Kernel Based Reconstruction for Generalized Graph Signal Processing

Xingchao Jian, *Graduate Student Member, IEEE*, Wee Peng Tay, *Senior Member, IEEE*, and Yonina C. Eldar, *Fellow, IEEE*

***Abstract*—In generalized graph signal processing (GGSP), the signal associated with each vertex in a graph is an element from a Hilbert space. In this paper, we study GGSP signal reconstruction as a kernel ridge regression (KRR) problem. By devising an appropriate kernel, we show that this problem has a solution that can be evaluated in a distributed way. We interpret the problem and solution using both deterministic and Bayesian perspectives and link them to existing graph signal processing and GGSP frameworks. We then provide an online implementation via random Fourier features. Under the Bayesian framework, we investigate the statistical performance under the asymptotic sampling scheme. Finally, we validate our theory and methods on real-world datasets.**

***Index Terms*—Graph signal processing, generalized graph signal processing, kernel ridge regression, signal reconstruction.**

I. Introduction

In real-world signal processing, data is often associated with a network. Graph signal processing (GSP) techniques have been proposed to perform filtering, sampling and reconstruction for this class of signals by accommodating the network structure [1], [2]. GSP models and exploits the relationship between signals and graphs through the definitions of the graph Fourier transform (GFT) and frequency. In practice, GSP can be utilized to analyze brain signals [3], [4], denoise an image [5], [6], and design recommendation systems [7].

Graph signal reconstruction aims to recover the entire graph signal based on observations from a subset of vertices. The major tasks in graph signal reconstruction are designing optimal sampling and recovery strategies [8], [9]. When the graph signal is bandlimited, [10] derived a least squares estimator. Based on this estimator, [11] formulates the sampling problem as an optimization problem. Assuming wide-sense stationary (WSS) and bandlimited signal and WSS noise, [12] studied a greedy sampling scheme, and derived a bound for its recovery mean-squared error (MSE). The paper [13] derived the Wiener filter for graph signal reconstruction under the assumption of WSS, while [14] studied the reconstruction problem for time-varying graph signals. By requiring smoothness in the vertex domain of the graph signals' first-order difference over time, reconstruction is formulated as an optimization problem. This optimization approach is generalized and accelerated in [15] through the Sobolev smoothness term. The work [16] studied the problem of recovering graph signals from nonlinear measurements.

Kernel-based GSP techniques have more flexibility in filtering and reconstruction, since they introduce nonlinearity and generalize existing approaches. In [17], the graph signal is modeled as a random nonlinear function of an arbitrary input with a specific covariance structure adapted to the graph, known as a Gaussian process over a graph (GPG). The covariance structure is based on a scalar-valued kernel for differentiating the inputs and the graph structure for regularizing the smoothness of the random graph signal. The papers [18], [19], [20] formulate a learning problem with a graph signal target. Besides the standard kernel ridge regression (KRR) fitness and regularization terms, this framework imposes smoothness on the output of the training set. The work [21] generalizes the graph-time linear filter [22, eq. (7)] to a nonlinear predictor via KRR. This model assumes the same nonlinear function on every vertex, hence can be made adaptive and distributed by random Fourier features (RFFs) [23], [24]. In the reconstruction problem, [25], [26] design the graph kernel by viewing the graph signal as a function on the vertex set. This approach generalizes the bandlimited graph signal reconstruction method. By implementing the multi-kernel learning (MKL) strategy, it does not require knowledge of the signal bandwidth.

The aforementioned techniques are developed in terms of the classical GSP framework, where each vertex signal is a *scalar*. In practice, the data associated with each vertex can have additional structure. For example, on each vertex, the observation may be a discrete-time signal of length T. This scenario is considered in the time-vertex framework [15], [27], [28], [29], where the spatial-time structure is modeled by a Cartesian product graph, and the Fourier transform and filters are then generalized to this graph. To be specific, the Cartesian

Manuscript received 14 August 2023; revised 5 March 2024 and 22 April 2024; accepted 25 April 2024. Date of publication 30 April 2024; date of current version 9 May 2024. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE-T2EP20220-0002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joao F. C. Mota. *(Corresponding author: Xingchao Jian.)*

Xingchao Jian and Wee Peng Tay are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: xingchao001@e.ntu.edu.sg; wptay@ntu.edu.sg).

Yonina C. Eldar is with the Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2024.3395021, provided by the authors.

Digital Object Identifier 10.1109/TSP.2024.3395021

product graph is constructed by the underlying graph and the cyclic graph with T vertices, the latter of which represents time steps. The data can then be embedded in this product graph as a standard graph signal. This framework is further extended to the generalized graph signal processing (GGSP) framework [30], [31], [32], where each vertex observation is an element from a Hilbert space, which can possibly be infinite-dimensional. An important example is the case where each vertex is associated with a continuous function on a bounded interval. This model allows for analyzing asynchronously sampled signals on each vertex, which is not possible under the time-vertex framework.

In this paper, we explore kernel-based signal reconstruction within the GGSP framework. Previous works like [26] have developed signal reconstruction methods within the traditional GSP or time-vertex frameworks. In GGSP, we consider vertex signals as elements in a *general Hilbert space*. In [31], a reconstruction method proposed for GGSP assumes that the signal lies in a finite-dimensional subspace (e.g., finitely many features of the signal spans the full space of interest). The work [32] proposed a reconstruction method that relies on knowledge of the signal's power spectral density (PSD) (e.g., this can be derived from a noiseless training set without missing values), which may not be applicable in practice. In this paper, we consider the case where the training set is small, noisy and incomplete, thus the method in [32] cannot be applied. Specifically, we consider the case where signal on each vertex is real-valued function. By utilizing a reasonable kernel, we are able to reconstruct the signal with good fidelity as long as the target signal is in the corresponding reproducing kernel Hilbert space (RKHS), which can be infinite-dimensional. Compared to the method in [31], our proposed method is able to utilize infinitely many features. Thus, it is more flexible and has better signal representation capability.

To motivate our work, consider the Intel lab temperature dataset[1] which consists of temperature records from 54 sensors in a lab, collected between February and April of 2004. The ground truth records and incomplete noisy observations on two connected sensors labeled as vertex 1 and 2 are shown in Fig. 1. Our goal is to reconstruct the signal at vertex 1. In the time interval $[0, 40000]$, there is a lack of observations on vertex 1. As shown in Fig. 1, the isolated KRR method fails to reconstruct this part. On the other hand, our proposed approach, referred to as KRR-GGSP, utilizes the graph structure to incorporate the observations from a vertex's neighbor to improve reconstruction. This example motivates the need for a new KRR framework under GGSP, which is the focus of this paper. Unlike the methods under WSS or joint wide-sense stationary (JWSS) assumptions [13], [29], KRR-GGSP does not require knowledge of the PSD of the signal, which can be hard to estimate when there are only noisy and incomplete samples in the training set. Further numerical experiments in Section V illustrate the utility of the approach presented in this paper.

[1]http://db.csail.mit.edu/labdata/labdata.html

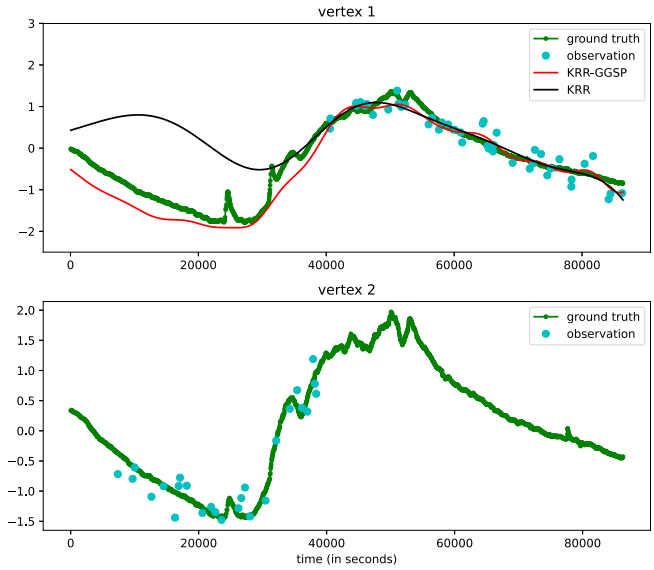


Fig. 1. The upper and lower plots represent observations and ground truths from two connected vertices 1 and 2, respectively. Green curves are ground truth signals, and the cyan dots represent the observations on each vertex. Note that reconstruction based on the single vertex 1's observations using KRR is much worse compared to the proposed KRR-GGSP approach.

Our main contributions are the following:

1) We construct an appropriate kernel and formulate the signal reconstruction in GGSP as a KRR problem. We interpret it as an extension of existing kernel-based frameworks.
2) We present an online approach for generalized graph signal reconstruction by utilizing RFF.
3) We compute the limit and asymptotic upper bound for conditional MSE of reconstruction under the Bayesian framework.
4) We present numerical case studies to illustrate the utility of KRR-GGSP in several applications.

This paper is related to our conference paper [33], whose goal was to learn a map from a generalized graph signal space to itself in *filtering*. We made use of the tensor product operator-valued kernel to formulate this filtering problem. In this paper, we instead study the *reconstruction* problem for generalized graph signal and our goal is to learn a function from the set of sample points to $\mathbb{R}$. Here the sample points are pairs of vertices and instances of the vertex function's domain. To achieve this, we consider a real-valued kernel defined on the set of sample points. We make use of the tensor product strategy to form a kernel.

The rest of this paper is organized as follows. In Section II, we formulate the signal reconstruction problem in GGSP. In Section III, we derive the solution to this problem, discuss its interpretation and compare it with existing methods. We also provide an online version of the reconstruction problem. In Section IV, we analyze the statistical performance of our reconstruction approach under the asymptotic case. In Section V, we validate our method on real-world datasets. We conclude in Section VI.

Notations. We use plain lower cases (e.g., x) to represent scalars and scalar-valued functions. We use bold lower cases (e.g., $\mathbf{x}$) to represent vectors and vector-valued functions. Note that in this paper, we consider generalized graph signals as scalar-valued functions. Although they are vectors in linear spaces, we use the functional view for ease of explanation. Bold upper cases (e.g., $\mathbf{S}$) are used to denote operators, including matrices. In particular, we write the N-dimensional identity operator or matrix as $\mathbf{I}_N$. We use calligraphic letters to represent spaces (e.g., $\mathcal{X}$), except for standard spaces like $\mathbb{R}$ and $\mathbb{N}$, which are the Euclidean space and space of natural numbers, respectively. For a Hilbert space $\mathcal{H}$, its inner product is $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and corresponding norm is $\|\cdot\|_{\mathcal{H}}$. For a set $\mathcal{X}$, we use $\mathcal{X}^{\mathsf{c}}$ to denote its complement. For two random variables (or elements) x and y, we write $x \in \sigma(y)$ if x is measurable with respect to (w.r.t.) the σ-algebra generated by y. We write $\mathrm{cov}(x,y)$ to denote the covariance between x and y, and $\mathrm{var}(x)$ to denote the variance of x. We use $\delta(\cdot,\cdot)$ to denote the Kronecker delta function, which equals 1 if its two arguments are the same and 0 otherwise. The tensor product is denoted by $\otimes$ and $\mathrm{diag}(\mathbf{v})$ is the diagonal matrix with its main diagonal given by the vector $\mathbf{v}$. The element-wise matrix multiplication is denoted by $\odot$, $(\cdot)^{\intercal}$ denotes transpose, $(\cdot)^*$ denotes conjugate transpose or the adjoint, and $(\cdot)^{\dagger}$ denotes the pseudo-inverse. We use $[m]$ to represent the set $\{1,\dots,m\}$.

II. Problem Formulation

In this section, we formulate the generalized graph signal reconstruction problem.

Consider a graph $G=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{1,\dots,N\}$ is the vertex set, and $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$ is the edge set. We use $\mathcal{N}_d(v)$ to denote the d-hop neighborhood of the vertex v and let $\overline{\mathcal{N}}_d(v)=\mathcal{N}_d(v)\cup\{v\}$. We assume that G is a connected undirected graph with no self-loops. In GSP theory, a typical graph signal is a function mapping from $\mathcal{V}$ to $\mathbb{R}$.[2] In the GGSP framework [31], the generalized graph signal f is defined as a function from $\mathcal{V}$ to a separable Hilbert space $\mathcal{H}$. The generalized graph signal space can then be identified with $\mathbb{R}^N\otimes\mathcal{H}$ via the map

$$f\mapsto\sum_{n=1}^{N}\mathbf{e}_n\otimes f(n),$$

where $\{\mathbf{e}_n : n=1,\dots,N\}$ is the standard basis of $\mathbb{R}^N$, i.e., $\mathbf{e}_n$ is the n-th column vector of $\mathbf{I}_N$.

One important case in GGSP is where $\mathcal{H}$ is a function space. Specifically, consider the domain of the functions to be a measure space $(\mathcal{T},\mathcal{A},\tau)$ and $\mathcal{H}=L^2(\mathcal{T})$ (i.e., the space of square-integrable functions on $\mathcal{T}$). For example, in Intel lab data mentioned in Section I, $\mathcal{T}=[0,86400]$, representing the time duration (in seconds) of one day. Then, a generalized graph signal f can be identified with the map

$$\begin{aligned} f' : \mathcal{V}\times\mathcal{T} &\to \mathbb{R}\\ (v,\mathbf{t}) &\mapsto f(v)(\mathbf{t}).\end{aligned}$$

[2]For simplicity, we consider only $\mathbb{R}$-valued signals instead of $\mathbb{C}$-valued signals.

Thus, the space of generalized graph signals can be also identified with $L^2(\mathcal{V}\times\mathcal{T})$. In this paper, we will mainly use $L^2(\mathcal{V}\times\mathcal{T})$ to denote the space of generalized graph signals, while references to $\mathbb{R}^N\otimes\mathcal{H}$ are used in explanations and proofs. We refer to $\mathcal{T}$ colloquially as the *time* domain. However, it is not restricted to subsets of $\mathbb{R}$ and can be a general measure space. Readers are referred to Appendix A and [31] for more details on GGSP.

Given noisy observation samples at a subset $\mathcal{S}\subset\mathcal{V}\times\mathcal{T}$ of vertices and time instances, our objective is to recover the generalized graph signal f. To avoid cluttered notations, denote $\mathcal{J}=\mathcal{V}\times\mathcal{T}$. Suppose the sampling set is $\mathcal{S}=\{(v_m,\mathbf{t}_m) : m=1,\dots,M\}\subset\mathcal{J}$, and the noisy observations are

$$y_m = f(v_m,\mathbf{t}_m)+\epsilon_m,\ m=1,\dots,M, \tag{1}$$

where ϵ_m are independent and identically distributed (i.i.d.) zero-mean noise with variance σ^2. In the Bayesian framework, f in (1) is further modeled as a Gaussian process. In this case, we will model f as a random element (cf. Appendix B). The noise terms ϵ_m are assumed to be Gaussian and independent of this process.

The GGSP signal reconstruction problem can be summarized in the following form:

$$\min_{\tilde{f}\in F(\mathcal{J},\mathbb{R})}\sum_{m=1}^{M}L(\tilde{f}(v_m,\mathbf{t}_m),y_m)+P(\tilde{f}), \tag{2}$$

where $F(\mathcal{J},\mathbb{R})$ is an appropriate space of functions from $\mathcal{J}$ to $\mathbb{R}$, $L(\cdot)$ is a loss function measuring the fitness of $\tilde{f}$ on the observations. Typical choices include the ℓ_1 and ℓ_2 losses. The regularization term $P(\tilde{f})$ imposes a smoothness constraint on $\tilde{f}$ over the vertex and time domains. To design proper $F(\mathcal{J},\mathbb{R})$ and $P(\tilde{f})$, we employ the KRR technique, which we briefly review in Appendix C.

The existing time-vertex methods [14], [15] have already addressed the reconstruction problem for time series on graphs. However, these methods are based on the assumption that the signals are evenly sampled with the same sampling rate on all vertices. In contrast, from (2), we observe that our formulation does not require synchronous samples from each vertex and applies even in the case where the sampling frequencies differ across vertices, or where the signal is not evenly sampled. In addition, compared to the time-vertex techniques, this formulation is not sensitive to the sampling rate since it makes use of the true time stamps. We refer the reader to the detailed discussion in Section III-B.

III. KRR Reconstruction in GGSP

In this section, we derive the KRR reconstruction solution for GGSP. We interpret this method under both deterministic and Bayesian models and connect our technique with existing kernel-based frameworks in GSP and graph signal reconstruction approaches. We also propose an online approach based on RFF that results in a distributed implementation.

To reconstruct a generalized graph signal $f \in L^2(\mathcal{J})$, we use a kernel $k : \mathcal{J} \times \mathcal{J} \to \mathbb{R}$ that is the multiplication of two kernels $k_G : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and $k_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$:

$$\begin{aligned} k &: \mathcal{J} \times \mathcal{J} \to \mathbb{R} \\ ((u, \mathbf{s}), (v, \mathbf{t})) &\mapsto k_G(u, v) k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}). \end{aligned} \tag{3}$$

The RKHS associated with the kernel (3) is $\mathcal{H}_k = \mathcal{H}_{k_G} \otimes \mathcal{H}_{k_{\mathcal{T}}}$ [34, Theorem 13]. In this paper, we construct k_G based on a graph shift operator (GSO) $\mathbf{A}_G$ of the graph G. A GSO is a $N \times N$ matrix representing the structure of the graph G such that its (u, v)-th entry is nonzero only if $(u, v) \in \mathcal{E}$ or $u = v$. Typical choices of GSO are graph adjacency and Laplacian matrices. In particular, we focus on the case where the matrix $\mathbf{K}_G := (k_G(i, j)) \in \mathbb{R}^{N \times N}$ takes the following form (cf. [26, (14)]):

$$\mathbf{K}_G = \mathbf{\Phi} \operatorname{diag}(r(\lambda_1), \ldots, r(\lambda_N)) \mathbf{\Phi}^{\mathsf{T}}, \tag{4}$$

where $\{\lambda_i\}$ are the eigenvalues of the GSO $\mathbf{A}_G$, $r(\cdot)$ is a non-negative function such that $r(\lambda_1) \geq \cdots \geq r(\lambda_N)$,[3] and $\mathbf{\Phi}$ is the matrix formed by the eigenvectors of $\mathbf{A}_G$. When $\mathcal{T}$ is a subset of Euclidean space, we can usually choose $k_{\mathcal{T}}$ as the radial basis function (RBF) kernel, e.g., $k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \exp(-\|\mathbf{s} - \mathbf{t}\|_2^2 / \beta_{\text{scale}})$ (Gaussian kernel) or $k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \exp(-\|\mathbf{s} - \mathbf{t}\|_1 / \beta_{\text{scale}})$ (Laplacian kernel), where β_{scale} is a tunable parameter.

Following the standard KRR formulation (36), we specify the reconstruction problem (2) as follows:

$$\hat{f} = \operatorname*{arg\,min}_{\tilde{f} \in \mathcal{H}_k} \sum_{m=1}^{M} |\tilde{f}(v_m, \mathbf{t}_m) - y_m|^2 + \mu \|\tilde{f}\|_{\mathcal{H}_k}^2. \tag{5}$$

Let $\mathbf{K}(\mathcal{S}, \mathcal{S}) = (k((v_m, \mathbf{t}_m), (v_{m'}, \mathbf{t}_{m'})))_{m,m'=1}^{M} \in \mathbb{R}^{M \times M}$ and $\mathbf{y}(\mathcal{S}) = (y_1, \ldots, y_M)^{\mathsf{T}}$. Using the representer theorem, the optimal solution to (5) is

$$\begin{aligned} \hat{f} &= \sum_{m=1}^{M} c_m k(\cdot, (v_m, \mathbf{t}_m)), \\ (c_1, \ldots, c_M)^{\mathsf{T}} &= (\mathbf{K}(\mathcal{S}, \mathcal{S}) + \mu \mathbf{I}_M)^{-1} \mathbf{y}(\mathcal{S}). \end{aligned} \tag{6}$$

Henceforth, we refer to the problem (5) and its solution (6) as KRR-GGSP. In this paper, we assume that all eigenvalues of $\mathbf{A}_G$ are distinct. By construction (4), $\mathbf{K}_G$ is a polynomial of $\mathbf{A}_G$ for some degree $L < N$, i.e., it suffices to consider $r(\cdot)$ as a polynomial whose degree is smaller than N, thus $k_G(u, v) = 0$ as long as $u \notin \overline{\mathcal{N}}_L(v)$. Therefore, the evaluation of $\hat{f}(v, \mathbf{t})$ only requires information from $\overline{\mathcal{N}}_L(v)$:

$$\begin{aligned} \hat{f}(v, \mathbf{t}) &= \sum_{m=1}^{M} c_m k((v, \mathbf{t}), (v_m, \mathbf{t}_m)) \\ &= \sum_{m=1}^{M} c_m k_G(v, v_m) k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_m) \\ &= \sum_{v_m \overline{\mathcal{N}}_L(v)} c_m k_G(v, v_m) k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_m). \end{aligned} \tag{7}$$

[3]Recall that $\{\lambda_i\}$ are indexed in increasing order of graph frequencies. Also note that [26, (14)] uses $r^{\dagger}(\mathbf{\Lambda})$ instead of $r(\mathbf{\Lambda})$ in the definition (4).

Note that when $\mathcal{T}$ is a singleton (i.e., the vertex signal space is one-dimensional), the KRR-GGSP framework degenerates to the GSP recovery problem [26]. In addition, when $\mathbf{K}_G = \mathbf{I}_N$, it degenerates to separately solving KRR problems on each vertex using the kernel $k_{\mathcal{T}}$. To see this, suppose on each vertex v we have M_v samples. We relabel $\mathcal{S}$ and $\{y_m\}$ such that $\mathcal{S} = \bigcup_{v \in \mathcal{V}} \{(v, \mathbf{t}_i^{(v)}) : i = 1, \ldots, M_v\}$, $\{y_m\} = \bigcup_{v \in \mathcal{V}} \{y_i^{(v)} : i = 1, \ldots, M_v\}$. We also relabel the coefficients as $c_i^{(v)}$, so that (6) can be rewritten as

$$\hat{f}(u, \mathbf{t}) = \sum_{v=1}^{N} \delta(u, v) \sum_{i=1}^{M_u} c_i^{(v)} k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_i^{(v)})$$

for each $u \in \mathcal{V}$ and $\mathbf{t} \in \mathcal{T}$, where $\delta(u, v) = 1$ when $u = v$ and $\delta(u, v) = 0$ otherwise. Note that $\hat{f}(u, \mathbf{t}) = \sum_{i=1}^{M_u} c_i^{(u)} k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_i^{(u)})$ and

$$\|\hat{f}\|_{\mathcal{H}_k}^2 = \sum_{u=1}^{N} \sum_{i,j=1}^{M_u} c_i^{(u)} k_{\mathcal{T}}(\mathbf{t}_i^{(u)}, \mathbf{t}_j^{(u)}) c_j^{(u)} = \sum_{u=1}^{N} \|\hat{f}(u, \cdot)\|_{\mathcal{H}_{k_{\mathcal{T}}}}^2.$$

Then problem (5) becomes

$$\hat{f} = \operatorname*{arg\,min}_{\tilde{f} \in \mathcal{H}_k} \sum_{u=1}^{N} \sum_{i=1}^{M_u} |\tilde{f}(u, \mathbf{t}_i^{(u)}) - y_i^{(u)}|^2 + \mu \sum_{u=1}^{N} \|\tilde{f}(u, \cdot)\|_{\mathcal{H}_{k_{\mathcal{T}}}}^2, \tag{8}$$

and each $\hat{f}(u, \cdot)$ can be solved separately using the samples on the vertex u.

In the rest of this paper, we make the following assumption.

Assumption 1: For the measure space $(\mathcal{T}, \mathcal{A}, \tau)$, $\mathcal{T}$ is a compact metric space, $\mathcal{A}$ is the Borel σ-algebra, and τ is a strictly positive finite Borel measure. The kernel $k_{\mathcal{T}}$ is a continuous symmetric positive definite kernel and $\mathbf{K}_G$ is a positive definite matrix.

A. Deterministic Interpretation

In this subsection, we consider the case where f in (1) is deterministic. Under Assumption 1, by Mercer's theorem [35], there exists an orthonormal sequence $\{\xi_i : i \geq 1\}$ in $L^2(\mathcal{T})$ such that:

$$\begin{aligned} \int_{\mathcal{T}} k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \xi_i(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}) &= \gamma_i \xi_i(\mathbf{t}), \\ \int_{\mathcal{T}} \xi_i(\mathbf{s}) \xi_j(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}) &= \delta(i, j), \\ k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) &= \sum_{i=1}^{\infty} \gamma_i \xi_i(\mathbf{s}) \xi_i(\mathbf{t}), \end{aligned}$$

where the sum converges absolutely and uniformly on $\mathcal{T}$ and γ_i, $i \geq 1$, are non-negative eigenvalues. Let $\phi_n(u)$ be the (n, u)-th element of $\mathbf{\Phi}$. Since k_G is given by (4), it can be decomposed in the same way:

$$k_G(u, v) = \sum_{n=1}^{N} r(\lambda_n) \phi_n(u) \phi_n(v).$$

By definition of k in (3), we then have

$$k((u,\mathbf{s}),(v,\mathbf{t}))=\sum_{n=1}^{N}\sum_{i=1}^{\infty} r(\lambda_n)\gamma_i\cdot\phi_n(u)\xi_i(\mathbf{s})\cdot\phi_n(v)\xi_i(\mathbf{t}).$$

Note that $\{\phi_n(\cdot)\xi_i(\cdot):n=1,\dots,N,i\geq 1\}$ is a orthonormal sequence in $L^2(\mathcal{J})$. Following the same argument as [36], $\mathcal{H}_k$ is a subset of $L^2(\mathcal{J})$ where the functions $\tilde{f}$ satisfy the following condition:

$$\tilde{f}(v,\mathbf{t})=\sum_{n=1}^{N}\sum_{i=1}^{\infty} c_{n,i}\cdot\phi_n(v)\xi_i(\mathbf{t})$$
$$\text{s.t. } \|\tilde{f}\|_{\mathcal{H}_k}^2=\sum_{n=1}^{N}\sum_{i=1}^{\infty}\frac{c_{n,i}^2}{r(\lambda_n)\gamma_i}<\infty. \tag{9}$$

By the definition of joint Fourier transform (JFT) (cf. (28)), it can be shown that $c_{n,i}=\mathfrak{F}_{n,i}(\tilde{f})$ where $\mathfrak{F}_{n,i}$ represents the (n,i)-th JFT coefficient. Therefore, penalizing on $\|\tilde{f}\|_{\mathcal{H}_k}$ is the same as penalizing on the energy of $\mathfrak{F}_{n,i}(\tilde{f})$ with weights $\frac{1}{r(\lambda_n)\gamma_i}$. Note that $r(\cdot)$ is non-increasing so that the Fourier coefficients associated with larger graph frequencies are more heavily penalized.

It is worth noting that if we construct $k_{\mathcal{T}}$ and k_G such that

$$k_{\mathcal{T}}(\mathbf{s},\mathbf{t})=\sum_{i=1}^{B'}\gamma_i\xi_i(\mathbf{s})\xi_i(\mathbf{t}) \tag{10}$$

for some $B'<\infty$, and $r(\lambda_n)=0$ for all $n>B''$ in (4), then problem (5) is equivalent to the bandlimited signal reconstruction in [31, Section VI.A] with an additional ridge penalty. To see this, we first note that $\mathcal{H}_k=\text{span}\{\phi_n(\cdot)\xi_i(\cdot):n=1,\dots,B'',i=1,\dots,B'\}$, i.e., the signal space used for reconstruction is a bandlimited space. We substitute (10) into (9) to obtain the optimization problem

$$\hat{f}(v,\mathbf{t})=\arg\min_{\tilde{f}\in\mathcal{H}_k}\sum_{m=1}^{M}|\tilde{f}(v_m,\mathbf{t}_m)-y_m|^2+\mu\sum_{n=1}^{B''}\sum_{i=1}^{B'}\frac{c_{n,i}^2}{r(\lambda_n)\gamma_i}, \tag{11}$$

which coincides with the bandlimited signal reconstruction problem formulated in [31] but with an additional penalty term. This indicates that if $k_{\mathcal{T}}$ is not a combination of finite functions, then $\dim(\mathcal{H}_k)=\infty$. This implies that the algorithm is able to capture more features than that of bandlimited signals. An example is the Gaussian kernel [37, Section 4.3.1].

Finally, we discuss the universality (see Appendix C for definition) of the kernel k in the following theorem. The definition of universality requires defining a topology on $\mathcal{J}$. In this paper, we equip $\mathcal{V}$ with the discrete topology and $\mathcal{J}=\mathcal{V}\times\mathcal{T}$ the product topology.

Theorem 1: If $k_{\mathcal{T}}$ is a universal kernel on $\mathcal{T}$, then k is universal on $\mathcal{J}$.

Proof: Consider an arbitrary compact set $\mathcal{Z}_J\subset\mathcal{V}\times\mathcal{T}$, and define $\mathcal{Z}_v$ such that $\{v\}\times\mathcal{Z}_v=\mathcal{Z}_J\cap(\{v\}\times\mathcal{T})$. By using the finite-cover definition of a compact set, we note that $\mathcal{Z}_v$ is compact in $\mathcal{T}$. Consider an arbitrary $h\in\mathcal{C}(\mathcal{Z}_J)$, where $\mathcal{C}(\mathcal{Z}_J)$ is the space of continuous functions on $\mathcal{Z}_J$ equipped with the supremum norm. Let $h_v:=h|_{\{v\}\times\mathcal{Z}_v}$. Due to the universality of $k_{\mathcal{T}}$, for any $\epsilon>0$, there exists $h_v'\in\text{span}\{k_{\mathcal{T}}(\cdot,\mathbf{t}):\mathbf{t}\in\mathcal{Z}_v\}$ such that $\|h_v'-h_v(v,\cdot)\|_{\mathcal{C}(\mathcal{Z}_v)}<\epsilon$. Let $h':=\sum_{v=1}^{N}\delta(v,\cdot)h_v'$. Then, we have $\|h'-h\|_{\mathcal{C}(\mathcal{Z}_J)}<\epsilon$. On the other hand, since $\mathbf{K}_G$ is positive definite, $\mathbf{K}_G$ is invertible. Therefore, there exists $\{a_{v,n}\}$ such that $\delta(v,\cdot)=\sum_{n=1}^{N}a_{v,n}k_G(n,\cdot)$, i.e., $\delta(v,\cdot)\in\text{span}\{k_G(n,\cdot):n=1,\dots,N\}$. Let $\mathcal{K}(\mathcal{Z}_J)$ be the closure of $\text{span}\{k(\cdot,(u,\mathbf{s})):(u,\mathbf{s})\in\mathcal{V}\times\mathcal{T}\}$ in $\mathcal{C}(\mathcal{Z}_J)$. By combining the above results, we conclude that $h'\in\mathcal{K}(\mathcal{Z}_J)$ and the universality of k follows by $\mathcal{K}(\mathcal{Z}_J)=\mathcal{C}(\mathcal{Z}_J)$. □

The universality discussed in Theorem 1 is different from that in [33, Theorem 2], which established universality for the following operator-valued kernels:

$$\mathbf{K}:\mathcal{X}\times\mathcal{X}\to\mathcal{L}(\mathcal{Y})$$
$$(\mathbf{x}_1,\mathbf{x}_2)\mapsto k_s(\mathbf{x}_1,\mathbf{x}_2)\mathbf{T}$$

where $\mathcal{X}\subset L^2(\mathcal{J})$ and $\mathcal{Y}\subset L^2(\mathcal{J})$, $k_s:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a real-valued RBF kernel, $\mathcal{L}(\mathcal{Y})$ is the space of linear operators on $\mathcal{Y}$, and $\mathbf{T}\in\mathcal{L}(\mathcal{Y})$.

We note that Theorem 1 cannot be derived from [33, Theorem 2] and vice versa. First, the kernel domain in this paper is in $\mathcal{J}\times\mathcal{J}$ instead of $L^2(\mathcal{J})\times L^2(\mathcal{J})$. For simplicity, consider the case where $\mathcal{T}$ is a singleton, so that $\mathcal{V}\times\mathcal{T}$ can be identified with $\mathcal{V}$. If we use the kernel in [33, Theorem 2] and let $\mathcal{X}=\mathcal{V}$, then it is required that $\mathcal{V}$ is a real (or complex) separable Hilbert space. However, as long as $1<|\mathcal{V}|<\infty$, this is impossible. Second, the output of the kernel in this paper is in $\mathbb{R}$ instead of an operator space, hence none of these two formulations encompasses the other.

B. Bayesian Interpretation

We now turn to the Bayesian interpretation where $f\sim\mathcal{GP}(0,k)$ and $\epsilon_m\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$ in (1). Let $(\Omega,\mathcal{F},\mathbb{P})$ be the underlying probability space, where $\mathcal{F}$ stands for the σ-algebra of the space. We regard $\mathcal{J}=\mathcal{V}\times\mathcal{T}$ as a measure space whose measure is the product measure of counting measure on $\mathcal{V}$ and the measure τ on $\mathcal{T}$. We denote this product measure as ζ. To be specific, f is a stochastic process $\{f((v,\mathbf{t}),\omega):(v,\mathbf{t})\in\mathcal{J},\omega\in\Omega\}$. We make the following assumptions:

Assumption 2:

i) $f((v,\mathbf{t}),\omega)$ is jointly measurable w.r.t. the product measure $\zeta\times\mathbb{P}$.

ii) $f(\cdot,\omega)\in L^2(\mathcal{J})$ for all $\omega\in\Omega$.

Under Assumption 2, f is a Gaussian random element (cf. Theorem B.1). Henceforth, we abbreviate $f((v,\mathbf{t}),\omega)$ as $f(v,\mathbf{t})$ for simplicity and consistent notations. First, we note that under the time-vertex framework, the Gaussian process (GP) prior $\mathcal{GP}(0,k)$ is a JWSS graph random process (GRP). A stochastic process f on $\mathcal{V}\times\mathcal{T}$ is said to be JWSS if its covariance operator commutes with the shift operator $\mathbf{S}:=\mathbf{A}_G\otimes\mathbf{A}_{\mathcal{H}}$ on $L^2(\mathcal{J})$ [32, Definition 2], where $\mathbf{A}_{\mathcal{H}}$ is the shift operator on $L^2(\mathcal{T})$. Consider the case where $\mathcal{T}=\{1,\dots,T\}$, and $\mathbf{K}_{\mathcal{T}}:=(k_{\mathcal{T}}(i,j))\in\mathbb{R}^{T\times T}$ is a symmetric positive-definite circulant matrix. Then the covariance operator of $\mathcal{GP}(0,k)$ is $\mathbf{C}_f=\mathbf{K}_G\otimes\mathbf{K}_{\mathcal{T}}$. Let $\mathbf{A}_{\mathcal{H}}$ be the shift operator

$$\mathbf{A}_{\mathcal{H}}(g)(t)=g((t+1)\bmod T),$$

which models the case where the vertex observation is a discrete-time signal with T time steps. Since $\mathbf{K}_{\mathcal{T}}$ is a circulant matrix, it commutes with $\mathbf{A}_{\mathcal{H}}$. On the other hand, by the construction of the kernel k_G in (4), we know that $\mathbf{K}_G$ commutes with $\mathbf{A}_G$. Therefore, $\mathbf{C}_f$ commutes with the shift operator $\mathbf{S} = \mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$, hence $\mathcal{GP}(0, k)$ is a JWSS prior.

Example 1: The GP prior generalizes the GPG framework [17], which defined a GPG as a vector-valued GP whose covariance matrix takes the form

$$\begin{aligned}\operatorname{cov}(\mathbf{s}, \mathbf{t}) &= k_{\mathcal{T}}(\mathbf{s}, \mathbf{t})\mathbf{B}(a),\\ \mathbf{B}(a) &= (\mathbf{I}_N + a\mathbf{L})^{-2} := (B(a)_{ij}),\end{aligned}$$

where $a > 0$ is a parameter and $\mathbf{L}$ is the graph Laplacian matrix. We see that this covariance structure corresponds to a GP prior in $L^2(\mathcal{J})$ with $k_G(i, j) = B(a)_{ij}$. The GPG also assumes that each observation is $(\mathbf{t}, \mathbf{x})$, where $\mathbf{x}$ is a complete graph signal, while in (5) we allow the observed graph signals to be incomplete. Therefore, this generalization allows us to reconstruct the generalized graph signal when the observations come from different subsets of vertices at different instances.

We next consider the posterior. The observations $\{(v_m, \mathbf{t}_m, y_m)\}$ are denoted as $\mathcal{D}_{\text{train}}$. According to Appendix C, the maximum a posteriori (MAP) estimator is given by (6) with $\mu = \sigma^2$. Since f is a GP, (6) is also the posterior expectation given $\mathcal{D}_{\text{train}}$, i.e., $\hat{f}(v, \mathbf{t}) = \mathbb{E}[f(v, \mathbf{t}) : \mathcal{D}_{\text{train}}]$. The posterior variance can be calculated by

$$\begin{aligned}&\operatorname{var}(f(v, \mathbf{t}) \mid \mathcal{D}_{\text{train}})\\ &\quad = k((v, \mathbf{t}), (v, \mathbf{t})) - \mathbf{k}^{\intercal}(\mathbf{K}(\mathcal{S}, \mathcal{S}) + \sigma^2\mathbf{I}_M)^{-1}\mathbf{k}, \qquad (12)\end{aligned}$$

where $\mathbf{k} := (k((v, \mathbf{t}), (v_1, \mathbf{t}_1)), \ldots, k((v, \mathbf{t}), (v_m, \mathbf{t}_m)))^{\intercal}$. This observation indicates that the time-vertex signal reconstruction approach is a special case of the KRR-GGSP approach.

Example 2: In the time-vertex signal reconstruction problem, the observed signal $\mathbf{X}_o \in \mathbb{R}^{N\times T}$ is an incomplete and noisy observation of the original signal $\mathbf{X}_r \in \mathbb{R}^{N\times T}$. The mask matrix is $\mathbf{\Pi}_{\mathcal{S}} \in \{0, 1\}^{N\times T}$. The paper [15] formulated the graph signal reconstruction via Sobolev smoothness (GTRSS) problem as follows:

$$\begin{aligned}\hat{\mathbf{X}}_r &= \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}^{N\times T}} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2\\ &\quad + \mu_{\text{TV}} \operatorname{tr}((\mathbf{X}\mathbf{D}_h)^{\intercal}(\mathbf{L} + \alpha\mathbf{I})^{\beta}\mathbf{X}\mathbf{D}_h)\\ &= \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}^{N\times T}} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2\\ &\quad + \mu_{\text{TV}} \operatorname{vec}(\mathbf{X})^{\intercal}(\mathbf{D}_h\mathbf{D}_h^{\intercal}) \otimes (\mathbf{L} + \alpha\mathbf{I})^{\beta} \operatorname{vec}(\mathbf{X}), \qquad (13)\end{aligned}$$

where $\mathbf{D}_h$ is the first order difference operator

$$\mathbf{D}_h = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{T\times(T-1)}.$$

For ease of further analysis, we slightly modify (13) to be

$$\begin{aligned}\hat{\mathbf{X}}_r &= \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}^{N\times T}} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2\\ &\quad + \mu_{\text{TV}} \operatorname{vec}(\mathbf{X})^{\intercal}(\mathbf{D}_h\mathbf{D}_h^{\intercal} + \delta_o\mathbf{I}) \otimes (\mathbf{L} + \alpha\mathbf{I})^{\beta} \operatorname{vec}(\mathbf{X}), \qquad (14)\end{aligned}$$

where $\delta_o > 0$. We also assume that $\operatorname{diag}(\operatorname{vec}(\mathbf{\Pi}_{\mathcal{S}})) + (\mathbf{D}_h\mathbf{D}_h^{\intercal}) \otimes (\mathbf{L} + \alpha\mathbf{I})^{\beta}$ is full-rank. It can be shown that the solution to (14) can approximate that of (13) arbitrarily well as long as δ_o is small enough.

We consider problem (14) under a Bayesian setting. Let the prior of $\operatorname{vec}(\mathbf{X}_r)$ be a Gaussian random vector with zero mean and covariance $((\mathbf{D}_h^{\intercal}\mathbf{D}_h + \delta_o\mathbf{I}) \otimes (\mathbf{L} + \alpha\mathbf{I})^{\beta})^{-1}$. In other words, if we let $k_{\mathcal{T}}(s, t) = (\mathbf{D}_h\mathbf{D}_h^{\intercal} + \delta_o\mathbf{I})^{-1}_{s,t}$, and $\mathbf{K}_G = (\mathbf{L} + \alpha\mathbf{I})^{-\beta}$, then $\mathbf{X}_r = (X_r(v, t))$ is a GP with covariance $\operatorname{cov}(X_r(u, s), X_r(v, t)) = k_{\mathcal{T}}(s, t)k_G(u, v)$. Suppose the noise is i.i.d. with variance μ_{TV}, then the objective function in (14) is the log-likelihood of the posterior $p(\mathbf{X}_r \mid \mathbf{X}_o)$ (up to a constant):

$$\begin{aligned}&-\log(p(\mathbf{X}_r \mid \mathbf{X}_o)) = -(\log(p(\mathbf{X}_r, \mathbf{X}_o)) - \log(p(\mathbf{X}_o)))\\ &= -(\log(p(\mathbf{X}_o \mid \mathbf{X}_r)) + \log(p(\mathbf{X}_r)) - \log(p(\mathbf{X}_o)))\\ &= \frac{1}{\mu_{\text{TV}}} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X}_r - \mathbf{X}_o\|_F^2\\ &\quad + \operatorname{vec}(\mathbf{X}_r)^{\intercal}(\mathbf{D}_h\mathbf{D}_h^{\intercal} + \delta_o\mathbf{I}) \otimes (\mathbf{L} + \alpha\mathbf{I})^{\beta} \operatorname{vec}(\mathbf{X}_r)\\ &\quad + \text{const},\end{aligned}$$

where const is a constant independent of $\mathbf{X}_r$. Therefore, the solution to this problem is the MAP of $\mathbf{X}_r$ given $\mathbf{X}_o$. According to the Bayesian interpretation in Appendix C, this MAP estimator $\hat{\mathbf{X}}_r = (\hat{X}_r(v, t))$ is the solution (6) of KRR-GGSP where $k_{\mathcal{T}}(s, t) = (\mathbf{D}_h\mathbf{D}_h^{\intercal} + \delta_o\mathbf{I})^{-1}_{s,t}$, $s, t \in \{1, 2, \ldots, T\}$, $\mathbf{K}_G = (\mathbf{L} + \alpha\mathbf{I})^{-\beta}$, and $\mu = \mu_{\text{TV}}$.

From Example 2, we see that the GTRSS problem can be understood as using a specific kernel in the time domain. In the following, we show that since this kernel depends on the number of discrete time steps, it is sensitive to the sampling rate.

Consider the case where $\mathcal{V}$ is a singleton and $\mathcal{T} = [a, b]$ is a closed interval, so that the signal $f : \mathcal{V} \times \mathcal{T} \to \mathbb{R}$ can be identified with a signal $f : [a, b] \to \mathbb{R}$. Without loss of generality, let $[a, b] = [0, 1]$. Suppose f is evenly sampled with interval length Δ. We denote the kernel from Example 2 as $k_{\text{GTRSS}}(s, t; \Delta) = (\mathbf{D}_h\mathbf{D}_h^{\intercal} + \delta_o\mathbf{I})^{-1}_{\frac{s}{\Delta}, \frac{t}{\Delta}}$, where $s, t \in \{0, \Delta, 2\Delta, \ldots, 1\}$. This leads to the problem that the prior distribution assigned to the signal relies on the sampling frequency. According to the Bayesian interpretation (cf. Appendix C), by using this kernel, we have assumed a prior distribution on f. We now examine the cross-correlation of the prior between $f(0)$ and $f(1)$, i.e.,

$$\begin{aligned}\operatorname{corr}(f(0), f(1); \Delta) &:= \frac{\operatorname{cov}(f(0), f(1))}{\sqrt{\operatorname{var}(f(0)) \operatorname{var}(f(1))}}\\ &= \frac{k_{\text{GTRSS}}(0, 1; \Delta)}{\sqrt{k_{\text{GTRSS}}(0, 0; \Delta)k_{\text{GTRSS}}(1, 1; \Delta)}}.\end{aligned}$$

By calculating this quantity with different values of Δ, we find that it is highly related to the sampling frequency (see Fig. 2). Specifically, when the sampling frequency is large

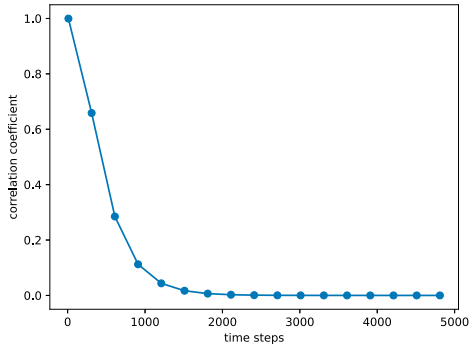


Fig. 2. The prior correlation coefficients $\mathrm{corr}(f(0), f(1); \Delta)$ as a function of $\frac{1}{\Delta} + 1$ (i.e., number of time steps) with $\delta_o = 10^{-5}$.

enough, the prior correlation between $f(0)$ and $f(1)$ tends to zero. Instead, if we use other kernels $k_{\mathcal{T}}$ which does not depend on Δ (e.g., the RBF kernel), then the prior cross-correlation $\frac{\tilde{k}_{\mathcal{T}}(a,b)}{\sqrt{\tilde{k}_{\mathcal{T}}(a,a)\tilde{k}_{\mathcal{T}}(b,b)}}$ does not depend on Δ. This accounts for the failure of GTRSS on datasets with high sampling frequency, while KRR-GGSP with RBF kernel works well (see Section V-B). This is essentially because the scale parameter in GTRSS kernel relies on the sampling frequency, while that in RBF kernel does not. Therefore the RBF kernel has one more degree of freedom than GTRSS kernel. Hence by using more flexible kernels, we can expect better reconstruction results.

C. Online and Distributed Implementation

We now consider the online learning problem where the data stream $\{(v_m, \mathbf{t}_m, y_m)\}$ arrives sequentially. Upon each arrival of $(v_m, \mathbf{t}_m)$, the learner provides a *distributed* prediction of $f(v_m, \mathbf{t}_m)$, i.e., the update and evaluation steps are implemented by each vertex exchanging information with its neighbors within a certain number of hops. After that, y_m is observed and the error is measured by comparing the prediction with y_m. The estimator of $f(v_m, \mathbf{t}_m)$ cannot depend on y_m, and the error is used to update the learner for the next prediction. Problem (5) can be adapted to this setting via RFFs [24] when $k_{\mathcal{T}}$ is a RBF kernel. Denote the columns of $\mathbf{K}_G^{\frac{1}{2}}$ by $[\mathbf{p}_1, \ldots, \mathbf{p}_N]$, and write $\mathbf{p}_v = (p_{1,v}, \ldots, p_{N,v})^\intercal$. For the kernel k, the RFF can be constructed as

$$\boldsymbol{\eta}(v, \mathbf{t}) = \mathbf{p}_v \otimes \mathbf{z}(\mathbf{t}),$$

where $\mathbf{z}(\mathbf{t}) \in \mathbb{R}^F$ is the RFF of the kernel $k_{\mathcal{T}}$, i.e., $\mathbb{E}[\mathbf{z}(\mathbf{s})^\intercal \mathbf{z}(\mathbf{t})] = k_{\mathcal{T}}(\mathbf{s}, \mathbf{t})$. By the construction of $\boldsymbol{\eta}(v, \mathbf{t})$, we have $\mathbb{E}[\boldsymbol{\eta}(u, \mathbf{s})^\intercal \boldsymbol{\eta}(v, \mathbf{t})] = k((u, \mathbf{s}), (v, \mathbf{t}))$. The reconstructed signal is then $\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \mathbf{c}^\intercal \boldsymbol{\eta}(v, \mathbf{t})$. Problem (5) is therefore converted to the linear regression problem [23, (7)]:

$$\min_{\mathbf{c} \in \mathbb{R}^{NF}} q(\mathbf{c}) = \sum_{m=1}^{M} (\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)^2 + \mu \|\mathbf{c}\|_2^2. \tag{15}$$

Alternatively, if we define $q_m(\mathbf{c}) := (\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)^2 + \frac{\mu}{M}\|\mathbf{c}\|_2^2$, then (15) turns out to be

$$\min_{\mathbf{c} \in \mathbb{R}^{NF}} q(\mathbf{c}) = \sum_{m=1}^{M} q_m(\mathbf{c}). \tag{16}$$

The evaluation of $\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \mathbf{c}^\intercal \boldsymbol{\eta}(v, \mathbf{t})$ can be distributed. To illustrate this, write $\mathbf{c} = (\mathbf{c}_1^\intercal, \ldots, \mathbf{c}_N^\intercal)^\intercal$ where $\mathbf{c}_n \in \mathbb{R}^F$, $n = 1, \ldots, N$. Since k_G takes the form (4), $\mathbf{K}_G^{\frac{1}{2}}$ can be represented as a polynomial of $\mathbf{A}_G$ of degree L_0, so that $p_{u,v} = 0$ for all $u \notin \mathcal{N}_{L_0}(v)$. Then for any input $(v, \mathbf{t})$, $\boldsymbol{\eta}(v, \mathbf{t}) = (p_{1,v}\mathbf{z}(\mathbf{t})^\intercal, \ldots, p_{N,v}\mathbf{z}(\mathbf{t})^\intercal)^\intercal$, $\hat{f}_{\mathrm{RFF}}$ is evaluated by

$$\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \sum_{u \in \mathcal{N}_{L_0}(v)} \mathbf{c}_u^\intercal p_{u,v} \mathbf{z}(\mathbf{t}),$$

which only requires information from $\mathcal{N}_{L_0}(v)$.

Problem (15) can be solved in an online and distributed way by stochastic gradient descent (SGD). To be specific, suppose the datastream is $\{(v_m, \mathbf{t}_m, y_m) : m = 1, 2, \ldots\}$. At the m-th step, we approximate ∇q with the instantaneous sample $(v_m, \mathbf{t}_m, y_m)$:

$$\nabla q_m = 2(\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)\boldsymbol{\eta}(v_m, \mathbf{t}_m) + 2\frac{\mu}{M}\mathbf{c}.$$

Note that $y_m - \mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) = y_m - \hat{f}_{\mathrm{RFF}}(v_m, \mathbf{t}_m) := \hat{e}_m$ is the approximation error at the current sample point $(v_m, \mathbf{t}_m)$. We can update $\mathbf{c}$ at the m-th iteration via

$$\mathbf{c}^{(m)} = \mathbf{c}^{(m-1)} - \theta \nabla q_m = \theta_1 \mathbf{c}^{(m-1)} + \theta_2 \hat{e}_m \boldsymbol{\eta}(v_m, \mathbf{t}_m), \tag{17}$$

where $\theta, \theta_1, \theta_2 > 0$. Note that:

- ∇q_m is Lipschitz continuous with Lipschitz constant $\mathrm{Lip}_m = 2\|\boldsymbol{\eta}(v_m, \mathbf{t}_m)\|^2 + 2\frac{\mu}{M}$. Define $\mathrm{Lip}_{\max} = \max_m \mathrm{Lip}_m$.
- q_m is convex.
- q is 2μ-strongly convex (cf. [38, Lemma 2.12]).

According to [38, Theorem 5.7], if $\theta \in (0, \frac{1}{2\mathrm{Lip}_{\max}})$, the convergence rate of SGD is linear when $\mu > 0$. Since $\mathbf{p}_{v_m}$ only has non-zero entries in $\mathcal{N}_{L_0}(v_m)$, and $\hat{e}_m$ can be evaluated in a distributed way, we see that (17) is an online and distributed update. This is always achievable when $k_{\mathcal{T}}$ is a RBF kernel.

IV. Conditional MSE of KRR-GGSP in the Bayesian Framework

In this section, we consider $f \sim \mathcal{GP}(0, k)$, i.e., the Bayesian framework considered in Section III-B. We derive the MSE of the estimate given by KRR-GGSP at a particular node $v_0 \in \mathcal{V}$ and time $\mathbf{t}_0 \in \mathcal{T}$, conditioned on an observation set $\{(v_m, \mathbf{t}_m, y_m) : m = 1, \ldots, M\}$. To be specific, we analyze

$$\begin{aligned} &\mathrm{var}(f(v_0, \mathbf{t}_0) \mid \{(v_m, \mathbf{t}_m, y_m)\}) \\ &\quad = \mathbb{E}[(\hat{f}(v_0, \mathbf{t}_0) - f(v_0, \mathbf{t}_0))^2 \mid \{(v_m, \mathbf{t}_m, y_m)\}] \end{aligned} \tag{18}$$

under the scenario when the noise energy is unknown, and the MSE is hard to compute when $M \to \infty$ as it involves taking the inverse of the kernel matrix of the observations. We study the dependence of the MSE on the graph structure when a subset of vertices have dense observation samples ($M \to \infty$). The asymptotic MSE and its upper bound can be used as a criterion to choose an optimal sampling vertex set.

We consider the asymptotic MSE of inference for $f(v_0, \mathbf{t}_0)$, i.e., the limit of (18) when $M \to \infty$. Note that if we allow uniform sampling on every vertex with an ever-growing sample

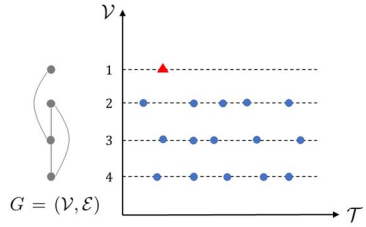


Fig. 3. The uniform exclusive sampling scheme with $M_0 = 5$. The blue circles denote $\mathcal{S}(M_0)$, and the red triangle is $(v_0, \mathbf{t}_0)$.

size, then it is known that the posterior variance will uniformly converge to 0 [39]. In order to examine the effect of leveraging information from other vertices in KRR-GGSP, we consider the case where there are no available sample points on $\{v_0\} \times \mathcal{T}$, and the value of $f(v_0, \mathbf{t})$ is to be estimated.

Mathematically, let $\mathcal{S}(v; M_0)$ be a set of M_0 samples i.i.d. from $\mathrm{Unif}(\{v\} \times \mathcal{T})$, where $v \in \{v_0\}^{\mathsf{c}}$. The sample set $\mathcal{S}(M_0)$ is then obtained by $\mathcal{S}(M_0) = \bigcup_{v \in \{v_0\}^{\mathsf{c}}} \mathcal{S}(v; M_0)$. This sampling scheme is illustrated in Fig. 3, and we call it *uniform exclusive sampling*. In practice, this scheme mimics the scene where only limited knowledge can be obtained from a certain vertex, and an inference for that is desired.

For ease of notation, we define $\mathcal{J}_S := \{v_0\}^{\mathsf{c}} \times \mathcal{T}$. We write $\mathbf{y}(M_0)$ to represent the observations $\mathbf{y}(\mathcal{S}(M_0))$ from the sampling set $\mathcal{S}(M_0)$, and $\mathbf{z}$ to represent the restriction of f on $\mathcal{J}_S$. We analyze $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ from two aspects: first, in Theorem 2 we analyze the integration of $\mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0))$ over $\mathbf{t}$; then in Theorem 3 we provide an asymptotic upper bound for $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$.

Let $\mathcal{T}_0$ be a subset of $\mathcal{T}$. We consider the following integration

$$\int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \, \mathrm{d}\tau(\mathbf{t}), \tag{19}$$

which represents the conditional MSE of the KRR-GGSP estimator over $\mathcal{T}_0$. Intuitively, when M_0 tends to infinity, the situation can be interpreted as f on $\mathcal{J}_S$ is known and can be utilized for inference. We formally address this in the following theorem:

Theorem 2: Under Assumption 1, the limit posterior covariance of $f(v_0, \mathbf{t})$ over $\mathcal{T}_0$ converges:

$$\begin{aligned} &\lim_{M_0 \to \infty} \int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \, \mathrm{d}\tau(\mathbf{t}) \\ &= \int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z}) \, \mathrm{d}\tau(\mathbf{t}). \end{aligned} \tag{20}$$

Proof: See Appendix D. □

From Theorem 2 we know the limiting posterior variance given an infinite number of sample points. This result can also be applied when only a subset of vertices have dense samples. In that case, the right-hand side (R.H.S.) of (20) becomes an asymptotic upper bound by letting $\mathbf{z}$ be the restriction of f on the vertices with dense samples. Moreover, we can get a rough idea of the behavior of $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ if we consider the following sequence of continuous functions

$$\rho_{M_0}(\alpha) := \begin{cases} \frac{1}{\alpha} \int_{B(\mathbf{t}_0, \alpha)} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \, \mathrm{d}\tau(\mathbf{t}), & \alpha > 0 \\ \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)), & \alpha = 0 \end{cases}$$

where $B(\mathbf{t}_0, \alpha)$ is the open ball centered at $\mathbf{t}_0$ with measure α. Specifically, by [39, Theorem 3] we note that $\rho_{M_0}(\alpha)$ is a monotonic sequence, i.e., $\rho_{M_0}(\alpha) \le \rho_{M_0'}(\alpha)$ if $M_0 > M_0'$. According to Theorem 2, the limit function of $\rho_{M_0}(\alpha)$ is

$$\rho(\alpha) = \lim_{M_0 \to \infty} \rho_{M_0}(\alpha) = \frac{1}{\alpha} \int_{B(\mathbf{t}_0, \alpha)} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z}) \, \mathrm{d}\tau(\mathbf{t})$$

when $\alpha > 0$, and

$$\rho(0) = \lim_{M_0 \to \infty} \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)).$$

Therefore, if we assume that the limit function of $\rho_{M_0}(\alpha)$ is continuous w.r.t. α and $\mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z})$ is continuous w.r.t. $\mathbf{t}$, then $\rho(0) = \lim_{\alpha \to 0} \rho(\alpha) = \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z})$, i.e.,

$$\lim_{M_0 \to \infty} \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)) = \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}). \tag{21}$$

From (21) we know that, although $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ is random due to the randomness of $\mathcal{S}(M_0)$, its limit $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z})$ is a deterministic quantity when $M_0 \to \infty$. In addition, it can be shown by Lemma D.3 that

$$\begin{aligned} &\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q})) \\ &\quad = \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}) + \mathrm{var}(\mathbb{E}[f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}] \,|\, f(\mathcal{Q})) \\ &\quad \ge \mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}), \end{aligned}$$

for arbitrary finite set $\mathcal{Q} \subset \mathcal{J}_S$. Therefore, according to (21), $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q}))$ can always serve as an upper bound for $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ when M_0 is large enough. Since $\mathcal{Q}$ is finite, $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q}))$ may be numerically computed. In contrast, We note that the quantities in (20) involve the pseudo-inverse of a possibly infinite-rank operator (cf. Lemma D.5), which may be difficult to numerically compute. Consider the case when $\mathcal{Q} = \mathcal{N}_d(v_0) \times \{\mathbf{t}_0\}$ where $d \in \mathbb{N}$ is the number of neighborhood hops. Let $N_d := |\mathcal{N}_d(v_0)|$. For simplicity, we introduce the following notations:

$$\begin{aligned} \mathbf{k}_G(v_0, \mathcal{N}_d) &:= (k_G(v_0, v))_{v \in \mathcal{V} \setminus \{v_0\}} \in \mathbb{R}^{N_d} \\ \mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d) &:= (k_G(u, v))_{u, v \in \mathcal{V} \setminus \{v_0\}} \in \mathbb{R}^{N_d \times N_d} \\ l(v_0, d) &:= k_G(v_0, v_0) \\ &\quad - \mathbf{k}_G(v_0, \mathcal{N}_d)^{\intercal} \mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1} \mathbf{k}_G(v_0, \mathcal{N}_d), \end{aligned}$$

so that

$$\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q})) = k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l(v_0, d).$$

To provide an explicit upper bound for (18), we derive an asymptotic bound with a convergence rate for the posterior variance which is locally computable.

Theorem 3: Suppose $\mathcal{T}$ is a compact subset of $\mathbb{R}^D$ whose boundary set has measure zero, and $\mathbf{t}_0$ is an interior point of $\mathcal{T}$. Suppose $k_{\mathcal{T}}$ is Lipschitz continuous on $\mathcal{T}$. Let $d \in \mathbb{N}_+$.

For any arbitrary $c_0 \in (0,1)$ we have

$$\begin{aligned}\operatorname{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{y}(M_0)) \le\ & k_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}_0)l(v_0,d)\\ &+ (C_1 c_0^{-1} + C_2 c_0^2) M_0^{-\frac{1}{3D+1}}\\ &+ C_3 c_0 M_0^{-\frac{2}{3D+1}} \end{aligned} \tag{22}$$

with probability at least

$$\left(1 - \frac{1}{2}\frac{1}{(1-c_0)^2 C_D M_0^{\frac{1}{3D+1}}}\right)^{N_d}. \tag{23}$$

Proof: As the proof is tedious and technical in nature, it is provided in Section SIII in the supplementary. □

We note that when $k_{\mathcal{T}}$ is RBF kernel, $k_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}_0)l(v_0,d)$ only depends on the graph structure. In other words, if we are allowed to select a subset of vertices $\mathcal{V}' \subset \mathcal{V}$ to recover the signal on v_0, then it is preferred that the subgraph with vertex set $\mathcal{V}' \bigcup \{v_0\}$ has a small $l(v_0,d)$. Theorem 3 indicates a trade-off between the quality and confidence of the upper bound (22). From the proof of Theorem 3, when the number of samples in a small neighborhood of every $(v,\mathbf{t}_0) \in \mathcal{N}_d(v_0) \times \{\mathbf{t}_0\}$ is larger than a threshold m_0, the conditional variance of $f(v_0,\mathbf{t}_0)$ given these samples is approximately $\operatorname{var}(f(v_0,\mathbf{t}_0)\,|\,f(\mathcal{Q}))$. However, the probability that this event happens is smaller when we require more vertices to at least m_0 samples in their neighborhoods. This explains why the probability lower bound (23) decreases as N_d increases. On the other hand, for a fixed $(v_0,\mathbf{t}_0)$, a larger $\mathcal{N}_d(v_0)$ indicates a better asymptotic upper bound for $\operatorname{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{y}(M_0))$, i.e, $l(v_0,d)$ decreases with a larger $\mathcal{N}_d(v_0)$. This is because $l(v_0,d)$ is a conditional variance of a Gaussian random variable by definition, and it is known that when we condition on a larger set of Gaussian random variables, the variance decreases [39, Lemma 9].

V. Numerical Experiments

In this section, we conduct experiments to illustrate the theory and methods of the KRR-GGSP approach. In the experiments, $\mathcal{T}$ is an interval, and the target signal is a function on $\mathcal{V} \times \mathcal{T}$. In the datasets, the target signal is downsampled on every vertex. We aim to reconstruct the target signal from the randomly selected samples with additive noise. We compare the following algorithms in the experiments:

1) KRR-GGSP. We reconstruct the signal using (5) with the tensor product kernel (3). We set $\mathbf{K}_G = a(\mathbf{L} - \lambda_N \mathbf{I})^2 + b\mathbf{I}$ such that
$$a(\lambda_1 - \lambda_N)^2 + b = 1, \tag{24}$$
and $0 \le b \le 1$ is a tunable parameter. This parameter setting ensures that $1 = r(\lambda_1) \ge \cdots \ge r(\lambda_N) = b$ (cf. (4)). We set $k_{\mathcal{T}}$ to be the RBF kernel $k_{\mathcal{T}}(s,t) = \exp(-|s-t|^2/\beta_{\text{scale}})$, where β_{scale} is a tunable parameter.
2) Isolated KRR. We recover the signal on each vertex separately using KRR (cf. (36) and (8)). In Section III, we have shown that this method is equivalent to using $\mathbf{K}_G = \mathbf{I}$ in KRR-GGSP, i.e., fixing $b = 1$ in (24).
3) GTRSS. We recover the signal using (13), where μ_{TV}, α and β are tunable parameters.
4) Graph recurrent imputation network (GRIN). We implement this method using the Spatiotemporal library [40].
5) Bandlimited-GGSP. We recover the signal using (11), where B', B'' and μ are tunable parameters. The eigenvalues $r(\lambda_n)$ and γ_i in (11) are set to be 1.

A. ECoG Dataset

We test the reconstruction performance of KRR-GGSP on an ECoG multivariate time series dataset.[4] This dataset contains measurements from 76 electrodes on an epilepsy patient during both ictal and pre-ictal periods [41]. We make use of the data from 2 ictal periods. Each period lasts 10 seconds with a sampling rate of 400 Hz. Therefore, the dataset we use is a 76×8000 matrix. We use the last 320 time steps for testing and the 160 time steps before the test set for training. We add additive white Gaussian noise (AWGN) to the dataset and randomly mask the data so that both training and test sets are incomplete and noisy. We test the recovery performances of KRR-GGSP, GTRSS, isolated KRR and GRIN on this dataset.

Except for the isolated KRR method, all other methods rely on a graph structure. To construct the graph, we first use the isolated KRR to roughly reconstruct the unknown signal values on 160 time steps in the training set, and then calculate the correlation coefficients of these recovered data. We regard two electrodes as connected if the correlation coefficients between them are larger than 0.5. We set the edge weights to be the correlation coefficients. For GRIN, the training set is used for model training and validation. Besides the small training set with 160 time steps, we also show its performance trained on all available training data from the dataset, i.e., 7680 time steps. For other methods, the training set is used for tuning parameters. The recovery performance is measured by the relative error

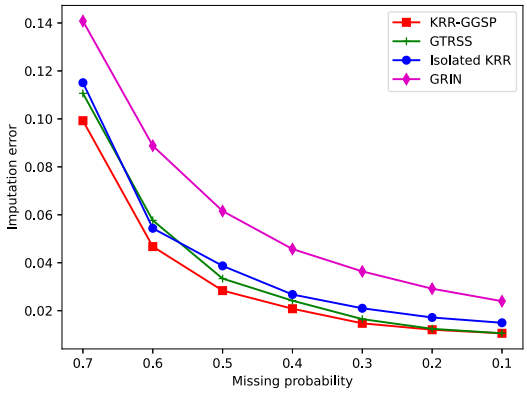
$$\frac{\mathbb{E}[(f(v,t) - \hat{f}(v,t))^2]}{\mathbb{E}[f(v,t)^2]}. \tag{25}$$

Similarly, we define the noise level to be

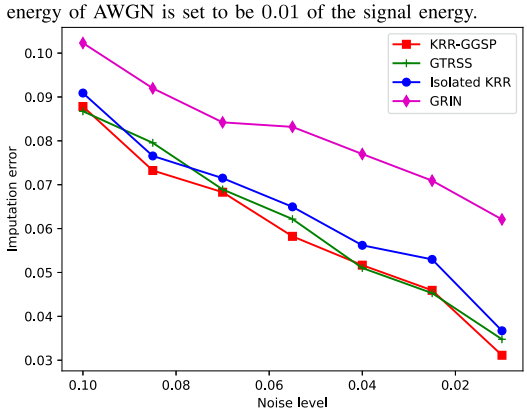
$$\frac{\mathbb{E}[\epsilon^2]}{\mathbb{E}[f(v,t)^2]}. \tag{26}$$

The recovery results are shown in Fig. 4. We observe that KRR-GGSP shows good recovery results and outperforms other methods. Since KRR-GGSP has a tunable kernel in the time domain, it shows better performance than GTRSS. This effect can be better observed in Section V-B. The isolated KRR method has a tunable kernel, but it is not able to take advantage of the graph structure, hence is outperformed by KRR-GGSP. Here, we show the performance of GRIN trained with 7680 time steps. We remark that the deep learning method GRIN requires a sufficiently large training set to obtain reasonable results. When the training set is as small as 160 time steps, GRIN does not yield reasonable reconstruction results. Since the bandlimited-GGSP method does not have comparable performances with

[4] https://math.bu.edu/people/kolaczyk/datasets.html



(a) Reconstruction performances under different missing value probabilities. The noise energy of AWGN is set to be 0.01 of the signal energy.



(b) Reconstruction performances under different noise levels (cf. (26)). The missing value probability is set to be 0.5.

Fig. 4. Comparison of different reconstruction methods on ECoG dataset. Each point in the figure is obtained by 20 repetitions.

the other methods (when noise level = 0.01 and missing value probability = 0.5, its imputation error is 0.28), we do not show its performance here.

B. Intel-Lab Temperature Data

We test the reconstruction performance of KRR-GGSP on the Intel lab temperature dataset illustrated in Fig. 1. In this experiment, we use the data from the first and second days. Since there are 86400 seconds in a day, the entire dataset we use is a 54×172800 matrix. Here we remark that since the sampling rate of each sensor is much smaller than 1 Hz and not uniform, only 1.93% of the entries are non-null. Therefore, this dataset is very sparse. We identify the temperature records outside the upper 99.92% quantile and lower 0.001% quantile as outliers and discard them. We subtract the mean value of all observed temperature records from the dataset. We treat each sensor as a vertex and construct a 5-NN graph using their locations. We use half of the first day's records for training and the second day's for testing. As in Section V-A, we add AWGN to the data and assign a random mask. In this experiment, the noise energy is set to be 5% of the signal energy. We compare the methods as described in Section V-A with performance measurement (25).

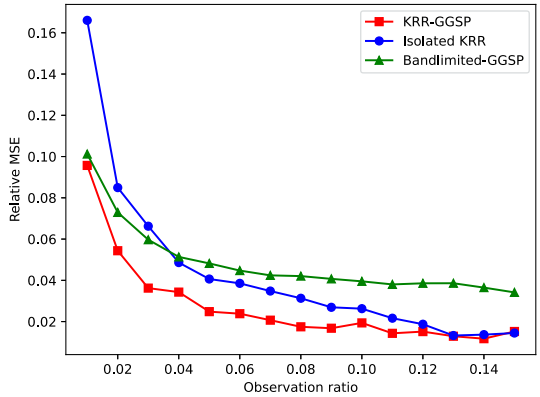


Fig. 5. Reconstruction error under different proportions of samples to be used for reconstruction. Each point in the figure is obtained by 10 repetitions.

From the result in Fig. 5, we observe that KRR-GGSP outperforms the isolated KRR and bandlimited-GGSP. This indicates that by utilizing infinitely many features, the reconstruction performance can be improved. On this dataset, GRIN and GTRSS fail to yield reasonable results. For example, when the observation ratio is 0.15, GTRSS has relative MSE around 0.8, and GRIN has relative MSE around 1.0. For GRIN, this is mainly due to the sparsity of the available data in the dataset. For GTRSS, this is due to the improper prior assumption on the dataset.

C. COVID-19 Case Prediction

We use the online reconstruction method in Section III-C to predict COVID-19 cases using only historical data. We use the data from The New York Times, based on reports from state and local health agencies[5]. From this dataset, we retrieve the records from California's 58 counties, starting from the first day when all counties have cases reported so that there are 886 days in total. We treat each county as a vertex and connect them if they are adjacent geographically. We set the datastream and prediction rule as follows: on each date t, we randomly choose a subset of vertices $\mathcal{V}_S = \{v_1, \ldots, v_Q\} \subset \mathcal{V}$ such that the learner is assumed to have access to $\mathbf{y}(\mathcal{V}_S \times \{t\})$. Besides, for each date t, the sample points $\{(v_i, t, y(v_i, t))\}$ are observed sequentially, one datum at a time.

We compare the online KRR-GGSP with several existing online and distributed reconstruction methods. The implementation details are the following:

1) Online KRR-GGSP. For each $(v_i, t) \in \mathcal{V}_S \times \{t\}$, we first calculate the prediction $\hat{f}_{\text{RFF}}(v_i, t)$. Then we compute the error $\hat{e}_i = y(v_i, t) - \hat{f}_{\text{RFF}}(v_i, t)$, and update the predictor by (17). Then for each $(v_j, t) \in \mathcal{V}_S^{\text{c}} \times \{t\}$, we also make predictions and compute the error, but will not update the predictor since the learner is not supposed to have access to the observations on them. We set $\mathbf{K}_G = g(\mathbf{L})^2$, where g is a polynomial of degree one such that $g(\lambda_1) = 1, g(\lambda_N) = 0.4$. We let $k_{\mathcal{T}}(s, t) =$

[5]https://github.com/TorchSpatiotemporal/tsl

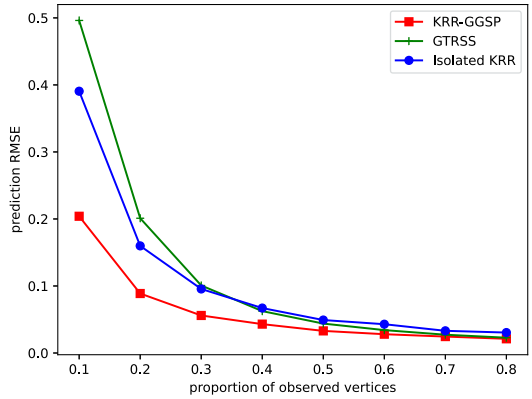


Fig. 6. Prediction error under different proportions of vertices to be sampled for learning. Each point in the figure is obtained by 10 repetitions.

$\exp(-(s-t)^2/\beta_{\text{scale}})$, where β_{scale} is an adjustable parameter. We set the dimension of $\mathbf{z}(t)$ to be 60.

2) Online isolated KRR. This is implemented by letting $\mathbf{K}_G = \mathbf{I}$ in the online KRR-GGSP method.
3) Online GTRSS. This method is a generalization of [14, (35)], by replacing $\mathbf{L}$ with $(\mathbf{L}+\alpha\mathbf{I})^\beta$. Let $\hat{\mathbf{f}}_t^l \in \mathbb{R}^N$ be the estimation of $\mathbf{f}_t = (y(1,t),\ldots,y(N,t))^\intercal$ after observing l samples on date t. The samples are denoted by $\mathbf{y}_t^l \in \mathbb{R}^N$ such that the unobserved entries are zero. We write $\mathbf{m}_t^l$ to denote the mask after observing l samples on date t. Let $\hat{\mathbf{f}}_{t-1}$ be the estimation of $\mathbf{f}_{t-1}$ after observing all available samples on date $t-1$. Then the update rule goes as follows:

$$\hat{\mathbf{f}}_t^l = \hat{\mathbf{f}}_t^{l-1} - \mu(\mathbf{m}_t^l \odot \hat{\mathbf{f}}_t^{l-1} - \mathbf{y}_t^l) - \mu\lambda(\mathbf{L}+\alpha\mathbf{I})^\beta(\hat{\mathbf{f}}_t^{l-1} - \hat{\mathbf{f}}_{t-1}). \tag{27}$$

When the $l+1$-th sample arrives, we evaluate the error $\hat{e}_{l+1} = y(v_{l+1},t) - \hat{f}(v_{l+1},t)$, where $\hat{f}(v_{l+1},t)$ is the v_{l+1}-th entry of $\hat{\mathbf{f}}_t^l$. λ, μ, α and β are adjustable parameters in this method.

We show the best performance of the methods with different parameters in Fig. 6. The error measurement is (25). We observe that the online KRR-GGSP method outperforms other online and distributed methods. We also tested the ARMA method on each vertex, but due to the missing values, it usually fails to converge and yields unstable results. For example, when the proportion of observed vertices is 80%, the ARMA$(2,0,2)$ model fails to converge on about 29% vertices, and the prediction error on each vertex varies from 0.004 to 665×10^4.

VI. Conclusion

In this paper, we devised a signal reconstruction approach for GGSP, yielding a predictor that can be computed in a distributed fashion. We interpreted this approach in both deterministic and Bayesian aspects and cast it as an extension of existing frameworks. In the former case where the signal is a deterministic function, we showed that the approach imposes smoothness on the reconstructed signal. In the latter case, the signal is regarded as a GP, and we analyzed its moments. By utilizing RFF, the reconstruction approach can be implemented online, and the evaluation is still distributed.

We provided statistical analysis on the predictor. Under the uniform exclusive sampling scheme, we derived the limit of the posterior variance and provided a numerically computable upper bound for it. We verified the KRR-GGSP approach by numerical experiments. By testing KRR-GGSP against existing methods on real datasets, we validated that introducing the graph structure and the product kernel improves reconstruction performance.

Appendix A
Preliminaries: GGSP

In GSP theory, typical choices of the GSO are the adjacency matrix, Laplacian matrix $\mathbf{L}$, and their normalized versions. We assume a normal GSO denoted as $\mathbf{A}_G$. Let $\mathbf{A}_G = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^\intercal$ be the eigendecomposition of $\mathbf{A}_G$, where $\mathbf{\Phi} = [\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_N]$ consists of orthonormal eigenvectors and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_N)$. Without loss of generality, we assume that $\{\lambda_i\}$ is indexed in increasing order of the graph frequencies, i.e., $\boldsymbol{\phi}_N$ is the eigenvector with the highest frequency. The GFT is then defined as the Euclidean inner product with the orthonormal basis $\mathbf{\Phi}$, i.e., the operator $\mathbf{\Phi}^\intercal$. In GGSP, due to the additional structure in $\mathcal{H}$, we further assume a shift operator (compact linear transformation) $\mathbf{A}_\mathcal{H}$ on $\mathcal{H}$. The shift operator $\mathbf{S}$ on $\mathbb{R}^N \otimes \mathcal{H}$ is then defined as $\mathbf{S} := \mathbf{A}_G \otimes \mathbf{A}_\mathcal{H}$. In $L^2(\mathcal{J})$, $\mathbf{S}$ operates as follows:

$$\mathbf{S} : L^2(\mathcal{J}) \to L^2(\mathcal{J})$$
$$f(v,\mathbf{t}) \mapsto \mathbf{S}(f)(v,\mathbf{t}) = \sum_{n=1}^N \mathbf{A}_G(v,n)\mathbf{A}_\mathcal{H}(f(n,\cdot))(\mathbf{t}),$$

Suppose we are given a complete orthonormal basis $\{\psi_i : i \geq 1\} \subset L^2(\mathcal{T})$. On the space $L^2(\mathcal{J})$, the JFT is defined as follows: for $n = 1,\ldots,N$ and $i \geq 1$,

$$\mathfrak{F}_{n,i} : L^2(\mathcal{J}) \to \mathbb{R}$$
$$f \mapsto \sum_{n'=1}^N \int_\mathcal{T} f(n',\mathbf{t})\phi_n(n')\psi_i(\mathbf{t})\,\mathrm{d}\tau(\mathbf{t}), \tag{28}$$

where $\phi_n(n')$ is the n'-th element of $\boldsymbol{\phi}_n$. Using the JFT, the signal is decomposed in the joint frequency domain indexed by $\{(n,i) : n = 1\ldots,N, i \geq 1\}$.

Appendix B
Preliminaries: Random Elements

In order to analyze the case where f is a stochastic process indexed by $(v,\mathbf{t})$, we model f as a random element [32], [42], [43]. Consider a probability space $(\Omega,\mathcal{F},\mu)$ where $\mathcal{F}$ stands for its σ-algebra, and a real separable Hilbert space $\mathcal{H}$ with its norm-induced Borel σ-algebra $\mathcal{B}$. A random element is defined as a measurable map $\mathbf{w} : \Omega \mapsto \mathcal{H}$, which induces a probability measure $\mathbb{P}$ on $(\mathcal{H},\mathcal{B})$ given by

$$\mathbb{P}(B) = \mu(\mathbf{w}^{-1}(B)),\ \forall B \in \mathcal{B}.$$

Assume that $\mathbb{E}[\|\mathbf{w}\|] < \infty$. The mean of $\mathbf{w}$ is defined as the element $m_\mathbf{w} \in \mathcal{H}$ such that

$$\langle m_\mathbf{w}, \mathbf{h}\rangle = \mathbb{E}[\langle \mathbf{w}, \mathbf{h}\rangle], \forall \mathbf{h} \in \mathcal{H}.$$

Assume that $\mathbb{E}[\|\mathbf{w}\|^2] < \infty$. The covariance of $\mathbf{w}$ is defined as the operator $\mathbf{C}_{\mathbf{ww}}$ on $\mathcal{H}$ such that

$$\langle \mathbf{C}_{\mathbf{ww}}\mathbf{h}, \mathbf{h}'\rangle = \mathbb{E}[\langle \mathbf{w} - m_{\mathbf{w}}, \mathbf{h}\rangle\langle \mathbf{w} - m_{\mathbf{w}}, \mathbf{h}'\rangle], \forall \mathbf{h}, \mathbf{h}' \in \mathcal{H}.$$

In this paper we alternatively write $m_{\mathbf{w}}$ and $\mathbf{C}_{\mathbf{ww}}$ as $\mathbb{E}[\mathbf{w}]$ and $\mathrm{cov}(\mathbf{w})$. It can be shown that $\mathrm{cov}(\mathbf{w})$ is always compact, self-adjoint, positive semi-definite and trace-class [43, Theorem 7.2.5]. For a pair of random elements $(\mathbf{w}_1, \mathbf{w}_2) : \Omega \to \mathcal{H}_1 \times \mathcal{H}_2$ which satisfies $\mathbb{E}[\|(\mathbf{w}_1, \mathbf{w}_2)\|^2] < \infty$, their cross-covariance operator is defined as the operator $\mathbf{C}_{\mathbf{w}_1\mathbf{w}_2} : \mathcal{H}_2 \to \mathcal{H}_1$ such that

$$\langle \mathbf{C}_{\mathbf{w}_1\mathbf{w}_2}\mathbf{h}_2, \mathbf{h}_1\rangle = \mathbb{E}[\langle \mathbf{w}_1 - \mathbb{E}[\mathbf{w}_1], \mathbf{h}_1\rangle\langle \mathbf{w}_2 - \mathbb{E}[\mathbf{w}_2], \mathbf{h}_2\rangle],$$

for all $\mathbf{h}_1 \in \mathcal{H}_1$, $\mathbf{h}_2 \in \mathcal{H}_2$. We alternatively write $\mathbf{C}_{\mathbf{w}_1\mathbf{w}_2}$ as $\mathrm{cov}(\mathbf{w}_1, \mathbf{w}_2)$. The mean element, covariance operator and cross-covariance operator can be alternatively defined by Bochner integral [43].

Let $\mathbf{h}_1 \in \mathcal{H}_1$ and $\mathbf{h}_2 \in \mathcal{H}_2$. We define $\mathbf{h}_1 \circledast \mathbf{h}_2$ as the following linear operator

$$\begin{aligned} \mathbf{h}_1 \circledast \mathbf{h}_2 : \mathcal{H}_2 &\to \mathcal{H}_1 \\ \mathbf{h} &\mapsto \langle \mathbf{h}, \mathbf{h}_2\rangle \mathbf{h}_1. \end{aligned}$$

Note that $\mathbf{h}_1 \circledast \mathbf{h}_2$ is in the space of Hilbert-Schmidt operators from $\mathcal{H}_2$ to $\mathcal{H}_1$, which is a Hilbert space [43, Theorem 4.4.5]. Then $\mathbf{C}_{\mathbf{w}_1\mathbf{w}_2}$ can be equivalently defined as $\mathbb{E}[(\mathbf{w}_1 - \mathbb{E}[\mathbf{w}_1]) \circledast (\mathbf{w}_2 - \mathbb{E}[\mathbf{w}_2])]$. The conditional expectation and covariance of a random element are defined as follows [42, Section II.4.1], [44]:

Definition B.1: Suppose the random element $\mathbf{w}$ takes values in a separable Hilbert space $\mathcal{H}$, $\mathbb{E}[\|\mathbf{w}\|] < \infty$, and $\mathcal{F}'$ is a sub σ-algebra of $\mathcal{F}$. The conditional expectation of $\mathbf{w}$ w.r.t. $\mathcal{F}'$ is the random element $\mathbf{w}_{\mathrm{cond}} \in \mathcal{F}'$ such that $\mathbb{E}[\|\mathbf{w}_{\mathrm{cond}}\|] < \infty$ and

$$\mathbb{E}[\mathbf{w}_{\mathrm{cond}} I_A] = \mathbb{E}[\mathbf{w} I_A], \ \forall A \in \mathcal{F}', \tag{29}$$

where I_A is the indicator function on the set A. We denote $\mathbf{w}_{\mathrm{cond}}$ by $\mathbb{E}[\mathbf{w} \,|\, \mathcal{F}']$. According to [42, Proposition 4.1], $\mathbb{E}[\mathbf{w} \,|\, \mathcal{F}']$ always exists.

The conditional covariance is defined as

$$\begin{aligned} &\mathrm{cov}(\mathbf{w}_1, \mathbf{w}_2 \,|\, \mathcal{F}') \\ &= \mathbb{E}[(\mathbf{w}_1 - \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}']) \circledast (\mathbf{w}_2 - \mathbb{E}[\mathbf{w}_2 \,|\, \mathcal{F}']) \,|\, \mathcal{F}'] \end{aligned}$$

We write $\mathrm{cov}(\mathbf{w}, \mathbf{w} \,|\, \mathcal{F}')$ as $\mathrm{cov}(\mathbf{w} \,|\, \mathcal{F}')$ for simplicity.

By the defining property (29) of conditional expectation it can be shown that

$$\begin{aligned} \langle \mathbb{E}[\mathbf{w} \,|\, \mathcal{F}'], \mathbf{h}\rangle &= \mathbb{E}[\langle \mathbf{w}, \mathbf{h}\rangle \,|\, \mathcal{F}'], \\ \langle \mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'](\mathbf{h}_2), \mathbf{h}_1\rangle &= \mathbb{E}[\langle \mathbf{w}_1, \mathbf{h}_1\rangle\langle \mathbf{w}_2, \mathbf{h}_2\rangle \,|\, \mathcal{F}'], \end{aligned} \tag{30}$$

for all $\mathbf{h} \in \mathcal{H}$, $\mathbf{h}_1 \in \mathcal{H}_1$, $\mathbf{h}_2 \in \mathcal{H}_2$. From (30) we know that $\mathbb{E}[\mathbf{w} \,|\, \mathcal{F}']$ is uniquely defined. Let $\mathcal{F}''$ be a sub σ-algebra of $\mathcal{F}'$. Like random variables, the random elements also satisfy the property [42, Section II.4.1]:

$$\mathbb{E}[\mathbb{E}[\mathbf{w} \,|\, \mathcal{F}'] \,|\, \mathcal{F}''] = \mathbb{E}[\mathbf{w} \,|\, \mathcal{F}''].$$

Let $(\mathcal{I}, \mathcal{F}_{\mathcal{I}}, \mu_{\mathcal{I}})$ be a σ-finite measure space. The stochastic process $\{f(\omega, \boldsymbol{\xi}) : \omega \in \Omega, \boldsymbol{\xi} \in \mathcal{I}\}$ can be modeled as a random element if it satisfies regularity conditions:

Theorem B.1: [45, Theorem 2] Suppose
1) f is a $\mu \times \mu_{\mathcal{I}}$-measurable stochastic process.
2) the paths of f are in $L^2(\mathcal{I})$.
Then the map

$$\begin{aligned} \Omega &\to L^2(\mathcal{I}) \\ \omega &\mapsto f(\omega, \cdot) \end{aligned} \tag{31}$$

is a random element with mean element $\mathbb{E}[f(\boldsymbol{\xi})] \in L^2(\mathcal{I})$. Its covariance operator $\mathbf{C}_f$ is the integral operator with kernel $\mathrm{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2))$. Specifically, if f is GP, then (31) is a Gaussian random element, i.e., composing any linear functional with it will yield a Gaussian random variable.

If we further assume that $\mathcal{I}$ is a compact metric space and $\mu_{\mathcal{I}}$ is a strictly positive Borel measure, and the function $\mathrm{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2))$ is continuous on $\mathcal{I} \times \mathcal{I}$, then it can be shown by Mercer's theorem [35] that

$$\mathrm{tr}(\mathbf{C}_f) = \int_{\mathcal{I}} \mathrm{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2)) \, \mathrm{d}\mu_{\mathcal{I}}. \tag{32}$$

In this paper we will make use of the following theorem which is more general than Theorem B.1. The proof of it is included in Section S1 in the supplementary for completeness.

Theorem B.2: Suppose a stochastic process f satisfies condition 1 and condition 2 in Theorem B.1. $\mathcal{F}'$ is a sub σ-algebra of the underlying probability space. Suppose f and $\mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'] \in L^2(\Omega \times \mathcal{I})$. Then

$$\mathbb{E}[f \,|\, \mathcal{F}'] = \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], \tag{33}$$

$$\begin{aligned} &\mathrm{cov}(f \,|\, \mathcal{F}') : L^2(\mathcal{I}) \to L^2(\mathcal{I}), \\ &\qquad g(\cdot) \mapsto \int_{\mathcal{I}} \mathrm{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) | \mathcal{F}') g(\boldsymbol{\xi}_2) \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_2). \end{aligned} \tag{34}$$

Let $\mathrm{var}(f(\boldsymbol{\xi}) \,|\, \mathcal{F}')$ be the conditional variance of the random variable $f(\boldsymbol{\xi})$. If we further assume that $\mathcal{I}$ is a compact metric space, and $\mathrm{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}')$ is continuous w.r.t. $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, then we have

$$\begin{aligned} \mathrm{tr}(\mathrm{cov}(f \,|\, \mathcal{F}')) &= \mathbb{E}[\|f - \mathbb{E}[f \,|\, \mathcal{F}']\|^2 \,|\, \mathcal{F}'] \\ &= \int_{\mathcal{I}} \mathrm{var}(f(\boldsymbol{\xi}) \,|\, \mathcal{F}') \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}). \end{aligned} \tag{35}$$

In the above formulas, the left-hand side (L.H.S.) are defined by moments of f as a random element. The moments in R.H.S. are defined pointwise, as functions on $\mathcal{I}$ or $\mathcal{I} \times \mathcal{I}$.

In this paper, the index set $\mathcal{I}$ can be $\mathcal{V} \times \mathcal{T}$ or a subset of $\mathcal{V} \times \mathcal{T}$. We always assume that the conditions in Theorem B.1 are met for the stochastic processes in concern. In this case, we call the stochastic process f as GRP [32]. In statistical GSP, a random graph signal is said to be WSS if its covariance commutes with $\mathbf{A}_G$ [46], [47]. Analogously, in the GGSP framework, a GRP f is said to be JWSS if its covariance operator $\mathbf{C}_f$ commutes with $\mathbf{S}$ [32].

Appendix C
Preliminaries: KRR Reconstruction and Interpretation

KRR is a supervised learning approach that aims to learn a map from $\mathcal{X}$ to $\mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$. Given a set of training inputs

and outputs, it searches for the best fitting function in a RKHS. Given a symmetric positive semi-definite kernel

$$k : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$$
$$(\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}'),$$

the associated RKHS $\mathcal{H}_k$ is defined as the Hilbert space satisfying [34, Definition 1]:

1) $k(\cdot, \mathbf{x}) \in \mathcal{H}_k$ for all $\mathbf{x} \in \mathcal{X}$.
2) $\langle g, k(\cdot, \mathbf{x})\rangle_{\mathcal{H}_k} = g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and $g \in \mathcal{H}_k$.

According to the Moore-Aronszajn theorem [34, Theorem 3], there exists a unique Hilbert space $\mathcal{H}_k$ satisfying these conditions. When $\mathcal{X}$ is a subset of Euclidean space, typical choices for k include the polynomial kernel ($k(\mathbf{x}, \mathbf{x}') = (a\mathbf{x}^\intercal\mathbf{x}' + 1)^b$ with parameters $a \in \mathbb{R}$, $b \in \mathbb{N}$), linear kernel (polynomial kernel with $a = 1, b = 1$), and RBF kernel ($k(\mathbf{x}, \mathbf{x}')$ is a function of $\|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}}$).

Given a training set $\{(\mathbf{x}_m, y_m) : \mathbf{x}_m \in \mathcal{X}, y_m \in \mathcal{Y}, m = 1, \ldots, M\}$, KRR searches for an optimal function in $\mathcal{H}_k$ to fit the data by solving for

$$\hat{f} = \arg\min_{\tilde{f} \in \mathcal{H}_k} \sum_{m=1}^{M} |\tilde{f}(\mathbf{x}_m) - y_m|^2 + \mu J(\|\tilde{f}\|_{\mathcal{H}_k}), \tag{36}$$

where $J(\cdot)$ is an increasing function, and μ is a penalty weight. The representer theorem [48, Theorem 4.2] states that the optimal solution to (36) takes the form

$$\hat{f} = \sum_{m=1}^{M} c_m k(\cdot, \mathbf{x}_m), \tag{37}$$

where c_m, $m = 1, \ldots, M$, are coefficients to be determined. By substituting (37) into (36), the problem (36) becomes an optimization over $\{c_m\}_{m=1}^M$. Specifically, when $J(\cdot) = (\cdot)^2$, problem (36) is quadratic and its solution is given by

$$(c_1, \ldots, c_M)^\intercal = (\mathbf{K} + \mu\mathbf{I}_M)^{-1}\mathbf{y}, \tag{38}$$

where $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^M \in \mathbb{R}^{M\times M}$ and $\mathbf{y} = (y_1, \ldots, y_M)^\intercal$. In the sequel, we assume $J(\cdot) = (\cdot)^2$ unless otherwise stated. When k is chosen as the linear kernel, (36) is equivalent to learning a linear function from $\mathcal{X}$ to $\mathcal{Y}$, i.e., linear regression.

It is natural to consider whether we can recover any continuous function pointwise to within arbitrary fidelity with a sufficiently large number of samples by KRR. This is achievable by employing a *universal* kernel k [49]. Let $\mathcal{X}$ be a Hausdorff topological space (e.g., $\mathbb{R}$) and $\mathcal{Z} \subset \mathcal{X}$ be a compact subset (e.g., $[a, b]$). Let $\mathcal{C}(\mathcal{Z})$ be the space of continuous functions on $\mathcal{Z}$ with the supremum norm. Define $\mathcal{K}(\mathcal{Z}) := \overline{\text{span}}\{k(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}\}$, where the closure is taken w.r.t. the norm in $\mathcal{C}(\mathcal{Z})$. The kernel k is said to be universal if $\mathcal{K}(\mathcal{Z}) = \mathcal{C}(\mathcal{Z})$ for any compact $\mathcal{Z} \subset \mathcal{X}$. In other words, $\text{span}\{k(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}\}$ is dense in $\mathcal{C}(\mathcal{Z})$.

Problem (36) has a Bayesian interpretation. Consider a GP w with mean function zero and covariance function $k(\mathbf{x}, \mathbf{x}')$, denoted as $w \sim \mathcal{GP}(0, k)$. Given the noisy observations $y_m = w(\mathbf{x}_m) + \epsilon_m$, $\epsilon_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mu)$, the MAP estimator of $w(\mathbf{x})$ is $\hat{f}(\mathbf{x})$ as defined in (37) and (38) for any $\mathbf{x} \in \mathcal{X}$.

The readers are referred to [34], [50] for more detailed discussions on RKHS and KRR.

Appendix D
Proof of Theorem 2

In order to prove Theorem 2, we introduce the following definitions and lemmas. The proofs of the lemmas are included in the supplementary for completeness.

Let $\mathbf{x}_0$ be the restriction of f on $\{v_0\} \times \mathcal{T}_0$, and $\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}} := \text{cov}(\mathbf{x}_0 \,|\, \mathbf{y}(M_0))$. Note that (19) can be equally written as $\text{tr}(\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}})$ (cf. (32)). Based on this observation, we analyze the asymptotic behavior of $\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}}$.

We compute the covariance operators $\mathbf{C}_{\mathbf{zz}}$ and $\mathbf{C}_{\mathbf{zx}_0}$ for later use:

$$\begin{aligned}
\mathbf{C}_{\mathbf{zz}} &: L^2(\mathcal{J}_S) \to L^2(\mathcal{J}_S)\\
&\quad g(\cdot) \mapsto \int_{\mathcal{T}} \sum_{u\in\{v_0\}^c} k_G(v,u) k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(u,\mathbf{s})\, \mathrm{d}\tau(\mathbf{s}),\\
\mathbf{C}_{\mathbf{zx}_0} &: L^2(\mathcal{T}_0) \to L^2(\mathcal{J}_S)\\
&\quad g(\cdot) \mapsto \int_{\mathcal{T}_0} k_G(v_0,v) k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(v_0,\mathbf{s})\, \mathrm{d}\tau(\mathbf{s}).
\end{aligned} \tag{39}$$

Define the integral operators

$$\begin{aligned}
\mathbf{H} &: L^2(\mathcal{T}) \to L^2(\mathcal{T})\\
&\quad g(\cdot) \mapsto \int_{\mathcal{T}} k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(\mathbf{s})\, \mathrm{d}\tau(\mathbf{s}),\\
\mathbf{H}_0 &: L^2(\mathcal{T}_0) \to L^2(\mathcal{T})\\
&\quad g(\cdot) \mapsto \int_{\mathcal{T}_0} k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(\mathbf{s})\, \mathrm{d}\tau(\mathbf{s}).
\end{aligned}$$

Define $\mathbf{K}_{G,**}$ as the submatrix of $\mathbf{K}_G$ without the v_0-th row and the v_0-th column. Let $\mathbf{k}_{G,0*}$ be the v_0-th column of $\mathbf{K}_G$ but without the v_0-th entry. Then we have

$$\begin{aligned}
\mathbf{C}_{\mathbf{zz}} &= \mathbf{K}_{G,**} \otimes \mathbf{H},\\
\mathbf{C}_{\mathbf{zx}_0} &= \mathbf{k}_{G,0*} \otimes \mathbf{H}_0.
\end{aligned} \tag{40}$$

Lemma D.1: Suppose a sequence of operators $\{\mathbf{C}_n\}$ on a separable Hilbert space $\mathcal{H}$, all of which are compact, self-adjoint, positive semi-definite and trace-class. Suppose $\mathbf{J}$ is a bounded linear operator from $\mathcal{H}$ to $\mathcal{G}$, where $\mathcal{G}$ is also a separable Hilbert space. If $\lim_{n\to\infty} \text{tr}(\mathbf{C}_n) = 0$, then $\lim_{n\to\infty} \text{tr}(\mathbf{J}\mathbf{C}_n\mathbf{J}^*) = 0$.

Lemma D.2: Suppose $\mathbf{w}_1$ is a random element in $\mathcal{H}_1$, and $\mathbf{w}_2$ is a random element in $\mathcal{H}_2$. $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable Hilbert spaces. $\mathcal{F}'$ is a sub σ-algebra of the underlying probability space. Suppose $\mathbf{w}_2 \in \mathcal{F}'$, then we have

$$\begin{aligned}
\mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'] &= \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'] \circledast \mathbf{w}_2,\\
\mathbb{E}[\mathbf{w}_2 \circledast \mathbf{w}_1 \,|\, \mathcal{F}'] &= \mathbf{w}_2 \circledast \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'].
\end{aligned}$$

Using Lemma D.2 we can simplify the definition of conditional covariance operator as

$$\text{cov}(\mathbf{w}_1, \mathbf{w}_2 \,|\, \mathcal{F}') = \mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'] - \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'] \circledast \mathbb{E}[\mathbf{w}_2 | \mathcal{F}'].$$

Lemma D.3: We have

$$\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}} = \mathbb{E}[\text{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) \,|\, \mathbf{y}(M_0)] + \text{cov}(\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] \,|\, \mathbf{y}(M_0)).$$

Lemma D.4: Let $\mathbf{C}_{\mathbf{z}\,|\,\mathbf{y}}$ be the conditional covariance operator of $\mathbf{z}$ given $\mathbf{y}(M_0)$. Then $\lim_{M_0\to\infty} \text{tr}(\mathbf{C}_{\mathbf{z}\,|\,\mathbf{y}}) = 0$ almost surely.

Lemma D.5: The conditional expectation and covariance of $\mathbf{x}_0$ given $\mathbf{z}$ are as follows:

$$\begin{aligned}\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] &= (\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^{*}\mathbf{z},\\ \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) &= \mathbf{C}_{\mathbf{x}_0} - \mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}^{*},\end{aligned} \tag{41}$$

where the operator $\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}$ is bounded.

Proof of Theorem 2: We can rewrite $\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}}$ as follows:

$$\begin{aligned}\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}} &= \mathbb{E}[\operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) \,|\, \mathbf{y}(M_0)] + \operatorname{cov}(\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] \,|\, \mathbf{y}(M_0))\\ &= \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) + \operatorname{cov}(\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{z} \,|\, \mathbf{y}(M_0))\\ &= \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) + \mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{z}\,|\,\mathbf{y}}(\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^{*}.\end{aligned} \tag{42}$$

The first equality holds by Lemma D.3. The second equality holds by the fact that $\operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z})$ is deterministic. By taking trace and limit on (42) we have

$$\begin{aligned}&\lim_{M_0\to\infty} \operatorname{tr}(\mathbf{C}_{\mathbf{x}_0\,|\,\mathbf{y}} - \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}))\\ &= \lim_{M_0\to\infty} \operatorname{tr}(\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{z}\,|\,\mathbf{y}}(\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^{*}).\end{aligned}$$

From Lemma D.1 and Lemma D.4 we know that the R.H.S. tends to zero. By writing $\operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z})$ as (41) and using (34) we conclude the proof.

REFERENCES

[1] 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 Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[2] A. Ortega, P. Frossard, J. Kovacević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[3] J. D. Medaglia et al., "Functional alignment with anatomical networks is associated with cognitive flexibility," *Nature Human Behav.*, vol. 2, no. 2, pp. 156–164, 2018.

[4] W. Huang, L. Goldsberry, N. F. Wymbs, S. T. Grafton, D. S. Bassett, and A. Ribeiro, "Graph frequency analysis of brain signals," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 7, pp. 1189–1203, Oct. 2016.

[5] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018.

[6] A. C. Yagan and M. T. Özgen, "Spectral graph based vertex-frequency wiener filtering for image and graph signal denoising," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 226–240, 2020.

[7] W. Huang, A. G. Marques, and A. R. Ribeiro, "Rating prediction via graph signal processing," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5066–5081, Oct. 2018.

[8] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, Oct. 2020.

[9] Y. Tanaka and Y. C. Eldar, "Generalized sampling on graphs with subspace and smoothness priors," *IEEE Trans. Signal Process.*, vol. 68, pp. 2272–2286, 2020.

[10] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Aug. 2015.

[11] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3775–3789, Mar. 2016.

[12] L. F. O. Chamon and A. Ribeiro, "Greedy sampling of graph signals," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 34–47, Jan. 2018.

[13] J. Hara, Y. Tanaka, and Y. C. Eldar, "Graph signal sampling under stochastic priors," *IEEE Trans. Signal Process.*, vol. 71, pp. 1421–1434, 2023.

[14] K. Qiu, X. Mao, X. Shen, X. Wang, T. Li, and Y. Gu, "Time-varying graph signal reconstruction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 870–883, Sep. 2017.

[15] J. H. Giraldo, A. Mahmood, B. Garcia-Garcia, D. Thanou, and T. Bouwmans, "Reconstruction of time-varying graph signals via Sobolev smoothness," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 201–214, 2022.

[16] A. Kroizer, T. Routtenberg, and Y. C. Eldar, "Bayesian estimation of graph signals," *IEEE Trans. Signal Process.*, vol. 70, no. 5, pp. 2207–2223, 2022.

[17] A. Venkitaraman, S. Chatterjee, and P. Händel, "Gaussian processes over graphs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, Spain, May 2020, pp. 5640–5644.

[18] A. Venkitaraman, S. Chatterjee, and P. Händel, "Multi-kernel regression for graph signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 4644–4648.

[19] A. Venkitaraman, S. Chatterjee, and P. Händel, "Predicting graph signals using kernel regression where the input signal is agnostic to a graph," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 4, pp. 698–710, Dec. 2019.

[20] V. R. M. Elias, V. C. Gogineni, W. A. Martins, and S. Werner, "Kernel regression over graphs using random Fourier features," *IEEE Trans. Signal Process.*, vol. 70, pp. 936–949, 2022.

[21] V. R. M. Elias, V. C. Gogineni, W. A. Martins, and S. Werner, "Adaptive graph filters in reproducing kernel Hilbert spaces: Design and performance analysis," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 7, pp. 62–74, 2021.

[22] E. Isufi, G. Leus, and P. Banelli, "2-dimensional finite impulse response graph-temporal filters," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Washington, DC, USA, Dec. 2016, pp. 405–409.

[23] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic, "Towards a unified analysis of random Fourier features," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 4887–4937, Jul. 2021.

[24] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., Vancouver, Canada: Curran Associates, Inc., 2007.

[25] D. Romero, M. Ma, and G. B. Giannakis, "Estimating signals over graphs via multi-kernel learning," in *Proc. IEEE Workshop Statist. Signal Process.*, Palma de Mallorca, Spain, Jun. 2016, pp. 1–5.

[26] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 764–778, Feb. 2017.

[27] A. Loukas and D. Foucard, "Frequency analysis of time-varying graph signals," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Washington, DC, USA, Dec. 2016, pp. 346–350.

[28] F. Grassi, A. Loukas, N. Perraudin, and B. Ricaud, "A time-vertex signal processing framework: Scalable processing and meaningful representations for time-series on graphs," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 817–829, Nov. 2018.

[29] A. Loukas and N. Perraudin, "Stationary time-vertex signal processing," *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, pp. 1–19, Aug. 2019.

[30] F. Ji and W. P. Tay, "Generalized graph signal processing," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Anaheim, CA, USA, Nov. 2018, pp. 708–712.

[31] F. Ji and W. P. Tay, "A Hilbert space theory of generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6188–6203, Nov. 2019.

[32] X. Jian and W. P. Tay, "Wide-sense stationarity in generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 70, pp. 3414–3428, 2022.

[33] X. Jian and W. P. Tay, "Kernel ridge regression for generalized graph signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[34] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York, NY, USA: Springer-Verlag, 2011.

[35] I. Steinwart and C. Scovel, "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs," *Constructive Approximation*, vol. 35, pp. 363–417, Feb. 2012.

[36] G. Wahba, "Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind," *J. Approximation Theory*, vol. 7, no. 2, pp. 167–185, 1973.

[37] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2005.

[38] G. Garrigos and R. M. Gower, "Handbook of convergence theorems for (stochastic) gradient methods," 2023, *arXiv:2301.11235*.

[39] P. Koepernik and F. Pfaff, "Consistency of Gaussian process regression in metric spaces," *J. Mach. Learn. Res.*, vol. 22, no. 244, pp. 1–27, 2021.

[40] A. Cini and I. Marisca, "Torch Spatiotemporal," GitHub. Accessed: Mar. 2022. [Online]. Available: https://github.com/TorchSpatiotemporal/tsl
[41] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, "Emergent network topology at seizure onset in humans," *Epilepsy Res.*, vol. 79, no. 2, pp. 173–186, May 2008.
[42] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan, *Probability Distributions on Banach Spaces*. New York, NY, USA: Springer-Verlag, 1987.
[43] T. Hsing and R. Eubank, *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*. Hoboken, NJ, USA: Wiley, 2015.
[44] I. Klebanov, B. Sprungk, and T. Sullivan, "The linear conditional expectation in Hilbert space," *Bernoulli*, vol. 27, no. 4, pp. 2267–2299, Nov. 2021.
[45] B. S. Rajput and S. Cambanis, "Gaussian processes and Gaussian measures," *Ann. Math. Statist.*, vol. 43, no. 6, pp. 1944–1952, 1972.
[46] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5911–5926, Aug. 2017.
[47] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, Jul. 2017.
[48] B. Scholkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.
[49] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J. Mach. Learn. Res.*, vol. 7, no. 95, pp. 2651–2667, 2006.
[50] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, US: Springer-Verlag, 2009.

Xingchao Jian (Graduate Student Member, IEEE) received the B.Sc. degree in statistics from Nankai University, Tianjin, China, in 2020. He is currently working toward the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include graph signal processing and statistical machine learning.

Wee Peng Tay (Senior Member, IEEE) received the B.S. degree in electrical engineering and mathematics, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2002, and the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is a Professor of signal and information processing with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics, in 1995, and the B.Sc. degree in electrical engineering, in 1996, both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science, in 2002, from Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel, where she holds the Dorothy and Patrick Gorman Professorial Chair and heads the Center for Biomedical Engineering. She was previously a Professor in the Department of Electrical Engineering with the Technion, where she held the Edwards Chair in Engineering. She is a member of the Israel Academy of Sciences and Humanities and of the Academia Europaea (elected 2023), a EURASIP Fellow, a Fellow of the Asia-Pacific Artificial Intelligence Association, and a Fellow of the 8400 Health Network. Her research interests include statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging and optics.