

# Kernel-based Reconstruction of Graph Signals

Daniel Romero, *Member, IEEE*, Meng Ma, Georgios B. Giannakis, *Fellow, IEEE*

**Abstract**—A number of applications in engineering, social sciences, physics, and biology involve inference over networks. In this context, graph signals are widely encountered as descriptors of vertex attributes or features in graph-structured data. Estimating such signals in all vertices given noisy observations of their values on a subset of vertices has been extensively analyzed in the literature of **signal processing on graphs (SPoG)**. This paper advocates kernel regression as a framework generalizing popular SPoG modeling and reconstruction and expanding their capabilities. Formulating signal reconstruction as a regression task on reproducing kernel Hilbert spaces of graph signals permeates benefits from statistical learning, offers fresh insights, and allows for estimators to leverage richer forms of prior information than existing alternatives. A number of SPoG notions such as bandlimitedness, graph filters, and the graph Fourier transform are naturally accommodated in the kernel framework. Additionally, this paper capitalizes on the so-called representer theorem to devise simpler versions of existing Tikhonov regularized estimators, and offers a novel probabilistic interpretation of kernel methods on graphs based on graphical models. Motivated by the challenges of selecting the bandwidth parameter in SPoG estimators or the kernel map in kernel-based methods, the present paper further proposes two multi-kernel approaches with complementary strengths. Whereas the first enables estimation of the unknown bandwidth of bandlimited signals, the second allows for efficient graph filter selection. Numerical tests with synthetic as well as real data demonstrate the merits of the proposed methods relative to state-of-the-art alternatives.

**Index Terms**—Graph signal reconstruction, kernel regression, multi-kernel learning.

## I. INTRODUCTION

Graph data play a central role in analysis and inference tasks for social, brain, communication, biological, transportation, and sensor networks [1], thanks to their ability to capture relational information. Vertex attributes or features associated with vertices can be interpreted as functions or signals defined on graphs. In social networks for instance, where a vertex represents a person and an edge corresponds to a friendship relation, such a function may denote e.g. the person's age, location, or rating of a given movie.

Research efforts over the last years are centered on estimating or processing functions on graphs; see e.g. [1]–[6]. Existing approaches rely on the premise that signals obey a certain form of parsimony relative to the graph topology. **For instance, it seems reasonable to estimate a person's age by looking at their friends' age.** The present paper deals with a general version of this task, where the goal is to estimate a graph signal given noisy observations on a subset of vertices.

The machine learning community has already looked at SPoG-related issues in the context of *semi-supervised learning*

under the term of *transductive* regression and classification [6]–[8]. Existing approaches rely on smoothness assumptions for inference of processes over graphs using *nonparametric* methods [2], [3], [6], [9]. Whereas some works consider estimation of real-valued signals [7]–[10], most in this body of literature have focused on estimating binary-valued functions; see e.g. [6]. **On the other hand, function estimation has also been investigated recently by the community of signal processing on graphs (SPoG) under the term *signal reconstruction* [11]–[18].** Existing approaches commonly adopt *parametric* estimation tools and rely on *bandlimitedness*, by which the signal of interest is assumed to lie in the span of the  **$B$  leading eigenvectors of the graph Laplacian or the adjacency matrix** [12]–[14], [16]–[19]. Different from machine learning works, SPoG research is mainly concerned with estimating real-valued functions.

The present paper cross-pollinates ideas and broadens both machine learning and SPoG perspectives under the *unifying* framework of kernel-based learning. The first part unveils the implications of adopting this standpoint and demonstrates how it naturally accommodates a number of SPoG concepts and tools. From a high level, this connection (i) brings to bear performance bounds and algorithms from transductive regression [8] and the extensively analyzed general kernel methods (see e.g. [20]); (ii) offers the possibility of reducing the dimension of the optimization problems involved in Tikhonov regularized estimators by invoking the so-called *representer theorem* [21]; and, (iii) it provides guidelines for systematically selecting parameters in existing signal reconstruction approaches by leveraging the connection with linear minimum mean-square error (LMMSE) estimation via *covariance kernels*.

Further implications of applying kernel methods to graph signal reconstruction are also explored. Specifically, it is shown that the finite dimension of graph signal spaces allows for an insightful proof of the representer theorem which, different from existing proofs relying on functional analysis, solely involves linear algebra arguments. Moreover, an intuitive probabilistic interpretation of graph kernel methods is introduced based on graphical models. These findings are complemented with a technique to deploy regression with Laplacian kernels in big-data setups.

It is further established that a number of existing signal reconstruction approaches, including the least-squares (LS) estimators for bandlimited signals from [11]–[16]; the Tikhonov regularized estimators from [4], [12], [22] and [23, eq. (27)]; and the maximum a posteriori estimator in [13], can be viewed as kernel methods on *reproducing kernel Hilbert spaces* (RKHSs) of graph signals. Popular notions in SPoG such as graph filters, the graph Fourier transform, and bandlimited signals can also be accommodated under the kernel framework. First, it is seen that a *graph filter* [4] is essentially a kernel smoother [24]. Second,

This work was supported by ARO grant W911NF-15-1-0492 and NSF grants 1343248, 1442686, and 1514056.

The authors are with the Dept. of ECE and the Digital Tech. Center, Univ. of Minnesota, USA. E-mail: {dromero,maxxx971,georgios}@umn.edu.

*bandlimited kernels* are introduced to accommodate estimation of bandlimited signals. Third, the connection between the so-called *graph Fourier transform* [4] (see [5], [15] for a related definition) and Laplacian kernels [2], [3] is delineated. Relative to methods relying on the bandlimited property (see e.g. [11]–[17]), kernel methods offer increased flexibility in leveraging prior information about the graph Fourier transform of the estimated signal.

The second part of the paper pertains to the challenge of model selection. On the one hand, a number of reconstruction schemes in SPoG [12]–[15], [17] require knowledge of the signal bandwidth, which is typically unknown [11], [16]. Existing approaches for determining this bandwidth rely solely on the set of sampled vertices, disregarding the observations [11], [16]. On the other hand, existing kernel-based approaches [1, Ch. 8] necessitate proper kernel selection, which is computationally inefficient through cross-validation.

The present paper addresses both issues by **means of two multi-kernel learning (MKL) techniques** having complementary strengths. Heed existing MKL methods on graphs are confined to estimating binary-valued signals [25]–[27]. This paper on the other hand, is concerned with MKL algorithms for real-valued graph signal reconstruction. The novel graph MKL algorithms optimally combine the kernels in a given dictionary and simultaneously estimate the graph signal by solving a single optimization problem.

The rest of the paper is structured as follows. Sec. II formulates the problem of graph signal reconstruction. Sec. III presents kernel-based learning as an encompassing framework for graph signal reconstruction, and explores the implications of adopting such a standpoint. Two MKL algorithms are then presented in Sec. IV. Sec. V complements analytical findings with numerical tests by comparing with competing alternatives via synthetic- and real-data experiments. Finally, concluding remarks are highlighted in Sec. VI.

**Notation.**  $(\cdot)_N$  denotes the remainder of integer division by  $N$ ;  $\delta[\cdot]$  the Kronecker delta, and  $\mathcal{I}[C]$  the indicator of condition  $C$ , returning 1 if  $C$  is satisfied and 0 otherwise. Scalars are denoted by lowercase letters, vectors by bold lowercase, and matrices by bold uppercase. The  $(i, j)$ th entry of matrix  $\mathbf{A}$  is  $(\mathbf{A})_{i,j}$ . Notation  $\|\cdot\|_2$  and  $\text{Tr}(\cdot)$  respectively represent Euclidean norm and trace;  $\mathbf{I}_N$  denotes the  $N \times N$  identity matrix;  $\mathbf{e}_i$  is the  $i$ -th canonical vector of  $\mathbb{R}^M$ , while  $\mathbf{0}$  ( $\mathbf{1}$ ) is a vector of appropriate dimension with all (ones). The span of the columns of  $\mathbf{A}$  is denoted by  $\mathcal{R}\{\mathbf{A}\}$ , whereas  $\mathbf{A} \succ \mathbf{B}$  (resp.  $\mathbf{A} \succeq \mathbf{B}$ ) means that  $\mathbf{A} - \mathbf{B}$  is positive definite (resp. semi-definite). Superscripts  $T$  and  $^\dagger$  respectively stand for transposition and pseudo-inverse, whereas  $\mathbb{E}$  denotes expectation.

## II. PROBLEM STATEMENT

A graph is a tuple  $\mathcal{G} := (\mathcal{V}, w)$ , where  $\mathcal{V} := \{v_1, \dots, v_N\}$  is the vertex set, and  $w : \mathcal{V} \times \mathcal{V} \rightarrow [0, +\infty)$  is a map assigning a weight to each vertex pair. For simplicity, it is assumed that  $w(v, v) = 0 \ \forall v \in \mathcal{V}$ . This paper focuses on *undirected* graphs, for which  $w(v, v') = w(v', v) \ \forall v, v' \in \mathcal{V}$ . A graph is said to be *unweighted* if  $w(v, v')$  is either 0 or 1. The edge set  $\mathcal{E}$  is

the support of  $w$ , i.e.,  $\mathcal{E} := \{(v, v') \in \mathcal{V} \times \mathcal{V} : w(v, v') \neq 0\}$ . Two vertices  $v$  and  $v'$  are *adjacent*, *connected*, or *neighbors* if  $(v, v') \in \mathcal{E}$ . **The  $n$ -th neighborhood  $\mathcal{N}_n$  is the set of neighbors of  $v_n$ , i.e.,  $\mathcal{N}_n := \{v \in \mathcal{V} : (v, v_n) \in \mathcal{E}\}$ .** The information in  $w$  is compactly represented by the  $N \times N$  weighted adjacency matrix  $\mathbf{W}$ , whose  $(n, n')$ -th entry is  $w(v_n, v_{n'})$ ; the  $N \times N$  diagonal *degree* matrix  $\mathbf{D}$ , whose  $(n, n)$ -th entry is  $\sum_{n'=1}^N w(v_n, v_{n'})$ ; and the *Laplacian* matrix  $\mathbf{L} := \mathbf{D} - \mathbf{W}$ , which is symmetric and positive semidefinite [1, Ch. 2]. The latter is sometimes replaced with its normalized version  $\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ , whose eigenvalues are confined to the interval  $[0, 2]$ .

A real-valued function (or signal) on a graph is a map  $f_0 : \mathcal{V} \rightarrow \mathbb{R}$ . As mentioned in Sec. I, the value  $f_0(v)$  represents an attribute or feature of  $v \in \mathcal{V}$ , such as age, political alignment, or annual income of a person in a social network. Signal  $f_0$  is thus represented by  $\mathbf{f}_0 := [f_0(v_1), \dots, f_0(v_N)]^T$ .

**Suppose that a collection of noisy samples (or observations)  $y_s = f_0(v_{n_s}) + e_s$ ,  $s = 1, \dots, S$ , is available, where  $e_s$  models noise and  $\mathcal{S} := \{n_1, \dots, n_S\}$  contains the indices  $1 \leq n_1 < \dots < n_S \leq N$  of the sampled vertices. In a social network, this may be the case if a subset of persons have been surveyed about the attribute of interest (e.g. political alignment). Given  $\{(n_s, y_s)\}_{s=1}^S$ , and assuming knowledge of  $\mathcal{G}$ , the goal is to estimate  $f_0$ . This will provide estimates of  $f_0(v)$  both at observed and unobserved vertices  $v \in \mathcal{V}$ . By defining  $\mathbf{y} := [y_1, \dots, y_S]^T$ , the observation model is summarized as**

$$\mathbf{y} = \mathbf{\Phi} \mathbf{f}_0 + \mathbf{e} \quad (1)$$

where  $\mathbf{e} := [e_1, \dots, e_S]^T$  and  $\mathbf{\Phi}$  is an  $S \times N$  matrix with entries  $(s, n_s)$ ,  $s = 1, \dots, S$ , set to one, and the rest set to zero.

## III. UNIFYING THE RECONSTRUCTION OF GRAPH SIGNALS

Kernel methods constitute the “workhorse” of statistical learning for nonlinear function estimation [20]. **Their popularity can be ascribed to their simplicity, flexibility, and good performance.** This section presents kernel regression as a novel unifying framework for graph signal reconstruction.

Kernel regression seeks an estimate of  $f_0$  in an RKHS  $\mathcal{H}$ , which is the space of functions  $f : \mathcal{V} \rightarrow \mathbb{R}$  defined as

$$\mathcal{H} := \left\{ f : f(v) = \sum_{n=1}^N \bar{\alpha}_n \kappa(v, v_n), \bar{\alpha}_n \in \mathbb{R} \right\}. \quad (2)$$

The *kernel map*  $\kappa : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$  is any function defining a symmetric and positive semidefinite  $N \times N$  matrix with entries  $[\bar{\mathbf{K}}]_{n,n'} := \kappa(v_n, v_{n'}) \ \forall n, n'$  [28]. Intuitively,  $\kappa(v, v')$  is a basis function in (2) measuring similarity between the values of  $f_0$  at  $v$  and  $v'$ . For instance, if a *feature vector*  $\mathbf{x}_n \in \mathbb{R}^D$  containing attributes of the entity represented by  $v_n$  is known for  $n = 1, \dots, N$ , one can employ the popular *Gaussian kernel*  $\kappa(v_n, v_{n'}) = \exp\{-\|\mathbf{x}_n - \mathbf{x}_{n'}\|^2 / \sigma^2\}$ , where  $\sigma^2 > 0$  is a user-selected parameter [20]. When such feature vectors  $\mathbf{x}_n$  are not available, the graph topology can be leveraged to construct graph kernels as detailed in Sec. III-B.

Different from RKHSs of functions  $f(\mathbf{x})$  defined over infinite sets, the expansion in (2) is finite since  **$\mathcal{V}$  is finite.** This implies

that RKHSs of graph signals are finite-dimensional spaces. From (2), it follows that any signal in  $\mathcal{H}$  can be expressed as:

$$\mathbf{f} := [f(v_1), \dots, f(v_N)]^T = \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}} \quad (3)$$

for some  $N \times 1$  vector  $\bar{\boldsymbol{\alpha}} := [\bar{\alpha}_1, \dots, \bar{\alpha}_N]^T$ . Given two functions  $f(v) := \sum_{n=1}^N \bar{\alpha}_n \kappa(v, v_n)$  and  $f'(v) := \sum_{n=1}^N \bar{\alpha}'_n \kappa(v, v_n)$ , their RKHS inner product is defined as<sup>1</sup>

$$\langle f, f' \rangle_{\mathcal{H}} := \sum_{n=1}^N \sum_{n'=1}^N \bar{\alpha}_n \bar{\alpha}'_{n'} \kappa(v_n, v_{n'}) = \bar{\boldsymbol{\alpha}}^T \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}}' \quad (4)$$

where  $\bar{\boldsymbol{\alpha}}' := [\bar{\alpha}'_1, \dots, \bar{\alpha}'_N]^T$ . The RKHS norm is defined by

$$\|f\|_{\mathcal{H}}^2 := \langle f, f \rangle_{\mathcal{H}} = \bar{\boldsymbol{\alpha}}^T \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}} \quad (5)$$

and will be used as a regularizer to control overfitting. As a special case, setting  $\bar{\mathbf{K}} = \mathbf{I}_N$  recovers the standard inner product  $\langle f, f' \rangle_{\mathcal{H}} = \mathbf{f}^T \mathbf{f}'$ , and Euclidean norm  $\|f\|_{\mathcal{H}}^2 = \|\mathbf{f}\|_2^2$ . Note that when  $\bar{\mathbf{K}} \succ \mathbf{0}$ , the set of functions of the form (3) equals  $\mathbb{R}^N$ . Thus, two RKHSs with strictly positive definite kernel matrices contain the same functions. They differ only in their RKHS inner products and norms. Interestingly, this observation establishes that any positive definite kernel is *universal* [29] for graph signal reconstruction.

The term *reproducing kernel* stems from the reproducing property. Let  $\kappa(\cdot, v_{n_0})$  denote the map  $v \mapsto \kappa(v, v_{n_0})$ , where  $n_0 \in \{1, \dots, N\}$ . Using (4), the reproducing property can be expressed as  $\langle \kappa(\cdot, v_{n_0}), \kappa(\cdot, v_{n'_0}) \rangle_{\mathcal{H}} = \mathbf{e}_{n_0}^T \bar{\mathbf{K}} \mathbf{e}_{n'_0} = \kappa(v_{n_0}, v_{n'_0})$ . Due to the linearity of inner products and the fact that all signals in  $\mathcal{H}$  are the superposition of functions of the form  $\kappa(\cdot, v_n)$ , the reproducing property asserts that inner products can be obtained just by evaluating  $\kappa$ . The reproducing property is of paramount importance when dealing with an RKHS of functions defined on *infinite* spaces (thus excluding RKHSs of graph signals), since it offers an efficient alternative to the costly multidimensional integration required by inner products such as  $\langle f_1, f_2 \rangle_{L^2} := \int_{\mathcal{X}} f_1(\mathbf{x}) f_2(\mathbf{x}) d\mathbf{x}$ .

Given  $\{y_s\}_{s=1}^S$ , RKHS-based function estimators are obtained by solving functional minimization problems formulated as

$$\hat{f}_0 := \arg \min_{f \in \mathcal{H}} \frac{1}{S} \sum_{s=1}^S \mathcal{L}(v_{n_s}, y_s, f(v_{n_s})) + \mu \Omega(\|f\|_{\mathcal{H}}) \quad (6)$$

where the regularization parameter  $\mu > 0$  controls overfitting, the increasing function  $\Omega$  is used to promote smoothness, and the loss function  $\mathcal{L}$  measures how estimates deviate from the data. The so-called *square loss*  $\mathcal{L}(v_{n_s}, y_s, f(v_{n_s})) := [y_s - f(v_{n_s})]^2$  constitutes a popular choice for  $\mathcal{L}$ , whereas  $\Omega$  is often set to  $\Omega(\zeta) = |\zeta|$  or  $\Omega(\zeta) = \zeta^2$ .

To simplify notation, consider loss functions expressible as  $\mathcal{L}(v_{n_s}, y_s, f(v_{n_s})) = \mathcal{L}(y_s - f(v_{n_s}))$ ; extensions to more general cases are straightforward. The vector-version of such a function is  $\mathcal{L}(\mathbf{y} - \Phi \mathbf{f}) := \sum_{s=1}^S \mathcal{L}(y_s - f(v_{n_s}))$ . Substituting (3) and (5) into (6) shows that  $\hat{f}_0$  can be obtained as  $\hat{\mathbf{f}}_0 = \bar{\mathbf{K}} \hat{\boldsymbol{\alpha}}$ , where

$$\hat{\boldsymbol{\alpha}} := \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{1}{S} \mathcal{L}(\mathbf{y} - \Phi \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}}) + \mu \Omega((\bar{\boldsymbol{\alpha}}^T \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}})^{1/2}). \quad (7)$$

<sup>1</sup>Whereas  $f$  denotes a *function*, symbol  $f(v)$  represents the *scalar* resulting from evaluating  $f$  at vertex  $v$ .

An alternative form of (7) that will be frequently used in the sequel results upon noting that  $\bar{\boldsymbol{\alpha}}^T \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}^T \bar{\mathbf{K}} \bar{\mathbf{K}}^\dagger \bar{\mathbf{K}} \bar{\boldsymbol{\alpha}} = \mathbf{f}^T \bar{\mathbf{K}}^\dagger \mathbf{f}$ . Thus, one can rewrite (7) as

$$\hat{f}_0 := \arg \min_{\mathbf{f} \in \mathcal{R}\{\bar{\mathbf{K}}\}} \frac{1}{S} \mathcal{L}(\mathbf{y} - \Phi \mathbf{f}) + \mu \Omega((\mathbf{f}^T \bar{\mathbf{K}}^\dagger \mathbf{f})^{1/2}). \quad (8)$$

If  $\bar{\mathbf{K}} \succ \mathbf{0}$ , the constraint  $\mathbf{f} \in \mathcal{R}\{\bar{\mathbf{K}}\}$  can be omitted, and  $\bar{\mathbf{K}}^\dagger$  can be replaced with  $\bar{\mathbf{K}}^{-1}$ . If  $\bar{\mathbf{K}}$  contains null eigenvalues, it is customary to remove the constraint by replacing  $\bar{\mathbf{K}}$  (or  $\bar{\mathbf{K}}^\dagger$ ) with a perturbed version  $\bar{\mathbf{K}} + \epsilon \mathbf{I}$  (respectively  $\bar{\mathbf{K}}^\dagger + \epsilon \mathbf{I}$ ), where  $\epsilon > 0$  is a small constant. Expression (8) shows that kernel regression unifies and subsumes the Tikhonov-regularized graph signal reconstruction schemes in [4], [12], [22] and [23, eq. (27)] by properly selecting  $\bar{\mathbf{K}}$ ,  $\mathcal{L}$ , and  $\Omega$  (see Sec. III-B).

#### A. Representer theorem

Although graph signals can be reconstructed from (7), such an approach involves optimizing over  $N$  variables. This section shows that a solution can be obtained by solving an optimization problem in  $S$  variables, where typically  $S \ll N$ .

The representer theorem [21], [28] plays an instrumental role in the non-graph setting of infinite-dimensional  $\mathcal{H}$ , where (6) cannot be directly solved. This theorem enables a solver by providing a finite parameterization of the function  $\hat{f}_0$  in (6). On the other hand, when  $\mathcal{H}$  comprises graph signals, (6) is inherently finite-dimensional and can be solved directly. However, the representer theorem can still be beneficial to reduce the dimension of the optimization in (7).

**Theorem 1** (Representer theorem). *The solution to the functional minimization in (6) can be expressed as*

$$\hat{f}_0(v) = \sum_{s=1}^S \alpha_s \kappa(v, v_{n_s}) \quad (9)$$

for some  $\alpha_s \in \mathbb{R}$ ,  $s = 1, \dots, S$ .

The conventional proof for the representer theorem involves tools from functional analysis [28]. However, when  $\mathcal{H}$  comprises functions defined on finite spaces, such as graph signals, an insightful proof can be obtained relying solely on linear algebra arguments (see Appendix A).

Since the solution  $\hat{f}_0$  of (6) lies in  $\mathcal{H}$ , it can be expressed as  $\hat{f}_0 = \sum_{n=1}^N \bar{\alpha}_n \kappa(v, v_n)$  for some  $\{\bar{\alpha}_n\}_{n=1}^N$ . Theorem 1 states that the terms corresponding to unobserved vertices  $v_n$ ,  $n \notin \mathcal{S}$ , play no role in the kernel expansion of the estimate; that is,  $\bar{\alpha}_n = 0$ ,  $\forall n \notin \mathcal{S}$ . Thus, whereas (7) requires optimization over  $N$  variables, Theorem 1 establishes that a solution can be found by solving a problem in  $S$  variables, where typically  $S \ll N$ . Clearly, this conclusion carries over to the signal reconstruction schemes in [4], [12], [22] and [23, eq. (27)], since they constitute special instances of kernel regression. The fact that the number of parameters to be estimated after applying Theorem 1 depends on (in fact, equals) the number of samples  $S$  justifies why  $\hat{f}_0$  in (6) is referred to as a *nonparametric estimate*.

Theorem 1 shows the form of  $\hat{f}_0$  but does not provide the optimal  $\{\alpha_s\}_{s=1}^S$ , which is found after substituting (9) into (6)



and solving the resulting optimization problem with respect to these coefficients. To this end, let  $\alpha := [\alpha_1, \dots, \alpha_S]^T$ , and write  $\bar{\alpha} = \Phi^T \alpha$  to deduce that

$$\hat{f}_0 = \bar{K} \bar{\alpha} = \bar{K} \Phi^T \alpha. \quad (10)$$

From (7) and (10), the optimal  $\alpha$  can be found as

$$\hat{\alpha} := \arg \min_{\alpha \in \mathbb{R}^S} \frac{1}{S} \mathcal{L}(\mathbf{y} - \mathbf{K} \alpha) + \mu \Omega((\alpha^T \mathbf{K} \alpha)^{1/2}) \quad (11)$$

where  $\mathbf{K} := \Phi \bar{\mathbf{K}} \Phi^T$ .

**Example 1 (kernel ridge regression).** For  $\mathcal{L}$  chosen as the square loss and  $\Omega(\zeta) = \zeta^2$ , the  $\hat{f}_0$  in (6) is referred to as the *kernel ridge regression* estimate. It is given by  $\hat{f}_{\text{RR}} = \bar{\mathbf{K}} \hat{\alpha}$ , where

$$\hat{\alpha} := \arg \min_{\alpha \in \mathbb{R}^S} \frac{1}{S} \|\mathbf{y} - \mathbf{K} \alpha\|^2 + \mu \alpha^T \mathbf{K} \alpha \quad (12a)$$

$$= (\mathbf{K} + \mu S \mathbf{I}_S)^{-1} \mathbf{y}. \quad (12b)$$

Therefore,  $\hat{f}_{\text{RR}}$  can be expressed as

$$\hat{f}_{\text{RR}} = \bar{\mathbf{K}} \Phi^T (\mathbf{K} + \mu S \mathbf{I}_S)^{-1} \mathbf{y}. \quad (13)$$

As seen in the next section, (13) generalizes a number of existing signal reconstructors upon properly selecting  $\bar{\mathbf{K}}$ . Thus, Theorem 1 can also be used to simplify Tikhonov-regularized estimators such as the one in [12, eq. (15)]. To see this, just note that (13) inverts an  $S \times S$  matrix whereas [12, eq. (16)] entails the inversion of an  $N \times N$  matrix.

**Example 2 (support vector regression).** If  $\mathcal{L}$  equals the so-called  $\epsilon$ -insensitive loss  $\mathcal{L}(v_{n_s}, y_s, f(v_{n_s})) := \max(0, |y_s - f(v_{n_s})| - \epsilon)$  and  $\Omega(\zeta) = \zeta^2$ , then (6) constitutes a support vector machine for regression (see e.g. [20, Ch. 1]).

## B. Graph kernels for signal reconstruction

When estimating functions on graphs, conventional kernels such as the aforementioned Gaussian kernel cannot be applied because **the underlying set where graph signals are defined is not a metric space**. Indeed, no vertex addition  $v_n + v_{n'}$ , scaling  $\beta v_n$ , or norm  $\|v_n\|$  can be naturally defined on  $\mathcal{V}$ . An alternative is to embed  $\mathcal{V}$  into an Euclidean space via a feature map  $\phi: \mathcal{V} \rightarrow \mathbb{R}^D$ , and apply a conventional kernel afterwards. However, for a given graph it is generally unclear how to design such a map or select  $D$ , which motivates the adoption of graph kernels [3]. **The rest of this section elaborates on three classes of graph kernels, namely Laplacian, bandlimited, and novel covariance kernels for reconstructing graph signals.**

1) *Laplacian kernels:* The term Laplacian kernel comprises a wide family of kernels obtained by applying a certain function to the Laplacian matrix  $\mathbf{L}$ . From a theoretical perspective, Laplacian kernels are well motivated since they constitute the graph counterpart of the so-called *translation invariant kernels* in Euclidean spaces [3]. This section reviews Laplacian kernels, provides novel insights in terms of interpolating signals, and highlights their versatility in capturing prior information about the *graph Fourier transform* of the estimated signal.

Let  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$  denote the eigenvalues of the graph Laplacian matrix  $\mathbf{L}$ , and consider the eigendecomposition

$\mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$ , where  $\mathbf{\Lambda} := \text{diag}\{\lambda_1, \dots, \lambda_N\}$ . A Laplacian kernel is a kernel map  $\kappa$  generating a matrix  $\bar{\mathbf{K}}$  of the form

$$\bar{\mathbf{K}} := r^\dagger(\mathbf{L}) := \mathbf{U} r^\dagger(\mathbf{\Lambda}) \mathbf{U}^T \quad (14)$$

where  $r(\mathbf{\Lambda})$  is the result of applying the user-selected non-negative map  $r: \mathbb{R} \rightarrow \mathbb{R}_+$  to the diagonal entries of  $\mathbf{\Lambda}$ . For reasons that will become clear, the map  $r(\lambda)$  is typically increasing in  $\lambda$ . Common choices include the diffusion kernel  $r(\lambda) = \exp\{\sigma^2 \lambda / 2\}$  [2], and the  $p$ -step random walk kernel  $r(\lambda) = (a - \lambda)^{-p}$ ,  $a \geq 2$  [3]. Laplacian regularization [3], [4], [9], [30], [31] is effected by setting  $r(\lambda) = 1 + \sigma^2 \lambda$  with  $\sigma^2$  sufficiently large.

Observe that obtaining  $\bar{\mathbf{K}}$  generally requires an eigendecomposition of  $\mathbf{L}$ , which is computationally challenging for large graphs ( $N \gg$ ). Two techniques to reduce complexity in these *big data* scenarios are proposed in Appendix B.

At this point, it is prudent to offer interpretations and insights into the principles behind the operation of Laplacian kernels. Towards this objective, note first that the regularizer from (8) is an increasing function of

$$\mathbf{f}^T \bar{\mathbf{K}}^\dagger \mathbf{f} = \mathbf{f}^T \mathbf{U} r(\mathbf{\Lambda}) \mathbf{U}^T \mathbf{f} = \tilde{\mathbf{f}}^T r(\mathbf{\Lambda}) \tilde{\mathbf{f}} = \sum_{n=1}^N r(\lambda_n) |\tilde{f}_n|^2 \quad (15)$$

where  $\tilde{\mathbf{f}} := \mathbf{U}^T \mathbf{f} := [\tilde{f}_1, \dots, \tilde{f}_N]^T$  comprises the projections of  $\mathbf{f}$  onto the eigenvectors of  $\mathbf{L}$ , and is referred to as the *graph Fourier transform* of  $\mathbf{f}$  in the SPoG parlance [4]. Before interpreting (15), it is worth elucidating the rationale behind this term. Since  $\mathbf{U} := [\mathbf{u}_1, \dots, \mathbf{u}_N]$  is orthogonal, one can decompose  $\mathbf{f}$  as

$$\mathbf{f} = \sum_{n=1}^N \tilde{f}_n \mathbf{u}_n. \quad (16)$$

Because vectors  $\{\mathbf{u}_n\}_{n=1}^N$ , or more precisely their signal counterparts  $\{u_n\}_{n=1}^N$ , are *eigensignals* of the so-called *graph shift operator*  $\mathbf{u} \mapsto \mathbf{L} \mathbf{u}$ , (16) resembles the classical Fourier transform in the sense that it expresses a signal as a superposition of *eigensignals* of a Laplacian operator [4]. Recalling from Sec. II that  $w(v_n, v_{n'})$  denotes the weight of the edge between  $v_n$  and  $v_{n'}$ , one can consider the smoothness measure for graph functions  $f$  given by

$$\partial f := \frac{1}{2} \sum_{n=1}^N \sum_{n' \in \mathcal{N}_n} w(v_n, v_{n'}) [f(v_n) - f(v_{n'})]^2 = \mathbf{f}^T \mathbf{L} \mathbf{f}$$

where the last equality follows from the definition of  $\mathbf{L} := \mathbf{D} - \mathbf{W}$ . Clearly, it holds  $\partial u_n = \lambda_n$ . Since  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ , it follows that  $0 = \partial u_1 \leq \dots \leq \partial u_N$ . In analogy to signal processing for time signals, where lower frequencies correspond to smoother eigensignals, the index  $n$ , or alternatively the eigenvalue  $\lambda_n$ , is interpreted as the *frequency* of  $u_n$ .

It follows from (15) that the regularizer in (8) strongly penalizes those  $\tilde{f}_n$  for which the corresponding  $r(\lambda_n)$  is large, thus promoting a specific structure in this frequency domain. Specifically, one prefers  $r(\lambda_n)$  to be large whenever  $|\tilde{f}_n|^2$  is

small and vice versa. The fact that  $|\tilde{f}_n|^2$  is expected to decrease with  $n$  for smooth  $f$ , motivates the adoption of an increasing  $r$  [3]. Observe that Laplacian kernels can capture richer forms of prior information than the signal reconstructors of bandlimited signals in [12]–[15], [17], [18], since the latter can solely capture the support of the Fourier transform whereas the former can also leverage magnitude information.

**Example 3 (circular graphs).** This example capitalizes on Theorem 1 to present a novel SPoG-inspired intuitive interpretation of nonparametric regression with Laplacian kernels. To do so, a closed-form expression for the Laplacian kernel matrix of a circular graph (or ring) will be derived. This class of graphs has been commonly employed in the literature to illustrate connections between SPoG and signal processing of time-domain signals [5].

Up to vertex relabeling, an unweighted circular graph satisfies  $w(v_n, v_{n'}) = \delta[(n - n')_N - 1] + \delta[(n' - n)_N - 1]$ . Therefore, its Laplacian matrix can be written as  $\mathbf{L} = 2\mathbf{I}_N - \mathbf{R} - \mathbf{R}^T$ , where  $\mathbf{R}$  is the rotation matrix resulting from circularly shifting the columns of  $\mathbf{I}_N$  one position to the right, i.e.,  $(\mathbf{R})_{n,n'} := \delta[(n' - n)_N - 1]$ . Matrix  $\mathbf{L}$  is *circulant* since its  $n$ -th row can be obtained by circularly shifting the  $(n - 1)$ -st row one position to the right. Hence,  $\mathbf{L}$  can be diagonalized by the standard Fourier matrix [32], meaning

$$\mathbf{L} = \tilde{\mathbf{U}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{U}}^H \quad (17)$$

where  $(\tilde{\mathbf{U}})_{m,m'} := (1/\sqrt{N}) \exp\{j2\pi(m - 1)(m' - 1)/N\}$  is the unitary inverse discrete Fourier transform matrix and  $(\tilde{\mathbf{\Lambda}})_{m,m'} := 2[1 - \cos(2\pi(m - 1)/N)]\delta[m - m']$ . Matrices  $\tilde{\mathbf{U}}$  and  $\tilde{\mathbf{\Lambda}}$  replace  $\mathbf{U}$  and  $\mathbf{\Lambda}$  since, for notational brevity, the eigendecomposition (17) involves complex-valued eigenvectors and the eigenvalues have not been sorted in ascending order.

From (14), a Laplacian kernel matrix is given by  $\tilde{\mathbf{K}} := \tilde{\mathbf{U}} r^\dagger(\tilde{\mathbf{\Lambda}}) \tilde{\mathbf{U}}^H := \tilde{\mathbf{U}} \text{diag}\{\mathbf{d}\} \tilde{\mathbf{U}}^H$ , where  $\mathbf{d} := [d_0, \dots, d_{N-1}]^T$  has entries  $d_n = r^\dagger(2[1 - \cos(2\pi n/N)])$ . It can be easily seen that  $(\tilde{\mathbf{K}})_{m,m'} = D_{m-m'}$ , where

$$D_m := \text{IDFT}\{d_n\} := \frac{1}{N} \sum_{n=0}^{N-1} d_n e^{j\frac{2\pi}{N}mn}.$$

If  $r(2[1 - \cos(2\pi n/N)]) > 0 \forall n$ , one has that

$$D_m = \frac{1}{N} \sum_{n=0}^{N-1} \frac{e^{j\frac{2\pi}{N}mn}}{r(2[1 - \cos(2\pi n/N)])}. \quad (18)$$

Recall that Theorem 1 dictates  $\hat{\mathbf{f}}_0 = \sum_{s \in \mathcal{S}} \bar{\alpha}_s \bar{\mathbf{k}}_s$ , where  $\bar{\mathbf{K}} := [\bar{\mathbf{k}}_1, \dots, \bar{\mathbf{k}}_N]$ . Since  $(\bar{\mathbf{K}})_{m,m'} = D_{m-m'}$  and because  $D_m$  is periodic in  $m$  with period  $N$ , it follows that the vectors  $\{\bar{\mathbf{k}}_n\}_{n=1}^N$  are all circularly shifted versions of each other. Moreover, since  $\bar{\mathbf{K}}$  is positive semidefinite, the largest entry of  $\bar{\mathbf{k}}_s$  is precisely the  $s$ -th one, which motivates interpreting  $\bar{\mathbf{k}}_s$  as an interpolating signal centered at  $s$ , which in turn suggests that the expression  $\hat{\mathbf{f}}_0 = \sum_{s \in \mathcal{S}} \bar{\alpha}_s \bar{\mathbf{k}}_s$  can be thought of as a reconstruction equation. From this vantage point, signals  $\{\bar{\mathbf{k}}_s\}_{s \in \mathcal{S}}$  play an analogous role to sinc functions in signal processing of time-domain signals. Examples of these interpolating signals are depicted in Fig. 1.

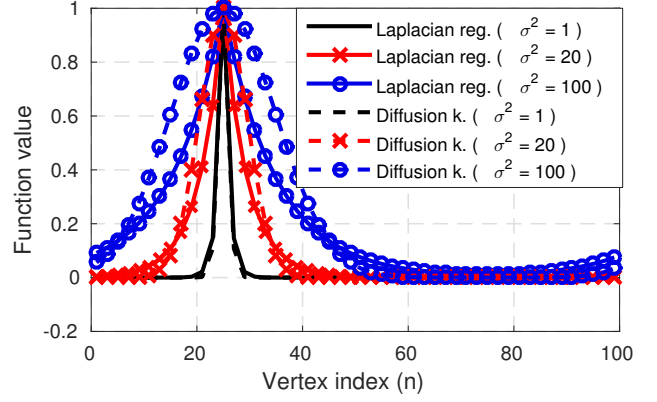


Fig. 1: 25-th column of  $\tilde{\mathbf{K}}$  for a circular graph with  $N = 100$  vertices. Different curves correspond to different parameter values for the Laplacian regularization and diffusion kernels.

2) *Bandlimited kernels:* A number of signal reconstruction approaches in the SPoG literature deal with graph bandlimited signals; see e.g. [11]–[18]. Here, the notion of *bandlimited kernel* is introduced to formally show that the LS estimator for bandlimited signals [11]–[16] is a limiting case of the kernel ridge regression estimate from (13). This notion will come handy in Secs. IV and V to estimate the bandwidth of a bandlimited signal from the observations  $\{y_s\}_{s=1}^S$ .

Signal  $f$  is said to be *bandlimited* if it admits an expansion (16) with  $\tilde{f}_n$  supported on a set  $\mathcal{B} \subset \{1, \dots, N\}$ ; that is,

$$\mathbf{f} = \sum_{n \in \mathcal{B}} \tilde{f}_n \mathbf{u}_n = \mathbf{U}_{\mathcal{B}} \tilde{\mathbf{f}}_{\mathcal{B}} \quad (19)$$

where  $\mathbf{U}_{\mathcal{B}}$  contains the columns of  $\mathbf{U}$  with indexes in  $\mathcal{B}$ , and  $\tilde{\mathbf{f}}_{\mathcal{B}}$  is a vector stacking  $\{\tilde{f}_n\}_{n \in \mathcal{B}}$ . The *bandwidth* of  $f$  can be defined as the cardinality  $B := |\mathcal{B}|$ , or, as the greatest element of  $\mathcal{B}$ .

If  $\mathbf{f}_0$  is bandlimited, it follows from (1) that  $\mathbf{y} = \Phi \mathbf{f}_0 + \mathbf{e} = \Phi \mathbf{U}_{\mathcal{B}} \tilde{\mathbf{f}}_{\mathcal{B}} + \mathbf{e}$  for some  $\tilde{\mathbf{f}}_{\mathcal{B}}$ . The LS estimate of  $\mathbf{f}_0$  is therefore given by [11]–[16]

$$\hat{\mathbf{f}}_{\text{LS}} = \mathbf{U}_{\mathcal{B}} \arg \min_{\tilde{\mathbf{f}}_{\mathcal{B}} \in \mathbb{R}^B} \|\mathbf{y} - \Phi \mathbf{U}_{\mathcal{B}} \tilde{\mathbf{f}}_{\mathcal{B}}\|^2 \quad (20a)$$

$$= \mathbf{U}_{\mathcal{B}} [\mathbf{U}_{\mathcal{B}}^T \Phi^T \Phi \mathbf{U}_{\mathcal{B}}]^{-1} \mathbf{U}_{\mathcal{B}}^T \Phi^T \mathbf{y} \quad (20b)$$

where the second equality assumes that  $\mathbf{U}_{\mathcal{B}}^T \Phi^T \Phi \mathbf{U}_{\mathcal{B}}$  is invertible, a necessary and sufficient condition for the  $B$  entries of  $\tilde{\mathbf{f}}_{\mathcal{B}}$  to be identifiable.

The estimate  $\hat{\mathbf{f}}_{\text{LS}}$  in (20) can be accommodated in the kernel regression framework by properly constructing a *bandlimited kernel*. Intuitively, one can adopt a Laplacian kernel for which  $r(\lambda_n)$  is large if  $n \notin \mathcal{B}$  (cf. Sec. III-B1). Consider the Laplacian kernel  $\bar{\mathbf{K}}_{\beta}$  with

$$r_{\beta}(\lambda_n) = \begin{cases} 1/\beta & n \in \mathcal{B} \\ \beta & n \notin \mathcal{B}. \end{cases} \quad (21)$$

For large  $\beta$ , this function strongly penalizes  $\{\tilde{f}_n\}_{n \notin \mathcal{B}}$  (cf. (15)), which promotes bandlimited estimates. The reason for setting  $r(\lambda_n) = 1/\beta$  for  $n \in \mathcal{B}$  instead of  $r(\lambda_n) = 0$  is to ensure that

$\bar{\mathbf{K}}_\beta$  is non-singular, a property that simplifies the statement and the proofs of some of the results in this paper.

**Proposition 1.** *Let  $\hat{\mathbf{f}}_{RR}$  denote the kernel ridge regression estimate from (13) with kernel  $\bar{\mathbf{K}}_\beta$  as in (21) and  $\mu > 0$ . If  $\mathbf{U}_B^T \Phi^T \Phi \mathbf{U}_B$  is invertible, as required by the estimator in (20b) for bandlimited signals, then  $\hat{\mathbf{f}}_{RR} \rightarrow \hat{\mathbf{f}}_{LS}$  as  $\beta \rightarrow \infty$ .*

*Proof:* See Appendix C. ■

Proposition 1 shows that the framework of kernel-based regression subsumes LS estimation of bandlimited signals. A non-asymptotic counterpart of Proposition 1 can be found by setting  $r_\beta(\lambda_n) = 0$  for  $n \in \mathcal{B}$  in (21), and noting that  $\hat{\mathbf{f}}_{RR} = \hat{\mathbf{f}}_{LS}$  if  $\mu = 0$ . Note however that imposing  $\mu = 0$  renders  $\hat{\mathbf{f}}_{RR}$  a degenerate kernel-based estimate.

3) *Covariance kernels:* So far, signal  $f_0$  has been assumed deterministic, which precludes accommodating certain forms of prior information that probabilistic models can capture, such as domain knowledge and historical data. A probabilistic interpretation of kernel methods on graphs will be pursued here to show that: (i) the optimal  $\bar{\mathbf{K}}$  in the MSE sense for ridge regression is the covariance matrix of  $\mathbf{f}_0$ ; and, (ii) kernel-based ridge regression seeks an estimate satisfying a system of local LMMSE estimation conditions on a Markov random field [33, Ch. 8].

Suppose without loss of generality that  $\{f_0(v_n)\}_{n=1}^N$  are zero-mean random variables. The LMMSE estimator of  $\mathbf{f}_0$  given  $\mathbf{y}$  is the linear estimator  $\hat{\mathbf{f}}_{LMMSE}$  minimizing  $\mathbb{E} \|\mathbf{f}_0 - \hat{\mathbf{f}}_{LMMSE}\|_2^2$ , where the expectation is over all  $\mathbf{f}_0$  and noise realizations. With  $\mathbf{C} := \mathbb{E}[\mathbf{f}_0 \mathbf{f}_0^T]$ , the LMMSE estimate is given by

$$\hat{\mathbf{f}}_{LMMSE} = \mathbf{C} \Phi^T [\Phi \mathbf{C} \Phi^T + \sigma_e^2 \mathbf{I}_S]^{-1} \mathbf{y} \quad (22)$$

where  $\sigma_e^2 := (1/S) \mathbb{E}[\|e\|_2^2]$  denotes the noise variance. Comparing (22) with (13) and recalling that  $\mathbf{K} := \Phi \bar{\mathbf{K}} \Phi^T$ , it follows that  $\hat{\mathbf{f}}_{LMMSE} = \hat{\mathbf{f}}_{RR}$  with  $\mu S = \sigma_e^2$  and  $\bar{\mathbf{K}} = \mathbf{C}$ . In other words, the similarity measure  $\kappa(v_n, v_{n'})$  embodied in the kernel map is just the covariance  $\text{cov}[f_0(v_n), f_0(v_{n'})]$ . A related observation was pointed out in [34] for general kernel methods.

In short, one can interpret kernel ridge regression as the LMMSE estimator of a signal  $\mathbf{f}_0$  with covariance matrix equal to  $\bar{\mathbf{K}}$ . This statement generalizes [13, Lemma 1], which requires  $\mathbf{f}_0$  to be Gaussian,  $\mathbf{C}$  rank-deficient, and  $\sigma_e^2 = 0$ .

Recognizing that kernel ridge regression is a linear estimator, readily establishes the following result.

**Proposition 2.** *If  $MSE(\bar{\mathbf{K}}, \mu) := \mathbb{E}[\|\mathbf{f}_0 - \hat{\mathbf{f}}_{RR}(\bar{\mathbf{K}}, \mu)\|^2]$ , where  $\hat{\mathbf{f}}_{RR}(\bar{\mathbf{K}}, \mu)$  denotes the estimator in (13), with kernel matrix  $\bar{\mathbf{K}}$ , and regularization parameter  $\mu$ , it then holds that*

$$MSE(\mathbf{C}, \sigma_e^2/S) \leq MSE(\bar{\mathbf{K}}, \mu)$$

for all kernel matrices  $\bar{\mathbf{K}}$  and  $\mu > 0$ .

Thus, for criteria aiming to minimize the MSE, Proposition 2 suggests  $\bar{\mathbf{K}}$  to be chosen *close* to  $\mathbf{C}$ . This observation may be employed for kernel selection and for parameter tuning in graph signal reconstruction methods of the kernel ridge regression family (e.g. the Tikhonov regularized estimators from [4], [12], [22] and [23, eq. (27)]), whenever an estimate of  $\mathbf{C}$  can

be obtained from historical data. For instance, the function  $r$  involved in Laplacian kernels can be chosen such that  $\bar{\mathbf{K}}$  resembles  $\mathbf{C}$  in some sense. Investigating such approaches goes beyond the scope of this paper.

A second implication of the connection between kernel ridge regression and LMMSE estimation involves signal estimation on Markov random fields [33, Ch. 8]. In this class of graphical models, an edge connects  $v_n$  with  $v_{n'}$  if  $f_0(v_n)$  and  $f_0(v_{n'})$  are *not* independent given  $\{f_0(v_{n''})\}_{n'' \neq n, n'}$ . Thus, if  $v_{n'} \notin \mathcal{N}_n$ , then  $f_0(v_n)$  and  $f_0(v_{n'})$  are independent given  $\{f_0(v_{n''})\}_{n'' \neq n, n'}$ . In other words, when  $f_0(v_{n''})$  is known for all neighbors  $v_{n''} \in \mathcal{N}_n$ , function values at non-neighboring vertices do not provide further information. This spatial Markovian property motivates the name of this class of graphical models. Real-world graphs obey this property when the topology captures direct interaction, in the sense that the interaction between the entities represented by two non-neighboring vertices  $v_n$  and  $v_{n'}$  is necessarily through vertices in a *path* connecting  $v_n$  with  $v_{n'}$ .

**Proposition 3.** *Let  $\mathcal{G}$  be a Markov random field, and consider the estimator in (13) with  $\bar{\mathbf{K}} = \mathbf{C} := \mathbb{E}[\mathbf{f}_0 \mathbf{f}_0^T]$ , and  $\mu = \sigma_e^2/S$ . Then, it holds that*

$$\hat{f}_{RR}(v_n) = \begin{cases} LMMSEE[f_0(v_n) | \{\hat{f}_{RR}(v)\}_{v \in \mathcal{N}_n}] & \text{if } n \notin \mathcal{S} \\ y_{s(n)} - \hat{e}_{s(n)} & \text{if } n \in \mathcal{S} \end{cases} \quad (23)$$

for  $n = 1, \dots, N$ , where  $s(n)$  denotes the sample index of the observed vertex  $v_n$ , i.e.,  $y_{s(n)} = f_0(v_n) + e_{s(n)}$ , and

$$\hat{e}_{s(n)} = \frac{\sigma_e^2}{\sigma_{n|\mathcal{N}_n}^2} \left[ \hat{f}_{RR}(v_n) - LMMSEE[f_0(v_n) | \{\hat{f}_{RR}(v)\}_{v \in \mathcal{N}_n}] \right].$$

Here,  $LMMSEE[f_0(v_n) | \{\hat{f}_{RR}(v)\}_{v \in \mathcal{N}_n}]$  is the LMMSE estimator of  $f_0(v_n)$  given  $f_0(v_{n'}) = \hat{f}_{RR}(v_{n'})$ ,  $v_{n'} \in \mathcal{N}_n$ , and  $\sigma_{n|\mathcal{N}_n}^2$  is the variance of this estimator.

*Proof:* See Appendix D. ■

If a (noisy) observation of  $f_0$  at  $v_n$  is not available, i.e.  $n \notin \mathcal{S}$ , then kernel ridge regression finds  $\hat{f}_{RR}(v_n)$  as the LMMSE estimate of  $f_0(v_n)$  given function values at the neighbors of  $v_n$ . However, since the latter are not directly observable, their ridge regression estimates are used instead. Conversely, when  $v_n$  is observed, implying that a sample  $y_{s(n)}$  is available, the sought estimator subtracts from this value an estimate  $\hat{e}_{s(n)}$  of the observation noise  $e_{s(n)}$ . Therefore, the kernel estimate on a Markov random field seeks an estimate satisfying the system of *local LMMSE conditions* given by (23) for  $n = 1, \dots, N$ .

**Remark 1.** In Proposition 3, the requirement that  $\mathcal{G}$  is a Markov random field can be relaxed to that of being a *conditional correlation graph*, defined as a graph where  $(v_n, v_{n'}) \in \mathcal{E}$  if  $f_0(v_n)$  and  $f_0(v_{n'})$  are correlated given  $\{f_0(v_{n''})\}_{n'' \neq n, n'}$ . Since correlation implies dependence, any Markov random field is also a conditional correlation graph. A conditional correlation graph can be constructed from  $\mathbf{C} := \mathbb{E}[\mathbf{f}_0 \mathbf{f}_0^T]$  by setting  $\mathcal{E} = \{(v_n, v_{n'}) : (\mathbf{C}^{-1})_{n, n'} \neq 0\}$  (see e.g. [35, Th. 10.2]).

**Remark 2.** Suppose that kernel ridge regression is adopted to estimate a function  $f_0$  on a certain graph  $\mathcal{G}$ , not necessarily a

Markov random field, using a kernel  $\bar{\mathbf{K}} \neq \mathbf{C} := \mathbb{E}[\mathbf{f}_0 \mathbf{f}_0^T]$ . Then it can still be interpreted as a method applying (23) on a conditional correlation graph  $\mathcal{G}'$  and adopting a signal covariance matrix  $\bar{\mathbf{K}}$ .

4) *Further kernels:* Additional signal reconstructors can be interpreted as kernel-based regression methods for certain choices of  $\bar{\mathbf{K}}$ . Specifically, it can be seen that [23, eq. (27)] is tantamount to kernel ridge regression with kernel

$$\bar{\mathbf{K}} = [(\mathbf{I}_N - \mathbf{W})^T (\mathbf{I}_N - \mathbf{W})]^{-1}$$

provided that the adjacency matrix  $\mathbf{W}$  is properly scaled so that this inverse exists. Another example is the Tikhonov regularized estimate in [12, eq. (15)], which is recovered as kernel ridge regression upon setting

$$\bar{\mathbf{K}} = [\mathbf{H}^T \mathbf{H} + \epsilon \mathbf{I}_N]^{-1}$$

and letting  $\epsilon > 0$  tend to 0, where  $\mathbf{H}$  can be viewed as a *high-pass filter* matrix. The role of the term  $\epsilon \mathbf{I}_N$  is to ensure that the matrix within brackets is invertible.

### C. Kernel-based smoothing and graph filtering

When an observation  $y_n$  is available per vertex  $v_n$  for  $n = 1, \dots, N$ , kernel methods can still be employed for denoising purposes. Due to the regularizer in (6), the estimate  $\hat{\mathbf{f}}_0$  will be a smoothed version of  $\mathbf{y}$ . This section shows how ridge regression smoothers can be thought of as graph filters, and vice versa. The importance of this two-way link is in establishing that kernel smoothers can be implemented in a decentralized fashion as graph filters [4].

Upon setting  $\Phi = \mathbf{I}_N$  in (13), one recovers the ridge regression smoother  $\hat{\mathbf{f}}_{\text{RRS}} = \bar{\mathbf{K}}(\bar{\mathbf{K}} + \mu N \mathbf{I}_N)^{-1} \mathbf{y}$ . If  $\bar{\mathbf{K}}$  is a Laplacian kernel, then

$$\hat{\mathbf{f}}_{\text{RRS}} = \mathbf{U} \tilde{r}(\Lambda) \mathbf{U}^T \mathbf{y} \quad (24)$$

where  $\tilde{r}(\lambda) := \mathcal{I}[r(\lambda) \neq 0] / [1 + \mu N r(\lambda)]$ .

To see how (24) relates to a graph filter, recall that the latter is an operator assigning  $\mathbf{y} \mapsto \mathbf{y}_F$ , where [4]

$$\mathbf{y}_F := \left( h_0 \mathbf{I}_N + \sum_{n=1}^{N-1} h_n \mathbf{L}^n \right) \mathbf{y} \quad (25a)$$

$$= \mathbf{U} \left( h_0 \mathbf{I}_N + \sum_{n=1}^{N-1} h_n \Lambda^n \right) \mathbf{U}^T \mathbf{y}. \quad (25b)$$

Graph filters can be implemented in a decentralized fashion since (25a) involves successive products of  $\mathbf{y}$  by  $\mathbf{L}$  and these products can be computed at each vertex by just exchanging information with neighboring vertices. Expression (25b) can be rewritten in the *Fourier domain* (cf. Sec. III-B1) as  $\tilde{\mathbf{y}}_F = [h_0 \mathbf{I}_N + \sum_{n=1}^{N-1} h_n \Lambda^n] \tilde{\mathbf{y}}$  upon defining  $\tilde{\mathbf{y}}_F := \mathbf{U}^T \mathbf{y}_F$  and  $\tilde{\mathbf{y}} := \mathbf{U}^T \mathbf{y}$ . For this reason, the diagonal of  $h_0 \mathbf{I}_N + \sum_{n=1}^{N-1} h_n \Lambda^n$  is referred to as the *frequency response* of the filter.

Comparing (24) with (25b) shows that  $\hat{\mathbf{f}}_{\text{RRS}}$  can be interpreted as a graph filter with frequency response  $\tilde{r}(\Lambda)$ . Thus, implementing  $\hat{\mathbf{f}}_{\text{RRS}}$  in a decentralized fashion using (25a) boils down to solving for  $\{h_n\}_{n=1}^N$  the system of linear equations

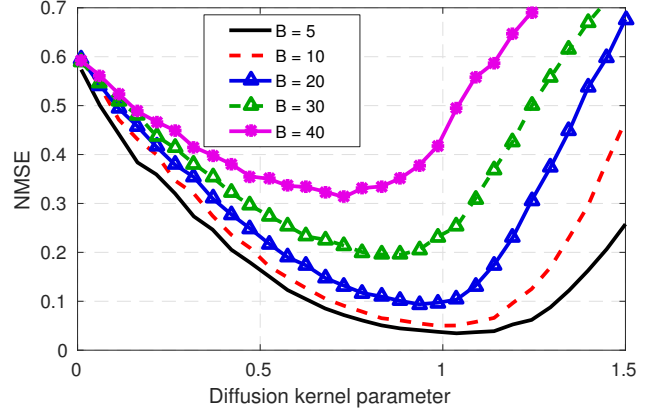


Fig. 2: Influence of the diffusion kernel parameter  $\sigma^2$  on NMSE for  $S = 40$  and several bandwidths  $B$  ( $N = 100$ ,  $\text{SNR} = 20$  dB,  $\mu = 10^{-4}$ ).

$\{h_0 + \sum_{n'=1}^{N-1} h_{n'} \lambda_n^{n'} = \tilde{r}(\lambda_n)\}_{n=1}^N$ . Conversely, given a filter, a Laplacian kernel can be found so that filter and smoother coincide. To this end, assume without loss of generality that  $\tilde{h}_n \leq 1 \forall n$ , where  $\tilde{h}_n := h_0 + \sum_{n'=1}^{N-1} h_{n'} \lambda_n^{n'}$ ; otherwise, simply scale  $\{h_n\}_{n=0}^{N-1}$ . Then, given  $\{\tilde{h}_n\}_{n=0}^{N-1}$ , the sought kernel can be constructed by setting

$$r(\lambda_n) = \frac{1 - \tilde{h}_n}{\mu N \tilde{h}_n} \mathcal{I}[\tilde{h}_n \neq 0].$$

### IV. MULTI-KERNEL GRAPH SIGNAL RECONSTRUCTION

One of the limitations of kernel methods is their sensitivity to the choice of the kernel. To appreciate this, Fig. 2 depicts the normalized mean-square error (NMSE)  $\mathbb{E} \|\mathbf{f}_0 - \hat{\mathbf{f}}_0\|_2^2 / \mathbb{E} \|\mathbf{f}_0\|_2^2$  when  $\mathcal{L}$  is the square loss and  $\Omega(\zeta) = |\zeta|$  across the parameter  $\sigma^2$  of the adopted diffusion kernel (see Sec. III-B1). The simulation setting is described in Sec. V. At this point though, it suffices to stress the impact of  $\sigma^2$  on the NMSE and the dependence of the optimum  $\sigma^2$  on the bandwidth  $B$  of  $\mathbf{f}_0$ .

Similarly, the performance of estimators for bandlimited signals degrades considerably if the estimator assumes a frequency support  $\mathcal{B}$  that differs from the actual one. Even for estimating *low-pass signals*, for which  $\mathcal{B} = \{1, \dots, B\}$ , parameter  $B$  is unknown in practice. Approaches for setting  $B$  were considered in [11], [16], but they rely solely on  $\mathcal{S}$  and  $\mathbf{L}$ , disregarding the observations  $\mathbf{y}$ . Note that by adopting the bandlimited kernels from Sec. III-B2, bandwidth selection boils down to kernel selection, so both problems will be treated jointly in the sequel through the lens of kernel-based learning.

This section advocates an MKL approach to kernel selection in graph signal reconstruction. Two algorithms with complementary strengths will be developed. Both select the most suitable kernels within a user-specified *kernel dictionary*.

#### A. RKHS superposition

Since  $\mathcal{H}$  in (6) is determined by  $\kappa$ , kernel selection is tantamount to RKHS selection. Therefore, a kernel dictionary

**Algorithm 1:** ADMM for multi-kernel regression

---

```

1: Input:  $\rho, \epsilon > 0, \check{\alpha}^{(0)}, \nu^0$ 
2: repeat
3:    $\check{\alpha}_m^{(k+1)} = \mathcal{T}_{\mu S/2\rho}(\beta_m^{(k)} + \nu_m^{(k)}) \quad m = 1, \dots, M$ 
4:    $\beta^{(k+1)} = (\Upsilon^T \Upsilon + \rho \mathbf{I})^{-1} [\Upsilon^T \mathbf{y} + \rho(\check{\alpha}^{(k+1)} - \nu^{(k)})]$ 
5:    $\nu_m^{(k+1)} = \nu_m^{(k)} + \beta_m^{(k+1)} - \check{\alpha}_m^{(k+1)}, \quad m = 1, \dots, M$ 
6:    $k \leftarrow k + 1$ 
7: until  $\|\beta^{(k+1)} - \check{\alpha}^{(k+1)}\| \leq \epsilon$ 

```

---

$\{\kappa_m\}_{m=1}^M$  can be equivalently thought of as an RKHS dictionary  $\{\mathcal{H}_m\}_{m=1}^M$ , which motivates estimates of the form

$$\hat{f} = \sum_{m=1}^M \hat{f}_m, \quad \hat{f}_m \in \mathcal{H}_m. \quad (26)$$

Upon adopting a criterion that controls sparsity in this expansion, the “best” RKHSs will be selected. A reasonable approach is therefore to generalize (6) to accommodate multiple RKHSs. With  $\mathcal{L}$  selected as the square loss and  $\Omega(\zeta) = |\zeta|$ , one can pursue an estimate  $\hat{f}$  by solving

$$\min_{\{f_m \in \mathcal{H}_m\}_{m=1}^M} \frac{1}{S} \sum_{s=1}^S \left[ y_s - \sum_{m=1}^M f_m(v_{n_s}) \right]^2 + \mu \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}. \quad (27)$$

Invoking Theorem 1 per  $f_m$  establishes that the minimizers of (27) can be written as

$$\hat{f}_m(v) = \sum_{s=1}^S \alpha_s^m \kappa_m(v, v_{n_s}), \quad m = 1, \dots, M \quad (28)$$

for some coefficients  $\alpha_s^m$ . Substituting (28) into (27) suggests obtaining these coefficients as

$$\arg \min_{\{\alpha_m\}_{m=1}^M} \frac{1}{S} \left\| \mathbf{y} - \sum_{m=1}^M \mathbf{K}_m \alpha_m \right\|^2 + \mu \sum_{m=1}^M (\alpha_m^T \mathbf{K}_m \alpha_m)^{1/2} \quad (29)$$

where  $\alpha_m := [\alpha_1^m, \dots, \alpha_S^m]^T$ , and  $\mathbf{K}_m = \Phi \bar{\mathbf{K}}_m \Phi^T$  with  $(\bar{\mathbf{K}}_m)_{n,n'} := \kappa_m(v_n, v_{n'})$ . Letting  $\check{\alpha}_m := \mathbf{K}_m^{1/2} \alpha_m$ , expression (29) becomes

$$\arg \min_{\{\check{\alpha}_m\}_{m=1}^M} \frac{1}{S} \left\| \mathbf{y} - \sum_{m=1}^M \mathbf{K}_m^{1/2} \check{\alpha}_m \right\|^2 + \mu \sum_{m=1}^M \|\check{\alpha}_m\|_2. \quad (30)$$

Note that the sum in the regularizer of (30) can be interpreted as the  $\ell_1$ -norm of  $[\|\check{\alpha}_1\|_2, \dots, \|\check{\alpha}_M\|_2]^T$ , which is known to promote sparsity in its entries and therefore in (26). Indeed, (30) can be seen as a particular instance of group Lasso [34].

As shown next, (30) can be efficiently solved using the alternating-direction method of multipliers (ADMM) [36]. To this end, rewrite (30) by defining  $\Upsilon := [\mathbf{K}_1^{1/2}, \dots, \mathbf{K}_M^{1/2}]$  and  $\check{\alpha} := [\check{\alpha}_1^T, \dots, \check{\alpha}_M^T]^T$ , and introducing the auxiliary variable  $\beta := [\beta_1^T, \dots, \beta_M^T]^T$ , as

$$\begin{aligned} \min_{\check{\alpha}, \beta} \quad & \frac{1}{2} \|\mathbf{y} - \Upsilon \beta\|^2 + \frac{S\mu}{2} \sum_{m=1}^M \|\check{\alpha}_m\|_2 \\ \text{s. to} \quad & \check{\alpha} - \beta = \mathbf{0}. \end{aligned} \quad (31)$$

**Algorithm 2:** Interpolated Iterative Algorithm

---

```

1: Input:  $\theta^{(0)}, \{\mathbf{K}_m\}_{m=1}^M, \mu, \theta_0, R, \eta, \epsilon$ .
2:  $\alpha^{(0)} = (\mathbf{K}(\theta^{(0)}) + \mu S \mathbf{I})^{-1} \mathbf{y}$ 
3:  $k = 0$ 
4: repeat
5:    $\xi^{(k)} = [\alpha^{(k),T} \mathbf{K}_0 \alpha^{(k)}, \dots, \alpha^{(k),T} \mathbf{K}_M \alpha^{(k)}]^T$ 
6:    $\theta^{(k)} = \theta_0 + (R/\|\xi^{(k)}\|_2) \xi^{(k)}$ 
7:    $\alpha^{(k+1)} = \eta \alpha^{(k)} + (1 - \eta)[\mathbf{K}(\theta^{(k)}) + \mu S \mathbf{I}]^{-1} \mathbf{y}$ 
8:    $k \leftarrow k + 1$ 
9: until  $\|\alpha^{(k+1)} - \alpha^{(k)}\| < \epsilon$ 

```

---

ADMM iteratively minimizes the *augmented Lagrangian* of (31) with respect to  $\check{\alpha}$  and  $\beta$  in a block-coordinate descent fashion, and updates the Lagrange multipliers associated with the equality constraint using gradient ascent (see [37] and references therein). The resulting iteration is summarized as Algorithm 1, where  $\rho$  is the augmented Lagrangian parameter,  $\nu := [\nu_1^T, \dots, \nu_M^T]^T$  is the Lagrange multiplier associated with the equality constraint, and

$$\mathcal{T}_\zeta(\mathbf{a}) := \frac{\max(0, \|\mathbf{a}\|_2 - \zeta)}{\|\mathbf{a}\|_2} \mathbf{a}$$

is the so-called *soft-thresholding* operator [36].

After obtaining  $\{\check{\alpha}_m\}_{m=1}^M$  from Algorithm 1, the wanted function estimate can be recovered as

$$\hat{f}_0 = \sum_{m=1}^M \bar{\mathbf{K}}_m \Phi^T \alpha_m = \sum_{m=1}^M \bar{\mathbf{K}}_m \Phi^T \mathbf{K}_m^{-1/2} \check{\alpha}_m. \quad (32)$$

It is recommended to normalize the kernel matrices in order to prevent imbalances in the kernel selection. Specifically, one can scale  $\{\bar{\mathbf{K}}_m\}_{m=1}^M$  such that  $\text{Tr}(\bar{\mathbf{K}}_m) = 1 \forall m$ . If  $\bar{\mathbf{K}}_m$  is a Laplacian kernel (see Sec. III-B1), where  $\bar{\mathbf{K}}_m = \mathbf{U} r_m^\dagger(\Lambda) \mathbf{U}^T$ , one can scale  $r_m$  to ensure  $\sum_{n=1}^N r_m^\dagger(\lambda_n) = 1$ .

**Remark 3.** Although criterion (27) is reminiscent of the MKL approach of [34], the latter differs markedly because it assumes that the right-hand side of (26) is uniquely determined given  $\hat{f}_0$ , which allows application of (6) over a direct-sum RKHS  $\mathcal{H} := \mathcal{H}_1 \oplus \dots \oplus \mathcal{H}_M$  with an appropriately defined norm. However, this approach cannot be pursued here since RKHSs of graph signals frequently overlap, implying that their sum is not a direct one (cf. discussion after (5)).

**B. Kernel superposition**

The MKL algorithm in Sec. IV-A can identify the best subset of RKHSs and therefore kernels, but entails  $MS$  unknowns (cf. (29)). This section introduces an alternative approach entailing only  $M + S$  variables at the price of not guaranteeing a sparse kernel expansion.

The approach is to postulate a kernel of the form  $\bar{\mathbf{K}}(\theta) = \sum_{m=1}^M \theta_m \bar{\mathbf{K}}_m$ , where  $\{\bar{\mathbf{K}}_m\}_{m=1}^M$  is given and  $\theta_m \geq 0 \forall m$ . The coefficients  $\theta := [\theta_1, \dots, \theta_M]^T$  can be found by jointly minimizing (11) with respect to  $\theta$  and  $\alpha$  [38]

$$(\theta, \hat{\alpha}) := \arg \min_{\theta, \alpha} \frac{1}{S} \mathcal{L}(\mathbf{y} - \mathbf{K}(\theta) \alpha) + \mu \Omega((\alpha^T \mathbf{K}(\theta) \alpha)^{1/2}) \quad (33)$$



where  $\mathbf{K}(\boldsymbol{\theta}) := \Phi \bar{\mathbf{K}}(\boldsymbol{\theta}) \Phi^T$ . Except for degenerate cases, problem (33) is not jointly convex in  $\boldsymbol{\theta}$  and  $\hat{\alpha}$ , but it is separately convex in these variables for a convex  $\mathcal{L}$  [38]. Criterion (33) generalizes the one in [39], which aims at combining Laplacian matrices of multiple graphs sharing the same vertex set.

A method termed *interpolated iterative algorithm* (IIA) was proposed in [40] to solve (33) when  $\mathcal{L}$  is the square loss,  $\Omega(\zeta) = \zeta^2$ , and  $\boldsymbol{\theta}$  is constrained to lie in a ball  $\Theta := \{\boldsymbol{\theta} : \boldsymbol{\theta} \geq \mathbf{0} \text{ and } \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq R\}$  for some user-defined center  $\boldsymbol{\theta}_0$  and radius  $R > 0$ . This constraint ensures that  $\boldsymbol{\theta}$  does not diverge. The first-order optimality conditions for (33) yield a nonlinear system of equations, which IIA solves iteratively. This algorithm is displayed as Algorithm 2, where  $\eta > 0$  is the step size.

As a special case, it is worth noting that Algorithm 2 enables kernel selection in ridge smoothing, which is tantamount to optimal filter selection for graph signal denoising (cf. Sec. III-C). In this case, Algorithm 2 enjoys a particularly efficient implementation for Laplacian kernels since their kernel matrices share eigenvectors. Specifically, recalling that for smoothing  $\mathbf{K}_m = \bar{\mathbf{K}}_m = \mathbf{U} r_m^\dagger(\boldsymbol{\Lambda}) \mathbf{U}^T$  and letting  $\tilde{\boldsymbol{\alpha}} := [\tilde{\alpha}_1, \dots, \tilde{\alpha}_N]^T := \mathbf{U}^T \boldsymbol{\alpha}$ , suggests that the  $\boldsymbol{\alpha}$ -update in Algorithm 2 can be replaced with its scalar version

$$\tilde{\alpha}_n^{(k+1)} = \eta \tilde{\alpha}_n^{(k)} + \frac{(1 - \eta) \tilde{y}_n}{\sum_{m=1}^M \theta_m^{(k)} r_m^\dagger(\lambda_n) + \mu S}, \quad n = 1, \dots, N$$

whereas the  $\boldsymbol{\xi}$ -update can be replaced with  $\xi_m^{(k)} = \sum_{n=1}^N r_m^\dagger(\lambda_n) (\tilde{\alpha}_n^{(k)})^2$ , where  $\boldsymbol{\xi} := [\xi_1, \dots, \xi_M]^T$ .

## V. NUMERICAL TESTS

This section compares the proposed methods with competing alternatives in synthetic- as well as real-data experiments. Monte Carlo simulation is used to average performance metrics across realizations of the signal  $\mathbf{f}_0$ , noise  $\mathbf{e}$  (only for synthetic-data experiments), and sampling set  $\mathcal{S}$ . The latter is drawn uniformly at random without replacement from  $\{1, \dots, N\}$ .

### A. Synthetic bandlimited signals

Three experiments were conducted on an Erdős-Rényi random graph with probability of edge presence 0.25 [1]. Bandlimited signals were generated as in (19) with  $\mathcal{B} = \{1, \dots, B\}$  for a certain  $B$ . The coefficients  $\{\tilde{f}_n\}_{n \in \mathcal{B}}$  are independent uniformly distributed over the interval  $[0, 1]$ . Gaussian noise was added to yield a target *signal-to-noise* ratio  $\text{SNR} := \|\mathbf{f}_0\|^2 / (N \sigma_e^2)$ .

The first experiment was presented in Fig. 2 and briefly described in Sec. IV to illustrate the strong impact of the kernel choice on the NMSE  $:= \mathbb{E} \|\mathbf{f}_0 - \hat{\mathbf{f}}_0\|_2^2 / \mathbb{E} \|\mathbf{f}_0\|_2^2$ .

The second experiment compares methods for estimating bandlimited signals. Fig. 3 depicts the NMSE in reconstructing a bandlimited signal with  $B = 20$  across  $S$ . The first two curves correspond to the MKL approaches proposed in Sec. IV, which employ a dictionary with 5 bandlimited kernels, where the  $m$ -th kernel has  $\beta = 10^4$  and bandwidth  $5m + 5$ ,  $m = 1, \dots, 5$ . The regularization parameter  $\mu$  was set to  $10^{-1}$  for RKHS superposition (RS), and to  $5 \cdot 10^{-3}$  for kernel superposition (KS). The next three curves correspond to the LS estimator for bandlimited (BL) signals in (20b) [11]–[16]. In order to illustrate

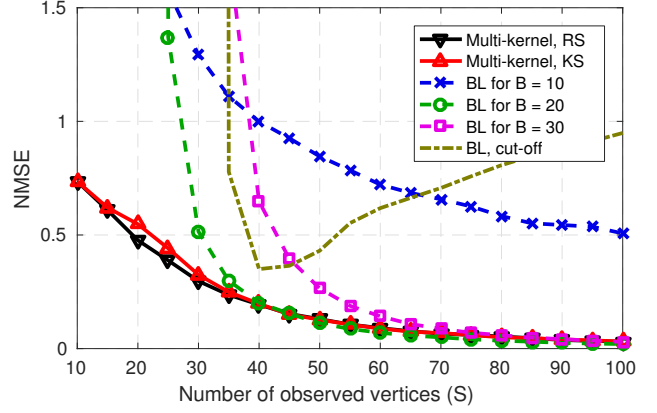


Fig. 3: Comparison of different algorithms for estimating bandlimited signals. Per Monte Carlo iteration, a bandlimited signal with  $B = 20$  is generated ( $N = 100$ ,  $\text{SNR} = 10$  dB).

the effects of the uncertainty in  $B$ , each curve corresponds to a different value of  $B$  used for *estimation* (all estimators observe the same synthetic signal of bandwidth  $B = 20$ ). The last curve pertains to the estimator in [11], [16], which is the LS estimator in (20b) with parameter  $B$  set to the *cut-off frequency* obtained from  $\mathbf{L}$  and  $\mathcal{S}$  by means of a *proxy* of order 5.

Observe in Fig. 3 that although the proposed MKL estimators do not know the bandwidth, their performance is no worse than that of the BL estimator with perfect knowledge of the signal bandwidth. Remarkably, the MKL reconstruction schemes offer a reasonable performance for  $S$  small, thus overcoming the need of the LS estimator for  $S \geq B$  samples.

The third experiment illustrates how the bandwidth of a graph signal can be estimated using the MKL scheme from Sec. IV-A. To this end, a dictionary of 17 bandlimited kernels was constructed with  $\beta = 10^3$  and uniformly spaced bandwidth between 10 and 90, i.e.,  $\bar{\mathbf{K}}_m$  is of bandwidth  $B_m := 5m + 5$ ,  $m = 1, \dots, 17$ . Fig. 4 depicts the *sparsity path* for a typical realization of a bandlimited signal with bandwidth  $B = 20$ . Each curve is obtained by executing Algorithm 1 for different values of  $\mu$  and represents the squared modulus of the vectors  $\{\boldsymbol{\alpha}_m\}_{m=1}^M$  in (32) for a different  $m$ . As expected, the sparsity effected in the expansion (26) increases with  $\mu$ , forcing Algorithm 1 to eventually rely on a single kernel. That kernel is expected to be the one leading to best data fit. Since the observed signal is bandlimited, such a kernel is in turn expected to be the one in the dictionary whose bandwidth is closest to  $B$ .

Constructing a rule that determines, without human intervention, which is the last curve  $\|\boldsymbol{\alpha}_m\|^2$  to vanish is not straightforward since it involves comparing  $\{\|\boldsymbol{\alpha}_m\|^2\}_{m=1}^M$  for a properly selected  $\mu$ . Thus, algorithms pursuing such objective fall out of the scope of this paper. However, one can consider the *naïve* approach that focuses on a prespecified value of  $\mu$  and estimates the bandwidth as  $\hat{B} = B_{m^*}$ , where  $m^* = \arg \max_{m \in \{1, \dots, M\}} \|\boldsymbol{\alpha}_m\|^2$ . Table I reports the performance of such estimator in terms of bias  $\mathbb{E}|B - \hat{B}|$  and standard deviation  $\sqrt{\mathbb{E}|B - \hat{B}|}$  for different values of  $B$  for a synthetically generated bandlimited signal.

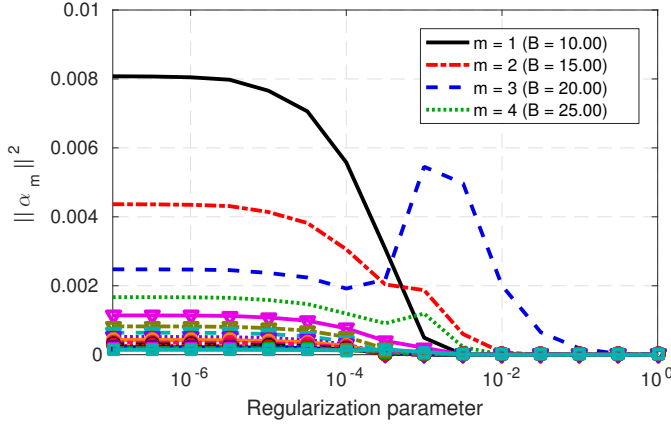


Fig. 4: Sparsity path of the estimate for a typical realization. The legend only displays the first four curves. The last curve to vanish indicates the bandwidth of the observed signal ( $S = 80$ ,  $N = 250$ ,  $\text{SNR} = 20$  dB).

	B = 10	B = 20	B = 30	B = 40	B = 50	B = 60
BIAS	0.0	0.6	0.5	0.4	0.4	3.6
STD	0.0	1.9	2.9	1.4	1.4	10.5

TABLE I: Bias and standard deviation for the *naïve* bandwidth estimator with  $\mu = 10^{-2}$  ( $S = 80$ ,  $N = 250$ ,  $\text{SNR} = 20$  dB).

### B. Real data

This section assesses the performance of the proposed methods with two real-data sets. In both experiments, the data set is split into a training set used to learn the edge weights, and a test set from which the observations  $\mathbf{y}$  are drawn for performance evaluation. Different from the synthetic-data experiments in Sec. V-A, where the generated noiseless function  $f_0$  is available and therefore the reconstruction NMSE can be measured on observed and unobserved vertices, the experiments in this section measure *generalization* NMSE solely at unobserved vertices.

The first data set comprises 24 signals corresponding to the average temperature per month in the intervals 1961-1990 and 1981-2010 measured by 89 stations in Switzerland [42]. The training set contains the first 12 signals, which correspond to the interval 1961-1990, whereas the test set contains the remaining 12. Each station is identified with a vertex and the graph is constructed by applying the algorithm in [43] with parameters  $\alpha = 1$  and  $\beta = 30$  to the training signals. Based on samples of a test signal on  $S$  vertices, the goal is to estimate the signal at the remaining  $N - S$  vertices. NMSE is averaged across the 12 test signals for a randomly chosen set  $\mathcal{S}$ . Fig. 5 compares the performance of the MKL schemes from Sec. IV along with single-kernel ridge regression (KRR) and estimators for bandlimited signals. The MKL algorithms employ a dictionary comprising 10 diffusion kernels with parameter  $\sigma^2$  uniformly spaced between 1 and 20. Single-kernel ridge regression uses diffusion kernels for different values of  $\sigma^2$ . Fig. 5 showcases the performance improvement arising from adopting the proposed multi-kernel approaches.

The second data set contains departure and arrival information for flights among U.S. airports [41], from which the  $3 \cdot 10^6$

flights in the months of July, August, and September of 2014 and 2015 were selected. A graph was constructed with vertices representing the  $N = 50$  airports with highest traffic. An edge was placed between a pair of vertices if the number of flights between the associated airports exceeds 100 within the observation window. A signal was constructed per day averaging the arrival delay of *all* inbound flights per selected airport. Thus, a total of 184 signals were considered, the first 154 were used for training (July, August, September 2014, and July, August 2015), and the remaining 30 for testing (September 2015).

Since it is reasonable to assume that the aforementioned graph approximately satisfies the Markovian property (cf. Sec. III-B3), a Markov random field was fit to the observations. To this end, the signals were assumed Gaussian so as to estimate the covariance matrix of the observations via maximum likelihood with constraints imposing the  $(n, n')$ -th entry of the inverse covariance matrix to be zero if  $(v_n, v_{n'}) \notin \mathcal{E}$ . Specifically,  $\mathbf{S} := \mathbf{C}^{-1}$  was found by solving the following convex program:

$$\begin{aligned} \min_{\mathbf{S} \in \mathbb{R}^{N \times N}} \quad & \text{Tr}(\mathbf{S} \hat{\mathbf{C}}^{-1}) - \log \det(\mathbf{S}) \\ \text{s.to} \quad & \mathbf{S} \succeq \mathbf{0}, (\mathbf{S})_{n,n'} = 0 \quad \forall (v_n, v_{n'}) \notin \mathcal{E} \end{aligned} \quad (34)$$

where  $\hat{\mathbf{C}}$  is the sample covariance matrix of the training signals after normalization to effect zero mean and unit variance per entry of  $\mathbf{f}_0$ . The inverse of  $\mathbf{S}$  was used as a covariance kernel (see Sec. III-B3). Note that such a kernel will only be nearly optimal since the *true* data covariance is unknown.

Employing Laplacian kernels or applying estimators for bandlimited signals requires a Laplacian matrix. Although the edge set  $\mathcal{E}$  has already been constructed, it is necessary to endow those edges with weights. Since our efforts to obtain a reasonable estimation performance over the graphs provided by the method in [43] turned out unsuccessful, a novel approach was developed. Specifically, the Laplacian matrix is sought as the minimizer of  $\|\mathbf{L} - \mathbf{S}\|_F^2$ , where  $\mathbf{S}$  is the solution to (34) and  $\mathbf{L}$  is a valid Laplacian with a zero at the  $(n, n')$ -th position if  $(v_n, v_{n'}) \notin \mathcal{E}$ . Due to space limitations, the rationale and details behind this approach are skipped.

Table II lists the NMSE and root mean-square error in minutes for the task of predicting the arrival delay at 40 airports when the delay at a randomly selected collection of 10 airports is observed. The second column corresponds to the ridge regression estimator that uses the nearly-optimal *estimated* covariance kernel. The next two columns correspond to the multi-kernel approaches in Sec. IV with a dictionary of 30 diffusion kernels with values of  $\sigma^2$  uniformly spaced between 0.1 and 7. The rest of columns pertain to estimators for bandlimited signals. Table II demonstrates the good performance of covariance kernels as well as the proposed multi-kernel approaches relative to competing alternatives.

## VI. CONCLUSIONS

This paper introduced kernel-based learning as a unifying framework subsuming a number of existing signal estimators. SPOG notions such as bandlimitedness, graph filters, and the graph Fourier transform were accommodated under this perspective. The notion of bandlimited kernels was invoked to establish

	KRR with cov. kernel	Multi-kernel, RS	Multi-kernel, KS	BL for $B = 2$	BL for $B = 3$	BL, cut-off
NMSE	0.34	0.44	0.43	1.55	32.64	3.97
RMSE [min]	3.95	4.51	4.45	8.45	38.72	13.50

TABLE II: Generalization NMSE and root mean square error for the experiment with the airport data set [41].

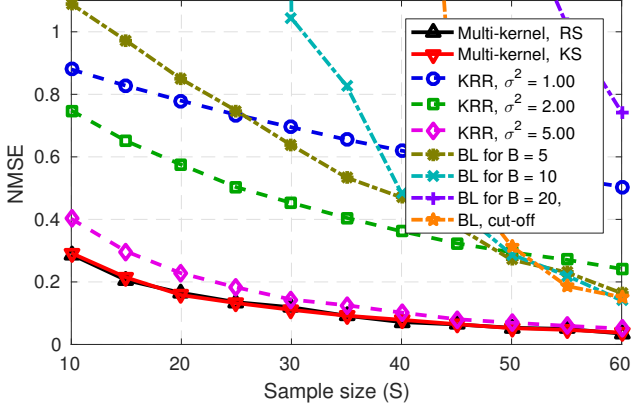


Fig. 5: Generalization NMSE for the data set in [42].

that LS estimators are limiting versions of the ridge regression estimator with Laplacian kernels. Optimality of covariance kernels was also revealed and a novel interpretation of kernel regression on graphs was presented in terms of Markov random fields. Graph filters are tantamount to kernel-based smoothers, which suggested applying the former to implement the latter in a decentralized fashion. Finally, numerical experiments corroborated the validity of the theoretical findings.

Future research will pursue algorithms for learning graph Laplacian matrices tailored for regression, broadening regression to directed graphs, and numerical experiments with further data sets.

#### APPENDIX A

##### PROOF OF THE REPRESENTER THEOREM

Theorem 1 can be proved upon decomposing  $\mathbf{f}$  according to the following result.

**Lemma 1.** *If  $\Phi$  is as in Sec. III and  $f$  belongs to  $\mathcal{H}$ , then  $\mathbf{f} := [f(v_1), \dots, f(v_N)]^T$  can be expressed as*

$$\mathbf{f} = \bar{\mathbf{K}}\Phi^T\alpha + \bar{\mathbf{K}}\beta \quad (35)$$

for some  $\alpha \in \mathbb{R}^S$  and  $\beta \in \mathbb{R}^N$  satisfying  $\Phi\bar{\mathbf{K}}\beta = \mathbf{0}$ .

*Proof:* Since  $f \in \mathcal{H}$ , there exists  $\bar{\alpha}$  such that  $\mathbf{f} = \bar{\mathbf{K}}\bar{\alpha}$ . Thus, one needs to show that, for a given  $\bar{\alpha}$ , it is possible to choose  $\alpha$  and  $\beta$  satisfying  $\bar{\mathbf{K}}\bar{\alpha} = \bar{\mathbf{K}}\Phi^T\alpha + \bar{\mathbf{K}}\beta$  and  $\Phi\bar{\mathbf{K}}\beta = \mathbf{0}$ . This is possible, for instance, if one fixes  $\beta = \bar{\alpha} - \Phi^T\alpha$ , and shows that there exists an  $\alpha$  such that  $\Phi\bar{\mathbf{K}}\beta = \Phi\bar{\mathbf{K}}(\bar{\alpha} - \Phi^T\alpha) = \mathbf{0}$ . This, in turn, follows if one establishes that  $\Phi\bar{\mathbf{K}}\bar{\alpha} = \Phi\bar{\mathbf{K}}\Phi^T\alpha$  always admits a solution in  $\alpha$ , which holds since  $\mathcal{R}\{\Phi\bar{\mathbf{K}}\} = \mathcal{R}\{\Phi\bar{\mathbf{K}}\Phi^T\}$ . To see this, consider the eigendecomposition  $\bar{\mathbf{K}} = \mathbf{U}_K\Lambda_K\mathbf{U}_K^T$  and note that

$$\mathcal{R}\{\Phi\bar{\mathbf{K}}\Phi^T\} = \mathcal{R}\{\Phi\mathbf{U}_K\Lambda_K\mathbf{U}_K^T\Phi^T\} = \mathcal{R}\{\Phi\mathbf{U}_K\Lambda_K^{1/2}\}$$

$$= \mathcal{R}\{\Phi\mathbf{U}_K\Lambda_K\mathbf{U}_K^T\} = \mathcal{R}\{\Phi\bar{\mathbf{K}}\}$$

which concludes the proof.  $\blacksquare$

Lemma 1 essentially states that for arbitrary  $\mathcal{S}$ , any  $f \in \mathcal{H}$  can be decomposed into two components as  $f = f_{\mathcal{S}} + f_{\perp}$ . The first can be expanded in terms of the vertices indexed by  $\mathcal{S}$  as  $f_{\mathcal{S}}(v) = \sum_{s=1}^S \alpha_s \kappa(v, v_{n_s})$ , whereas the second vanishes in the sampling set, i.e.,  $f_{\perp}(v_s) = 0 \forall s \in \mathcal{S}$ . Conversely, it is clear that any function that can be written as in (35) for arbitrary  $\alpha$  and  $\beta$  belonging to  $\mathcal{H}$ . Hence, Lemma 1 offers an alternative parameterization of  $\mathcal{H}$  in terms of  $\alpha$  and  $\beta$ . Thus, the minimizer  $\hat{\alpha}$  of (7) can be obtained as  $\hat{\alpha} = \Phi^T\hat{\alpha} + \hat{\beta}$ , where

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) := \arg \min_{\alpha, \beta} & \frac{1}{S} \mathcal{L}(\mathbf{y} - \Phi\bar{\mathbf{K}}(\Phi^T\alpha + \beta)) \\ & + \mu\Omega \left( [(\Phi^T\alpha + \beta)^T \bar{\mathbf{K}}(\Phi^T\alpha + \beta)]^{1/2} \right) \end{aligned} \quad (36)$$

s. to  $\Phi\bar{\mathbf{K}}\beta = \mathbf{0}$ .

Since  $\mathcal{L}(\mathbf{y} - \Phi\bar{\mathbf{K}}(\Phi^T\alpha + \beta)) = \mathcal{L}(\mathbf{y} - \Phi\bar{\mathbf{K}}\Phi^T\alpha)$ , the first term in the objective does not depend on  $\beta$ . On the other hand, since  $\Omega$  is increasing and

$$(\Phi^T\alpha + \beta)^T \bar{\mathbf{K}}(\Phi^T\alpha + \beta) = \alpha^T \Phi\bar{\mathbf{K}}\Phi^T\alpha + \beta^T \bar{\mathbf{K}}\beta$$

it follows that the objective of (36) is minimized for  $\beta = \mathbf{0}$ , which shows that  $\hat{f}_0$  in (6) can be written as  $\hat{f}_0 = \bar{\mathbf{K}}\hat{\alpha} = \bar{\mathbf{K}}\Phi^T\hat{\alpha}$ , thus completing the proof.

#### APPENDIX B

##### BIG DATA SCENARIOS

Evaluating the  $N \times N$  Laplacian kernel matrix in (14)) incurs complexity  $\mathcal{O}(N^3)$ , which does not scale well with  $N$ . This appendix explores two means of reducing this complexity. Both rely on solving (8) rather than (11) since the former employs  $\bar{\mathbf{K}}^\dagger = \mathbf{U}_r(\Lambda)\mathbf{U}^T$ , whereas the latter needs  $\bar{\mathbf{K}}$ .

Recall from Sec. III-B1 that Laplacian kernels control the smoothness of an estimate by regularizing its Fourier coefficients  $|\tilde{f}_n|$  via  $r$ . Computational savings can be effected if one is willing to finely tune the regularization only for large  $n$ , while allowing a coarse control for small  $n$ . Specifically, the key idea here is to adopt a function of the form

$$r(\lambda_n) = \begin{cases} d\lambda_n & \text{if } 1 < n \leq B \\ d_n & \text{if } n > B \text{ or } n = 1 \end{cases} \quad (37)$$

where  $d$  and  $d_n$  are constants freely selected over the ranges  $d, d_1 > 0$  and  $d_n > -\lambda_n$  for  $n > B$ . Note that (37) can be employed, in particular, to promote bandlimited estimates of bandwidth  $B$  by setting  $\{d_n\}_{n=B+1}^N$  sufficiently large. Defining  $\mathbf{U}_B^c$  as the matrix whose columns are the  $N - B$  principal eigenvectors of  $\mathbf{L}$ , one obtains

$$\bar{\mathbf{K}}^{-1} = d\mathbf{L} + \mathbf{U}_B^c(\Delta - d\Lambda_B^c)\mathbf{U}_B^{cT} + d_1\mathbf{1}\mathbf{1}^T + \epsilon\mathbf{I}_N \quad (38)$$

where  $\Delta := \text{diag}\{d_{B+1}, \dots, d_N\}$  and  $\epsilon \mathbf{I}_N$  with  $\epsilon > 0$  is added to ensure that  $\bar{\mathbf{K}}$  is invertible in case that the multiplicity of the zero eigenvalue of  $\mathbf{L}$  is greater than one, which occurs when the graph has multiple connected components.

Alternative functions that do not require eigenvector computation are low-order polynomials of the form

$$r(\lambda) = \sum_{p=0}^P a_p \lambda^p. \quad (39)$$

In this case, the resulting  $\bar{\mathbf{K}}^{-1}$  reads as

$$\bar{\mathbf{K}}^{-1} = a_0 \mathbf{I}_N + \sum_{p=1}^P a_p \mathbf{L}^p.$$

The cost of obtaining this matrix is reduced since powers of  $\mathbf{L}$  can be efficiently computed when  $\mathbf{L}$  is sparse, as is typically the case. In the extreme case where  $P = 1$ ,  $a_1 > 0$ , and  $a_0 \rightarrow 0$ , the regularizer becomes  $\mathbf{f}^T \mathbf{L} \mathbf{f}$ , which corresponds to the Laplacian regularization (cf. Sec. III-B1).

#### APPENDIX C

##### PROOF OF PROPOSITION 1

Without loss of generality, let  $\mathcal{B} = \{1, \dots, B\}$ ; otherwise, simply permute the order of the eigenvalues. Define also the  $N \times B$  matrix  $\Psi = [\mathbf{I}_B, \mathbf{0}]^T$  and the  $N \times (N-B)$  matrix  $\Psi_c = [\mathbf{0}, \mathbf{I}_{N-B}]^T$ , whose concatenation clearly gives  $[\Psi, \Psi_c] = \mathbf{I}_N$ . Since in this case  $\mathbf{U}_B = \mathbf{U}\Psi$ , (20b) becomes

$$\hat{\mathbf{f}}_{\text{LS}} = \mathbf{U}\Psi[\Psi^T \mathbf{U}^T \Phi^T \Phi \mathbf{U}\Psi]^{-1} \Psi^T \mathbf{U}^T \Phi^T \mathbf{y}. \quad (40)$$

On the other hand, the ridge regression version of (8) is

$$\hat{\mathbf{f}}_0 := \arg \min_{\mathbf{f}} \frac{1}{S} \|\mathbf{y} - \Phi \mathbf{f}\|^2 + \mu \mathbf{f}^T \bar{\mathbf{K}}^{-1} \mathbf{f} \quad (41)$$

where the constraint has been omitted since  $r_\beta(\lambda_n) > 0 \forall n$ . The minimizer of (41) is

$$\hat{\mathbf{f}}_0 = (\Phi^T \Phi + \mu S \bar{\mathbf{K}}^{-1})^{-1} \Phi^T \mathbf{y} \quad (42a)$$

$$= \mathbf{U}(\mathbf{U}^T \Phi^T \Phi \mathbf{U} + \mu S r_\beta(\Lambda))^{-1} \mathbf{U}^T \Phi^T \mathbf{y}. \quad (42b)$$

Establishing that  $\hat{\mathbf{f}}_0 \rightarrow \hat{\mathbf{f}}_{\text{LS}}$  therefore amounts to showing that the right-hand side of (40) converges to that of (42b). For this, it suffices to prove that

$$(\mathbf{G} + \mu S r_\beta(\Lambda))^{-1} \rightarrow \Psi[\Psi^T \mathbf{G} \Psi]^{-1} \Psi^T \quad (43)$$

where  $\mathbf{G} := \mathbf{U}^T \Phi^T \Phi \mathbf{U}$ . Note that  $\Psi^T \mathbf{G} \Psi = \mathbf{U}_B^T \Phi^T \Phi \mathbf{U}_B$  is invertible by hypothesis. With  $\Lambda_B := (1/\beta) \mathbf{I}_B$  and  $\Lambda_B^c := \beta \mathbf{I}_{N-B}$  representing the *in-band* and *out-of-band* parts of  $r_\beta(\Lambda)$ , the latter can be written as  $r_\beta(\Lambda) = \text{diag}\{\Lambda_B, \Lambda_B^c\}$ . With this notation, (43) becomes

$$\left( \begin{bmatrix} \Psi^T \mathbf{G} \Psi & \Psi^T \mathbf{G} \Psi_c \\ \Psi_c^T \mathbf{G} \Psi & \Psi_c^T \mathbf{G} \Psi_c \end{bmatrix} + \mu S \begin{bmatrix} \Lambda_B & \mathbf{0} \\ \mathbf{0} & \Lambda_B^c \end{bmatrix} \right)^{-1} \rightarrow \begin{bmatrix} (\Psi^T \mathbf{G} \Psi)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \quad (44)$$

Using block matrix inversion formulae, it readily follows that the left-hand side equals the following matrix product

$$\begin{bmatrix} \mathbf{I}_B & -(\Psi^T \mathbf{G} \Psi + \mu S \Lambda_B)^{-1} \Psi^T \mathbf{G} \Psi_c \\ \mathbf{0} & \mathbf{I}_{N-B} \end{bmatrix} \begin{bmatrix} (\Psi^T \mathbf{G} \Psi + \mu S \Lambda_B)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_B & \mathbf{0} \\ -\Psi_c^T \mathbf{G} \Psi (\Psi^T \mathbf{G} \Psi + \mu S \Lambda_B)^{-1} & \mathbf{I}_{N-B} \end{bmatrix} \quad (45)$$

where

$$\mathbf{M} := \Psi_c^T \mathbf{G} \Psi_c + \mu S \Lambda_B^c - \Psi_c^T \mathbf{G} \Psi (\Psi^T \mathbf{G} \Psi + \mu S \Lambda_B)^{-1} \Psi^T \mathbf{G} \Psi_c.$$

Recalling that  $\Psi^T \mathbf{G} \Psi$  is invertible and letting  $\beta \rightarrow \infty$ , it follows that  $\mathbf{M}^{-1} \rightarrow \mathbf{0}$  and  $\Lambda_B \rightarrow \mathbf{0}$  as  $\beta \rightarrow \infty$ , which implies that (45) converges to the right-hand side of (44) and concludes the proof.

#### APPENDIX D

##### PROOF OF PROPOSITION 3

The first-order optimality condition for (12a) is given by

$$\sigma_e^2 \mathbf{C}^{-1} \mathbf{f} = \Phi^T (\mathbf{y} - \Phi \mathbf{f}). \quad (46)$$

Without loss of generality, one can focus on the relation implied by the first row of (46). To this end, partition  $\mathbf{C}$  as

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & \mathbf{c}_{2:N,1}^T \\ \mathbf{c}_{2:N,1} & \mathbf{C}_{2:N,2:N} \end{bmatrix}$$

and apply block matrix inversion formulae to obtain

$$\mathbf{C}^{-1} = \frac{1}{c_{1|2:N}} \begin{bmatrix} 1 & -\mathbf{c}_{2:N,1}^T \mathbf{C}_{2:N,2:N}^{-1} \\ -\mathbf{C}_{2:N,2:N}^{-1} \mathbf{c}_{2:N,1} & \tilde{\mathbf{C}}^{-1} \end{bmatrix}$$

where  $c_{1|2:N} := c_{1,1} - \mathbf{c}_{2:N,1}^T \mathbf{C}_{2:N,2:N}^{-1} \mathbf{c}_{2:N,1}$  and

$$\tilde{\mathbf{C}}^{-1} := \mathbf{C}_{2:N,2:N}^{-1} - \mathbf{c}_{2:N,1}^T \mathbf{C}_{2:N,2:N}^{-1} \mathbf{c}_{2:N,1} + c_{1|2:N} \mathbf{C}_{2:N,2:N}^{-1}.$$

Note that  $\sigma_{n|\mathcal{N}_n}^2 := c_{1|2:N}$  is in fact the variance of the LMMSE predictor for  $f_0(v_1)$  given  $f_0(v_2), \dots, f_0(v_N)$ .

Two cases can be considered for the first row of (46). First, if  $1 \notin \mathcal{S}$ , then the first row of  $\Phi$  is zero, and the first row of (46) becomes

$$f(v_1) = \mathbf{c}_{2:N,1}^T \mathbf{C}_{2:N,2:N}^{-1} \mathbf{f}_{2:N} \quad (47a)$$

$$= \sum_{n:v_n \in \mathcal{N}_1} (-c_{1|2:N} \gamma_{1,n}) f(v_n) \quad (47b)$$

where  $\mathbf{f}_{2:N} := [f(v_2), \dots, f(v_N)]^T$  and  $\gamma_{n,n'} = (\mathbf{C}^{-1})_{n,n'}$ . The sum in (47b) involves only the neighbors of  $v_1$  since the graph is a Markov random field, for which if there is no edge between  $v_n$  and  $v_{n'}$ , then  $f_0(v_n)$  and  $f_0(v_{n'})$  are conditionally independent given the rest of vertices, which in turn implies that  $\gamma_{n,n'} = 0$ . Note that the right-hand side of (47a) is the LMMSE predictor of  $f(v_1)$  given the estimated function value at its neighbors. Since this argument applies to all vertices  $v_n$ ,  $n \notin \mathcal{S}$ , it follows that the optimality condition (46) seeks values of  $f(v_n)$  so that the function value at unobserved vertices



agrees with its LMMSE estimate given the estimated value at its neighbors.

On the other hand, if  $1 \in \mathcal{S}$ , the first row of  $\Phi$  has a 1 at the  $(1, 1)$  position, which implies that the first row of (46) is

$$y_1 = f(v_1) + \frac{\sigma_e^2}{c_{1|2:N}}(f(v_1) - \mathbf{c}_{2:N,1}^T \mathbf{C}_{2:N,2:N}^{-1} \mathbf{f}_{2:N}).$$

The second term on the right can be thought of as an estimate of the noise  $e_1$  present in  $y_1$ . Therefore, the optimality condition imposes that each observation  $y_{s(n)}$  agrees with the estimated noisy version of the function given the neighbors of  $v_n$ .

## REFERENCES

- [1] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, Springer New York, 2009.
- [2] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, Jul. 2002, pp. 315–322.
- [3] A. J. Smola and R. I. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines*, pp. 144–158. Springer, 2003.
- [4] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Sig. Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
- [5] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Sig. Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.
- [6] O. Chapelle, B. Schölkopf, A. Zien, et al., *Semi-supervised Learning*, MIT press Cambridge, 2006.
- [7] O. Chapelle, V. Vapnik, and J. Weston, "Transductive inference for estimating values of functions," in *Proc. Advances Neural Inf. Process. Syst.*, Denver, Colorado, 1999, vol. 12, pp. 421–427.
- [8] C. Cortes and M. Mohri, "On transductive regression," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, Canada, 2007, pp. 305–312.
- [9] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
- [10] J. Lafferty and L. Wasserman, "Statistical analysis of semi-supervised regression," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, Canada, 2007.
- [11] S. K. Narang, A. Gadde, and A. Ortega, "Signal processing techniques for interpolation in graph structured data," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, Vancouver, Canada, 2013, IEEE, pp. 5445–5449.
- [12] S. K. Narang, A. Gadde, E. Sanou, and A. Ortega, "Localized iterative methods for interpolation in graph structured data," in *Global Conf. Sig. Inf. Process.*, Austin, Texas, 2013, IEEE, pp. 491–494.
- [13] A. Gadde and A. Ortega, "A probabilistic interpretation of sampling theory of graph signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, Brisbane, Australia, Apr. 2015, pp. 3257–3261.
- [14] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *arXiv preprint arXiv:1507.08822*, 2015.
- [15] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Sig. Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015.
- [16] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *arXiv preprint arXiv:1510.00297*, 2015.
- [17] X. Wang, P. Liu, and Y. Gu, "Local-set-based graph signal reconstruction," *IEEE Trans. Sig. Process.*, vol. 63, no. 9, pp. 2432–2444, May 2015.
- [18] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Sampling of graph signals with successive local aggregations," *IEEE Trans. Sig. Process.*, vol. 64, no. 7, pp. 1832–1843, Apr. 2016.
- [19] S. Segarra, A. G. Marques, G. Leus, and A. Ribeiro, "Reconstruction of graph signals through percolation from seeding nodes," *arXiv preprint arXiv:1507.08364*, 2015.
- [20] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.
- [21] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Mathematical Analysis Appl.*, vol. 33, no. 1, pp. 82–95, 1971.
- [22] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Proc. Annual Conf. Learning Theory*, Banff, Canada, Jul. 2004, Springer, vol. 3120, pp. 624–638.
- [23] S. Chen, A. Sandryhaila, J. M. F. Moura, and J. Kovačević, "Signal recovery on graphs: Variation minimization," *IEEE Trans. Sig. Process.*, vol. 63, no. 17, pp. 4609–4624, Sep. 2015.
- [24] G. Wahba, *Spline Models for Observational Data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, 1990.
- [25] X. Zhu, J. Kandola, Z. Ghahramani, and J. D. Lafferty, "Nonparametric transforms of graph kernels for semi-supervised learning," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, Canada, 2004, pp. 1641–1648.
- [26] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.
- [27] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, "On kernel-target alignment," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, Canada, 2001, pp. 367–373.
- [28] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory*. Springer, 2001, pp. 416–426.
- [29] C. Carmeli, E. De Vito, A. Toigo, and V. Umanita, "Vector valued reproducing kernel Hilbert spaces and universality," *Analysis Appl.*, vol. 8, no. 1, pp. 19–61, 2010.
- [30] D. Zhou and B. Schölkopf, "A regularization framework for learning from graph data," in *ICML Workshop Stat. Relational Learn. Connections Other Fields*, Banff, Canada, Jul. 2004, vol. 15, pp. 67–68.
- [31] P. A. Forero, K. Rajawat, and G. B. Giannakis, "Prediction of partially observed dynamical processes over networks via dictionary learning," *IEEE Trans. Sig. Process.*, vol. 62, no. 13, pp. 3305–3320, Jul. 2014.
- [32] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*, Now Pub, 2006.
- [33] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics. Springer, 2006.
- [34] J.-A. Bazerque and G. B. Giannakis, "Nonparametric basis pursuit via kernel-based learning," *IEEE Sig. Process. Mag.*, vol. 28, no. 30, pp. 112–125, Jul. 2013.
- [35] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. I: Estimation Theory*, Prentice-Hall, 1993.
- [36] J.-A. Bazerque, G. Mateos, and G. B. Giannakis, "Group-lasso on splines for spectrum cartography," *IEEE Trans. Sig. Process.*, vol. 59, no. 10, pp. 4648–4663, Oct. 2011.
- [37] G. B. Giannakis, Q. Ling, G. Mateos, I. D. Schizas, and H. Zhu, "Decentralized learning for wireless communications and networking," *arXiv preprint arXiv:1503.08855*, 2016.
- [38] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," in *J. Mach. Learn. Res.*, 2005, pp. 1099–1125.
- [39] A. Argyriou, M. Herbster, and M. Pontil, "Combining graph Laplacians for semi-supervised learning," in *Proc. Advances Neural Inf. Process. Syst.*, Vancouver, Canada, Dec. 2005, pp. 67–74.
- [40] C. Cortes, M. Mohri, and A. Rostamizadeh, " $l_2$  regularization for learning kernels," in *Conf. Uncertainty Artificial Intell.*, Montreal, Canada, Jun. 2009, pp. 109–116.
- [41] "Bureau of Transportation, United States," [Online]. Available: <http://www.transtats.bts.gov/>, 2016.
- [42] "Federal Office of Meteorology and Climatology MeteoSwiss," [Online]. Available: <http://www.meteoswiss.admin.ch/home/climate/past/climate-normals/climate-diagrams-and-normal-values-per-station.html>, 2014.
- [43] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning laplacian matrix in smooth graph signal representations," *arXiv preprint arXiv:1406.7842*, 2014.