On the numerical rank of radial basis function kernel matrices in high dimension

Ruoxi Wang * Yingzhou Li * Eric Darve †

June 27, 2017

Abstract

Low-rank approximations are popular techniques to reduce the high computational cost of large-scale kernel matrices, which are of significant interest in many applications. The success of low-rank methods hinges on the matrix rank, and in practice, these methods are effective even for high-dimensional datasets. The practical success has elicited the theoretical analysis of the rank in this paper. We will consider radial basis functions (RBF) and present theorems on the rank and error bounds. Our three main results are as follows. First, the rank of RBFs grows polynomially with the data dimension, in the worst case; second, precise approximation error bounds in terms of function properties and data structures are derived; and last, a group pattern in the decay of singular values for RBF kernel matrices is analyzed, and is explained by a grouping of the expansion terms. Empirical results verify and illustrate the theoretical results.

1 Introduction

With the increase in the size and dimensionality of datasets, large-scale high-dimensional dense matrices have become a central part of many linear systems. Kernel matrices, especially those related to radial basis functions (RBF), have attracted attention in a variety of fields including machine learning, inverse problems, graph theory and PDEs [30, 35, 25, 18, 20, 19, 29] due to their good interpolation properties. Examples of interest in machine learning include kernel ridge regression [34, 6] and kernel support vector machines [5]. Regrettably, kernel methods often involve operations on large-scale matrices, which can be computationally challenging.

The broad applications of kernel matrices and the high computational cost have made algorithms that accelerate matrix computations particularly important. There have been many algebraic algorithms proposed to reduce the computational burden, mostly based on matrix approximations. One popular class of algorithms relies on low-rank approximations of the matrix or certain submatrices [35]. The singular value decomposition (SVD) [16] is optimal but has an undesirable cubic complexity. Many methods have been proposed to accelerate the low-rank construction with an acceptable sacrifice in the accuracy [22, 17, 21, 26, 9, 15, 8, 38].

The success of these low-rank algorithms hinges on a large spectrum gap, or a fast decay of the spectrum of the matrix or some of its submatrices. However, to our knowledge, there is no theoretical guarantee that this is always the case. For d-variable functions with bounded derivatives up to the q-th order, a classical analysis based on tensor-product grids [7] shows that to achieve an approximation error ϵ, the rank needed is $R = O(\epsilon^{-d/q})$.

The exponential growth with d is striking because even for a moderate d, a reasonable accuracy should be expected to be difficult to achieve. However, in practice, people have observed much lower ranks. Algebraic low-rank techniques work effectively in many cases where the data dimension ranges from moderate to high. One high-level explanation is that both the functions and the data of practical interest enjoy special properties and latent structures.

Despite the popularity of applying low-rank algorithms to RBF kernels, the relation between the rank and data dimension has not been described analytically. The aim of this paper is to provide a theoretical foundation for the practical success of low-rank matrix algorithms. In this

*Institute for Computational and Mathematical Engineering, Stanford University, Email: {ruoxi, ryanlee}@stanford.edu

†Department of Mechanical Engineering, Stanford University, Email: darve@stanford.edu

paper, we present two main results. First, we show that under common smoothness assumptions, the rank of RBF kernels grows polynomially with the dimension d in the worst case. We also derive explicit error bounds. Second, we examine the decay of the singular values of RBF kernel matrices, and explain the observed "decay-plateau" behavior.

1.1 Related Work

There has been an extensive interest in kernel properties in a high-dimensional setting.

One line of research focuses on the spectrum of kernel random matrices. There is a rich literature on the smallest eigenvalues, mainly concerning the matrix conditioning. [1, 23, 27, 28] provided lower bounds for the smallest eigenvalues, and Ball [2] discussed the upper bound. Some work further studied the eigenvalue distributions. Karoui [10] obtained the spectral distribution in the limit by applying a second-order Taylor expansion to the kernel function. In particular, Karoui considered kernel matrices with the (i, j)-th entry $K(\mathbf{x}_i^T\mathbf{x}_j/h^2)$ and $K(\|\mathbf{x}_i - \mathbf{x}_j\|^2/h^2)$, and showed that as data dimension $d \to \infty$, the spectral property is the same as that of the covariance matrix $\frac{1}{d}XX^T$. Wathen [36] described the eigenvalue distribution of RBF kernel matrices more explicitly. Specifically, the authors provided formulas to calculate the number of eigenvalues that decay like $(1/h)^{2k}$ as $h \to \infty$, for a given k. This group pattern in eigenvalues was observed earlier in [12] but with no explanation. The same pattern also occurs in the coefficients of the orthogonal expansion in the RBF-QR method proposed in [11].

Another line of research is on developing efficient methods for function expansion and interpolation. The goal is to diminish the exponential dependence on the data dimension introduced by the tensor-product based approach. Barthelmann [3] considered polynomial interpolation on a sparse grid [13]. Sparse grids are based on a high-dimensional multiscale basis and involves only $O(N(\log N)^{d-1})$ degrees of freedom, where N is the number of grid points in one coordinate direction at the boundary. This is in contrast with the $O(N^d)$ degrees of freedom from tensor-product grids. Barthelmann showed that when $d \to \infty$, the number of selected points grows as $O(d^k)$, where k is related to the smoothness of the function.

Trefethen [33] commented that most methods (including sparse grid) aiming to mitigate the curse of dimensionality have taken advantage of the rotational or translational non-uniformity of the underlying functions. This may cause over- or under- sampling in one direction to achieve a certain accuracy. To ensure a uniform resolution in all directions, Trefethen suggested that the Euclidean degree may be helpful. He investigated the complexity of polynomials with degree defined by 1-, 2- and $\infty-$ norms and concluded that by using the 2-norm we achieve similar accuracy as with the ∞-norm, but with $d!$ fewer points.

1.2 Main Results

In this paper, we study radial basis functions (RBF). An RBF is of the form $\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|^2)$. Such functions include but are not limited to the Gaussian kernel and the Cauchy kernel. The numerical rank of an RBF related to error ϵ is defined as

$$R_\epsilon = \min \left\{ r \mid \exists\ \{h_i\}_{i=1}^r, \{g_i\}_{i=1}^r, \text{ s.t. } \forall\ \mathbf{x}, \mathbf{y}, \left| \mathcal{K}(\mathbf{x}, \mathbf{y}) - \sum_{i=1}^{r} g_i(\mathbf{x}) h_i(\mathbf{y}) \right| \le \epsilon \right\}$$

Our two main results are as follows. First, we show that under common smoothness assumptions of RBFs, the rank for RBF kernels is a polynomial function of the data dimension d. In particular, for RBF kernel matrices, the rank $R = O(d^q)$ where q is related to the approximation error. We will prove precise error bounds.

Second, we observe that the singular values for RBF kernel matrices form groups with plateaus. A pictorial example is in Figure 1.

There are 5 groups (plateau) of singular values with a sharp drop in magnitude between groups; the group cardinalities depend on the data dimension, but are independent of the data size. We explain this phenomenon by applying an appropriate analytic expansion of the function and grouping expansion terms appropriately.

1.3 Organization

This paper is organized as follows. Section 2 presents our theorems on the function rank and the error bound. Section 3 provides the theorem proofs. Section 4 shows the optimality of the

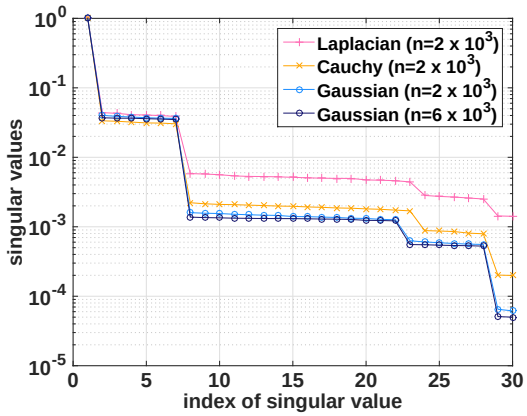


Figure 1: Group patterns in singular values. The singular values are normalized and ordered s.t. $1 = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. The data were randomly generated in dimension 6. The legend shows the data size and the kernel functions: Laplacian ($\exp(-\|x-y\|)$), Cauchy ($1/(1+\|x-y\|^2)$), and Gaussian ($\exp(-\|x-y\|^2)$).

derived rank. Section 5 verifies our theorems experimentally. Section 6 investigates and discusses the group pattern in the singular values for RBF kernel matrices.

2 Main Theorems

In this section, we present three theorems on the rank of an RBF function, and the approximation error bound. The first two theorems rely on the Chebyshev expansion. They consider functions with different smoothness assumptions, and show that the function rank grows polynomially with increasing data dimension, and that the approximation error decreases with decreasing dataset radius. The third theorem relies on the Fourier expansion. It provides a separable form that captures more subtle relations between the error and rank, and shows that the error decreases with either the source or target dataset radius.

Before stating our theorems, we introduce some notations. Let $\mathbf{E}(x)$ and $\mathbf{Var}(x)$ denote the expectation and variance of x, respectively. Let

$$E_{\rho^2} =: \left\{ z = \frac{\rho^2 e^{\imath\theta} + \rho^{-2} e^{-\imath\theta}}{2} \,|\, \theta \in [0, 2\pi) \right\}$$

be the *Bernstein ellipse* defined on $[-1, 1]$ with parameter ρ^2, an open region bounded by an ellipse. For an arbitrary interval, the ellipse is scaled and shifted.

In the following theorems, we assume that the bandwidth parameter $h = 1$. In fact, h is tightly related to the scale of point sets; we can always adjust h indirectly by scaling the data points.

Theorem 2.1. *Consider kernel function $\mathcal{K}$ and point sets $\Omega \subset \mathbb{R}^d$. We assume that*

1. *$\mathcal{K}$ is a shift-invariant kernel,* i.e., *$\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|^2)$.*
2. *f is analytic in $[0, D^2]$ where D is the diameter for Ω, and is analytically continuable to a Bernstein ellipse defined on $[0, D^2]$ with parameter ρ^2, and $|f(x)| \leq C$ for some C.*
3. *Vectors from Ω are statistically independent with their infinity norms bounded by M.*

Under conditions 1 and 2, there exist sequences of real-valued functions, $\{g_i\}_{i=1}^R$ and $\{h_i\}_{i=1}^R$, such that

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{R} g_i(\mathbf{x}) h_i(\mathbf{y}) + \epsilon, \quad \forall\, \mathbf{x}, \mathbf{y} \in \Omega$$

where $R = \binom{n+d+2}{d+2}$ is the approximation rank of $\mathcal{K}(\mathbf{x}, \mathbf{y})$, n is the degree of Chebyshev partial sum of $f(r)$, and for $n \geq 0$ the approximation error ϵ is bounded as follows:

$$|\epsilon| \leq \frac{2CD^2}{2\rho^2 - D^2} \left(\frac{2\rho^2}{D^2} \right)^{-n}$$

Further, if condition 3 holds, then with probability at least

$$1 - 2\exp\left(\frac{-d\delta^2}{2\sigma^2 + 8M^2\delta/3}\right),$$

the approximation error ϵ is bounded as follows:

$$|\epsilon| \leq \frac{2Cd\delta}{2\rho^2 - d\delta}\left(\frac{2\rho^2}{d\delta}\right)^{-n}$$

and

$$\|\mathbf{x} - \mathbf{y}\|^2 \in [E_d - d\delta, E_d + d\delta]$$

where $\sigma^2 = \frac{1}{d}\sum_{i=1}^{d} \mathbf{Var}[(x_i - y_i)^2]$ and $E_d = \sum_{i=1}^{d} \mathbf{E}[(x_i - y_i)^2]$.

Theorem 2.1 describes many insightful relations between dimension d, rank R, and approximation error ϵ. First, the rank grows polynomially with increasing data dimension d, *i.e.*, $R = O(d^n)$. This can be seen from $R = \binom{n+d+2}{d+2}$ with n fixed and $d \to \infty$.

Second, the error bound decreases with decreasing data diameter D. This suggests that decreasing the dataset radius will increase the effectiveness of low-rank algorithms. In other words, if the data are tightly clustered together, we will observe a small approximation error; if the data spread out uniformly in the space, we will likely encounter a large approximation error.

Third, the error bound decreases with increasing dimension d in the probabilistic sense. When d increases, the phenomenon of concentration of measures starts to reduce the variance in the pairwise distances $\|\mathbf{x} - \mathbf{y}\|$. That is, most values of $\|\mathbf{x} - \mathbf{y}\|$'s will concentrate in a small-sized subinterval of $[0, D]$, which in turn shrinks the data radius.

The analytic assumption in Theorem 2.1 is very strong because many RBFs are not infinitely differentiable when the domain contains zero. However, most RBFs of practical interest are q-times differentiable. In the following theorem, we weaken the analytic assumption to a differentiable assumption and compute the estimation of the corresponding error bound.

Theorem 2.2. *Consider the same kernel function and point sets in Theorem 2.1. Suppose the assumptions 1 and 3 in Theorem 2.1 hold, but with condition 2 replaced by*

2' let f and its derivatives through $f^{(q-1)}$ be absolutely continuous in $[0, D^2]$ and the q-th derivative has bounded total variation on $[0, D^2]$, $V(\frac{\mathrm{d}^q f}{\mathrm{d}x^q}) \leq V_q$.

Then we arrive at the same conclusions, but with a different upper bound. That is, for $n > q$ the approximation error ϵ is bounded as follows:

$$|\epsilon| \leq \frac{2V_q D^{2q}}{\pi q[2(n-q)]^q}$$

Further, with a probability of at least $1 - 2\exp(\frac{-d\delta^2}{2\sigma^2 + 8M^2\delta/3})$,

$$|\epsilon| \leq \frac{2V_q (d\delta)^{2q}}{\pi q[2(n-q)]^q}$$

Compared to Theorem 2.1, the major difference in Theorem 2.2 lies in the convergence rate and effect of data diameter D. The convergence rate slows down from a nice geometric convergence rate $O((\rho^2/D^2)^{-n})$ to an algebraic convergence rate $O(n^{-q})$. Each time the function becomes one derivative smoother (q increased by 1), the convergence rate will also become one order faster. The dataset diameter D affects the error bound through D^{2q} instead of D^{2n}, where q is determined by the function but n is adjustable from the function expansion.

One drawback of the above two theorems is the lack of consideration of the source/target dataset radius. As a result, the error bound fails to explain a known fact that a kernel matrix with underlying data tightly clustered around k points has an approximation rank that is approximately k. In the following theorem, we include the radius of target point set $\Omega_{\mathbf{x}}$ and source point set $\Omega_{\mathbf{y}}$ in the error bound. Our result theoretically supports the motivation behind many algorithms that take advantage of the low-rank property of submatrices.

Theorem 2.3. *Consider kernel function $\mathcal{K}$ and point sets $\Omega_{\mathbf{x}} \subset \mathbb{R}^d, \Omega_{\mathbf{y}} \subset \mathbb{R}^d$. Suppose the assumptions in Theorem 2.1 hold, but with condition 2 replaced by*

2" Denote by $G(x) = \mathcal{T}f(x)$ a pattern function[1] of f. Let $g(x) = \sum_n G(x + n\rho^2)$ be the periodic extension of $f(x)$ where $\sup_{x\in[0,D^2]} |f(x) - p(x)| \le \epsilon_P$, with its derivatives through $p^{(q-1)}$ be continuous, and the q-th derivative be piecewise continuous with the total variation over one period ρ^2 bounded by V_q.

Then, we obtain the following error bound:

$$|\epsilon| \le \frac{\|p\|}{e\rho}\left(\frac{4\pi e D_{\mathbf{x}} D_{\mathbf{y}}}{\rho^2}\right)^{(M_t+1)}\left(\frac{M_f+1}{M_t+1}\right)^{M_t+3/2} + \frac{V_q T^q}{\pi q(2\pi M_f)^q} + \epsilon_P$$

where $\|\cdot\| = \|\cdot\|_{L_2[0,\rho^2]}$, $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ are the radius of $\Omega_{\mathbf{x}}$ and $\Omega_{\mathbf{y}}$, respectively, and M_f is the number of terms in the Fourier expansion of $p(x)$, M_t is the number of terms in the Taylor expansion of $\exp(-\imath x)$, both of which are independent of the point sets.

$$R = 4M_f\binom{M_t+d}{d}$$

is the approximation rank of $\mathcal{K}(\mathbf{x},\mathbf{y})$.

The error bound in Theorem 2.3 includes three sources: the function periodic extension, the Fourier expansion, and the Taylor expansion. The radius of the two point sets occur in the error bound through term $(D_{\mathbf{x}}D_{\mathbf{y}})^{M_t+1}$, suggesting that shrinking the radius of either point set will decrease the error. As mentioned above, this property has motivated people to approach matrix approximation problems by identifying low-rank blocks in a matrix, which is partially achieved by partitioning the data into clusters of small radius.

The rank grows as $R = O(d^{M_t})$, when M_f and M_t are fixed and $d \to \infty$. Finer rank and error analysis can be performed by decreasing the number of Taylor expansion terms M_t with the Fourier expansion orders. For simplicity we assume M_t to be fixed.

The periodic extension error ϵ_P depends on the extension methods (see [4] for details). When the bell function $\mathcal{T} \equiv 1$ is in the physical domain of f, $\epsilon_P = 0$.

In the end, considering that the derived error bounds correspond to function's infinity norm, we relate the function's infinity norm to matrix norms of primary interest.

Lemma 2.4. *Assume that a real-valued function $\mathcal{K}$ can be a approximated by two sequences of real-valued functions $\{g_i(\mathbf{x})\}_{i=1}^R$ and $\{h_i(\mathbf{y})\}_{i=1}^R$,* i.e.*,*

$$\left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R} g_i(\mathbf{x})h_i(\mathbf{y})\right| \le \epsilon, \quad \mathbf{x}\in\Omega_X, \mathbf{y}\in\Omega_Y$$

Consider any kernel matrix $K \in \mathbb{R}^{n\times n}$ with entry $K_{ij} = \mathcal{K}(\mathbf{x}_i,\mathbf{y}_j)$ where $\mathbf{x}_i \in \Omega_{\mathbf{x}}$ and $\mathbf{y}_j \in \Omega_{\mathbf{y}}$. Let $G, H \in \mathbb{R}^{n\times R}$ be two matrices with entry $G_{ij} = g_j(\mathbf{x}_i)$ and $H_{ij} = h_j(\mathbf{y}_i)$, respectively. Then, K can be approximated by GH^T with error bounded by the following:

$$\|K - GH^T\|_F \le \epsilon n, \quad \|K - GH^T\|_2 \le \epsilon n, \quad \text{and} \quad \|K - GH^T\|_{\max} \le \epsilon$$

where $\|\cdot\|_F, \|\cdot\|_2$ are the Frobenius norm and two norm, respectively. $\|\cdot\|_{\max}$ is the maximum norm, and $\|K\|_{\max} = \max_{i,j}|K_{ij}|$.

Remark. We summarize the assumptions, error bound and rank estimation of each theorem in Table 1, and discuss two main differences from the perspective of the rank and the error bound. The rank in the Fourier expansion approach grows with the expansion order at a slower rate than that in the Chebyshev approach. The error bound in the Fourier approach considered both $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$, whereas the Chebyshev approach only considered the radius of the entire point set. That $(D_{\mathbf{x}}D_{\mathbf{y}})^{M_t+1}$ occurs in the error bound explains why point clustering and local interpolations generally leads to a more memory efficient approximation.

[1] A pattern function G related to f is defined by f multiplied by a bell function $\mathcal{T}$. $\mathcal{T}$ is approximately 1 in the physical region and decays to zero smoothly outside of the region. See details in [4].

Table 1: Theorem Summary

Approach	Chebyshev expansion + Exact expansion of $\|\mathbf{x}-\mathbf{y}\|^{2l}$		Fourier expansion + Taylor expansion of $\exp(\imath\mathbf{x}^T\mathbf{y})$
Condition	• f is analytic in $[0, D^2]$, and is analytically continuable to a Bernstein ellipse defined on $[0, D^2]$ with parameter ρ^2 • $\|f(x)\| \leq C$ for some $C < \infty$	• The first $q-1$ derivatives of f are absolutely continuous on $[0, D^2]$ • The q-th derivative on $[0, D^2]$ has bounded total variation V_q	• Let $p(x)$ be the ρ^2-periodic extension of f and $\sup_{x\in[0,D^2]} \|f(x) - p(x)\| \leq \epsilon_P$ • The first $q-1$ derivatives of p are continuous • The q-th derivative on $[0, \rho^2]$ is piece-wise continuous with bounded total variation V_q
Error	$\frac{2CD^2}{2\rho^2 - D^2}\left(\frac{2\rho^2}{D^2}\right)^{-n}$	$\frac{2V_q D^{2q}}{\pi q[2(n-q)]^q}$	$\left(\frac{4\pi e D_{\mathbf{x}} D_{\mathbf{y}}}{\rho^2}\right)^{M_t+1}\left(\frac{M_f+1}{M_t+1}\right)^{M_t+\frac{3}{2}} \times \frac{\|p\|}{e\rho^2} + \frac{V_q \rho^{2q}}{\pi q(2\pi M_f)^q} + \epsilon_P$
Rank	$\binom{n+d+2}{d+2}$		$4M_f\binom{M_t+d}{d}$
Notation	n: degree of Chebyshev expansion D: diameter of Ω		$D_{\mathbf{x}}$: radius of $\Omega_{\mathbf{x}}$; $D_{\mathbf{y}}$: radius of $\Omega_{\mathbf{y}}$ M_f: # terms in Fourier expansion M_t: # terms in Taylor expansion

3 Theorem Proof

In this section, we provide proofs for the theorems in section 2. All the proofs aim to separate $\mathcal{K}(\mathbf{x}, \mathbf{y})$ into functions of $\mathbf{x}$ and of $\mathbf{y}$. The proofs of Theorem 2.1 and Theorem 2.2 rely on a Chebyshev expansion of $\mathcal{K}$ followed by an expansion of the terms $\|\mathbf{x}-\mathbf{y}\|^{2l}$. The proof of Theorem 2.3 first applies a Fourier expansion on $\mathcal{K}$ to extract the cross term $\exp(\mathbf{x}^T\mathbf{y})$, then it applies a Taylor expansion on the cross term to complete the separation.

Notations. For multi-index $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_d] \in \mathbb{N}^d$ and vector $\mathbf{x} = [x_1, \cdots, x_d] \in \mathbb{R}^d$, we define $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$, $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ and the multinomial coefficient with $|\boldsymbol{\alpha}| = m$ to be $\binom{m}{\boldsymbol{\alpha}} = \frac{m!}{\alpha_1!\alpha_2!\cdots\alpha_d!}$.

3.1 Proof of Theorem 2.1

Proof. The proof consists of four components. First, we apply a linear transformation of the data. Second, we apply the Chebyshev expansion theorem on the kernel function $\mathcal{K}$. Third, we write the approximated function in a separable form and compute the corresponding error bound and approximation rank. And last, we consider and the phenomenon of concentration of measures.

1. *Linear transformation of data.* $\forall\ \mathbf{x}, \mathbf{y} \in \Omega$, let $D = \max_{\mathbf{x},\mathbf{y}\in\Omega}\|\mathbf{x}-\mathbf{y}\|$. We denote by $\mathbf{x}' = \frac{\sqrt{2}}{D}\mathbf{x}$ the transformed data. Then

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|^2) = f\left(\frac{D^2}{2}\|\mathbf{x}'-\mathbf{y}'\|^2\right) = \tilde{f}(\|\mathbf{x}'-\mathbf{y}'\|^2) \tag{1}$$

2. *Chebyshev approximation.* We approximate $\tilde{f}(r)$ by its Chebyshev partial sum of degree n:

$$\tilde{f}(r) = \sum_{k=0}^{n} c_k T_k(r) + \epsilon_n \tag{2}$$

where $c_k = \frac{2}{\pi}\int_{-1}^{1} \frac{\tilde{f}(r)T_k(r)}{\sqrt{1-r^2}} dr$, and $T_n(r)$ is the Chebyshev polynomial of the first kind of degree n defined by the relation:

$$T_n(x) = \cos(n\theta), \text{ with } x = \cos(\theta) \tag{3}$$

Because f is analytic in $[0, D^2]$, and is analytically continuable to the Bernstein ellipse defined on $[0, D^2]$ with parameter ρ^2, g is also analytic in $[0, 2]$ and is analytically continuable to the open Bernstein ellipse defined on $[0, 2]$ with parameter $\tilde{\rho}^2 = \frac{2}{D^2}\rho^2$. According to Theorem 8.2 in [32], $\forall\ n \geq 0$, the error term ϵ_n satisfies:

$$|\epsilon_n| \leq \frac{2C\tilde{\rho}^{-2n}}{\tilde{\rho}^2 - 1} = \frac{2CD^2}{2\rho^2 - D^2}\left(\frac{2\rho^2}{D^2}\right)^{-n} \tag{4}$$

Because $T_k(r)$ is a Chebyshev polynomial with degree k, by rearranging the terms we obtain

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \tilde{f}(\|\mathbf{x}' - \mathbf{y}'\|^2) = \sum_{k=0}^{n} a_k \|\mathbf{x}' - \mathbf{y}'\|^{2k} + \epsilon_n = \sum_{k=0}^{n} \frac{a_k}{D^{2k}} \|\mathbf{x} - \mathbf{y}\|^{2k} + \epsilon_n \tag{5}$$

where a_k depends on c_k but is independent of $\mathbf{x}$ and $\mathbf{y}$.

3. *Separable form.* We separate each term $\|\mathbf{x} - \mathbf{y}\|^{2l}$ in (5) into functions of $\mathbf{x}$ and $\mathbf{y}$:

$$\|\mathbf{x} - \mathbf{y}\|^{2l} = \sum_{k=0}^{l} \sum_{j=0}^{k} \sum_{|\boldsymbol{\alpha}|=l-k} C_{l,k,\boldsymbol{\alpha}} \left(\|\mathbf{x}\|^{2j}\mathbf{x}^{\boldsymbol{\alpha}}\right)\left(\|\mathbf{y}\|^{2(k-j)}\mathbf{y}^{\boldsymbol{\alpha}}\right) \tag{6}$$

where $C_{l,k,\boldsymbol{\alpha}} = (-2)^{l-k}\binom{l}{k}\binom{k}{j}\binom{l-k}{\boldsymbol{\alpha}}$. Substituting (6) to (5), we obtain a separate form of $\mathcal{K}$:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_{l=0}^{n} \sum_{k=0}^{l} \sum_{j=0}^{k} \sum_{|\boldsymbol{\alpha}|=l-k} D_{l,k,\boldsymbol{\alpha}} \left(\|\mathbf{x}\|^{2j}\mathbf{x}^{\boldsymbol{\alpha}}\right)\left(\|\mathbf{y}\|^{2(k-j)}\mathbf{y}^{\boldsymbol{\alpha}}\right) + \epsilon_n \tag{7}$$

where $D_{l,k,\boldsymbol{\alpha}} = \frac{a_l}{D^{2l}}(-2)^{l-k}\binom{l}{k}\binom{k}{j}\binom{l-k}{\boldsymbol{\alpha}}$ is a constant independent of the point set. The rank of $\mathcal{K}$ can be bounded by the total number of separate terms, which is:

$$\sum_{l=0}^{n}\sum_{k=0}^{l}(k+1)\binom{l-k+d-1}{d-1} = \binom{n+d+2}{d+2}$$

To summarize, $\mathcal{K}(\mathbf{x}, \mathbf{y})$ can be written in a separable form with rank

$$R = \binom{n+d+2}{d+2} \tag{8}$$

and approximation error

$$|\epsilon| \leq \frac{2CD^2}{2\rho^2 - D^2}\left(\frac{2\rho^2}{D^2}\right)^{-n} \tag{9}$$

4. *Concentration of measure.* We consider the concentration of measure phenomenon. When dimension d increases, with a high probability the values of $\|\mathbf{x} - \mathbf{y}\|$ will concentrate in a small-sized subinterval of $[0,\ D]$. Therefore, the upper bound from the above analysis would be loose in high-dimensional settings.

   We apply concentration inequalities to achieve a sharper bound in the probability sense. We assume that the vectors in the dataset are statistically independent, with their entries bounded by $M < \infty$. If the vectors are statistically dependent, we can obtain a better rank result. We denote $z_i = (x_i - y_i)^2 - \mathbf{E}[(x_i - y_i)^2]$. From the assumptions, z_i's are statistically independent with mean zero and are bounded by $4M^2$. Let us denote $\sigma^2 = \frac{1}{d}\sum_{i=1}^{d}\mathbf{Var}(z_i)$. By applying the Bernstein's inequality on z_i's, we conclude that $\forall \delta \geq 0$,

$$P\left(\left|\|\mathbf{x} - \mathbf{y}\|^2 - E_d\right| \leq d\delta\right) \geq 1 - 2\exp\left(\frac{-d\delta^2}{2\sigma^2 + 8M^2\delta/3}\right) \tag{10}$$

   where $E_d = \sum_{i=1}^{d}\mathbf{E}[(x_i - y_i)^2]$. In other words, with probability at least

$$1 - 2\exp\left(\frac{-d\delta^2}{2\sigma^2 + 8M^2\delta/3}\right)$$

$\|\mathbf{x}-\mathbf{y}\|^2 \in [E_d - d\delta, E_d + d\delta]$. If we adopt the same analysis as above, and denote $\mathbf{x}' = \sqrt{\frac{2}{d\delta}}\,\mathbf{x}$, then

$$\mathcal{K}(\mathbf{x},\mathbf{y}) = f\left(\|\mathbf{x}-\mathbf{y}\|^2\right) = \tilde{f}\left(\|\mathbf{x}'-\mathbf{y}'\|^2 - \frac{2E_d}{d\delta}\right)$$

Because $\tilde{f}$ also satisfies the smoothness assumptions, the analysis in item 2 applies here and the error term satisfies:

$$|\epsilon_n| \le \frac{2Cd\delta}{2\rho^2 - d\delta}\left(\frac{2\rho^2}{d\delta}\right)^{-n}$$

This sharper bound can be achieved with the same function rank as in (8). Rewriting the separable form for $\mathcal{K}(\mathbf{x},\mathbf{y})$

$$\begin{aligned}\mathcal{K}(\mathbf{x},\mathbf{y}) &= \sum_{k=0}^{n} a_k \left(\|\mathbf{x}'-\mathbf{y}'\|^2 - \frac{2E_d}{d\delta}\right)^k + \epsilon_n \\ &= \sum_{i=0}^{n}\left(\sum_{k=i}^{n}\frac{a_k}{c}\binom{k}{i}\left(\frac{-2E_d}{d\delta}\right)^{k-i}\right)\|\mathbf{x}-\mathbf{y}\|^{2i} + \epsilon_n\end{aligned}$$

we arrive at the same form as in (7) except for the coefficients. Therefore, the function rank related to error ϵ_n remains $\binom{n+d+2}{d+2}$, and we have proved our result.

□

3.2 Proof of Theorem 2.2

Proof. The proof follows the same steps as that in Theorem 2.1; we only need to establish that the error term in the Chebyshev expansion is bounded by $\frac{2D^{2q}V_q}{\pi q[(n-q)]^q}$. Consider (1). Because $f^{(q)}$ is piecewise continuous with its total variation on $[0, D^2]$ bounded by V_q, $\tilde{f}^{(q)}$ is piecewise continuous on $[0, 2]$, with its total variation on $[0, 2]$ bounded as follows

$$V\left(\frac{\mathrm{d}^q\tilde{f}}{\mathrm{d}x^q}\right) = V\left(\frac{D^{2q}}{2^q}\frac{\mathrm{d}^q f}{\mathrm{d}^q x^q}\right) = \frac{D^{2q}}{2^q}V\left(\frac{\mathrm{d}^q f}{\mathrm{d}^q x^q}\right) \le \frac{D^{2q}}{2^q}V_q$$

Therefore, for $n > q$, the convergence rate of the order-n Chebyshev expansion of $\tilde{f}(r)$ to $\tilde{f}(r)$ is given by the result from Chebyshev approximation theory [32]:

$$|\epsilon_n| \le \frac{2V_q(\tilde{f})}{\pi q(n-q)^q} \le \frac{2D^{2q}V_q}{\pi q(2(n-q))^q}$$

The rest of the proof is identical to that of Theorem 2.1. □

3.3 Proof of Theorem 2.3

We first introduce a lemma regarding the rank of complex functions.

Lemma 3.1. *If a real-valued function $\mathcal{K}$ can be approximated by two sequences of complex-valued functions, i.e.,*

$$\left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R_c}\Psi_i(\mathbf{x})\Phi_i(\mathbf{y})\right| \le \epsilon, \quad \mathbf{x}\in\Omega_X, \mathbf{y}\in\Omega_Y$$

where $\{\Psi_i(\mathbf{x})\}_{i=1}^{R_c}$ and $\{\Phi_i(\mathbf{y})\}_{i=1}^{R_c}$ are complex-valued functions, then there exist two sequences of real-valued functions, $\{g_i(\mathbf{x})\}_{i=1}^{R}$ and $\{h_i(\mathbf{y})\}_{i=1}^{R}$, such that for $R = 2R_c$,

$$\left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R} g_i(\mathbf{x})h_i(\mathbf{y})\right| \le \epsilon, \quad \mathbf{x}\in\Omega_X, \mathbf{y}\in\Omega_Y$$

Proof. Let $\mathrm{Re}\,(\cdot)$ and $\mathrm{Im}\,(\cdot)$ denote the real and imaginary part of a complex value, respectively. For each term, $\Psi_i(\mathbf{x})\Phi_i(\mathbf{y})$, we rewrite it as

$$\begin{aligned}\Psi_i(\mathbf{x})\Phi_i(\mathbf{y}) = &\left(\mathrm{Re}\left(\Psi_i(\mathbf{x})\right)\mathrm{Re}\left(\Phi_i(\mathbf{y})\right) - \mathrm{Im}\left(\Psi_i(\mathbf{x})\right)\mathrm{Im}\left(\Phi_i(\mathbf{y})\right)\right)\\ &+ \imath\left(\mathrm{Re}\left(\Psi_i(\mathbf{x})\right)\mathrm{Im}\left(\Phi_i(\mathbf{y})\right) + \mathrm{Im}\left(\Psi_i(\mathbf{x})\right)\mathrm{Re}\left(\Phi_i(\mathbf{y})\right)\right)\end{aligned} \tag{11}$$

We can then construct the sequences of real-valued functions as follows

$$\begin{cases} g_{2i-1}(\mathbf{x}) = \mathrm{Re}\left(\Psi_i(\mathbf{x})\right), g_{2i}(\mathbf{x}) = -\mathrm{Im}\left(\Psi_i(\mathbf{x})\right) \\ h_{2i-1}(\mathbf{y}) = \mathrm{Re}\left(\Phi_i(\mathbf{y})\right), h_{2i}(\mathbf{y}) = \mathrm{Im}\left(\Phi_i(\mathbf{y})\right) \end{cases}, \quad i = 1, 2, \ldots, R_c \tag{12}$$

The approximation error holds for the real-valued approximation:

$$\left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R} g_i(\mathbf{x})h_i(\mathbf{y})\right| \leq \left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R_c} \Psi_i(\mathbf{x})\Phi_i(\mathbf{y})\right| \leq \epsilon \tag{13}$$

□

We now start the proof for Theorem 2.3.

Proof. The proof consists of two main parts. The first part derives a separable form of $\mathcal{K}(\mathbf{x},\mathbf{y})$. This includes applying a Fourier expansion to the periodic function $p(z)$ followed by a Taylor expansion. The second part computes the approximation rank and the upper bound for the truncation error.

1. *Fourier expansion.* Let the Fourier expansion of p with error term ϵ_{M_f} be

$$p(z) = \sum_{n_1=-M_f}^{M_f} C_{n_1} \exp(\imath n_1 \omega z) + \epsilon_{M_f} \tag{14}$$

where $C_{n_1} = \frac{1}{\rho^2}\int_{-\rho^2/2}^{\rho^2/2} p(t)\exp(-\imath n_1 \omega t)dt$ and $\omega = \frac{2\pi}{\rho^2}$. The error term ϵ_{M_f} has a fast decay because $p(z)$ is a smooth periodic function. Let $\mathbf{x}^*$ and $\mathbf{y}^*$ be the center of point sets Ω_X and Ω_Y, respectively. We rewrite z in terms of $\boldsymbol{\rho}_\mathbf{x} = \mathbf{x} - \mathbf{x}^*$, $\boldsymbol{\rho}_\mathbf{y} = \mathbf{y} - \mathbf{y}^*$ and $\boldsymbol{\rho} = \mathbf{x}^* - \mathbf{y}^*$:

$$z = \|\mathbf{x}-\mathbf{y}\|^2 = \|\boldsymbol{\rho}_\mathbf{x}\|^2 + \|\boldsymbol{\rho}_\mathbf{y}\|^2 + \|\boldsymbol{\rho}\|^2 - 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y} + 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho} - 2\boldsymbol{\rho}_\mathbf{y}^T\boldsymbol{\rho} \tag{15}$$

Substituting (15) into $\exp(\imath n_1 \omega z)$, we obtain

$$\exp(\imath n_1\omega z) = \underbrace{\exp(\imath n_1\omega(\|\boldsymbol{\rho}_\mathbf{x}+\boldsymbol{\rho}\|^2))}_{\text{function of } \mathbf{x} \text{ only}}\underbrace{\exp(\imath n_1\omega(\|\boldsymbol{\rho}_\mathbf{y}\|^2 - 2\boldsymbol{\rho}_\mathbf{y}^T\boldsymbol{\rho}))}_{\text{function of } \mathbf{y} \text{ only}}\underbrace{\exp(-\imath n_1\omega 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y})}_{\text{function of } \mathbf{x} \text{ and } \mathbf{y}} \tag{16}$$

2. *Taylor expansion.* To further separate the terms in (16) involving both $\mathbf{x}$ and $\mathbf{y}$, we apply a Taylor expansion to $\exp(-\imath n_1\omega 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y})$ with error term $\epsilon_{M_t}^{n_1}$:

$$\exp\left(-\imath n_1\omega 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y}\right) = \sum_{n_2=0}^{M_t} \frac{(-i)^{n_2}}{n_2!}\left(2n_1\omega\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y}\right)^{n_2} + \epsilon_{M_t}^{n_1} \tag{17}$$

Adopting the multi-index notation,

$$\left(\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y}\right)^m = \sum_{|\boldsymbol{\alpha}|=m}\binom{m}{\boldsymbol{\alpha}}\boldsymbol{\rho}_\mathbf{x}^{\boldsymbol{\alpha}}\boldsymbol{\rho}_\mathbf{y}^{\boldsymbol{\alpha}} \tag{18}$$

Substituting (18) into (17),

$$\exp\left(-\imath n_1\omega 2\boldsymbol{\rho}_\mathbf{x}^T\boldsymbol{\rho}_\mathbf{y}\right) = \sum_{n_2=0}^{M_t}\frac{(-i\ 2n_1\omega)^{n_2}}{n_2!}\sum_{|\boldsymbol{\alpha}|=n_2}\binom{n_2}{\boldsymbol{\alpha}}\boldsymbol{\rho}_\mathbf{x}^{\boldsymbol{\alpha}}\boldsymbol{\rho}_\mathbf{y}^{\boldsymbol{\alpha}} + \epsilon_{M_t}^{n_1} \tag{19}$$

3. *Separable form derivation.* Combining (19), (16) and (14) and rearranging the terms, we obtain

$$\begin{aligned} p(z) = &\sum_{n_1=-M_f}^{M_f} \sum_{n_2=0}^{\lfloor M_t/2 \rfloor} \sum_{|\boldsymbol{\alpha}|=2n_2} A(n_1, n_2, \boldsymbol{\alpha}) h(n_1, \boldsymbol{\alpha}, \mathbf{x}) g(n_1, \boldsymbol{\alpha}, \mathbf{y}) \\ &+ \sum_{n_1=-M_f}^{M_f} \sum_{n_2=0}^{\lfloor M_t/2 \rfloor} \sum_{|\boldsymbol{\alpha}|=2n_2+1} B(n_1, n_2, \boldsymbol{\alpha}) h(n_1, \boldsymbol{\alpha}, \mathbf{x}) g(n_1, \boldsymbol{\alpha}, \mathbf{y}) + \epsilon \end{aligned} \tag{20}$$

where

$$\begin{aligned} A(n_1, n_2, \boldsymbol{\alpha}, \mathbf{x}) &= C_{n_1} \frac{(-1)^{n_2} (2n_1\omega)^{2n_2}}{(2n_2)!} \binom{2n_2}{\boldsymbol{\alpha}} \\ B(n_1, n_2, \boldsymbol{\alpha}, \mathbf{y}) &= -\imath\, C_{n_1} \frac{(-1)^{n_2} (2n_1\omega)^{2n_2+1}}{(2n_2+1)!} \binom{2n_2+1}{\boldsymbol{\alpha}} \end{aligned} \tag{21}$$

are independent of the point distribution in Ω_X and Ω_Y,

$$\begin{aligned} h(n_1, \boldsymbol{\alpha}, \mathbf{x}) &= \exp\left(\imath n_1 \omega \|\boldsymbol{\rho}_{\mathbf{x}} + \boldsymbol{\rho}\|^2\right) \boldsymbol{\rho}_{\mathbf{x}}^{\boldsymbol{\alpha}} \\ g(n_1, \boldsymbol{\alpha}, \mathbf{y}) &= \exp\left(\imath n_1 \omega (\|\boldsymbol{\rho}_{\mathbf{y}}\|^2 - 2\boldsymbol{\rho}_{\mathbf{y}}^T \boldsymbol{\rho})\right) \boldsymbol{\rho}_{\mathbf{y}}^{\boldsymbol{\alpha}} \end{aligned} \tag{22}$$

are functions of $\mathbf{x}$ only and $\mathbf{y}$ only, respectively, and

$$\epsilon = \sum_{n_1=-M_f}^{M_f} C_{n_1} \exp\left(\imath n_1 \omega \|\boldsymbol{\rho}_{\mathbf{x}} + \boldsymbol{\rho}\|^2\right) \exp\left(\imath n_1 \omega (\|\boldsymbol{\rho}_{\mathbf{y}}\|^2 - 2\boldsymbol{\rho}_{\mathbf{y}}^T \boldsymbol{\rho})\right) \epsilon_{M_t}^{n_1} + \epsilon_{M_f}$$

4. *Rank computation.* Equation (20) is a separable form of $\mathcal{K}(\mathbf{x}, \mathbf{y})$ with rank

$$R_c = 2M_f \sum_{k=0}^{\lfloor M_t/2 \rfloor} \left(\binom{2k+d-1}{d-1} + \binom{2k+1+d-1}{d-1} \right) \leq 2M_f \binom{M_t+d}{d} \tag{23}$$

By Lemma 3.1, the kernel function can be approximated by two sequences of real-valued functions $\{g_i\}_{i=1}^R$ and $\{h_i\}_{i=1}^R$ with rank

$$R = 2R_c \leq 4M_f \binom{M_t+d}{d}$$

When M_f and M_t are fixed and $d \to \infty$, the rank grows as $O(d^{M_t})$.

5. *Error analysis.* The total error consists of that from the Fourier expansion, the Taylor expansion, and the periodical extension. According to Theorem 2 in [14], the Fourier expansion error ϵ_{M_f} is bounded as follows

$$|\epsilon_{M_f}| \leq \frac{V_q(p)}{\pi q (\omega M_f)^q} = \frac{V_q \rho^{2q}}{\pi q (2\pi M_f)^q} \tag{24}$$

where $V_q(p)$ is the upper bound for the total variation of $p^{(q)}$ over one period. Using the Lagrange form of the remainder, the Taylor expansion error $\epsilon_{M_t}^{n_1}$ is bounded as

$$|\epsilon_{M_t}^{n_1}| \leq \left| \frac{1}{(M_t+1)!} \left(2n_1 \omega \boldsymbol{\rho}_{\mathbf{x}}^T \boldsymbol{\rho}_{\mathbf{y}}\right)^{(M_t+1)} \right| \leq \frac{(2|n_1|\omega)^{(M_t+1)}}{(M_t+1)!} (D_{\mathbf{x}} D_{\mathbf{y}})^{(M_t+1)} \tag{25}$$

Therefore, the total error ϵ in (20) can be bounded as

$$\begin{aligned} \epsilon &\leq \frac{\|p\|}{\rho} \frac{(2\omega)^{(M_t+1)}}{(M_t+1)!} (D_{\mathbf{x}} D_{\mathbf{y}})^{(M_t+1)} \sqrt{2 \sum_{n_1=1}^{M_f} n_1^{2(M_t+1)}} + \epsilon_{M_f} \\ &\leq \frac{\|p\|}{e\rho} \left(\frac{4\pi e (M_f+1) D_{\mathbf{x}} D_{\mathbf{y}}}{\rho^2 (M_t+1)} \right)^{(M_t+1)} \sqrt{\frac{M_f+1}{(M_t+1)}} + \frac{V_q \rho^{2q}}{\pi q (2\pi M_f)^q} \end{aligned} \tag{26}$$

where $\|\cdot\| = \|\cdot\|_{L_2[0,\rho^2]}$. The first inequality comes from the Cauchy-Schwarz inequality and the Parseval's theorem, and the last inequality comes from an upper bound for the discrete sum and the inequality $e(\frac{n}{e})^n \leq n!$.

Therefore, we have found two sequences of real-valued functions that approximate the kernel function with rank

$$R = (4M_f + 2)\binom{M_t + d}{d}$$

and approximation error

$$|\epsilon| \leq \frac{\|p\|}{e\rho}\left(\frac{4\pi e D_{\mathbf{x}} D_{\mathbf{y}}}{\rho^2}\right)^{M_t+1}\left(\frac{M_f+1}{M_t+1}\right)^{M_t+3/2} + \frac{V_q \rho^{2q}}{\pi q (2\pi M_f)^q} + \epsilon_P \tag{27}$$

where ϵ_P is the periodic-extension error of f and $\sup_{x\in[0,D^2]} |f(x) - p(x)| \leq \epsilon_P$.

□

4 Optimality

The analytic error bound in Theorem 2.1 is optimal in the sense that will be defined below. Schaback [37] provided a lower bound on the minimal number of interpolation points needed for a linear operator to reach a required accuracy. Specifically, let $L_n(f) = \sum_{j=1}^n f(x_j)a_j$ be an interpolation operator in dimension d with $a_j \in C([-1,1]^d)$, let $\|f\|_k^2 := \sum_{l\in\mathbb{N}_0}(1+l^2)^k c_l^2[f]$ with $c_l[f]$ denoting the Fourier coefficient of f, and let the ϵ-complexity $n(\epsilon, d)$ of $L_n(f)$ be defined as

$$n(\epsilon, d) = \min\{n : \exists\ L_n, \|I_d - L_n\| \leq \epsilon\} \tag{28}$$

where $\|I_d - L_n\| = \sup\{\|f - L_n(f)\|_{L_2} \mid \|f\|_k \leq 1\}$. Then, $n(\epsilon, d)$ satisfies:

$$n(\epsilon, d) \geq c_\epsilon d^{c\log(\epsilon^{-1})} \tag{29}$$

The lower bound in (29) is the minimal rank needed to achieve approximation error ϵ, considering that the number of interpolation points is equivalent to the rank of a function.

In the following, we show that the rank in Theorem 2.1 reaches the lower bound in (29) asymptotically. We assume that the volume of data space is bounded by 1, *i.e.*, $\mathrm{Vol}(\Omega) \leq 1$, and that $\|\mathcal{K}\|_k \leq 1$. The latter is implied by the smoothness assumption in Theorem 2.1. Under these conditions,

$$\sup\left\{\|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^R h_i(\mathbf{x})g_i(\mathbf{y})\|_{L_2} \mid \|\mathcal{K}\|_k \leq 1\right\} \leq \frac{2CD^2}{\rho^2 - D^2}\left(\frac{\rho^2}{D^2}\right)^{-n} = \epsilon \tag{30}$$

For simplicity, we denote $c_1 = \frac{2CD^2}{\rho^2 - D^2}$ and $c_2 = \frac{\rho^2}{D^2}$, where c_1 and c_2 are constants for a fixed kernel function. Then, rewriting the equality in (30), we obtain

$$n = \frac{-\log\epsilon + \log c_1}{\log c_2} \tag{31}$$

Finally, we relate rank R to error ϵ and dimension d. When $d \to \infty$,

$$R = \binom{n+d+2}{d+2} \approx \frac{d^n}{n!} = c_\epsilon d^{\frac{-\log\epsilon + \log c_1}{\log c_2}} \tag{32}$$

where $c_\epsilon = \frac{1}{n!}$, and "$\approx$" denotes the strong equivalence of sequences, *i.e.*, $v_n \approx w_n$ iff $\lim_{n\to\infty}\frac{v_n}{w_n} = 1$. In (32), "$\approx$" applies when n is fixed and $d \to \infty$. Therefore, (32) is the same as (29) asymptotically. We conclude that our result for the rank achieves optimal dependence on dimension d and error ϵ.

Related work. There has been work achieved similar results using the sparse grid. Barthelmann [3] considered a polynomial interpolation on a sparse grid, and showed that such interpolation

can achieve an acceptable accuracy with the number of interpolation points growing polynomially on data dimension. Specifically, consider function f in the function class

$$F_d^k = \{f : [-1,1]^d \mapsto \mathbb{R} \mid D^{\boldsymbol{\alpha}} f \text{ continuous if } \alpha_i \leq k \text{ for all } i\} \tag{33}$$

If we apply an interpolation operator $A(k+d,d)$, the Smolyak formula [31], on f, then the interpolation error is bounded as follows

$$\|f - A(k+d,d)(f)\|_{L_2} \leq c_{d,k} N^{-k} (\log N)^{(k+1)(d-1)} \|f\|_k \tag{34}$$

where

$$N = N(k+d,d) = \sum_{s=0}^{\min(k,d)} \binom{k}{s} \binom{k+d-s}{k} \leq \binom{2k+1+d}{d} \tag{35}$$

is the number of interpolation points for $A(k+d,d)$ (see [24]), and

$$A(k+d,d) = \sum_{k+1 \leq |\mathbf{i}| \leq k+d} (-1)^{k+d-|\mathbf{i}|} \binom{d-1}{k+d-|\mathbf{i}|} (U_1^{i_1} \otimes \cdots \otimes U_d^{i_d}) \tag{36}$$

where $\mathbf{i} \in \mathbb{N}^d$ is a multi-index, and $\otimes$ denotes the tensor product operator.

Consider $N(k+d,d)$. When k is fixed and $d \to \infty$, $N(k+d,d) \approx \frac{2^k}{k!} d^k$, suggesting that the rank grows polynomially with d for large d, *i.e.*, $R = O(d^k)$. This polynomial dependence is consistent with our result in (7) that the rank is approximately $O(d^n)$. The two notations k and n are both related to the smoothness of a function: k is the degree of polynomials and n is the degree of the Chebyshev expansion. Further, both $A(k+d,d)$ and Equation (7) are exact for all polynomials of the same degree.

5 Numerical Experiments

In this section, we experimentally verify two main results suggested by our theorems: the polynomial growth of the numerical rank with the data dimension, and the effect of data radius on the approximation error. Considering that the error bounds in our theorems correspond to the worst-case error, we report the highest rank from varying data distributions to be the numerical rank.

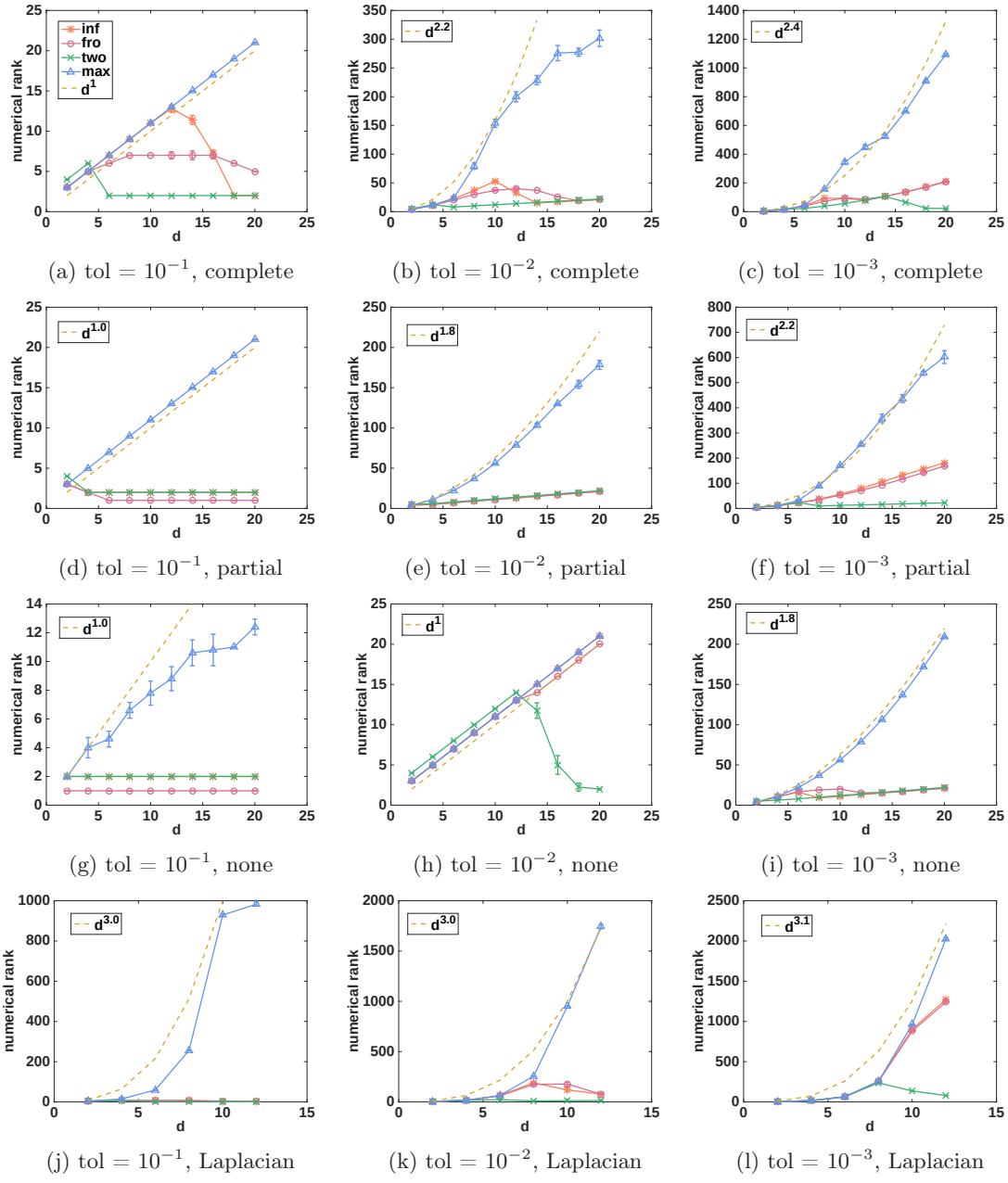
One difficulty of generating data which is representative of the worst case lies in the concentration of measures, that is, a peak will occur in the pairwise-distance distribution as the data dimension increases, leading to all the elements in the matrix being nearly identical. To mitigate this issue and achieve a larger variance in the pairwise distances, we project those randomly generated points to the boundary. The detailed procedure is as follows. First, we uniformly generate data in the cube $[a,b]^d$; second, we randomly sample a fraction p of points and set k randomly selected dimensions of each sampled data point to be a or b. For our experiments, we take $p \in \{0.1, 0.4, 0.7, 1.0\}$ and $k \in \{0, \cdots, d\}$.

The numerical rank R associated with tolerance *tol* is defined as

$$R = \min\left\{r \mid \|K - U_r S_r V_r^T\| \leq tol\, \|K\|\right\}$$

where U_r, S_r, V_r are factors from the singular value decomposition (SVD) of matrix K. Depending on the choice of the norm, the value of R will vary; however, our main focus is on the max norm, which is consistent with the function infinity norm in the theorems. Theoretically, the max error does not decrease monotonically with the rank; however, we find that for the RBF kernel matrices, the max error decreases in general with the rank, except for certain small short-lived increases.

Throughout our experiments, we fix the number of points at 3,000. The kernels used are the Gaussian kernel $\exp(-\|\mathbf{x}-\mathbf{y}\|^2/h^2)$ and the Laplacian kernel $\exp(-\|\mathbf{x}-\mathbf{y}\|/h)$ with $h = \max_{\mathbf{x}_i,\mathbf{x}_j \in \Omega} \|\mathbf{x}_i - \mathbf{x}_j\|$. For each setting of dimension and tolerance, we report the mean and standard deviation of the numerical rank out of 5 independent runs. Figure 2 shows the numerical rank as a function of data dimension subject to a fixed tolerance on 3 different data overlapping scenarios: source and target data both in $[0,1]^d$; source data in $[0,2/3]^d$ and target data in $[1/3,1]^d$; and source data in $[0,1/2]^d$ and target data in $[1/2,1]^d$. By design, the ratio between the data radius $D_{\mathbf{x}}$ (or $D_{\mathbf{y}}$) of these scenarios is roughly $6:4:3$ and they are shown from top to bottom for each fixed tolerance in Figure 2.



(a) tol $= 10^{-1}$, complete (b) tol $= 10^{-2}$, complete (c) tol $= 10^{-3}$, complete

(d) tol $= 10^{-1}$, partial (e) tol $= 10^{-2}$, partial (f) tol $= 10^{-3}$, partial

(g) tol $= 10^{-1}$, none (h) tol $= 10^{-2}$, none (i) tol $= 10^{-3}$, none

(j) tol $= 10^{-1}$, Laplacian (k) tol $= 10^{-2}$, Laplacian (l) tol $= 10^{-3}$, Laplacian

Figure 2: Numerical rank vs. data dimension. The data size was fixed at 3,000. The rank is defined as $\min\{r \mid \|K - U_r S_r V_r^T\| \leq tol \, \|K\|\}$ with the choice of norm listed in the legend. "inf" is infinity norm; "fro" is Frobenius norm; "two" is two norm; and "max" is max norm. Subplots (a) to (i) used the Gaussian kernel with data completely overlapped for (a) to (c), partially overlapped for (d) to (f), and not overlapped for (g) to (i). Subplots (j) to (l) used the Laplacian kernel with data completely overlapped.

One observation on the growth of rank verifies that for a fixed degree k of the Chebyshev expansion, the rank grows in the order of $O(d^k)$ with dimension d. In our experiments, we increase k by decreasing the approximation tolerance, according to the relation between k and error ϵ in Theorem 2.1. We observe results consistent with the order $O(d^k)$.

Another observation on the data radius and the error bound verifies that shrinking the radius of either point set decreases the error bound. Theorem 2.3 suggests that the radius of target points $D_{\mathbf{x}}$ and source points $D_{\mathbf{y}}$ affect the error in the form of $\left(\frac{D_{\mathbf{x}} D_{\mathbf{y}}}{\rho^2}\right)^{M_t+1}$ where $\frac{D_{\mathbf{x}}}{\rho} < 1$ and $\frac{D_{\mathbf{y}}}{\rho} < 1$. That is, to achieve a certain tolerance threshold, a smaller data radius $D_{\mathbf{x}}$ (or $D_{\mathbf{y}}$) allows M_t to be smaller. This is verified by our experimental results. Figure 2 shows that with a fixed tolerance, the numerical rank that is positively related to M_t, decreases with the decrease of the data radius from top to bottom.

The numerical rank reported, however, grows at a slower rate than the theoretical value. This is expected. Even though the data was generated to represent the worst case scenario, it is difficult to avoid the concentration of measures and achieve such worst case with a limited number of points and data distributions.

The rare occurrence of the worst-case data distribution provides part of the reason for the success of low-rank approximations for high-dimensional problems. The real-world data are often more structured and most of them in fact live in a lower-dimensional manifold. Consequently, the kernel matrices of practical interest are much simpler than the striking number of points n and large data dimension d suggest.

6 Group Pattern of Singular Values

In this section, we reveal and explain a group pattern in the singular values of kernel matrices generated by RBFs. Specifically, the singular values form groups by their magnitudes, with the group cardinalities dependent on the data dimension and independent of the data size.

We explain this phenomenon based on Theorem 2.3 by an appropriate grouping of the number of terms in the function's separable form. For any RBF, consider the number of separable terms n in its function expansion:

$$n = \sum_{f=0}^{M_f} \sum_{t=0}^{M_t} n_t = \sum_{f=0}^{M_f} \sum_{t=0}^{M_t} \binom{t+d-1}{t-1} \tag{37}$$

The two summations correspond to the Fourier expansion of the kernel function, and the Taylor expansion of $\exp\left(\imath \boldsymbol{\rho}_{\mathbf{x}}^T \boldsymbol{\rho}_{\mathbf{y}}\right)$, respectively. n_t denotes the number of separable terms in $\left(\boldsymbol{\rho}_{\mathbf{x}}^T \boldsymbol{\rho}_{\mathbf{y}}\right)^t$. The observed group cardinalities are consistent with an appropriate grouping of the terms in (37). The order of these terms is governed by the error term in the truncation. One grouping example is

$$\underbrace{n_0, n_1, n_2}_{\text{1st term of Fourier expansion}} \;\Big|\; \underbrace{n_0, n_1}_{\text{2nd term of Fourier expansion}} \;\Big|\; \underbrace{n_3, n_4}_{\text{1st term of Fourier expansion}}$$

The cardinality of the 1st, 2nd and 3rd group is $n_0 + n_1 + n_2$, $n_0 + n_1$ and $n_3 + n_4$, respectively.

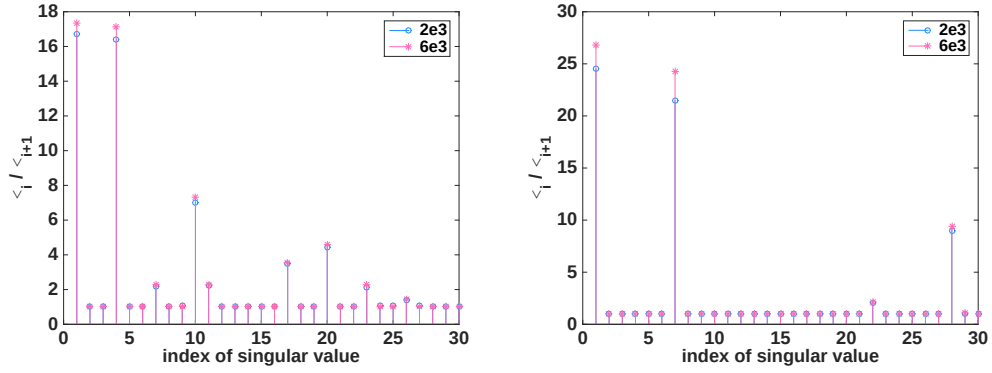
6.1 Experimental Verification

We experimentally verify the above claim and assume that the singular values are ordered such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$.

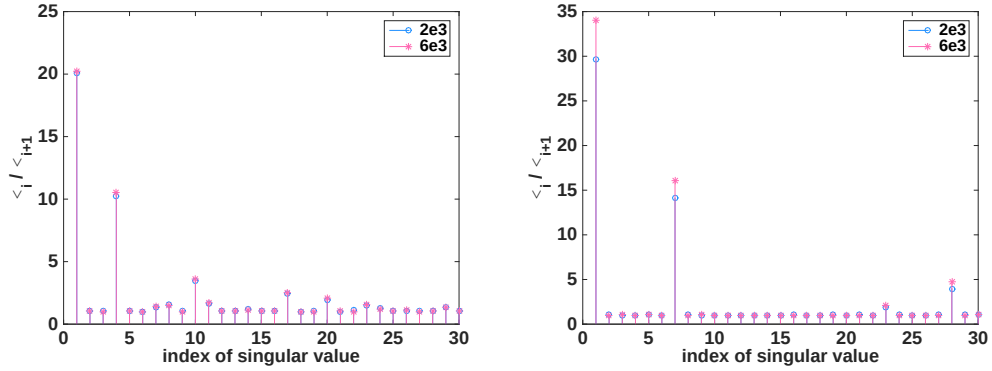
Figure 3 shows σ_i / σ_{i+1}, the ratio of the i-th largest singular value to the next smaller one. One observation is that both the group cardinalities indicated by the differences between the high-ratio indices, and the singular value decay amount indicated by the magnitudes of those ratios, are independent of the data size. This observation suggests that the numerical rank is independent of the data size.

We study the group cardinality in detail. Consider Figure 3a. The indices with ratios above 4 are as follows with the ratio shown in parenthesis,

$$1\ (17.3),\ 4\ (17.1),\ 10\ (7.3),\ 20\ (4.5)$$



(a) $d = 3$, $\exp(-\|\mathbf{x}-\mathbf{y}\|^2/h^2)$ (b) $d = 6$, $\exp(-\|\mathbf{x}-\mathbf{y}\|^2/h^2)$



(c) $d = 3$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|^2/h^2}$ (d) $d = 6$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|^2/h^2}$

Figure 3: Singular value ratio σ_i/σ_{i+1} vs. index i. The singular values are ordered such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, and the legend represents the data size (matrix dimension). Subplot (a) and (b) used Gaussian kernel $\exp(-\|\mathbf{x}-\mathbf{y}\|^2/h^2)$ and subplot (c) and (d) used Cauchy kernel $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|^2/h^2}$.

The values of these indices correspond to a cumulative sum of the number of separable terms in the following Taylor expansion terms,

$$\underbrace{0th}_{1\text{ term}},\ \underbrace{1st}_{3\text{ terms}},\ \underbrace{2nd}_{6\text{ terms}},\ \underbrace{3rd}_{10\text{ terms}}$$

The above grouping suggests that the separable terms contributing to significant decays in singular values come from the polynomial approximation for the first-order Fourier term. We note that the higher-order Fourier terms contribute as well, but the accuracy gained from those terms is smaller than that from the first-order Fourier term.

The indices with ratios above 2 are

$$1\ (17.3),\ 4\ (17.1),\ 7\ (2.3),\ 10\ (7.3),\ 11\ (2.3),\ 17\ (3.5),\ 20\ (4.6)$$

These subtler gains in accuracy may come from the contributions of other higher-order expansion terms. One possible grouping is as follows, with the order of Fourier expansion and the order of Taylor expansion shown in order in parenthesis,

$$\underbrace{(1,0)}_{1\text{ term}},\ \underbrace{(2,0),(3,0),(4,0)}_{3\text{ terms}},\ \underbrace{(1,1)}_{3\text{ terms}},\ \underbrace{(2,1)}_{3\text{ terms}},\ \underbrace{(5,0)}_{1\text{ term}},\ \underbrace{(1,2)}_{6\text{ terms}},\ \underbrace{(3,1)}_{3\text{ terms}}$$

Applying a cumulative sum of the number of these terms yields the above indices.

Our explanation adopts the idea of the Fourier approach instead of the Chebyshev approach. The key reason is that the Fourier approach allows us to group separable terms into finer sets that contributes to subtler error decay. The Chebyshev approach considers $\|\mathbf{x}-\mathbf{y}\|^{2l}$ as a unit, which has $\binom{l+d+1}{d+1}$ separable terms; whereas the Fourier approach considers $\left(\mathbf{x}^T\mathbf{y}\right)^l$ as a unit, which only involves $\binom{l+d-1}{d-1}$ separable terms. This flexibility in grouping benefits from the Fourier expansion, which raises the term $\|\mathbf{x}-\mathbf{y}\|^2$ to the exponent of e so that we are able to directly expand the cross term $\exp(\mathbf{x}^T\mathbf{y})$.

A Fourier expansion is unnecessary if the cross term involving $\mathbf{x}^T\mathbf{y}$ is naturally separated from $\|\mathbf{x}-\mathbf{y}\|^2$. One example is the Gaussian kernel $\exp(-\|\mathbf{x}-\mathbf{y}\|^2) = \exp(-\|\mathbf{x}\|^2)\exp(-\|\mathbf{y}\|^2)\exp(2\mathbf{x}^T\mathbf{y})$. This in fact explains our observations for Gaussian kernels, that the threshold ranks are represented by the number of separable terms in $\sum_l \left(\mathbf{x}^T\mathbf{y}\right)^l$.

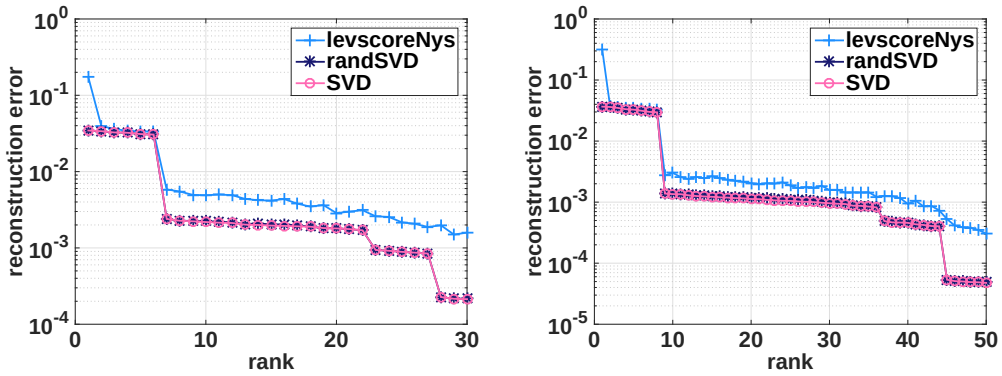
6.2 Practical Guidance

The group pattern in the singular values theoretically explains many phenomenons in practice. One common example is the threshold ranks in matrix approximations. In other words, we often need to increase the rank above a certain threshold to observe a further decay in the matrix approximation error. We can take advantage of the group pattern when applying low-rank algorithms. In practice, most low-rank approximation algorithms take input as a desired rank; however, it is unclear what rank is reasonable. The group cardinalities provide candidate ranks for the inputs of those algorithms.

We examine the effectiveness of our guidance on two popular RBF kernel matrices. We expect significant decay in the reconstruction error around rank $R = \binom{n+d}{d}$, the number of separate terms in the n-th order Taylor expansion of $\exp(h(\mathbf{x})^T g(\mathbf{y}))$. For Gaussian kernel, $h(\mathbf{x}) = g(\mathbf{x}) = \mathbf{x}$; for Laplacian kernel and other inseparable kernels, $h(\mathbf{x}) = \boldsymbol{\rho}_\mathbf{x}, g(\mathbf{y}) = \boldsymbol{\rho}_\mathbf{y}$. Figure 4 shows the reconstruction error as a function of the rank. For leverage-score Nyström method, we oversample 30 and 60 columns for $d = 6$ and $d = 8$, respectively, and report the mean of reconstruction error out of 5 independent runs. For all the methods, a significant decay in the error occurs at rank 1, 7, and 28 for $d = 6$, and at rank 1, 9 and 45 for $d = 8$. The trend in the decay pattern meets our expectation, except for several subtle perturbations that may be caused by the data distribution and contributions from other expansion terms.

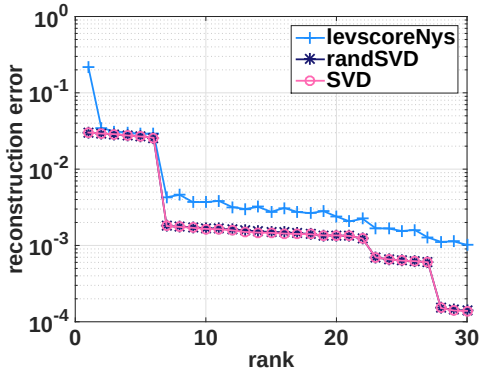
7 Conclusions

Approximation algorithms that take advantage of the low-rank structures of a matrix or certain submatrices are widely applied to RBF kernel matrices. The success of such algorithms, especially for high-dimensional datasets, has motivated us to study how the rank grows with the data dimension. In this paper, we provided three key results on the rank and the approximation error.

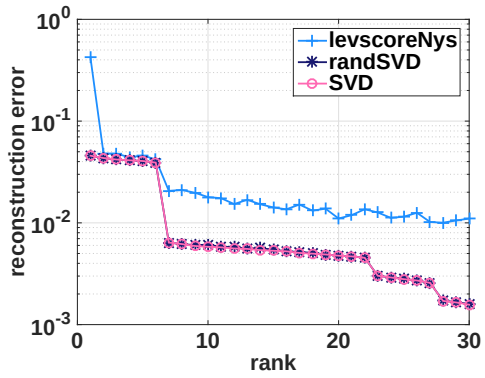


(a) $d = 6$, $\exp(-\|\mathbf{x} - \mathbf{y}\|^2/h^2)$

(b) $d = 8$, $\exp(-\|\mathbf{x} - \mathbf{y}\|^2/h^2)$



(c) $d = 6$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|^2/h^2}$



(d) $d = 6$, $\exp(-\|\mathbf{x} - \mathbf{y}\|/h)$

Figure 4: Reconstruction error vs. approximation rank. The legend represents low-rank algorithms, "levscoreNys" is the leverage-score Nyström method, "randSVD" is the randomized SVD with iteration parameter to be 2, and "SVD" is the exact SVD. The bandwidth parameter h was set to be the maximum pairwise distance. A significant decay in the error occurs at rank $= \binom{n+d}{d}$ $(n = 1, 2, 3)$ for all experiments.

First, the rank of an RBF grows polynomially, in the worst case, with data dimension d as $d \to \infty$. The exponential growth for multivariate functions from a classical analysis is absent for RBFs. Empirical studies show that a much slower growth than polynomial is even possible in some cases; part of the reason is the concentration of measures and the structures in the data of interest.

Second, reducing the radius of either the source or target point set decreases the approximation error, assuming a fixed bandwidth parameter.

Third, we observed groups in the singular values of RBF kernel matrices. We explained this group pattern by our analytic expansion of the kernel function. Specifically, the number of singular values of the same magnitude can be computed by an appropriate grouping of the separable terms of the function. Very commonly, the cardinality of the i-th group is $\binom{i+d-1}{d-1}$, the number of separable terms in the i-th order term in the Taylor expansion of $\exp(\mathbf{x}^T\mathbf{y})$.

References

[1] Keith Ball. Eigenvalues of Euclidean distance matrices. *J. Approx. Theory*, 68(1):74–82, jan 1992.

[2] Keith Ball, Natarajan Sivakumar, and Joseph D. Ward. On the sensitivity of radial basis interpolation to minimal data separation distance. *Constr. Approx.*, 8(4):401–426, dec 1992.

[3] Volker Barthelmann, Erich Novak, and Klaus Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000.

[4] John P. Boyd. A comparison of numerical algorithms for Fourier extension of the first, second, and third kinds. *J. Comput. Phys.*, 178(1):118–160, may 2002.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, sep 1995.

[6] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines: And other kernel-based learning methods.* Cambridge University Press, New York, NY, USA, 2000.

[7] Philip J. Davis and Philip. Rabinowitz. *Methods of numerical integration.* Dover Publication Inc., 2nd edition, 2007.

[8] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, dec 2012.

[9] Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, dec 2005.

[10] Noureddine El Karoui. The spectrum of kernel random matrices. *Ann. Stat.*, 38(1):1–50, feb 2010.

[11] Bengt Fornberg, Elisabeth Larsson, and Natasha Flyer. Stable computations with Gaussian radial basis functions. *SIAM J. Sci. Comput.*, 33(2):869–892, jan 2011.

[12] Bengt Fornberg and Julia Zuev. The Runge phenomenon and spatially variable shape parameters in RBF interpolation. *Comput. Math. with Appl.*, 54(3):379–398, 2007.

[13] Thomas Gerstner and Michael Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3/4):209–232, 1998.

[14] Charles R. Giardina and Paul M. Chirlian. Bounds on the truncation error of periodic signals. *IEEE Trans. Circuit Theory*, 19(2):206–207, 1972.

[15] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):3977–4041, jan 2016.

[16] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 4th edition, 2013.

[17] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, jan 2011.

[18] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Stat.*, 36(3):1171–1220, jun 2008.

[19] Peter K. Kitanidis. Compressed state Kalman filter for large systems. *Adv. Water Resour.*, 76:120–126, feb 2015.

[20] Judith Yue Li, Sivaram Ambikasaran, Eric F. Darve, and Peter K. Kitanidis. A Kalman filter powered by H2-matrices for quasi-continuous data assimilation problems. *Water Resour. Res.*, 50(5):3734–3749, may 2014.

[21] Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proc. Natl. Acad. Sci. U. S. A.*, 104(51):20167–72, dec 2007.

[22] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends® Mach. Learn.*, 3(2):123–224, 2011.

[23] Francis J. Narcowich and Joseph D. Ward. Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices. *J. Approx. Theory*, 69(1):84–109, apr 1992.

[24] Erich Novak and Klaus Ritter. Simple cubature formulas with high polynomial exactness. *Constr. Approx.*, 15(4):499–522, oct 1999.

[25] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2005.

[26] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci.*, pages 143–152. IEEE, 2006.

[27] Robert Schaback. Lower bounds for norms of inverses of interpolation matrices for radial basis functions. *J. Approx. Theory*, 79(2):287–306, nov 1994.

[28] Robert Schaback. Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.*, 3(3):251–264, apr 1995.

[29] Bernhard. SchoÌĹlkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[30] Si Si, Cho-Jui Hsieh, and Inderjit Dhillon. Memory efficient kernel approximation. In Eric P Xing and Tony Jebara, editors, *Proc. 31st Int. Conf. Mach. Learn.*, volume 32 of *Proceedings of Machine Learning Research*, pages 701–709, Bejing, China, 2014. PMLR.

[31] Sergey A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Sov. Math. Dokl.*, volume 4, pages 240–243, 1963.

[32] Lloyd N. Trefethen. *Approximation theory and approximation practice*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2013.

[33] Lloyd N. Trefethen. Cubature, approximation, and isotropy in the hypercube. *SIAM Rev. to Appear*, 2016.

[34] Vladimir Naumovich Vapnik. *Statistical learning theory*. Wiley, 1998.

[35] Ruoxi Wang, Yingzhou Li, Michael W. Mahoney, and Eric Darve. Structured block basis factorization for scalable kernel matrix evaluation. Technical report, 2015.

[36] Andrew J. Wathen and Shengxin Zhu. On spectral distribution of kernel matrices related to radial basis functions. *Numer. Algorithms*, 70(4):709–726, dec 2015.

[37] Henryk Woźniakowski. Tractability and strong tractability of linear multivariate problems. *J. Complex.*, 10(1):96–128, mar 1994.

[38] Kai Zhang and James T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Trans. Neural Networks*, 21(10):1576–1587, 2010.