered at e_k with radius $R_k \geq 1$ (e_k is the standard k-th basis vector in $\mathbb{R}^d$). Assume that there are $R \geq 1$ and $\delta \in (0, 1/d)$ such that $|R_k^2 - R^2| < \delta$ for $k = 1, \ldots, d$. Then $|\bigcap_{k=1}^d S_k| = 2$.*

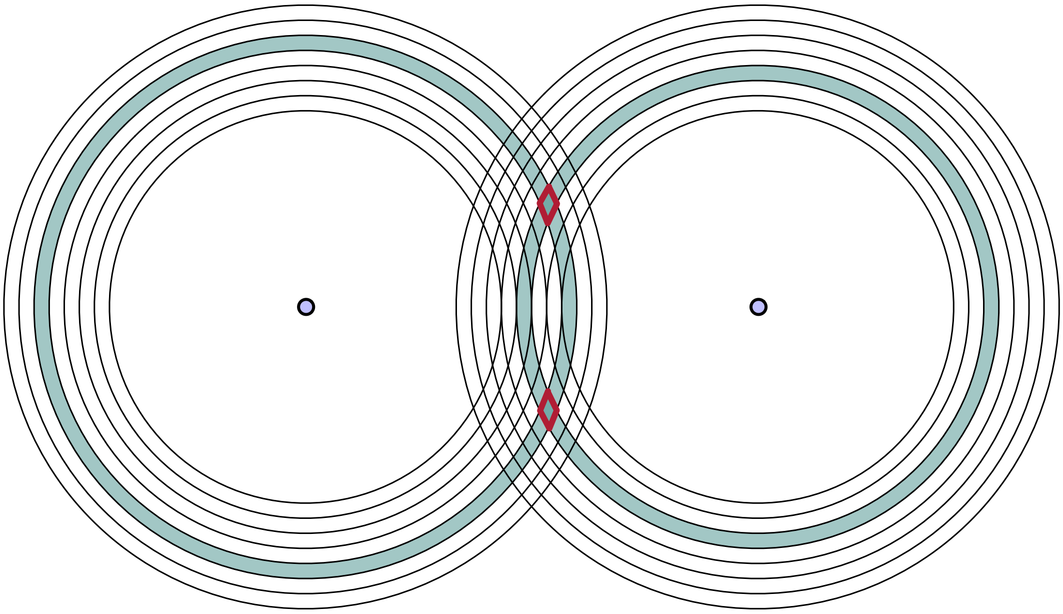

Figure 1 Illustration of intersection of annuluses around two blue points. The two green annuluses intersect forming 2 grid-like cells in red.

Crucially, the restriction that $\delta < 1/d$ in Lemma 22 will lead the next lemma to only obtain a $(1/\varepsilon)^{\Omega((1-\lambda)d)}$ size bound when $d = O(1/\varepsilon^{\lambda})$ for any $\lambda \in (0,1)$.

▶ **Lemma 23.** *Let $\varepsilon \in (0,1/3)$ and $d < \frac{1}{e^{3-\lambda}}\frac{1}{\varepsilon^{\lambda}} - \frac{1}{e}$ for some $\lambda \in (0,1)$. Then there are $\Omega(1/\varepsilon^{d})$ d-way intersections among d-spheres S_k in $\mathbb{R}^d$, where S_k is centered at e_k with radius $R_k = \ln(\frac{1}{i\varepsilon})$ for some $k = 1, \ldots, d$ and any integer i in $[\frac{1}{(e+1/d)\varepsilon}, \frac{1}{e\varepsilon}]$. Hence, there are at least $(\frac{2}{\varepsilon})^{(1-\lambda)d}$ grid-like cells obtained through the intersection of annuluses obtained by these spheres. Moreover, if d is constant, then the lower bound can be improved to be $\Omega(1/\varepsilon^{d})$.*

Proof. We will show that d-spheres in this setting satisfy the assumptions of Lemma 22. For any integer $i \in \left[\frac{1}{(e+1/d)\varepsilon}, \frac{1}{e\varepsilon}\right]$, using $R_i^2 = \ln(\frac{1}{i\varepsilon})$ and $R = 1$, we can infer

$$|R^2 - R_i^2| = \left|1 - \ln\frac{1}{i\varepsilon}\right| \leq \left|1 - \ln\left(e + \frac{1}{d}\right)\right| < \frac{1}{d},$$

where the last step follows by $x < e^x$. Therefore, by Lemma 22, each sphere with radii R_i centered at e_k will intersect any sphere with radii R_j centered at e_ℓ, where i, j are integers in $[\frac{1}{(e+1/d)\varepsilon}, \frac{1}{e\varepsilon}]$ and $k \neq \ell$. Note

$$\frac{1}{e\varepsilon} - \frac{1}{(e+\frac{1}{d})\varepsilon} = \frac{1}{e(ed+1)\varepsilon} > \frac{e^{1-\lambda}\varepsilon^{\lambda}}{\varepsilon} = \left(\frac{e}{\varepsilon}\right)^{1-\lambda}.$$

It means that there would be at least $(\frac{e}{\varepsilon} - 1)^{(1-\lambda)d}$ d-way intersections. Consequently, there would be at least $(\frac{e}{\varepsilon} - 2)^{(1-\lambda)d} > (\frac{2}{\varepsilon})^{(1-\lambda)d}$ grid-like cells obtained through intersection of annuluses formed by consecutive spheres on each axis with annuluses centered on other axes, as illustrated in Figure 1.

If d is fixed, then the length of the interval $[\frac{1}{(e+\frac{1}{d})\varepsilon}, \frac{1}{e\varepsilon}]$ would be a constant fraction of $1/\varepsilon$, leading to the lower bound of $\Omega(1/\varepsilon^{d})$. ◀

We remark that in the proof of Lemma 23 if we use $R_i = \ln(\frac{1}{i\varepsilon})$, the lemma holds true. The following theorem gives a lower bound on the kernel ε-cover size.

▶ **Theorem 24.** *Let $\varepsilon \in (0, 1/3)$ and $d < \frac{1}{e^{3-\lambda}}\frac{1}{\varepsilon^{\lambda}} - \frac{1}{e}$ for some constant $\lambda \in (0, 1)$ and let $X = \{e_1, \ldots, e_d\} \subset \mathbb{R}^d$ be the vertices of the standard (d-1)-simplex (i.e. e_k is the k-th basis vector). Let also $K(x, y) = e^{-\|x-y\|^2}$ (or $K(x, y) = e^{-\|x-y\|}$). Then the size of any ε-cover for (X, K) is at least $(1/\varepsilon)^{\Omega(1/\varepsilon^{\lambda})}$. If d is a constant, then the size of any ε-cover for (X, K) is at least $\Omega(1/\varepsilon^d)$.*

Proof. We need to start by constructing objects on the grid-like structure we call super-cells. Let the d-way intersections of the spheres be the nodes in a graph. Two nodes are connected by an edge if they share $d - 1$ spheres and for the last dimension of the nodes, the associated spheres are consecutive in the radius ordering. Then given a node p (which corresponds with a point in $\mathbb{R}^d$), we can define an L_1-distance to other nodes in the graph as the minimum number of edges one needs to traverse to get from p to another node q. A super-cell S_p contains all nodes within an L_1-distance of εd. However, the super cell S_p contains not just nodes, but also the part of $\mathbb{R}^d$ that one can reach from any node in a super-cell without crossing a sphere boundary. So one can think of a super-cell as a collection of cells from the grid-like structure that are reachable by moving into one of the 2^d cells incident to p, and then including face-incident cells to those within εd additional steps.

Next we want to argue that if there are no points q in a set Q that intersect a super-cell S_p, then for all $q \in Q$ that $d^X_{\triangle}(p, q) > \varepsilon$. Thus, if Q is an ε-cover it must hit (there must exist some $q \in Q$ so $q \in S_p$) the super-cell S_p. To see this, consider any point $q \in S_p$. One can reach q in a sequence of moves from $p = p_0$ to p_1 and recursively from p_j to p_{j+1}. The first j_* steps for $j_* \leq \varepsilon d$ moves must be along the edges of the grid-graph. These steps have the property that for some dimension k we change $|K(e_k, p_j) - K(e_k, p_{j+1})| = \varepsilon$, and for every other dimension k' we have $|K(e_{k'}, p_j) - K(e_{k'}, p_{j+1})| = 0$. The last step from p_{j_*} to q must have $|K(e_k, p_{j_*}) - K(e_k, q)| \leq \varepsilon$ for all dimensions k. Now $d^X_{\triangle}(p, q) = \frac{1}{d}\|R^X_p - R^X_q\|_1$ is $\frac{1}{d}$ times the sum of all changes in $|(R^X_p)_k - (R^X_q)_k| = |K(e_k, p) - K(e_k, q)|$. The first $j_* \leq \varepsilon d$ steps of the path captures any one-dimensional change in increments of ε, and the last step all residual changes in any coordinate less than ε. If the sum of these changes is less than εd, then it must be captured by some path. Therefore, a point q belongs to S_p if and only if it can be captured by some path starting at p with the sum of above-mentioned changes less than εd. The latter is equivalent to $d^X_{\triangle}(p, q) \leq \varepsilon$.

Now we need to provide an upper-bound of the size of a super-cell in terms of cells in the grid-like structure. Then we can lower bound the size of the ε-cover by the number of cells of the grid-like structure divided by the size of a super-cell. To do so, we divide a super-cell into 2^d orthants from p, so each path in some orthant only allows each dimension k to either increment or decrement. In addition, for each of the 2^d orthant choices, this determines which cell q moves into in the last step where it deviates from the grid-graph: it moves into the same-oriented incident orthant from p_{j_*}. By vector addition commutativity, we can take this step first to move from p to one of the incident 2^d cells, and then the remaining steps move to face-incident cells in the grid-like structure. The number of steps j_* can be between 0 and εd. Each step has d choices. So after fixing one of the 2^d orthants, the total number of distinct paths is $\sum_{j=0}^{\varepsilon d} d^j \leq d^{\varepsilon d+1}$ for $d > 1$. Note this is an overcount since two paths may end up in the same location by changing the order of steps. Regardless, we will use $2^d \cdot d^{\varepsilon d+1}$ as an upper bound on the size of a super-cell.

Finally, by a volume argument we can lower bound the size of an ε-cover as the number of cells in a grid-like structure, which is at least $(2/\varepsilon)^{(1-\lambda)d}$ by Lemma 23, divided by the number of cells in a super-cell which is at most $2^d \cdot d^{\varepsilon d+1}$. Further, by Lemma 23 we can use dimension as large as $d = \frac{1}{e^{3-\lambda}} \frac{1}{\varepsilon^\lambda} - \frac{1}{e} - 1$. We have

$$\begin{aligned}
\frac{(2/\varepsilon)^{(1-\lambda)d}}{2^d \cdot d^{\varepsilon d+1}} &= \frac{(2/\varepsilon)^{(1-\lambda)(\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1)}}{2^{\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1} \cdot (\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1)^{\varepsilon(\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1)+1}} \\
&\geq \frac{(2/\varepsilon)^{(1-\lambda)(\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1)}}{2^{1/\varepsilon^\lambda} \cdot (1/\varepsilon^\lambda)^{\varepsilon(1/\varepsilon^\lambda)+1}} \\
&= 2^{(1-\lambda)(\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1) - 1/\varepsilon^\lambda} \cdot (1/\varepsilon)^{(1-\lambda)(\frac{1}{e^{3-\lambda}}1/\varepsilon^\lambda - \frac{1}{e} - 1) - \lambda\varepsilon^{1-\lambda} + \lambda} \\
&= 2^{\varepsilon^{-\lambda}((1-\lambda)e^{\lambda-3} - 1) - (1-\lambda)\frac{1}{e} - (1-\lambda)} \cdot (1/\varepsilon)^{\varepsilon^{-\lambda}((1-\lambda)e^{\lambda-3}) - \lambda\varepsilon^{1-\lambda} + 2\lambda - (1-\lambda)\frac{1}{e} - 1} \\
&= (1/\varepsilon)^{\Omega(1/\varepsilon^\lambda)}.
\end{aligned}$$

The bound for constant d can be obtained similarly. ◀

Finally, notice that assuming a constant dimension d, the upper bound of $O(\log^{d/2}(1/\varepsilon)/\varepsilon^d)$ in Theorem 1 is up to logarithmic factors tight with respect to the lower bound of $\Omega(1/\varepsilon^d)$.

6 A lower bound on the ε-cover for combinatorial range spaces

The lower bound given by Haussler for ε-cover size is $\left(\frac{n}{2e(k+d)}\right)^d$, which is designed for a special range space $(X, \mathcal{R})$ with VC-dimension d, where $n = sd$ for some integer s and $1 \leq k \leq n$, where $\varepsilon = k/n$. But it does not apply specifically to any common geometric range spaces like those defined for points in $\mathbb{R}^d$ and by half-spaces, balls, or fixed-radius balls.

We address this for large d case for half-spaces by providing a new ε-cover size lower bound that is roughly $1/\varepsilon^d$, and thus cannot be similar to what we obtained for kernel range spaces. We then, via a discussion in Appendix A.1, argue this lower bound also holds for ball and fixed-radius ball range spaces.

▶ **Theorem 25.** *Let $\varepsilon \in (0, 0.3)$, $n = d$ and $X = \{e_1, \ldots, e_d\}$ be the vertices of the standard $(d-1)$-simplex in $\mathbb{R}^d$. Then one needs at least M points as an ε-cover for $(X, \mathcal{H})$, where $\mathcal{H}$ denotes the half-space ranges and*

(i) $M = 2^d$ *if* $d \leq 1/\varepsilon$,

(ii) $M = 2^{(1-\varepsilon\log_2(e/\varepsilon))d}$ *if* $d > 1/\varepsilon$. *Notice, in this case,* $1 - \varepsilon\log_2(e/\varepsilon) \to 1$ *and so* $M \to 2^d$ *as* $\varepsilon \to 0$.

Proof. (i) If $d \leq 1/\varepsilon$, in order to get an ε-cover, one needs to consider all 2^d subsets of X, since any two half-spaces h and h' which contain a different subset of points, say A and B, have $d^X_\triangle(h, h') = \frac{1}{n}|A \triangle B| > \varepsilon$. Therefore, in this case there is only one ε-cover $Q = 2^X$.

(ii) Consider a half-space h and a length d binary vector $a = (a_1, \ldots, a_d)$ associated to h by means of $a_i = 1$ if $x_i \in h$ and $a_i = 0$ if $x_i \notin h$. We need to count the number of ranges that differ in at least $\varepsilon d + 1$ points. It means that we are allowed to flip a_i's up to εd times and this flipping does not affect our counting process. Thus there are at most

$$N = \binom{d}{0} + \binom{d}{1} + \cdots + \binom{d}{\varepsilon d}$$

other ranges within an $\varepsilon n = \varepsilon d$ distance of h. By a counting argument, any ε-cover Q will need at least $2^d/N$ elements. We can upper bound N by the well-known inequality $N \leq (\frac{ed}{\varepsilon d})^{\varepsilon d} = (\frac{e}{\varepsilon})^{\varepsilon d}$. Hence,

$$|Q| \geq \frac{2^d}{N} \geq \frac{2^d}{(e/\varepsilon)^{\varepsilon d}} = 2^{(1-\varepsilon\log_2(e/\varepsilon))d}.$$ ◀

▶ **Corollary 26.** *When $n = d$ and $d > 1/\varepsilon$, and $\varepsilon \in (0, c)$ for some constant c that goes to* 0 *(as n and d grow accordingly), then the size of any ε-cover needs to be at least* $\Omega((1/\varepsilon)^{d^{1-o(1)}})$ *where the $o(1)$ shrinks as d grows for* $\log_{1/\varepsilon}(d) \geq 1$.

That is if $n = d$ and both d and $1/\varepsilon$ grow, then an ε-cover requires nearly $1/\varepsilon^d$ ranges, and how close it is to this bound depends on how much faster d grows than $\log_2(1/\varepsilon)$.

Proof. Let $\lambda \in (0, 1/2)$, $\eta \in (0.9, 1)$ and let $c = 1/a$ such that $1 - \varepsilon \log_2(e/\varepsilon) \geq \eta$ and a is chosen such that for any $x > a$ the inequality $\log_2 x \leq \eta x^\lambda$ holds. Let also $X = \{e_1, \ldots, e_d\}$ be the vertices of the $(d-1)$-simplex in $\mathbb{R}^d$, where $n = d = 1/\varepsilon^k$ so $k = \log_{1/\varepsilon}(d)$. By changing the base 2 in case (ii) of Theorem 25 to $1/\varepsilon$ we obtain

$$|Q| \geq 2^{d(1-\varepsilon\log_2(e/\varepsilon))} = \left(\frac{1}{\varepsilon}\right)^{d(1-\varepsilon\log_2(e/\varepsilon))\log_{1/\varepsilon}(2)} \geq \left(\frac{1}{\varepsilon}\right)^{\eta d/\log_2(1/\varepsilon)} \geq \left(\frac{1}{\varepsilon}\right)^{\varepsilon^\lambda d}. \tag{5}$$

Now, if $\lambda/k = o(1)$ hence $\lambda = o(k)$ and thus $(\frac{1}{\varepsilon})^\lambda = (\frac{1}{\varepsilon})^{o(k)}$ or equivalently $\varepsilon^\lambda \geq d^{-o(1)}$. Therefore, by (5), $|Q| \geq (\frac{1}{\varepsilon})^{\varepsilon^\lambda d} \geq (\frac{1}{\varepsilon})^{d^{1-o(1)}}$. ◀

References

1 Pankaj K Agarwal. Range searching. In *Handbook of discrete and computational geometry*, pages 1057–1092. Chapman and Hall/CRC, 2017.

2 Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

3 Carlos Améndola, Alexander Engström, and Christian Haase. Maximum number of modes of Gaussian mixtures. *Information and Inference: A Journal of the IMA*, 9:587–600, 2020.

4 Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 2009.

5 Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, pages 1355–1362, 2012.

6 Miguel A Carreira-Perpinán and Christopher KI Williams. On the number of modes of a Gaussian mixture. In *International Conference on Scale-Space Theories in Computer Vision*, pages 625–640. Springer, 2003.

7 Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 172–183. IEEE, 2020.

8 Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE, 2017.

9 Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel hearding. In *Conference on Uncertainty in Artificial Intellegence*, 2010.

10 Yeshwanth Cherapanamjeri and Jelani Nelson. Terminal embeddings in sublinear time. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2021.

11 Mónika Csikós and Nabil H Mustafa. Optimal approximations made easy. *Information Processing Letters*, 176:106250, 2022.

12 Richard M Dudley, Evarist Giné, and Joel Zinn. Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

13 Herbert Edelsbrunner, Brittany Terese Fasy, and Günter Rote. Add isotropic Gaussian kernels at own risk: More and more resiliant modes in higher dimensions. In *Proceedings 28th Annual Symposium on Computational Geometry (SoCG)*, pages 91–100, 2012.

14 Dylan Fitzpatrick, Yun Ni, and Daniel B Neill. Support vector subset scan for spatial pattern detection. *Computational Statistics & Data Analysis*, 157:107149, 2021.

15 Mingxuan Han, Michael Matheny, and Jeff M Phillips. The kernel spatial scan statistic. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 349–358, 2019.

16 Sariel Har-Peled. *Geometric Approximation Algorithms.* American Mathematical Society, 2011.

17 David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

18 William Herlands, Edward McFowland, Andrew Wilson, and Daniel Neill. Gaussian process subset scanning for anomalous pattern detection in non-iid data. In *International Conference on Artificial Intelligence and Statistics*, pages 425–434. PMLR, 2018.

19 Mark A. Iwen and Mark Philip Roach. On outer bi-lipschitz extensions of linear johnson-lindenstrauss embeddings of low-dimensional submanifolds of $\mathbb{R}^n$. *arXiv:2206.03376v*, pages 1–19, 2022.

20 Haotian Jiang, Tarun Kathuria, Yin Tat Lee, and Swati Padmanabhan. A faster interior point method for semidefinite programming. In *IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, 2020.

21 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*, pages 47–56, 2011.

22 Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993. PMLR, 2019.

23 Matti Karppa, Martin Aumüller, and Rasmus Pagh. Deann: Speeding up kernel-density estimation using approximate nearest neighbor search. In *International Conference on Artificial Intelligence and Statistics*, pages 3108–3137. PMLR, 2022.

24 Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

25 Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (PMLR)*, pages 544–552, 2015.

26 Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the samples complexity of learning. *Journal of Computer ans System Science*, 62:516–527, 2001.

27 David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461. PMLR, 2015.

28 Michael Matheny and Jeff M. Phillips. Computing approximate statistical discrepancy. In *International Symposium on Algorithm and Computation (ISAAC)*, pages 32:1–32:13, 2018.

29 Michael Matheny, Raghvendra Singh, Liang Zhang, Kaiqiang Wang, and Jeff M Phillips. Scalable spatial scan statistics through sampling. In *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2016.

30 Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, pages 839–860, 2003.

31 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* MIT Press, Second Edition, 2018.

32 Nabil H Mustafa. *Sampling in Combinatorial and Geometric Set Systems.* American Mathematical Society (AMS), Mathematical surveys and monographs, 2022.

33 Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *ACM Symposium on Theory of Computing (STOC), ACM*, pages 1064–1069, 2019.

34 Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

35 Jeff M Phillips. ε-samples for kernels. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms (SIAM)*, 2013.

36 Jeff M Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. *Discrete & Computational Geometry*, 63(4):867–887, 2020.

37 Jeff M. Phillips and Pingfan Tang. Sketched mindist. In *International Symposium on Computational Geometry*, 2020.

38 Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2018.

39 Clayton Scott. Rademacher complexity of kernel classes, 2014. URL: `https://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/notes/15_rademacher_kernel.pdf`.

40 Wai Ming Tai. Optimal Coreset for Gaussian Kernel Density Estimation. In *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224, 2022. `doi:10.4230/LIPIcs.SoCG.2022.63`.

41 Michel Talagrand. Sharper bounds for Gaussian and emperical processes. *Annals of Probability*, 22(1):28–76, 1994.

42 Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

43 VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

44 Yan Zheng and Jeff M Phillips. Coresets for kernel regression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654, 2017.

A Appendix

A.1 Equivalence between half-spaces, fixed-radius-balls, and any-radius-balls

Recall that $(X, \mathcal{H})$, $(X, \mathcal{B})$ and $(X, \mathcal{B}_r)$ denote the half-space, ball and fixed-radius-r-ball range spaces respectively, where $X \subset \mathbb{R}^d$ is finite. To be precise, $(X, \mathcal{B})$ shows all possible ranges defined by all possible balls with any radii, while $(X, \mathcal{B}_r)$ shows all possible ranges defined by all possible balls with radius r. Our goal here is to show that these range spaces are roughly equivalent when considering large d.

Consider a range space $(X, \mathcal{H})$. Then for any range $R = X \cap h$ defined by a half-space h, we can choose a large enough radius r_R and identify a radius-r_R ball B_{r_R} so $B_{r_R} \cap X = R = X \cap h$. Indeed, for a sufficiently large radius r for each $R \in (X, \mathcal{H})$, there exists a ball B_r which corresponds to that range. Therefore, if we choose $r = \max_{R \in (X,\mathcal{H})} r_R$ with appropriate centers for each ball, then $(X, \mathcal{H}) \subset (X, \mathcal{B}_r)$. Thus any lower bound for $(X, \mathcal{H})$ also applies to $(X, \mathcal{B}_r)$.

The inclusion $(X, \mathcal{B}_r) \subset (X, \mathcal{B})$ is trivial as $\mathcal{B}_r \subset \mathcal{B}$.

Now let $R \in (X, \mathcal{B})$. Then there is a ball $B_s(p)$ such that $R = B_s(p) \cap X$. We consider the Veronese map $\Psi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ by $x \mapsto (x, \|x\|^2)$. We show that the $\Psi(R)$ can be obtained via the intersection of a half-space h in $\mathbb{R}^{d+1}$. Let $x = (x_1, \ldots, x_d) \in R$. Then $\|x - p\|^2 \leq s^2$. Rewriting this we get $\langle (-2p_1, \ldots, -2p_d, 1), (x_1, \ldots, x_d, \|x\|^2) \rangle \leq s^2 - \|p\|^2$, where $p = (p_1, \ldots, p_d)$. This means that $\Psi(R) = \Psi(X) \cap h_{p,s}$, where $h_{p,s}$ is a half-space in $\mathbb{R}^{d+1}$ defined by $h_{p,s}(y) = \langle (-2p_1, \ldots, -2p_d, 1), (y_1, \ldots, y_d, y_{d+1}) \rangle + \|p\|^2 - s^2$. Therefore, $(X, \mathcal{B})$ can have its corresponding in $(\Psi(X), \mathcal{H})$ where the Veronese map Ψ lifts the dimension d to $d + 1$. Hence any lower bound for $(X, \mathcal{B})$ in $\mathbb{R}^d$ also applies to $(X, \mathcal{H})$ in $\mathbb{R}^{d+1}$.

A.2 Relation between ε-KDE-samples and ε-cover-samples

The following lemma shows that an ε-cover-sample is a $(1+c)\varepsilon$-KDE-sample for any $c>0$.

▶ **Lemma 27.** *Let $S \subset X$ be such that for any $p, q \in \mathbb{R}^d$, $\left|d^X_\triangle(p,q) - d^S_\triangle(p,q)\right| \le \varepsilon$. Then S is a $(1+c)\varepsilon$-KDE-sample for any $c>0$ if the critical radius of K is finite for any $\varepsilon > 0$.*

Proof. Let $q \in \mathbb{R}^d$. Take a point $p \in \mathbb{R}^d$ at infinity (i.e. is very far from all data points). More precisely, take $p \in \mathbb{R}^d$ in such a way that $K(p,x) \le \min\{c\varepsilon/2, K(q,x)\}$ for all $x \in X$. Then $|K(q,x) - K(p,x)| = K(q,x) - K(p,x)$ and by setting $K_X(y) = \frac{1}{|X|}\sum_{x\in X} K(y,x)$ we have

$$\begin{aligned}
|K_X(q) - K_S(q)| &\le |K_X(q) - K_X(p) + K_S(p) - K_S(q)| + |K_X(p) - K_S(p)| \\
&\le |K_X(q) - K_X(p) + K_S(p) - K_S(q)| + K_X(p) + K_S(p) \\
&\le \left|\frac{1}{|X|}\sum_{x\in X}|K(q,x) - K(p,x)| - \frac{1}{|S|}\sum_{s\in S}|K(q,s) - K(p,s)|\right| + c\varepsilon \\
&= \left|d^X_\triangle(p,q) - d^S_\triangle(p,q)\right| + c\varepsilon \\
&\le (1+c)\varepsilon.
\end{aligned}$$

◀

Employing Lemma 27 along with Theorem 4 we obtain the following corollary.

▶ **Corollary 28.** *Let S be an ε-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to a kernel K, where the critical radius of K is finite for any $\varepsilon > 0$. Then S is a $(1+c)\varepsilon$-KDE-sample for any $c > 0$.*

A.3 More on ε-KDE-samples

Corollary 16 gives an input-size- and dimension-free upper bound on the ε-KDE-samples. We remark that if K_1, K_2 satisfy the conditions of Corollary 16, $K_1 + cK_2$ and K_1K_2 also meet the conditions too, for any constant $c>0$. So, a wide variety of kernels work for Corollary 16. We introduce some non-characteristic kernels that do the job.

Consider the truncated Gaussian kernel. That is, let $K(x,p) = e^{-\|x-p\|^2/\sigma^2}$ for x in the super-level set $R_{p,\tau} = \{x : e^{-\|x-p\|^2/\sigma^2} \ge \tau\}$ for some $\tau \in (0,1)$, and 0 elsewhere. Then the modified truncated Gaussian is $(e^{-\|x-p\|^2/\sigma^2} - \tau)/(1-\tau)$ for $x \in R_{p,\tau}$ and 0 elsewhere. Note, the modified truncated Gaussian kernel is $(1, \sqrt{\ln(1/\varepsilon)})$-standard 2-simply computable like the Gaussian kernel.

▶ **Corollary 29.** *With probability $\ge 1-\delta$, a random sample from X is an ε-KDE-sample when K is an Epanechnikov, quartic, triweight, triangle or modified truncated Gaussian kernel and the sample size is $O(\frac{1}{\varepsilon^6}\log^2\frac{1}{\varepsilon\delta})$.*

A.4 Constructing ε-cover-samples from ε-samples

In the following theorem we restate and prove Theorem 4. As we mentioned before Theorem 4, the proof technique mostly is borrowed from Joshi *et al.* [21].

▶ **Theorem 30** (Restatement of Theorem 4)**.** *Let S be an ε-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to a kernel K, where $K \le 1$. Then S is an ε-cover-sample for X.*

Proof. We need to show that for any $p, q \in \mathbb{R}^d$, $\left|d^X_\triangle(p,q) - d^S_\triangle(p,q)\right| \le \varepsilon$.

Suppose $X = \{x_1, \ldots, x_n\}$, $S = \{s_1, \ldots, s_m\}$ and $k = n/m$, where without loss of generality we can assume that k is an integer (otherwise we can work with fractional assignments). Moreover, for the sake of convenience let $E = d^X_\triangle(p,q) - d^S_\triangle(p,q)$. In order to get the desired inequality, we design two different partitions for X and show that $E \le \varepsilon$ and $E \ge -\varepsilon$. Given p and q, again for simplicity let $f(x) = |K(p,x) - K(q,x)|$ (note that $f(x) \le 1$).

Undercounts. For the first partition we sort X and S in decreasing way by their f value, i.e.

$$f(x_1) \ge f(x_2) \ge \cdots \ge f(x_n) \quad \text{and} \quad f(s_1) \ge f(s_2) \ge \cdots \ge f(s_m).$$

Without loss of generality one may assume that $f(s_1) > f(s_2) > \cdots > f(s_m)$ by a tiny perturbation of S. Then any semi-super-level set containing x_i (s_i respectively) will also contain all x_j (s_j respectively) for $j < i$. Then we consider $2m$ (possibly empty) sets $\{P_1, \ldots, P_m\} \cup \{Q_1, \ldots, Q_m\}$ using the sorted order by f. Starting with x_1 (the point with highest f value) we place points in P_j or Q_j following their sorted order. Starting at $i = j = 1$, we place x_i in Q_j as long as $f(x_i) > f(s_j)$ (this can be empty). Then we place the next k points into P_j. After these k points we start by Q_{j+1} and place points in Q_{j+1} as long as $f(x_i) > f(s_{j+1})$. Then we put the next k points of X into P_{i+1}. We continue this process until all of X has been placed in some set. Let $t \le m$ be the index of last set P_j such that $|P_j| = k$. Then $|P_{t+1}| < k$ and $P_j = Q_j = \emptyset$ for all $j > t+1$. We also observe that for all $x_i \in P_j$ (for $j \le t$) we have $f(s_j) \ge f(x_i)$ and so $kf(s_j) \ge \sum_{x_i \in P_j} f(x_i)$ or equivalently, $\frac{1}{m} f(s_j) \ge \frac{1}{n} \sum_{x_i \in P_j} f(x_i)$. We can now bound the undercounts as

$$\begin{aligned} E &= \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(s_j) = \sum_{j=1}^{m}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) + \frac{1}{n}\sum_{x_i \in Q_j} f(x_i)\Big) - \frac{1}{m}\sum_{j=1}^{m} f(s_j) \\ &= \underbrace{\sum_{j=1}^{m}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) - \frac{1}{m} f(s_j)\Big)}_{\le 0} + \sum_{j=1}^{m}\Big(\frac{1}{n}\sum_{x_i \in Q_j} f(x_i)\Big) \le \sum_{j=1}^{t+1}\Big(\frac{1}{n}\sum_{x_i \in Q_j} f(x_i)\Big) \\ &\le \frac{1}{n}\sum_{j=1}^{t+1} |Q_j| = \frac{1}{n}\sum_{j=1}^{t+1} |Q_j \cap A|, \end{aligned} \tag{6}$$

where A is the semi-super-level set $A = \{x \in \mathbb{R}^d : f(x) \ge \tau\} \in \mathcal{A}$ with $\tau = f(x_l)$, where l is the largest index such that $f(x_l) > f(s_{t+1})$. Then A contains s_t but not s_{t+1}. Therefore, $s_j \in A$ for $j \le t$ and $s_j \notin A$ for $j \ge t+1$, and so $|P_j \cap A| = k$ for $j \le t$ and $|P_j \cap A| = 0$ for $j \ge t+1$. Since S is an ε-sample for $(X, \mathcal{A})$, then

$$\sum_{j=1}^{t+1} |Q_j \cap A| = \Big(\sum_{j=1}^{t+1} |Q_j \cap A| + \sum_{j=1}^{t} |P_j \cap A|\Big) - k|S \cap A| = |X \cap A| - k|S \cap A| \le n\varepsilon. \tag{7}$$

Therefore, using (6) and (7) we can write $E \le \varepsilon$.

Overcounts. For the second partition we do overcounts analysis similar to undercounts. In this partitioning we sort X and S in increasing way by their f value, i.e.

$$f(x_1) \le f(x_2) \le \cdots \le f(x_n) \quad \text{and} \quad f(s_1) \le f(s_2) \le \cdots \le f(s_m).$$

Again without loss of generality we assume that $f(s_1) < f(s_2) < \cdots < f(s_m)$. Then we consider $2m$ (possibly empty) sets $\{P_1, \ldots, P_m\} \cup \{Q_1, \ldots, Q_m\}$ using the sorted order by f. Starting with x_1 (the point with lowest f value) we place points in P_j or Q_j following their sorted order. Starting at $i = j = 1$, we place x_i in Q_j as long as $f(x_i) < f(s_j)$ (this may be empty). Then we place the next k points x_i into P_j. After k points are placed in P_j, we begin with Q_{j+1} until all of X has been placed in some set. Let $t \le m$ be the index of last set P_j such that $|P_j| = k$. Then $|P_{t+1}| < k$ and $P_j = Q_j = \emptyset$ for all $j > t+1$. We also observe that for all $x_i \in P_j$ (for $j \le t$), $f(s_j) \le f(x_i)$, and thus $\frac{1}{m} f(s_j) \le \frac{1}{n} \sum_{x_i \in P_j} f(x_i)$. We can now bound the overcounts as

$$\begin{aligned} E &= \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(s_j) = \sum_{j=1}^{m}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) + \frac{1}{n}\sum_{x_i \in Q_j} f(x_i)\Big) - \frac{1}{m}\sum_{j=1}^{m} f(s_j) \\ &= \underbrace{\sum_{j=1}^{t}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) - \frac{1}{m} f(s_j)\Big)}_{\ge 0} + \sum_{j=t+1}^{m}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) - \frac{1}{m} f(s_j)\Big) + \underbrace{\sum_{j=1}^{m}\Big(\frac{1}{n}\sum_{x_i \in Q_j} f(x_i)\Big)}_{\ge 0} \\ &\ge \sum_{j=t+1}^{m}\Big(\frac{1}{n}\sum_{x_i \in P_j} f(x_i) - \frac{1}{m} f(s_j)\Big) = \Big(\frac{1}{n}\sum_{x_i \in P_{t+1}} f(x_i) - \frac{1}{m} f(s_{t+1})\Big) - \frac{1}{m}\sum_{j=t+2}^{m} f(s_j) \\ &\ge \frac{1}{n}\sum_{x_i \in P_{t+1}} f(x_i) - \frac{1}{m} f(s_{t+1}) - \frac{(m-t-1)}{m}. \end{aligned} \tag{8}$$

Now let $A \in \mathcal{A}$ be a semi-super-level set containing no point from $\cup_{j=1}^{m} Q_j$. For instance, A can be $R_{p,q,f(s_u)}$, where u is the largest index such that $Q_u \neq \emptyset$ (note $u \le t+1$). Then $X \cap A = P_u \cup \cdots \cup P_{t+1}$ and $S \cap A = \{s_u, \ldots, s_m\}$, and so $|X \cap A| = (t+1-u)k + |P_{t+1}|$ and $|S \cap A| = m - u + 1$. Hence, because S is an ε-sample for $(X, \mathcal{A})$, we have

$$\begin{aligned} \varepsilon &\ge \frac{1}{m}|S \cap A| - \frac{1}{n}|X \cap A| = \frac{m-u+1}{m} - \frac{(t+1-u)k + |P_{t+1}|}{n} \\ &= \frac{m-u+1}{m} - \frac{(t+1-u)}{m} - \frac{|P_{t+1}|}{n} = \frac{m-t}{m} - \frac{|P_{t+1}|}{n}, \end{aligned}$$

which implies $m - t - 1 \le m\varepsilon + \frac{m}{n}|P_{t+1}| - 1$. Let f attain its minimum on P_{t+1} at $x_i \in P_{t+1}$, so $f(x_i) \ge f(s_{t+1})$. Then by applying (8) we can infer

$$\begin{aligned} E &\ge \frac{1}{n}\sum_{x_i \in P_{t+1}} f(x_i) - \frac{1}{m} f(s_{t+1}) - \frac{(m\varepsilon + \frac{m}{n}|P_{t+1}| - 1)}{m} \\ &= -\varepsilon + \Big(\frac{k - |P_{t+1}|}{n}\Big) - \frac{1}{n}\Big(\sum_{x_i \in P_{t+1}} f(x_i) - k f(s_{t+1})\Big) \\ &\ge -\varepsilon + \Big(\frac{k - |P_{t+1}|}{n}\Big) - \Big(\frac{k - |P_{t+1}|}{n}\Big) f(x_i) \ge -\varepsilon. \end{aligned}$$

◀

A.5 Missing proof elements for lower bound

▶ **Lemma 31** (Restatement of Lemma 22). *Let $S_1, \ldots, S_d$ be d-spheres in $\mathbb{R}^d$, where S_k is centered at e_k with radius $r_k \ge 1$ (e_k is the standard k-th basis vector in $\mathbb{R}^d$). Assume that there are $r \ge 1$ and $\delta \in (0, 1/d)$ such that $|r_k^2 - r^2| < \delta$ for $k = 1, \ldots, d$. Then $|\bigcap_{k=1}^{d} S_k| = 2$.*

Proof. We are looking for two points like $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ such that $\|x - e_k\|^2 = r_k^2$ for $k = 1, \ldots, d$. Writing each equation and gathering similar terms yields

$$x_k = \tfrac{1}{2}(1 + \|x\|^2 - r_k^2), \quad k = 1, \ldots, d. \tag{9}$$

If the system of equations (9) has a solution, then $y = \|x\|^2$ would be one of the solutions to the quadratic equation

$$y = \frac{1}{4}\sum_{k=1}^{d}(1 + y - r_k^2)^2 = \frac{d}{4}y^2 + \frac{1}{2}\sum_{k=1}^{d}(1 - r_k^2)y + \frac{1}{4}\sum_{k=1}^{d}(1 - r_k^2)^2, \tag{10}$$

or equivalently,

$$p(y) = \frac{d}{4}y^2 + \Big(\frac{1}{2}\sum_{k=1}^{d}(1 - r_k^2) - 1\Big)y + \frac{1}{4}\sum_{k=1}^{d}(1 - r_k^2)^2 = 0. \tag{11}$$

Now we consider the discriminant to obtain the criteria to guarantee existence of 2 distinct solutions to the quadratic equation (11):

$$\Delta = \Big(\frac{1}{2}\sum_{k=1}^{d}(1-r_k^2)-1\Big)^2 - \frac{d}{4}\sum_{k=1}^{d}(1-r_k^2)^2 = \frac{1}{4}\Big(\sum_{k=1}^{d}(1-r_k^2)\Big)^2 + 1 + \sum_{k=1}^{d}(r_k^2-1) - \frac{d}{4}\sum_{k=1}^{d}(1-r_k^2)^2.$$

By setting $B = (b_1, \ldots, b_d)$, where $b_i = r_i^2 - 1$ and $b = r^2 - 1$, and using $|b_k - b| < \delta$ we get

$$\begin{aligned} 4\Delta &= \|B\|_1^2 + 4 + 4\|B\|_1 - d\|B\|_2^2 \geq d^2(b-\delta)^2 - d^2(b+\delta)^2 + 4d(b-\delta) + 4 \\ &= -4d^2\delta b + 4db - 4d\delta + 4 = 4(db+1)(1-d\delta). \end{aligned} \tag{12}$$

The condition $\delta < 1/d$ implies $\Delta > 0$, which shows that the quadratic equation (11) has two roots, say y_1, y_2. Moreover, these two roots are positive. (Notice the roots of the quadratic equation $y^2 - by + c = 0$ are real and positive if and only if $b > 0$ and $b^2 \geq 4c > 0$.) Substituting y_1, y_2 in (9) we obtain two points in the intersection of spheres $S_1, \ldots, S_d$ as

$$x_k = \tfrac{1}{2}(1 + y_1 - r_k^2), \quad k = 1, \ldots, d, \quad \text{and} \quad x_k = \tfrac{1}{2}(1 + y_2 - r_k^2), \quad k = 1, \ldots, d.$$

Therefore, $|\bigcap_{k=1}^{d} S_k| = 2$. ◀

A.6 Simple computability of triangle kernel

The following theorem shows that, like the Gaussian and Epanechnikov kernels, triangular kernels are 2-simply computable.

▶ **Theorem 32.** *The triangular kernel $K(x, y) = \max(0, 1 - \|x - y\|)$ is 2-simply computable.*

Proof. Let $p, q, x \in \mathbb{R}^d$, $\tau \in \mathbb{R}^+$ and consider the inequality $|K(p, x) - K(q, x)| \geq \tau$. There are three cases:

1. $\|p - x\| < 1$ and $\|q - x\| \geq 1$. Then verifying $|K(p, x) - K(q, x)| \geq \tau$ is equivalent to verifying $\|x - p\| \leq 1 - \tau$.
2. $\|p - x\| \geq 1$ and $\|q - x\| < 1$. Then verifying $|K(p, x) - K(q, x)| \geq \tau$ is equivalent to verifying $\|x - p\| > 1 - \tau$.
3. $\|p - x\| < 1$ and $\|q - x\| < 1$. Then writing down the inequality $|K(p, x) - K(q, x)| \geq \tau$, we get

$$\sqrt{\textstyle\sum_{i=1}^{d}(x_i - p_i)^2} \geq \tau + \sqrt{\textstyle\sum_{i=1}^{d}(x_i - q_i)^2} \quad \text{or} \quad \sqrt{\textstyle\sum_{i=1}^{d}(x_i - q_i)^2} \geq \tau + \sqrt{\textstyle\sum_{i=1}^{d}(x_i - p_i)^2}.$$

Consider the left hand side inequality (the other comes by symmetry). Squaring both sides and simplifying the equation we obtain the following equation:

$$\Big[-\tau^2 + 2\sum_{i=1}^{d}((q_i - p_i)x_i + p_i^2 - q_i^2)\Big]^2 \geq 4\tau^2\sum_{i=1}^{d}(x_i - q_i)^2, \tag{13}$$

where $x = (x_1, \ldots, x_d)$, $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$.

The cases 1 and 2 can be computed with $O(d)$ arithmetic operations or jumps conditions. The equation (13) shows that Case 3 can also be computed in $O(d)$ time applying the same operations. Therefore, K is 2-simply computable. ◀

B Terminal JL

The algorithm to compute terminal JL, taken almost verbatim from [19], is as follows:

Listing 1 Terminal JL.

```
Input: $\varepsilon \in (0,1)$, $X \subset \mathbb{R}^d$, $|X| = n$, $Q \subset \mathbb{R}^d \setminus X$ and $|Q| = k$.
For any $x \in X$ set $f(x) = (\Pi x, 0)$, where $\Pi$ is a JL map (explained below).
For $q \in Q$:
    (1) Compute $x_{NN} = \operatorname{argmin}_{x \in X} \|x - q\|$,
    (2) Solve the following constrained optimization problem:
        Minimize      $h_{q,x_{NN}}(z) = \|z\|^2 + 2\langle \Pi(q - x_{NN}), z\rangle$
        Subject to    $\|z\|^2 \leq \|q - x_{NN}\|$
                      $|\langle z, \Pi(x - x_{NN})\rangle - \langle q - x_{NN}, x - x_{NN}\rangle| \leq \varepsilon \|q - x_{NN}\| \|x - x_{NN}\|$
                      $(\forall x \in X)$,
    (3) Let $q'$ be the solution to the minimization problem in (2). Set
        $f(q) = (\Pi x_{NN} + q', \sqrt{\|q - x_{NN}\|^2 - \|q'\|^2})$.
Return $f$.
```

Run time analysis. First we need to construct a JL map, i.e. a random matrix $\Phi \in \mathbb{R}^{m \times d}$ with entries from normal distribution $N(0,1)$ normalized by $1/\sqrt{m}$, say $\Pi = 1/\sqrt{m}\Phi$, where $m = O(\log(n)/\varepsilon^2)$. So, we can consider $O(dm)$ as its run time. Calculating Πx for any $x \in X$ needs $O(md)$ times and so for the whole X we need $O(nmd)$ time. Therefore, the run time for embedding X, i.e. computing ΠX, is $O(dm + nmd) = O(dn \log(n)/\varepsilon^2))$.

Now we need to compute the run time for embedding a single $q \in \mathbb{R}^d \setminus X$ in Terminal JL algorithm. Step 1 for finding the nearest neighbor x_{NN} of q needs at most $O(nd)$ time.

The optimization problem for finding $f(q)$ is a semidefinite programming since we can rewrite it as

$$\begin{array}{rl}
\textbf{Minimize} & t \\
\textbf{Subject to} & \langle z, z\rangle + \langle 2\Pi(q - x_{NN}), z\rangle \leq t \\
 & \langle z, z\rangle \leq \|q - x_{NN}\| \\
 & \langle z, \Pi(x - x_{NN})\rangle \leq \varepsilon \|q - x_{NN}\| \|x - x_{NN}\| + \langle q - x_{NN}, x - x_{NN}\rangle \quad \forall x \in X \\
 & \langle z, \Pi(x_{NN} - x)\rangle \leq -\varepsilon \|q - x_{NN}\| \|x - x_{NN}\| - \langle q - x_{NN}, x - x_{NN}\rangle \quad \forall x \in X
\end{array}$$

So, we have $2n + 2$ constraints on $z \in \mathbb{R}^d$. Therefore, according to the literature, the running time for computing $f(u)$, given x_{NN}, is $O(\sqrt{d}(n^3 + d^3)\log(1/\varepsilon))$, see [20]. Therefore, $O(dn\log(n)/\varepsilon^2 + \sqrt{d}(n^3 + d^3)\log(1/\varepsilon))$ can be a (non-optimal) upper bound on the running time of embedding $X \cup \{q\}$ via terminal JL.

C Rademacher complexity of unit ball in RKHS

The following theorem is well known but we could not find it in the literature to cite. So, we present the theorem here and include the proof from [39] for completeness. According to Definition 3.1 of [31] The empirical Rademacher complexity of a function class $\mathcal{F}$ (functions mapping X to $[a,b]$) with respect to the sample $S=\{s_1,\dots,s_m\}$ from X is defined as

$$\hat{\mathfrak{R}}_S(\mathcal{F})=\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(s_i)\right],$$

where $\sigma=(\sigma_1,\dots,\sigma_n)$ and σ_i's are independent uniform random variables from $\{-1,1\}$.

▶ **Theorem 33.** *Let K be a positive definite bounded kernel with $\sup_x\sqrt{K(x,x)}=B$ and let $\mathcal{H}$ be its RKHS. Then for any sample $S=\{s_1,\dots,s_m\}$, $\hat{\mathfrak{R}}_S(B_1^{\mathcal{H}}(0))\le\frac{B}{\sqrt{m}}$, where $B_1^{\mathcal{H}}(0)=\{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\le 1\}$.*

Proof. Fix $S=\{s_1,\dots,s_m\}$. Then

$$\begin{aligned}
\hat{\mathfrak{R}}_S(B_1^{\mathcal{H}}(0)) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in B_1^{\mathcal{H}}(0)}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(s_i)\right]=\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in B_1^{\mathcal{H}}(0)}\sum_{i=1}^{m}\sigma_i\langle f,K(\cdot,s_i)\rangle\right]\\
&=\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in B_1^{\mathcal{H}}(0)}\left\langle f,\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\right\rangle\right]\\
&=\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\left\langle\frac{\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)}{\|\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\|_{\mathcal{H}}},\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\right\rangle\right]\\
&=\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\right\|_{\mathcal{H}}\right]=\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sqrt{\left\|\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\right\|_{\mathcal{H}}^2}\right]\\
&\le\frac{1}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left\|\sum_{i=1}^{m}\sigma_i K(\cdot,s_i)\right\|_{\mathcal{H}}^2}=\frac{1}{m}\sqrt{\sum_{i=1}^{m}\|K(\cdot,s_i)\|_{\mathcal{H}}^2}\\
&=\frac{1}{m}\sqrt{\sum_{i=1}^{m}K(s_i,s_i)}\le\frac{1}{m}\sqrt{mB^2}=\frac{B}{\sqrt{m}}.
\end{aligned}$$

We used Jensen's inequality, reproducing property, $\mathbb{E}_\sigma[\sigma_i\sigma_j]=0$ for $i\neq j$, and the fact that a bounded linear functional obtains its norm in its normalized representer according to the Riesz representation theorem. ◀