On the Role of Eigendecomposition in Kernel Embedding

J. Jon Ryu Jiun-Ting Huang Young-Han Kim
Department of Electrical and Computer Engineering
University of California, San Diego

***Abstract*—This paper proposes a special variant of Laplacian eigenmaps, whose solution is characterized by the underlying density and the eigenfunctions of the associated Hilbert–Schmidt operator of a similarity kernel function. In contrast to existing kernel-based spectral methods such as kernel principal component analysis and Laplacian eigenmaps, the new embedding algorithm only involves estimating density at each query point without any eigendecomposition of a matrix. A concrete example of dot-product kernels over hypersphere is provided to illustrate the applicability of the proposed framework.**

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517746

I. Introduction

Finding a good embedding of data for discovering *meaningful structures* is one of the fundamental problems in machine learning and data science, with important applications such as clustering, dimensionality reduction, and data visualization. Among a myriad of algorithms which have been proposed in the last few decades, we particularly focus on a class of kernel-based spectral embedding algorithms, which find embedding of data based on eigenvectors of data-dependent similarity kernel matrices [2, 8]—this class subsumes kernel principal component analysis (PCA) [19], Laplacian eigenmaps [1], spectral clustering [15, 20], multidimensional scaling [4], locally linear embedding [17], and Isomap [22]. Proven to be extremely powerful in various applications, the common disadvantage of such methods is the computational complexity of eigendecomposition of a kernel matrix, which could be prohibitively large in big data analysis.

As an attempt to resolve the computational bottleneck, in this paper, we propose a new kernel embedding framework, which suggests a sample based embedding algorithm without eigendecomposition of a matrix for special choices of kernels. To motivate our approach, we first review kernel PCA and introduce Laplacian eigenmaps as a special case of kernel PCA framework with a kernel with density regularization in Section II. In Section III, we then propose and study a new density-regularized kernel, which separates the underlying density and spectral decomposition of the kernel operator. We describe the resulting sample based algorithm, which simply combines density estimates given sample and known eigenfunctions of a kernel operator. In Section IV, dot-product kernels over hypersphere are discussed as a concrete example to which the proposed embedding framework may apply. We briefly discuss relevant literature in Section VI.

Notation Throughout the paper, we assume that a random vector $\mathbf{X}$ is drawn from density p over a closed subset $\mathcal{X} \subset \mathbb{R}^d$, and data points $\mathbf{x}_{1:N} := \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ are independently and identically distributed (i.i.d.) random variables drawn from p. Given $\mathcal{X} \subset \mathbb{R}^d$ and a density μ on $\mathcal{X}$, we consider a Hilbert space $L^2_\mu(\mathcal{X}) := \{f\colon \mathcal{X} \to \mathbb{C} | \int |f(\mathbf{x})|^2 \,\mathrm{d}\mu(\mathbf{x}) < \infty\}$ with inner product $\langle f, g\rangle_\mu := \int f(\mathbf{x})\overline{g(\mathbf{x})}\,\mathrm{d}\mu(\mathbf{x})$. For a kernel function $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, we denote the *associated Hilbert–Schmidt integral operator* in boldface $\mathbf{K}\colon L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$, which is defined as $(\mathbf{K}f)(\mathbf{x}) := \int_{\mathcal{X}} k(\mathbf{x},\mathbf{t}) f(\mathbf{t})\,\mathrm{d}\mu(\mathbf{t})$. In what follows, we always assume that a kernel is symmetric, i.e., $k(\mathbf{x},\mathbf{t}) = k(\mathbf{t},\mathbf{x})$, and satisfies $\iint_{\mathcal{X}\times\mathcal{X}} k^2(\mathbf{x},\mathbf{t})\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(\mathbf{t}) < \infty$, so that the operator $\mathbf{K}$ is self-adjoint and compact.

II. Review of kerenl PCA and Laplacian eigenmaps

A. Kernel PCA

1) Feature space formulation: Kernel PCA [19] was proposed as an efficient method to perform PCA over transformed samples with a given nonlinear mapping. Let $|\phi(\cdot)\rangle\colon \mathcal{X} \to \mathcal{F}$ be a feature map that maps a data point $\mathbf{x}$ to a point in a feature space $|\phi(\mathbf{x})\rangle \in \mathcal{F}$, where $\mathcal{F}$ is a vector space with inner product $\langle\cdot|\cdot\rangle$.[1] For simplicity, assume for now that $\mathsf{E}[|\phi(\mathbf{X})\rangle] = |0\rangle$. Kernel PCA aims to perform PCA over the lifted random vector $|\phi(\mathbf{X})\rangle$, that is, to solve

$$\begin{aligned} \underset{|u_\ell\rangle\in\mathcal{F}}{\text{maximize}} \quad & \sum_{\ell=1}^{L} \langle u_\ell|\mathbf{C}_\phi|u_\ell\rangle \\ \text{subject to} \quad & \langle u_\ell|u_{\ell'}\rangle = \delta_{\ell\ell'} \end{aligned} \tag{1}$$

Here, $\mathbf{C}_\phi := \mathsf{E}[|\phi(\mathbf{X})\rangle\langle\phi(\mathbf{X})|]$ denotes the covariance operator of $|\phi(\mathbf{X})\rangle$. We call this the *(population) feature space problem* of kernel PCA.[2] When $\mathcal{F}$ is high- or infinite-dimensional, it is often not feasible to directly solve this problem.

2) Function space formulation: To avoid the issue with high-dimensionality of the feature space $\mathcal{F}$, we can convert the feature space problem (1) into an equivalent optimization problem over a *function space* by the so-called *kernel trick* as follows. Define a symmetric kernel function $k(\mathbf{x},\mathbf{t}) := \langle\phi(\mathbf{x})|\phi(\mathbf{t})\rangle$. Consider the following optimization problem

$$\begin{aligned} \underset{f_\ell\in L^2_p(\mathcal{X})}{\text{maximize}} \quad & \sum_{\ell=1}^{L} \langle f_\ell, \mathbf{K}f_\ell\rangle_p \\ \text{subject to} \quad & \langle f_\ell, f_{\ell'}\rangle_p = \delta_{\ell\ell'} \end{aligned} \tag{2}$$

[1] We use the bra-ket notation to note that $\mathcal{F}$ may be infinite dimensional.
[2] If $|\phi(\mathbf{x})\rangle = \mathbf{x}$, it boils down to the original PCA.

Since $\mathbf{K}$ is self-adjoint and compact, the solution is characterized by the top-L eigenfunctions and eigenvalues of $\mathbf{K}$; see, e.g., [3, Proposition A.2.10]. The following proposition establishes the equivalence between (1) and (2); the proof is easy and thus omitted.

Proposition 1. *Let $\lambda_1, \dots, \lambda_L$ and $|u_1^\star\rangle, \dots, |u_L^\star\rangle$ be the top-L eigenvalues and orthonormal eigenvectors of the operator $\mathbf{C}_\phi$, respectively. Let $\mu_1, \dots, \mu_L$ and $f_1^\star, \dots, f_L^\star$ be the top-L eigenvalues and orthonormal eigenfunctions of the operator $\mathbf{K}$, respectively. Then, $\lambda_\ell = \mu_\ell$,*

$$f_\ell^\star(\mathbf{x}) = \frac{1}{\sqrt{\lambda_\ell}} \langle \phi(\mathbf{x}) | u_\ell^\star \rangle, \text{ and} \tag{3}$$

$$|u_\ell^\star\rangle = \frac{1}{\sqrt{\lambda_\ell}} \int f_\ell^\star(\mathbf{x}) |\phi(\mathbf{x})\rangle p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{4}$$

for each $\ell \in [L]$.

Hence, we call this problem (6) as the *(population) function space problem* of kernel PCA. If the top-L eigenfunctions $f_1^\star, \dots, f_L^\star$ of the operator $\mathbf{K}$ are given, then the embedding of a query point $\mathbf{x}$ by kernel PCA is the projection of the lifted data $|\phi(\mathbf{x})\rangle$ onto the principal directions $|u_1^\star\rangle, \dots, |u_L^\star\rangle$, or equivalently in view of (3),

$$\psi_{\mathsf{KPCA}}(\mathbf{x}) := [\sqrt{\lambda_1} f_1^\star(\mathbf{x}), \dots, \sqrt{\lambda_L} f_L^\star(\mathbf{x})]^T. \tag{5}$$

3) Sample solution: The spectral decomposition of $\mathbf{K}$ in $L_p^2(\mathcal{X})$ cannot be performed directly in general even if the density p is known. Given sample $\mathbf{x}_{1:N}$, we can approximately solve (2) in practice. Let $\mathsf{K} \in \mathbb{R}^{N \times N}$ denote the *sample kernel matrix* whose (m,n)-th entry is $(\mathsf{K})_{mn} = k(\mathbf{x}_m, \mathbf{x}_n)$. Then, we can solve

$$\begin{aligned} \underset{\mathbf{f}_\ell \in \mathbb{R}^N}{\text{maximize}} \quad & \sum_{\ell=1}^{L} \frac{\mathbf{f}_\ell^T}{\sqrt{N}} \frac{\mathsf{K}}{N} \frac{\mathbf{f}_\ell}{\sqrt{N}} \\ \text{subject to} \quad & \frac{\mathbf{f}_\ell^T}{\sqrt{N}} \frac{\mathbf{f}_{\ell'}}{\sqrt{N}} = \delta_{\ell\ell'} \end{aligned} \tag{6}$$

as a proxy to (2), which is equivalent to the eigendecomposition of K. The optimal solution is characterized by the top-L eigenvectors $\mathbf{f}_1^\star, \dots, \mathbf{f}_L^\star \in \mathbb{R}^N$ of the normalized sample kernel matrix K/N with eigenvalues $\lambda_1, \dots, \lambda_L$ with norm $\|\mathbf{f}_\ell^\star\|_2 = \sqrt{N}$. The L-dimensional embedding of a point $\mathbf{x}$ is then

$$\hat{\psi}_{\mathsf{KPCA}}(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} k(\mathbf{x}, \mathbf{x}_i) \Big[\frac{(\mathbf{f}_1^\star)_i}{\sqrt{\lambda_1}}, \dots, \frac{(\mathbf{f}_L^\star)_i}{\sqrt{\lambda_L}} \Big]^T. \tag{7}$$

This is often referred to the *Nyström formula*; see, e.g., [2]. In particular, for a sample point $\mathbf{x}_n$, the embedding is simply

$$\hat{\psi}_{\mathsf{KPCA}}(\mathbf{x}_n) := [\sqrt{\lambda_1} (\mathbf{f}_1^\star)_n, \dots, \sqrt{\lambda_L} (\mathbf{f}_L^\star)_n]^T.$$

We refer to kernel PCA as the procedure consisting of the eigendecomposition of the kernel matrix K and the embedding (7).

Remark 2 (Centering)**.** In (1), (2), (6), and (7), we assume $\mathsf{E}[|\phi(\mathbf{X})\rangle] = |0\rangle$. Hence, given sample $\mathbf{x}_{1:N}$, we need to center the sample kernel matrix K as $\mathsf{K}_c = (I_N - 1_N)\mathsf{K}(I_N - 1_N) \in \mathbb{R}^{N \times N}$, where $1_N := \frac{1}{N} \mathbb{1}_N \mathbb{1}_N^T \in \mathbb{R}^{N \times N}$.

Remark 3. In practice, any choice of symmetric kernel function k can be deployed in kernel PCA. Note, however, that the feature space formulation and PCA interpretation via Proposition 1 remain valid if and only if the kernel is in the form $k(\mathbf{x}, \mathbf{t}) = \langle \phi(\mathbf{x}) | \phi(\mathbf{t}) \rangle$ for some inner product space $\mathcal{F}$ and function $\phi \colon \mathcal{X} \to \mathcal{F}$. Mercer's theorem [12] establishes positive definiteness of a kernel as an equivalent condition for the existence of such a mapping.

B. Laplacian eigenmaps

Laplacian eigenmaps [1] is one of the most popular embedding algorithm, which can be justified as an approximation of the Laplacian–Beltrami operator or a relaxed solution to the graph min-cut problem [20]. Here, we introduce Laplacian eigenmaps as a special instance of kernel PCA. Given a base symmetric kernel function k, we first define the *kernelized density* $p_k(\mathbf{x}) := \int k(\mathbf{x}, \mathbf{t}) p(\mathbf{t}) \, \mathrm{d}\mathbf{t}$ and define a new kernel function as

$$\overline{k}_p(\mathbf{x}, \mathbf{t}) := \frac{k(\mathbf{x}, \mathbf{t})}{\sqrt{p_k(\mathbf{x}) p_k(\mathbf{t})}}.$$

The (kernelized) Laplacian eigenmaps with the base kernel k is characterized by the population function space optimization problem (2) of kernel PCA with the kernel $\overline{k}_p$ which is the function space optimization problem (6) of kernel PCA with the kernel $\overline{k}_p(\mathbf{x}, \mathbf{t})$. Let $f_1^\star, \dots, f_L^\star$ denote the top-L orthonormal eigenfunctions of the operator $\overline{\mathbf{K}}_p$. Then, the Laplacian eigenmaps of a point $\mathbf{x}$ is defined as the evaluations of the eigenfunctions:

$$\psi_{\mathsf{LE}}(\mathbf{x}) := [f_1^\star(\mathbf{x}), \dots, f_L^\star(\mathbf{x})]^T. \tag{8}$$

As in kernel PCA, given samples $\mathbf{x}_{1:N}$, we perform eigendecomposition of the sample kernel matrix $\overline{\mathsf{K}}_p$ defined as

$$(\overline{\mathsf{K}}_p)_{ij} := \frac{k(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\hat{p}_k(\mathbf{x}_i) \hat{p}_k(\mathbf{x}_j)}},$$

where $\hat{p}_k(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^N k(\mathbf{x}, \mathbf{x}_i)$ denotes the empirical estimate of the kernelized density $p_k(\mathbf{x})$. The embedding by Laplacian eigenmaps of a sample point $\mathbf{x}_n$ is then

$$\hat{\psi}_{\mathsf{LE}}(\mathbf{x}_n) := [(\mathbf{f}_1^\star)_n, \dots, (\mathbf{f}_L^\star)_n]^T.$$

Remark 4. Despite the apparent mathematical equivalence, kernel PCA embedding with the kernel $\overline{k}_p$ may differ from Laplacian eigenmaps embedding significantly due to centering in kernel PCA (see Remark 2) and the different definitions of embeddings (see (5) and (8)).

Remark 5. Laplacian eigenmaps may also use the k-th neighborhood adjacency matrix instead of kernel-based weight matrix given samples—however, it does not fit to the current population optimization framework, and thus studying this version is beyond the scope of the paper.

Remark 6. Eigendecomposition of $\overline{\mathsf{K}}_p$ is approximately equivalent to the that of the *symmetric normalized graph Laplacian*,

which is typically used in spectral embedding. Define a weight matrix $\mathsf{W} \in \mathbb{R}^{N\times N}$ as $(\mathsf{W})_{ij} = (1-\delta_{ij})k(\mathbf{x}_i,\mathbf{x}_j)$ and define a degree matrix D as the diagonal matrix with entry $(\mathsf{D})_{ii} = \sum_{j=1}^{N}(\mathsf{W})_{ij}$. The symmetric normalized graph Laplacian is then defined as $\mathsf{L}_{\mathsf{sym}} := \mathsf{D}^{-1/2}\mathsf{W}\mathsf{D}^{-1/2}$. Since the difference

$$\overline{\mathsf{K}}_p - \mathsf{L}_{\mathsf{sym}} = \frac{1}{N}\operatorname{diag}\Big(\frac{k(\mathbf{x}_1,\mathbf{x}_1)}{\hat{p}_k(\mathbf{x}_1)},\dots,\frac{k(\mathbf{x}_N,\mathbf{x}_N)}{\hat{p}_k(\mathbf{x}_N)}\Big)$$

vanishes in the operator norm as $N\to\infty$, eigendecomposition of $\overline{\mathsf{K}}_p$ becomes equivalent to that of $\mathsf{L}_{\mathsf{sym}}$ in the sample limit.

III. Kernel embedding without eigendecomposition

A. A new density-regularized kernel

So far, we reviewed the two important kernel-based embedding frameworks, kernel PCA and Laplacian eigenmaps: Laplacian eigenmaps fits into the framework of kernel PCA with a specific form of density-regularized kernel $\overline{k}_p$; see Table I. The population problems cannot be solved directly, but given sample, we can approximately solve them via eigendecomposition of a matrix of possibly large size.

In this section, motivated by the form of the kernel $\overline{k}_p$ of Laplacian eigenmaps, we introduce a new kernel function

$$k_p(\mathbf{x},\mathbf{t}) := \frac{k(\mathbf{x},\mathbf{t})}{\sqrt{p(\mathbf{x})p(\mathbf{t})}} \tag{9}$$

and propose the population function space optimization problem (2) of kernel PCA with k_p, that is,

$$\begin{aligned}\underset{f_\ell\in L_p^2(\mathcal{X})}{\text{maximize}} \quad & \sum_{\ell=1}^{L}\langle f_\ell, \mathbf{K}_p f_\ell\rangle_p \\ \text{subject to} \quad & \langle f_\ell, f_{\ell'}\rangle_p = \delta_{\ell\ell'}\end{aligned} \tag{10}$$

as a new criterion for kernel embedding. Compared to the kernel $\overline{k}_p$ of Laplacian eigenmaps, the base kernel function k is now regularized by the true underlying density p instead of the kernelized density p_k.

With the new kernel k_p, we can reshape the population optimization problem (10) into a much simpler form. For a weighting function $w\colon \mathcal{X}\to\mathbb{R}_+$ whose support subsumes the support of p, we define the density-scaled function

$$g_\ell(\mathbf{x}) := \sqrt{\frac{p(\mathbf{x})}{w(\mathbf{x})}} f_\ell(\mathbf{x}). \tag{11}$$

Note that if $f_\ell \in L_p^2(\mathcal{X})$, then $g_\ell \in L_w^2(\mathcal{X})$. If we define $k_w(\mathbf{x},\mathbf{t}) := k(\mathbf{x},\mathbf{t})/\sqrt{w(\mathbf{x})w(\mathbf{t})}$, we have

$$\langle f_\ell, \mathbf{K}_p f_\ell\rangle_p = \langle g_\ell, \mathbf{K}_w g_\ell\rangle_w \quad \text{and} \quad \langle f_\ell, f_{\ell'}\rangle_p = \langle g_\ell, g_{\ell'}\rangle_w,$$

which imply that the new problem (10) can be recast as

$$\begin{aligned}\underset{g_\ell\in L_w^2(\mathcal{X})}{\text{maximize}} \quad & \sum_{\ell=1}^{L}\langle g_\ell, \mathbf{K}_w g_\ell\rangle_w \\ \text{subject to} \quad & \langle g_\ell, g_{\ell'}\rangle_w = \delta_{\ell\ell'}.\end{aligned} \tag{12}$$

We remark that (12) solely depends on the choice of kernel k and the weighting function w. Provided that $\mathbf{K}_w$ is compact, the solution of this optimization problem is characterized by the top-L eigenfunctions $g_1^\star,\dots,g_L^\star$ of the operator $\mathbf{K}_w$. Somewhat surprisingly, for a few special cases, the eigenexpansion of $\mathbf{K}_w$ is given in an analytical form; see Section IV. The eigenfunctions of $\mathbf{K}_p$ are then given as the functions $f_1^\star,\dots,f_L^\star$, where $f_\ell^\star(\mathbf{x}) := \sqrt{w(\mathbf{x})/p(\mathbf{x})}g_\ell^\star(\mathbf{x})$. Provided that the density $p(\mathbf{x})$ can be evaluated, the L-dimensional embedding of a query point $\mathbf{x}$ is

$$\psi_{\mathsf{KE}}(\mathbf{x}) := \sqrt{\frac{w(\mathbf{x})}{p(\mathbf{x})}}[g_1^\star(\mathbf{x}),\dots,g_L^\star(\mathbf{x})]^T. \tag{13}$$

B. A new sample based kernel embedding

Provided that spectral decomposition of $\mathbf{K}_w$ is known for a choice of k and w, the only unknown in the embedding (13) is the density p. Hence, given sample $\mathbf{x}_{1:N}$, we only need to estimate the density, without any spectral decomposition of a matrix. This yields the following kernel embedding algorithm.

Algorithm 1 Kernel embedding without spectral decomposition

Input a base kernel k, a weighting function w, a density estimator $\hat{p}(\cdot)$, sample $\{\mathbf{x}_n\}_{n=1}^N$, a target dimension $L\in\mathbb{N}$.

1: Find the top-L orthonormal eigenfunctions $g_1^\star,\dots,g_L^\star$ of the integral operator $\mathbf{K}_w\colon L_w^2(\mathcal{X})\to L_w^2(\mathcal{X})$.
2: Given a query point $\mathbf{x}\in\mathcal{X}$, output the L-dimensional embedding of $\mathbf{x}$ as

$$\hat{\psi}_{\mathsf{KE}}(\mathbf{x}) := \sqrt{\frac{w(\mathbf{x})}{\hat{p}(\mathbf{x})}}[g_1^\star(\mathbf{x}),\dots,g_L^\star(\mathbf{x})]^T.$$

IV. Dot-product kernels over hypersphere

In this section, we focus on a special class of kernel functions of the form of $k_w(\mathbf{x},\mathbf{t}) = f(\mathbf{x}^T\mathbf{t})$ for some function $f\colon\mathbb{R}\to\mathbb{R}$, which are called *dot-product kernels*. This class contains many interesting kernels including homogeneous polynomial $f(u)=u^p$ $(p>0)$, inhomogeneous polynomial $f(u)=(1+u)^p$ $(p>0)$, Vovk's real polynomial $f(u)=(1-u^p)/(1-u)$ $(p>0)$, Vovk's infnite polynomial $f(u)=1/(1-u)$, and hyperbolic tangent $f(u)=\tanh(a+u)$ $(a\in\mathbb{R})$ kernels [21].

Further, we consider a special domain, the unit hypersphere $\mathbb{S}^{d-1} := \{\mathbf{x}\in\mathbb{R}^d\colon \|\mathbf{x}\|_2=1\}$ in $\mathbb{R}^d$. On $\mathbb{S}^{d-1}$, the class of dot-product kernels include additional popular kernels such as Gaussian kernels $f(u)=e^{-(1+u)/\sigma^2}$ $(\sigma>0)$ and arccosine kernel $f(u)=1-(2/\pi)\cos^{-1}(u)$. Note that some real-world data such as images approximately lie on a hypersphere [13, 21] and dot-product kernels may work best on $\mathbb{S}^{d-1}$ by nature. The key property of $\mathbb{S}^{d-1}$ is that with uniform weighting function w, the eigensystem of $\mathbf{K}_w$ is characterized by *spherical harmonics*.

Definition 7 (Spherical harmonics)**.** Let $\Delta = -\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ denote the Laplacian operator on $\mathbb{R}^d$. Let $\mathcal{P}_n$ denote the space of $\mathbb{C}$-valued homogeneous polynomials of degree n in d real variables. Let $\mathcal{Y}_n(d)$ denote the subspace of all homogeneous harmonics of order n, that is, $\mathcal{Y}_n(d) := \{p\in\mathcal{P}_n\colon \Delta p = 0\}$.

	Kernel PCA		Laplacian eigenmaps	Proposed
	Feature space	Function space		
Population	$\underset{\lvert u_\ell\rangle \in \mathcal{F}}{\text{maximize}} \sum_{\ell=1}^{L} \langle u_\ell \rvert \mathbf{C}_\phi \rvert u_\ell \rangle$ (1) subjecto to $\langle u_\ell \rvert u_{\ell'} \rangle = \delta_{\ell\ell'}$	$\underset{f_\ell \in L_p^2(\mathcal{X})}{\text{maximize}} \sum_{\ell=1}^{L} \langle f_\ell, \mathbf{K} f_\ell \rangle_p$ (2) subjecto to $\langle f_\ell, f_{\ell'} \rangle_p = \delta_{\ell\ell'}$	$\mathbf{K} \leftarrow \overline{\mathbf{K}}_p$	$\mathbf{K} \leftarrow \mathbf{K}_p$ (10)
Sample	$\underset{\lvert u_\ell\rangle \in \mathcal{F}}{\text{maximize}} \sum_{\ell=1}^{L} \langle u_\ell \rvert \hat{\mathbf{C}}_\phi \rvert u_\ell \rangle$ subjecto to $\langle u_\ell \rvert u_{\ell'} \rangle = \delta_{\ell\ell'}$	$\underset{\mathbf{f}_\ell \in \mathbb{R}^N}{\text{maximize}} \sum_{\ell=1}^{L} \frac{\mathbf{f}_\ell^T}{\sqrt{N}} \frac{\mathsf{K}}{N} \frac{\mathbf{f}_\ell}{\sqrt{N}}$ (6) subjecto to $\frac{\mathbf{f}_\ell^T}{\sqrt{N}} \frac{\mathbf{f}_{\ell'}}{\sqrt{N}} = \delta_{\ell\ell'}$	$\mathsf{K} \leftarrow \overline{\mathsf{K}}_p$	–

TABLE I
OVERVIEW OF POPULATION AND SAMPLE PROBLEMS OF KERNEL PCA, LAPLACIAN EIGENMAPS, AND THE PROPOSED KERNEL EMBEDDING.

The *spherical harmonics* of order n and dimension d are defined as the functions in $\mathcal{Y}_n(d)$ restricted over $\mathbb{S}^{d-1}$.

Remark 8. The dimension of the subspace $\mathcal{Y}_n(d)$ is $N(d,n) := \dim \mathcal{Y}_n(d) = \frac{2n+d-2}{n}\binom{n+d-3}{n-1}$ for $n \geq 0$.

The following elegant theorem, which is often referred to as the Funk–Hecke formula, shows that spherical harmonics fully characterize the eigenfunctions of any dot-product kernel over $\mathbb{S}^{d-1}$. Let $P_\ell^m(t)$ denote the *associated Legendre polynomial* of degree ℓ and order m for integers $0 \leq m \leq \ell$. Let $|\mathbb{S}^{d-1}| := (2\pi^{d/2})/\Gamma(d/2)$ denote the surface area of $\mathbb{S}^{d-1}$.

Theorem 9 (Funk–Hecke [14])**.** *Let $f\colon [-1,1] \to \mathbb{R}$ be a continuous function. For $Y_n \in \mathcal{Y}_n(d)$ for $n \geq 0$, we have*

$$\int_{\mathbb{S}^{d-1}} f(\mathbf{x}^T\mathbf{t}) Y_n(\mathbf{t}) \, \mathrm{d}\mathbb{S}^{d-1}(\mathbf{t}) = \lambda_n Y_n(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{S}^{d-1},$$

where $\lambda_n = |\mathbb{S}^{d-2}| \int_{-1}^{1} f(u) P_n^d(u) (1-u^2)^{\frac{d-3}{2}} \, \mathrm{d}u$.

Corollary 10. *Let $\mathcal{X} = \mathbb{S}^{d-1}$ for $d \geq 2$ and let w be the uniform density on $\mathbb{S}^{d-1}$. For any dot-product kernel of the form $k_w(\mathbf{x},\mathbf{y}) = f(\mathbf{x}^T\mathbf{y})$ for some continuous function $f\colon [-1,1] \to \mathbb{R}$, there is an orthonormal basis of $\mathcal{Y}_n(d)$ comprised by the eigenfuntions of $\mathbf{K}_w$ with eigenvalue λ_n defined in Theorem 9.*

Remark 11. Minh et al. [13, Theorems 2 and 3] computed the nonzero eigenvalues of gaussian kernels $f(u) = \exp(-(1+u)/\sigma^2)$ ($\sigma > 0$) and polynomial kernels $f(u) = (1+u)^p$ ($p \in \mathbb{N}$) in terms of special hypergeometric functions. In particular, the eigenvalues $(\lambda_n)_{n=0}^{\infty}$ of gaussian kernels are decreasing in n if $\sigma^2 \geq 2/d$, and those of the polynomial kernel of degree p are always decreasing in n and $\lambda_n = 0$ for $n \geq p+1$.

Hence, if we choose Gaussian or polynomial kernels, we only need to evaluate the first L real spherical harmonics to compute the kernel embedding (13) on $\mathcal{X} = \mathbb{S}^{d-1}$. For practical implementation, here we present a version of real orthonormal basis $\{Y_{n,j}^d(\mathbf{x})\}_{j=1}^{N(d,n)}$ of spherical harmonics of order n and dimension d [10, Section 2]. Given a point $\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{S}^{d-1}$, define a hyperspherical coordinate system $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d-1}) \in [0, 2\pi) \times [0,\pi]^{d-2}$ as

$$\theta_1 := \begin{cases} \cos^{-1} \dfrac{x_2}{\sqrt{x_1^2 + x_2^2}} & \text{if } x_1 \geq 0, \\ 2\pi - \cos^{-1} \dfrac{x_2}{\sqrt{x_1^2 + x_2^2}} & \text{if } x_1 < 0, \end{cases}$$

$$\theta_i := \cos^{-1} \frac{x_{i+1}}{\sqrt{x_1^2 + \ldots + x_{i+1}^2}}, \quad 2 \leq i \leq d-1.$$

Here, θ_1 and $\theta_2, \ldots, \theta_{d-1}$ are called the *azimuthal angle* and the *polar angles*, respectively. For integers $|\ell_1| \leq \ell_2 \leq \ldots \leq \ell_{d-1} = n$, we define a *canonical spherical harmonics of degree $\ell_{d-1} = n$ and order $(\ell_1, \ldots, \ell_{d-2})$* as

$$Y_{\ell_1,\ldots,\ell_{d-1}}(\boldsymbol{\theta}) := \frac{1}{\sqrt{2\pi}} e^{i\ell_1\theta_1} \prod_{j=2}^{d-1} {}_j\overline{P}_{\ell_j}^{\ell_{j-1}}(\theta_j), \quad \text{where}$$

$${}_j\overline{P}_{\ell'}^{\ell}(\theta) := \sqrt{\Big(\ell + \frac{j-1}{2}\Big) \frac{(\ell+\ell'+j-2)!}{(\ell'-\ell)!}} \frac{P_{\ell'+\frac{j-2}{2}}^{-(\ell+\frac{j-2}{2})}(\cos\theta)}{\sin^{\frac{j-2}{2}}(\theta)}$$

for $\ell \leq \ell'$ and $j \geq 2$. Here, $P_\lambda^\mu(z)$ denotes the *Legendre functions of the first kind* for $z \in \mathbb{C}$ such that $|1-z| < 2$. Finally, we define a real-valued version, often called the *tesseral harmonics*, as

$$\tilde{Y}_{\ell_1,\ell_2,\ldots,\ell_{d-1}} := \begin{cases} \sqrt{2}(-1)^{\ell_1} \operatorname{Im}(Y_{|\ell_1|,\ell_2,\ldots,\ell_{d-1}}) & \text{if } \ell_1 < 0, \\ Y_{0,\ell_2,\ldots,\ell_{d-1}} & \text{if } \ell_1 = 0, \\ \sqrt{2}(-1)^{\ell_1} \operatorname{Re}(Y_{|\ell_1|,\ell_2,\ldots,\ell_{d-1}}) & \text{if } \ell_1 > 0. \end{cases}$$

Then, $\mathcal{B}_n(d) := \{\tilde{Y}_{\ell_1,\ldots,\ell_{d-1}} \colon |\ell_1| \leq \ell_2 \leq \ldots \leq \ell_{d-1} = n\}$ forms a real orthonormal eigenbasis of $\mathcal{Y}_n(d)$.

A few remarks on related results are in order.

Remark 12 (Multiplicative dot-product kernels over a torus)**.** Since Corollary 10 remains valid for $d = 2$, the eigenfunctions of any dot-product kernels over $\mathbb{S}^1$ are the Fourier basis $\{e^{i\ell\theta}/\sqrt{2\pi}\}_{\ell=0}^{\infty}$ for $\theta \in [0, 2\pi)$. Hence, any dot-product kernel of a multiplicative form such as Gaussian kernels over the d-dimensional torus $\mathbb{T}^d := \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ (with d products) has the product of 1-dimensional Fourier bases as eigenfunctions.

This result may be of particular interest for real-world data naturally lying on the torus such as RNA structure data [6].

Remark 13 (Dot-product kernels over a ball)**.** Smola et al. [21, Section 6] provided a version of the eigensystem of a dot-product kernel over the unit ball $\mathbb{B}^d := \{\mathbf{x} \in \mathbb{R}^d \colon \|\mathbf{x}\|_2 \leq 1\}$ by the separation of variables trick. Here we present the idea with a minor correction. Let $f \colon \mathbb{R} \to \mathbb{R}$ be an analytic function such that $f(t) = \sum_{m=0}^{\infty} f_m t^m$. Plugging in the expansion of monomial u^m ($u \in [-1, 1], m \geq 0$) with respect to the associated Legendre polynomials $(P_n^d(u))_{n \geq d}$ of dimension d, we can write

$$f(\mathbf{x}^T \mathbf{t}) = \sum_{n=d}^{\infty} \kappa_n(\|\mathbf{x}\|\|\mathbf{t}\|) P_n^d\Big(\frac{\mathbf{x}^T \mathbf{t}}{\|\mathbf{x}\|\|\mathbf{t}\|}\Big),$$

where we define $\kappa_n(u) := \sum_{m=0}^{\infty} f_m c_m(d, n) u^m$ and $c_m(d, n) := \frac{2n+1}{2} \frac{(n-d)!}{(n+d)!} \int_{-1}^{1} u^m P_n^d(u) \, \mathrm{d}u$ for $n \geq d$. Now, for each $n \geq d$, let $(\varphi_{nm} \in L^2_{r \mapsto r^{d-1}}([0, 1]))_{m=1}^{\infty}$ and $(\rho_{nm})_{m=1}^{\infty}$ be the eigensytem of the 1D kernel κ_n, i.e.,

$$\int_0^1 \kappa_n(r\tilde{r}) \varphi_{nm}(\tilde{r}) \tilde{r}^{d-1} \, \mathrm{d}\tilde{r} = \rho_{nm} \varphi_{nm}(r). \tag{14}$$

With the addition theorem [14, p. 18] on the expansion of P_n^d with $(Y_{n,j}^d)_{j=1}^{N(d,n)}$, it is then easy to check that $\{\varphi_{nm}(r) Y_{n,j}^d(\boldsymbol{\theta}) \colon m \geq 1, n \geq d, 1 \leq j \leq N(d, n)\}$ forms an orthonormal eigenbasis of $\mathbf{K}$ over $\mathbb{B}^d$ with eigenvalues $\frac{\mathbb{S}^{d-1}}{N(d,n)} \rho_{nm}$ of multiplicity $N(d, n)$. We note, in practice, that the integral equation (14) can be solved by eigendecomposition of a matrix with approximation of κ_n with finite terms.

Remark 14 (Gaussian kernels with Gaussian weighting)**.** For $\mathcal{X} = \mathbb{R}^d$, when $\mathbf{K}_w$ is a Gaussian kernel with a Gaussian weighting function w, the eigensystem of $\mathbf{K}_w$ is characterized by Hermite polynomials [7, 16]. Note, however, that since $w(\mathbf{x})$ is non-uniform being Gaussian, the base kernel $k(\mathbf{x}, \mathbf{t}) = \sqrt{w(\mathbf{x})} k_w(\mathbf{x}, \mathbf{t}) \sqrt{w(\mathbf{t})}$ becomes a Gaussian kernel with an additional attenuation term.

V. Experiments

To illustrate the applicability of the proposed framework, we consider the following simple image segmentation procedure. Suppose that we are given an (color) image $Y \in [0, 1]^{H \times W \times 3}$. For each pixel $Y^{(i)} \in \mathbb{R}^3$, we consider the $P \times P \times 3$ patch centered at $Y^{(i)}$, denoted as $\mathbf{y}^{(i)} \in [0, 1]^{P \times P \times 3} \cong [0, 1]^{3P^2}$, as its representation. We apply a kernel embedding algorithm such as Laplacian eigenmaps or the proposed kernel embedding to the patches $\{\mathbf{y}^{(i)}\}_{i=1}^{HW}$, and apply the k-means algorithm [9, 11] as in spectral clustering [20]; the resulting labels can be viewed as a segmentation of the image.

We present a sample image segmentation result with $P = 2$ in Fig. 1. For kernel PCA and Laplacian eigenmaps, we used isotropic Gaussian kernels with bandwidth selected as median of all pairwise Euclidean distances. For the proposed kernel embedding, we applied the kernel embedding based on spherical harmonics in Section IV, by mapping the data onto a unit hypersphere and used the Gaussian kernel density estimator with the same bandwidth. The number of clusters used in the k-means algorithm was 8. We remark that the spherical embedding has orders-of-magnitude lesser time complexity (∼2s) than the other kernel-based embeddings (∼100s), while providing a comparable result.

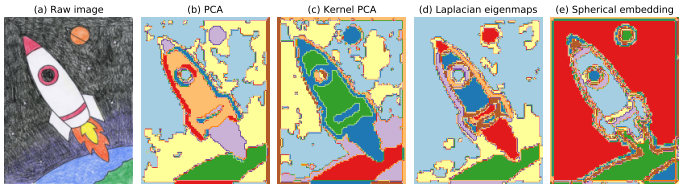

Fig. 1. An illustrative example with image segmentation.

VI. Related work

Spectral clustering [15, 20, 23] has many versions depending on the form of graph Laplacian in the procedure, and Laplacian eigenmaps [1] is equivalent to the spectral embedding used in the version of spectral clustering by Shi and Malik [20]. Schiebinger et al. [18] analyzed the normalized kernel operator $\overline{\mathbf{K}}_p$ to establish the performance of spectral clustering.

The mathematical equivalence between Laplacian eigenmaps and kernel PCA established in Section II is not entirely new. For example, Ng et al. [15] pointed out a link between spectral clustering and kernel PCA. More generally, Ham et al. [8] and Bengio et al. [2] interpreted Laplacian eigenmaps, multidimensional scaling, Isomap, and locally linear embedding as specific instantiations of kernel PCA. Note, however, that they only considered the sample based algorithms not the underlying population optimization problems, while this paper crucially relies on the population formulation.

Dot-product kernels have been studied in the context its regularization property for support vector machines [21] and their feature functions [13]. For a more detailed account on spherical harmonics, we refer an interested reader to [5, 14].

VII. Concluding remarks

In this paper, we proposed a rather unorthodox perspective on kernel-based spectral embedding. We introduced a new criterion for kernel embedding with a new density-regularized kernel, which results in a kernel embedding algorithm without spectral decomposition of a matrix. The advantage comes from the special structure of the kernel k_p in (9), which allows the separation of the density from the eigendecomposition of the kernel operator.

We emphasize that the proposed algorithm is not proposed to replace the existing spectral methods; instead, it should be viewed as an extremely low-cost kernel-based embedding, which may be particularly advantageous when a dataset is large and computational resource is limited. A deeper investigation and more extensive experiments including its variations in Remarks 12, 13, and 14 will be reported elsewhere.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant CCF-1911238.

REFERENCES

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[2] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Comput.*, 16(10):2197–2219, 2004.

[3] Marianna Bolla. *Spectral clustering and biclustering: Learning large graphs and contingency tables*. John Wiley & Sons, 2013.

[4] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[5] Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014.

[6] Benjamin Eltzner, Stephan Huckemann, and Kanti V Mardia. Torus principal component analysis with applications to RNA structure. *Ann. Appl. Statist.*, 12(2):1332–1359, 2018.

[7] Gregory E Fasshauer. Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation*, 4:21–63, 2011.

[8] Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proc. Int. Conf. Mach. Learn.*, page 47, 2004.

[9] John A Hartigan and Manchek A Wong. A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat*, 28(1):100–108, 1979.

[10] Atsushi Higuchi. Symmetric tensor spherical harmonics on the n-sphere and their application to the de sitter group so (n, 1). *Journal of mathematical physics*, 28(7): 1553–1566, 1987.

[11] Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[12] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philos. Trans. R. Soc. A*, 209(441-458):415–446, 1909.

[13] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *Int. Conf. Comput. Learn. Theory*, pages 154–168. Springer, 2006.

[14] Claus Müller. *Analysis of spherical symmetries in Euclidean spaces*, volume 129. Springer Science & Business Media, 2012.

[15] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Proc. Syst.*, 14:849–856, 2001.

[16] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

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

[18] Geoffrey Schiebinger, Martin J Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *Ann. Statist.*, 43(2):819–846, 2015.

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

[20] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[21] Alex J Smola, Zoltan L Ovari, and Robert C Williamson. Regularization with dot-product kernels. *Adv. Neural Inf. Proc. Syst.*, pages 308–314, 2001.

[22] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[23] Yair Weiss. Segmentation using eigenvectors: a unifying view. In *IEEE Int. Conf. Comp. Vis.*, volume 2, pages 975–982. IEEE, 1999.