Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

A unified framework for harmonic analysis of functions on directed graphs and changing data

H.N. Mhaskar [a,b,1]

[a] *Department of Mathematics, California Institute of Technology, Pasadena, CA 91125, United States*
[b] *Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, United States*

ARTICLE INFO

Article history:
Received 1 July 2015
Accepted 19 June 2016
Available online xxxx
Communicated by Charles K. Chui

Keywords:
Kernel construction
Directed graphs
Changing data problems
Extension problems
Wavelet-like representation
Tauberian theorem

ABSTRACT

We present a general framework for studying harmonic analysis of functions in the settings of various emerging problems in the theory of diffusion geometry. The starting point of the now classical diffusion geometry approach is the construction of a kernel whose discretization leads to an undirected graph structure on an unstructured data set. We study the question of constructing such kernels for directed graph structures, and argue that our construction is essentially the only way to do so using discretizations of kernels. We then use our previous theory to develop harmonic analysis based on the singular value decomposition of the resulting non-self-adjoint operators associated with the directed graph. Next, we consider the question of how functions defined on one space evolve to another space in the paradigm of changing data sets recently introduced by Coifman and Hirn. While the approach of Coifman and Hirn requires that the points on one space should be in a known one-to-one correspondence with the points on the other, our approach allows the identification of only a subset of landmark points. We introduce a new definition of distance between points on two spaces, construct localized kernels based on the two spaces and certain interaction parameters, and study the evolution of smoothness of a function on one space to its lifting to the other space via the landmarks. We develop novel mathematical tools that enable us to study these seemingly different problems in a unified manner.

1. Introduction

There are many approaches developed during the last decade or so in order to analyze large, unstructured, possibly high dimensional data. The basic idea is to embed the high dimensional data on a low dimensional sub-manifold of the ambient Euclidean space. The main theme of the research is then to understand the data geometry in terms of the geometric properties of this sub-manifold. Well known techniques in this direction, dimensionality reduction in particular, are Isomaps [63], maximum variance unfolding (MVU)

E-mail address: hrushikesh.mhaskar@cgu.edu.
[1] The research of this author is supported in part by ARO Grant W911NF-15-1-0385.

http://dx.doi.org/10.1016/j.acha.2016.06.007
1063-5203/

(also called semidefinite programming (SDP)) [66], locally linear embedding (LLE) [57], local tangent space alignment method (LTSA) [67], Laplacian eigenmaps (Leigs) [3], Hessian locally linear embedding (HLLE) [19], diffusion maps (Dmaps) [16], and randomized anisotropic transform [11]. A recent survey of these methods is given by Chui and Wang in [12]. An excellent introduction to the subject of diffusion geometry can be found in the special issue [8] of Applied and Computational Harmonic Analysis, 2006. The application areas are too numerous to mention exhaustively. They include, for example, document analysis [17], face recognition [38,46,11], hyperspectral imaging [10], semi-supervised learning [2,4], image processing [22,6], cataloguing of galaxies [23], and social networking [64].

The starting point in diffusion geometry is the point cloud. A point cloud is a set $\mathcal{P}$ of points $\{x_i\}$ in a Euclidean space, together with a similarity relation W, viewed as the matrix of edge weights in an undirected graph. In the absence of any known structure on the set of points, the similarity relation is constructed by the so-called diffusion matrix. For example, a standard construction for W is given by

$$W_{i,j} = \exp(-\|x_i - x_j\|^2/\epsilon),$$

where $\{x_i\}$ is the set of points, $\|\cdot\|$ is the Euclidean norm, and ϵ is a judiciously chosen variance parameter. The first few eigenvalues and eigenvectors of this matrix (or some related matrix, such as $\mathsf{diag}(\text{row sums of } W) - W$) provide the desired low dimensional embedding. Usually, these vectors codify certain identifiable features of the data set. For example, in [11], the point cloud consists of thumbnail images of the faces of the same person in different orientations. The components of the first non-trivial eigenvector order the collection according to the angle of rotation. Many such examples can be found in the literature; indeed, the well known concept of kernel PCA is based on this fact. It is proved in [3,5,45,61] that as the data becomes dense, the so-called graph Laplacian based on a judiciously chosen (weighted) adjacency matrix, such as W, converges to the Laplace–Beltrami operator on the **unknown** manifold from which the data is sampled, and likewise for the corresponding eigenvalues/eigenfunctions. A deeper insight into the intimate connection between the eigenfunctions and the manifold is explained in [40,41], where it is shown that some of the eigenfunctions define a local coordinate system on the unknown manifold so that the Euclidean distance between the points represented with these coordinates is proportional to the geodesic distance between the points.

The applications of this theory to semi-supervised learning can be formulated as problems of function extension, going beyond the important and difficult question of understanding the data geometry. For example, in semi-supervised learning for classification, the class labels are known only on a small *training subset* $\mathcal{C} \subset \mathcal{P}$, and the objective is to find the class labels for all points in $\mathcal{P}$. The class label for a point x can be viewed as the value of a *target function* f at x, and the question is then to extend f from $\mathcal{C}$ to $\mathcal{P}$. Formulated in this manner, one can even think of extending f to the entire manifold $\mathbb{X}$, including those points which are not in the original data set $\mathcal{P}$. Clearly, all the regression problems in learning theory are also problems of function extension/approximation, for example [30]. Indeed, one of the main reasons for the popularity of neural and radial basis function networks in learning theory is their universal approximation property. An important problem in this paradigm is to estimate the fidelity of the approximation scheme at unseen data points. Within the context of diffusion geometry, a rigorous analytic theory to address this question is developed in [48,25,26,51,50], based on an abstract framework formulated in [52]. The constructive algorithm that emerges from this theory is tested in the proof-of-concept experiments on recognition of hand-written digits [48], the Cleveland heart disease data set, the Wisconsin breast cancer data set, and a new experiment at NIH to predict automatically the local classifications of drusen in patients with age related macular degeneration (AMD) based on multi-spectral fundus images of the retina [24].

The goal of the present paper is to develop a general framework in which function extension/approximation problems can be studied in the context of two of the newly emerging paradigms in the theory of

Please cite this article in press as: H.N. Mhaskar, A unified framework for harmonic analysis of functions on directed graphs and changing data, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.007