For Kernel Range Spaces a Constant Number of Queries Are Sufficient

Jeff M. Phillips and Hasan Pourmahmood-Aghababa
University of Utah
jeffp@cs.utah.edu and h.pourmahmoodaghababa@utah.edu

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 Gaussian kernel $K(p, \cdot) = \exp(-\|p - \cdot\|^2)$. 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 the dimension d. We obtain a bound of $(1/\varepsilon)^{\tilde{O}(1/\varepsilon^2)}$, where $\tilde{O}(f(1/\varepsilon))$ hides log factors in $(1/\varepsilon)$. This implies that when one relaxes 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-dimensional settings. 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.

1 Introduction

Given a data set X a *range space* $(X, \mathcal{R})$ is the collection of possible ways that set can be queried; it is a set of subsets of X defined by ranges $\mathcal{R}$. For a data structure, it specifies the shape of any range query [1]. For machine learning, it categorizes the function class of possible classifiers [13]. For spatial scan statistics, it restricts the family of regions which might form an anomalous hotspot [6]. 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* 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 ranges spaces with bounded VC-dimension [4]. In particular, if the VC dimension is ν, then there exist ε-covers of size $O(1/\varepsilon^\nu)$ and may require that size. It is worth considering the two most common ranges spaces, which includes points $X \subset \mathbb{R}^d$ and subsets defined by either halfspaces, or by balls; both have VC-dimension $d + 1$. Note that through a Veronese map, one can represent any ball query in $\mathbb{R}^d$ by a halfspace query in $\mathbb{R}^{d+1}$. Observe that while the size of the ε-cover $1/\varepsilon^{O(d)}$ does not depend on $n = |X|$ it does however depend exponentially on d.

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, and kernelized versions of data structures queries and scan statistics are also common. Partially motivated by these cases, the complexity of kernel range spaces have also been studied, and in particular coresets for density approximation. These are samples $S \subset X$ (called ε-KDE-*sample* or *ε-sample for* (X, K))) so for every query $p \in \mathbb{R}^d$ that

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

While it is known that for positive and symmetric kernels, a bound of $O(d/\varepsilon^2)$ for such an ε-KDE coreset can be derived based on bounds for ball range spaces [5], more remarkably, these coresets can be constructed of size $O(1/\varepsilon^2)$ [9, 7, 2], 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.In this setting, any *range* in the kernel range space is defined by a point $p \in \mathbb{R}^d$, and reports $(K(p, x_1), K(p, x_2), \ldots, K(p, x_n)) = R_p \in \mathbb{R}^n$, a scalar value (we consider when $K(p, x) \in [0, 1]$) for each $x_i \in X$. This generalizes the notion of a set, where these values are 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

$$d_{\triangle}(R_p, R_q) = \tfrac{1}{|X|} \textstyle\sum_{x_i \in X} |K(p, x_i) - K(q, x_i)| = \tfrac{1}{|X|} \|R_p - R_q\|_1 \leq \varepsilon.$$

Note, in this paper we consider the Gaussian kernel $K(x, y) = e^{-\|x-y\|^2/\sigma^2}$ in $\mathbb{R}^d$; the parameter $\sigma > 0$ for simplicity is elsewhere assumed $\sigma = 1$.

2 Our Results

Our main result is that ε-covers for kernel ranges 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 $O((1/\varepsilon)^{\tilde{O}(1/\varepsilon^2)})$; where $\tilde{O}(f(1/\varepsilon))$ hides polylogarithmic factors in $1/\varepsilon$.

Theorem 2.1. *Let $\varepsilon > 0$ and $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$. There exist a set of size $O\big((\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2} \ln^2(\frac{1}{\varepsilon}))}\big)$ that is an ε-cover of (X, K).*

Moreover, we show that this $(1/\varepsilon)^{\mathsf{poly}(1/\varepsilon)}$ is necessary. In particular, 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 $\mathbb{R}^{\Omega(1/\varepsilon^\lambda)}$.

Theorem 2.2. *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 $X = \{e_1, \ldots, e_d\} \subseteq \mathbb{R}^d$ be the vertices of the standard (d-1)-simplex. 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 $\Omega(1/\varepsilon^d)$.*

When viewed in comparison to the ε-cover size bound for traditional range spaces, e.g. for halfspaces or balls, where the size grows exponentially in d, we believe this result is quite surprising. Almost all learning or data structure bounds, even approximate ones, have exponential dependence on d in the space 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. This should be especially relevant in data analysis applications where a complete trust in data coordinates is rare, it is common to have high dimensional data, and this sort of ε-error is tolerated if not expected. We hope this sheds a bit of light onto why learning in such high-dimensional space is not as challenging as traditional curse-of-dimensionality bounds may suggest.

3 Overview of Techniques and Selected Theorems

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-coresets. 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-coreset rely on properties of reproducing kernels, namely that the kernel density estimate 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 [7, 2] or sampling [9]. However, the ε-cover is a 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 $\mathbb{R}^d$.

Theorem 3.1. *Consider a point set X of size n in $\mathbb{R}^d$ and the Gaussian kernel K. One can construct an ε-cover of size $O(n \ln^{d/2}(1/\varepsilon)/\varepsilon^d)$ for the kernel range space (X, K).*

Reduction of input size n. For the reduction of the size n, we connect this to ε-samples of (traditional) ranges 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 the Gaussian kernel that this VC-dimension bound is $O(d^2)$. So this reduction eliminates the dependence on size n, but increases the dependence on dimension d.

An *ε-cover-sample* for X is a set $S \subseteq X$ such that for any $p, q \in \mathbb{R}^d$, $\left|d_\triangle^X(R_p, R_q) - d_\triangle^S(R_p, R_q)\right| \leq \varepsilon$. 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

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

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 $p, q \in \mathbb{R}^d$, $\tau \in \mathbb{R}^+$. This is extending the idea of super-level sets and linking kernels to range spaces introduced in Joshi *et al.* [5]. As an extension of the linking-based result in [5] to ε-cover-samples that are now semi-linked to an appropriate range space we prove that an ε-sample for $(X, \mathcal{A})$ (i.e. a set $S \subseteq X$ such that $\max_{A \in \mathcal{A}} \left|\frac{|X \cap A|}{|X|} - \frac{|S \cap A|}{|S|}\right| \leq \varepsilon$ [3]), where $\mathcal{A}$ is semi-linked to K, is an ε-cover-sample for X too. The proof mostly follows the strategy of the similar one in [5]. This allows us to remove the dependency on the size of X from Theorem 3.1.

Theorem 3.2. *One can construct an ε-cover with $O\big(s \ln^{d/2}(1/\varepsilon)/\varepsilon^d\big)$ points for the kernel range space (X, K), where s is the size of an ε-sample for $(X, \mathcal{A})$, where $\mathcal{A}$ is semi-linked to K.*

Removing the dependence on dimension d. Similarly, we also show we can eliminate the dependence on dimension d by invoking terminal JL [11]; this reduces the dimension to $O((1/\varepsilon^2) \ln n)$. So while it eliminates the dependence on d, it increases it with respect to the number of points. Luckily, however, we can combine these reductions together in an iterative inductive framework that shows we can eliminate dependence on n and d entirely. The first step is calculating the VC-dimension of the semi-super-level sets.

Let $\mathcal{A}_d = \{R_{p,q,\tau} : p, q \in \mathbb{R}^d, \tau > 0\}$. We have shown that $\dim_{\mathrm{VC}}(\mathcal{A}_d) = O(d^2)$. For a range space $(X, \mathcal{A})$ with VC-dimension ν, a random sample from X of size $O((1/\varepsilon^2)(\nu + \ln 1/\delta))$ is an ε-sample with probability at least $1 - \delta$ [12, 8]. Therefore, we get the following corollary.

Corollary 3.3. *A random sample from the ground set $X \subset \mathbb{R}^d$ of size $O((1/\varepsilon^2)(d^2 + \ln 1/\delta))$ is an ε-cover-sample for X with probability at $1 - \delta$. Hence, $O\big((d^2 + \ln(1/\delta)) \ln^{d/2}(1/\varepsilon)/\varepsilon^{d+2}\big)$ points suffice to construct an ε-cover for (X, K) with probability at least $1 - \delta$.*

The next tool we will make use of to prove Theorem 3.4 (and thus Theorem 2.1) is the concept of *ε-terminal dimensionality reduction* from [11]. Let $\varepsilon \in (0, 1)$ and $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ be arbitrary with $n > 1$. Then there exists a function $f : \mathbb{R}^d \to \mathbb{R}^m$ with $m = O(\ln(n)/\varepsilon^2)$ such that for all $x_i \in \mathbb{R}^d$ and all $p \in \mathbb{R}^d$, $\|p - x_i\| \leq \|f(p) - f(x_i)\| \leq (1 + \varepsilon)\|p - x_i\|$.

Now we can prove our main result that gives an input-size-free and dimension-free upper bound on the size of ε-cover-sample size, which will result in an input-size- and dimension-free upper bound for ε-cover size (Theorem 2.1).

Theorem 3.4. *Let $\varepsilon > 0$, $\delta \in (0, 1)$ and consider a finite point set $X \subset \mathbb{R}^d$. Then with probability at least $1 - \delta$, a random sample of size $O\big(\frac{1}{\varepsilon^6} \ln^2(\frac{1}{\varepsilon\delta})\big)$ from X is an ε-cover-sample for X. Consequently, there exist a set of size $O\big((\frac{1}{\varepsilon})^{O(\frac{1}{\varepsilon^2} \ln^2(\frac{1}{\varepsilon}))}\big)$ that is an ε-cover of (X, K).*

The intuition behind the proof of Theorem 3.4 is recursively 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 $d_i' = O((1/\varepsilon^2)\ln n_i)$, and then Corollary 3.3 to create an ε-cover-sample of size (roughly) $n_{i+1} = O((d_i'/\varepsilon)^2) = O((1/\varepsilon)^6 \ln n_i)$. Combining these steps does not immediately remove the dependence on n (or the initial d), but it does, for instance, 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 $\ln^*(n)$ error factor (since we accumulate ε-error at each recursive step), which ultimately needs to be folded back into the size bound, adjusting $\varepsilon' = \varepsilon/\ln^*(n)$. Instead we apply an inductive argument (inspired by the proof of Theorem 12.3 of [10]), so we only need to argue about one step. That is, we show if we apply the reductions with sufficiently small error parameter ε it can be independent of n and d. It again uses above simple observation but only once. However, this argument is complicated by the two-stage approach because the depdence on n and d are linked, and reducing one relies on the other. However, like the recursive method sketched above, by combining them we can reduce the dependence on both terms.

A Lower Bound on ε-Cover Size. 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 halfspaces. The size is trivially $\Omega(1/\varepsilon)$ in $\mathbb{R}^1$, and as we add each dimension we add an "orthogonal" point 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 $(1/\varepsilon)^d$ bound for fixed-radius disks. However, interestingly, this construction stops working for kernels as we approach $1/\varepsilon$ dimensions. Indeed, we complement the upper bound of $(1/\varepsilon)^{\tilde{O}(1/\varepsilon^2)}$ with an example that requires an ε-cover of size $(1/\varepsilon)^{\Omega(1/\varepsilon^\lambda)}$ for any $\lambda \in (0,1)$ in $\mathbb{R}^{\Omega(1/\varepsilon^\lambda)}$.

As an overview of the proof, we prove a lemma to find some criteria on d spheres in $\mathbb{R}^d$ that can generate exactly two points in their intersections. Then we use this in another lemma to design a point set that provides us with the desired lower bound (Theorem 2.2). These two lemmas are not presented here for the sake of page limit constraint. Finally, notice that assuming a constant dimension d, the upper bound of $O(\ln^{d/2}(1/\varepsilon)/\varepsilon^d)$ in Theorem 3.1 is up to logarithmic factors tight with respect to the lower bound of $\Omega(1/\varepsilon^d)$.

References

[1] Pankaj K Agarwal. Range searching. In *Handbook of discrete and computational geometry*, pages 1057–1092. 2017.

[2] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

[3] Sariel Har-Peled. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.

[4] 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.

[5] 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.

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

[7] Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552. PMLR, 2015.

[8] 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.

[9] 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.

[10] Nabil H Mustafa. *Sampling in Combinatorial and Geometric Set Systems*. AMS, Mathematical surveys and monographs, 2022.

[11] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *STOC*, pages 1064–1069, 2019.

[12] Michel Talagrand. Sharper bounds for gaussian and emperical processes. *Annals of Probability*, 22(1):28–76, 1994.

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