Chapter 20
Construction of Tight Frames on Graphs and Application to Denoising

Franziska Göbel, Gilles Blanchard, and Ulrike von Luxburg

Abstract Given a neighborhood graph representation of a finite set of points $x_i \in \mathbb{R}^d, i = 1, \ldots, n$, we construct a frame (redundant dictionary) for the space of real-valued functions defined on the graph. This frame is adapted to the underlying geometrical structure of the x_i, has finitely many elements, and these elements are localized in frequency as well as in space. This construction follows the ideas of Hammond et al. (Appl Comput Harmon Anal 30:129–150, 2011), with the key point that we construct a tight (or Parseval) frame. This means we have a very simple, explicit reconstruction formula for every function f defined on the graph from the coefficients given by its scalar product with the frame elements. We use this representation in the setting of denoising where we are given noisy observations of a function f defined on the graph. By applying a thresholding method to the coefficients in the reconstruction formula, we define an estimate of f whose risk satisfies a tight oracle inequality.

Keywords Neighborhood graph · Tight frame · Dictionary learning · Denoising · Thresholding · Oracle inequality

20.1 Introduction

20.1.1 Motivation

When dealing with high-dimensional data, a general principle is that the curse of dimensionality can be efficiently fought if one assumes the data points to lie on a structure of smaller intrinsic dimensionality, typically a manifold. Some well-known

F. Göbel · G. Blanchard (✉)
Institute of Mathematics, University of Potsdam, Potsdam, Germany
e-mail: goebel@uni-potsdam.de; gilles.blanchard@math.uni-potsdam.de

U. von Luxburg
Department of Computer Science, University of Tübingen, Tübingen, Germany
e-mail: luxburg@informatik.uni-tuebingen.de

W. K. Härdle et al. (eds.), *Handbook of Big Data Analytics*, Springer Handbooks of Computational Statistics, https://doi.org/10.1007/978-3-319-18284-1_20

methods to discover such a lower dimensional structure include Isomap (Tenenbaum et al. 2000), LLE (Roweis and Saul 2000), and Laplacian Eigenmaps (Belkin and Niyogi 2003).

In this work, our main interest is not in visualizing or representing by an explicit mapping the underlying structure of the observed data points; rather, we want to represent or estimate efficiently a real-valued function on these points. More specifically, we focus on the following *denoising* problem: assuming we observe a noisy version of the function f, $y_i = f(x_i) + \varepsilon_i$ at points $(x_1, \ldots, x_n)$, we would like to recover the values of f at these points. An important step for solving this problem is to find a dictionary of functions to represent the signal f, which is adapted to the structure of the data. Ideally, we would like this dictionary to exhibit the features of a wavelet basis. In traditional signal processing on a flat space, with data points on a regular grid, orthogonal wavelet bases offer a very powerful tool to sparsely represent signals with inhomogeneous regularity (such as a signal that is very smooth everywhere except at a few singular points where it is discontinuous). Such bases are in particular well suited to the denoising task. Can this be generalized to irregularly scattered data on a manifold?

We present such a method to construct a so-called *Parseval frame* of functions exhibiting wavelet-like properties while adapting to the intrinsic geometry of the data. Furthermore, we use this dictionary for the denoising task using a simple coefficient thresholding method.

This work is organized as follows. In the coming section, we discuss the relationship to previous work on which the present chapter is built, as well as pointing out our new contributions. In Sect. 20.2, we recall important notions of frame theory as well as of neighborhood graphs needed for our construction. The construction of the frame and its properties is presented in Sect. 20.3. In Sect. 20.4, we develop a coefficient thresholding strategy for the denoising problem. In Sect. 20.5, we present numerical results and method comparison on testbed data.

20.1.2 Relation to Previous Work

Regression methods that adapt to an underlying lower dimension of the data have been considered by Bickel and Li (2007), Kpotufe and Dasgupta (2012) and Kpotufe (2011) using local polynomial estimates, random projection trees, and nearest-neighbors, respectively. However, these methods are not constructed to adapt to an inhomogeneous regularity of the target function: in these three cases, the smoothing scale (determined by the smoothing kernel bandwidth, the tree partition's average data diameter, or the number of neighbors, respectively) is fixed globally. In the experimental Sect. 20.5, for data lying on a smooth manifold but a target function exhibiting a sharp discontinuity, we demonstrate the advantage of our method over kernel smoothing.

Based on the motivations similar to ours, a method for constructing a wavelet-like basis on scattered data was proposed by Gavish et al. (2010). It is based on a hierarchical tree partition of the data, on which a Haar-like basis of 0–1 functions is constructed. However, the performance of that method is then adapted to the geometry of the *tree*, in the sense that the distance of two points is measured through tree path distance. This can strongly distort the original distance: two close points in original distance can find themselves in very separated subtrees.

The construction proposed here, based on a transform of the spectral decomposition of the graph Laplacian, follows closely the ideas of Hammond et al. (2011). Two important contributions brought forth in the present work are that we construct a Parseval (or tight) frame, rather than a general frame; and we consider an explicit thresholding method for the denoising problem. The former point is crucial to obtain sharp bounds for the thresholding method, and also eliminates the computational problem of signal reconstruction from the frame coefficients, since Parseval frames enjoy a reconstruction formula similar to that of an orthonormal basis. The choice of multiscale bandpass filter functions leading to the tight frame is inspired by the recent work of Coulhon et al. (2012), where the spectral decomposition principle is also studied, albeit in the setting of a quite general metric space.

20.2 Notation and Basics

20.2.1 Setting

We consider a sample of n points $x_i \in \mathbb{R}^d$. These points are assumed to belong to an unknown low-dimensional submanifold $\mathscr{M} \subset \mathbb{R}^d$. We denote the design by $\mathfrak{D} = \{x_1, \ldots, x_n\} \subset \mathscr{M}$. Furthermore, we observe on these points the (noisy) value of a function $f : \mathfrak{D} \to \mathbb{R}$. Since $\mathfrak{D}$ is finite, we can represent the function f as vector $f = (f(x_1), \ldots, f(x_n))^t \in \mathbb{R}^n$. The space of all (square-integrable) functions f defined on $\mathfrak{D}$ is denoted $L^2(\mathfrak{D})$ and endowed with the usual Euclidean inner product.

We denote by $y_i = f(x_i) + \epsilon_i$ the noisy observation of f at x_i, where ϵ_i are independent identically distributed centered random variables. The problem we consider in this work is that of denoising, that is, try to recover the underlying value of the function f at the points x_i.

While the existence of a low-dimensional supporting manifold $\mathscr{M}$ for the design points motivates the construction of the proposed method, we underline (again) that $\mathscr{M}$ is not known to the user and the method only uses the knowledge of the design points. In such a setting, a key idea to recover implicitly some information on the geometry of $\mathscr{M}$ is to construct a neighborhood graph based on the design points (see Sect. 20.2.3 for details).

20.2.2 Frames

For the construction in Sect. 20.3, we rely on the notion of a *vector frame*, for which we recall here some important properties (see, e.g., Casazza et al. 2013; Han 2007; Christensen 2008). A frame is an overcomplete dictionary with particular properties allowing it to act almost as basis.

Definition 1 Let $\mathscr{H}$ be a Hilbert space. Then a countable set $\{z_i\}_{i\in I} \subset \mathscr{H}$ is a frame with frame bounds A and B for $\mathscr{H}$ if there exists constants $0 < A \leq B < \infty$ such that

$$\forall z \in \mathscr{H} : \quad A\,\|z\|^2 \leq \sum_{i\in I} |\langle z, z_i\rangle|^2 \leq B\,\|z\|^2\,. \tag{20.1}$$

A frame is called tight if $A = B$, in particular the frame is called Parseval if $A = B = 1$.

In the remainder of this work we consider the case of a Euclidean space $\mathscr{H} = \mathbb{R}^n$, and assume that $\{z_i\}_{i\in I}$ is a frame with a finite number of elements. Two important operators associated to the frame are the *analysis* operator

$$T : \mathbb{R}^n \to \mathbb{R}^I,\ Tz := (\langle z, z_i\rangle)_{i\in I} \tag{20.2}$$

(sequence of frame coefficients), and its adjoint the *synthesis* operator:

$$T^* : \mathbb{R}^I \to \mathbb{R}^n,\ T^*a = T^*(a_i)^t_{i\in I} = \sum_{i\in I} a_i z_i. \tag{20.3}$$

Further, the *frame* operator is defined as $S = T^*T$:

$$S : \mathbb{R}^n \to \mathbb{R}^n,\ Sz = T^*Tz = \sum_{i\in I} \langle z, z_i\rangle\, z_i, \tag{20.4}$$

and finally the *Gramian* operator as $U = TT^*$,

$$U : \mathbb{R}^I \to \mathbb{R}^I,\ Ua = TT^*a = \left\{\left\langle \sum_{i\in I} a_i z_i, z_k \right\rangle\right\}_{k\in I}. \tag{20.5}$$

In matrix form, the columns of T^* are the vectors $z_i, i \in I$, T is its transpose and $U_{ij} = \langle z_i, z_j\rangle$.

The definition of a frame implies that S is invertible, and it is possible to reconstruct any z from its frame coefficients by $z = \sum_{i\in I} \langle z, z_i\rangle\, z_i^* = \sum_{i\in I} \langle z, z_i^*\rangle\, z_i$, where $z_i^* := S^{-1} z_i, i \in I$ is called the *canonical dual* frame of $(z_i)_{i\in I}$.

We recall some properties of finite Parseval frames over Euclidean spaces (see, e.g., Han 2007, chapter 3).

Theorem 1 (Properties of Parseval frames) *Let $\mathscr{H}$ be a Hilbert space with* $\dim \mathscr{H} = n < \infty$. *The following statements are equivalent:*

1. $\{z_i\}_{1 \leq i \leq k} \subset \mathscr{H}$ *is a Parseval frame.*
2. $\forall y \in \mathscr{H} : \ \ y = \sum_{i=1}^{k} \langle y, z_i \rangle z_i$.
3. *The frame operator S is the identity on $\mathbb{R}^n$.*
4. *The Gramian operator U is an orthogonal projector of rank n in $\mathbb{R}^k$.*

Furthermore if $\{z_i\}_{1..k} \subset \mathscr{H}$ is a Parseval frame, then

- $\|z_i\| \leq 1$ *for* $i \in \{1, \ldots, k\}$;
- $\dim \mathscr{H} = n = \sum_{i=1}^{k} \|z_i\|^2$;
- *the canonical dual frame is the frame itself.*

For the present work, the two most important points of this theory are the following: first, the reconstruction formula (point two above), where we see that a Parseval frame acts similarly to an orthonormal basis; secondly, if we construct a vector $v = T^* a = \sum_i a_i z_i$ from an arbitrary vector of coefficients (a_i), then

$$\left\| \sum_i a_i z_i \right\|^2 = \langle T^* a, T^* a \rangle = \langle a, Ua \rangle = \|Ua\|^2 \leq \|a\|^2 , \tag{20.6}$$

which follows from property 4 above.

20.2.3 Neighborhood Graphs

In order to exploit the structure and geometry of the unknown submanifold $\mathscr{M}$ on which the sample $\mathfrak{D}$ is supposed to lie, a powerful idea is to use a graph-based representation of the data $\mathfrak{D}$ through a *neighborhood graph*. The points in $\mathfrak{D}$ correspond to the vertices of the graph, and two vertices of the graph are joined by an edge when the two corresponding points are neighbors (in some appropriate sense) in $\mathbb{R}^n$. The underlying idea is that the local geometry of $\mathbb{R}^n$ is reflected in the local connectivity of the graph, while the long-range geometry of the graph reflects the geometrical properties of the manifold $\mathscr{M}$, rather than those of $\mathbb{R}^n$.

Formally, a finite graph $G = (V, E)$ is given by a finite set of vertices V and a set of edges $E \subset V \times V$. The $|V| \times |V|$ adjacency matrix A of the graph is defined by $A_{i,j} = 1$ if $(v_i, v_j) \in E$ and $A_{i,j} = 0$ otherwise. An undirected graph is such that its adjacency matrix is symmetric.

The graph is called weighted if every edge $e \in E$ has a positive weight $w(e) \in \mathbb{R}_+$. In this case the notion of adjacency matrix is extended to $A_{i,j} = w((v_i, v_j))$ if $(v_i, v_j) \in E$ and $A_{i,j} = 0$ otherwise. The degree of a vertex v_i in a (possibly weighted) graph is defined as $d_i = d(i) = \sum_{j=1}^{|V|} A_{i,j}$.

As announced, we focus on geometric graphs, which (can) approximate the structure of the unknown $\mathscr{M}$. Each point x_i is represented by a vertex, say v_i. An edge between two vertices represents a small distance, or a high similarity, of the two associated points. The weight of an edge can quantify the similarity more finely.

We use the Euclidean distance $d(x_i, x_j) = \left\| x_i - x_j \right\|$. We recall three usual ways to construct the edges of a neighborhood graph:

- (undirected) k-nearest-neighbor graph: an undirected edge connects the two vertices v_i and v_j iff x_i belongs to the k nearest neighbors of x_j, or x_j belongs to the k nearest neighbors of x_i ("the k-NN-graph").
- ϵ-graph: an undirected edge connects two vertices v_i and v_j iff $d(x_i, x_j) \leq \epsilon$.
- complete weighted neighborhood graph: for each pair of vertices there exists an undirected edge with a weight depending on the distance/similarity of the two vertices.

A k-NN graph or an ϵ-graph can be made weighted by additionally assigning weights to the edges depending on $d(x_i, x_j)$, for instance by choosing Gaussian weights $w(\{i, j\}) = \exp(-d^2(x_i, x_j)/2\lambda^2)$.

20.2.4 Spectral Graph Theory

If one considers real-valued functions $f : \mathscr{M} \to \mathbb{R}$ defined on a submanifold $\mathscr{M} \subset \mathbb{R}^d$, it is known that under some regularity assumptions on the submanifold $\mathscr{M}$, the eigenfunctions of the Laplace-Beltrami-operator give a basis of the space of squared-integrable functions on $\mathscr{M}$. Since $\mathscr{M}$ is unknown in our setting, the principle of the *Laplacian Eigenmaps* method (Belkin and Niyogi 2003) is to use a discrete analogon, namely the graph Laplace operator L on a neighborhood graph.

Given a finite weighted undirected graph with adjacency matrix A ($n \times n$) and vertex degrees $(d_i)_i$, as introduced in the previous section, we will either use the unnormalized graph Laplace operator L^u or the normalized (symmetric) graph Laplace operator L^{norm} defined by

$$\begin{aligned} L^u &= D - A \qquad (20.7) \\ L^{\text{norm}} &= \mathbf{I}_n - D^{-1/2} A D^{-1/2}, \end{aligned}$$

where $D = \operatorname{diag}(d_1, \ldots, d_n)$ is a diagonal matrix with entries d_i on the diagonal. By construction L^u and L^{norm} are symmetric matrices. The positive semidefiniteness follows from

$$f^t L^u f = 0.5 \sum_{(i,j)} A_{i,j} (f_i - f_j)^2 \text{ and} f^t L^{\text{norm}} f = 0.5 \sum_{(i,j)} A_{i,j} \Big(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \Big)^2,$$

respectively. The spectral theorem for matrices indicates that the normalized eigenvectors Φ_i of the graph Laplace operator L (L^u resp. L^{norm}) form an orthonormal basis of $\mathbb{R}^n$ and all eigenvalues are nonnegative. Furthermore the number of components of the graph is given by the number of eigenvalues equal to 0.

20.3 Construction and Properties

20.3.1 Construction of a Tight Graph Frame

As discussed earlier, the principle of Laplacian Eigenmaps is to use the basis $(\Phi_i)_{1\leq i\leq n}$ to represent and process the data. An important advantage of this basis as compared with the natural basis of $\mathbb{R}^d$ is that it will be *adapted* to the geometry of the underlying submanifold $\mathscr{M}$ supporting the data distribution. For instance, in the denoising problem, a reasonable estimator of f could be a truncated expansion of the noisy vector of observations Y in the basis $(\Phi_i)_{1\leq i\leq n}$.

On the other hand, a disadvantage of this basis is that it is not *spatially localized*. To get an intuitive view, consider the simple case of the interval [0, 1] with uniformly distributed data. In the population view, the eigenbasis of the Laplacian is the Fourier basis. While a truncated expansion in this basis is well-adapted to represent functions that are uniformly regular, it is not well-suited for functions exhibiting locally varying regularity (as an extreme example, a signal that is very smooth everywhere except at a few singular points where it is discontinuous). By contrast, wavelet bases, because they are localized both in space and frequency, allow for an efficient (i.e., sparse) representation of signals with locally varying regularity.

If we now think of data supported on a one-dimensional submanifold (curve) of $\mathbb{R}^d$, we can expect that the Laplacian eigenmaps method will discover a warped Fourier basis following the curve; and, for a more general submanifold $\mathscr{M}$, "harmonics" on $\mathscr{M}$.

In order to go from this basis to a spatially localized dictionary, following ideas of Coulhon et al. (2012) and Hammond et al. (2011), we use the principle of the *Littlewood-Paley* decomposition.

Let G be an undirected geometric neighborhood graph with adjacency matrix A constructed from $\mathfrak{D}$, and L be an associated symmetric graph Laplace operator with increasing eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and normalized eigenvectors $\Phi_i \in \mathbb{R}^n, i = 1 \ldots n$.

We first define a set of vectors using a decomposition of unity and a splitting operation and we will show that this vector set is a Parseval frame.

Definition 2 Let $\{\zeta_k\}_{k\in\mathbb{N}}$ be a sequence of functions $\zeta_k : \mathbb{R}_+ \to [0, 1]$ satisfying

(DoU) $\sum_{j\geq 0} \zeta_j(x) = 1$ for all $x \geq 0$;
(FD) $\#\{\zeta_k : \zeta_k(\lambda_i) \neq 0\} < \infty$ for $i = 1, \ldots, n$.

Then we define the set of column vectors $\{\Psi_{kl} \in \mathbb{R}^n, 0 \leq k \leq Q, 1 \leq j \leq n\}$ by

$$\Psi_{kl} = \sum_{i=1}^{n} \sqrt{\zeta_k(\lambda_i)}\Phi_i(x_l)\Phi_i. \tag{20.8}$$

with $Q := \max\{k : \exists i \in \{1, \ldots, n\} \text{ with } \zeta_k(\lambda_i) > 0\}$.

Theorem 2 *$\{\Psi_{kl}\}_{k,l}$ is a Parseval frame for $\mathscr{H} = \mathbb{R}^n$, that is for all $x \in \mathbb{R}^n$:*

$$\sum_{k,l} |\langle x, \Psi_{kl}\rangle|^2 = \|x\|^2 \ . \tag{20.9}$$

Proof If we can show that $\sum_{(k,l)} \Psi_{kl}\Psi_{kl}^t = \mathbf{I}_n$, we get immediately

$$y = \mathbf{I}_n y = \left(\sum_{(k,l)} \Psi_{kl}\Psi_{kl}^t\right) y = \sum_{(k,l)} \langle y, \Psi_{kl}\rangle\, \Psi_{kl}, \tag{20.10}$$

for $y \in \mathbb{R}^n$. According to Theorem 1 this equation is equivalent to the condition (20.1) with $A = B = 1$. So we are done. It remains to show $\sum_{k,l} \Psi_{kl}\Psi_{kl}^t = \mathbf{I}_n$. We have (since we sum over a finite number of elements)

$$\begin{aligned}\sum_{(k,l)} \Psi_{kl}\Psi_{kl}^t &= \sum_{k,l,i,j} \sqrt{\zeta_k(\lambda_i)}\sqrt{\zeta_k(\lambda_j)}\Phi_i(x_l)\Phi_j(x_l)\Phi_i\Phi_j^t \\ &= \sum_{i=1}^{n}\sum_{k=0}^{Q} \zeta_k(\lambda_i)\, \Phi_i\Phi_i^t \\ &= \sum_{i=1}^{n} \Phi_i\Phi_i^t = \mathbf{I}_n. \end{aligned} \tag{20.11}$$

For the second equality, we have used that $\sum_l \Phi_i(x_l)\Phi_j(x_l) = \langle \Phi_i, \Phi_j\rangle = \mathbf{1}\{\, i = j\}$, since $\{\Phi_i\}_i$ is an orthonormal basis (onb). For the third equality, we used (DoU), and for the last again the onb property. □

We now choose a special sequence of functions satisfying the decomposition of unity (DoU) condition while also ensuring (a) a spectral localization property for the frame elements and (b) a multiscale decomposition interpretation of the resulting decomposition. This construction follows Coulhon et al. (2012), and is known in the context of functional analysis as a smooth Littlewood-Paley decomposition.

Definition 3 (Multiscale Bandpass Filter) Let $g \in C^\infty(\mathbb{R}_+)$, $\operatorname{supp} g \subset [0, 1]$, $0 \le g \le 1$, $g(u) = 1$ for $u \in [0, 1/b]$ (for some constant $b > 1$). For $k \in \mathbb{N} = \{0, 1, \ldots\}$ the functions $\zeta_k : \mathbb{R}_+ \to [0, 1]$ are defined by

$$\zeta_k(x) := \begin{cases} g(x) & \text{if}\, k = 0 \\ g(b^{-k}x) - g(b^{-k+1}x) & \text{if}\, k > 0 \end{cases} \tag{20.12}$$

The sequence $\{\zeta_k\}_{k\ge 0}$ is called multiscale bandpass filter.

This definition leads to the following properties: $\zeta_k(x) = \zeta_1(b^{-k}x)$ for $k \ge 1$ (multiscale decomposition), $\zeta_k \in C^\infty(\mathbb{R}_+)$, $0 \le \zeta_k \le 1$, $\operatorname{supp}\zeta_0 \subset [0, 1]$,

$\operatorname{supp}\zeta_k \subset [b^{k-2}, b^k]$ for $k \geq 1$ (spectral localization property). Moreover, one can check readily

$$\sum_{j\geq 0} \zeta_j(x) = 1, \tag{20.13}$$

i.e., the (DoU) condition holds. In practice, we use a dyadic bandpass filter, that is, $b = 2$. The functions $\zeta_0, \ldots, \zeta_5$ with $b = 2$ are displayed in Fig. 20.1b. By construction, the parameter k in Ψ_{kl} is naturally a spectral scale parameter, while l is a spatial localization parameter: the frame element Ψ_{kl} is localized around the point x_l, as we discuss next.

20.3.2 Spatial Localization

By construction, the elements of the frame are band-limited, i.e. localized in the spectral scale, in the sense that for a fixed k, the frame elements Ψ_{kl} ($l = 1, \ldots, n$) are linear combinations of the eigenvectors of the graph Laplacian ("graph harmonics") corresponding to eigenvalues in the range $[b^{k-2}, b^k]$ only.

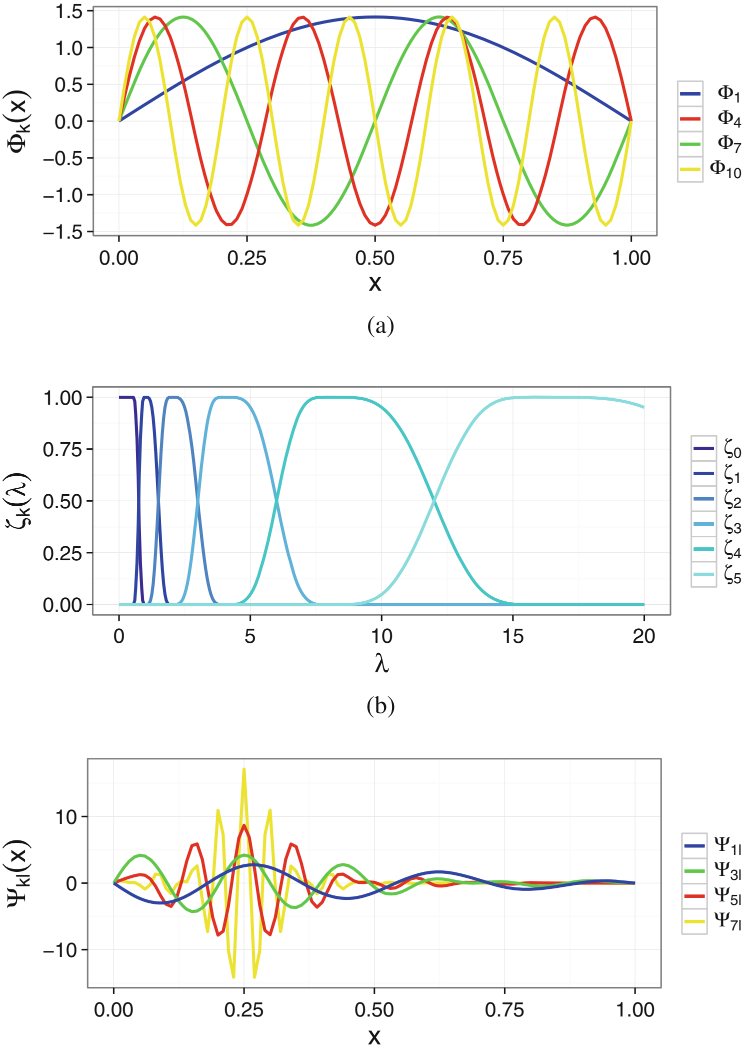
From our initial motivations, it is desirable that in contrast with the eigenfunctions of the Laplace operator, the frame elements Ψ_{kl} are spatially localized functions. In the classical Littlewood-Paley construction for the usual Laplacian on the interval $[0,1]$, this is a well-known fact: the use of linear combination of trigonometric functions $\Psi_{kl}(y) := \sin(kl)\sin(ky)$ via smooth multiscale bandpass filters weights as described in Definition 3 gives rise to strongly localized functions (as illustrated in Fig. 20.1).

Regarding the corresponding discrete construction based on the graph Laplacian, this localization property is certainly observed in practice (as illustrated in Figs. 20.2 and 20.3, see Sect. 20.5 for the setup of the numerical experiments).

Concerning the theoretical perspective, we first review briefly the existing results of Hammond et al. (2011), denote d the shortest path distance in the graph. Theorem 5.5 of Hammond et al. (2011) gives the following localization result for graph frames:

$$\frac{\Psi_{kl}(x)}{\|\Psi_{kl}\|_2} \leq Cb^{-k}, \tag{20.14}$$

for all x with $d(x, x_l) \geq K$, under the assumption that the scaling function ζ_1 is K-times differentiable with vanishing first $(K-1)$ derivatives in 0, non-vanishing K-th derivative, and the scale parameter k is big enough. This says that Ψ_{kl} is "localized" around the point x_l. Unfortunately, this result is not informative in our framework for two reasons: first, we chose a function ζ_1 (see (20.12)) vanishing in a neighborhood of zero, so that all derivatives vanish in the origin, contradicting one of the above assumptions. Secondly, and independently of this first issue, the condition "k is big



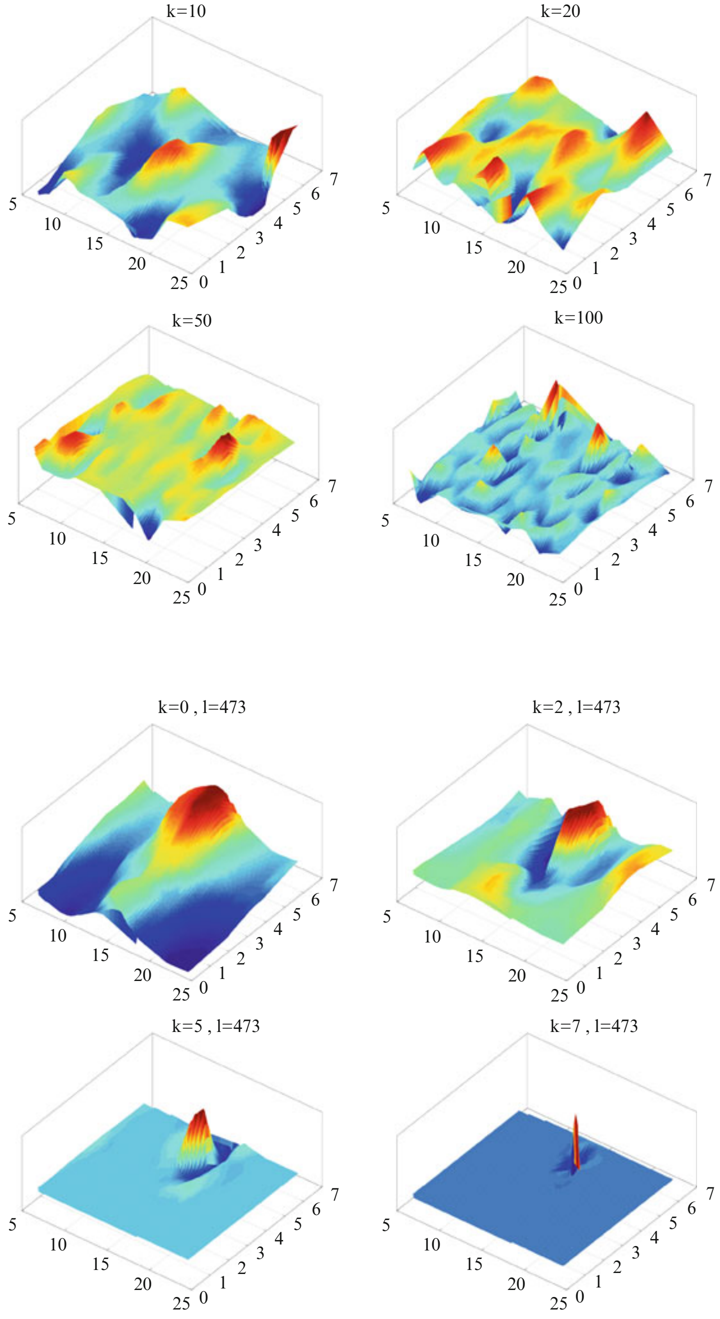
(a)

(b)

(c)

Fig. 20.1 Littlewood-Paley on $L^2(0,1)$: (**a**) eigenfunctions; (**b**) multiscale bandpass filter; (**c**) frame elements

enough," and the factor C depend on the size n of the graph and of the largest eigenvalue of the Laplacian. As a consequence it is unclear if this bound covers any interesting part of the spectrum (for k too large, the spectral support $[b^{k-2}, b^k]$ does not contain any eigenvalues, so that Ψ_{kl} is trivial). Finally, for fixed k the bound also does not give information on the behavior of $\Psi_{kl}(x)$ when the path distance of x to x_l becomes very large.


k=10

k=20

k=50

k=100

k=0 , l=473

k=2 , l=473

k=5 , l=473

k=7 , l=473

Fig. 20.2 Swiss roll data: top: eigenvectors Φ_j for $j = 10, 30, 50, 100$; bottom: frame elements Ψ_{kl} for l fixed and $k = 0, 2, 5, 7$ (construction from actual swiss roll data, then "unrolled" for clearer graphical representation)

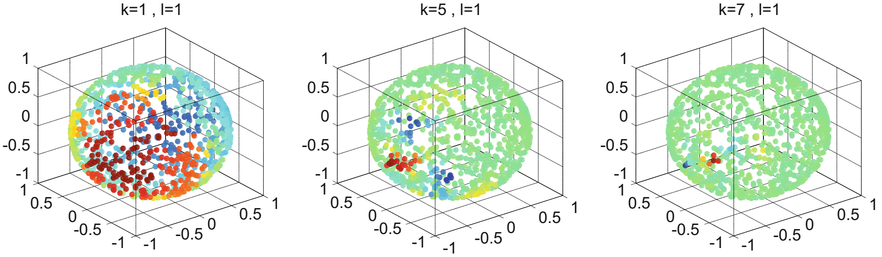


Fig. 20.3 Sphere data: frame elements Ψ_{kl} for l fixed and $k = 1, 5, 7$ (color encodes the value of the function)

On the other hand, the form of the scaling function ζ_1 used in the present work is based on Coulhon et al. (2012) where a theory of multiscale frame analysis is developed on very general metric spaces under certain geometrical assumptions. In a nutshell, it is proved there that using this construction, the obtained frame functions $\Psi_{kl}(x)$ are upper bounded by $O((d(x, x_l)/b^k)^{-\nu})$ for ν arbitrary large. We observe that this type of localization estimate is sharper than (20.14) for fixed x and growing k, as well as for fixed scale k and varying x. We conjecture that these theoretical results apply meaningfully in the discrete setting considered here, under the assumption that $x_1, \ldots, x_n$ are iid from a sufficiently regular distribution $\mathbf{P}_0$ on a regular manifold $\mathscr{M}$, but it is out of the intended scope of the present chapter to establish this formally. In particular "meaningfully" means that the constants involved in the bounds should be independent of the graph size (otherwise the bounds could potentially be devoid of interest for any particular graph, as pointed out above), a question that we are currently investigating.

20.4 Denoising

We consider the regression model for fixed design points $\mathfrak{D} = \{x_i, i = 1 \ldots n\}$ and observations $y_i = f(x_i) + \epsilon_i$ (ϵ_i are independent and identically distributed random variables with $\mathbf{E}(\epsilon_i) = 0$ and $\mathbf{Var}(\varepsilon)_i = \sigma^2$). The aim of denoising is to recover the function $f : \mathfrak{D} \to \mathbb{R}$ at the design points themselves. We will use the proposed Parseval frame in order to define an estimate $\widehat{f}$ of the function f. In what follows, since the $\mathfrak{D}$ is fixed, we identify f with the vector $(f(x_1), \ldots, f(x_n))$ and denote $y = (y_1, \ldots, y_n)$.

Given the frame $\mathscr{F}$ associated to the data points $\mathfrak{D}$ with a multiscale bandpass filter as from Definitions 2 and 3, we denote the frame coefficients $a_{kl} = \langle \Psi_{kl}, f \rangle$ for f and $b_{kl} = \langle \Psi_{kl}, y \rangle$ for y. Due to the linearity of the inner product we get $a_{kl} = b_{kl} - \langle \Psi_{kl}, \epsilon \rangle$. We estimate the unknown coefficients a_{kl} by adjusting the known coefficients b_{kl} by soft-thresholding:

$$S_s(z, c) = \operatorname{sgn}(z)\,(|z| - c)_+. \tag{20.15}$$

In order to take into account that the frame elements Ψ_{kl} are not normalized, and generally have different norms, we use element-adapted thresholds of the form $c_{kl} = \sigma \, \|\Psi_{kl}\| \, t$ which depend on the variance of $\langle \epsilon, \Psi_{kl} \rangle$ and some global parameter t. Equivalently, this corresponds to first normalizing the observed coefficients b_{kl} by dividing by their variance, then applying a global threshold to the normalized coefficients, and finally inverting the normalization.

The estimator of f is then the plug-in estimator

$$\widehat{f}_{S_s} = \sum_{k,l} S_s\,(b_{kl}, c_{kl})\,\Psi_{kl} = T^* S_s(b, c), \tag{20.16}$$

where $S_s(b, c)$ denotes the vector of thresholded coefficients, and T^* is the synthesis operator of the frame as introduced in Sect. 20.2.2.

To measure the performance of this estimator, we use the risk measure

$$Risk(\widehat{f}, f) = \mathbf{E}_\epsilon \left(\left\| \widehat{f} - f \right\|^2 \right), \tag{20.17}$$

that is, the expected quadratic norm at the sampled points (where $\|f\|^2 = \sum_{i=1}^n f(x_i)^2$ is the Euclidean vector norm of f on the observation points), for the performance analysis of an estimator $\widehat{f} \in \mathbb{R}^n$.

For bounding the risk of the thresholding estimator $\widehat{f}_{S_s}$, rather than assuming some specific regularity properties on the function f, it is useful to compare the performance of $\widehat{f}_{S_s}$ to that of a group of reference estimators. This is called the *oracle* approach (Candès 2006; Donoho and Johnstone 1994): can the proposed estimator have a performance (almost) as good as the best estimator (for this specific f) in a reference family (that is to say, as good as if an oracle would have given us advance knowledge of which reference estimator is the best for this function f). We review here briefly some important results.

A suitable class of simple reference estimators consists of "keep or kill" (or diagonal projection) estimators, that keep without changes the observed coefficients $b_{k,l}$ for (k, l) in some subset I, and put to zero the coefficients for indices outside of I:

$$\widehat{f}_I := \sum_{(k,l)\in I} b_{kl} \Psi_{kl} = T^* \widehat{a}^I_{kl}, \tag{20.18}$$

where $\widehat{a}^I_{kl} = b_{kl} \mathbf{1}\{(k, l) \in I\}$. Now using the frame reconstruction formula and (20.6), we obtain

$$\begin{aligned} \mathbf{E}_\epsilon \left(\left\| \widehat{f}_I - f \right\|^2 \right) &= \mathbf{E}_\epsilon \left(\left\| T^*(a - \widehat{a}^I) \right\|^2 \right) \\ &\leq \mathbf{E}_\epsilon \left(\left\| a - \widehat{a}^I \right\|^2 \right) \end{aligned}$$

$$= \sum_{(k,l)} \Big(a_{kl}^2 \mathbf{1}\{(k,l) \notin I\} \\ + \sigma^2 \left\|\psi_{kl}\right\|^2 \mathbf{1}\{(k,l) \in I\}\Big) . \qquad (20.19)$$

Therefore, the optimal (oracle) choice of the index set I^* obtained by minimizing the above upper bound is given by

$$\begin{aligned} (k,l) \in I^* \quad &\Leftrightarrow \quad \langle f, \Psi_{kl}\rangle^2 \geq \sigma^2 \left\|\Psi_{kl}\right\|^2 \quad \text{(keep)} \\ (k,l) \notin I^* \quad &\Leftrightarrow \quad \langle f, \Psi_{kl}\rangle^2 \leq \sigma^2 \left\|\Psi_{kl}\right\|^2 \quad \text{(kill)} \ . \end{aligned} \qquad (20.20)$$

One deduces from this that

$$\inf_I \mathbf{E}_\epsilon \left(\left\| \widehat{f}_I - f \right\|^2 \right) \leq \sum_{(k,l)\in N} \min \left(\langle f, \Psi_{kl}\rangle^2 , \sigma^2 \left\|\Psi_{kl}\right\|^2 \right) =: OB(f) . \qquad (20.21)$$

The relation of soft thresholding estimators to the collection of keep-or-kill estimators on a Parseval frame is captured by the following oracle-type inequality (see Candès 2006, Section 9)[1]:

Theorem 3 *Let $\{\Psi_{kl}\}_{k,l}$ be a Parseval frame and consider the denoising observation model with Gaussian noise. Let $\widehat{f}_{S_s} = \sum_{k,l} S_s \left(\langle y, \Psi_{kl} \rangle , t_{kl} \right) \Psi_{kl}$ be the soft-threshold frame estimator from (20.16). Then with $t_{kl} = \sigma \left\|\Psi_{kl}\right\| \sqrt{2\log(n)}$ the following inequality holds:*

$$\mathbf{E}_\epsilon \left(\left\| \widehat{f}_{S_s} - f \right\|^2 \right) \leq (2\log(n) + 1) \left(\sigma^2 + OB(f) \right) . \qquad (20.22)$$

To interpret this result, observe that if we renormalize the squared norm by $\frac{1}{n}$, so that it represents averaged squared error per point, we expect (depending on the regularity of f) the order of magnitude of $n^{-1}OB(f)$ to be typically a polynomial rate $O(n^{-\nu})$ for some $\nu < 1$. Then the term σ^2/n is negligible in comparison, and the oracle inequality states that the performance of $\widehat{f}_{S_s}$ is only worse by a logarithmic factor than the performance obtained with the optimal, f-dependent choice of I in a keep-or-kill estimator.

For this tight oracle inequality to hold, it is particularly important that a Parseval frame is used. While thresholding strategies can also be applied to the coefficients of a frame that is not Parseval, the reconstruction step is less straightforward (the canonical dual frame must be computed for reconstruction from the thresholded coefficients, see Sect. 20.2.2); furthermore, an additional factor B/A comes into the bound ($A \leq 1 \leq B$ being the frame bounds from definition (20.1)) (see, for instance, Haltmeier and Munk 2014, Prop. 3.10). Therefore, the performance of simple thresholding estimates deteriorates when used with a non-Parseval frame.

[1] Candès (2006) only hints at the proof; we provide a proof in the appendix for completeness.

20.5 Numerical Experiments

We investigate the performance of the proposed method for denoising on two testbed datasets where the ground truth is known and the design points are drawn randomly iid from a distribution on a manifold. More precisely, we will consider one example where the design points $\mathfrak{D}$ are drawn uniformly ($n = 500$) on the unit square, which is then rolled up into a "swiss roll" shape in 3D. We consider a very simple target function represented (on the original unit square) as a piecewise constant function (with values 5 and -3) on two triangles, displaying a sharp discontinuity along one diagonal of the square and very smooth regularity elsewhere. This function is observed with an additional Gaussian noise of variance $\sigma^2 = 1$. In the second example the design points $\mathfrak{D}$ are drawn uniformly ($n = 500$) on the unit sphere in $\mathbb{R}^3$. The target function remains a piecewise constant function, defined on the two parts of the sphere when intersecting it with a chosen plane. Again, this function is observed with an additional Gaussian noise of variance $\sigma^2 = 1$. For the swiss roll example as well as for the sphere example, one sample consisting of design points and noisy function values is displayed in Fig. 20.4.

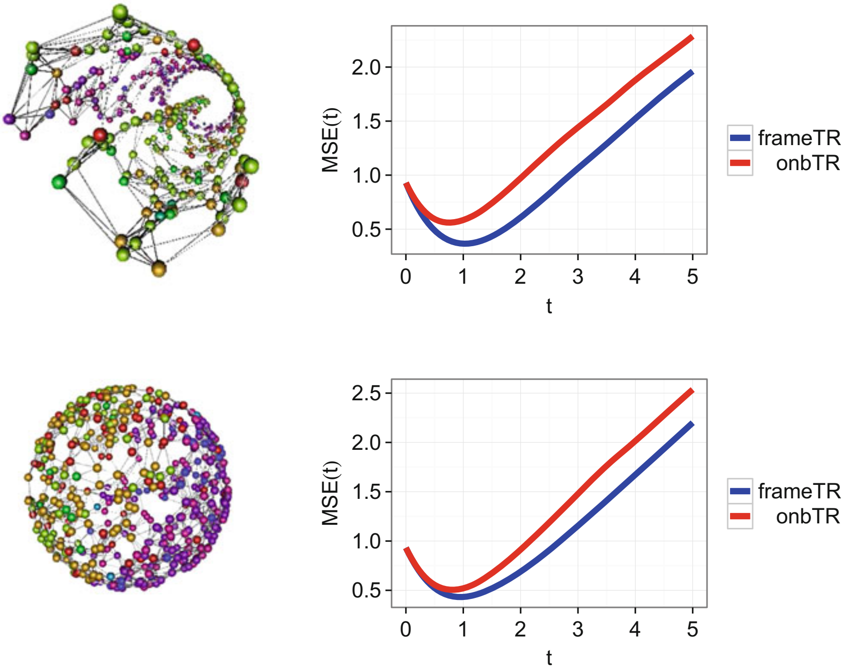


Fig. 20.4 Left: noisy function on swiss roll data (top) and sphere data (bottom), graph representation. Right: MSE for two representative settings (weighted ε-Graph and k-NN-Graph) as a function of threshold level. Red is thresholding in the original Laplacian Eigenmaps ONB, blue is thresholding of frame coefficients

In each example, we consider the different types of neighborhood graphs described in Sect. 20.2.3. Following usual heuristics, for the construction of the k-NN graph we take $k = 7 \approx \log n$; for the ε-graph, we take for ε the average distance to the $k = 7$th nearest neighbor, and for weighted graphs we take Gaussian weights, where the bandwidth λ is calibrated so that points at the distance ε defined above are given weight 0.5.

After constructing the (weighted or unweighted) graph Laplacian, we compute explicitly its eigendecomposition. For the construction of the frame via the multiscale bandpass filter, we use a $\mathscr{C}^3$ piecewise polynomial plateau function g satisfying the support constraints of Definition 3 for $b = 2$ (i.e., constant equal to 1 for $x \leq 0.5$, and zero for $x \geq 1$). While this function is not $\mathscr{C}^\infty$, it has the advantage of fast computation.

We compare the denoising performance of the following competitors: Parseval frame with soft thresholding, soft thresholding applied to the Laplacian Eigenmaps orthonormal basis, and truncated expansion in the Laplacian Eigenmaps basis (only the k coefficients corresponding to the first eigenvalues are kept, without thresholding). The latter method is in the spirit of Belkin and Niyogi (2002). It is well-known (from the regular grid case) that the "universal" theoretical threshold $\sigma\sqrt{\log n}$ is often too conservative in practice. For a fair comparison, we therefore compute the mean squared error (MSE) of both thresholding methods for varying threshold t (still modulated by $\|\Psi_{kl}\|$ for the Parseval frame). Comparison of the MSE for one sample across the t-range for two particular settings is plotted in Fig. 20.4. For all studied settings (different graph and graph Laplacian types), for the same threshold level t we observed that the frame-based method systematically shows a noticeable improvement.

In Table 20.1 we report the minimum MSEs and their standard error (averaged over $m = 50$ samples of design points and independent noise) for different methods over the possible range of the parameter (threshold level t, resp. number of coefficients for truncated expansion), both for the swissroll and for the sphere example. We observe an improvement of 20–25% across the different settings (the best overall results being obtained with weighted graphs and the unnormalized Laplacian). We also compared to the more traditional methods of kernel smoothing (Nadaraya-Watson estimator) and kernel ridge regression, using a Gaussian kernel (also with optimal choices of bandwidth and regularization parameter), and observed a comparable performance improvement. While it is not realistic to assume that the optimal parameter choice is known in practice, it is fair to compare all methods under their respective optimal parameter settings, as parameter selection methods will induce a comparable performance hit with respect to the best setting.

20.6 Outlook

Following the recently introduced idea of generalizing the Littlewood-Paley spectral decomposition, we constructed explicitly a Parseval frame of functions on a neighborhood graph formed on the data points. We established that a thresholding strategy

Table 20.1 MSE performance under optimal parameter choice

Graph	L	FrTh	LETh	LETr
Example 1: sphere, jump function,$\sigma^2 = 1, n = 500, m = 50$				
kNN	U	0.510 (0.050)	0.693 (0.061)	0.905 (0.108)
kNN	N	0.538 (0.046)	0.712 (0.055)	0.931 (0.094)
WkNN	U	0.521 (0.049)	0.652 (0.050)	0.800 (0.097)
WkNN	N	0.530 (0.049)	0.674 (0.057)	0.749 (0.091)
CGK	U	0.520 (0.055)	0.638 (0.065)	0.821 (0.107)
CGK	N	0.530 (0.052)	0.670 (0.050)	0.725 (0.081)
ϵG	U	0.505 (0.058)	0.650 (0.068)	0.865 (0.115)
ϵG	N	0.557 (0.052)	0.710 (0.059)	0.902 (0.106)
WϵG	U	0.482 (0.055)	0.622 (0.064)	0.787 (0.111)
WϵG	N	0.530 (0.049)	0.674 (0.057)	0.749 (0.091)
Smoothing Kernel Regression: min. MSE = 0.612 (0.066)				
Kernel Ridge Regression: min. MSE = 0.594 (0.051)				
Example 2: swiss roll, jump function,$\sigma^2 = 1, n = 500, m = 50$				
kNN	U	0.462 (0.043)	0.647 (0.039)	0.876 (0.079)
kNN	N	0.494 (0.043)	0.676 (0.043)	0.902 (0.071)
WkNN	U	0.443 (0.045)	0.600 (0.050)	0.790 (0.102)
WkNN	N	0.500 (0.043)	0.659 (0.045)	0.775 (0.079)
CGK	U	0.491 (0.053)	0.625 (0.057)	0.844 (0.096)
CGK	N	0.520 (0.047)	0.648 (0.049)	0.713 (0.079)
ϵG	U	0.459 (0.049)	0.610 (0.053)	0.872 (0.095)
ϵG	N	0.532 (0.045)	0.681 (0.050)	0.884 (0.089)
WϵG	U	0.441 (0.049)	0.574 (0.049)	0.793 (0.113)
WϵG	N	0.503 (0.045)	0.643 (0.051)	0.744 (0.089)
Smoothing Kernel Regression: min. MSE = 0.589 (0.082)				
Kernel Ridge Regression: min. MSE = 0.779 (0.052)				

FrTh: Frame Thresholding; LETh/LETr: Laplacian Eigenmaps Thresholding/Truncated expansion. Prefix W indicates edge weighting in the graph. CGK is the complete graph with Gaussian weights. U/N is un/normalized graph Laplacian. Standard error in brackets. Top: Sphere example. Bottom: Swiss roll example

on the frame coefficients has superior performance for the denoising problem as compared to usual, spectral or non-spectral, approaches. Future developments include extension of this methodology to the semisupervised learning setting, and a stronger theoretical basis for spatial localization.

Acknowledgements The authors acknowledge the financial support of the German DFG, under the Research Unit FOR-1735 “Structural Inference in Statistics—Adaptation and Efficiency.”

Appendix

Proof of Theorem 3

Theorem 3 states a oracle-type inequality which captures the relation of soft thresholding estimators $\hat{f}_{S_s} = \sum_{k,l} S_s\left(\langle y, \Psi_{kl}\rangle, t_{kl}\right)\Psi_{kl}$ defined in (20.16) to the collection of keep-or-kill estimators on a Parseval frame. This result is known in the literature (see Candès 2006, Section 9), but we provide a short self-contained proof for completeness, modulo a technical result from Donoho and Johnstone (1994) for soft thresholding of a single one-dimensional Gaussian variable, which is basic for the Proof of Theorem 3.

Lemma 1 *For* $0 \le \delta \le 1/2$, $t = \sqrt{2\log(\delta^{-1})}$ *and* $X \sim \mathcal{N}(\mu, 1)$

$$\begin{aligned}\mathbf{E}_X\left((S_s(X,t)-\mu)^2\right) &\le (2\log(\delta^{-1})+1)(\delta+\min(1,\mu^2))\\ &= (t^2+1)\left(\exp\left(-\frac{t^2}{2}\right)+\min(1,\mu^2)\right). \qquad (20.23)\end{aligned}$$

The proof of this lemma can be found in appendix 1 of Donoho and Johnstone (1994). Now we are able to prove Theorem 3.

Proof First note that for $y = \tau x$, $\tau > 0$, we have

$$S_s(y,u) = \tau S_s\left(x, \frac{u}{\tau}\right). \qquad (20.24)$$

Secondly we remark that

$$\frac{\langle y, \Psi_{kl}\rangle}{\sigma\left\|\Psi_{kl}\right\|} \sim \mathcal{N}\left(\frac{a_{kl}}{\sigma\left\|\Psi_{kl}\right\|}, 1\right). \qquad (20.25)$$

Considering now the risk of the soft thresholding estimator $\hat{f}_{S_s}$ we get

$$\begin{aligned}\mathbf{E}\left(\left\|\hat{f}_{S_s} - f\right\|^2\right) &= \mathbf{E}\left(\left\|\sum_{k,l}\left(S_s\left(\langle y, \Psi_{kl}\rangle, t_{kl}\right) - a_{kl}\right)\Psi_{kl}\right\|^2\right)\\ &\le \mathbf{E}\left(\sum_{k,l}\left(S_s\left(\langle y, \Psi_{kl}\rangle, t_{kl}\right) - a_{kl}\right)^2\right)\\ &= \sum_{k,l}\mathbf{E}\left(\left(S_s\left(\langle y, \Psi_{kl}\rangle, t_{kl}\right) - a_{kl}\right)^2\right). \qquad (20.26)\end{aligned}$$

by using inequality (20.6). By applying (20.24) and then (20.23) with $t = \sqrt{2\log(n)}$ it follows that

$$\begin{aligned}
\mathbf{E}\left(\left\|\hat{f}_{S_s} - f\right\|^2\right) &\leq \sum_{k,l} \sigma^2 \left\|\Psi_{kl}\right\|^2 \mathbf{E}\left(\left(S_s\left(\frac{\langle y, \Psi_{kl}\rangle}{\sigma \left\|\Psi_{kl}\right\|}, \sqrt{2\log(n)}\right) - \frac{a_{kl}}{\sigma \left\|\Psi_{kl}\right\|}\right)^2\right) \\
&\leq \sum_{k,l} \sigma^2 \left\|\Psi_{kl}\right\|^2 (2\log(n) + 1) \left(\exp\left(-\frac{2\log(n)}{2}\right) + \min\left(1, \frac{a_{kl}^2}{\sigma^2 \left\|\Psi_{kl}\right\|^2}\right)\right) \\
&= \sum_{k,l} (2\log(n) + 1) \left(\frac{1}{n}\sigma^2 \left\|\Psi_{kl}\right\|^2 + \min\left(\sigma^2 \left\|\Psi_{kl}\right\|^2, a_{kl}^2\right)\right) \\
&= (2\log(n) + 1) \left(\frac{1}{n}\sum_{k,l} \sigma^2 \left\|\Psi_{kl}\right\|^2 + \sum_{k,l} \min\left(\sigma^2 \left\|\Psi_{kl}\right\|^2, a_{kl}^2\right)\right). \qquad (20.27)
\end{aligned}$$

Recalling the Parseval frame property $\sum_{k,l} \left\|\Psi_{kl}\right\|^2 = n$, we finally obtain

$$\begin{aligned}
\mathbf{E}\left(\left\|\hat{f}_{S_s} - f\right\|^2\right) &\leq (2\log(n) + 1) \left(\frac{1}{n} n\sigma^2 + \sum_{k,l} \min\left(\sigma^2 \left\|\Psi_{kl}\right\|^2, a_{kl}^2\right)\right) \\
&= (2\log(n) + 1) \left(\sigma^2 + \sum_{k,l} \min\left(\sigma^2 \left\|\Psi_{kl}\right\|^2, a_{kl}^2\right)\right). \qquad (20.28)
\end{aligned}$$

where we recognize the upper bound $\sum_{k,l} \min\left(\sigma^2 \left\|\Psi_{kl}\right\|^2, a_{kl}^2\right) = OB(f)$ for the oracle. □

References

Belkin M, Niyogi P (2002) Using manifold structure for partially labeled classification. In: NIPS, pp 929–936

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396

Bickel P, Li B (2007) Local polynomial regression on unknown manifolds. In: Complex datasets and inverse problems: tomography, networks and beyond. IMS lecture notes, vol 54. Institute of Mathematical Statistics, Bethesda, pp 177–186

Candès E (2006) Modern statistical estimation via oracle inequalities. Acta Numer 15:257–325

Casazza P, Kutyniok G, Philipp F (2013) Introduction to finite frame theory. In: Casazza PG, Kutyniok G (eds) Finite frames, applied and numerical harmonic analysis. Birkhäuser, Boston, pp 1–53

Christensen O (2008) Frames and bases: an introductory course. In: Applied and numerical harmonic analysis. Birkhäuser, Boston

Coulhon T, Kerkyacharian G, Petrushev P (2012) Heat kernel generated frames in the setting of Dirichlet spaces. J Fourier Anal Appl 18(5):995–1066

Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81(3):425–455

Gavish M, Nadler B, Coifman RR (2010) Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning. In: Fürnkranz J, Joachims T (eds) ICML. Omnipress, Madison, pp 367–374

Haltmeier M, Munk A (2014) Extreme value analysis of empirical frame coefficients and implications for denoising by soft-thresholding. Appl Comput Harmon Anal 36(3):434–460. https://doi.org/10.1016/j.acha.2013.07.004

Hammond DK, Vandergheynst P, Gribonval R (2011) Wavelets on graphs via spectral graph theory. Appl Comput Harmon Anal 30(2):129–150

Han D (2007) Frames for undergraduates. In: Student mathematical library. American Mathematical Society, Providence

Kpotufe S (2011) k-NN regression adapts to local intrinsic dimension. In: NIPS, pp 729–737

Kpotufe S, Dasgupta S (2012) A tree-based regressor that adapts to intrinsic dimension. J Comput Syst Sci 78(5):1496–1515

Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326

Tenenbaum J, de Silva V, Langford J (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323