Entrywise error bounds for low-rank approximations of kernel matrices

Alexander Modell
Department of Mathematics
Imperial College London, U.K.
a.modell@imperial.ac.uk

Abstract

In this paper, we derive *entrywise* error bounds for low-rank approximations of kernel matrices obtained using the truncated eigen-decomposition (or singular value decomposition). While this approximation is well-known to be optimal with respect to the spectral and Frobenius norm error, little is known about the statistical behaviour of individual entries. Our error bounds fill this gap. A key technical innovation is a delocalisation result for the eigenvectors of the kernel matrix corresponding to small eigenvalues, which takes inspiration from the field of Random Matrix Theory. Finally, we validate our theory with an empirical study of a collection of synthetic and real-world datasets.

1 Introduction

Low-rank approximations of kernel matrices play a central role in many areas of machine learning. Examples include kernel principal component analysis [Schölkopf et al., 1998], spectral clustering [Ng et al., 2001, Von Luxburg, 2007] and manifold learning [Roweis and Saul, 2000, Belkin and Niyogi, 2001, Coifman and Lafon, 2006], where they serve as a core component of the algorithms, and support vector machines [Cortes and Vapnik, 1995, Fine and Scheinberg, 2001] and Gaussian process regression [Williams and Rasmussen, 1995, Ferrari-Trecate et al., 1998] where they serve to dramatically speed up computation times.

In this paper, we derive entrywise error bounds for low-rank approximations of kernel matrices obtained using the truncated eigen-decomposition (or singular value decomposition). Entrywise error bounds are important for a number of reasons. The first is practical: in applications where individual errors carry a high cost, such as system control and healthcare, we should seek methods with low entrywise error. The second is theoretical: good entrywise error bounds can lead to improved analyses of learning algorithms which use them.

For this reason, a wealth of literature has emerged establishing entrywise error bounds for a variety of matrix estimation problems, such as covariance estimation [Fan et al., 2018, Abbe et al., 2022], matrix completion [Candes and Recht, 2012, Chi et al., 2019], phase synchronisation [Zhong and Boumal, 2018, Ma et al., 2018], reinforcement learning [Stojanovic et al., 2023], community detection [Balakrishnan et al., 2011, Lyzinski et al., 2014, Eldridge et al., 2018, Lei, 2019, Mao et al., 2021] and graph inference [Cape et al., 2019, Rubin-Delanchy et al., 2022] to name a few.

1.1 Contributions

- Our main result (Theorem 1) is an entrywise error bound for the low-rank approximation of a kernel matrix. Under regularity conditions, we find that for kernels with polynomial eigenvalue decay, $\lambda_i = \mathcal{O}(i^{-\alpha})$, we require a polynomial-rank approximation, $d = \Omega(n^{1/\alpha})$,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4027

to achieve entrywise consistency. For kernels with exponential eigenvalue decay, $\lambda_i = \mathcal{O}(e^{\beta i^\gamma})$, we require a (poly)log-rank approximation, $d > \log^{1/\gamma}(n^{1/\beta})$.

- The main technical contribution of this paper is to establish a delocalisation result for the eigenvectors of the kernel matrix corresponding to small eigenvalues (Theorem 3), the proof of which draws on ideas from the Random Matrix Theory literature. To our knowledge, this is the first such result for a random matrix with non-zero mean and dependent entries.
- Along the way, we prove a novel concentration inequality for the distance between a random vector (with a potentially non-zero mean) and a subspace (Lemma 1), which may be of independent interest.
- We complement our theory with an empirical study on the entrywise errors of low-rank approximations of the kernel matrices on a collection of synthetic and real datasets.

1.2 Related work

Some complementary results to ours use the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1982] to bound the entrywise error of low-rank matrix approximations obtained via random projections [Srebro and Shraibman, 2005, Alon et al., 2013, Udell and Townsend, 2019, Budzinskiy, 2024a,b]. In Section 3.2 we discuss these results in more detail and compare them with ours.

Our proof strategy draws heavily on ideas from the Random Matrix Theory literature, where delocalisation results have been established for certain classes of zero-mean random matrices with independent entries [Erdős et al., 2009b,a, Tao and Vu, 2011, Rudelson and Vershynin, 2015, Vu and Wang, 2015]. In addition, our result is made possible by recent *relative* eigenvalue concentration bounds for kernel matrices [Braun, 2006, Valdivia, 2018, Barzilai and Shamir, 2023], which improve upon classical *absolute* concentration bounds [Rosasco et al., 2010] which would not provide sufficient control for our purposes.

1.3 Big-$\mathcal{O}$ notation and frequent events

We use the standard big-$\mathcal{O}$ notation where $a_n = \mathcal{O}(b_n)$ (resp. $a_n = \Omega(b_n)$) means that for sufficiently large n, $a_n \leq Cb_n$ (resp. $a_n \geq Cb_n$) for some constant C which doesn't depend on the parameters of the problem. We will occasionally write $a_n \lesssim b_n$ to mean that $a_n = \mathcal{O}(b_n)$.

In addition, we say that an event E_n holds *with overwhelming probability* if for *every* $c > 0$, $\mathbb{P}(E_n) \geq 1 - \mathcal{O}(n^{-c})$, where the hidden constant is allowed to depend on c.

2 Setup

We begin by describing the setup of our problem. We suppose that we observe n, p-dimensional data points $\{x_i\}_{i=1}^n$, which we assume were drawn i.i.d. from some probability distribution ρ, supported on a set $\mathcal{X}$. Given a symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we construct the $n \times n$ kernel matrix K, with entries

$$K(i,j) := k(x_i, x_j).$$

We will assume throughout that the kernel is positive-definite, continuous and bounded. Let $\widehat{K}_d$ denote the "best" rank-d approximation of K, in the sense that $\widehat{K}_d$ satisfies

$$\widehat{K}_d := \underset{K' : \mathrm{rank}(K') = d}{\arg\min} \|K - K'\|_\xi , \tag{1}$$

where $\|\cdot\|_\xi$ is a rotation-invariant norm, such as the spectral or Frobenius norm. Then by the Eckart-Young-Mirsky theorem [Eckart and Young, 1936, Mirsky, 1960], $\widehat{K}_d$ is given by the truncated eigen-decomposition of K, i.e.

$$\widehat{K}_d = \sum_{i=1}^{d} \widehat{\lambda}_i \widehat{u}_i \widehat{u}_i^\top$$

where $\{\widehat{\lambda}_i\}_{i=1}^n$ are the eigenvalues of K (counting multiplicities) in decreasing order, and $\{\widehat{u}_i\}_{i=1}^n$ are corresponding eigenvectors.

We now introduce some population quantities which will form the basis of our theory. Let L^2_ρ denote the Hilbert space of real-valued square integrable functions with respect to ρ, with the inner product defined as $\langle f, g \rangle_\rho = \int f(x)g(x)\mathrm{d}\rho(x)$. We define the integral operator $\mathcal{K} : L^2_\rho \to L^2_\rho$ by

$$(\mathcal{K}f)(x) = \int k(x,y)f(y)\mathrm{d}\rho(y)$$

which is the infinite sample limit of $\frac{1}{n}K$. The operator $\mathcal{K}$ is self-adjoint and compact [Hirsch and Lacombe, 1999], so by the spectral theorem for compact operators, there exists a sequence of eigenvalues $\{\lambda_i\}_{i=1}^\infty$ (counting multiplicities) in decreasing order, with corresponding eigenfunctions $\{u_i\}_{i=1}^\infty$ which are orthonormal in L^2_ρ, such that

$$\mathcal{K}u_i = \lambda_i u_i.$$

Moreover, by the classical Mercer's theorem [Mercer, 1909, Steinwart and Scovel, 2012], the kernel k can be decomposed into

$$k(x,y) = \sum_{i=1}^\infty \lambda_i u_i(x) u_i(y) \tag{2}$$

where the series converges absolutely and uniformly in x, y.

3 Entrywise error bounds

This section is devoted to our main theoretical result. We begin by discussing our assumptions, before presenting our main theorem and giving some special cases of kernels which fit within our framework.

In our asymptotics, we will assume that k and ρ are fixed, and that the number of samples n goes to infinity. This places us in the so-called low-dimensional regime, in which the dimension p of the input space is considered fixed.

We shall assume that the eigenvalues of the kernel exhibit either polynomial decay, i.e. $\lambda_i = \mathcal{O}(i^{-\alpha})$ for some $\alpha > 1$, or (nearly) exponential decay, i.e. $\lambda_i = \mathcal{O}(e^{-\beta i^\gamma})$ for some $\beta > 0$ and $0 < \gamma \leq 1$. We will refer to these two hypotheses as (P) and (E) respectively. We also assume a corresponding hypothesis on the supremum-norm growth of the eigenfunctions. Under (P), we assume that $\|u_i\|_\infty = \mathcal{O}(i^r)$ with $\alpha > 2r + 1$, and under (E), we assume that $\|u_i\|_\infty = O(e^{si^\gamma})$ with $\beta > 2s$.

Our eigenvalue decay hypothesis is commonplace in the kernel literature [Braun, 2006, Ostrovskii and Rudi, 2019, Xu, 2018, Lei, 2021], and can be related to the smoothness of the kernel. For example, the decay of the eigenvalues is directly implied by a Hölder or Sobolev-type smoothness hypothesis on the kernel (see, for example, Nicaise [2000], Belkin [2018], Section 2.2 of Xu [2018], Section 5 of Valdivia [2018], Scetbon and Harchaoui [2021] and Proposition 2 in this paper). We don't consider a finite-rank (say, D) hypothesis, since in this case the maximum entrywise error is trivially zero whenever $d \geq D$.

Our hypothesis on the supremum norm of the eigenfunctions is necessary to control the deviation of the sample eigenvectors from their corresponding population eigenfunctions, and is a requirement of eigenvalue bounds we employ. In the literature, it is common to see much stronger assumptions, such as uniformly bounded eigenfunctions [Williamson et al., 2001, Lafferty et al., 2005, Braun, 2006], which do not hold for many commonly-used kernels (see Mendelson and Neeman [2010], Steinwart and Scovel [2012], Zhou [2002] and Barzilai and Shamir [2023] for discussion). This assumption is reminiscent of the *incoherence* assumption [Candes and Recht, 2012] in the low-rank matrix estimation literature — a supremum norm bound on population eigenvectors — which governs the hardness of many compressed sensing and eigenvector estimation problems [Candès and Tao, 2010, Keshavan et al., 2010, Chi et al., 2019, Abbe et al., 2020, Chen et al., 2021].

In addition, we introduce a regularity hypothesis, which we will refer to as (R), which relates to the following two quantities:

$$\Delta_i = \max_{j \geq i} \{\lambda_j - \lambda_{j+1}\}$$

which measures the largest eigengap after a certain point in the spectrum, and

$$\Gamma_i = \sum_{j=i+1}^\infty \left(\int u_j(x)\mathrm{d}\rho(x) \right)^2$$

 https://doi.org/10.52202/079017-4027

	λ_i	$\|u_i\|_\infty$	
(P)	$\mathcal{O}\left(i^{-\alpha}\right)$	$\mathcal{O}\left(i^{r}\right)$	$\alpha > 2r+1$
(E)	$\mathcal{O}\left(e^{-\beta i^{\gamma}}\right)$	$\mathcal{O}\left(e^{si^{\gamma}}\right)$	$\beta > 2s, 0 < \gamma \leq 1$

	Δ_i	Γ_i	
(R)	$\mathcal{O}(\lambda_i^a)$	$\mathcal{O}(\lambda_i^b)$	$1 \leq a < b/16$

Table 1: Summary of the hypotheses (P), (E) and (R).

which measures the squared residual after projecting the unit-norm constant function onto the first i eigenfunctions. Under (R), we assume that $\Delta_i = \Omega\left(\lambda_i^a\right)$ and $\Gamma_i = \mathcal{O}\left(\lambda_i^b\right)$ with $1 \leq a < b/16 \leq \infty$.

A sufficient condition for (R) to hold, is that the first eigenfunction is constant. This holds, for example, when k is a dot-product kernel and ρ is a uniform distribution on a hypersphere. In such scenarios, $\Gamma_i = 0$ for all $i \geq 1$ and it is not necessary to make any assumptions on the eigengap quantity Δ_i. We remark that (R) permits repeated eigenvalues in the spectrum of $\mathcal{K}$, which occur for many commonly-used kernels, but which are often precluded in the literature [Hall and Horowitz, 2007, Meister, 2011, Lei, 2014, 2021].

The hypotheses (P), (E) and (R) are summarised in Table 1. We are now ready to state our main theorem.

Theorem 1. *Suppose that k is a symmetric, positive-definite, continuous and bounded kernel and ρ is a probability measure which satisfy (R) and one of either (P) or (E). If the hypothesis (P) holds and $d = \Omega\left(n^{1/\alpha}\right)$, then*

$$\left\|\widehat{K}_d - K\right\|_{\max} = \mathcal{O}\left(n^{-\frac{\alpha-1}{\alpha}} \log(n)\right) \tag{3}$$

with overwhelming probability. If the hypothesis (E) holds and $d > \log^{1/\gamma}(n^{1/\beta})$, then

$$\left\|\widehat{K}_d - K\right\|_{\max} = \mathcal{O}\left(n^{-1}\right)$$

with overwhelming probability.

3.1 Special cases

In this section, we provide some examples of kernels which satisfy the assumptions of Theorem 1. Proofs of the propositions in this section are given in Section A of the appendix. We start with a canonical example of a radial basis kernel.

Proposition 1. *Suppose $k(x,y) = \exp\left(-\|x-y\|^2/2\omega^2\right)$ is a* radial basis kernel*, and $\rho \sim \mathcal{N}(0, \sigma^2 I_p)$ is a isotropic Gaussian distribution on $\mathbb{R}^p$. Then the hypotheses (E) and (R) are satisfied with*

$$\beta = \log\left(\frac{1+\upsilon+\sqrt{1+2\upsilon}}{\upsilon}\right), \qquad \gamma = 1$$

where $\upsilon := 2\sigma^2/\omega^2$.

For this example, the eigenvalues and eigenfunctions were explicitly calculated in Zhu et al. [1997] (see also Shi et al. [2008] and Shi et al. [2009]), and we are able to verify the assumptions by direct calculation.

For our second example, we consider the case that ρ is the uniform distribution on a hypersphere $\mathbb{S}^{p-1}$, and k is a dot-product kernel. In this setting, we are able to replace our assumptions with a smoothness hypothesis on the kernel. Note that this class of kernels includes those which are functions of Euclidean distance, since on the sphere we have the identity $\|x-y\|^2 = 2 - 2\langle x, y\rangle$.

Proposition 2. *Suppose that*

$$k(x,y) = f(\langle x,y\rangle) \equiv \sum_{i=0}^{\infty} b_i \left(\langle x,y\rangle\right)^i$$

is a dot-product kernel *and ρ is the uniform distribution on the hypersphere $\mathbb{S}^{p-1}$ with $p \geq 3$. If there exists $a > (p^2-4p+5)/2$ such that $b_i = \mathcal{O}(i^{-a})$, then (P) and (R) are satisfied with*

$$\alpha = \frac{2a+p-3}{p-2}.$$

Alternatively, if there exists $0 < r < 1$ such that $b_i = \mathcal{O}(r^i)$, then (E) and (R) are satisfied with

$$\beta = \frac{(p-1)!}{C}\log(1/r), \qquad \gamma = \frac{1}{p-1}$$

for some universal constant $C > 0$.

In this example, the eigenfunctions posses the property that they do not depend on the choice of kernel, and are made up of *spherical harmonics* [Smola et al., 2000]. In particular, the first eigenfunction is constant, and therefore (R) is satisfied automatically. The eigenvalue bounds are derived in Scetbon and Harchaoui [2021], and we make use of a supremum norm bound for spherical harmonics in Minh et al. [2006].

3.2 Comparison with random projections and the Johnson-Lindenstrauss lemma

We pause here to consider how our entrywise bounds compare with existing bounds in the literature for low-rank matrix obtained via random projections [Srebro and Shraibman, 2005, Alon et al., 2013, Udell and Townsend, 2019, Budzinskiy, 2024a,b].

For an $n \times n$ symmetric, positive semi-definite matrix M with bounded entries, the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1982] can be used to show the existence of a rank-d approximation $\widehat{M}_d$ whose entrywise error is bounded by ε when $d = \Omega(\varepsilon^{-2}\log(n))$.

The proof is via a probabilistic construction. Let X be an $n \times n$ matrix such that $M = XX^\top$, and for some $d \leq n$, let R be an $n \times d$ matrix with i.i.d. entries from $\mathcal{N}(0, 1/d)$. Then, the randomised low-rank approximation

$$\widehat{M}_d := XRR^\top X^\top \tag{4}$$

achieves the desired bound with high probability. Here, we state a adaptation of Theorem 1.4 of Alon et al. [2013] which makes the probabilistic construction from the proof explicit.

Theorem 2. *Let M be an $n \times n$ positive semi-definite matrix with bounded entries, and $\widehat{M}_d$ be a randomised rank-d approximation of M described in* (4)*. Then*

$$\left\|\widehat{M}_d - M\right\|_{\max} = \mathcal{O}\left(\sqrt{\frac{\log(n)}{d}}\right)$$

with overwhelming probability.

To obtain a polynomial entrywise error rate, i.e. $\mathcal{O}(n^{-c})$ for some $c > 0$, with Theorem 2, requires the rank d to be polynomial in n. In contrast, under our hypothesis (E), we are able to obtain a polynomial entrywise error rate using a spectral low-rank approximation with only poly-logarithmic rank. In addition, while our entrywise error bounds are $o(n^{-1/2})$ for the cases we consider, this rate can never be achieved, regardless of the choice of rank d, by (4) with Theorem 2.

On the flip side, Theorem 2 holds for arbitrary positive semi-definite matrix with bounded entries, whereas our theorem only holds for kernel matrices satisfying the hypotheses in Table 1.

4 Proof of Theorem 1

In this section, we outline the proof of Theorem 1. Without loss of generality, we will assume that k is upper bounded by one. We cover the main details here, and defer some of the technical details to the appendix. By the Eckart-Young-Mirsky theorem, we have that

$$\begin{aligned}\left\|\widehat{K}_d - K\right\|_{\max} &:= \max_{1\leq i,j\leq n}\left|\widehat{K}_d(i,j) - K(i,j)\right| = \left|\max_{1\leq i,j\leq n}\sum_{l=d+1}^{n}\widehat{\lambda}_l\widehat{u}_l(i)\widehat{u}_l(j)\right| \\ &\leq \sum_{l=d+1}^{n}\left|\widehat{\lambda}_l\right|\cdot\max_{d<l\leq n}\left\|\widehat{u}_l\right\|_\infty^2.\end{aligned}$$

Using a concentration bound due to Valdivia [2018], we are able to show that

$$\sum_{l=d+1}^{n}\left|\widehat{\lambda}_l\right| = \begin{cases}\mathcal{O}\left(n^{1/\alpha}\log(n)\right) & \text{under (P) with } d = \Omega\left(n^{1/\alpha}\right) \\ \mathcal{O}\left(1\right) & \text{under (E) with } d > \log^{1/\gamma}(n^{1/\beta}).\end{cases} \tag{5}$$

 https://doi.org/10.52202/079017-4027

with overwhelming probability, the details of which are given in Section B of the appendix. Then, the proof boils down to showing the following result, which we state as an independent theorem.

Theorem 3. *Assume the setting of Theorem 1, then simultaneously for all $d+1 \leq l \leq n$,*

$$\|\widehat{u}_l\|_\infty = \mathcal{O}\left(n^{-1/2}\right) \tag{6}$$

with overwhelming probability.

When a unit eigenvector satisfies (6) (up to log factors), it is said to be *completely delocalised*. There is a now expansive literature in the field of Random Matrix Theory proving the delocalisation of the eigenvectors of certain mean-zero random matrices with independent entries [Erdős et al., 2009b,a, Tao and Vu, 2011, Rudelson and Vershynin, 2015, Vu and Wang, 2015]. Theorem 3 may be of independent interest since to our knowledge, it is the first eigenvector delocalisation result for a random matrix with non-zero mean and dependent entries.

To prove Theorem 3, we take inspiration from a proof strategy employed in Tao and Vu [2011] (see also Erdős et al. [2009b]) which makes use of an identity relating the eigenvalues and eigenvectors of a matrix with that of its principal minor. The non-zero mean, and dependence between the entries of a kernel matrix present new challenges which require novel technical insights and tools and make up the bulk of our technical contribution.

Proof of Theorem 3. By symmetry and a union bound, to prove Theorem 3, it suffices to establish the bound for the first coordinate of $\widehat{u}_l$ for some an arbitrary index $d < l \leq n$. We shall let $\widetilde{K}$ denote the bottom right principal minor of K , that is the $n-1 \times n-1$ matrix such that

$$K = \begin{pmatrix} z & y^\top \\ y & \widetilde{K} \end{pmatrix}$$

where $z = k(x_1, x_1)$ and $y = (k(x_1, x_2), \ldots, k(x_1, x_n))^\top$. We will denote the ordered eigenvalues and corresponding eigenvectors of $\widetilde{K}$ by $\{\widetilde{\lambda}_l\}_{l=1}^{n-1}$ and $\{\widetilde{u}_l\}_{l=1}^{n-1}$ respectively. By Lemma 41 of Tao and Vu [2011], we have the following remarkable identity:

$$\widehat{u}_l(1)^2 = \frac{1}{1 + \sum_{j=1}^{n-1} \left(\widetilde{\lambda}_j - \widehat{\lambda}_i\right)^{-2} \left(\widetilde{u}_j^\top y\right)^2}. \tag{7}$$

In addition, Cauchy's interlacing theorem tells us that the eigenvalues of K and $\widetilde{K}$ interlace, i.e. $\widehat{\lambda}_i \leq \widetilde{\lambda}_i \leq \widehat{\lambda}_{i+1}$ for all $1 \leq i \leq n-1$. By (5) we have that $|\widehat{\lambda}_i| = \mathcal{O}(1)$ for all $d+1 \leq i \leq n$ with overwhelming probability and so by Cauchy's interlacing theorem, we can find a set of indices $J \subset \{d+1, \ldots, n-1\}$ with $|J| \geq (n-d)/2$ such that $|\widetilde{\lambda}_j - \widehat{\lambda}_i| = \mathcal{O}(1)$ for all $j \in J$. Combining this observation with (7), we have that

$$\sum_{j=1}^{n-1} \left(\widetilde{\lambda}_j - \widehat{\lambda}_i\right)^{-2} \left(\widetilde{u}_j^\top y\right)^2 \geq \sum_{j \in J} \left(\widetilde{\lambda}_j - \widehat{\lambda}_i\right)^{-2} \left(\widetilde{u}_j^\top y\right)^2 \gtrsim \|\pi_J(y)\|^2. \tag{8}$$

where π_J denotes the orthogonal projection onto the subspace spanned by $\{\widetilde{u}_j\}_{j \in J}$. So, to establish (6), it suffices to show that

$$\|\pi_J(y)\|^2 = \Omega(n) \tag{9}$$

with overwhelming probability. We condition on x_1, so that y is a vector of independent random variables and denote its conditional mean by $\bar{y}$, which is a constant vector whose entries are less than one. In addition, each entry of y has common conditional variance which we denote by $\sigma^2 = \mathbb{E}_{x \sim F}\{k^2(x_1, x)\}$.

To obtain the lower bound (9), we prove a novel concentration inequality for the distance between a random vector and a subspace, which may be of independent interest. Our lemma generalises a similar result in Tao and Vu [2011, Lemma 43] which holds only for random vectors with zero mean and unit variance. The proof is provided in Section C of the appendix.

Lemma 1. *Let $y \in \mathbb{R}^n$ be a random vector with mean $\bar{y} := \mathbb{E}y$ whose entries are independent, have common variance σ^2 and are bounded in $[0,1]$ almost surely. Let H be a subspace of dimension $q \geq 64/\sigma^2$ and π_H the orthogonal projection onto H. If H is such that $\|\pi_H(\bar{y})\| \leq 2(\sigma^2 q)^{1/4}$, then for any $t \geq 8$*

$$\mathbb{P}\left(\left|\|\pi_H(y)\| - \sigma q^{1/2}\right| \geq t\right) \leq 4\exp\left(-t^2/32\right).$$

In particular, one has

$$\|\pi_H(y)\| = \sigma q^{1/2} + \mathcal{O}\left(\log^{1/2}(n)\right) \tag{10}$$

with overwhelming probability.

Returning to the main thread, we claim for the moment that $\|\pi_J(\bar{y})\| \leq 2|J|^{1/4}$. Then, by Lemma 1 we have that, conditional on x_1,

$$\|\pi_J(y)\|^2 \gtrsim |J| \geq (n-d)/2 = \Omega(n)$$

with overwhelming probability. This holds for all $x_1 \in \mathcal{X}$ and therefore establishes (9). To complete the proof, then, it remains to prove our claim, and it is here where we require the regularity hypothesis (R). The proof of the claim is quite involved, so we defer the details to Section D of the appendix, given which, the proof of Theorem 3 is complete.

□

5 Experiments

5.1 Datasets and setup

To see how our theory translates into practice, we examine the maximum entrywise error of the low-rank approximations of kernel matrices derived from a synthetic dataset and a collection of five real-world data sets, which are summarised in the following table[1].

Dataset	Description	# Instances	# Dimensions
GMM	simulated data from a Gaussian mixture model	1,000	10
Abalone	physical measurements of Abalone trees	4,177	7
Wine Quality	physicochemical measurements of wines	6,497	11
MNIST	handwritten digits	10,000	784
20 Newsgroups	tf-idf vectors from newsgroup messages	11,314	21,108
Zebrafish	gene expression in zebrafish embryo cells	6,346	5,434

Additional details about the dataset are provided in Section E of the appendix.

For the purpose of our experiment, we employ kernels in the class of *Matérn kernels*, of the form

$$k_\nu(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|x-y\|}{\omega}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{\|x-y\|}{\omega}\right)$$

where Γ denotes the gamma function, and K_ν is the modified Bessel function of the second kind. The class of Matérn kernels is a generalisation of the radial basis kernel, with an additional parameter ν which governs the smoothness of the resulting kernel. When $\nu = 1/2$, we obtain the non-differentiable exponential kernel, and in the $\nu \to \infty$ limit, we obtain the infinitely-differentiable radial basis kernel. For the intermediate values $\nu = 3/2$ and $\nu = 5/2$, we obtain, respectively, once and twice-differentiable functions.

The optimal choice of the bandwidth parameter is problem-dependent, and in supervised settings is typically chosen using cross-validation. In unsupervised settings, it is necessary to rely on heuristics,

[1] Code to reproduce the experiments in this section can be found at `https://gist.github.com/alexandermodell/b16b0b29b6d0a340a23dab79219133f2`.

 https://doi.org/10.52202/079017-4027

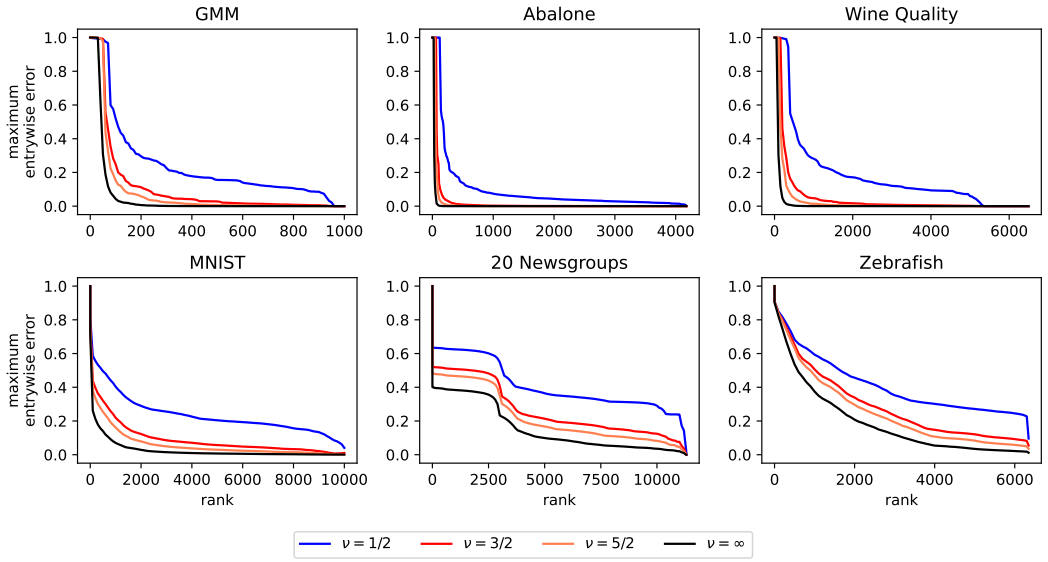

Figure 1: The maximum entrywise error against rank for low-rank approximations of kernel matrices constructed from a collection of datasets. The kernel matrices are constructed using Matérn kernels with a range of smoothness parameters, each of which is represented by a line in each plot. Details of the experiment are provided in Section 5.

and for this experiment, we use the popular *median heuristic* [Flaxman et al., 2016, Mooij et al., 2016, Mu et al., 2016, Garreau et al., 2017], which has been shown to perform well in practice.

For each dataset, we construct four kernel matrices using Matérn kernels with smoothness parameters $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \infty$, each time selecting the bandwidth using the median heuristic. For each kernel, we compute the best rank-d low-rank approximation of the kernel matrix using the `svds` function in the SciPy library for Python [Virtanen et al., 2020]. We do this for a range of ranks d from 1 to n, where n is the number of instances in the dataset, and record the entrywise errors.

5.2 Interpretation of the results

Figure 1 shows the maximum entrywise errors for each dataset and kernel. For comparison, the Frobenius norm errors are plotted in Figure 2 in Section E of the appendix.

As predicted by our theory, for the four "low-dimensional" datasets, *GMM*, *Abalone*, *Wine Quality* and *MNIST*, the maximum entrywise decays rapidly as we increase the rank of the approximation, with the exception of the highly non-smooth $\nu = \frac{1}{2}$ kernel, for which the maximum entrywise error decays much more slowly. In addition, the decay rates of the maximum entrywise error are in order of the smoothness of the kernels.

For the "high-dimensional" datasets, *20 Newsgroups* and *Zebrafish*, a different story emerges. Even for the smooth radial basis kernel ($\nu = \infty$), the maximum entrywise error decays very slowly. This would suggests that our theory does potentially *not* carry over to the high-dimensional setting, and that caution should be taken when employing low-rank approximations for such data. Interestingly, the *20 Newsgroups* dataset exhibits a sharp drop in maximum entrywise error between $d = 2500$ and $d = 3000$ which *is not* seen in the decay of the Frobenius norm error (Figure 2 in Section E).

6 Limitations and open problems

To conclude, we discuss some of the limitations of our theory, as well as some of the open problems.

6.1 Limitations of our theory

Positive semi-definite kernels. One significant limitation of our theory is the assumption that the kernel is positive semi-definite and continuous. This condition is known as Mercer's condition in the literature and ensures that the spectral decomposition of the kernel (2) converges uniformly, however we don't actually require such a strong notion of convergence for our theory. Valdivia [2018, Lemma 22] show that the decomposition converges *almost surely* under a much weaker condition which is implied by our hypotheses (P) and (E). The only other places we need this assumption is to make use of results in Rosasco et al. [2010] and Tang et al. [2013]. These results make heavy use of reproducing kernel Hilbert space technology though it seems plausible that they could be generalised to the indefinite setting using the framework of Krein spaces [Ong et al., 2004, Lei, 2021].

Low-dimensional setting. In our asymptotics, we explicitly assume that the dimension of the input space remains fixed as the number of sample increases, which places us in the so-called low-dimensional setting. We do not consider the high-dimensional setting, however our empirical experiments suggest that our conclusions may not carry over.

Verification of the assumptions. While there is a established literature studying the eigenvalue decay of kernels under general probability measures [Kühn, 1987, Cobos and Kühn, 1990, Ferreira and Menegatto, 2013, Belkin, 2018, Li et al., 2024], except in very specialised settings (such as Propositions 1 and 2), control of the eigenfunctions is typically out of reach. This makes verifying the assumptions of our theory under general probability distributions quite challenging. This is a widespread limitation of many theoretical analyses in the kernel literature, and for an extended discussion, we refer the reader to Barzilai and Shamir [2023].

6.2 Open problems

Randomised low-rank approximations. While the truncated spectral decomposition provides the "ideal" low-rank approximation, it requires computing the whole kernel matrix which can be prohibitive for very large datasets. Randomised low-rank approximations, such as the *randomised SVD* [Halko et al., 2011], the *Nyström* method [Williams and Seeger, 2000, Drineas et al., 2005] and *random Fourier features* [Rahimi and Recht, 2007, 2008], have emerged as efficient alternatives, and there is an extensive body of literature examining their statistical performance [Drineas et al., 2005, Rahimi and Recht, 2007, Belabbas and Wolfe, 2009, Boutsidis et al., 2009, Kumar et al., 2009a,b, Gittens, 2011, Gittens and Mahoney, 2013, Altschuler et al., 2016, Derezinski et al., 2020]. However, their primary focus is on classical error metrics such as the spectral and Frobenius norm errors and an entrywise analysis would presumably provide greater insights into these approximations, particularly given recently observed multiple-descent phenomena [Derezinski et al., 2020].

Lower bounds. At present, it is unclear whether the bounds we obtain are tight, or indeed whether the truncated spectral decomposition itself is optimal with respect the the entrywise error. An interesting direction for future research would be to investigate lower bounds to understand the fundamental limits of this problem.

Acknowledgements

The author thanks Nick Whiteley, Yanbo Tang and Mahmoud Khabou for helpful discussions and Annie Gray for providing code to preprocess the *20 Newsgroups* and *Zebrafish* datasets.

This work was supported by the Engineering and Physical Sciences Research Council [grant EP/X002195/1].

References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452, 2020.

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

 https://doi.org/10.52202/079017-4027

Noga Alon, Troy Lee, Adi Shraibman, and Santosh Vempala. The approximate rank of a matrix and its algorithmic applications: approximate rank. In *Proceedings of the forty-fifth annual ACM Symposium on Theory of Computing*, pages 675–684, 2013.

Jason Altschuler, Aditya Bhaskara, Gang Fu, Vahab Mirrokni, Afshin Rostamizadeh, and Morteza Zadimoghaddam. Greedy column subset selection: New bounds and distributed algorithms. In *International Conference on Machine Learning*, pages 2539–2548. PMLR, 2016.

Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. *Advances in Neural Information Processing Systems*, 24, 2011.

Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. *arXiv preprint arXiv:2312.15995*, 2023.

Mohamed-Ali Belabbas and Patrick J Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361. PMLR, 2018.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14, 2001.

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 968–977. SIAM, 2009.

Mikio L Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *The Journal of Machine Learning Research*, 7:2303–2328, 2006.

Stanislav Budzinskiy. On the distance to low-rank matrices in the maximum norm. *Linear Algebra and its Applications*, 688:44–58, 2024a.

Stanislav Budzinskiy. Entrywise tensor-train approximation of large tensors via random embeddings. *arXiv preprint arXiv:2403.11768*, 2024b.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Fernando Cobos and Thomas Kühn. Eigenvalues of integral operators with positive definite kernels satisfying integrated hölder conditions over metric compacta. *Journal of Approximation Theory*, 63(1):39–55, 1990.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Michal Derezinski, Rajiv Khanna, and Michael W Mahoney. Improved guarantees and a multiple-descent curve for column subset selection and the nystrom method. *Advances in Neural Information Processing Systems*, 33:4953–4964, 2020.

Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(12), 2005.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic learning theory*, pages 321–358. PMLR, 2018.

László Erdős, Benjamin Schlein, and Horng-Tzer Yau. Local semicircle law and complete delocalization for wigner random matrices. *Communications in Mathematical Physics*, 287(2):641–655, 2009a.

László Erdős, Benjamin Schlein, and Horng-Tzer Yau. Semicircle law on short scales and delocalization of eigenvectors for wigner random matrices. 2009b.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

Gregory E Fasshauer and Michael J McCourt. Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

Giancarlo Ferrari-Trecate, Christopher Williams, and Manfred Opper. Finite-dimensional approximation of gaussian processes. *Advances in Neural Information Processing Systems*, 11, 1998.

JC Ferreira and VA3128739 Menegatto. Eigenvalue decay rates for positive integral operators. *Annali di Matematica Pura ed Applicata*, 192(6):1025–1041, 2013.

Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. *arXiv preprint arXiv:1603.02160*, 2016.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Alex Gittens. The spectral norm error of the naive nystrom extension. *arXiv preprint arXiv:1110.5305*, 2011.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pages 567–575. PMLR, 2013.

I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 8 edition, 2014. ISBN 978-0123849335.

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

Peter Hall and Joel L Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 2007.

 https://doi.org/10.52202/079017-4027

Francis Hirsch and Gilles Lacombe. *Elements of Functional Analysis*. Springer, 1999.

Jack Indritz. An inequality for hermite polynomials. *Proceedings of the American Mathematical Society*, 12(6):981–983, 1961.

William Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Conference in Modern Analysis and Probability*, 26:189–206, 01 1982.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

Thomas Kühn. Eigenvalues of integral operators with smooth positive definite kernels. *Archiv der Mathematik*, 49:525–534, 1987.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble nystrom method. *Advances in Neural Information Processing Systems*, 22, 2009a.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 553–560, 2009b.

John Lafferty, Guy Lebanon, and Tommi Jaakkola. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(1), 2005.

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

Michel Ledoux. *The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs*. Number 89. American Mathematical Soc., 2001.

Jing Lei. Adaptive global testing for functional linear models. *Journal of the American Statistical Association*, 109(506):624–634, 2014.

Jing Lei. Network representation using graph root distributions. *The Annals of Statistics*, 2021.

Lihua Lei. Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*, 2019.

Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

Vince Lyzinski, Daniel L Sussman, Minh Tang, Avanti Athreya, and Carey E Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Statist.*, 2014.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 116(536):1928–1940, 2021.

Alexander Meister. Asymptotic equivalence of functional linear regression and a white noise inverse problem. *The Annals of Statistics*, pages 1471–1495, 2011.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 2010.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A*, 209(441-458): 415–446, 1909.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Krikamol Mu, Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, et al. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41, 2016.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, 2001.

Serge Nicaise. Jacobi polynomials, weighted sobolev spaces and approximation results of some singularities. *Mathematische Nachrichten*, 213(1):117–140, 2000.

Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J Smola. Learning with non-positive kernels. In *Proceedings of the twenty-first International Conference on Machine Learning*, page 81, 2004.

Dmitrii M Ostrovskii and Alessandro Rudi. Affine invariant covariance estimation for heavy-tailed distributions. In *Conference on Learning Theory*, pages 2531–2550. PMLR, 2019.

David Pollard. Empirical processes: theory and applications. IMS, 1990.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1446–1473, 2022.

Mark Rudelson and Roman Vershynin. Delocalization of eigenvectors of random matrices with independent entries. *Duke Math. J.*, 2015.

Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 3394–3402. PMLR, 2021.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In *Proceedings of the 25th International Conference on Machine Learning*, pages 936–943, 2008.

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pages 3960–3984, 2009.

Alex Smola, Zoltán Ovári, and Robert C Williamson. Regularization with dot-product kernels. *Advances in Neural Information Processing Systems*, 13, 2000.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

Stefan Stojanovic, Yassir Jedra, and Alexandre Proutiere. Spectral entry-wise matrix estimation for low-rank reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 77056–77070, 2023.

Michel Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1 – 34, 1996.

Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406 – 1430, 2013.

Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127 – 204, 2011. doi: 10.1007/s11511-011-0061-3.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Ernesto Araya Valdivia. Relative concentration bounds for the spectrum of kernel matrices. *arXiv preprint arXiv:1812.02108*, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms*, 47(4):792–821, 2015.

Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8, 1995.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13, 2000.

Robert C Williamson, Alexander J Smola, and Bernhard Scholkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE transactions on Information Theory*, 47(6):2516–2532, 2001.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, 2018.

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Huaiyu Zhu, Christopher KI Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. 1997.

Appendix

A Proof of Propositions 1 and 2

For notational simplicity, in this section we will assume in this section that the eigenvalues and eigenfunctions are indexed from 0 rather than 1.

A.1 Proof of Proposition 1

We will begin by reducing the problem of verifying our assumptions under ρ to verifying them under the probability measure associated with the univariate Gaussian distribution $\mathcal{N}(0, \sigma^2)$, which we will denote by μ.

Let $\underline{\mathcal{K}} : L^2_\rho \to L^2_\rho$ denote the integral operator associated with the kernel k and the measure ρ, and let $\{\underline{\lambda}_i\}$ denote its eigenvalues, arranged in descending order, and $\{\underline{u}_i\}$ denote their corresponding eigenfunctions. By the rotation invariance of both k and ρ, the operator $\underline{\mathcal{K}}$ may be written as the p-fold tensor product

$$\underline{\mathcal{K}} = \mathcal{K} \otimes \cdots \otimes \mathcal{K}$$

where $\mathcal{K} : L^2_\mu \to L^2_\mu$ denotes the integral operator associated with the kernel k and the univariate Gaussian measure μ. Let $\{\lambda_i\}$ denote its eigenvalues, arranged in descending order, and $\{u_i\}$ denote their corresponding eigenfunctions. Then, the eigenvalues and eigenfunctions of $\underline{\mathcal{K}}$ and $\mathcal{K}$ are related in the following way (see Shi et al. [2008] or Fasshauer and McCourt [2012]). For every i, there exists $i_1, \ldots, i_p$ satisfying $\sum_{j=1}^p i_j = i$ such that

$$\underline{\lambda}_i = \prod_{j=1}^p \lambda_{i_j} \qquad \text{and} \qquad \underline{u}_i(x) = \prod_{j=1}^p u_{i_j}(x^j) \tag{11}$$

for all $x = (x^1, \ldots, x^p)^\top \in \mathbb{R}^p$. Now suppose that $\lambda_i = \Theta\left(e^{-\beta i}\right)$, then

$$\underline{\lambda}_i = \prod_{j=1}^p \lambda_{i_j} \asymp \prod_{j=1}^p e^{-\beta i_j} = e^{-\beta \sum_j i_j} = e^{-\beta i},$$

and suppose that $\|u_i\|_\infty = \mathcal{O}(e^{si})$ for some $s < \beta/2$. Then

$$\|\underline{u}_i\|_\infty \leq \prod_{j=1}^p \|u_{i_j}\|_\infty \lesssim \prod_{j=1}^p e^{si_j} = e^{si}$$

Therefore to prove that (E) hold under ρ, it suffices to show that it holds under μ.

Shi et al. [2008] provide an explicit formula for the eigenvalues and eigenfunctions of $\mathcal{K}$, which is a refinement of an earlier result of Zhu et al. [1997].

Let $\upsilon := 2\sigma^2/\omega^2$ and let $H_i(x)$ be the ith order Hermite polynomial. Then the eigenvalues and eigenfunctions of $\mathcal{K}$ are given by

$$\lambda_i = \sqrt{\frac{2}{1+\upsilon+\sqrt{1+2\upsilon}}}\left(\frac{\upsilon}{1+\upsilon+\sqrt{1+2\upsilon}}\right)^i$$

$$u_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^i i!}} \exp\left(-\frac{x^2}{2\sigma^2}\frac{\sqrt{1+2\beta}-1}{2}\right) H_i\left(\left(\frac{1}{4}+\frac{\beta}{2}\right)^{1/4}\frac{x}{\sigma}\right).$$

Therefore, we have $\lambda_i = C_1 e^{-\beta i}$ where

$$\beta = \log\left(\frac{1+\upsilon+\sqrt{1+2\upsilon}}{\upsilon}\right).$$

We will now show that each u_i is uniformly bounded. By a change of variables, we can write u_i as

$$u_i(x) = \frac{C_2}{\sqrt{2^i i!}} e^{-y^2} H_i(y)$$

 https://doi.org/10.52202/079017-4027

for some $y \in \mathbb{R}$. On the other hand, we have the following inequality due to Indritz [1961]. For all $x \in \mathbb{R}$,

$$e^{-x^2} H_i(x) \leq 1.09\sqrt{2^i i!}.$$

Therefore $u_i(x) \leq 1.09 C_2$ for all x. We can use the fact that $H_i(x)$ is either odd or even to obtain an analogous lower bound. Therefore $\|u_i\|_\infty \leq 1.09 C_2$ for all i, so $s = 0$ and (E) holds.

We will now show that $\Gamma_i = 0$ for all $i \geq 1$ so that (R) holds with $b = +\infty$ and there is no requirement on the eigengaps Δ_i.

Expanding $\int u_i(x) \mathrm{d}\mu(x)$, collecting exponential terms and applying a change-of-variables, one can calculate that

$$\int u_i(x) \mathrm{d}\mu(x) = C_3 \int_{-\infty}^{+\infty} e^{-y^2} H_i(y) \mathrm{d}y.$$

It is a standard result that $e^{-y^2} H_i(y) = 0$ as long as $i \neq 0$ [Gradshteyn and Ryzhik, 2014], and therefore $\left| \int u_i(x) \mathrm{d}\mu(x) \right| = 0$ for all $i \geq 1$, and by (11), we have that $\Gamma_i = 0$ for all $i \geq 1$.

A.2 Proof of Proposition 2

For any $p \geq 3$, dot-product kernels with respect to the uniform measure on the sphere exhibit the spectral decomposition

$$k(x, y) = \sum_{l=0}^{\infty} \lambda_l^* \sum_{m=1}^{N_l} u_{l,m}^*(x) u_{l,m}^*(y)$$

where the eigenfunctions $\{u_{l,m}^*\}$ are the mth spherical harmonic of degree l, $N_l = \frac{2l+p-2}{l} \binom{l+p-3}{p-2} = \mathcal{O}(l^{p-2})$ is the number of harmonics of each degree, and $\{\lambda_l^*\}$ are the distinct eigenvalues [Smola et al., 2000].

The first spherical harmonic is a constant function, and therefore by the orthogonality of the eigenfunctions in L_ρ^2, $\int u_{l,m}^*(x) \mathrm{d}\rho(x) = 0$ for all $l \geq 1$, and therefore $\Gamma_i = 0$ for all $i \geq 1$. Therefore (R) holds with $s = +\infty$, and there are no requirements on the eigengaps.

In addition, Lemma 3 of Minh et al. [2006] shows that the supremum-norm of a spherical harmonic is upper bounded by

$$\|u_{l,m}^*\|_\infty \leq \sqrt{\frac{N_l}{|\mathbb{S}^{p-1}|}} = \mathcal{O}\left(i^{\frac{p-2}{2}}\right). \tag{12}$$

The eigenvalue decay rates are obtained from Propositions 2.3 and 2.4 of Scetbon and Harchaoui [2021], and given (12), the condition $a > (p^2 - 4p + 5)/2$ ensures that the conditions for (P) are met.

B Proof of Equation (5)

We begin this section with two upper bounds on polynomial and exponential series, which we prove in Section B.1, and which we will use throughout this proof.

Lemma 2. *Let $\alpha > 1$, $\beta > 0$ and $0 < \gamma \leq 1$ for fixed constants. Then the following upper bounds hold:*

$$\sum_{i=d+1}^{\infty} i^{-\alpha} = \mathcal{O}\left(d^{-\alpha+1}\right), \qquad \sum_{i=d+1}^{\infty} e^{-\beta i^\gamma} = \mathcal{O}\left(e^{-\beta d^\gamma} d^{1-\gamma}\right).$$

Throughout this proof, $\varepsilon > 0$ will denote some constant which may change from line to line, and even within lines.

To show equation (5), we first note that under (P) with $d = \Omega(n^{1/\alpha})$

$$\sum_{l=d+1}^{n} \lambda_l \lesssim \sum_{i=d+1}^{n} i^{-\alpha} \lesssim d^{-\alpha+1} = \mathcal{O}\left(n^{\frac{-\alpha-1}{\alpha}}\right)$$

where we have used Lemma 2. In addition, under (E) with $d > \log^{1/\gamma}(n^{1/\beta})$, we have that

$$\sum_{l=d+1}^{n} \lambda_l \lesssim \sum_{l=d+1}^{n} e^{-\beta l^\gamma} \lesssim e^{-\beta d^\gamma} d^{1-\gamma} = n^{-(1+\varepsilon)} \log^{(1-\gamma)/\gamma}(n^{1/\beta}) = \mathcal{O}(n^{-1})$$

where we have again used Lemma 2. Now, by the triangle inequality we have that

$$\frac{1}{n} \sum_{l=d+1}^{n} \left|\widehat{\lambda}_l\right| \leq \sum_{l=d+1}^{n} \lambda_l + \sum_{l=d+1}^{n} \left|\frac{\widehat{\lambda}_l}{n} - \lambda_l\right|$$

and therefore we are left to show that

$$\sum_{l=d+1}^{n} \left|\frac{\widehat{\lambda}_l}{n} - \lambda_l\right| = \begin{cases} \mathcal{O}\left(n^{-(\alpha-1)/\alpha} \log(n)\right) & \text{under (P) with } d = \Omega\left(n^{1/\alpha}\right); \\ \mathcal{O}\left(n^{-1}\right) & \text{under (E) with } d > \log^{1/\gamma}(n^{1/\beta}). \end{cases} \tag{13}$$

To bound (13), we employ a fine-grained concentration bound due to Valdivia [2018]. We begin with the polynomial hypothesis. The authors only consider the cases that α, r are natural numbers, since they draw a comparison between between these values and a Sobolev-type notion of regularity. Inspecting their proofs, they treat the cases $r = 0$ and $r \geq 1$ separately, however their proofs follow through in exactly the same way when the $r \geq 1$ case is replaced with $r > 0$, in order to cover all values of $\alpha > 2r + 1$, $r \geq 0$. For the $r > 0$ case, they derive the following result.

Lemma 3. *Suppose that the hypothesis (P) holds with $r > 0$. Then, with overwhelming probability*

$$\left|\frac{\widehat{\lambda}_i}{n} - \lambda_i\right| \lesssim B(i,n) \log(n)$$

where

$$B(i,n) = \begin{cases} i^{-\alpha + \frac{\alpha}{\alpha-1}(r+\frac{1}{2})} n^{-1/2} & \text{if } 1 \leq i \leq n^{\frac{\alpha-1}{\alpha}\frac{1}{2r+1}}; \\ i^{-\alpha+1+\frac{\alpha-1}{\alpha}(r+\frac{1}{2})} n^{-1/2} & \text{if } n^{\frac{\alpha-1}{\alpha}\frac{1}{2r+1}} \leq i \leq n^{\frac{1}{2r}}; \\ i^{-\alpha+r+1} n^{-1/2} & \text{if } n^{\frac{1}{2r}} \leq i \leq n; \end{cases}$$

Via some rearrangement we can show that

$$B(i,n) = \mathcal{O}\left(i^{-\alpha}\right), \qquad \text{if } 1 \leq i \leq n^{\frac{1}{2r}},$$

and by Lemma 2 we have that

$$\sum_{l=d+1}^{\lfloor n^{1/2r} \rfloor} \left|\frac{\widehat{\lambda}_l}{n} - \lambda_l\right| \lesssim \log(n) \sum_{l=d+1}^{\lfloor n^{1/2r} \rfloor} i^{-\alpha} \lesssim n^{-\frac{\alpha-1}{\alpha}} \log(n). \tag{14}$$

In addition, we have that

$$\begin{aligned} \sum_{l=\lceil n^{1/2r} \rceil}^{n} \left|\frac{\widehat{\lambda}_l}{n} - \lambda_l\right| &\lesssim n^{-1/2} \log(n) \sum_{l=\lceil n^{1/2r} \rceil}^{n} i^{-\alpha+r+1} \\ &\lesssim n^{-1/2} \log(n) \cdot \left(n^{\frac{1}{2r}}\right)^{-\alpha+r+2} \\ &\lesssim n^{-1} \log(n) \\ &\lesssim n^{-\frac{\alpha-1}{\alpha}} \end{aligned} \tag{15}$$

where we have used that $0 < r < (\alpha - 1)/2$. Combining (14) with (15) establishes (13) under (P) assuming $r > 0$. The case with $r = 0$ follows analogous fashion so we omit the details.

We now turn to the hypothesis (E). The authors only explicitly derive a result for the case that $\gamma = 1$, however following through their proof with Lemma 2 to hand, we obtain the following for the general case that $0 < \gamma \leq 1$.

Lemma 4. *Suppose that the hypothesis (E) holds with $s > 0$. Then, with overwhelming probability*

$$\left|\frac{\widehat{\lambda}_i}{n} - \lambda_i\right| \lesssim e^{(-\beta+\delta)i^\gamma} i^{1-\gamma} n^{-1/2} \log(n)$$

for all $1 \leq i \leq n$.

https://doi.org/10.52202/079017-4027

Therefore we have that

$$\begin{aligned}\sum_{i=d+1}^{\infty}\left|\frac{\widehat{\lambda}_i}{n}-\lambda_i\right| &\lesssim n^{-1/2}\log(n)\sum_{i=d+1}^{n} e^{(-\beta+s)i} i^{1-\gamma}\\ &\leq n^{-1/2}\log(n)\sum_{i=d+1}^{n} e^{-(\beta/2+\varepsilon)i^{\gamma}} i^{1-\gamma}\\ &\lesssim n^{-1/2}\log(n)\sum_{i=d+1}^{n} e^{-(\beta i^{\gamma}/2+\varepsilon)}\\ &\lesssim n^{-1/2}\log(n) n^{-(1/2+\varepsilon)}\\ &= \mathcal{O}(n^{-1})\end{aligned}$$

where in the second inequality we have used the assumption that $s < \beta/2$ and in the fourth we have used Lemma 2. The case for $s = 0$ follows similarly, so we omit the details. Then Equation (5) is established.

B.1 Proof of Lemma 2

To bound the polynomial series, we upper bound it by an integral as

$$\sum_{i=d+1}^{\infty} i^{-\alpha} \leq \int_{d}^{\infty} t^{-\alpha}\mathrm{d}t = \frac{d^{-\alpha+1}}{-\alpha+1} = \mathcal{O}\left(d^{-\alpha+1}\right).$$

To bound the exponential series, we again employ an integral approximation, and upper bound it as

$$\sum_{i=d+1}^{\infty} e^{-\beta i^{\gamma}} \leq \int_{d}^{\infty} e^{-\beta t^{\gamma}}\mathrm{d}t.$$

We then apply the substitution $u = \beta t^{\gamma}$ to obtain

$$\int_{d}^{\infty} e^{-\beta t^{\gamma}}\mathrm{d}t = \frac{1}{\gamma}\int_{\beta d^{\gamma}}^{\infty} e^{-u}u^{(1-\gamma)/\gamma}\mathrm{d}u = \frac{1}{\gamma}\Gamma\left(\frac{1}{\gamma}, \beta d^{\gamma}\right)$$

where Γ denotes the incomplete Gamma function. We can then use the fact that $\Gamma(s,x) \leq e^{-x}x^{s-1}$ for $s > 0$ to obtain the upper bound

$$\frac{1}{\gamma}\Gamma\left(\frac{1}{\gamma}, \beta d^{\gamma}\right) \leq \frac{1}{\gamma}e^{-\beta d^{\gamma}}\left(\beta d^{\gamma}\right)^{1/\gamma-1} = \frac{\beta}{\gamma}e^{\beta d^{\gamma}}d^{1-\gamma},$$

from which we can conclude that

$$\sum_{i=d+1}^{\infty} e^{-\beta i^{\gamma}} = \mathcal{O}\left(e^{\beta d^{\gamma}}d^{1-\gamma}\right),$$

as required.

C Proof of Lemma 1

The proof of Lemma 1 generalises the proof of Lemma 43 of Tao and Vu [2011]. We will make use of the following theorem due to Ledoux [2001] which is a corollary of Talagrand's inequality [Talagrand, 1996].

Theorem 4 (Talagrand's inequality)**.** *Let $y = (y_1, \ldots, y_n)^{\top}$ be a vector of independent random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex 1-Lipschitz function. Then, for all $t \geq 0$,*

$$\mathbb{P}\left(|f(y) - M(f)| \geq t\right) \leq 4\exp\left(-t^2/16\right)$$

where $M(f)$ denotes the median of f.

It is easy to verify that the map $y \to \|\pi_H(y)\|$ is convex and 1-Lipschitz, and so by Talagrand's inequality we have that

$$\mathbb{P}\left(|\|\pi_H(y)\| - M(\|\pi_H(y)\|)| \geq t\right) \leq 4\exp\left(-t^2/16\right). \tag{16}$$

For $t \geq 8$, we have that

$$4\exp\left(-(t-4)^2/16\right) \leq 4\exp\left(-t^2/32\right),$$

so to conclude the proof, it suffices to show that

$$|M(\|\pi_H(x)\|) - \sigma\sqrt{q}| \leq 4. \tag{17}$$

Let $\mathcal{E}_+$ denote the event that $\|\pi_H(x)\| \geq \sigma\sqrt{q} + 4$ and let $\mathcal{E}_-$ denote the event that $\|\pi_H(x)\| \leq \sigma\sqrt{q} - 4$. By the definition of a median, (17) is established if we can show that $\mathbb{P}(\mathcal{E}_+) < 1/2$ and $\mathbb{P}(\mathcal{E}_-) < 1/2$.

Let ε be the mean-zero random vector such that $y = \bar{y} + \varepsilon$, and let $P = (p_{ij})_{1 \leq i,j \leq n}$ be the orthogonal projection matrix onto H. We have that $\operatorname{tr} P^2 = \operatorname{tr} P = \sum_i p_{ii} = q$ and $|p_{ii}| \leq 1$. Furthermore

$$\|\pi_H(\varepsilon)\|^2 - \sigma^2 q = \sum_{1 \leq i,j \leq n} p_{ij}\varepsilon_i\varepsilon_j - \sigma^2 q = S_1 + S_2.$$

where $S_1 = \sum_{i=1}^n p_{ii}(\varepsilon_i^2 - \sigma^2)$ and $S_2 = \sum_{1 \leq i \neq j \leq n} p_{ij}\varepsilon_i\varepsilon_j$. We now upper bound the expectations of S_1^2 and S_2^2 which we will use later on for bounding the probabilities of $\mathcal{E}_+$ and $\mathcal{E}_-$ using Markov's inequality. Before we do, note that since $\varepsilon \in [-\bar{y}, 1-\bar{y}]$ almost surely, Popoviciu's inequality implies that $\sigma^2 \leq 1/4$. Therefore, we also have that $\mathbb{E}(\varepsilon_i^4) \leq \sigma^2$, and so

$$\begin{aligned}\mathbb{E}\left(S_1^2\right) &= \sum_{i,j=1}^n p_{ii}p_{jj}\mathbb{E}\left\{\left(\varepsilon_i^2 - \sigma^2\right)\left(\varepsilon_j^2 - \sigma^2\right)\right\} = \sum_{i=1}^n p_{ii}^2\mathbb{E}\left\{\left(\varepsilon_i^2 - \sigma^2\right)^2\right\} \\ &\leq \sum_{i=1}^n p_{ii}^2\left\{\mathbb{E}\varepsilon_i^4 - 2\sigma^2\mathbb{E}\left(\varepsilon_i^2\right) + (\sigma^2)^2\right\} \leq \sum_{i=1}^n p_{ii}^2\left\{\sigma^2 - 2(\sigma^2)^2 + (\sigma^2)^2\right\} \leq \sigma^2 q,\end{aligned} \tag{18}$$

where the second inequality follows from the independence of $\left(\varepsilon_i^2 - \sigma^2\right)$ and $\left(\varepsilon_j^2 - \sigma^2\right)$ for all $i \neq j$. In addition, we have that

$$\mathbb{E}\left(S_2^2\right) = \mathbb{E}\left(\sum_{i \neq j} p_{ij}\varepsilon_i\varepsilon_j\right)^2 = \sum_{i \neq j} p_{ij}^2\mathbb{E}(\varepsilon_i^2)\mathbb{E}(\varepsilon_j^2) \leq (\sigma^2)^2 q \leq \frac{\sigma^2 q}{4}. \tag{19}$$

To upper bound the probability of $\mathcal{E}_+$, we first observe that by assumption

$$\|\pi_H(y)\|^2 = \|\pi_H(\bar{y})\|^2 + \|\pi_H(\varepsilon)\|^2 \leq 4\sigma\sqrt{q} + \|\pi_H(\varepsilon)\|^2$$

and therefore we have that

$$\mathbb{P}(\mathcal{E}_+) = \mathbb{P}\left(\|\pi_H(y)\| \geq \sigma\sqrt{q} + 4\right) \leq \mathbb{P}\left(\|\pi_H(y)\|^2 \geq \sigma^2 q + 8\sigma\sqrt{q}\right) \leq \mathbb{P}\left(\|\pi_H(\varepsilon)\|^2 \geq \sigma^2 q + 4\sigma\sqrt{q}\right).$$

Using the definitions of S_1 and S_2, it follows that

$$\mathbb{P}(\mathcal{E}_+) \leq \mathbb{P}\left(S_1 + S_2 \geq 4\sigma\sqrt{q}\right) \leq \mathbb{P}\left(S_1 \geq 2\sigma\sqrt{q}\right) + \mathbb{P}\left(S_2 \geq 2\sigma\sqrt{q}\right)$$

By Markov's inequality, we have that

$$\mathbb{P}\left(S_1 \geq 2\sigma\sqrt{q}\right) = \mathbb{P}\left(S_1^2 \geq 4\sigma^2 q\right) \leq \frac{\mathbb{E}\left(S_1^2\right)}{4\sigma^2 q} \leq \frac{1}{4}$$

and similarly that

$$\mathbb{P}\left(S_2 \geq 2\sigma\sqrt{q}\right) = \mathbb{P}\left(S_2^2 \geq 4\sigma^2 q\right) \leq \frac{\mathbb{E}\left(S_2^2\right)}{4\sigma^2 q} \leq \frac{1}{16}.$$

It therefore follows that $\mathbb{P}(\mathcal{E}_+) < 1/2$. We upper bound $\mathbb{P}(\mathcal{E}_-)$ in a similar fashion. Since $\|\pi_H(x)\| \geq \|\pi_H(\varepsilon)\|$, we have that

$$\mathbb{P}\left(\mathcal{E}_-\right) = \mathbb{P}(\|\pi_H(y)\| \leq \sigma\sqrt{q} - 4) \leq \mathbb{P}(\|\pi_H(\varepsilon)\| \leq \sigma\sqrt{q} - 4) = \mathbb{P}(\|\pi_H(\varepsilon)\|^2 \leq \sigma^2 q - 8\sigma\sqrt{q} + 16).$$

https://doi.org/10.52202/079017-4027

Again, recalling the definitions of S_1 and S_2 we have that

$$\mathbb{P}\left(\mathcal{E}_-\right) \leq \mathbb{P}\left(S_1 + S_2 \leq -8\sigma\sqrt{q} + 16\right) \leq \mathbb{P}(S_1 \leq 8 - 4\sigma\sqrt{q}) + \mathbb{P}(S_2 \leq 8 - 4\sigma\sqrt{q}).$$

As before, applying Markov's inequality we have that

$$\mathbb{P}(S_1 \leq 8 - 4\sigma\sqrt{q}) \leq \mathbb{P}(S_1^2 \geq 64 - 64\sigma\sqrt{q} + 16\sigma^2 q) \leq \mathbb{P}\left(S_1^2 \geq 8\sigma^2 q\right) \leq \frac{\mathbb{E}\left(S_1^2\right)}{8\sigma^2 q} \leq \frac{1}{8}$$

where we have twice used the assumption that $q \geq 64/\sigma^2$. Similarly

$$\mathbb{P}(S_2 \leq 8 - 4\sigma\sqrt{q}) \leq \mathbb{P}\left(S_2^2 \geq 8\sigma^2 q\right) \leq \frac{\mathbb{E}\left(S_2^2\right)}{8\sigma^2 q} \leq \frac{1}{32}.$$

It therefore follows that $\mathbb{P}\left(\mathcal{E}_-\right) < 1/2$, thereby establishing (17) and concluding the proof.

D Proof of the claim that $\|\pi_J(\bar{y})\| \leq 2|J|^{1/4}$

In this section, we prove the claim made in the proof of Theorem 1 that

$$\|\pi_J(\bar{y})\| \leq 2|J|^{1/4}, \tag{20}$$

with overwhelming probability, which we require in order to invoke Lemma 1. For notational convenience, we shall assume we are working in an n-dimensional space, rather than an $(n-1)$-dimensional space as this is immaterial in our big-$\mathcal{O}$ bounds.

Recall that $\bar{y}$ is a constant vector with entries in $[0,1]$, and let $\mathbf{1}$ denote the all-ones vector. Let ξ be some value such that $8a < \xi < b/2$, which exists by assumption (R), and let d' denote the smallest index such that

$$\lambda_{d'+1} = \mathcal{O}\left(n^{-1/\xi}\right).$$

This implies that under (P), $d' = \mathcal{O}\left(n^{1/\xi\alpha}\right)$, and under (E), $d' < \log^{1/\gamma}(n^{1/\xi\beta})$. Clearly $d' \leq d$ and so $J \subset \{d'+1, \ldots, n\}$, and therefore

$$\|\pi_J(\bar{y})\| \leq \left\|\pi_{\{d'+1,\ldots,n\}}(\mathbf{1})\right\|.$$

In addition, we observe that

$$\left\|\pi_{\{d'+1,\ldots,n\}}(\mathbf{1})\right\|^2 = n - \left\|\pi_{\{1,\ldots,d'\}}(\mathbf{1})\right\|^2.$$

Since $|J| = \Omega(n)$, it will suffice to show that

$$\frac{1}{n}\left\|\pi_{\{1,\ldots,d'\}}(\mathbf{1})\right\|^2 = 1 - \Omega\left(n^{-1/2-\Omega(1)}\right) \tag{21}$$

with overwhelming probability.

Let $\widehat{U}_{d'}$ and $U_{d'}$ denote the $n \times d'$ matrices with entries

$$\widehat{U}_{d'}(i,j) = \widehat{u}_j(i), \qquad U_{d'}(i,j) = \frac{u_j(x_i)}{n^{1/2}}$$

respectively. Then we have

$$\frac{1}{n}\left\|\pi_{\{1,\ldots,d'\}}(\mathbf{1})\right\|^2 = \frac{1}{n}\left\|\widehat{U}_{d'}\widehat{U}_{d'}^\top \mathbf{1}\right\|^2 = \frac{1}{n}\left\|\widehat{U}_{d'}^\top \mathbf{1}\right\|^2 = \frac{1}{n}\left\|(\widehat{U}_{d'}W)^\top \mathbf{1}\right\|^2,$$

where W is an orthogonal matrix which we will define later on. By the triangle inequality, we have that

$$\frac{1}{n}\left\|(\widehat{U}_{d'}W)^\top \mathbf{1}\right\|^2 \geq \frac{1}{n}\left\|U_{d'}^\top \mathbf{1}\right\|^2 - \frac{1}{n}\left\|\widehat{U}_{d'}W - U_{d'}^\top\right\|^2 \|\mathbf{1}\|^2 = \frac{1}{n}\left\|U_{d'}^\top \mathbf{1}\right\|^2 - \left\|\widehat{U}_{d'}W - U_{d'}^\top\right\|^2$$

and therefore to show (21), we need to show that

$$n^{-1}\left\|U_{d'}^\top \mathbf{1}\right\|^2 = 1 - \mathcal{O}\left(n^{-1/2-\Omega(1)}\right) \tag{22}$$

and

$$\left\|\widehat{U}_{d'}W - U_{d'}^\top\right\| = O\left(n^{-1/4-\Omega(1)}\right) \tag{23}$$

with overwhelming probability.

D.1 Bounding (22)

To bound (22), we begin by using the the inequality $(c_1 + c_2)^2 \leq 2(c_1^2 + c_2^2)$ to write

$$n^{-1} \left\| U_{d'}^\top \mathbf{1} \right\|^2 = n^{-1} \sum_{l=1}^{d'} \left(\sum_{i=1}^{n} \frac{u_l(x_i)}{n^{1/2}} \right)^2 = \sum_{l=1}^{d'} \left(\sum_{i=1}^{n} \frac{u_l(x_i)}{n} \right)^2$$
$$\geq \sum_{l=1}^{d'} \left(\int u_l(x) \mathrm{d}\rho(x) \right)^2 - 2 \sum_{l=1}^{d'} \left(\sum_{i=1}^{n} \frac{u_l(x_i)}{n} - \int u_l(x) \mathrm{d}\rho(x) \right)^2.$$

To bound the first term, we observe that since $\{u_i\}_{i=1}^{\infty}$ forms an orthonormal basis for $L^2_\rho(\mathcal{X})$, we have that

$$\sum_{l=1}^{\infty} \left(\int u_l(x) \mathrm{d}\rho(x) \right)^2 = 1$$

and therefore

$$\sum_{l=1}^{d'} \left(\int u_l(x) \mathrm{d}\rho(x) \right)^2 = 1 - \sum_{l=d'+1}^{\infty} \left(\int u_l(x) \mathrm{d}\rho(x) \right)^2 =: 1 - \Gamma_{d'+1}$$

By assumption,

$$\Gamma_{d'+1} = \mathcal{O}\left(\lambda_{d'+1}^b\right) = \mathcal{O}\left(n^{-b/\xi}\right) = \mathcal{O}\left(n^{-1/2-\Omega(1)}\right)$$

where we have used the assumption that $b > \xi/2$, and therefore

$$\sum_{l=1}^{d'} \left(\int u_l(x) \mathrm{d}\rho(x) \right)^2 = 1 - \mathcal{O}\left(n^{-b/\xi}\right) = 1 - \mathcal{O}\left(n^{-1/2-\Omega(1)}\right),$$

as required.

Now to bound the second term, we use Hoeffding's inequality to obtain that

$$\left| \sum_{i=1}^{n} \frac{u_l(x_i)}{n} - \int u_l(x) \mathrm{d}\rho(x) \right| = \mathcal{O}\left(\|u_l\|_\infty \frac{\log^{1/2}(n)}{n^{1/2}} \right).$$

Under (P), we have that for all $l \leq d'$,

$$\|u_l\|_\infty \lesssim {d'}^r \lesssim n^{r/\xi\alpha} = \mathcal{O}\left(n^{1/2\xi}\right) = \mathcal{O}(n^{1/16-\Omega(1)})$$

where we have used that $r < (\alpha - 1)/2 \leq \alpha/2$, and that $\xi > 8$, and under (E)

$$\|u_l\|_\infty \lesssim e^{s{d'}^\gamma} \lesssim e^{s\log(n)/\xi\beta} = \mathcal{O}\left(n^{1/2\xi}\right) = \mathcal{O}(n^{1/16-\Omega(1)})$$

where we have used that $s < \beta/2$. Therefore, since $d' = \mathcal{O}\left(n^{1/8}\right)$, we have that

$$\sum_{l=1}^{d'} \left(\sum_{i=1}^{n} \frac{u_l(x_i)}{n} - \int u_l(x) \mathrm{d}\rho(x) \right)^2 \lesssim d' \left(n^{1/16-1/2} \right)^2 = \mathcal{O}\left(n^{-3/4}\right)$$

which is $\mathcal{O}\left(n^{-1/2-\Omega(1)}\right)$ as required.

D.2 Bounding (23)

To bound (23), we begin by defining the $d' \times d'$ diagonal matrices $\widehat{\Lambda}_{d'}$ and $\Lambda_{d'}$ with diagonal entries

$$\widehat{\Lambda}_{d'}(i,i) := \frac{\widehat{\lambda}_i}{n}, \qquad \Lambda_{d'}(i,i) := \lambda_i,$$

respectively, and the $n \times d'$ matrices $\widehat{\Phi}_{d'}$ and $\Phi_{d'}$ with entries

$$\widehat{\Phi}_{d'}(i,j) = \widehat{\lambda}_j^{1/2} \widehat{u}_j(i), \qquad \Phi_{d'}(i,j) = \lambda_j^{1/2} u_j(i),$$

https://doi.org/10.52202/079017-4027

respectively. Then in matrix notation we have that

$$\widehat{\Phi}_{d'} = n^{1/2}\widehat{U}_{d'}\widehat{\Lambda}_{d'}^{1/2}, \qquad \Phi_{d'} = n^{1/2}U_{d'}\Lambda_{d'}^{1/2}$$

Now, we decompose $\widehat{U}_{d'}W_{d'} - U_{d'}$ as

$$\widehat{U}_{d'}W_{d'} - n^{-1/2}U_{d'} = \left\{n^{-1/2}\left(\widehat{\Phi}_{d'}W_{d'} - \Phi_{d'}\right) + \widehat{U}_{d'}\left(W_{d'}\Lambda_{d'}^{1/2} - \widehat{\Lambda}_{d'}^{1/2}W_{d'}\right)\right\}\Lambda^{-1/2}$$

and so we have that

$$\left\|\widehat{U}_{d'}W_{d'} - U_{d'}\right\| \leq \lambda_{d'}^{-1/2}\left\{n^{-1/2}\left\|\widehat{\Phi}_{d'}W_{d'} - \Phi_{d'}\right\|_{\mathrm{F}} + \left\|W_{d'}\Lambda_{d'}^{1/2} - \widehat{\Lambda}_{d'}^{1/2}W_{d'}\right\|\right\}.$$

By the construction of d' we have that $\lambda_{d'} = \Omega(n^{-1/\xi}) = \Omega(n^{-1/8})$ and therefore $\lambda_{d'}^{-1/2} = \mathcal{O}\left(n^{1/16}\right)$. Therefore to show (23), it will suffice to show that

$$\left\|\widehat{\Phi}_{d'}W_{d'} - \Phi_{d'}\right\|_{\mathrm{F}} = \mathcal{O}\left(n^{\frac{3}{16}-\Omega(1)}\right) \tag{24}$$

and

$$\left\|W_{d'}\Lambda_{d'}^{1/2} - \widehat{\Lambda}_{d'}^{1/2}W_{d'}\right\| = \mathcal{O}\left(n^{-\frac{5}{16}-\Omega(1)}\right). \tag{25}$$

To obtain the bound (24), we make use of the following result which is Equation 3.8 of Tang et al. [2013].

Lemma 5. *The exists an orthogonal matrix W_d (constructed as in* (29)*) such that*

$$\left\|\widehat{\Phi}_d W_d - \Phi_d\right\|_{\mathrm{F}} = \mathcal{O}\left(\frac{\log^{1/2}(n)}{\lambda_d - \lambda_{d+1}}\right)$$

with overwhelming probability.

Now, let

$$d^\star := \arg\max_{i \geq d'}\left\{\lambda_i - \lambda_{i+1}\right\}, \tag{26}$$

then by the assumption (R), we have that

$$\lambda_{d^\star} - \lambda_{d^\star+1} = \Omega(\lambda_{d'}^a) = \Omega(n^{-a/\xi}) = \Omega\left(n^{-1/8+\Omega(1)}\right). \tag{27}$$

Using Lemma 5, we obtain that

$$\left\|\widehat{\Phi}_{d'}W_{d'} - \Phi_{d'}\right\|_{\mathrm{F}} \leq \left\|\widehat{\Phi}_{d^\star}W_{d^\star} - \Phi_{d^\star}\right\|_{\mathrm{F}} = \mathcal{O}\left(\frac{\log^{1/2}(n)}{\lambda_{d^\star} - \lambda_{d^\star+1}}\right) = \mathcal{O}\left(n^{1/8-\Omega(1)}\right),$$

which establishes (24). Obtaining the bound (25) requires some new concepts, so we dedicate it its own section.

D.3 Bounding (25)

We now turn our attention to the bound (25). We begin by defining $\mathcal{H}$, the reproducing kernel Hilbert space associated with the kernel k, and define the operators $\mathcal{K}_{\mathcal{H}}, \widehat{\mathcal{K}}_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ given by

$$\begin{aligned}\mathcal{K}_{\mathcal{H}}f &= \int \langle f, k_x\rangle_{\mathcal{H}}\, k_x \mathrm{d}\rho(x), \\ \widehat{\mathcal{K}}_{\mathcal{H}} &= \frac{1}{n}\sum_{i=1}^n \langle f, k_{x_i}\rangle_{\mathcal{H}}\, k_{x_i}\end{aligned} \tag{28}$$

where $k_x(\cdot) = k(\cdot, x)$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. These operators are known as the "extension operators" of $\mathcal{K}$ and $\frac{1}{n}K$, respectively, and it may be shown that each has the same eigenvalues as its corresponding operator, possibly up to zeros (see e.g. Propositions 8 and 9 of Rosasco et al. [2010]). We will make use of the following concentration inequality for $\mathcal{K}_{\mathcal{H}} - \widehat{\mathcal{K}}_{\mathcal{H}}$ which is due to De Vito et al. [2005] and appears as Theorem 7 of Rosasco et al. [2010].

Lemma 6 (Theorem 7 of Rosasco et al. [2010]). *The operators $\mathcal{K}_{\mathcal{H}}$ and $\widehat{\mathcal{K}}_{\mathcal{H}}$ are Hilbert-Schmidt, and*

$$\left\|\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}}\right\|_{HS} = \mathcal{O}\left(\frac{\log^{1/2}(n)}{n^{1/2}}\right)$$

with overwhelming probability.

Let $\{u_{\mathcal{H},i}\}_{i\leq d}$ denote the eigenfunctions of $\mathcal{K}_{\mathcal{H}}$ corresponding to the eigenvalues $\{\lambda_i\}_{i\leq d}$, and let $\{\widehat{u}_{\mathcal{H},i}\}_{i\leq d}$ denote the eigenfunctions of $\widehat{\mathcal{K}}$ corresponding to the eigenvalues $\{\widehat{\lambda}_i/n\}_{i\leq d}$. We define the (infinite-dimensional) "matrices" $U_{\mathcal{H},d}$ and $\widehat{U}_{\mathcal{H},d}$ whose columns contain the eigenfunctions $\{u_{\mathcal{H},i}\}_{i\leq d}$ and $\{\widehat{u}_{\mathcal{H},i}\}_{i\leq d}$, respectively. These "matrices" are well-defined with matrix multiplication is defined as usual with inner products taken in $\mathcal{H}$. We will refer to $U_{\mathcal{H},d}, \widehat{U}_{\mathcal{H},d}$ and the subspaces spanned by their columns interchangably.

Let $H_d = U_{\mathcal{H},d}^{\top}\widehat{U}_{\mathcal{H},d}$ with entries $H_d(i,j) = \langle u_{\mathcal{H},i}, \widehat{u}_{\mathcal{H},j}\rangle_{\mathcal{H}}$ and denote its singular values by $\xi_1, \ldots, \xi_d$. Then the principal angles between the subspaces $U_{\mathcal{H},d}$ and $\widehat{U}_{\mathcal{H},d}$, which we will denote by $\theta_1, \ldots, \theta_d$, are define as via $\xi_i = \cos(\theta_i)$. Define the matrix

$$\sin\Theta\left(\widehat{U}_{\mathcal{H},d}, U_{\mathcal{H},d}\right) = \operatorname{diag}\left(\sin(\theta_1), \ldots, \sin(\theta_d)\right).$$

Let $d^\star$ be as in (26) so that $\lambda_{d^\star} - \lambda_{d^\star+1} = \Omega\left(n^{-1/8+\Omega(1)}\right)$. Since $d^\star \geq d'$, by the Davis-Kahan theorem, we have that

$$\begin{aligned}\left\|\sin\Theta\left(\widehat{U}_{\mathcal{H},d'}, U_{\mathcal{H},d'}\right)\right\| \leq \left\|\sin\Theta\left(\widehat{U}_{\mathcal{H},d^\star}, U_{\mathcal{H},d^\star}\right)\right\| &\leq \frac{\sqrt{2}\left\|\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}}\right\|}{\lambda_{d^\star} - \lambda_{d^\star+1}} \\ &\leq \frac{\sqrt{2}\left\|\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}}\right\|_{\text{HS}}}{\lambda_{d^\star} - \lambda_{d^\star+1}} = \mathcal{O}\left(n^{-3/8-\Omega(1)}\right).\end{aligned}$$

where the second inequality comes from the relationship $\|\cdot\| \leq \|\cdot\|_{\text{HS}}$, and the final inequality follows from Lemma 6 and (27).

We now come to constructing the matrix W_d. We denote the singular value decomposition of H as $H = W_{d,1}\Xi W_{d,2}^{\top}$ and define the matrix W_d by

$$W = W_{d,1}W_{d,2}^{\top}, \tag{29}$$

which is known as the "matrix sign" of H. Let $\widehat{\mathcal{P}}_d$ and $\mathcal{P}_d$ to be the projections onto the subspaces $\widehat{U}_{\mathcal{H},d}$ and $U_{\mathcal{H},d}$, respectively. Then we have the following decomposition.

$$\begin{aligned}W_{d'}\Lambda_{d'}^{1/2} - \widehat{\Lambda}_{d'}^{1/2}W_{d'} &= (W_{d'} - H_{d'})\widehat{\Lambda}_{d'} + \Lambda_{d'}(H_{d'} - W_{d'}) \\ &\quad + U_{\mathcal{H},d'}^{\top}(\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}})(\widehat{\mathcal{P}}_{d'} - \mathcal{P}_{d'})\widehat{U}_{\mathcal{H},d'} \\ &\quad + U_{\mathcal{H},d'}^{\top}(\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}})U_{\mathcal{H},d'}H_{d'}\end{aligned} \tag{30}$$

We first observe that $\|H_{d'}\| \leq \|\widehat{U}_{\mathcal{H},d'}\|\|U_{\mathcal{H},d'}\| = 1$, and by (5), $\left\|\widehat{\Lambda}_{d'}\right\|, \|\Lambda_{d'}\| = \mathcal{O}(1)$. In addition, following the same steps as in the proof Lemma 6.7 of Cape et al. [2019] (who prove similar (in)equalities for finite-dimensional matrices) we have that

$$\begin{aligned}\left\|\widehat{\mathcal{P}}_{d'} - \mathcal{P}_{d'}\right\| &= \left\|\sin\Theta\left(\widehat{U}_{\mathcal{H},d'}, U_{\mathcal{H},d'}\right)\right\| = \mathcal{O}\left(n^{-3/8-\Omega(1)}\right) \\ \|H_{d'} - W_{d'}\| &\leq \left\|\sin\Theta\left(\widehat{U}_{\mathcal{H},d'}, U_{\mathcal{H},d'}\right)\right\|^2 = \mathcal{O}\left(n^{-3/4-\Omega(1)}\right).\end{aligned}$$

We now turn to bounding $\|U_{\mathcal{H},d}^{\top}(\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}})U_{\mathcal{H},d}\|$. To condense notation, let $Q = U_{\mathcal{H},d}^{\top}(\widehat{\mathcal{K}}_{\mathcal{H}} - \mathcal{K}_{\mathcal{H}})U_{\mathcal{H},d}$. We will bound $\|Q\|$ using a classical ε-net argument. Let $\mathbb{S}_{\varepsilon}^{d-1}$ be an ε-net of the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1} := \{v : \|v\| = 1\}$, that is, a subset of $\mathbb{S}^{d-1}$ such that for any $v \in \mathbb{S}^{d-1}$,

https://doi.org/10.52202/079017-4027

there exists some $w_v \in \mathbb{S}_\varepsilon^{d-1}$ such that $\|v - w_v\| < \varepsilon$. Then, we have that

$$\begin{aligned}\|Q\| &= \max_{v:\|v\|\leq 1} \left|v^\top Q v\right| \\ &= \max_{v:\|v\|\leq 1} \left|(v - w_v + w_v)^\top Q (v - w_v + w_v)\right| \\ &\leq \left(\varepsilon^2 + 2\varepsilon\right)\|Q\| + \max_{w\in\mathbb{S}_\varepsilon^{d-1}} \left|w^\top Q w\right|.\end{aligned}$$

With $\varepsilon = 1/3$, we have

$$\|Q\| \leq \frac{9}{2} \max_{w\in\mathbb{S}_\varepsilon^{d-1}} \left|w^\top Q w\right|.$$

Using the definitions (28), its (l, m)th entry can be calculated as

$$Q(l,m) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)u_l(x_i)u_m(x_j) - \iint k(x,y)u_l(x)u_m(y)\mathrm{d}\rho(x)\mathrm{d}\rho(y), \tag{31}$$

and so for a given $w \in \mathbb{S}_\varepsilon^{d-1}$, we have that

$$\begin{aligned}\left|w^\top Q w\right| &= \left|\sum_{l,m=1}^{d} Q(l,m)w(l)w(m)\right| \\ &= \left|\sum_{l,m=1}^{d} w(l)w(m)\left(\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j)u_l(x_i)u_m(x_j) - \iint k(x,y)u_l(x)u_m(y)\mathrm{d}\rho(x)\mathrm{d}\rho(y)\right)\right| \\ &\leq \max_{1\leq l\leq d}\|u_l\|_\infty \left|\sum_{i=1}^{n}\left\{\sum_{l,m=1}^{d} w(l)w(m)\left(\frac{u_m(x_i)}{n} - \int u_m(x)\mathrm{d}\rho(x)\right)\right\}\right|\end{aligned}$$

where we have used that $k(x, y) \leq 1$ for all $x, y \in \mathcal{X}$. This is a sum of independent random variables, so by Hoeffding's inequality, we that that

$$\mathbb{P}\left(\left|w^\top Q w\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{\max_{1\leq l\leq d}\|u_l\|_\infty^4}\right)$$

where we have used that $\sum_{l,m=1}^{d} w(l)w(m) = 1$. The set $\mathbb{S}_{1/3}^{d-1}$ can be selected so that its cardinality is no greater than 18^d (see, for example, Pollard [1990]), so using a union bound we have that

$$\begin{aligned}\mathbb{P}\left(\|Q\| \geq t\right) &\leq \mathbb{P}\left(\max_{w\in\mathbb{S}_{1/3}^{d-1}}\left|w^\top Q w\right| \geq \frac{2}{9}t\right) \\ &\leq \sum w \in \mathbb{S}_{1/3}^{d-1}\mathbb{P}\left(\left|w^\top Q w\right| \geq \frac{2}{9}t\right) \\ &\leq 2\cdot 18^{d'}\exp\left(-\frac{8nt^2}{81\cdot\max_{1\leq l\leq d'}\|u_l\|_\infty^4}\right) \\ &\leq 2\exp\left(d'\log(18) - \frac{8nt^2}{81\cdot\max_{1\leq l\leq d'}\|u_l\|_\infty^4}\right) \\ &\leq 2\exp\left(C\left(n^{1/8-\Omega(1)} - n^{3/4}t^2\right)\right)\end{aligned}$$

where we have used that $d' = \mathcal{O}(n^{1/8-\Omega(1)})$ and $\max_{1\leq l\leq d'}\|u_l\|_\infty = \mathcal{O}(n^{1/16})$. Choosing $t = 2n^{-5/16-\Omega(1)}$ we have that

$$\mathbb{P}\left(\|Q\| \geq 2n^{-5/16-\Omega(1)}\right) \leq 2\exp\left(-n^{1/8}\right) \leq n^{-c}$$

for any $c > 0$ for large enough n. Therefore

$$\|Q\| = \mathcal{O}\left(n^{-5/16-\Omega(1)}\right)$$

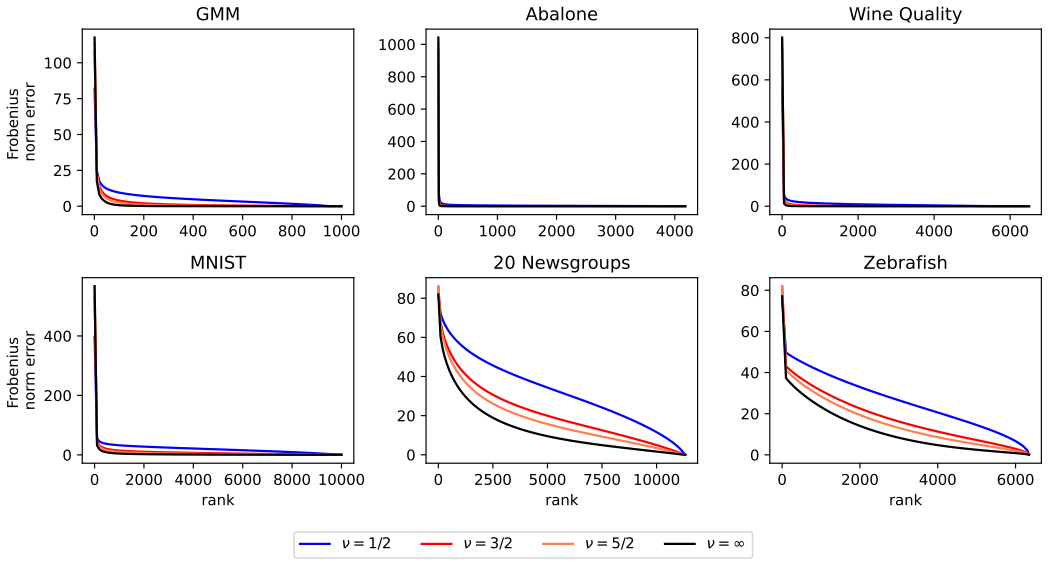

Figure 2: The Frobenius-norm error against rank for low-rank approximations of kernel matrices constructed from a collection of datasets. The kernel matrices are constructed using Matérn kernels with a range of smoothness parameters, each of which is represented by a line in each plot. Details of the experiment are provided in Section 5.

with overwhelming probability.

Combining the above bounds with (30) we obtain that

$$\left\| W_{d'} \Lambda_{d'}^{1/2} - \widehat{\Lambda}_{d'}^{1/2} W_{d'} \right\| = \mathcal{O}\left(n^{-5/16 - \Omega(1)} \right)$$

which establishes (25) and therefore completes the proof.

E Additional details about the experiments

In this section, we provide some additional details and plots relating to the experiments in Section 5.

E.1 Details about the datasets

GMM is a synthetic dataset of 1,000 simulated data points from a 10-component Gaussian mixture model with unit isotropic covariances and means of size ten on the axes. *Abalone* [Nash et al., 1995] and *Wine Quality* [Cortez et al., 2009] are popular benchmark datasets which we standarised in the usual way by centering and rescaling each feature to have unit variance. We drop the binary *Sex* feature from the *Abalone* dataset to retain only the continuous features. *MNIST* [Deng, 2012] is a dataset of handwritten digits, represented as 28x28 gray-scale pixels which we concatenate into 784 dimensional vectors. These four datasets might be described as "low-dimensional", and are thus representative of the theory we present in Section 3.

In addition, we consider two "high-dimensional" datasets. *20 Newsgroups* [Lang, 1995] is a popular natural language dataset of messages collected from twenty different "newnews" newsgroups. We remove stop-word and words which appear in fewer than 5 documents or more than 80% of them, and convert each document into a vector using term frequency-inverse document frequency (tf-idf) features for each word. *Zebrafish* [Wagner et al., 2018] is a dataset of single-cell gene expression in a zebrafish embryo taken during their first day of development. We subsample 10% of the cells, and process the data following the steps in Wagner et al. [2018].

https://doi.org/10.52202/079017-4027

E.2 Frobenius norm errors

Figure 2 shows the Frobenius norm error of the low-rank approximations. For the low-dimensional datasets *GMM*, *Abalone*, *Wine Quality* and *MNIST*, the Frobenius norm error decays very quickly. However for the high-dimensional datasets *20 Newsgroups* and *Zebrafish*, the Frobenius norm error decays much more slowly. As pointed out in the main text, the Frobenius norm error of the *20 Newsgroups* dataset does not exhibit the sharp drop between $d = 2500$ and $d = 3000$ that the maximum entrywise error exhibits.

E.3 Implementation details

The experiments were performed on the HPC cluster at Imperial College London with 8 cores and 16GB of RAM. The *GMM* experiment took less than 1 minute; the *Abalone*, *Wine Quality*, *MNIST* and *Zebrafish* experiments took less than 8 hours; and the *20 Newsgroups* experiment took less than 24 hours.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.1. Contributions

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our theory are discussed extensively in Section 6.1.

https://doi.org/10.52202/079017-4027

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions for the main theorem, Theorem 1, are provided in Section 3 and are summarised in Table 1. The full proof is provided in Section 4 and additonal technical details are provided in Sections B, C and D. Proofs of Propositions 1 and 2 are provided in Section A.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Code to reproduce the experiments is provided with the submission.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used for the experiments are openly available and a link to the code to reproduce the experiments is provided in Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

 https://doi.org/10.52202/079017-4027

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Preprocessing of data is explained and code provided. There are no hyperparameters to tune or data splits in our experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments are deterministic, so there are no error bars to show.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are provided in Section E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the introduction, we discuss the importance of seeking methods with low entrywise error bounds for applications in which individual errors carry a high cost. Our work is foundational theory, and we don't foresee any negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

https://doi.org/10.52202/079017-4027

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Licenses for the datasets used in the experiments are described with the provided code.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-4027