On the numerical rank of radial basis function kernels in high dimensions

Ruoxi Wang * Yingzhou Li † Eric Darve ‡

April 2, 2018

Abstract

Low-rank approximations are popular methods to reduce the high computational cost of algorithms involving large-scale kernel matrices. 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 function rank in this paper, which is an upper bound of the matrix rank. The concept of function rank will be introduced to define the number of terms in the minimal separate form of a kernel function. We consider radial basis functions (RBF) in particular, and approximate the RBF kernel with a low-rank representation that is a finite sum of separate products, and provide explicit upper bounds on the function rank and the L_∞ error for such approximation. Our three main results are as follows. First, for a fixed precision, the function rank of RBFs, in the worst case, grows polynomially with the data dimension. Second, precise error bounds for the low-rank approximations in the L_∞ norm are derived in terms of the function smoothness and the domain diameters. And last, a group pattern in the magnitude of singular values for RBF kernel matrices is observed and analyzed, and is explained by a grouping of the expansion terms in the kernel's low-rank representation. Empirical results verify 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 attentions in a variety of fields including machine learning, inverse problems, graph theory and PDEs [33, 39, 27, 19, 21, 20, 32]. Examples of interest in machine learning include kernel ridge regression [38, 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 associated 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 [39]. The singular value decomposition (SVD) [17] is optimal but has an undesirable cubic complexity. Many methods [23, 18, 22, 28, 8, 16, 7, 42] have been proposed to accelerate the low-rank constructions by sacrificing the accuracy to an acceptable extend. The success of these low-rank algorithms hinges on a large spectrum gap, or a fast decay of the spectrum of the matrix itself or of its submatrices. However, to our knowledge, there is no theoretical guarantee that this is always the case.

Nonetheless, algebraic low-rank techniques work effectively in many cases where the data dimension ranges from moderate to high, motivating us to look into the growth rate of matrix ranks in high dimensions. A precise analysis of the matrix rank is nontrivial, and we turn to analyzing its upper bound, that is, the function rank of kernels that will be defined in what follows. The *function rank* is the number of terms in the minimal separate form of $\mathcal{K}(\mathbf{x}, \mathbf{y})$, when $\mathcal{K}$ is approximated by a finite sum of separate products $h_i(\mathbf{x})g_i(\mathbf{y})$ where h_i and g_i are real-valued functions.

*Institute for Computational and Mathematical Engineering, Stanford University, Email: ruoxi@stanford.edu
†Department of Mathematics, Duke University, Email: yingzhou.li@duke.edu
‡Department of Mechanical Engineering, Stanford University, Email: darve@stanford.edu

If the function rank does not grow exponentially with the data dimension, neither will the matrix rank.

If, however, we expand the function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ into a separate form in a classic way, then the number of terms will grow exponentially with the data dimension. The exponential growth is striking in the sense that even for a moderate dimension, a reasonable accuracy would be difficult to achieve. However, in practice, people have observed much lower matrix ranks. A plausible reason is that both the functions and the data of practical interest enjoy some special properties, which should be considered when carrying out the analysis.

The aim of this paper, therefore, is to analytically describe the relation between the function rank and the properties of the function and the data, including measures of function smoothness, the data dimension, and the domain diameter. Such relation has not been described before. We hope the conclusions of this paper on functions can provide some theoretical foundations for the practical success of low-rank matrix algorithms.

In this paper, we present three main results. First, we show that under common smoothness assumptions and up to some precision, the function rank of RBF kernels grows polynomially with increasing dimension d in the worst case. Second, we provide explicit L_∞ error bounds for the low-rank approximations of RBF kernel functions. And last, we explain the observed "decay-plateau" behavior of the singular values of smooth RBF kernel matrices.

1.1 Related Work

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

One line of research focuses on the spectrum of kernel matrices. There is a rich literature on the smallest eigenvalues, mainly concerning the matrix conditioning. Several papers [1, 25, 29, 30] provided lower bounds for the smallest eigenvalues, and Ball *et al.* discussed the upper bound in [1]. 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 $K(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / h^2)$, and showed that as data dimension $d \to \infty$, the spectral property is the same as that of the covariance matrix $\frac{1}{d} X X^T$. Wathen [40] 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 [13] but with no explanation. The same pattern also occurs in the coefficients of the orthogonal expansion in the RBF-QR method proposed in [11]. There have also been studies focusing on the "flat-limit" situation where $h \to \infty$ [9, 31, 12].

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 *et al.* [2] considered polynomial interpolation on a sparse grid [14]. Sparse grids are based on a high-dimensional multiscale basis and involve 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 the 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 function smoothness.

Trefethen [37] commented that most methods (including sparse grids) 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 that is of the form $\|\boldsymbol{\alpha}\|_2$ for a multi-index $\boldsymbol{\alpha}$ 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 $\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x} - \mathbf{y}\|_2^2)$. Such functions include but are not limited to the Gaussian kernel and the Cauchy kernel. We define the numerical *function rank* of a kernel $\mathcal{K}(\mathbf{x}, \mathbf{y})$ related to error ϵ, to which we will frequently refer.

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

where h_i's and g_i's are real functions on $\mathbb{R}^d$, and the separate form $\sum_{i=1}^{r} g_i(\mathbf{x})h_i(\mathbf{y})$ will be referred as a *low-rank kernel*, or a *low-rank representation* of rank at most r. Note that the rank definition concerns the function rank instead of the matrix rank.

Our two main results are as follows. First, we show that under common smoothness assumptions of RBFs and for a fixed precision, the function rank for RBF kernels is a polynomial function of the data dimension d. Specifically, the function rank $R = O(d^q)$ where q is related to the low-rank approximation error. Furthermore, Precise error bounds in their explicit form will be proved.

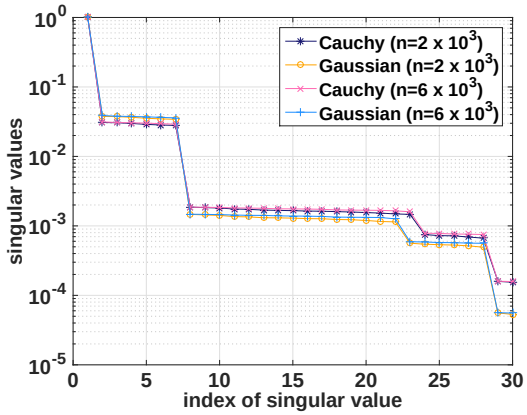


Figure 1: Group patterns in singular values. The singular values are normalized and ordered s.t. $1 = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. The data were randomly generated in dimension 6 with default random seed in MATLAB. The legend shows the data size and the kernel functions: Cauchy $(1/(1 + \|x - y\|_2^2))$ and Gaussian $(\exp(-\|x - y\|_2^2))$.

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 are dependent on the data dimension, but 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 concerning the function rank of the approximation of the RBF kernel function, and the L_∞ error bound of the approximations. Section 3 provides the theorem proofs. Section 4 shows that for a fixed precision, the polynomial growth of the derived rank cannot be improved. Section 5 verifies our theorems experimentally. Finally, in Section 6, we investigate and discuss the group pattern in the singular values for RBF kernel matrices.

2 Main theorems

In this section, we present theorems concerning the function rank and function and data properties. Each theorem approximates the RBF kernels in the L_∞ norm with low-rank kernels where the function rank and the error bound are given in explicit formulas. We briefly describe the theorems and then delve into further details.

The first four theorems consider kernels with two types of smoothness assumptions, and for each type, we present the deterministic result and the probabilistic result separately in two theorems. The probabilistic results take into account the concentration of measures for large data dimensions. The separate form is obtained by applying a Chebyshev expansion of $f(z)$ followed by a further expansion of $z = \|\mathbf{x} - \mathbf{y}\|_2^2$.

The key advantage of this approach is that the accuracy of the expansion only depends on $\|\mathbf{x} - \mathbf{y}\|_2^2$ and is independent of $(\mathbf{x}, \mathbf{y})$, which lie in a d-dimensional space. Assume we have expanded $f(z)$ to order n with accuracy ϵ. Then, we substitute $z = \|\mathbf{x} - \mathbf{y}\|_2^2$, expand the result, and re-arrange the terms to identify the number of distinct separate products of the form $h(\mathbf{x})g(\mathbf{y})$ in the final representation. This number becomes our upper bound on the function rank.

The theorems show that for a fixed precision, the function rank grows polynomially with data dimension d, and that the L_∞ error for low-rank approximations decreases with decreasing diameter of the domain that contains $\mathbf{x}$ and $\mathbf{y}$.

The last theorem considers kernels with finite smoothness assumptions. The separate form is obtained by applying a Fourier expansion of $f(z)$ followed by a Taylor expansion on each Fourier term. Additional to what the previous theorems suggest, the formulas for the error and the function rank capture more subtle relations of different parameters, and the theorem shows that the error decreases when either the diameter of the domain that contains $\mathbf{x}$ or that contains $\mathbf{y}$ decreases. Before presenting our theorems, we introduce some notations.

Notations. Let $\mathbf{E}(\cdot)$ and $\mathbf{Var}(\cdot)$ denote the expectation and variance, respectively. Let

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

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 and is referred as the transformed Bernstein ellipse. For instance, given an interval $[a, b]$, let $\phi(x)$ be a linear mapping from $[a, b]$ to $[-1, 1]$. And the *transformed Bernstein ellipse* for $[a, b]$ is defined to be $\phi^{-1}(E_{\rho^2})$. In this case, the parameter ρ^2 still characterizes the shape of the transformed Bernstein ellipse. Therefore, throughout this paper, when we say a transformed Bernstein ellipse with parameter ρ^2, we refer to the parameter of the Bernstein ellipse defined on [-1, 1]. Let the function domain be $\Omega_{\mathbf{x}} \times \Omega_{\mathbf{y}} \subset \mathbb{R}^d \times \mathbb{R}^d$, and we refer to $\Omega_{\mathbf{x}}$ as the *target domain* and $\Omega_{\mathbf{y}}$ as the *source domain*. We assume the domain is not a manifold, where lower ranks can be expected. Let the sub-domain containing data of interest be $\widetilde{\Omega}_{\mathbf{x}} \times \widetilde{\Omega}_{\mathbf{y}} \subset \Omega_{\mathbf{x}} \times \Omega_{\mathbf{y}}$.

The following theorems assume the bandwidth parameter h in $\mathcal{K}_h(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|_2^2/h^2)$ to be fixed at 1. A scaled kernel $\mathcal{K}_h(\mathbf{x}, \mathbf{y})$ will not be considered because it can be handled by rescaling the data points instead. Now that we have introduced the notations and made the assumptions, we can present our theorems.

Theorem 2.1. *Consider a function f and kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x} - \mathbf{y}\|_2^2)$ with $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$. We assume that $x_i \in [0, D/\sqrt{d}]$, $y_i \in [0, D/\sqrt{d}]$, where D is a constant independent of d. And this implies $\|\mathbf{x} - \mathbf{y}\|_2^2 \le D^2$.*

We further assume that f is analytic in $[0, D^2]$, and is analytically continuable to a transformed Bernstein ellipse with parameter $\rho_D^2 > 1$, and $|f(x)| \le C_D$ inside the ellipse.

Then, $\forall\ n \ge 0$, the kernel $\mathcal{K}$ can be approximated in the L_∞ norm by a low-rank kernel $\widetilde{\mathcal{K}}$ of function rank at most $R(n, d) = \binom{n+d+2}{d+2}$

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{R} g_i(\mathbf{x}) h_i(\mathbf{y}) + \epsilon_n = \widetilde{\mathcal{K}}(\mathbf{x}, \mathbf{y}) + \epsilon_n, \tag{1}$$

where $\{g_i\}_{i=1}^R$ and $\{h_i\}_{i=1}^R$ are two sequences of d-variable polynomials. And the error term $\epsilon_n = \epsilon_n(D)$ is bounded as

$$|\epsilon_n(D)| \le \frac{2 C_D \rho_D^{-2n}}{\rho_D^2 - 1}.$$

Remark. If an approximation with a given maximal rank r is requested, we need to select an $n(r, d)$ such that $\binom{n(r,d)+d+2}{d+2} \le r$. Then, we obtain an approximation with error $|\epsilon_n(D)| \le \frac{2C_D\rho_D^{-2n}}{\rho_D^2-1}$ and function rank at most $\binom{n(r,d)+d+2}{d+2} \le r$. The low-rank kernel $\widetilde{\mathcal{K}}$ is of order $2n$, which can be revealed from the explicit form of $\widetilde{\mathcal{K}}$ in the proof (see Section 3.2). For the space of d-variate polynomials with maximum total degree $2n$, the dimension is $\binom{2n+d}{d}$. In contrast, our upper bound is $\binom{n+d+2}{d+2}$. When $d \ge 4$, our formula becomes favorable for a large range of k.

Corollary 2.2. *Under the same assumptions in Theorem 2.1 and with n fixed, the low-rank kernel approximation, for a fixed precision ϵ, is achievable with a rank proportional to $d^{\frac{-\log c_1 \epsilon}{c_2}}$, where c_1 and c_2 are positive constants.*

The proofs of Theorem 2.1 and Corollary 2.2 can be found in Section 3.1 and Section 3.2, respectively.

Theorem 2.1 suggests that for some precision ϵ, the function rank grows polynomially with increasing data dimension d, *i.e.*, $R = O(d^n)$, where n is determined by the desired precision ϵ, D, ρ_D and C_D. This can be seen from $R = \binom{n+d+2}{d+2}$ with n fixed and $d \to \infty$.

For a fixed n and for a sub-domain $\widetilde{\Omega}_{\mathbf{x}} \cup \widetilde{\Omega}_{\mathbf{y}}$ of diameter $\tilde{D} < D$, the error bound decreases. In this case, the same function f on the sub-domain can be analytically extended to a Bernstein ellipse whose parameter is larger than ρ_D^2, reducing the error bound. Therefore, when the diameter of the domain that contains our data decreases, we will observe a lower approximation error for low-rank approximations with a fixed function rank, and similarly, we will observe a lower function rank for low-rank approximations with a fixed accuracy.

Along the same line of reasoning, for a fixed kernel on a fixed domain, when the point sets become more dense, we should expect the function rank to remain unchanged for a fixed precision. The result for function ranks turns out to be in perfect agreement with the observations in practical situations on matrix ranks, assuming there are sufficiently many points to make the matrix rank visible before reaching a given precision.

We now turn to the case when d is large. Because we have assumed x_i and y_i to be in $[0, D/\sqrt{d}]$, by the concentration of measures, the values of $\|\mathbf{x} - \mathbf{y}\|_2^2$ will fall into a small-sized subinterval of $[0, D^2]$ with high probability. Therefore, we are interested in quantifying this probabilistic error bound.

In the next theorem, we will consider i.i.d. random variables $x_i\sqrt{d}$ and $y_i\sqrt{d}$, with $|x_i\sqrt{d}| < D$ and $|y_i\sqrt{d}| < D$, and their second moments exist. The assumption on the infinity norm is consistent with Theorem 2.1. For example, if

$$E_d = \Big(\sum_{i=1}^{d} \mathbf{E}[(x_i - y_i)^2]\Big)^{1/2} = \mathbf{E}[\Big(x_i\sqrt{d}\Big)^2]$$

then $E_d \in \Theta(1)$ with respect to d, *i.e.*, the mean distance between pairs of points neither goes to 0 nor ∞ with d. Also, we let $\sigma_d^2 = \sum_{i=1}^{d} \mathbf{Var}[(x_i - y_i)^2]$, and from our assumptions, $\sigma_d^2 \in \Theta(\frac{1}{d})$ (a concentration of measure).

Theorem 2.3. *Consider the same kernel function in Theorem 2.1, and let function $\tilde{f}(x - E_d^2) = f(x)$. Then, $\tilde{f}$ is analytic in $[-E_d^2, D^2 - E_d^2]$, with the parameter of its transformed Bernstein ellipse to be $\tilde{\rho}_D^2 > 1$, and $|\tilde{f}(x)| \leq \tilde{C}_D$ inside the ellipse. Suppose x_i and y_i are i.i.d. random variables, with $|x_i\sqrt{d}| < D$ and $|y_i\sqrt{d}| < D$, and have their second moments exist. Assuming $\mathbf{x}$ and $\mathbf{y}$ are sampled under the above probability distribution involving D, σ_d and E_d. Then, $\forall\ 0 < \delta < D$, with probability at least*

$$1 - 2\exp\left(\frac{-\delta^4\, d}{2\sigma_d^2\, d + 8D^2\delta^2/3}\right), \tag{2}$$

the low-rank representation in Theorem 2.1 has an error bounded by

$$|\epsilon_n(D, \delta)| \leq \frac{2C_D\delta^2}{D^2(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2}) - \delta^2}\left(\frac{D^2(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2})}{\delta^2}\right)^{-n}.$$

And with the same probability, the distance of a sampled pair will fall into the following interval

$$\|\mathbf{x} - \mathbf{y}\|_2^2 \in [E_d^2 - \delta^2, E_d^2 + \delta^2].$$

The proof of Theorem 2.3 can be found in Section 3.3.

In Theorem 2.3, as $d \to \infty$, δ need to decrease with d to maintain the same probability. If we choose $\delta = \left(\frac{C}{d}\right)^{1/4}$ with C being a very large number, then the probability remains close to 1 because $\sigma_d^2 = \Theta(\frac{1}{d})$. Moreover, we can keep ϵ_n small while reducing n, because $\delta \to 0$. This means that for sufficiently large d and for a given error, n goes down as d increases. Asymptotically, n reaches 0 and the function rank reaches 1. On the other hand, for a fixed n, the error bound decreases when d increases.

Note that $2\delta^2$ is the size of the subinterval where the values of $\|\mathbf{x}-\mathbf{y}\|_2^2$'s fall into with probability given by (2) and, by concentration of measures, with the same probability, the interval size $2\delta^2$ shrinks with increasing d. This is consistent with what we have discussed that δ need to decrease with d to maintain the same probability.

The analytic assumption in Theorem 2.1 and Theorem 2.3 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 finite-smoothness assumption, and compute the corresponding error bound.

Theorem 2.4. *Consider the same kernel function and domain in Theorem 2.1. We assume that f and its derivatives through $f^{(q-1)}$ are absolutely continuous on $[0, D^2]$ and the q-th derivative has bounded total variation on $[0, D^2]$, $V\left(\frac{\mathrm{d}^q f}{\mathrm{d}x^q}\right) \leq V_q$. Then for $n > q$, the low-rank representation in Theorem 2.1 has an L_∞ error bounded by*

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

Remark. We can weaken the assumption to $f^{(q-1)}$ being Lipschitz continuous and obtain the same error rate $O(n^{-q})$; however, we will not have the explicit constants in the error. Moreover, for functions encountered in practice, it is rare that they are Lipschitz continuous but lack the derivative of bounded total variation.

The proof of Theorem 2.4 can be found in Section 3.4.

Compared to Theorem 2.1, the convergence rate slows down from a nice geometric convergence rate $O(\rho_D^{-2n})$ 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 domain diameter D affects the error bound by D^{2q}, where q represents the smoothness of the function. For a sub-domain with diameter $\widetilde{D}$, it is straightforward to obtain that the error is bounded by $\frac{2V_q \widetilde{D}^{2q}}{\pi q[2(n-q)]^q}$, and for a fixed n, a decrease in $\widetilde{D}$ will reduce the error.

We also consider the phenomenon of concentration of measures and present the probabilistic result in the following theorem.

Theorem 2.5. *Consider the same kernel function and domain in Theorem 2.1. Suppose the assumptions in Theorem 2.4 hold, and $x_i\sqrt{d}$ and $y_i\sqrt{d}$ satisfy the same probability distribution as in Theorem 2.3 involving D, σ_d and E_d. Then, $\forall\ 0 < \delta < D$, the low-rank representation in Theorem 2.1 has an L_∞ error bounded by*

$$|\epsilon_n(V_q, \delta, q)| \leq \frac{2V_q \delta^{4q}}{\pi q[2(n-q)]^q}$$

with probability at least

$$1 - 2\exp\left(\frac{-\delta^4 d}{2\sigma_d^2 d + 8D^2\delta^2/3}\right).$$

The proof of Theorem 2.5 can be found in Section 3.4.

Up to now, we have only considered a single parameter D that characterizes the domain, to make the error bound more informative as in response to more subtle changes of the domain, we also consider the diameters of the target domain $D_{\mathbf{x}}$ and of the source domain $D_{\mathbf{y}}$. The following theorem nicely quantifies the influences of $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ on the error. Our result theoretically offers critical insights and motivation for many algorithms that take advantage of the low-rank property of sub-matrices, where these sub-matrices usually relate to data clusters that are of small diameters.

Theorem 2.6. *Consider the same kernel function and domain in Theorem 2.1. We assume further that there are $D_{\mathbf{x}} < D$ and $D_{\mathbf{y}} < D$, such that $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq D_{\mathbf{x}}$ and $\|\mathbf{y}_i - \mathbf{y}_j\|_2 \leq D_{\mathbf{y}}$.*

Let $f_p(x) = \sum_n \mathcal{T} \circ f(x + 4nD^2)$ be a $4D^2$-periodic extension of $f(x)$, where $\mathcal{T}(\cdot)$[1] is 1 on $[-D^2, D^2]$ and smoothly decays to 0 outside of the interval. We assume that f_p and its derivatives through $f_p^{(q-1)}$ are continuous, and the q-th derivative is piecewise continuous with the total variation over one period bounded by V_q.

Then, $\forall\ M_f, M_t > 0$ with $9M_f \leq M_t$, the kernel $\mathcal{K}$ can be represented in the form of Equation 1 of rank at most $R(M_f, M_t, d) = 4M_f\binom{M_t+d}{d}$ with L_∞ error bounded by

$$|\epsilon_{M_f, M_t}(D_{\mathbf{x}}, D_{\mathbf{y}}, q, \rho)| \leq \|f\|_\infty \left(\frac{D_{\mathbf{x}} D_{\mathbf{y}}}{D^2}\right)^{M_t+1} + \frac{V_q}{\pi q}\left(\frac{2D^2}{\pi M_f}\right)^q$$

[1] see details in [4]

The proof of Theorem 2.6 can be found in Section 3.5.

Different than the previous theorems where the domain information only enters the error as D, in Theorem 2.6, the diameters of the source domain $D_{\mathbf{y}}$ and target domain $D_{\mathbf{x}}$ also appear in error. The form $\left(\frac{D_{\mathbf{x}}D_{\mathbf{y}}}{D^2}\right)^{M_t+1}$ suggests that a decrease in $\frac{D_{\mathbf{x}}D_{\mathbf{y}}}{D^2}$ will reduce the error, which can be achieved when either the source or the target domain has a smaller diameter. 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 diameters.

The function rank still remains a polynomial growth and it grows as $R = O(d^{M_t})$, when M_f and M_t are fixed and $d \to \infty$. M_f represents the Fourier expansion order of f, and each term in the expansion is further expanded into Taylor terms up to order M_t. We assumed M_t to be the same across all the Fourier terms for simplicity. If we decrease the Taylor order M_t with increasing Fourier order to preserve more information of low-order Fourier terms, then a lower error bound can be attained for the same function rank.

Remark. We summarize the assumptions, error bounds and function ranks of the deterministic theorems in Table 1, and discuss the similarities and differences in the function rank and the error bound. We refer to Theorem 2.1 and Theorem 2.4 as the Chebyshev approach and Theorem 2.6 as the Fourier-Taylor approach based on their proof techniques. The function rank is determined by the data dimension and the expansion order, and it is a power of the dimension, where the power is the expansion order and is different in the Chebyshev approach and the Fourier-Taylor approach. The error bounds quantify the influences from the expansion order and the domain diameter: a higher expansion order reduces the error bound, so does a smaller domain diameter. The domain diameter occurs as a single parameter D in the Chebyshev approach but as $D_{\mathbf{x}}$, $D_{\mathbf{y}}$ and D in the Fourier-Taylor approach.

From the practical viewpoint, the absence of exponential growth for the function rank agrees with the practical situation where people observe lower matrix ranks for high dimension data. And, the fact that decreasing $D_{\mathbf{x}}$ or $D_{\mathbf{y}}$ reduces the error is also in agreement with practice and moreover, it provides an insight of why point clusterings followed by local interpolations often leads to a more memory efficient approximation.

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 transformed Bernstein ellipse with parameter $\rho_D^2 > 1$. • $\|f(x)\| \le C < \infty$	• The first $q-1$ derivatives of f are absolutely continuous on $[0, D^2]$, and the q-th derivative on $[0, D^2]$ has bounded total variation V_q	• Let $f_p(x)$ be the $4D^2$-periodic extension of f and $\|f(x) - f_p(x)\|_{\infty,[-D^2,D^2]} = 0$ • The first $q-1$ derivatives of f_p are continuous, and the q-th derivative on $[-D^2, D^2]$ is piece-wise continuous with bounded total variation V_q • $9M_f \le M_t$
Error	$\frac{2C_D\rho_D^{-2n}}{\rho_D^2-1}$	$\frac{2V_qD^{2q}}{\pi q[2(n-q)]^q}$	$\|f\|_\infty\left(\frac{D_{\mathbf{x}}D_{\mathbf{y}}}{D^2}\right)^{M_t+1} + \frac{V_q}{\pi q}\left(\frac{2D^2}{\pi M_f}\right)^q$
Rank	$\binom{n+d+2}{d+2}$		$4M_f\binom{M_t+d}{d}$
Notation	n: Chebyshev expansion order D: $\|\mathbf{x}-\mathbf{y}\|_2 < D$		$D_{\mathbf{x}}$: $\|\mathbf{x}_i-\mathbf{x}_j\|_2 < D_{\mathbf{x}}$ $D_{\mathbf{y}}$: $\|\mathbf{y}_i-\mathbf{y}_j\|_2 < D_{\mathbf{y}}$ M_f: Fourier expansion order M_t: Taylor expansion order

3 Theorem Proof

In this section, we prove the theorems in section 2. All the proofs consist of three components: separating $\mathcal{K}(\mathbf{x}, \mathbf{y})$ into a finite sum of products of real valued functions $h_i(\mathbf{x})g_i(\mathbf{y})$, counting the terms to obtain an upper bound for the function rank, and calculating the error bound. Similar techniques can be found in [31, 43, 24]. We describe the high-level procedure of the separation step; the rest steps should be straightforward.

In the proofs of Theorem 2.1 and Theorem 2.4, the separate form was obtained by first expanding the kernel into polynomials of $z = \|\mathbf{x}-\mathbf{y}\|^2$ of a certain order to settle the error bound, and then expanding the terms $\|\mathbf{x}-\mathbf{y}\|^{2l}$. The key advantage of this approach has been discussed at the

beginning of section 2. We seek approximation theorems in 1D that provide optimal convergence rate and explicit error bounds. Chebyshev theorems (Theorem 8.2 and Theorem 7.2 in [35]) are ideal choices. Analogous results also exist, *e.g.*, the classic Bernstein and Jackson's approximation theorems, but the downside is that they only provide an error rate rather than an explicit formula, and moreover, they will not improve our results or simplify the proofs.

We discuss briefly the differences of those approximation theorems. The classic Bernstein approximation theorem is analogous to Theorem 8.2 in [35], and they have the same assumptions and error rate (with no explicit constant). The Jackson's approximation theorem is analogous to Theorem 7.2 in [35], and it weakens the assumption from $f^{(q)}$ having bounded total variation to $f^{(q-1)}$ being Lipschitz continuous and provides the same error rate; however, it is not common in practice to see a function that is Lipschitz continuous but lacks derivative of bounded variation.

In the proof of Theorem 2.6, the separate form was obtained by first applying a Fourier expansion on $\mathcal{K}$ to separate the cross term $\exp(\mathbf{x}^T\mathbf{y})$, then applying a Taylor expansion on the cross term.

Before stating the detailed proofs, we introduce some notations that will be used.

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 two components. First, we map the domain of f to $[0, 2]$ (for the convenience of the proof) and approximate f with a Chebyshev polynomial, and this settles the error. Second, we further separate terms $\|\mathbf{x} - \mathbf{y}\|^2$ in the polynomial and count the number of distinct terms to be the upper bound of the function rank.

1. *Polynomial approximation with Chebyshev.* We first linearly map the domain of f to $[0, 2]$ and denote the new function as $\tilde{f}$:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = f(\|\mathbf{x} - \mathbf{y}\|_2^2) = \tilde{f}\left(\frac{2}{D^2}\|\mathbf{x} - \mathbf{y}\|_2^2\right) = \tilde{f}(z). \tag{3}$$

Because $\|\mathbf{x} - \mathbf{y}\|^2 \in [0, D^2]$, it follows that $r \in [0, 2]$. From our assumptions, $\tilde{f}$ is analytic in $[0, 2]$ and is analytically continuable to the open Bernstein ellipse with parameter ρ_D^2 (consider a shifted ellipse).

According to Theorem 8.2 in [36], $\forall\ n \geq 0$, we can approximate $\tilde{f}$ by its Chebyshev truncations $\tilde{f}_n$ in the L_∞ norm with error

$$|\epsilon_n| \leq \frac{2C_D\rho_D^{-2n}}{\rho_D^2 - 1}, \tag{4}$$

and

$$\tilde{f}_n(z) = \sum_{k=0}^{n} c_k T_k(z) + \epsilon_n, \tag{5}$$

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

$$T_k(x) = \cos(k\theta), \text{ with } x = \cos(\theta). \tag{6}$$

Rearranging the terms in (5) we obtain a polynomial of $z = \|\mathbf{x} - \mathbf{y}\|^2$:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \tilde{f}\left(\frac{2}{D^2}\|\mathbf{x} - \mathbf{y}\|_2^2\right) = \sum_{k=0}^{n} \frac{a_k}{D^{2k}}\|\mathbf{x} - \mathbf{y}\|^{2k} + \epsilon_n, \tag{7}$$

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

2. *Separable form.* We separate each term $\|\mathbf{x}-\mathbf{y}\|^{2l}$ in (7) into a finite sum of separate products:

$$\|\mathbf{x}-\mathbf{y}\|^{2l} = \sum_{k=0}^{l}\sum_{j=0}^{k}\sum_{|\boldsymbol{\alpha}|=l-k} C_{l,k,\boldsymbol{\alpha}} \left(\|\mathbf{x}\|^{2j}\mathbf{x}^{\boldsymbol{\alpha}}\right)\left(\|\mathbf{y}\|^{2(k-j)}\mathbf{y}^{\boldsymbol{\alpha}}\right), \tag{8}$$

where $C_{l,k,\boldsymbol{\alpha}} = (-2)^{l-k}\binom{l}{k}\binom{k}{j}\binom{l-k}{\boldsymbol{\alpha}}$. Substituting (8) into (7), we obtain a separate form of $\mathcal{K}$:

$$\mathcal{K}(\mathbf{x},\mathbf{y}) = \sum_{l=0}^{n}\sum_{k=0}^{l}\sum_{j=0}^{k}\sum_{|\boldsymbol{\alpha}|=l-k} D_{l,k,\boldsymbol{\alpha}} \left(\|\mathbf{x}\|^{2j}\mathbf{x}^{\boldsymbol{\alpha}}\right)\left(\|\mathbf{y}\|^{2(k-j)}\mathbf{y}^{\boldsymbol{\alpha}}\right) + \epsilon_n, \tag{9}$$

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 $\mathbf{x}$ and $\mathbf{y}$. Therefore, the function rank of $\mathcal{K}$ can be upper bounded by the total number of separate terms:

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

To summarize, we have proved that $\mathcal{K}(\mathbf{x},\mathbf{y})$ can be approximated by the separable form in (9) in the L_∞ norm with rank at most

$$R(n,d) = \binom{n+d+2}{d+2}, \tag{10}$$

and approximation error

$$|\epsilon_n(D)| \le \frac{2C_D\rho_D^{-2n}}{\rho_D^2-1}. \tag{11}$$

□

3.2 Proof of Corollary 2.2

Proof. For a fixed kernel function and fixed n, we define two constants $c_1 = \frac{\rho_D^2-1}{2C_D}$ and $c_2 = \log\rho_D^2$. Then, the truncation error ϵ can be rewritten as

$$\epsilon = \frac{2C_D\rho_D^{-2n}}{\rho_D^2-1} = \frac{e^{-nc_2}}{c_1},$$

and equivalently,

$$n = \frac{-\log c_1\epsilon}{c_2}. \tag{12}$$

We relate function rank R to error ϵ and dimension d. When $d \ge n+2$, we have,

$$R = \binom{n+d+2}{d+2} \le \frac{2^n d^n}{n!} = c_n d^{\frac{-\log c_1\epsilon}{c_2}}, \tag{13}$$

where $c_n = \frac{2^n}{n!}$ is a constant for a fixed n. Therefore, an ϵ error is achievable with the function rank R proportional to $d^{\frac{-\log c_1\epsilon}{c_2}}$. □

3.3 Proof of Theorem 2.3

Proof. We consider the concentration of measure phenomenon and apply concentration inequalities to obtain a probabilistic error bound. The proof mostly follows the proof of Theorem 2.1, and we will focus on computing the error bound for a smaller domain.

To simplify the proof, we consider a function $\tilde{f}$ that is shifted by E_d^2 such that

$$f(\|\mathbf{x}-\mathbf{y}\|_2^2) = \tilde{f}(\|\mathbf{x}-\mathbf{y}\|_2^2 - E_d^2),$$

and we will see later that this shift ensures the inputs for $\tilde{f}$ to fall into an interval that centers around 0 with some probability. $\tilde{f}$ inherits the analyticity of f, therefore, it is analytic on

$[-E_d^2, D^2 - E_d^2]$, and can be analytically extended to a transformed Bernstein Ellipse with parameter $\tilde{\rho}_D^2$.

Let us denote $z_i = (x_i - y_i)^2 - \mathbf{E}[(x_i - y_i)^2]$ and we will shortly apply concentration inequality to this variable. With the assumptions that $x_i\sqrt{d}$ and $y_i\sqrt{d}$ are i.i.d. random variables where $|x_i\sqrt{d}| < D$ and $|y_i\sqrt{d}| < D$, it follows that z_i's are statistically independent with mean zero and are bounded by $\frac{4D^2}{d}$. By applying the Bernstein's inequality [3] on the sum of z_i's, we conclude that $\forall\ \delta \geq 0$,

$$P\left(|\|\mathbf{x} - \mathbf{y}\|_2^2 - E_d^2| \leq \delta^2\right) \geq 1 - 2\exp\left(\frac{-\delta^4\, d}{2\sigma_d^2\, d + 8D^2\delta^2/3}\right), \tag{14}$$

where $E_d^2 = \sum_{i=1}^{d} \mathbf{E}[(x_i - y_i)^2] = 2\mathbf{E}[\left(x_i\sqrt{d}\right)^2]$ is a constant. In other words, $\|\mathbf{x} - \mathbf{y}\|_2^2 \in [E_d^2 - \delta^2, E_d^2 + \delta^2]$ with probability at least

$$1 - 2\exp\left(\frac{-\delta^4\, d}{2\sigma_d^2\, d + 8D^2\delta^2/3}\right).$$

This also means that with the same probability in (14), the inputs for $\tilde{f}$ will fall into the interval $[-\delta^2, \delta^2]$.

Therefore, for a probability associated with δ, we can turn to considering $\tilde{f}$ on domain $[-\delta^2, \delta^2]$. We assume that $\tilde{f}$ is analytically extended to a transformed Bernstein ellipse with parameter ρ_δ^2, with the value of $\tilde{f}(z)$ inside the ellipse bounded by C_δ. Following the same argument in the proof for Theorem 2.1, we obtain that $\forall\ \delta > 0$ and with probability in (14), the approximation error for $\mathbf{x}$ and $\mathbf{y}$ sampled from the above distribution is bounded by

$$|\epsilon_n| \leq \frac{2C_\delta}{\rho_\delta^2 - 1}\rho_\delta^{-2n}. \tag{15}$$

This sharper bound can be achieved with the same function rank as in (10) and with the same low-rank representation as in (9) except for coefficients.

Next, we rewrite the upper bound in (15) with the parameters $\tilde{\rho}_D$, $\tilde{C}_D$, and δ. If we linearly map the domain of $\tilde{f}$ from $[-\delta^2, \delta^2]$ to [-1,1], then the Bernstein ellipse with parameter $\tilde{\rho}_D^2$ will be scaled by $\frac{1}{\delta^2}$. We seek the largest ρ_δ^2 such that the Bernstein ellipse with parameter ρ_δ^2 will be contained in the transformed Bernstein ellipse with parameter $\tilde{\rho}_D^2$. In that case, the lengths of their semi-minor axises match and the largest ρ_δ^2 satisfies

$$\rho_\delta^2 - \rho_\delta^{-2} = \frac{D^2}{\delta^2}\left(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2}\right), \tag{16}$$

and we obtain $\rho_\delta^2 = \frac{D^2}{\delta^2}(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2}) + \left(\frac{D^4}{4\delta^4}(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2})^2 + 1\right)^{\frac{1}{2}}$. In the special case where $\delta^2 = D^2$, $\rho_\delta^2 = \tilde{\rho}_D^2$, we recover the error bound $\frac{2C_D}{\tilde{\rho}_D^2 - 1}\tilde{\rho}_D^{-2n}$. To simplify the bound, we use the relation that $\rho_\delta^2 > \frac{D^2}{\delta^2}(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2})$. Substituting this into (15), along with the fact that $C_\delta \leq C_D$, we obtain

$$|\epsilon_n| \leq \frac{2C_D\delta^2}{D^2(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2}) - \delta^2}\left(\frac{D^2(\tilde{\rho}_D^2 - \tilde{\rho}_D^{-2})}{\delta^2}\right)^{-n}. \tag{17}$$

Therefore, the function rank related to error ϵ_n remains $\binom{n+d+2}{d+2}$, and we have proved our result. □

3.4 Proof of Theorem 2.4 and Theorem 2.5

Proof. The proof follows the same steps as that in Theorem 2.1 and Theorem 2.3; 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 (3). Because $f^{(q)}$ is piecewise continuous with its total variation on $[0, D^2]$ bounded by V_q, it follows that $\tilde{f}^{(q)}$ in (3) is piecewise continuous on $[0, 2]$, with its total variation on $[0, 2]$ bounded as follows

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

Therefore, by Theorem 7.2 in [36], for $n > q$, the order-n Chebyshev expansion $\tilde{f}_n$ approximates $\tilde{f}$ in the L_∞ norm with error bounded by

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

The rest of the proof is identical to that of Theorem 2.1 for the deterministic result, and identical to that of Theorem 2.3 for the probabilistic result. □

3.5 Proof of Theorem 2.6

We first introduce a lemma concerning the function 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.,*

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

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 $R = 2R_c$,

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

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

$$\begin{aligned}\Psi_i(\mathbf{x})\Phi_i(\mathbf{y}) = &\left(\mathrm{Re}\,(\Psi_i(\mathbf{x}))\,\mathrm{Re}\,(\Phi_i(\mathbf{y})) - \mathrm{Im}\,(\Psi_i(\mathbf{x}))\,\mathrm{Im}\,(\Phi_i(\mathbf{y}))\right)\\ &+ \imath\left(\mathrm{Re}\,(\Psi_i(\mathbf{x}))\,\mathrm{Im}\,(\Phi_i(\mathbf{y})) + \mathrm{Im}\,(\Psi_i(\mathbf{x}))\,\mathrm{Re}\,(\Phi_i(\mathbf{y}))\right)\end{aligned} \tag{18}$$

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

$$\begin{cases} g_{2i-1}(\mathbf{x}) = \mathrm{Re}\,(\Psi_i(\mathbf{x})), g_{2i}(\mathbf{x}) = -\mathrm{Im}\,(\Psi_i(\mathbf{x})) \\ h_{2i-1}(\mathbf{y}) = \mathrm{Re}\,(\Phi_i(\mathbf{y})), h_{2i}(\mathbf{y}) = \mathrm{Im}\,(\Phi_i(\mathbf{y})) \end{cases}, \quad i = 1, 2, \ldots, R_c \tag{19}$$

The approximation error holds for the real-valued approximation:

$$\left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R} g_i(\mathbf{x})h_i(\mathbf{y})\right| \leq \left|\mathcal{K}(\mathbf{x},\mathbf{y}) - \sum_{i=1}^{R_c} \Psi_i(\mathbf{x})\Phi_i(\mathbf{y})\right| \leq \epsilon \tag{20}$$

□

We now start the proof for Theorem 2.6.

Proof. The proof consists of three major parts: derivation of a separable form for $\mathcal{K}(\mathbf{x},\mathbf{y})$, analysis on the truncation error, and estimation of the number of separable terms. The first part is proceeded in three steps: Fourier expansion of the periodic input function, Taylor expansion of each Fourier component, and finalization on the overall separable form.

We denote by $\Omega_\mathbf{x}$ the domain of $\mathbf{x}$, and $\Omega_\mathbf{y}$ the domain of $\mathbf{y}$, with their centers to be $\mathbf{x}_c$ and $\mathbf{y}_c$, respectively. To simplify the notations, we use $f(\cdot)$ to represent the periodic function $f_p(\cdot)$.

1.1. Fourier expansion. Let the Fourier expansion of f with error term ϵ_F be

$$f(z) = \sum_{j=-M_f}^{M_f} a_j \exp(\imath\omega j z) + \epsilon_F, \tag{21}$$

where $a_j = \frac{1}{4D^2}\int_{-2D^2}^{2D^2} f(z)\exp(-\imath\omega j z)\,\mathrm{d}z$ is the Fourier coefficient and $\omega = \frac{2\pi}{4D^2}$ is a constant. Obviously, each Fourier coefficient can be bounded by the infinity norm of the function $f(z)$, *i.e.*, $|a_j| \leq \|f\|_\infty$. Detailed analysis of the error ϵ_F will be discussed in the second major part of the proof. The fact that $\mathcal{K}(\mathbf{x},\mathbf{y}) = f(\|\mathbf{x}-\mathbf{y}\|_2^2)$ is a function of $z = \|\mathbf{x}-\mathbf{y}\|_2^2$ naturally requires a

separation of z in order to proceed with the separation of $\mathcal{K}(\mathbf{x},\mathbf{y})$. Adopting notations, $\boldsymbol{\rho}_{\mathbf{x}} = \mathbf{x}-\mathbf{x}_c$, $\boldsymbol{\rho}_{\mathbf{y}} = \mathbf{y}-\mathbf{y}_c$ and $\boldsymbol{\rho}_c = \mathbf{x}_c-\mathbf{y}_c$, we rewrite $z = \|\mathbf{x}-\mathbf{y}\|_2^2 = \|\boldsymbol{\rho}_{\mathbf{x}}+\boldsymbol{\rho}_c\|^2+\|\boldsymbol{\rho}_{\mathbf{y}}\|^2-2\boldsymbol{\rho}_{\mathbf{y}}^T\boldsymbol{\rho}_c-2\boldsymbol{\rho}_{\mathbf{x}}^T\boldsymbol{\rho}_{\mathbf{y}}$ and, therefore,

$$\exp(\imath\omega j z) = \underbrace{\exp(\imath\omega j\|\boldsymbol{\rho}_{\mathbf{x}}+\boldsymbol{\rho}_c\|^2)}_{\text{function of } \mathbf{x} \text{ only}}\underbrace{\exp(\imath\omega j(\|\boldsymbol{\rho}_{\mathbf{y}}\|^2-2\boldsymbol{\rho}_{\mathbf{y}}^T\boldsymbol{\rho}_c))}_{\text{function of } \mathbf{y} \text{ only}}\underbrace{\exp(-\imath\omega j 2\boldsymbol{\rho}_{\mathbf{x}}^T\boldsymbol{\rho}_{\mathbf{y}})}_{\text{function of } \mathbf{x} \text{ and } \mathbf{y}}. \tag{22}$$

1.2. Taylor expansion. The last term in (22) still involves both $\mathbf{x}$ and $\mathbf{y}$ and need to be further separated. We apply the Taylor expansion to this term,

$$\begin{aligned}\exp\left(-\imath\omega j 2\boldsymbol{\rho}_{\mathbf{x}}^T\boldsymbol{\rho}_{\mathbf{y}}\right) &= \sum_{k=0}^{M_t}\frac{(-\imath\omega j 2\boldsymbol{\rho}_{\mathbf{x}}^T\boldsymbol{\rho}_{\mathbf{y}})^k}{k!}+\epsilon_T(j)\\ &= \sum_{k=0}^{M_t}\frac{(-\imath 2j\omega)^k}{k!}\sum_{|\boldsymbol{\alpha}|=k}\binom{k}{\boldsymbol{\alpha}}\boldsymbol{\rho}_{\mathbf{x}}^{\boldsymbol{\alpha}}\boldsymbol{\rho}_{\mathbf{y}}^{\boldsymbol{\alpha}}+\epsilon_T(j),\end{aligned} \tag{23}$$

where M_t is the order of the Taylor expansion, $\epsilon_T(j)$ is the corresponding truncation error, and the last equality adopts the multi-index notation introduced earlier.

1.3. Separable form. Combining (23), (22) and (21), we obtain

$$f(z) = \sum_{j=-M_f}^{M_f}\sum_{k=0}^{M_t}\sum_{|\boldsymbol{\alpha}|=k} h_{j,\boldsymbol{\alpha}}(\mathbf{x})g_{j,\boldsymbol{\alpha}}(\mathbf{y})+\epsilon, \tag{24}$$

where

$$\begin{aligned}h_{j,\boldsymbol{\alpha}}(\mathbf{x}) &= a_j\frac{(-\imath 2j\omega)^k}{k!}\binom{k}{\boldsymbol{\alpha}}\exp\left(\imath\omega j\|\boldsymbol{\rho}_{\mathbf{x}}+\boldsymbol{\rho}_c\|^2\right)\boldsymbol{\rho}_{\mathbf{x}}^{\boldsymbol{\alpha}} \text{ and}\\ g_{j,\boldsymbol{\alpha}}(\mathbf{y}) &= \exp\left(\imath\omega j(\|\boldsymbol{\rho}_{\mathbf{y}}\|^2-2\boldsymbol{\rho}_{\mathbf{y}}^T\boldsymbol{\rho}_c)\right)\boldsymbol{\rho}_{\mathbf{y}}^{\boldsymbol{\alpha}}\end{aligned} \tag{25}$$

are functions of $\mathbf{x}$ only and $\mathbf{y}$ only respectively, and ϵ is the overall error

$$\epsilon = \sum_{j=-M_f}^{M_f} a_j\exp\left(\imath\omega j\|\boldsymbol{\rho}_{\mathbf{x}}+\boldsymbol{\rho}_c\|^2\right)\exp\left(\imath\omega j(\|\boldsymbol{\rho}_{\mathbf{y}}\|^2-2\boldsymbol{\rho}_{\mathbf{y}}^T\boldsymbol{\rho}_c)\right)\epsilon_T(j)+\epsilon_F. \tag{26}$$

A naïve bound on ϵ is given as,

$$|\epsilon| \le \sum_{j=-M_f}^{M_f}|a_j||\epsilon_T(j)|+|\epsilon_F| \le 2M_f\|f\|_\infty\max_j|\epsilon_T(j)|+|\epsilon_F|, \tag{27}$$

where the first inequality used the fact that the absolute values of both exponential terms are one.

2. Error analysis. According to (27), the total error consists of two parts, the truncation errors of the Taylor expansion and of the Fourier expansion. We consider first the Taylor expansion errors. Applying the Lagrange remainder form, we bound the Taylor part of the total error as where the second inequality adopts the inequality $e(\frac{n}{e})^n \le n!$ with e being the Euler's constant, and the third inequality can be varified with our assumption $9M_f \le M_t$.

We then consider the Fourier expansion errors. According to Theorem 2 in [15], the truncation error of the Fourier expansion, ϵ_F can be bounded as follows

$$|\epsilon_F| \le \frac{V_q}{\pi q(\omega M_f)^q} = \frac{V_q}{\pi q}\left(\frac{2D^2}{\pi M_f}\right)^q, \tag{28}$$

where V_q is the total variation of the q-th derivative of $f(z)$ over one period.

Therefore, the total error ϵ in (24) can be bounded as

$$|\epsilon| \le \|f\|_\infty\left(\frac{D_{\mathbf{x}}D_{\mathbf{y}}}{D^2}\right)^{M_t+1}+\frac{V_q}{\pi q}\left(\frac{2D^2}{\pi M_f}\right)^q. \tag{29}$$

3. Rank computation. Equation (24) is a separable form of $\mathcal{K}(\mathbf{x}, \mathbf{y})$ in its complex form with rank at most

$$R_c = 2M_f \sum_{\ell=0}^{M_t} \binom{\ell + d - 1}{d - 1} \leq 2M_f \binom{M_t + d}{d}. \tag{30}$$

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 at most

$$R(M_f, M_t, d) = 2R_c \leq 4M_f \binom{M_t + d}{d}. \tag{31}$$

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

□

4 Optimality of the polynomial growth of the function rank

Corollary 2.2 shows that asymptotically, for a given error ϵ and dimension d, the function rank needed for a low-rank representation to approximate an analytic function with error ϵ is proportional to $d^{\frac{-\log c_1 \epsilon}{c_2}}$. We will show that up to some constant, this asymptotic rank has achieved the lower bound on the minimal number of interpolation points needed for a linear operator to reach a required accuracy [41].

Wozniakowski stated in [41] that for a given ϵ and d, the minimal number of interpolation points $n = n(\epsilon, d)$, for a linear interpolation operator $L_n(f) = \sum_{j=1}^n f(x_j)c_j$ to approximate function f that satisfies $\|f\|_k \leq 1$ in the L_2 norm with precision ϵ, is bounded by

$$n(\epsilon, d) \geq c_\epsilon d^{c \log(\epsilon^{-1})}, \tag{32}$$

where $c_j \in C([-1, 1]^d)$, and $\|f\|_k^2 := \sum_{l \in \mathbb{N}_0} (1 + l^2)^k a_l^2[f]$ with $a_l[f]$ denoting the Fourier coefficient of f.

We establish that the function rank in Theorem 2.1 is equivalent to $n(\epsilon, d)$ described above. We start with the assumptions. In Theorem 2.1, the analytic assumption implies that $\|f\|_k \leq 1$, and the L_∞-norm error suggests the same results hold for L_2-norm error if we assume the volume of the domain is bounded by 1. We then connect the number of points from a function interpolation to the number of terms from a function expansion by the following formula:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \mathcal{K}(\mathbf{x}_i, \mathbf{y}) c_i(\mathbf{x}) + \epsilon. \tag{33}$$

Therefore, we have established the equivalence of the function rank in Theorem 2.1 and $n(\epsilon, d)$, and we conclude that our function rank reaches the lower bound in (32) asymptotically.

Related work. Barthelmann [2] considered a polynomial interpolation on a sparse grid, and showed that such interpolation could reach an acceptable accuracy with the number of interpolation points growing polynomially with the data dimension. Specifically, consider a real-valued function f defined on $[-1, 1]^d$ with its derivative $D^{\boldsymbol{\alpha}} f$ being continuous for $\|\boldsymbol{\alpha}\|_\infty \leq k$. If we interpolate f using the Smolyak formula [34], then the interpolation error in the 0-norm is bounded by

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

where the norm $\|\cdot\|_0$ and $\|\cdot\|_k$ adopt the same notations as above. The number of interpolation points used (see [26]) is

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

Consider $N(k + d, d)$. When k is fixed and $d \to \infty$, the number of points used in the Smolyak technique roughly behaves as $O(d^k)$. We use the same argument between the lines of (33) to connect the number of function interpolation points and the number of function expansion terms, and conclude that the polynomial dependence on d is consistent with our result in (9).

In the following section, we use the matrix rank to verify our theoretical results on the function rank. We have mentioned in section 1 that the function rank is an upper bound of the matrix rank. Hence, we would expect the matrix rank related to the max norm to grow polynomially with d as well. The low-rank representation of a kernel function and its approximation error can be related to those of a kernel matrix defined on the same domain in the following way. If a kernel function $\mathcal{K}$ can be approximated by the separate form $\sum_{i=1}^{R} h_i(\mathbf{x})g_i(\mathbf{y})$ with L_∞ error ϵ. Then, for an n by n kernel matrix K with entries $K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{y}_j)$, it is straightforward to construct a low-rank representation GH^T of K with rank at most R, where $G_{ij} = g_j(\mathbf{x}_i)$ and $H_{ij} = h_j(\mathbf{y}_i)$. And, the matrix approximation error in the Frobenius-, two-, and max- norm is bounded by ϵn, ϵn, and ϵ, respectively.

Now that the connections between matrix rank and function rank have been established explicitly, we can move on to the numerical experiments.

5 Numerical experiments

In this section, we experimentally verify two main results from our theorems: the polynomial growth of the numerical function rank with the data dimension, and the influence of the diameters of $\Omega_{\mathbf{x}}$ and $\Omega_{\mathbf{y}}$ on the approximation error. By the arguments before the beginning of this section, we will use the matrix rank to verify the behavior on the function rank. We report the matrix rank for various data distributions, considering that the error bounds in our theorems is for the worst-case.

We first consider the data distribution in the experiments. Generating data which is representative of the worst case is difficult. On the one hand, if we sample randomly from some common distributions, then the empirical variance of the pair-wise distances will decrease with d due to concentration of measures; on the other hand, if we design the points, such that they are correlated to some extent, to achieve a large empirical variance, then the points will likely lie on a manifold. Both methods will yield matrices with lower matrix ranks. Because the functions we consider are functions of the distances, we seek distributions of points in a unit cube of dimension d such that the pairwise distances between points follow a probability distribution whose variance decreases slower with d, and the point layout does not lie approximately on a manifold of the domain.

For a limited number of points that is imposed by the computational limit and for large d, the fast decay of the empirical variance is observed for quasi-uniform distributions of points, *e.g.*, using data generated from perturbed grid points or Halton points. The pair-wise distances of Halton points and uniform sampled points fell into a small-sized subinterval of $[0, \sqrt{d}]$ that is away from the endpoint $\sqrt{d}$, reducing the range of observed distances, leading to spurious low-ranks.

We propose a sampling distribution to encourage the occurrence of large distances that would otherwise not be covered with a high probability. Specifically, for random variable X, $\Pr(X = a) = \Pr(X = b) = p_d$, $\Pr(a < X < b) = 1 - 2p_d$, where p_d was selected by a grid search to yield the largest rank for each d. The range of the covered domain from this distribution is much wider than either using Halton points or uniform sampling.

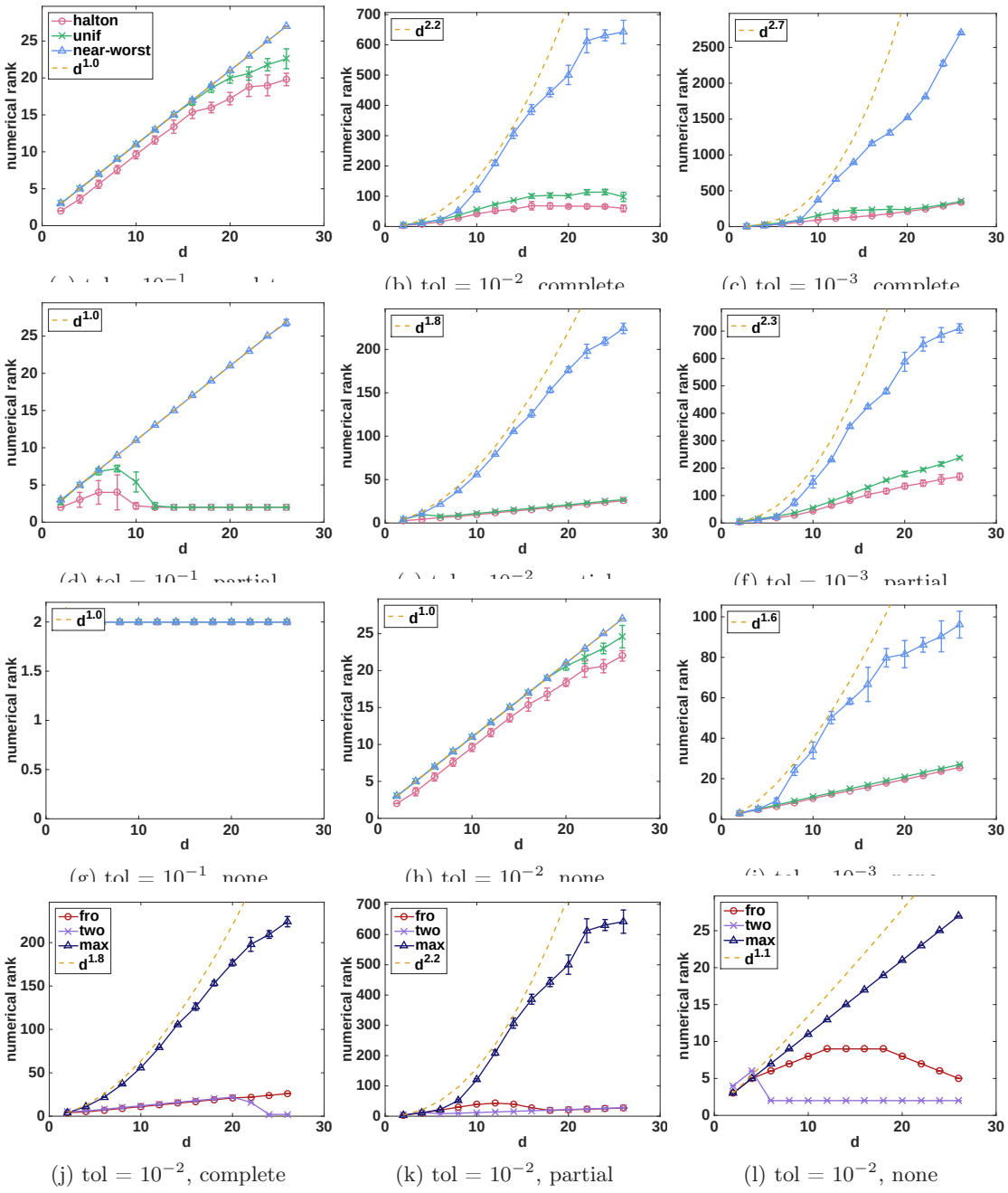
We consider next the numerical matrix rank that will be reported in the results. The numerical matrix rank R associated with tolerance tol is

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

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. 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 matrix rank; however, we find that for the RBF kernel matrices, the max error decreases in general with the matrix rank, except for certain small, short-lived increases.

Throughout our experiments, we fix the number of points at 10,000. The kernel used is the Gaussian kernel $\exp(-\|\mathbf{x} - \mathbf{y}\|^2/h^2)$ with $h = \sqrt{d}$. For each set 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 matrix 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 $D_{\mathbf{x}}$ (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.



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

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

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

(j) tol $= 10^{-2}$, complete (k) tol $= 10^{-2}$, partial (l) tol $= 10^{-2}$, none

Figure 2: Numerical rank vs. data dimension. The data size was fixed at 10,000. Subplots (a) to (i) reported results from different sampling method with the rank related to max norm; subplots (j) to (ℓ) reported results from different rank definition with worst-case sampling. Subplots (a) to (i) share d the same legend, where "halton" is Halton set; "unif" is uniform sampling; and "end point" is our proposed sampling. Subplots (j) to (ℓ) shared the same legend listing the choice of norm in the rank definition $\min\{r \mid \|K - U_r S_r V_r^T\| \leq tol\, \|K\|\}$, and "fro" is Frobenius norm; "two" is two norm; and "max" is max norm. Subplots considered different data scenarios, in which the regions containing the source and target points either completely overlap [(a) to (c)], partially overlap [(d) to (f)], or do not overlap [(g) to (i)].

The plots along each row verify that for a fixed n that represents the polynomial order in the low-rank representation, the function rank grows as $O(d^n)$ with d. In our experiments, we increase n by decreasing the approximation tolerance, according to the relation between order n and error ϵ in Theorem 2.1. We observe results consistent with the order $O(d^n)$.

The plots along each column verify that decreasing the domain diameter for either $\Omega_{\mathbf{x}}$ or $\Omega_{\mathbf{y}}$ reduces the error bound. Theorem 2.6 suggests that $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ influence the error in the form of $\left(\frac{D_{\mathbf{x}} D_{\mathbf{y}}}{D^2}\right)^{M_t+1}$. That is, to maintain a certain precision, a smaller domain diameter allows M_t to be smaller, and consequently allows the rank to be smaller. This relation of domain diameter and error bound is verified by our experimental results when observing from top to bottom.

The plots in the bottom row suggest that the matrix rank related to the Frobenius norm and the two norm increases with d in the small-d regime, and in the large-d regime, it decreases. This is an interesting observation. Regretfully, we cannot provide a clear explanation based on our theorems; we will only describe our observation in the paper and leave the theory for future work.

To summarize, up to some precision, smooth RBF kernels behave like kernels constructed by summations of products of functions of $\mathbf{x}$ and of $\mathbf{y}$. For a fixed kernel on a fixed domain, the maximal total degree of those products and the dimension altogether determine the observed function rank in practice. And, the dimension influence on the function rank is only a power of the dimension, and the power depends on the accuracy. In addition, this is still the worst-case scenario, attained for large and regular point sets. The real-world data are often more structured and rarely realize the worst-case, and for a fixed kernel and the practical data, the low-rank approximations would have lower function ranks, and hence the corresponding kernel matrices would have lower matrix ranks.

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, and the group cardinalities are dependent on the data dimension and independent of the data size.

If we order the singular values from large to small, then the indices where significant decays occur can be described as $\binom{k+d}{d}$. This number is a cumulative sum of the dimensions of the d-variate polynomial spaces arising in the terms of the truncated power series kernel $\sum_{|\boldsymbol{\alpha}|\leq k} c_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \mathbf{y}^{\boldsymbol{\alpha}}$ up to order k, which is close to our separate form of the kernel in a loose sense.

However, this formula fails to capture those less significant decays. We, therefore, explain the group pattern based on Theorem 2.6 by an appropriate grouping of the number of terms in the function's separable form. For any RBF, consider the number of separable terms $n(M_f, M_j)$ in its separate form:

$$n(M_f, M_j) = \sum_{j=0}^{M_f} \sum_{k=0}^{M_j} n_k = \sum_{j=0}^{M_f} \sum_{k=0}^{M_j} \binom{k+d-1}{k} = \sum_{j=0}^{M_f} \binom{M_j+d}{d} \tag{36}$$

The two summations correspond to the Fourier expansion of the kernel function, and the Taylor expansion of each Fourier term, respectively. n_k denotes the number of separable terms in $\left(\boldsymbol{\rho}_{\mathbf{x}}^T \boldsymbol{\rho}_{\mathbf{y}}\right)^k$ that occurs in the k-th order Taylor term. The observed group cardinalities are described by a grouping of the terms in (36), where the order of these terms is governed by the error term in the truncation. One grouping example is

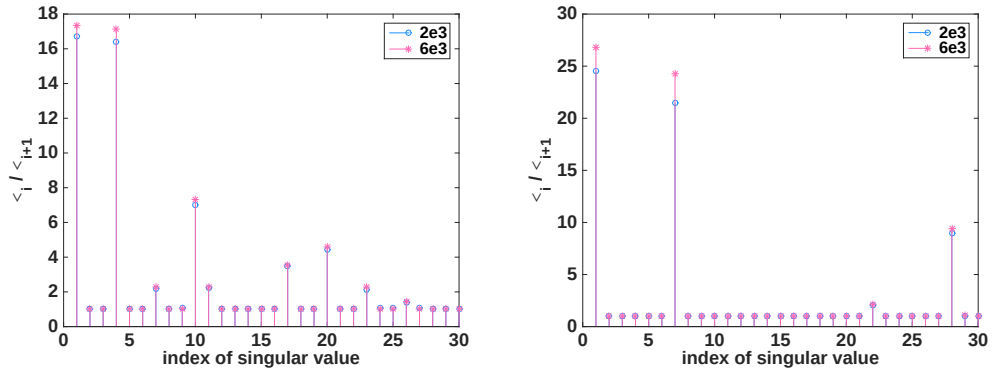
$$\underbrace{n_0, n_1, n_2}_{\text{1st term of Fourier expansion}} \;\Big|\; \underbrace{n_0, n_1}_{\text{2nd term of Fourier expansion}} \;\Big|\; \underbrace{n_3, n_4}_{\text{1st term of Fourier expansion}}$$

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. And the decay indices would be a cumulative sum of these numbers. Note that the formula in (36) is a generalization of the formula given by the dimension of the polynomial space. In the special case where only the first-order Fourier term is considered, the summation of the number of Taylor terms up to the k-th order is the same as the dimension of the d-variate polynomial space of maximum degree k.

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. We are interested in two measures: the group cardinality and the singular value decay amount, and these two measures together determine the matrix rank. In the plot, the group cardinality is indicated by the distance between two adjacent high-ratio indices, and the singular value decay amount is indicated by the magnitudes of the ratio (the spike). We observe these two measures are independent of the data size. This observation suggests that for a fixed kernel and a fixed precision, the numerical matrix rank is independent of the data size, assuming the data does not lie in a manifold. This also verifies an earlier statement that as the point sets in a fixed domain become denser, the rank and the error remain unchanged.



(a) $d = 3$, $\exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$ (b) $d = 6$, $\exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$

(c) $d = 3$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|_2^2/h^2}$ (d) $d = 6$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|_2^2/h^2}$

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 $\exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$ and subplot (c) and (d) used Cauchy kernel $1/(1 + \|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$.

We study the group cardinality in detail. Consider Figure 3a. We consider first the groups separated by significant decays. The indices with ratios above 4 are as follows with the ratio shown in parenthesis,

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

The indices can be accurately described as the cumulative sum of the number of separable terms in the following Taylor expansion terms from the first-order Fourier term,

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

This term arrangement suggests that the polynomial approximation for the first-order Fourier term contributes to the significant gains in accuracy. We note that the higher-order Fourier terms

contribute as well, but the accuracy gained from those terms is less than that from the first-order Fourier term.

We consider next the groups separated by less significant decays. 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 Fourier order and the Taylor order 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-Taylor approach instead of the Chebyshev approach. The key reason is that the Fourier approach allows us to group separable terms into finer sets that contribute to subtler error decays. 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 $\left(\boldsymbol{\rho}_{\mathbf{x}}^T\boldsymbol{\rho}_{\mathbf{y}}\right)^l$ as a unit, which only involves $\binom{l+d-1}{d-1}$ separable terms.

6.2 Practical Guidance

The group pattern in the singular values offers insights to many phenomena in practice. One example is the threshold matrix ranks in matrix approximations where the input matrix rank has to increase beyond some 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 request matrix rank; however, it is unclear what matrix rank is reasonable. Our quantification for the group cardinalities provides candidate matrix rank input for the those algorithms.

We examine the effectiveness of our guidance on two popular RBF kernel matrices with varying low-rank algorithms. We expect significant decays in the reconstruction error around matrix rank $R = \binom{n+d}{d}$. 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. Figure 4 shows the reconstruction error as a function of the approximation matrix rank. For all the algorithms, a significant decay in error occurs at rank 1, 7, and 28 for $d = 6$, and at rank 1, 9 and 45 for $d = 8$, in perfect agreements with our expectation. Note there exist several subtle perturbations, and they may be caused by the data layouts and contributions from other expansion terms.

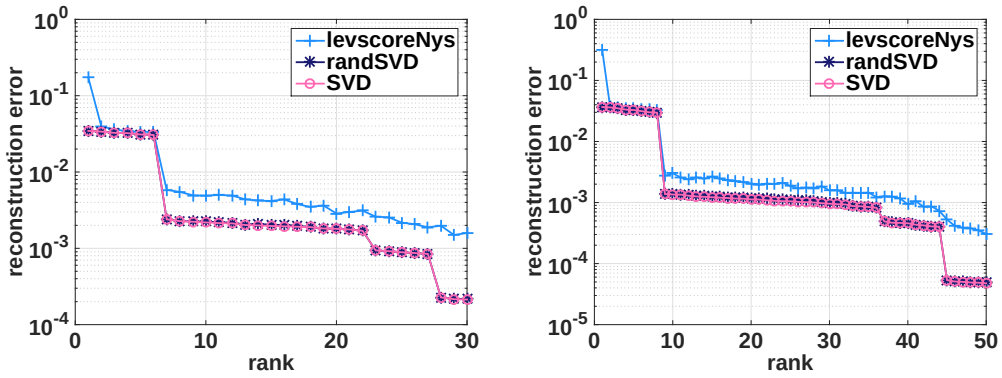
7 Conclusions

Motivated by the practical success of low-rank algorithms for RBF kernel matrices with high-dimensional datasets, we study the behavior of matrix rank of RBF kernel matrices by analyzing its upper bound, that is, the function rank of RBF kernels. Specifically, we approximate the RBF kernel by a finite sum of separate products, and explicitly describe the upper bounds on the function ranks and the L_∞ error for such approximations. Our three main results are as follows.

First, for a fixed precision, the function rank of an RBF is a power of data dimension d in the worst case, and the power is related to the precision. The exponential growth for multivariate functions from a classical analysis is absent for RBFs.

Second, for a fixed function rank, the approximation error will be reduced when the diameters of either the target domain or the source domain decrease.

Third, we observed group patterns in the magnitude of singular values of RBF kernel matrices. We explained this 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 in the function's separate form. Very commonly, the cardinality of the i-th group is $\binom{i+d-1}{d-1}$.



(a) $d = 6$, $\exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$

(b) $d = 8$, $\exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/h^2)$

(c) $d = 6$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|_2^2/h^2}$

(d) $d = 8$, $\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|_2^2/h^2}$

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 error occurs at rank $= \binom{n+d}{d}$ $(n = 1, 2, 3)$ for all experiments.

References

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

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

[3] SN Bernstein. Probability theory, moscow-leningrad, 1946.

[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] 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.

[8] 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.

[9] Tobin A Driscoll and Bengt Fornberg. Interpolation in the limit of increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 43(3-5):413–422, 2002.

[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, GRADY Wright, and Elisabeth Larsson. Some observations regarding interpolants in the limit of flat radial basis functions. *Computers & mathematics with applications*, 47(1):37–55, 2004.

[13] 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.

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

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

[16] 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.

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

[18] 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.

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

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

[21] 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.

[22] 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.

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

[24] Charles A Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, pages 143–145. Springer, 1984.

[25] 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.

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

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

[28] 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.

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

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

[31] Robert Schaback. Limit problems for interpolation by analytic radial basis functions. *Journal of Computational and Applied Mathematics*, 212(2):127–149, 2008.

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

[33] 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.

[34] 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.

[35] Lloyd N Trefethen. *Approximation theory and approximation practice*, volume 128. Siam, 2013.

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

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

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

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

[40] 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.

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

[42] 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.

[43] Barbara Zwicknagl. Power series kernels. *Constructive Approximation*, 29(1):61–84, 2009.