Online Kernel-Based Graph Topology Identification with Partial-Derivative-Imposed Sparsity

Mircea Moscu[⋆] Ricardo Borsoi[⋆,†] Cédric Richard[⋆]

[⋆] Université Côte d'Azur, CNRS, OCA, Nice, France
[†] Federal University of Santa Catarina, Florianópolis, Brazil

***Abstract*—In many applications, such as brain network connectivity or shopping recommendations, the underlying graph explaining the different interactions between participating agents is unknown. Moreover, many of these interactions may be based on nonlinear relationships, rendering the topology inference problem more complex. This paper presents a new topology inference method that estimates a possibly directed adjacency matrix in an online manner. In contrast to previous approaches which are based on additive models, the proposed model is able to explain general nonlinear interactions between the agents. Partial-derivative-imposed sparsity is implemented, while reproducing kernels are used to model nonlinearities. The impact of the increasing number of data points is alleviated by using dictionaries of kernel functions. A comparison with a previously developed method showcases the generality of the new model.**

***Index Terms*—topology inference, partial derivative sparsity, brain connectivity, nonlinear interactions, kernels**

I. Introduction

In analyzing networks such as gene regulation systems [1], socio-economical interactions [2], or brain activity [3], graphs have proven to be a useful tool, given their inherently distributed nature. Most graph signal processing algorithms, however, assume the graph topology as known beforehand. Recently, significant interest has been dedicated to the estimation of the graph topology from available data. Most of these works assume linear dependencies between the agents, e.g., brain regions, genes in a network, or sectors of a market economy. However, the presence of nonlinear interactions in real-world applications imposes the need of developing more general algorithms. As such, the ability of reproducing kernels to model nonlinear relationships between nodal signals makes them a powerful tool to the graph inference process.

In this article, we consider a setting where online nodal measurements are acquired and used in order to infer the topology of the underlying network. In the developed approach, it is desired to estimate a possibly directed adjacency matrix while accounting for general nonlinear dependencies between nodal signals, and to distribute it over the different agents. Since many real world examples, such as social graphs, show considerable edge sparsity, a sparsity-inducing framework based on partial derivatives is employed.

Definitions: A graph $\mathcal{G}$ consists of a set $\mathcal{N}$ of $(N+1)$ nodes, and a set $\mathcal{E}$ of edges such that if nodes m and n are linked, then $(m,n) \in \mathcal{E}$. For undirected graphs, these node pairs are unordered. At node level, we collect a real-valued signal $\boldsymbol{y}(i) \triangleq [y_1(i), \ldots, y_{N+1}(i)]^\top$, where $y_n(i)$ is the sample of the signal $\boldsymbol{y}(i)$ at node n and time instant i. The adjacency matrix $\boldsymbol{A}$ [4], [5], is defined as an $(N+1) \times (N+1)$ matrix whose entries a_{nm} are zero if $(m,n) \notin \mathcal{E}$ and set to one otherwise. Throughout the paper we consider $a_{nn} = 0, \forall n$.

Prior works: Early advances in topology identification are put forward in [6], where a covariance estimation based method of inferring links is introduced. On the same line, in [7] the graphical Lasso is employed in order to estimate the inverse covariance matrix from data. In [8], the authors advocate that connectivity can be recovered from estimated spectral templates, while the authors in [9] propose an online adaptive algorithm for learning the topology from streaming graph signals driven by a diffusion process. Another online and adaptive algorithm is developed in [10], where the authors assume a linear model. Linearity of interactions and signal stationarity are assumed in [11], while developing an ADMM algorithm.

In modeling non-linear phenomena, works such as [12], [13] focus on polynomial structural equation models, while the authors of [14] use their nonlinear counterparts. They, however, have some limitations, such as assuming knowledge of certain connections or the form of the non-linear functions. Reproducing kernels have seen widespread use in topology inference problems. One of these works is [15] where kernels model nonlinear relationships between nodes. In [16], a kernel-based approach was considered in an online algorithm to deal with dynamical settings. The multi-kernel approach in [17] uses partial correlations to encode graph topology and ℓ_p-norm regression to enhance performance. In [18], a thorough analysis of the kernel-based topology inference problem is given, focusing on capturing non-linear and dynamic links. A review of state-of-the art methods in topology inference is in [19]. All these works, however, share a common limitation: the interactions between the different nodes are restricted to be additive.

Encoding the graph topology with the adjacency matrix $\boldsymbol{A}$ is ubiquitous in graph signal models. This operator describes the interactions between entities and, by extension, it can be considered as a tool for representing relationships between data. In this work, we propose an online approach able to estimate an adjacency matrix based on a general nonlinear model. This model ensures a better representativity of nonlinear interactions, without assuming a particular manner on how

The work of C. Richard was funded in part by the ANR under grant ANR-19-CE48-0002, and by the 3IA Côte d'Azur Senior Chair program.

agents in a network influence each other. Moreover, due to the online nature of the method, the data acquisition process can be stopped exactly when an estimate is obtained, a property which can prove useful in domains such as medical research.
Notations: Normal font letters denote scalars, while boldface lowercase and uppercase letters stand for column vectors and matrices, respectively. Operator $\|\cdot\|_0$ returns the number of their non-zero entries. Uppercase calligraphic letters denote sets, of cardinality $|\cdot|$. Finally, $\mathbb{E}\{\cdot\}$ is the expectation operator.

II. Local problem formulation

Consider an $(N+1)$-node graph with adjacency matrix $\boldsymbol{A}$ which models a system such as the brain network or a power grid. In this setting, the electrical activity of different brain-regions [20], [21], or the voltage angle per bus [17], numbered from 1 through $(N+1)$, can be measured at different time instants $i \in \mathbb{N}$, leading to a dynamic graph signal $\boldsymbol{y}(i)$. The signal at each node influences and is influenced by the signals at the other nodes, with nonlinear relationships being reported in many applications such as, e.g., in the case of brain connectivity [22], [23].

The distributed nature of graphs allows for a local problem formulation. As such, we focus on a single node n, while keeping in mind that the following reasoning can be applied for any other particular node. For ease of notation, we assume that $n \equiv (N+1)$ (i.e., we identify n with the $(N+1)^{\text{th}}$ node of the graph), which allows us to denote $\mathcal{N} \setminus \{n\} = \{1, \ldots, N\}$. Recent methods have considered models of the form [16], [24]:

$$y_n(i) = \sum_{m=1}^{N} a_{nm} g_m(\boldsymbol{y}_{L_m}(i)) + v_n(i)\,, \tag{1}$$

where $v_n(i)$ represents innovation noise, and a_{nm} is the $(n,m)^{\text{th}}$ entry of the graph adjacency matrix $\boldsymbol{A}$. This matrix models how each function g_m, $m = 1, \ldots, N$ influences the signal observed at node n. Let $g_m : \mathbb{R}^{L_m} \to \mathbb{R}$ be a nonlinear function whose possibly vector-valued argument is $\boldsymbol{y}_{L_m}(i) = [y_m(i), \ldots, y_m(i - L_m + 1)]^\top$, for $L_m \geq 1$. Therefore, the signal at each node depends non-linearly on the signals at all the other nodes, up to the past L_m samples. Given nodal measurements $\boldsymbol{y}(i)$ acquired online and model (1), the goal is to estimate the adjacency matrix $\boldsymbol{A}$ locally at each node:

$$\begin{aligned} \underset{\boldsymbol{a}_n, g_1, \ldots, g_N}{\operatorname{argmin}} \; & \frac{1}{2}\mathbb{E}\Big\| y_n(i) - \sum_{m=1}^{N} a_{nm} g_m(\boldsymbol{y}_{L_m}(i)) \Big\|^2 + \phi(\boldsymbol{a}_n) \\ \text{subject to } & a_{nm} \in \{0, 1\}\,, \end{aligned} \tag{2}$$

where $\boldsymbol{a}_n$ is the n^{th} row of $\boldsymbol{A}$, and function ϕ is a sparsity promoting regularizer. However, models such as (1) do not consider the nonlinear interactions between multiple nodes, as they assume an additive model for $y_n(i)$ [25]. To overcome this issue, we propose to consider the following general nonlinear model:

$$y_n(i) = f_n(\boldsymbol{y}_{L_1}(i), \ldots, \boldsymbol{y}_{L_N}(i)) + v_n(i)\,. \tag{3}$$

Model (3), compared to (1), captures much more complex relationships between the different nodes without relying on an assumed additive model, therefore rendering it more general.

A. Nonparametric sparsity

Although kernel-based and other nonlinear regression frameworks can be applied to estimate the function f_n described in equation (3), there remains a challenge to relate f_n to the underlying graph topology $\boldsymbol{A}$. Although the lack of an additive model precludes a straightforward relationship such as in (1), the influence of a certain variable can be quantified by the norm of the corresponding partial derivative, i.e.:

$$\text{node } m \text{ does not influence } n \iff \left\| \frac{\partial f_n}{\partial \boldsymbol{y}_{L_m}} \right\| = 0\,, \tag{4}$$

under the assumption that f_n is continuously differentiable.

If we do not assume the additive model, we can generalize problem (2) as:

$$\begin{aligned} \underset{f_n}{\operatorname{argmin}} \; & \frac{1}{2}\mathbb{E}\Big\| y_n(i) - f_n(\boldsymbol{y}_{L_1}(i), \ldots, \boldsymbol{y}_{L_N}(i)) \Big\|^2 \\ \text{subject to } & \operatorname{col}\left\{ \left\| \frac{\partial f_n}{\partial \boldsymbol{y}_{L_m}} \right\| \right\}_{m=1}^{N} \text{ being sparse}\,. \end{aligned} \tag{5}$$

Let us denote $\widetilde{\boldsymbol{y}}(i) = [\boldsymbol{y}_{L_1}(i)^\top, \ldots, \boldsymbol{y}_{L_N}(i)^\top]^\top$. As thoroughly explained in [26], in order to define a nonparametric notion of sparsity that leads to a convex optimization problem, one can define the sparsity through the following functional:

$$\boldsymbol{\Omega}_{\mathbb{E}}(f_n) = \sum_{m=1}^{N} \left\| \frac{\partial f_n}{\partial \boldsymbol{y}_{L_m}} \right\|_{\mathbb{E}} = \sum_{m=1}^{N} \sqrt{\mathbb{E}_{\widetilde{\boldsymbol{y}}} \left\{ \left\| \frac{\partial f_n(\widetilde{\boldsymbol{y}})}{\partial \boldsymbol{y}_{L_m}} \right\|^2 \right\}}\,. \tag{6}$$

The expectation involving the derivatives in (6) can be approximated by the empirical average on all the data samples available up to instant i. Employing the ℓ_2-norm, as proposed in the aforementioned paper, and approximating the expectation, we obtain the following sampled version functional:

$$\boldsymbol{\Omega}(f_n) = \sum_{m=1}^{N} \left\| \frac{\partial f_n}{\partial \boldsymbol{y}_{L_m}} \right\|_i = \sum_{m=1}^{N} \sqrt{\frac{1}{i} \sum_{p=1}^{i} \left\| \frac{\partial f_n(\widetilde{\boldsymbol{y}}(p))}{\partial \boldsymbol{y}_{L_m}} \right\|^2}\,. \tag{7}$$

By including $\boldsymbol{\Omega}(f_n)$ as an additive term in the cost function of (5), we are able to obtain a convex optimization problem, which allow us to obtain more efficient algorithms, and cater to real-word graphs, which tend to be sparse.

B. Sparsity in Reproducing Kernel Hilbert Spaces

The penalty term proposed in the previous section allows us to promote sparsity in the estimated topology without the restrictive constraint of an additive model. However, there remains a fundamental step to constrain f_n to an adequate class of functions that is flexible but allows for an efficient, finite dimensional implementation. Several solutions exist in the literature, including non-linear and polynomial Structural Equation Models [27] and function selection from function sets defined *a priori* [28]. In this work we will consider kernel methods, which address the presence of nonlinearities in classification or regression problems by applying linear algorithms to a high-dimensional feature space obtained by mapping the input data to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\kappa$ associated with a positive definite reproducing kernel $\kappa(\cdot, \cdot)$. RKHS-based

solutions have been applied in the context of nonlinear additive models for online topology estimation in, e.g., [16].

For simplicity, consider from now on $L_m = 1, \forall m \in \mathcal{N}$, which means that $\boldsymbol{y}_{L_m}(i) = y_m(i)$. Using the sparsity penalty in (7), constraining f_n to belong to an RKHS $\mathcal{H}_\kappa$ and approximating the expectation in (5) by an empirical average leads to the following optimization problem:

$$\underset{f_n \in \mathcal{H}_\kappa}{\operatorname{argmin}} \ \frac{1}{2i} \sum_{\ell=1}^{i} \left\| y_n(\ell) - f_n(\widetilde{\boldsymbol{y}}(\ell)) \right\|^2 + \eta_n \left(\sum_{m=1}^{N} \sqrt{\frac{1}{i} \sum_{p=1}^{i} \left(\frac{\partial f_n(\widetilde{\boldsymbol{y}}(p))}{\partial y_m} \right)^2} + \psi\big(\|f_n\|_{\mathcal{H}_\kappa}\big) \right) . \tag{8}$$

In (8), parameter $\eta_n > 0$ controls the relative importance of respecting the constraint on f_n, and $\psi : \mathbb{R} \to [0, \infty[$ is a monotonically increasing function.

Despite allowing us to introduce sparsity in f_n without constraining it to an additive model, the cost function in (8) also contains a significant challenge to an RKHS-based solution: the sparsity-promoting penalty term precludes the direct application of traditional representation theorems to obtain a finite dimensional representation. Fortunately, if we suppose that the kernel $\kappa(\cdot, \cdot)$ is at least twice differentiable, then the following relation holds [29]:

$$\mathcal{H}_\kappa \ni \frac{\partial f_n(\widetilde{\boldsymbol{y}})}{\partial y_m} = \langle f_n n, \kappa_{\partial_m}(\cdot, \widetilde{\boldsymbol{y}}) \rangle_{\mathcal{H}_\kappa} , \tag{9}$$

where

$$\kappa_{\partial_m}(\cdot, \widetilde{\boldsymbol{y}}(q)) = \left. \frac{\partial \kappa(\cdot, \boldsymbol{s})}{\partial s_m} \right|_{\boldsymbol{s} = \widetilde{\boldsymbol{y}}(q)} . \tag{10}$$

This means that for sufficiently smooth kernels, the derivative of functions in $\mathcal{H}_\kappa$ also belong to $\mathcal{H}_\kappa$, and can be evaluated in the form of simple inner products. This makes it possible to obtain a finite dimensional representation of the solution of (8), as previously demonstrated in [26]. As such, the representer theorem holds for our approach [26], and a solution of (8) is:

$$f_n^* = \sum_{p=1}^{i} \alpha_p \kappa(\cdot, \widetilde{\boldsymbol{y}}(p)) + \sum_{m=1}^{N} \sum_{q=1}^{i} \beta_{m,q} \kappa_{\partial_m}(\cdot, \widetilde{\boldsymbol{y}}(q)) , \tag{11}$$

which can be substituted in (8) in order to obtain a finite dimensional optimization problem.

III. An online algorithm

An immediate observation concerning solution (11) is that the number of coefficients α_p and $\beta_{m,q}$ can become prohibitive as i increases, since each acquired measurement increases the number of kernel functions. A solution to this problem is the use of kernel dictionaries which can be defined *a priori* [30] or which can admit a new candidate kernel function only if the candidate function passes a certain sparsification rule based on, e.g., the coherence criterion [31]. Under this framework, each node n in the network creates, updates, and stores a dictionary of kernel functions and their derivatives, $\mathcal{D}_n = \{\{\kappa(\cdot, \widetilde{\boldsymbol{y}}(\omega_j)), \kappa_{\partial_1}(\cdot, \widetilde{\boldsymbol{y}}(\omega_j)), \ldots, \kappa_{\partial_N}(\cdot, \widetilde{\boldsymbol{y}}(\omega_j))\} : \omega_j \in \mathcal{I}_n^i \subset \{1, \ldots, i-1\}\}$, where $\mathcal{I}_n^i$ represents the set of time indices of elements selected for the dictionary, before instant i. This entails the fact that, after a sufficient number of samples i has been acquired, only a number $|\mathcal{I}_n^i| \ll i$ of coefficient couples will be needed. A candidate kernel function $\kappa(\cdot, \widetilde{\boldsymbol{y}}(i))$ is added to $\mathcal{D}_n$ if the following sparsification condition holds [31]:

$$\max_{\omega_j \in \mathcal{I}_n^i} |\kappa(\widetilde{\boldsymbol{y}}(i), \widetilde{\boldsymbol{y}}(\omega_j))| \leq \xi_n , \tag{12}$$

where $\xi_n \in [0, 1[$ determines the level of sparsity and coherence of the dictionary. The number of entries in the dictionary satisfies $|\mathcal{I}_n^i| < \infty$ when $i \to \infty$ [31]. We rewrite (11) as:

$$f_n^* = \sum_{p=1}^{|\mathcal{I}_n^i|} \alpha_p \kappa(\cdot, \widetilde{\boldsymbol{y}}(\omega_p)) + \sum_{m=1}^{N} \sum_{q=1}^{|\mathcal{I}_n^i|} \beta_{m,q} \kappa_{\partial_m}(\cdot, \widetilde{\boldsymbol{y}}(\omega_q)) . \tag{13}$$

Let vectors $\boldsymbol{\alpha} = \operatorname{col}\{\{\alpha_p\}_{p=1}^{|\mathcal{I}_n^i|}\}$, $\boldsymbol{\beta} = \operatorname{col}\{\{\boldsymbol{\beta}_m\}_{m=1}^{N}\}$, with $\boldsymbol{\beta}_m = \operatorname{col}\{\{\beta_{m,q}\}_{q=1}^{|\mathcal{I}_n^i|}\}$, group the coefficients in (13). Considering the online version of the batch cost function (8) with the instantaneous MSE estimate (measured only at instant i), setting $\psi = 0$, and using the dictionary-based representation of f_n^* in (13), we obtain the following finite-dimensional optimization problem:

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \ \frac{1}{2} \left\| y_n(i) - \boldsymbol{\gamma}^\top \boldsymbol{s}(i) \right\|^2 + \eta_n \sum_{m=1}^{N} \sqrt{\frac{1}{i} \sum_{p=1}^{i} \big(\boldsymbol{\gamma}^\top \boldsymbol{t}_m(p) \big)^2} , \tag{14}$$

with $\boldsymbol{s}(i) = \begin{bmatrix} \boldsymbol{z}(i) \\ \boldsymbol{k}(i) \end{bmatrix}$, $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}$, $\boldsymbol{t}_m(p) = \begin{bmatrix} \boldsymbol{\ell}_m(p) \\ \boldsymbol{z}_m(p) \end{bmatrix}$, and:

$$\boldsymbol{k}(i) = \operatorname{col}\Big\{ \kappa(\widetilde{\boldsymbol{y}}(i), \widetilde{\boldsymbol{y}}(\omega_q)) \Big\}_{q=1}^{|\mathcal{I}_n^i|} , \tag{15}$$

$$\boldsymbol{z}(i) = \big[\boldsymbol{z}_1^\top(i), \ldots, \boldsymbol{z}_N^\top(i) \big]^\top , \tag{16}$$

$$[\boldsymbol{z}_m(i)]_q = \left. \frac{\partial \kappa(\widetilde{\boldsymbol{y}}(i), \widetilde{\boldsymbol{y}}(\omega_q))}{\partial y_m(\omega_q)} \right|_{q=1,\ldots,|\mathcal{I}_n^i|} ,$$

$$\boldsymbol{\ell}_m(i) = \big[\boldsymbol{\ell}_{1,m}^\top(i), \ldots, \boldsymbol{\ell}_{N,m}^\top(i) \big]^\top , \tag{17}$$

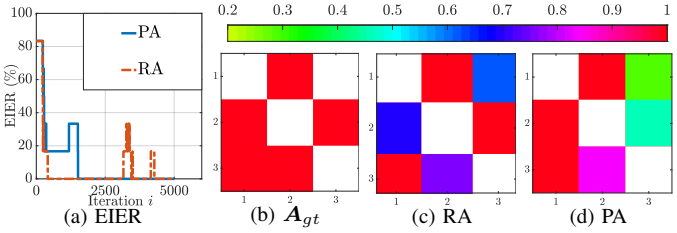
$$[\boldsymbol{\ell}_{m_1,m_2}(i)]_q = \left. \frac{\partial^2 \kappa(\widetilde{\boldsymbol{y}}(i), \widetilde{\boldsymbol{y}}(\omega_q))}{\partial y_{m_1}(\omega_q) \partial y_{m_2}(\omega_q)} \right|_{q=1,\ldots,|\mathcal{I}_n^i|} .$$

The quantities (15), (16), and (17) can be computed in closed form when an explicit expression of the continuously differentiable kernel $\kappa(\cdot, \cdot)$ is chosen.

One difficulty with problem (14) is that the summation of $\boldsymbol{t}_m(p)$, $p = 1, \ldots, i$ in the regularization term grows linearly with i, and is thus not scalable in this form. Note that we can write the finite-dimensional regularizer in (14) as:

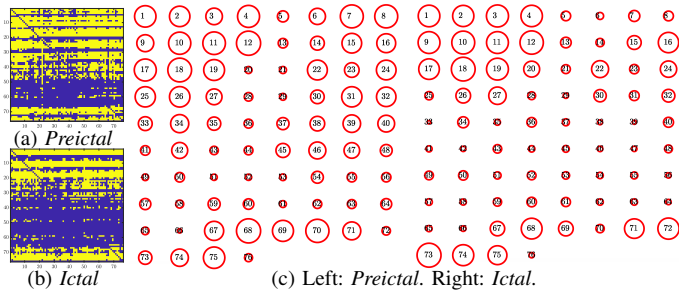
$$\boldsymbol{\Omega}(f_n) = \sum_{m=1}^{N} \sqrt{\frac{1}{i} \sum_{p=1}^{i} \big(\boldsymbol{\gamma}^\top \boldsymbol{t}_m(p) \big)^2} = \sum_{m=1}^{N} \sqrt{\boldsymbol{\gamma}^\top \left(\frac{1}{i} \sum_{p=1}^{i} \boldsymbol{T}_m(p) \right) \boldsymbol{\gamma}} = \sum_{m=1}^{N} \sqrt{\boldsymbol{\gamma}^\top \overline{\boldsymbol{T}}_m(i) \boldsymbol{\gamma}} , \tag{18}$$

with $\boldsymbol{T}_m(p) = \boldsymbol{t}_m(p) \boldsymbol{t}_m^\top(p)$ and $\overline{\boldsymbol{T}}_m(i) = \frac{1}{i} \sum_{p=1}^{i} \boldsymbol{T}_m(p)$, for



(a) EIER (b) $\boldsymbol{A}_{gt}$ (c) RA (d) PA

Fig. 1. EIER, ground truth and estimates. White represents 0.



(a) *Preictal* (b) *Ictal* (c) Left: *Preictal*. Right: *Ictal*.

Fig. 2. (a) and (b): Estimated adjacency matrices. (c): Summed in- and out-degrees for the estimated graphs. The larger the radius corresponding to node n, the larger the summed degree of node n.

$m = 1, \ldots, N$. Since the following relation is satisfied:

$$\overline{\boldsymbol{T}}_m(i) = \frac{1}{i}\boldsymbol{T}_m(i) + \frac{i-1}{i}\overline{\boldsymbol{T}}_m(i-1)\,, \quad \forall i \geq 2\,, \tag{19}$$

we can compute $\overline{\boldsymbol{T}}_m(i)$ recursively for all i with a fixed complexity. Optimization problem (14) now becomes:

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\, \frac{1}{2}\Big\|y_n(i) - \boldsymbol{\gamma}^\top \boldsymbol{s}(i)\Big\|^2 + \eta_n \sum_{m=1}^{N} \sqrt{\boldsymbol{\gamma}^\top \overline{\boldsymbol{T}}_m(i)\boldsymbol{\gamma}}\,. \tag{20}$$

Cost function (20) can then be optimized iteratively using the subgradient descent algorithm as:

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(i+1) = \hat{\boldsymbol{\gamma}}(i) + \mu_n \boldsymbol{s}(i)\big(y_n(i) - \boldsymbol{s}^\top(i)\hat{\boldsymbol{\gamma}}(i)\big) \\ - \mu_n\eta_n \sum_{m=1}^{N} \frac{\overline{\boldsymbol{T}}_m(i)\hat{\boldsymbol{\gamma}}(i)}{\hat{\Lambda}_m(i)}\,,\end{aligned} \tag{21}$$

with $\hat{\Lambda}_m(i) = \sqrt{\hat{\boldsymbol{\gamma}}^\top(i)\overline{\boldsymbol{T}}_m(i)\hat{\boldsymbol{\gamma}}(i)}$. In (20) and (21), each Λ_m represents the estimate of the partial derivative of f_n with respect to y_m. The proposed method, summarized in Algorithm 1, has a per-iteration complexity of $\mathcal{O}(N^2)$. Approximate strategies can be considered to obtain a scalable implementation. Parameter τ_n acts as an edge identification threshold. It is used to identify the topology from the estimated coefficients $\hat{\Lambda}_m(i)$, determining whether there exist links from each node $m \in \mathcal{N} \setminus \{n\}$ towards n. When processing real data, τ_n can be set as to obtain an estimated topology which realistically explains the studied process.

Algorithm 1: Kernel-based online graph inference

Inputs: For every node n: μ_n, η_n, $\kappa(\cdot,\cdot)$, ξ_n, and τ_n
Initialization: Set all entries of $\hat{\boldsymbol{\gamma}}(0)$ to 0
Algorithm: At each time instant $i \geq 1$
Update $\mathcal{D}_n$ if $\kappa(\cdot, \widetilde{\boldsymbol{y}}(i))$ satisfies condition (12)
Compute $\boldsymbol{s}(i)$ and $\overline{\boldsymbol{T}}_m(i)$ with (15), (16), (17), (19)
Update $\hat{\boldsymbol{\gamma}}(i)$ using (21)
Set $\hat{a}_{nm}(i)$ to 1 if $\hat{\Lambda}_m(i) \geq \tau_n$, to 0 otherwise

IV. Experimental results

In this section, the performance of the proposed method is evaluated by two experiments: one considering synthetic data, and another considering real epilepsy data. The Gaussian kernel $\kappa(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(-\|\boldsymbol{a} - \boldsymbol{b}\|^2/2\sigma_n^2\right)$ was used in both cases.

Lorenz graph: Consider the first experiment in [16] (hereafter called RA for *reference algorithm*), based on the discretized version of the Lorenz attractor [32]. The system parameters were set to the same values as in [33], with initial conditions $y_n(0) = 10^{-10}$, $\forall n$, for both RA and the proposed algorithm (PA). This system contains nonlinear interactions which cannot be completely characterized by additive models. We set μ_n and τ_n as to achieve the fastest convergence. Parameters σ_n and ξ_n were set as to obtain the same number of dictionary entries per node, nine, in order to achieve a meaningful comparison. Finally, τ_n were set as to achieve the best performance in terms of edge identification error rate, EIER $\triangleq \|\boldsymbol{A}_{gt} - \hat{\boldsymbol{A}}\|_0/(N(N-1)) \cdot 100\%$, where $\boldsymbol{A}_{gt}$ is the ground truth depicted in Fig. 1(b). Fig. 1(a) depicts the EIER of both algorithms as a function of the iterations. In Fig. 1(c) and (d), each entry represents the mean of the normalized $\hat{\Lambda}_m(i)$, per n, over $i = 5000$ iterations, which encodes the strength of a link from node m towards n before thresholding. It is desired that the amplitude of the elements for which $a_{nm} = 1$ be larger and as separated as possible from the amplitude of the elements for which $a_{nm} = 0$, since this makes distinguishing the active and inactive links easier. On the first row, the more general model of the current method is able to better differentiate between the absence and presence of a link. This is seen through the larger difference between $\hat{\Lambda}_2$ and $\hat{\Lambda}_3$ corresponding to $\hat{a}_{12} = 1$ and $\hat{a}_{13} = 0$, respectively, in the case of PA, while for the RA the strength of the link corresponding to $\hat{a}_{13}$ is much closer to that of the active link. The lower performance of RA was expected since it constrains the interactions to obey an additive model.

Real data: The data for this experiment come from a 39-year-old female subject suffering from intractable epilepsy. The data acquisition and pre-processing information is provided in [3]. The data-set contains 8 instances of electrocorticography (ECoG) time series, each instance representing one seizure and contains voltage measurements from 76 different regions on and inside the brain, during the 10 seconds before the epilepsy seizure (*preictal* interval) and the first 10 seconds during the seizure (*ictal* interval). We set $\mu_n = 10^{-3}, \eta_n = 10^2, \xi_n = 0.8$ for each n. Since there was no ground truth, parameters σ_n and τ_n were set as to obtain coherent results.

Fig. 2(a) and (b) show the estimated connectivity of the brain, for each interval, averaged over the 8 instances. Fig. 2(c) depicts the degree (sum of in- and out-degree), encoded in the radii of the circles, relative to each interval. Interestingly, our online estimate reveals roughly the same behavior before and during the seizure as the estimate obtained using the method batch developed in [21] and the online method in [16]. More precisely, the number of total connections decreases from one

interval to the other, especially due to the variation of in-degrees for nodes 30 to 50. Further analyzing the connections, nodes 75 and 76 have a small in-degree, however they present a more important out-degree. Observe the decrease of the degree of node 26 or the major increase for node 73. This behavior is consistent with the findings of the aforementioned papers.

The algorithm is therefore able to obtain results similar to those obtained in previous works, based on a general data model. For reference, the number of kernel functions inserted in the dictionaries was at most 9, after 4000 samples. These results show how a small number of kernel functions are actually needed in order to obtain a satisfactory topology estimate. This fact, alongside the online approach and general model, can translate in reduced computational complexity due to the drastically reduced number of needed kernel functions.

V. Conclusion

In this paper, a new kernel-based online topology estimation method was proposed accounting for general nonlinear interactions between the agents in a network. Previous work only considered models based on additive interactions between the signals at the different nodes. However, such a simplifying model is not entirely justified in many practical applications. Following a more general approach, we consider arbitrary nonlinear interactions between the nodes, which render our model much more general. By encoding links as partial derivatives of the nonlinear functions, we are able to benefit from the kernel machinery framework to estimate a possibly directed, sparse adjacency matrix. An online algorithm is proposed, using kernel dictionaries and recursive computations of the regularization terms to operate with bounded complexity. Preliminary results indicate the proposed method can lead to more accurate estimates for more general nonlinear systems.

References

[1] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851 EP –, 10 2007. [Online]. Available: https://doi.org/10.1038/nature06258

[2] R. H. Heiberger, "Predicting economic growth with stock networks," *Physica A: Statistical Mechanics and its Applications*, vol. 489, pp. 102–111, 2018.

[3] M. Kramer, E. D. Kolaczyk, and H. Kirsch, "Emergent network topology at seizure onset in humans," *Epilepsy research*, vol. 79, pp. 173–86, 2008.

[4] N. Biggs, *Algebraic Graph Theory*. Cambridge University Press, 1993.

[5] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 80–90, 2014.

[6] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[7] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–41, 2008.

[8] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 467–483, 2017.

[9] S. Vlaski, H. P. Maretić, R. Nassif, P. Frossard, and A. H. Sayed, "Online graph learning from sequential data," in *Proc. IEEE Data Science Workshop*, Lausanne, Switzerland, 2018, pp. 190–194.

[10] M. Moscu, R. Nassif, F. Hua, and C. Richard, "Learning causal networks topology from streaming graph signals," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[11] R. Shafipour, A. Hashemi, G. Mateos, and H. Vikalo, "Online topology inference from streaming stationary graph signals," *IEEE Data Science Workshop (DSW)*, pp. 140–144, 2019.

[12] J. Harring, B. Weiss, and J.-C. Hsu, "A comparison of methods for estimating quadratic effects in nonlinear structural equation models," *Psychological methods*, vol. 17, pp. 193–214, 2012.

[13] W. H. Finch, "Modeling nonlinear structural equation models: A comparison of the two-stage generalized additive models and the finite mixture structural equation model," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 22, no. 1, pp. 60–75, 2015.

[14] N. Lim, F. D'Alché-Buc, C. Auliac, and G. Michailidis, "Operator-valued kernel-based vector autoregressive models for network inference," *Machine learning*, vol. 99, no. 3, pp. 489–513, 2015.

[15] Y. Shen, B. Baingana, and G. B. Giannakis, "Topology inference of directed graphs using nonlinear structural vector autoregressive models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 6513–6517.

[16] M. Moscu, R. Borsoi, and C. Richard, "Online graph topology inference with kernels for brain connectivity estimation," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2020.

[17] L. Zhang, G. Wang, and G. B. Giannakis, "Going beyond linear dependencies to unveil connectivity of meshed grids," in *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Curaçao, Dutch Antilles, 2017, pp. 1–5.

[18] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology Identification and Learning over Graphs: Accounting for Nonlinearities and Dynamics," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 787–807, 2018.

[19] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 44–63, 2019.

[20] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[21] Y. Shen, B. Baingana, and G. B. Giannakis, "Nonlinear structural vector autoregressive models for inferring effective brain network connectivity," *arXiv preprint arXiv:1610.06551*, 2016.

[22] W. J. Freeman, "EEG analysis gives model of neuronal template-matching mechanism for sensory search with olfactory bulb," *Biological cybernetics*, vol. 35, no. 4, pp. 221–234, 1979.

[23] J. A. de Zwart, P. van Gelderen, J. M. Jansma, M. Fukunaga, M. Bianciardi, and J. H. Duyn, "Hemodynamic nonlinearities affect BOLD fMRI response timing and amplitude," *Neuroimage*, vol. 47, no. 4, pp. 1649–1658, 2009.

[24] Y. Shen, B. Baingana, and G. B. Giannakis, "Kernel-based structural equation models for topology identification of directed networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2503–2516, 2017.

[25] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models," *The Annals of Statistics*, pp. 453–510, 1989.

[26] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri, "Nonparametric sparsity and regularization," *Journal of Machine Learning Research*, vol. 14, pp. 1665–1714, 2013.

[27] K. G. Jöreskog, F. Yang, G. Marcoulides, and R. Schumacker, "Nonlinear structural equation models: The Kenny-Judd model with interaction effects," *Advanced structural equation modeling: Issues and techniques*, no. 3, pp. 57–88, 1996.

[28] X.-Y. Song, Z.-H. Lu, J.-H. Cai, and E. H.-S. Ip, "A bayesian modeling approach for generalized semiparametric structural equation models," *Psychometrika*, vol. 78, no. 4, pp. 624–647, 2013.

[29] D.-X. Zhou, "Derivative reproducing properties for kernel methods in learning theory," *Journal of computational and Applied Mathematics*, vol. 220, no. 1-2, pp. 456–463, 2008.

[30] J. Chen, W. Gao, C. Richard, and J.-C. M. Bermudez, "Convergence analysis of kernel LMS algorithm with pre-tuned dictionary," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[31] C. Richard, J.-C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, 2009.

[32] E. N. Lorenz, "Computational chaos-a prelude to computational instability," *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 299 – 317, 1989.

[33] F. A. Tobar, S. Kung, and D. P. Mandic, "Multikernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 265–277, 2014.