Appl. Comput. Harmon. Anal. 38 (2015) 420–451

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

A class of Laplacian multiwavelets bases for high-dimensional data

Nir Sharon *, Yoel Shkolnisky

School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

Article history:
Received 29 October 2012
Received in revised form 30 April 2014
Accepted 4 July 2014
Available online 9 July 2014
Communicated by Amit Singer

Keywords:
High dimensional data
Graph Laplacian
Multiwavelets
Multiresolution analysis

ABSTRACT

We introduce a framework for representing functions defined on high-dimensional data. In this framework, we propose to use the eigenvectors of the graph Laplacian to construct a multiresolution analysis on the data. We assume the dataset to have an associated hierarchical tree partition, together with a function that measures the similarity between pairs of points in the dataset. The construction results in a one parameter family of orthonormal bases, which includes both the Haar basis as well as the eigenvectors of the graph Laplacian, as its two extremes. We describe a fast discrete transform for the expansion in any of the bases in this family, and estimate the decay rate of the expansion coefficients. We also bound the error of non-linear approximation of functions in our bases. The properties of our construction are demonstrated using various numerical examples.

1. Introduction

Expanding a function in an orthonormal basis is one of the most basic tools in mathematics. Such an expansion transforms a given function, given either as a mathematical object or as a set of samples stored in a computer, into a set of coefficients. Then, instead of analyzing the original function, one analyzes the resulting expansion coefficients. This procedure arises in many areas such as harmonic analysis, numerical analysis, partial differential equations, and signal processing, to name a few. Examples for tasks that are easily implemented using this approach include denoising, compression, and extrapolation. This approach goes back to classical harmonic analysis, where the properties of a function are analyzed by inspecting its expansion coefficients into a Fourier series. Other more modern bases include wavelet bases, splines, and orthogonal polynomials.

Although this approach turned out to be very powerful, most tools and theory are only available in one dimension. Extensions to higher dimensions are commonly derived by tensor products of one-dimensional elements, which usually do not exploit any special structure of the underlying domain. Moreover, even the

* Corresponding author.
E-mail address: nirsharo@post.tau.ac.il (N. Sharon).

http://dx.doi.org/10.1016/j.acha.2014.07.002

one-dimensional theory cannot be applied to arbitrary sets on the line, and is typically restricted to an interval or the entire line. For example, it cannot be applied when our domain is simply a finite scattered set of points on the real line.

In recent years, the need to analyze data in high dimensions has grown rapidly. Moreover, most often the data is simply some finite set of high-dimensional vectors, without any of the rich Euclidean structure inherent into the classical tools. Analyzing functions defined on such general datasets arises naturally, for example, in meteorology [17], gene research [32], medical imaging [31], etc.

This practical need for analyzing large high-dimensional datasets has motivated extensive research in past years. These studies resulted in many methods for data representation and dimensionality reduction such as [6,39]. In other cases, these studies gave rise to generalizations of the Euclidean space construction of wavelets and wavelet packets. Such a wavelets analogue on manifolds and graphs, based on the diffusion operator, is suggested in [10,15,42]. These diffusion based constructions exploit the decay of the spectrum of the diffusion operator and its powers. A different approach for graphs was introduced in [27], where the wavelet operator is defined using the localization and scaling of the graph Laplacian. Another construction for non-Euclidean setting is presented in [3] for general spaces of homogeneous type.

The problem of basis construction is also known as dictionary learning in the machine learning community, where one often uses a given hierarchy tree on the data to construct a multiresolution analysis [8,30]. Thus, constructing a basis is related to finding optimal trees [9], and is also known as data-adaptive signal representation [22].

Two other important constructions available today for high-dimensional data in almost arbitrary domains, modelled as graphs, are the graph Laplacian (e.g., [16,38]) and the Haar basis on graphs [24]. The graph Laplacian approach uses as basis functions the eigenvectors of the graph Laplacian over the data. If the data are uniformly sampled from some Euclidean domain, then, under certain conditions, as the number of data points goes to infinity, these eigenvectors converge to the eigenfunctions of the Laplacian over the underlying domain [41]. For example, if the data are sampled from the unit circle, then the eigenvectors would converge to sine and cosine functions. Thus, this approach results in an analogue of Fourier basis for general domains [37]. This approach is described in details in Section 2.1.

The second basis available for general data is the Haar basis on graphs. The continuous Haar basis is considered as the simplest wavelet basis, and is defined using dilation and translation of a piecewise constant function. To extend the Haar basis to general datasets, one is required to replace the standard dyadic partition of the interval with some hierarchical partition of the data (see Subsection 2.3). Once such a partition is given, the basis is constructed by taking the characteristic functions over the elements of the partition, and applying an orthogonalization procedure to these functions. Such a construction is described in [24].

On one hand we have the eigenvectors of the graph Laplacian whose support cannot be controlled and can be shown to be "smooth" under an appropriate definition [4]. On the other hand we have the Haar basis whose basis elements are localized piecewise constant functions. An open question is whether there is any basis "in between".

In this paper, we suggest a family of bases, named Laplacian multiwavelets, which can be constructed for any set of N data points. We require the dataset to have an associated hierarchical tree partition, together with a function that measures the similarity between pairs of points in the dataset. These requirements will be made precise in Section 3. The family of bases is parameterized by one integer parameter $1 \leq k \leq N$, where $k = 1$ corresponds to the Haar basis defined in [24], and $k = N$ corresponds to the eigenvectors of the graph Laplacian (Fourier basis). Intermediate values of k correspond to various degrees of "generalized vanishing moments" as explained in Section 5 and Subsection 6.1. In particular, we show that generalized vanishing moments are related to the decay rate of the expansion coefficients in our basis. We also show through numerical examples that by tuning the value of k, our basis is capable of efficiently representing both slowly varying functions, just like the graph Laplacian basis [37], as well as highly oscillatory functions, just