FAST SPECTRUM ESTIMATION OF SOME KERNEL MATRICES*

MIKHAIL LEPILOV†

Abstract. In data science, individual observations are often assumed to come independently from an underlying probability space. Kernel matrices formed from large sets of such observations arise frequently, for example during classification tasks. It is desirable to know the eigenvalue decay properties of these matrices without explicitly forming them, such as when determining if a low-rank approximation is feasible. In this work, we introduce a new eigenvalue quantile estimation framework for some kernel matrices. This framework gives meaningful bounds for all the eigenvalues of a kernel matrix while avoiding the cost of constructing the full matrix. The kernel matrices under consideration come from a kernel with quick decay away from the diagonal applied to uniformly-distributed sets of points in Euclidean space of any dimension. We prove the efficacy of this framework given certain bounds on the kernel function, and we provide empirical evidence for its accuracy. In the process, we also prove a very general interlacing-type theorem for finite sets of numbers. Additionally, we indicate an application of this framework to the study of the intrinsic dimension of data, as well as several other directions in which to generalize this work.

Key words. eigenvalue algorithm, kernel matrix, interlacing theorem, quantile bounds

AMS subject classifications. 60B20, 62R07, 65F15, 68T09

1. Introduction.

Background. Kernel matrices that result from applying a positive-definite function pairwise to a finite set of points $X \subseteq \mathbb{R}^d$ arise in several areas of computational mathematics such as image processing and machine learning. In the latter field especially, common methods involve performing expensive computations with a kernel matrix, such as inverting it or finding its eigenvalues [18, 19]. The kernel matrix involved, however, may be of a prohibitively large size to even form, let alone to do computations with. On the other hand, if the matrix has quick eigenvalue decay relative to its norm, then we may be able to efficiently carry out computations on its low-rank approximation instead. A good overview of such computations and their complexity is found in [6]. Hence, it is useful to study *a priori* the eigenvalue decay of a kernel matrix. Given the n data points with which the kernel matrix is formed, we would like to find ways to estimate all of its eigenvalues faster than by having to form the matrix first. That is, we would like to do so in a sub-quadratic number of operations relative to n.

We consider a setting common in data science, which is when the points in X are assumed to be independent and identically-distributed, coming from some latent distribution. In the past, the study of eigenvalue decay of such kernel matrices often focused on asymptotic eigenvalue behavior as the number of distribution samples in X was taken to infinity, after making some appropriate assumptions on the distribution and kernel function involved [15, 4]. However, as the examples in [4] suggest, these bounds rely on the kernel function having its truncated eigendecomposition (in some appropriate function space) readily available. Furthermore, it is unclear exactly how many terms to keep when computing and truncating such an eigendecomposition in order to obtain an eigenvalue decay bound within some tolerance. Thus, it is impractical to use such ideas for our purposes of estimating eigenvalue decay of a given kernel matrix.

*Preprint.

Funding: The research of Mikhail Lepilov was supported in part by NSF grant DMS-2038118.

†Department of Mathematics, Emory University, Atlanta, GA 30306 (mlepilo@emory.edu).

These difficulties are sidestepped when empirical methods are used to obtain bounds on eigenvalues, such as matrix sketching. However, most sketching techniques typically require not only forming the kernel matrix but also finding matrix-vector products with sets of specially-crafted vectors. For some examples and an overview, see [22, 20]. Such techniques applied to an $n \times n$ matrix, therefore, would require a number of operations that scales at least quadratically in n, so most sketching approaches do not provide a way to achieve our goal. One exception is the class of techniques known as Nyström methods, which can be thought of as matrix sketching methods that do not require forming the entire kernel matrix. In Nyström methods, a random subsample of the points in X, and hence of the kernel matrix, is used to obtain a low-rank decomposition of the full matrix. The spectrum of this randomly-subsampled matrix is shown to be correlated pointwise with the first few eigenvalues of the full matrix [21]. Various strategies for sampling the matrix and obtaining theoretical pointwise accuracy guarantees for this correlation have been implemented over the years. Such guarantees depend on performing additional computations with the data points informing the choice of samples; see, for example, [8]. An in-depth empirical exploration of such guarantees, and especially their limitations, is given in [12]. However, since the goal of such methods is to find the best low-rank approximation, and not to find whether or not a good such approximation exists, these accuracy guarantees only apply to give eigenvalue estimates for the first few eigenvalues. Furthermore, in practice, the low-cost "naive" Nyström method of [21] actually does not work to give a subsampled matrix with similar eigenvalues if the matrix has high numerical rank; see Figure 1.1 for an illustration of this phenomenon.

Even more recently, related work comes from approximating graph spectra in subquadratic time, such as in [2, 5]. In this approach, the kernel matrix can be regarded as the Laplacian of a particular weighted complete graph. Specifically, each vertex corresponds to one point, and each edge has weight equal to the kernel evaluated at the points corresponding to the vertices that the edge connects. Methods based on this are different from Nyström methods and instead give bounds in the Wasserstein-1 metric, often referred to as the "earth-mover distance." From this, however, it is difficult to obtain pointwise estimates of the matrix spectrum. The reference [2] does contain such estimates for the first few eigenvalues but not for the later eigenvalues.

Finally, something close to being fit for our purpose may be found in [3]. This is a result for general symmetric matrices that, in its basic form, gives additive bounds unrelated to the magnitude of each eigenvalue for the later eigenvalues. This makes controlling errors difficult for the later eigenvalues, and it prevents us from using the approach if the numerical rank of the matrix is not already low.

Hence, to obtain accurate pointwise estimates for all the eigenvalues of a given kernel matrix in subquadratic time, we must find a new empirical approach that avoids the issues of the methods above. To do so, we first note that all of the methods we mention so far use no more information than just the fact that the matrix is symmetric. Thus, using more information about the distribution underlying X, as well as the kernel involved in forming the matrix, may enable us to find a better approximation for its spectrum.

Our contribution. In this work, we use this information to design a fundamentally new eigenvalue estimation technique based on finding bounds for the expected k quantiles of the eigenvalue distribution of a kernel matrix, for the case that $k \ll n$. This is done, in turn, by matching the moments of this eigenvalue distribution with that of a smaller, $k \times k$ matrix formed specifically for the purpose. Empirically, it

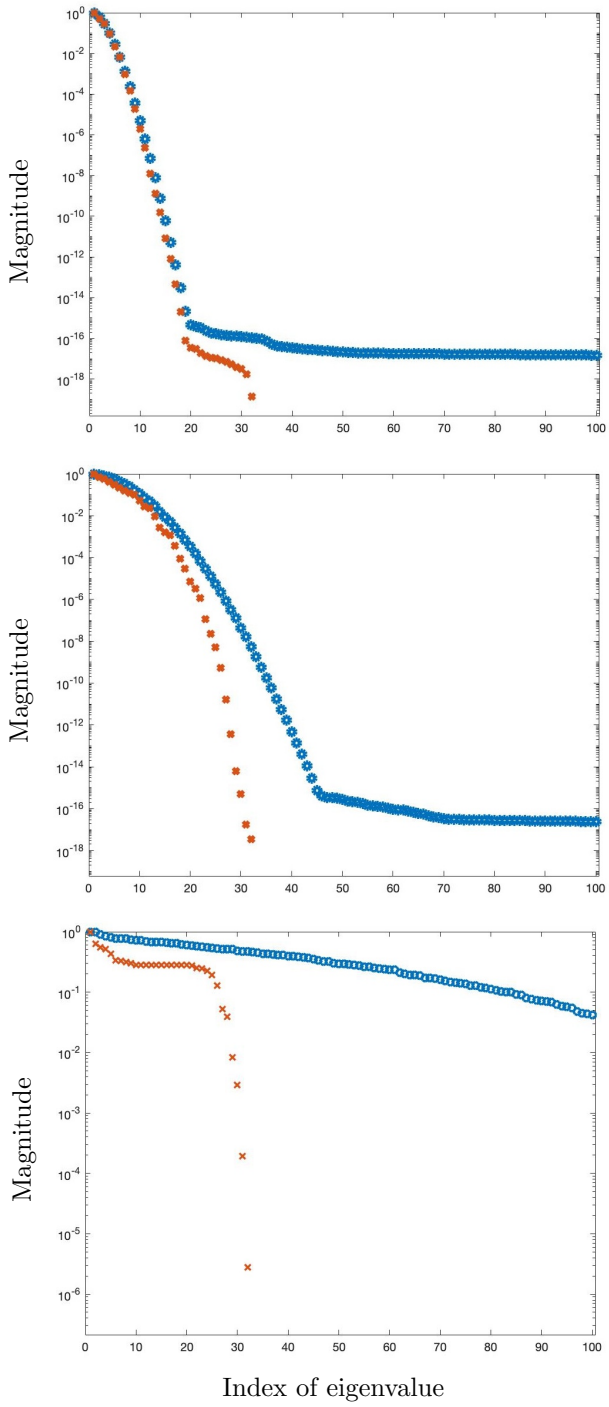

FIG. 1.1. *The first 100 eigenvalues of the kernel matrix (blue) formed when X consists of 512 points taken from the standard uniform distribution in one dimension, as well as those of its "naive" Nyström approximation (red) with 32 points. Here, the kernel used is $\kappa(x,y) = \exp(-10(x-y)^2)$ (top figure), $\kappa(x,y) = exp(-100(x-y)^2)$ (middle figure), and $\kappa(x,y) = exp(-10000(x-y)^2)$ (bottom figure). It is evident that, in the top figure, the eigenvalue decay of the subsampled matrix corresponds well with the eigenvalue decay of the full matrix, but in the center and especially bottom figures, this is no longer the case. This indicates that the Nyström method only works to give an estimate of numerical rank if we know a priori that it is low for our given kernel matrix, as in the top figure.*

turns out that this technique works precisely when the kernel in question has quick decay away from the diagonal, which corresponds to the case that the matrix is of high numerical rank. This complements the existing methods mentioned above, which do not give good accuracy guarantees in such cases (again, see Figure 1.1). Although it is true that in the case of a one-dimensional kernel, such matrices may be approximated by banded matrices, this is no longer the case when X is in Euclidean space of moderate or high dimension. Our framework, on the other hand, still applies even in the moderate- or high-dimensional setting.

This new framework requires $O(mk^2)$ computations, where m is a constant that depends on the desired approximation accuracy. Thus, for certain distributions giving rise to X and kernels used to compute A, our new framework allows for the only subquadratic method to find bounds on the later eigenvalues of the resulting kernel matrix, after a preprocessing step that does not depend on the matrix or kernel. In addition, since this is an entirely new approach, it provides a natural set of questions for further study that could allow subquadratic eigenvalue estimates for wider classes of kernel matrices. Along the way, we also show a very general result concerning the interlacing of sets of real numbers which, to our knowledge, has never been shown before. Finally, we propose an application of this work to the problem of finding the so-called intrinsic dimension of a dataset.

The rest of the paper is structured as follows: in Section 2, we detail our approach. In the process, we prove several new results that show its efficacy in kernel matrix eigenvalue quantile estimation. Among these results is the aforementioned new, general interlacing result about finite sets of real numbers. In Section 3, we give some numerical experiments showing the strengths and limitations of our new framework. Finally, in Section 4, we pose a number of questions for further study that could improve the framework. We also suggest an application to the problem of dimension reduction in data science.

Throughout the paper, we use the following notation. Let $d, n \in \mathbb{N}$, and let $X \subseteq \mathbb{R}^d$ with $|X| = n$. Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric, positive-definite function. Fix an indexing $X = \{x_1, \dots, x_n\}$. By $\kappa(X, X)$, we mean the kernel matrix $A \in \mathbb{R}^{n\times n}$ with entries $A_{ij} = \kappa(x_i, x_j)$. For a symmetrix matrix $A \in \mathbb{R}^{n\times n}$ and some $1 \le j \le n$, we denote by $\sigma_j(A)$ the jth largest eigenvalue of A. Finally, for $a, b \in \mathbb{R}$ with $a \le b$, we denote by $U[a, b]$ the uniform distribution on the interval $[a, b]$.

2. Theoretical results. Fix X and A as above. We will assume throughout the paper that each $x_i \sim U[0, 1]$, but we will comment later on how we may relax this assumption to obtain more general analogs of our main ideas. We concern ourselves with finding bounds for the eigenvalues of A.

We do so by finding another kernel matrix $B \in \mathbb{R}^{k\times k}$, for $k \ll n$, formed using k points sampled from among the x_is. We wish for the k eigenvalues of B to then give bounds for the k quantiles of the eigenvalue distribution of A in the following way. Without loss of generality, we may assume $k|n$. We wish for B to have the property that

$$\sigma_{\lceil \frac{jk}{n} \rceil - 1}(B) \ge \sigma_j(A) \ge \sigma_{\lceil \frac{jk}{n} \rceil + 1}(B), \tag{2.1}$$

for $1 \le j \le n$, where we define "$\sigma_0(B) = \infty$" and "$\sigma_{k+1}(B) = 0$." In other words, we wish for each n/k consecutive eigenvalues, ordered of A to be "sandwiched" between two of the k eigenvalues of B, which we may compute in $O(k^2)$ time [13]. We may look ahead to Figure 2.1 for a picture of this, but we first state our motivation. The reason we wish to find another matrix B using a subsample of the original points,

heuristically, is to preserve information about the geometry of the distribution that gives rise to the x_is. An implicit assumption is that n is so large compared to k that picking k of the x_is is the same thing as drawing from the original distribution, so that B contains some information about the geometry of the points forming A.

2.1. Interlacing property of sets of real numbers. We may expect to get something like the bounds in (2.1) if we match each of the k moments of the empirical spectral distributions of A and B, which are defined as the discrete uniform distributions $\mathcal{A} = \{\sigma_1(A), \dots, \sigma_n(A)\}$ and $\mathcal{B} = \{\sigma_1(B), \dots, \sigma_k(B)\}$, respectively. This is because of the usual notion that the moments of a distribution convey its "shape." In the case of the discrete uniform distribution $\mathcal{B}$, we know that such shape information is contained entirely in its first k moments, since $\mathcal{B}$ contains only k points. Hence, we may informally think of matching each of the k moments of $\mathcal{A}$ and $\mathcal{B}$ as the best we can do in terms of estimating quantiles. Formally, we have the proposition below, which is a very general property of sets of real numbers.

Note that, for convenience of notation, we assume henceforth that all the eigenvalues of A and B are distinct. In practice, this assumption holds if the underlying distribution of X is continuous and the kernel is strictly decreasing away from the diagonal. However, the following proposition and corollary can be easily modified to hold even in the case of repeated eigenvalues.

PROPOSITION 2.1. *Let $S, T \subseteq \mathbb{R}_{\geq 0}$ with $|S| = n$, $|T| = k$, and $k|n$. Denote by a_i and b_j the ith and jth largest elements of S and T, respectively, and suppose $\sum_{i=1}^{n} \frac{a_i^r}{n} = \sum_{i=1}^{k} \frac{b_i^r}{k}$ for all $r = 1, \dots, k$. Then*

$$b_{\lceil \frac{jk}{n} \rceil - 1} \leq a_j \leq b_{\lceil \frac{jk}{n} \rceil + 1}$$

for all $j = 1, \dots, n$, where we define $b_0 = 0$ and $b_{k+1} = \infty$. (See Figure 2.1 for an illustration of this.)

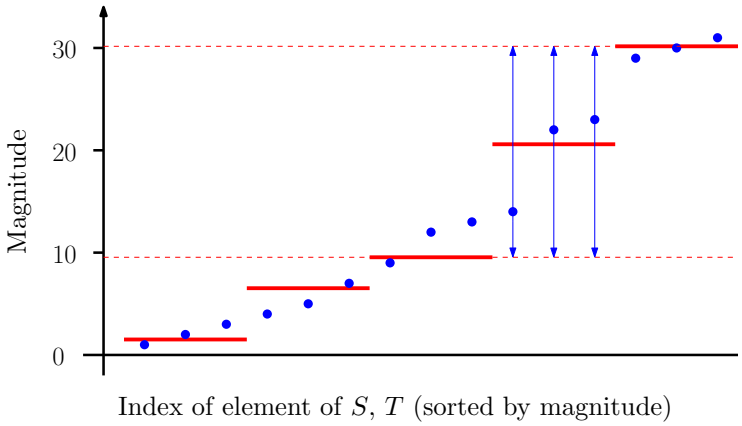

FIG. 2.1. *The sets $S = \{1, 2, 3, 4, 5, 7, 9, 12, 13, 14, 22, 23, 29, 30, 31\}$ (blue dots) and T (solid red dashes), where T is picked such that $\sum_{i=1}^{15} a_i^r/15 = \sum_{i=1}^{5} b_i^r/5$ for $r = 1, \dots, 5$. Hence, T is approximately $\{1.51216, 6.52312, 9.54601, 20.5897, 30.1624\}$. Proposition 2.1 shows, for example, that $b_3 \leq a_{10}, a_{11}, a_{12} \leq b_5$. This is illustrated with the blue arrows above.*

Proof. Consider the discrete uniform probability distributions on S and T, with the former having cumulative distribution function F_S. Then denoting by μ_i and

ν_i the ith moments of these distributions on S and T for $i = 0, \ldots, k$, respectively, our assumptions are equivalent to requiring that $\mu_i = \nu_i$ for each $i = 1, \ldots, k$, and therefore for each $i \in \mathbb{N}$. The statement follows as a quick corollary to some classical results on the bounds for F_S in terms of its moments, which we reproduce here.

Following the notation and presentation of [1]—in particular, note the relationships in Equations 1.3 and 1.4 of Chapter 1—we construct the set of polynomials $P_0, \ldots, P_k$ by the explicit formulas $P_0 = 1$ and

$$P_j = \frac{1}{\sqrt{D_{j-1} D_j}} \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_j \\ \mu_1 & \mu_2 & \cdots & \mu_{j+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{j-1} & \mu_j & \cdots & \mu_{2j-1} \\ 1 & x & \cdots & x^j \end{vmatrix}$$

for $j = 1, \ldots, k$, where

$$D_j = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_j \\ \mu_1 & \mu_2 & \cdots & \mu_{j+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_j & \mu_{j+1} & \cdots & \mu_{2j} \end{vmatrix}$$

for $j = 0, \ldots, k$. These polynomials satisfy a number of properties, but here we note only the following: if we write the product $P_i P_j$ as $(P_i P_j)(x) = \sum_{l=0}^{\deg(P_i)\deg(P_j)} c_{i,j,l} x^l$ for some coefficients $c_{i,j,l}$, then

$$\sum_{l=0}^{\deg(P_i)\deg(P_j)} c_{i,j,l} \mu_l = \delta_{i,j}.$$

Since $\nu_i = \mu_i$ for all i and because T has the discrete uniform distribution, this is equivalent to

$$\sum_{l=1}^{k} \frac{(P_i P_j)(b_l)}{k} = \delta_{i,j}, \text{ or}$$

$$\sum_{l=1}^{k} (P_i P_j)(b_l) = k\delta_{i,j}, \tag{2.2}$$

where $\delta_{i,j}$ is the Kronecker delta. (In other words, in our case, $(P_i)_{i=0,\ldots,k}$ is a sequence of polynomials orthogonal with respect to the average of the evaluation functionals at the b_js for $j = 1, \ldots, k$.) Furthermore, following [11], we construct the "empirical Christoffel function"

$$\lambda_k = \frac{1}{\sum_{i=0}^{k} P_i^2}$$

Now, let x_i for $i = 1, \ldots, k$ be the roots of P_k. Using the function λ, the authors in [11] note the following bounds on F_S:

$$1 - \sum_{j=i}^{k} \lambda(x_j) \leq F_S(x_i) \leq \sum_{j=1}^{i} \lambda(x_j).$$

By our definition of F_S, the proposition therefore follows if we show that (1) the b_i's are precisely the roots x_i of P_k, and (2) $\lambda(x_i) = 1/k$ for each $i = 1, \dots, k$. To see (1), we note that since $\mu_i = \nu_i$ for all $i \in \mathbb{N}$, for each b_i we have

$$P_k(b_i) = \frac{1}{\sqrt{D_{k-1}D_k}} \begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^k \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^2 & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{k-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{k} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2k-1} \\ 1 & b_l & \cdots & b_l^k \end{vmatrix}$$

$$= \frac{1}{\sqrt{D_{k-1}D_k}} \left| \begin{bmatrix} \left(\frac{1}{k}\right) & \left(\frac{1}{k}\right) & \cdots & \left(\frac{1}{k}\right) & \cdots & \left(\frac{1}{k}\right) \\ \left(\frac{1}{k}\right) b_1 & \left(\frac{1}{k}\right) b_2 & \cdots & \left(\frac{1}{k}\right) b_i & \cdots & \left(\frac{1}{k}\right) b_k \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right) b_1^{k-1} & \left(\frac{1}{k}\right) b_2^{k-1} & \cdots & \left(\frac{1}{k}\right) b_i^{k-1} & \cdots & \left(\frac{1}{k}\right) b_k^{k-1} \\ 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 1 & b_1 & b_1^2 & \cdots & b_1^k \\ 1 & b_2 & b_2^2 & \cdots & b_2^k \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 1 & b_k & b_k^2 & \cdots & b_k^k \end{bmatrix} \right|$$

$$= 0.$$

Now, note that by fact (1), we see that (2) is equivalent to the condition that $\sum_{i=0}^{k} P_i^2(b_i) = k$ for each $i = 1, \dots, k$. Define the matrix C by

$$C_{j,m} = \sum_{i=0}^{k-1} \frac{\begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} v_l^j \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^2 & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2j-1} \\ 1 & b_m & \cdots & b_m^i \end{vmatrix} \begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i-1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2i-1} \end{vmatrix}}{\begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i} \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^2 & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i+1} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2i} \end{vmatrix} \begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i-1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{i} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2i-2} \end{vmatrix}};$$

then (2) follows once we show that $C_{j,j} = k$ for $j = 1, \dots, k$. To see this, we note that $C = A^T A$, where

$$A_{j,m} = \frac{\begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} v_l^j \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^2 & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2j-1} \\ 1 & b_m & \cdots & b_m^i \end{vmatrix}}{\left(\begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j} \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^2 & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j+1} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2j} \end{vmatrix} \begin{vmatrix} \left(\frac{1}{k}\right)\sum_{l=1}^{k} 1 & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j-1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j-1} & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{j} & \cdots & \left(\frac{1}{k}\right)\sum_{l=1}^{k} b_l^{2j-2} \end{vmatrix} \right)^{1/2}}.$$

On the other hand, we see that

$$\begin{aligned} (AA^T)_{j,m} &= \sum_{i=1}^{k} P_j(b_i) P_m(b_i) \\ &= k\delta_{j,m}, \end{aligned}$$

with the last equality by Equation (2.2). Hence, $AA^T = kI_{k\times k}$, and therefore we have $C = A^T A = kI_{k\times k}$. Thus, $C_{j,j} = k$ for $j = 1, \dots, k$, as desired. ∎

Since the sum of the rth powers of all the eigenvalues of a matrix is equivalent to the trace of its rth power, Proposition 2.1 implies the following statement:

COROLLARY 2.2. *Suppose $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{k\times k}$ have distinct, nonnegative eigenvalues, and suppose we have $\operatorname{tr}(B^r/k) = \operatorname{tr}(A^r/n)$ for $1 \le r \le k$. Then $\sigma_{\lceil \frac{jk}{n} \rceil - 1}(B) \ge \sigma_j(A) \ge \sigma_{\lceil \frac{jk}{n} \rceil + 1}(B)$ for $1 \le j \le n$, where we define "$\sigma_0(B) = \infty$" and "$\sigma_{k+1}(B) = 0$."*

Proof. Defining the empirical spectral distributions $\mathcal{A} = \{\sigma_1(A) \ldots, \sigma_n(A)\}$ and $\mathcal{B} = \{\sigma_1(B), \ldots, \sigma_k(B)\}$ as above, apply Proposition 2.1 while setting $S = \mathcal{A}$ and $T = \mathcal{B}$. The result follows since $\operatorname{tr}(A^r) = \sum_{i=1}^{n}(\sigma_i(A))^r$ and $\operatorname{tr}(B^s) = \sum_{j=1}^{k}(\sigma_j(B))^s$ for $1 \le r \le n$ and $1 \le s \le k$, which follows, in turn, because A and B are positive-definite. ∎

2.2. Matching traces in expectation. Hence, given A, finding B such that Equation (2.1) holds requires us to match the traces of the rth powers of A and B for $r = 1, \ldots, k$. Since A is a random matrix, we will concentrate on understanding the expected traces of A^r and B^r. Here, we assume that n is large enough such that $\sigma_i(A)$ does not vary very much from its expected value in relative terms. However, since k is small, we would need to form B repeatedly m times, where m depends on k and the desired approximation accuracy, and empirically compute the average value of $\sigma_i(B)$. These $\sigma_i(B)s$ would then be used in the way of (2.1).

While we do not know of a way of matching these expected traces exactly, in the next proposition we show a way of matching them approximately if (1) κ is "close to the Kronecker delta"; that is, if κ has very quick decay away from the diagonal; and (2) we have access to a special probability distribution Ξ on $\mathbb{R}^k$. More precisely, κ must satisfy the condition of Equation (2.4) below for some $\epsilon > 0$ to give the relative moment bound (2.5), and Ξ must satisfy (2.3). See Figure 2.2 for an illustration of the condition on κ.

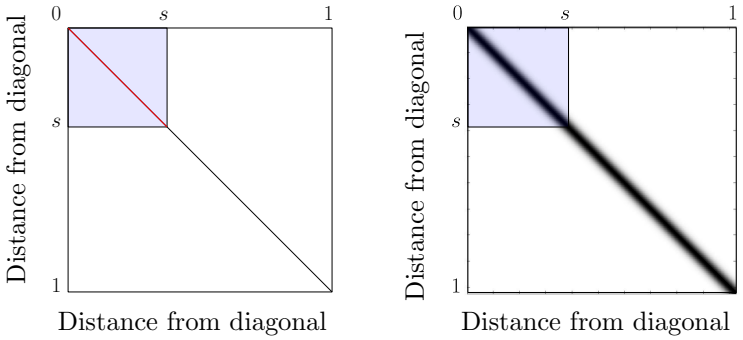

FIG. 2.2. *The condition in (2.4): the left figure is a heatmap of the Kronecker delta on the region $[0,1]\times[0,1]$, and the right figure is a heatmap of the Gaussian kernel $\kappa_1(x,y) = e^{-1000(x-y)^2}$ on the same region. Informally, we may think of the integral of the Kronecker delta over the blue subregion $[0,s]\times[0,s]$ (the length of the red diagonal) as s times its integral over the entire region $[0,1]\times[0,1]$ (the length of the entire diagonal). Of course, both integrals are formally 0. Similarly, we can see that the integral of κ_1 over $[0,s]\times[0,s]$ is approximately s times its integral over $[0,1]\times[0,1]$. This is contrasted with the case, for example, of the Gaussian kernel $\kappa_2(x,y) = e^{-(x-y)^2/10000}$, whose integral over $[0,s]\times[0,s]$ is approximately s^2 times its integral over $[0,1]\times[0,1]$. Thus, the condition (2.4) makes precise the way in which κ_1 does and κ_2 does not have fast decay away from the diagonal.*

In general, as we see in [14], approximate moment matching for a guarantee of pointwise closeness of two cumulative distribution functions may require prohibitively close tolerances. This is likely the main theoretical reason for the requirement that κ

decays quickly away from the diagonal. We will see in Section 3 how quick the decay has to be in practice. The recent work of [16, Theorem 1] suggests that we may be, however, be able to bound our approximate quantile estimates in the Wasserstein-1 metric by a perturbative bound from the "true" quantile estimate.

Once we have (1) and (2), we use the following strategy for picking B such that $\mathbb{E}(\mathrm{tr}(A^r)/n) = \mathbb{E}(\mathrm{tr}(B^r)/k)$ for all $r = 1, \dots, k$:

1. we pick a set Y of some points $y_1, \dots, y_k$ at random from X;
2. we scale each y_i by a random number z_i, where the z_is are picked from a distribution Ξ such that the random vector $\mathbf{z} = (z_1, \dots, z_k)$ satisfies (2.3); and
3. we set $B = \kappa(Y, Y)$ and find its eigenvalues.

We then repeat these steps m times to find the average $\sigma_j(B)$ for $j = 1, \dots, k$.

In order to prove that this works, for technical reasons, we need to fix notation for a walk on the complete graph on n vertices K_n. Namely, we identify a function $\pi : \{0, \dots, r\} \to \{1, \dots, n\}$ with a walk of length r starting (and ending) at a vertex m of the complete graph K_n, where the value of $\pi(i)$ is the index of the vertex of K_n visited at the ith step. (In particular, note that since π is a walk, $\pi(0) = \pi(r) = m$.) We denote by $|\pi|$ the cardinality of the image of π. Then we have the following proposition:

PROPOSITION 2.3. *Let $d, k, n \in \mathbb{N}$ with $k \mid n$. Suppose $\mathbf{z} = (z_1, \dots, z_k)$ is a vector in $\mathbb{R}^k$ with distribution Ξ such that*

$$\begin{cases} P(z_i \neq z_j) = 0 \text{ for all } 1 \leq i, j \leq k, \text{ and} \\ \mathbb{E}\left(\prod_{i \in C} z_i\right) = \frac{k}{n} \frac{\binom{n}{|C|+1}}{\binom{k}{|C|+1}} \end{cases} \tag{2.3}$$

for all nontrivial subsets $C \subseteq \{1, \dots, n\}$ of cardinality at most $k-1$. Suppose that $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a positive-definite function such that, for some $t \in (0,1)$ and any walk π with $|\pi| = l$ on K_n,

$$\frac{\int_{[0,s]^l} \prod_{i=1}^{l} \kappa(x_{\pi(i-1)}, x_{\pi(i)}) d\mathbf{x}_\pi}{\int_{[0,1]^l} \prod_{i=1}^{l} \kappa(x_{\pi(i-1)}, x_{\pi(i)}) d\mathbf{x}_\pi} = s + \epsilon, \tag{2.4}$$

for all $s \in [t, 1]$. Define $x_i \sim U[0,1]$ and $y_j \sim (1/z_j)^{1/d} U[0,1]$ for $1 \leq i \leq n$ and $1 \leq j \leq k$; define $Y = \{y_1, \dots, y_k\}$; and set $A = \kappa(X, X)$ and $B = \kappa(Y, Y)$. Then

$$1 - \epsilon \leq \frac{\mathbb{E}\left(\mathrm{tr}\left(B^r/k\right)\right)}{\mathbb{E}\left(\mathrm{tr}\left(A^r/n\right)\right)} \leq 1 + \epsilon \tag{2.5}$$

for $r = 1, \dots, k$.

Proof. First, note that

$$(A^r)_{mm} = \sum_{l=1}^{r} \sum_{\pi} \prod_{i=1}^{r} A_{\pi(i-1)\pi(i)},$$

where the inner sum ranges over all walks π of length r that visit l distinct vertices on the complete graph K_n, starting at the vertex labeled m. Denote the set of all such walks, starting at any vertex, by $W_l^r(K_n)$. This bookkeeping of walks will be important for our argument to follow. Similarly, we have

$$(B^r)_{mm} = \sum_{l=1}^{r} \sum_{\psi_l^m} \prod_{i=1}^{r} B_{\psi(i-1)\psi(i)},$$

where the inner sum ranges over all walks ψ_m of length r that visit l distinct vertices on the complete graph K_k, starting at the vertex labeled m. Again, denote the set of all such walks, starting at any vertex, by $W_l^r(K_k)$.

Now, note that by linearity of expectation,

$$\begin{aligned}\mathbb{E}\left(\operatorname{tr}(A^r)\right) &= \sum_{m=1}^{n} \mathbb{E}\left((A^r)_{mm}\right) \\ &= \sum_{m=1}^{n} \mathbb{E}\left(\sum_{l=1}^{r} \sum_{\pi \in W_l^r(K_n)} \prod_{i=1}^{r} A_{\pi(i-1)\pi(i)}\right) \\ &= \sum_{l=1}^{r} \sum_{\pi \in W_l^r(K_n)} \mathbb{E}\left(\prod_{i=1}^{r} A_{\pi(i-1)\pi(i)}\right).\end{aligned}$$

By the definition of expectation and the variables A_{ij}, for each $\pi \in W_l^r(K_n)$, we have

$$\begin{aligned}\mathbb{E}\left(\prod_{i=1}^{r} A_{\pi(i-1)\pi(i)}\right) &= \int_{\mathbf{x}_{\mathbf{S}_\pi}} \prod_{i=1}^{r} A_{\pi(i-1)\pi(i)} \\ &= \int_{\mathbf{x}_{\mathbf{S}_\pi}} \prod_{i=1}^{r} \kappa(x_{\pi(i-1)}, x_{\pi(i)}) \\ &= \int_{\mathbb{R}^l} \prod_{i=1}^{r} \kappa(x_{\pi(i-1)}, x_{\pi(i)}) f_\pi(\mathbf{x}_{\mathbf{S}_\pi}) d\mathbf{x}_{\mathbf{S}_\pi} \\ &= \int_{[0,1]^l} \prod_{i=1}^{r} \kappa(x_{\pi(i-1)}, x_{\pi(i)}) d\mathbf{x}_{\mathbf{S}_\pi},\end{aligned}$$

where S_π is the set of vertices visited on the walk π and f_π is the probability density function of the joint distribution of the random variable $\mathbf{x}_{\mathbf{S}_\pi} = (x_{\pi(1)}, \ldots, x_{\pi(r)})$. Similarly, for $\psi \in W_l^r(K_k)$, we have

$$\begin{aligned}\mathbb{E}\left(\prod_{i=1}^{r} B_{\psi(i-1),\psi(i)}\right) &= \int_{\mathbb{R}^k} \int_{\mathbb{R}^l} \prod_{i=1}^{r} \kappa(y_{\psi(i-1)}, y_{\psi(i)}) g_\psi(\mathbf{y}_{\mathbf{S}_\psi}) d\mathbf{y}_{\mathbf{Z},\mathbf{S}_\psi} d\mathbf{z} \\ &= \int_{\mathbb{R}^k} \int_{\mathbb{R}^l} \prod_{i=1}^{r} \kappa(y_{\psi(i-1)}, y_{\psi(i)}) f_\psi(F_\psi^{-1}(\mathbf{y}_{\mathbf{S}_\psi})) |\operatorname{Jac}(F_\psi^{-1})| d\mathbf{y}_{\mathbf{Z},\mathbf{S}_\psi} d\mathbf{z} \\ &= \int_{\mathbb{R}^k} \int_{\prod_{j \in S_\psi}[0,1/z_j]} \prod_{i=1}^{r} \kappa(y_{\psi(i-1)}, y_{\psi(i)}) \prod_{j \in S_\psi} z_j d\mathbf{y}_{\mathbf{S}_\psi} d\mathbf{z} \\ &= \int_{\mathbb{R}^k} \int_{[0,1/z_1]^l} \prod_{i=1}^{r} \kappa(y_{\psi(i-1)}, y_{\psi(i)}) z_1^l d\mathbf{y}_{\mathbf{S}_\psi} d\mathbf{z},\end{aligned}$$

where $F_\psi : \mathbb{R}^l \to \mathbb{R}^l$ is the projection onto the indices S_ψ of the function defined by $F(\mathbf{x}) = (1/Z_1, \ldots, 1/Z_k) \cdot \mathbf{x}$, restricted to the indices S_ψ. Note that the second equality follows from the change-of-variables formula for probability density functions applied to the variable $\mathbf{y} = (y_{\psi(1)}, \ldots, y_{\psi(l)})$, and the third equality follows from the definition of F and the fact that $f_\psi = 1$ for every ψ. Finally, the fourth equality follows from the fact that $P(Z_i \neq Z_j) = 0$ for all $1 \leq i, j \leq k$.

Hence, we see that

$$\frac{\mathbb{E}(\mathrm{tr}(B^r))}{\mathbb{E}(\mathrm{tr}(A^r))} = \frac{\sum_{l=1}^{r}\sum_{\psi\in W_l^r(K_k)}\int_{\mathbb{R}^k}\int_{[0,1/z_1]^l}\prod_{i=1}^{r}\kappa(y_{\psi(i-1)},y_{\psi(i)})z_1^l d\mathbf{y}_{\mathbf{S}_\psi}d\mathbf{z}}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}}$$

$$= \frac{\sum_{l=1}^{r}\frac{\binom{k}{l}}{\binom{n}{l}}\sum_{\pi\in W_l^r(K_n)}\int_{\mathbb{R}^k}\int_{[0,1/z_1]^l}\prod_{i=1}^{r}\kappa(y_{\pi(i-1)},y_{\pi(i)})z_1^l d\mathbf{y}_{\mathbf{S}_\pi}d\mathbf{z}}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}},$$

where the second equality follows from the fact that, for every walk of length r with $1 \le r \le k$ visiting l distinct vertices on K_k, there are $\binom{n}{l}/\binom{k}{l}$ such walks on K_n. Then, by our assumption on κ in Equation (2.4),

$$(1-\epsilon)\frac{k}{n} = \frac{(1-\epsilon)\left(\sum_{l=1}^{r}\frac{\binom{k}{l}}{\binom{n}{l}}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(y_{\pi(i-1)},y_{\pi(i)})d\mathbf{y}_{\mathbf{S}_\pi}\left(\frac{k}{n}\frac{\binom{n}{l}}{\binom{k}{l}}\right)\right)}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}}$$

$$= \frac{(1-\epsilon)\sum_{l=1}^{r}\frac{\binom{k}{l}}{\binom{n}{l}}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(y_{\pi(i-1)},y_{\pi(i)})d\mathbf{y}_{\mathbf{S}_\pi}\int_{\mathbb{R}^k}z_1^{l-1}d\mathbf{z}}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}}$$

$$= \frac{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\frac{\binom{k}{l}}{\binom{n}{l}}\int_{\mathbb{R}^k}\frac{1-\epsilon}{z_1}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(y_{\pi(i-1)},y_{\pi(i)})z_1^l d\mathbf{y}_{\mathbf{S}_\pi}d\mathbf{z}}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}}$$

$$\le \frac{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\frac{\binom{k}{l}}{\binom{n}{l}}\int_{\mathbb{R}^k}\int_{[0,1/z_1]^l}\prod_{i=1}^{r}\kappa(y_{\pi(i-1)},y_{\pi(i)})z_1^l d\mathbf{y}_{\mathbf{S}_\pi}d\mathbf{z}}{\sum_{l=1}^{r}\sum_{\pi\in W_l^r(K_n)}\int_{[0,1]^l}\prod_{i=1}^{r}\kappa(x_{\pi(i-1)},x_{\pi(i)})d\mathbf{x}_{\mathbf{S}_\pi}}$$

$$= \frac{\mathbb{E}(\mathrm{tr}(B^r))}{\mathbb{E}(\mathrm{tr}(A^r))}.$$

By linearity of trace and expectation, we thus get $1-\epsilon \le \mathbb{E}(\mathrm{tr}(B^r/k))/\mathbb{E}(\mathrm{tr}(A^r/n))$. The second inequality in Equation (2.5) follows from Equation (2.4) in a similar way.□

Two questions immediately arise from this last proposition. First, it is not clear which functions κ satisfy Equation (2.4). We explore this topic empirically in Section 3. For the Gaussian kernel $\kappa(x,y) = e^{-\lambda(x-y)^2}$ in particular, we note that for each $\pi \in W_l^r(K_n)$ and $s \in (0,1]$,

$$\lim_{\lambda\to\infty}\frac{\int_{[0,s]^l}e^{-\lambda\sum_{i=1}^{r}(x_{\pi(i-1)},x_{\pi(i)})^2}d\mathbf{x}_{\mathbf{S}_\pi}}{\int_{[0,1]^l}e^{-\lambda\sum_{i=1}^{r}(x_{\pi(i-1)},x_{\pi(i)})^2}d\mathbf{x}_{\mathbf{S}_\pi}} = s.$$

Hence, there exists a length scale λ that makes κ satisfy (2.4). Analogous results may be obtained for other radial basis function (RBF) kernels by finding appropriate limits with respect to the length scale (as with respect to λ above). However, the exact relationship of s, l, and λ in the previous display to a given tolerance ϵ as in (2.4) warrants further study, since it may allow for a more precise formulation of moment bounds. This may be done in combination with studies similar to [14, 16].

Second, it is not clear *a priori* whether or not any distribution Ξ that satisfies (2.3) in the above proposition exists, and if it does, where its support lies. If such a distribution exists, then the method outlined at the beginning of this section should work. It turns out that such a distribution does exist; we next give an example.

EXAMPLE 1. *Fix $n = 49$, $k = 7$. We construct a distribution Ξ such that the random vector $\mathbf{z}$ sampled from it satisfies the mixed moment condition in Proposition 2.3. To do so, we assume that $\mathbf{z}$ takes the same value in each coordinate; for such distributions, the first equation of (2.3) is automatically satisfied. Then, to simplify the search for Ξ, we assume that it has finite support. This assumption makes the second equation of (2.3) equivalent to the system of 7 equations in 8 unknowns*

$$\begin{aligned}
a + b + c + d &= (k/n)\tbinom{n}{1}/\tbinom{k}{1} = 1 \\
a\alpha + b\beta + c\gamma + d\delta &= (k/n)\tbinom{n}{2}/\tbinom{k}{2} = 8 \\
a\alpha^2 + b\beta^2 + c\gamma^2 + d\delta^2 &= (k/n)\tbinom{n}{3}/\tbinom{k}{3} = \tfrac{376}{5} \\
a\alpha^3 + b\beta^3 + c\gamma^3 + d\delta^3 &= (k/n)\tbinom{n}{4}/\tbinom{k}{4} = \tfrac{4324}{5} \\
a\alpha^4 + b\beta^4 + c\gamma^4 + d\delta^4 &= (k/n)\tbinom{n}{5}/\tbinom{k}{5} = 12972 \\
a\alpha^5 + b\beta^5 + c\gamma^5 + d\delta^5 &= (k/n)\tbinom{n}{6}/\tbinom{k}{6} = 285384 \\
a\alpha^6 + b\beta^6 + c\gamma^6 + d\delta^6 &= (k/n)\tbinom{n}{7}/\tbinom{k}{7} = 12271512.
\end{aligned}$$

We picked $\mathbf{z}$ to have four distinct values $\alpha, \beta, \gamma, \delta$ to give enough degrees of freedom for it to satisfy the moment conditions of (2.3); that is, otherwise, we would not have enough unknowns to satisfy the 7 equations above. The values $a \approx 0.41166$, $b \approx 0.56810$, $c \approx 0.020241$, $d \approx 1.4709 \cdot 10^{-6}$, $\alpha \approx 4.8651$, $\beta \approx 9.6827$, $\gamma \approx 24.519$, and $\delta \approx 130.90$ form a solution to this system. Hence, taking Ξ to be the distribution that gives the vector $\mathbf{z}$ with all entries equal to α, all entries equal to β, all entries equal to γ, and all entries equal to δ with probabilities a, b, c, and d, respectively, we find that Ξ satisfies the mixed moment conditions of Equation (2.3). Note that this is equivalent to simply letting Y be a random subset of points in X scaled by α, β, γ, and δ, with probabilities a, b, c, and d, respectively.

We found a distribution in Example 1 that we may use to build a matrix B from A such that (2.5) holds, but only for the case that $n = 49$ and $k = 7$. We did so by looking for a distribution Ξ which gives a random vector $\mathbf{z}$ that can only take the same value in all of its entries. For such distributions, the first equation of (2.3) is automatically satisfied. Furthermore, we assumed Ξ is discrete, which yielded a straightforward system of polynomial equations we could use to find Ξ from the second equation of (2.3).

This construction naturally leads to two questions: first, can we use this technique to find such a distribution for every n, k such that $k|n$? And second, will the support of such a distribution take values that are "too large" to truncate κ in such a manner as to make (2.5) provide a meaningfully-small ϵ? To answer these last two questions, we prove the following proposition. It states that we may always find a distribution with nonnegative support satisfying (2.3), although further questions about its support may be harder to answer.

PROPOSITION 2.4. *Let $k, n \in \mathbb{N}$ such that k is odd and $k|n$. There exists a distribution Ξ on the random variable $\mathbf{z} = (z_1, \ldots, z_k)$, with nonnegative support in each coordinate, such that (2.3) holds for all nontrivial subsets $C \subseteq \{1, \ldots, n\}$ of cardinality at most $k - 1$.*

Proof. If we restrict ourselves to the case that the support of Ξ takes the same value in each coordinate, the moment conditions become equivalent to k prescribed moment conditions for a univariate probability distribution Z with nonnegative sup-

port:

$$\mathbb{E}(Z^r) = (k/n)\tbinom{n}{2}/\tbinom{k}{2} = \frac{\frac{k}{n}\binom{n}{l+1}}{\binom{k}{l+1}}, \quad l = 0, \dots, k-1.$$

This is the approach we had taken for specific values of k and n in Example 1 above. (Note that these moment conditions are largely unrelated to any moment conditions we considered in Proposition 2.1.) But this is just the so-called Stieltjes moment problem, which is well-known to have a solution if certain moment matrices are positive semidefinite and full-rank (or, equivalently, positive definite). For a complete treatment of this question and questions on related moment problems, see the treatise of Curto and Fialkow on the subject [7, Theorem 5.3]. From that result, we see that showing the Proposition comes down to showing that the Hankel matrices

$$H_{k,n} = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{(k-1)/2} \\ \mu_1 & \mu_2 & \cdots & \mu_{(k-1)/2+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{(k-1)/2} & \mu_{(k-1)/2+1} & \cdots & \mu_{k-1} \end{bmatrix} \quad \text{and}$$

$$H'_{k,n} = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_{(k-1)/2} \\ \mu_2 & \mu_3 & \cdots & \mu_{(k-1)/2+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{(k-1)/2} & \mu_{(k-1)/2+1} & \cdots & \mu_{k-1} \end{bmatrix}$$

are positive definite, where $\mu_l = \frac{k}{n}\binom{n}{l+1}/\binom{k}{l+1}$ for $l = 0, \dots, k-1$. We see this once we realize both $H_{k,n}$ and $H'_{k,n}$ as Gram matrices associated to linearly independent sets of vectors in a Hilbert space. In particular, consider the space V of square-integrable functions on the compact interval $[0, 1]$ with respect to the Radon-Nikodym derivative $x^{n-k-1}(1-x)^{k+1}$. For $i = 0, \dots, (k-1)/2$ define $v_i = \sqrt{n-k}(1/(1-x))^{i+1/2}$; and for $j = 0, \dots, (k-1)/2 - 1$, define $w_i = \sqrt{n-k}(1/(1-x))^{i+1}$. Clearly, we have $v_i, w_j \in V$ for $i = 0, \dots, (k-1)/2$ and $j = 0, \dots, (k-1)/2 - 1$. Furthermore, the sets $\{v_i\}_{i=0}^{(k-1)/2}$ and $\{w_j\}_{j=0}^{(k-1)/2-1}$ are linearly independent, and we see that

$$\begin{aligned}
\frac{1}{\frac{k}{n}\binom{n}{k}}\mu_l &= \frac{1}{\frac{k}{n}\binom{n}{k}} \frac{\frac{k}{n}\binom{n}{l+1}}{\binom{k}{l+1}} \\
&= \frac{1}{\binom{n}{k}} \frac{\binom{n}{k}}{\binom{n-(l+1)}{k-(l+1)}} \\
&= \frac{(n-k)!(k-(l+1))!}{(n-(l+1))!} \\
&= (n-k) \int_0^1 x^{n-(l+2)-(k-(l+1))}(1-x)^{k-l} dx \\
&= \int_0^1 \left(\sqrt{n-k} \frac{1}{(1-x)^{i+1/2}} \right) \left(\sqrt{n-k} \frac{1}{(1-x)^{j+1/2}} \right) x^{n-k-1}(1-x)^{k+1} dx \\
&= \langle v_i, v_j \rangle_V .
\end{aligned}$$

whenever $i + j = l$ for $l = 0, \dots, k-1$. Hence, the Gram matrix $H_{k,n}$ associated to

$\{v_i\}_{i=0}^{(k-1)/2}$ in V is positive definite. Similarly,

$$\begin{aligned}\frac{1}{\frac{k}{n}\binom{n}{k}}\mu_{l+1} &= \int_0^1 \left(\sqrt{n-k}\frac{1}{(1-x)^{i+1}}\right)\left(\sqrt{n-k}\frac{1}{(1-x)^{j+1}}\right)x^{n-k-1}(1-x)^{k+1}dx \\ &= \langle w_i, w_j\rangle_V\end{aligned}$$

whenever $i+j=l$ for $l=0,\ldots,k-2$, so $H'_{k,n}$ associated to $\{w_j\}_{j=0}^{(k-1)/2-1}$ in V is also positive definite. ∎

Here, we note two things: first, we assumed k is odd in showing the existence of Ξ. The case when k is even is handled similarly, so we omit it for brevity. The main theoretical difference is that we use Theorem 5.1 of [7] (and therefore that the distribution Ξ thus obtained is actually unique, but that is irrelevant for our examples) instead of Theorem 5.3. Second, computing a distribution as in Example 1 may be no small task for large values of k and may take a lot of computing power. Nevertheless, since Ξ does not depend on the specific choice of κ as long as κ satisfies the condition of Equation (2.4), we may precompute the values Ξ for each combination of values of k, n. This is the "preprocessing step" alluded to in the introduction.

3. Numerical experiments. The last proposition thus completes an answer for how, given $X=\{x_1,\ldots,x_n\}$ with $x_i \in U[0,1]$ for $1\le i\le n$ and $A=\kappa(X,X)$, we may design a framework for obtaining a matrix B such that Corollary 2.2 applies in expectation. Namely, we will fix k and n, precompute Ξ as in Proposition 2.4 above, and then take $B=\kappa(Y,Y)$, where the $Y=\{y_1,\ldots,y_k\}$ is defined as in Proposition 2.3 using the distribution of Proposition 2.4. That is, Y is the set obtained by multiplying a random subsample of X by a random scalar picked using Ξ. Because this way of obtaining Y is probabilistic and only guarantees moment matching in expectation, we thus need to find the average of the jth largest eigenvalue of B, for $1\le j\le k$, for a number of trials m of forming such matrices B. Even though Ξ depends on n and k, empirically m seems to depends on k alone. The average $\sigma_j(B)$s should then correspond to bounds for the k quantiles of the eigenvalues of A as in (2.1). First, we look at the performance of this framework for Ξ as computed in Example 1 (that is, we set $n=49$ and $k=7$):

EXAMPLE 2. *Let $n=49$, $k=7$, $d=1$, and $\kappa:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be defined by $\kappa(x,y)=e^{-1000(x-y)^2}$. Since n is so small in this case, we perform 10 trials of forming $A=(X,X)$ and average the jth largest eigenvalue for $1\le j\le n$. We then perform $m=256000$ trials of forming $B=(Y,Y)$ according to the scheme in Proposition 2.3 using the distribution from Proposition 2.4, and we average the jth largest eigenvalue thus obtained for $1\le j\le k$. The resulting averaged eigenvalues of A are plotted in Figure 3.1, along with the eigenvalue quantile bounds obtained from the averaged eigenvalues of B. (We repeat each eigenvalue of B $49/7=7$ times in order to better visualize the quantile bounds given for the eigenvalues of A in Corollary 2.2, as in Figure 2.1.)*

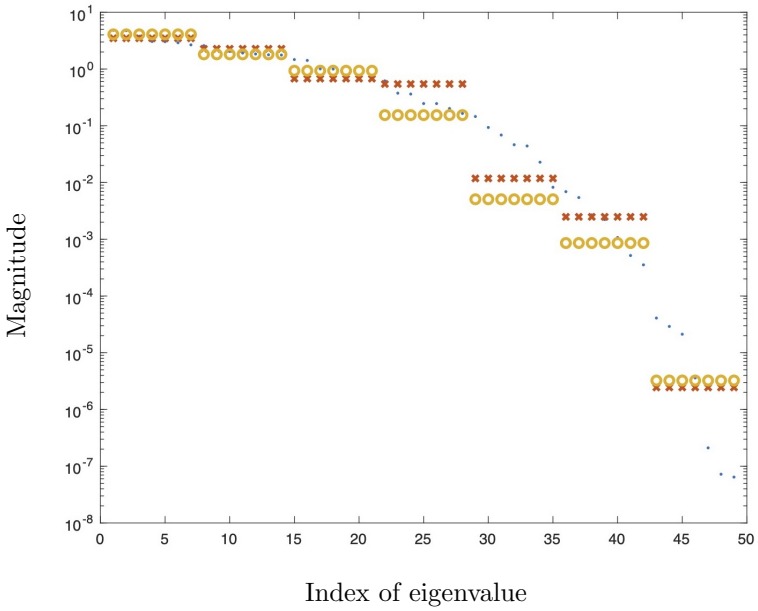

FIG. 3.1. *The averaged eigenvalues of A (blue dots) together with the repeated, averaged eigenvalues of B (red crosses and yellow circles), formed as in Example 2. Two averages over $m = 256000$ runs of finding B are shown, illustrating the variation inherent to our framework.*

Note the length scale of κ: setting κ to have such quick decay away from the diagonal seems to be necessary to have a meaningful correlation between the quantile bounds obtained from the eigenvalue distribution of B for the eigenvalue distribution of A. We will see in Example 5 what happens with our framework if this is not the case. Also, as we noted in Section 1, setting $d = 1$ as in Example 2 obviates the need for our approximation, since the resulting matrix A may be approximated for our choice of Gaussian kernel using a banded matrix. Therefore, it may be more illustrative to set d to something larger than one to better showcase the strengths of the framework. We do so in the next two examples.

EXAMPLE 3. *We set $n = 729$, $k = 9$, $d = 3$, and $\kappa : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ be defined by $\kappa(\mathbf{x}, \mathbf{y}) = e^{-500(|\mathbf{x}-\mathbf{y}|)^2}$. As before, we perform 10 trials of forming $A = (X, X)$ and average the jth largest eigenvalue for $1 \leq j \leq n$. We perform $m = 128000$ trials of forming $B = (Y, Y)$ as in the previous examples and average the jth largest eigenvalue thus obtained for $1 \leq j \leq k$. The resulting averaged eigenvalues of A are plotted in Figure 3.2, along with the eigenvalue quantile bounds obtained from the averaged eigenvalues of B. (As before, we repeat each eigenvalue of B $729/9 = 81$ times in order to visualize the quantile bounds given for the eigenvalues of A in Corollary 2.2.)*

In this previous example, setting d equal to 3 means that it is impossible to approximate A by a (singly) banded matrix. We will continue showing the efficacy of our framework for points X with an even higher dimension in the next example. Finally, we note that the kernel used does not have to have any particular form (i.e. we take κ to be the Cauchy kernel instead of the Gaussian kernel), as long as the steep decay away from the diagonal is maintained.

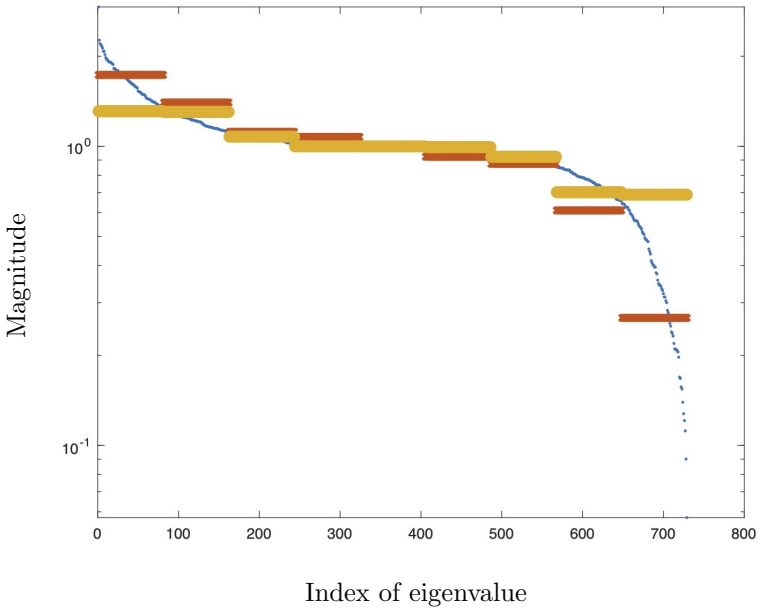

FIG. 3.2. *The averaged eigenvalues of A (blue dots) together with the repeated, averaged eigenvalues of B (red crosses and yellow circles), formed as in Example 3. Two averages over $m = 128000$ runs of finding B are shown, illustrating the variation inherent to our framework.*

EXAMPLE 4. *We set $n = 729$, $k = 9$, $d = 6$, and $\kappa : \mathbb{R}^7 \times \mathbb{R}^7 \to \mathbb{R}$ be defined by $\kappa(\mathbf{x}, \mathbf{y}) = 1/(1 + 10000(|\mathbf{x} - \mathbf{y}|)^2)$. As before, we perform 10 trials of forming $A = (X, X)$ and average the jth largest eigenvalue for $1 \leq j \leq n$. We perform $m = 128000$ trials of forming $B = (Y, Y)$ and average the jth largest eigenvalue thus obtained for $1 \leq j \leq k$. The resulting averaged eigenvalues of A are plotted in Figure 3.3, along with the eigenvalue quantile bounds obtained from the averaged eigenvalues of B. (We repeat each eigenvalue of B $729/9 = 81$ times in order to visualize the quantile bounds given for the eigenvalues of A in Corollary 2.2.)*

Finally, we will note what happens if the fast decay away from the diagonal in Equation (2.4) is not satisfied: in the next example, we set all parameters equal to those of Example 3, except the dimension of the points X is set to be 1 instead of 3.

EXAMPLE 5. *Figure 3.4 shows what happens when the setup is kept exactly the same as in Example 3, except for setting $d = 1$. Observe that there seems to be no correlation whatsoever between the eigenvalues of B and quantile bounds for A, which we may attribute to a lack of decay of κ away from the diagonal as required by (2.4). (Note that A has low numerical rank here.)*

We thus note here that, for higher dimensions, Examples 3 and 5 indicate that the length scale involved in κ does not have to be quite as small in higher dimensions as in does in lower dimensions for fast decay to be satisfied. This corresponds to the well-known (but unintuitive) heuristic that unit balls in high dimension are "concentrated near the axes." This last example therefore also illustrates the limitations of our framework.

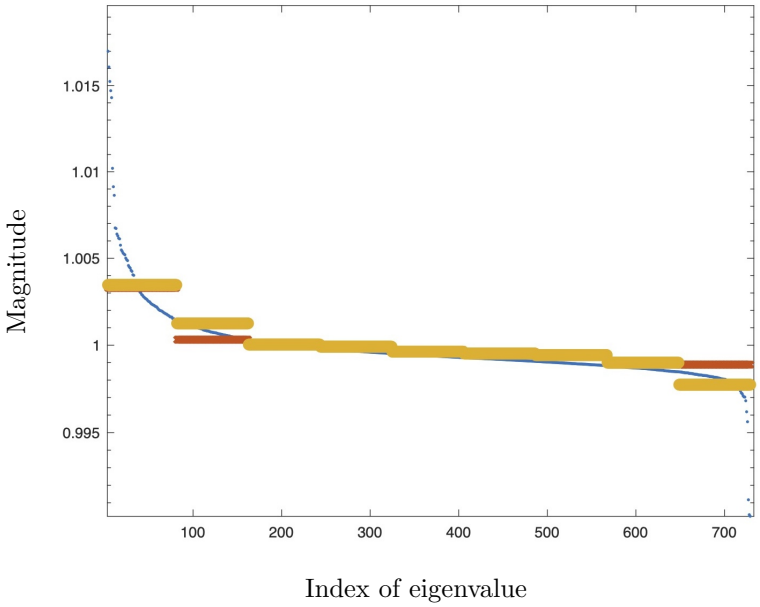

FIG. 3.3. *The averaged eigenvalues of A (blue dots) together with the repeated, averaged eigenvalues of B (red crosses and yellow circles), formed as in Example 4. Two averages over $m = 8000$ runs of finding B are shown, illustrating the variation inherent to our framework.*

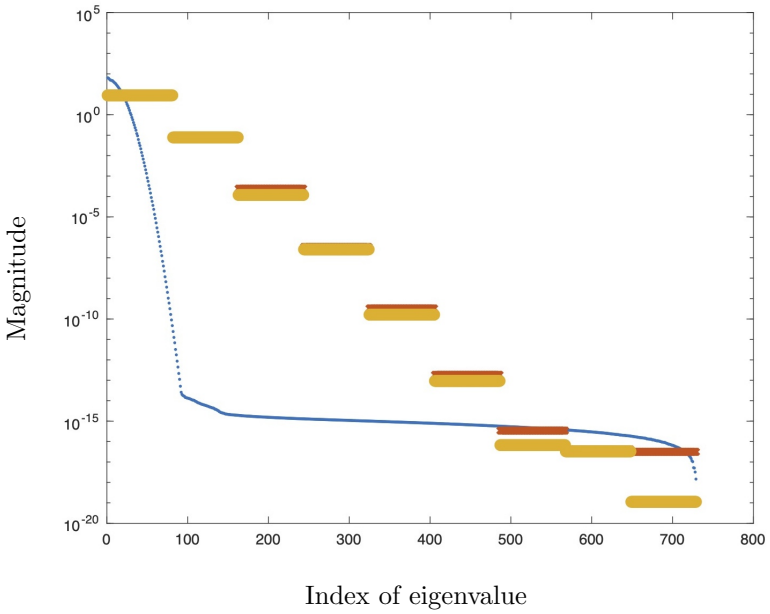

FIG. 3.4. *The averaged eigenvalues of A (blue dots) together with the repeated, averaged eigenvalues of B (red crosses and yellow circles), formed as in Example 5. Two averages over $m = 128000$ runs of finding B are shown, illustrating the variation inherent to our framework.*

4. Conclusion and future work. We have introduced a new framework that aims to provide a way to approximate the eigenvalues of a kernel matrix evaluated at sets of n points X which come from standard uniform distributions on $\mathbb{R}^d$ without having to form the full kernel matrix itself. In particular, after fixing k, our framework provides bounds in expectation on the k spectrum quantiles of the kernel matrix A. Since we do not require forming the full matrix A, for $k \ll n$, this new framework allows us to find such bounds in subquadratic time relative to n. In particular, it requires $O(mk^2)$ steps, where m is the number of times we form B. However, our work includes a number of limitations that we aim to overcome in the future. We go over these limitations one by one, and mention which directions to take to address them.

First, our work so far concerned only points which come from the uniform distribution on $\mathbb{R}^d$. However, we may extend this work to consider any compactly-supported, absolutely continuous distribution Ω by composing κ with an appropriate coordinate transformation, which in turn may be obtained from the CDF of Ω. In doing so, for our framework to work, we must ensure that an analog of the condition of Equation (2.4) is adequately satisfied on this composition of functions. A future study of commonly-used distributions (for example, the multivariate normal distribution) will be useful in finding empirical and analytic evidence for when this is the case.

Second, the distribution Ξ provided by Proposition 2.4 seems to require a lot of trials of forming, finding the eigenvalues of, and then averaging B in order to get a good approximation for the quantiles of A. In other words, the constant m is high, even if it does not depend explicitly on n. This seems to be because the probabilities of some of the scalar multiples appear to be quite low in general. For example, in Example 1, we require each coordinate of $\mathbf{x}$ to be multiplied by $\delta = 130.90$ with probability $d = 1.4709 \cdot 10^{-6}$. Another disadvantage of Ξ from Proposition 2.4 is that precomputing the relevant values of $\mathbf{z}$ and their probabilities is computationally expensive and becomes infeasible for large k. This distribution, however, is only one distribution that satisfies (2.3). We know from [7] that there is not even a unique discrete distribution satisfying Equation (2.3); furthermore, there may potentially be continuous distributions satisfying Equation (2.3) that are easier to compute with for our purpose. Thus, we would like to know if such distributions exist which cause our quantile estimates to converge to their expectation with fewer trials than Ξ requires. If we obtain such distributions which require asymptotically fewer than $O(n^2)$ trials, we would be guaranteed to find bounds for the quantiles of the eigenvalues of A in provably subquadratic time. Furthermore, the approach of [16] may allow us solve the moment problem for Ξ approximately and with less computational cost, and then to then find perturbative bounds from a "true solution" in the Wasserstein-1 distance.

In its present form, however, our work may already be applicable to the question of locally finding the so-called intrinsic dimension of data. Namely, the *manifold hypothesis* in data science is that real-world data embedded in high-dimensional space, such as collections of 64-by-64-pixel images with certain properties (for example, containing a dog) embedded in the space of all 64-by-64-pixel images, actually reside on some kind of lower-dimensional manifold. Often, this is stated up to some perturbation from the addition of "noise." This idea, taken literally for the case of a $\mathcal{C}^2$-manifold, was tested in [10]. Less literal but more practical mathematical formulations of this idea of a "latent dimension" are explored in [17], as well as the resulting estimates for each notion of dimension.

Here, we propose a new such formulation. Until now, we have not paid much attention to the parameter d used in the definition of the y_j's in Proposition 2.3.

However, this parameter is key to obtaining good quantile bounds for the kernel matrix A formed from X. See, for example, Figure 4.1 for the result of setting $d = 2$ or $d = 4$ instead of $d = 3$ when forming B in the setup of Example 3.

Therefore, if we start with the collection of points X restricted to a small volume V in $\mathbb{R}^d$ and wish to find the (local) dimension of the piece of a manifold where that part of X "truly lives," as the manifold hypothesis stipulates, we can use our eigenvalue quantile estimation technique to see if we get accurate bounds after setting d to several candidate values. That is, we could sample e.g. $n = 49$ and $k = 7$ points and sees which value of d works best to give quantile estimates. In doing so, we would be assuming that our points are "locally uniformly" distributed (i.e. uniform on an appropriate, small-enough chart of some manifold), and that the embedding generating X restricted to V guarantees that $\kappa(x, y)$ is far from 0 only for points x and y that are close within the latent manifold. In making these assumptions, this setup could effectively test a "local" manifold hypothesis. In addition, because of its locality, this notion of dimension is likely related to various existing k-nearest-neighbor-type estimators for intrinsic dimension [9].

Acknowledgements. The author sincerely thanks Levon Nurbekyan for his support and ideas related to the theoretical aspects of this work, as well as Yuanzhe Xi and Shifan Zhao for the overall direction of the work and several key observations motivating it.

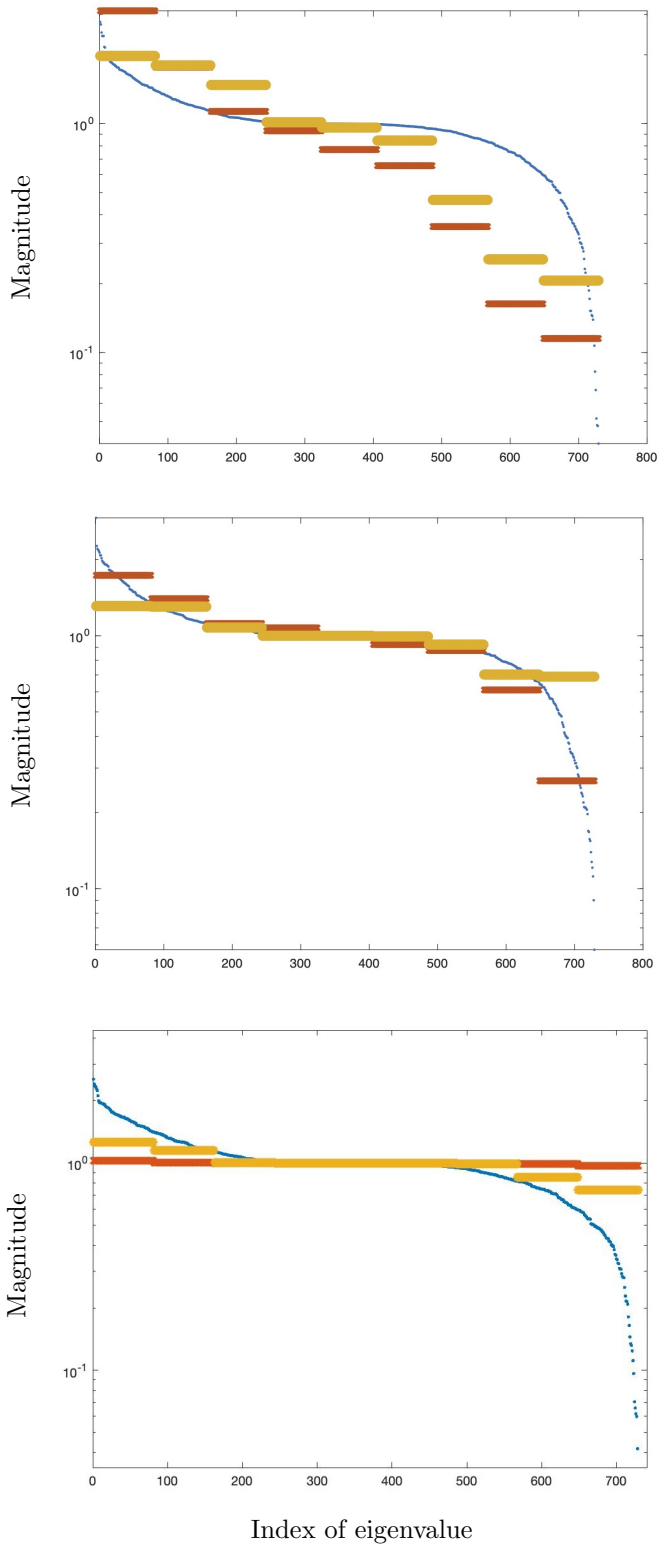

FIG. 4.1. *Here, we reproduce Figure 3.2 of Example 3 as the middle figure, which shows good quantile estimates. In the top and bottom figures, our setup is exactly the same as in Example 3, except we set $d = 2$ (top figure) and $d = 4$ (bottom figure) when forming B. Since these are the wrong values of d, we get worse quantile estimates in the top and bottom figures.*

REFERENCES

[1] N. I. Akhiezer *The Classical Moment Problem and Some Related Questions in Analysis*, 1st ed., Oliver & Boyd, Ltd., London, England, 1965.

[2] A. Bakshi, P. Indyk, P. Kacham, S. Silwal, and S. Zhou, *Subquadratic algorithms for kernel matrices via kernel density estimation*, preprint, arXiv:2212.00642v1 [cs.LG], 2022.

[3] R. Bhattacharjee, G. Dexter, P. Drineas, C. Musco, and A. Ray *Sublinear time eigenvalue approximation via random sampling*, Algorithmica 86 (2024), pp. 1764–1829.

[4] M. Braun *Accurate error bounds for the eigenvalues of the kernel matrix*, J. Mach. Learn. Res. 7 (2006), pp. 2303–2328.

[5] D. Cohen-Steiner and W. Kong, C. Sohler *Approximating the spectrum of a graph*, KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018), pp. 1263–1271.

[6] N. Cesa-Bianchi, Y. Mansour, and O. Shamir *On the complexity of learning with kernels*, Proceedings of Machine Learning Research 40 (2015), pp. 297-395.

[7] R. E. Curtow and L. A. Fiaklow *Recursiveness, positivity, and truncated moment problems*, Houston J. Math. 17(4) (1991), pp. 603–635.

[8] P. Drineas and M. W. Mahoney *On the Nyström method for approximating a Gram matrix for improved kernel-based learning*, J. Mach. Learn. Res. 6 (2005), pp. 2153–2175.

[9] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*, Sci. Rep. 7:12140 (2017).

[10] C. Fefferman, S. Mitter, H. Narayanan *Testing the manifold hypothesis*, J. Amer. Math. Soc. 29 (2016), pp. 983-1049.

[11] P.N. Gavriliadis, G.A. Athanassoulis *Moment information for probability distributions, without solving the moment problem, II: Main-mass, tails, and shape approximation*, J. Comput. Appl. Math. 229 (2009), pp. 7–15.

[12] A. Gittens and M. W. Mahoney *Revisiting the Nyström method for improved large-scale machine learning*, J. Mach. Learn. Res. 17 (2016), pp. 1–65.

[13] G. H. Golub, C. F. Van Loan *Matrix Computations*, the Johns Hopkins University Press, Baltimore (1996), 3rd ed.

[14] Y. Jin, C. Musco, A. Sidford, and A. V. Singh *Moments, random walks, and limits for spectrum approximation*, Proceedings of Machine Learning Research 195 (2023), pp. 1–22.

[15] V. Koltchinskii and E. Giné *Random matrix approximation of spectra of integral operators*, Bernoulli, 6(1) (2000), pp. 113–167.

[16] C. Musco, C. Musco, L. Rosenblatt, A. V. Singh *Sharper bounds for Chebyshev moment matching with applications to differential privacy and beyond*, pre-print.

[17] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldblum, *The intrinsic dimension of images and its impact on learning*, ICLR 2021: 9th International Conference on Learning Representations (2021).

[18] B. Schölkopf, A. J. Smola *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge (2001).

[19] J. Shaw-Taylor, N. Cristianini *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge (2004).

[20] W. Swatsworth and D. P. Woodruff *Optimal Eigenvalue Approximation via Sketching*, STOC 2023: Proceedings of the 55th Annual ACM Symposium on Theory of Computing (2023), pp. 145–155.

[21] C. Williams and M. Seeger *Using the Nyström method to speed up kernel machines*, Advances in Neural Information Processing Systems 13 (2001), pp. 682–688.

[22] D. P. Woodruff *Sketching as a tool for numerical linear algebra*, Found. Trends Signal Process. 10(1-2) (2014), pp. 1–157.