Convergence Analysis of the Graph-Topology-Inference Kernel LMS Algorithm

Mircea Moscu, Ricardo Borsoi, Cédric Richard

CONVERGENCE ANALYSIS OF THE GRAPH-TOPOLOGY-INFERENCE KERNEL LMS ALGORITHM

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

[*] Université Côte d'Azur, CNRS, OCA, Nice, France

[†] Federal University of Santa Catarina, Florianópolis, Brazil

ABSTRACT

Identifying directed connectivity patterns from nodal measurements is an important problem in network analysis. Recent works proposed to leverage the performance and flexibility of strategies operating in reproducing kernel Hilbert spaces (RKHS) to model nonlinear interactions between network agents. Moreover, several applications require online and efficient solutions, which motivated the consideration of distributed adaptive learning strategies inspired by algorithms such as the kernel least mean square (KLMS). Despite showing good performance, a thorough theoretical understanding of the behavior of such algorithms is still missing. This makes applying them in practice challenging, especially because the set-up of adaptive algorithms involves additional parameters like the step size and a dictionary of kernel functions. In this paper, we present a convergence analysis of the graph-topology-inference KLMS algorithm. Monte Carlo simulations demonstrate the accuracy of the theoretical models.

Index Terms— convergence analysis, kernel least mean squares, topology inference, online processing

1. INTRODUCTION

Graphs have proven to be a fundamental tool in several applications, such as the analysis of socio-economical interactions [1] or brain activity [2]. In this context, information on the network structure is important and required by traditional graph signal processing algorithms [3]. It however appears that the topology of the graph is usually unknown beforehand. Recent works have addressed this issue by estimating the network topology directly from a set of measurements acquired at each node of the graph [4]. Although most topology identification algorithms assume linear dependencies between the nodal signals, nonlinear interactions between measurements at different nodes can be observed in many real-world applications, e.g., brain activity [2]. Many big data-oriented applications also require algorithms which are distributed, scalable, and can operate online with adaptive capabilities.

Although several solutions have been proposed for nonlinear online topology identification, a theoretical evaluation of the performance of such methods remain a challenging but important issue. These analyses are important both in understanding and designing the algorithm. In this paper, we present a theoretical model for the convergence behavior graph-topology-inference KLMS algorithm.

Background and prior work: Many approaches have been proposed for topology identification in the presence of linear interactions. Such works include inverse covariance estimation based on the graphical Lasso [5], or the recovery of the topology from spectral templates [6]. More recently, online strategies have been proposed to address this problem, including diffusion-based [7] and distributed [8] adaptive algorithms, and variable-splitting based methods for stationary data [9]. However, these works do not consider nonlinear interactions that may happen.

Several strategies have been proposed to address the topology inference problem with nonlinear interactions, such as polynomial structural equation models [10, 11] and their nonlinear counterparts [12]. However, these methods require the nonlinear interaction models to be specified *a priori*. Kernel-based strategies, on the other hand, offer an efficient means of representing nonlinear functions by selecting an appropriate RKHS associated to a positive definite reproducing kernel [13]. The flexibility, elegance and mathematical tractability of kernel-based methods made their use widespread in topology identification. An in-depth overview of kernel-based topology inference can be found in [4], where several solutions are discussed. For instance, in [14], reproducing kernels are used to represent nodal measurements acquired at different time instants. An auto-regressive framework is employed to model graph connectivity over time, and a variable splitting-based batch algorithm is proposed to estimate the topology. A multi-kernel is also proposed in [15], where the graph topology is encoded via partial correlations. Despite performing well when the nonlinearities in the data are unknown, these strategies are not adaptive and their computational cost can be high.

An important aspect of kernel-based machinery is that it allows one to develop strategies that can address the online and adaptive topology identification problem in nonlinear settings. In [16], a batch strategy is modified to operate in an online fashion by considering a moving window with a fixed amount of samples. A stochastic gradient-based solution using a multi-kernel model was later proposed in [17] by considering a coherence-based online dictionary construction [18]. This method provides a competitive performance for estimating brain connectivity at a very small computational complexity. These models, however, are based on additive nonlinear interactions. Recently, a strategy was proposed to consider more general nonlinearities in the data [19] by employing a nonparametric topology identification framework.

However, to this date no theoretical evaluation of the performance of the online topology estimation approaches is available. This is an important issue since adaptive algorithms require the careful design of, e.g., a dictionary construction strategy [18] and an optimized step size in order to obtain a satisfactory performance. In this work, we provide a statistical performance analysis of the graph-topology-inference KLMS with a Gaussian kernel and a preselected dictionary. The presented model characterizes the mean and mean-squared behavior of the algorithm, both in transient and steady-state. Although previous works have analyzed the behavior of the KLMS

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.

algorithm [20, 21, 22], we consider a distributed setting, with separate kernels and dictionaries for each node, and a temporal dependence in the input data, which allows the algorithm to have memory and make use of past data. This last characteristic is important in applications such as functional brain topology estimation, where there is a 10–20 ms delay in signal propagation between nodes [23]. Moreover, considering the dictionary as a fixed parameter of the algorithm allows one to analyze its influence on the performance of the model, and to adjust it so as to maximize the performance [20]. Simulation results illustrate the validity of the model. In future works, we will explore the use of the presented model to devise parameter selection strategies, providing guaranteed means of inferring the topology of a graph over time.

Definitions: A graph $\mathcal{G}$ consists of a set $\mathcal{N}$ of $(N+1)$ nodes, and a set $\mathcal{E}$ of edges such that $(m,n) \in \mathcal{E}$ if, and only if node n is linked to node m. 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 $\boldsymbol{A}$ [24], is defined as an $(N+1) \times (N+1)$ matrix whose entries a_{nm} are zero if $(m,n) \notin \mathcal{E}$ and one otherwise. We also consider $a_{nn} = 0$, for all n.

Notations: Normal font letters denote scalars, while boldface lowercase and uppercase letters stand for column vectors and matrices, respectively. Uppercase calligraphic letters denote sets, of cardinality $|\cdot|$. Finally, $\mathbb{E}\{\cdot\}$ is the expectation operator.

2. PROBLEM FORMULATION

Let us consider measurements $\boldsymbol{y}(i) \in \mathbb{R}^{N+1}$ acquired over an $(N+1)$-node graph with adjacency matrix $\boldsymbol{A}$. We suppose the measurements $\boldsymbol{y}(i)$ to be acquired sequentially at all time instants $i \geq 0$. Signal $y_n(i)$ at each node $n = 1, \ldots, N$ of the graph is non-linearly coupled to the signals at all nodes in its neighborhood, according to the topology described in $\boldsymbol{A}$. By considering an additive nonlinear model, the local measurements at the n-th node can be represented as:

$$y_n(i) = \sum_{m \in \mathcal{N}\setminus\{n\}} f_{nm}(\boldsymbol{y}_{L_m}(i)) + v_n(i)\,, \tag{1}$$

where $\boldsymbol{y}_{L_m} = \big[y_m(i), \ldots, y_m(i - L_m + 1)\big]^\top$, functions $f_{nm} : \mathbb{R}^{L_m} \to \mathbb{R}$ represent the interactions between the different nodes in the network, and $v_n(i)$ denotes innovation noise. By relating how each node $m \in \mathcal{N} \setminus \{n\}$ influences node n, functions f_{nm} encode the connectivity in $\boldsymbol{A}$ directly as $a_{nm} = 0$ if and only if $f_{nm} \equiv 0$. For ease of notation, we assume that $n \equiv (N+1)$, i.e., we identify n with the $(N+1)$-th node of the graph, which allows us to denote $\mathcal{N}\setminus\{n\} = \{1, \ldots, N\}$. Thus, considering the available graph signal measurements $\boldsymbol{y}(i)$ for $1 \leq \ell \leq i$, the non-parametric local topology estimation problem at node n can be written as [16, 17]:

$$\min_{f_{n1},\ldots,f_{nN} \in \mathcal{H}} \; \frac{1}{2i} \sum_{\ell=1}^{i} \mathbb{E}\left\{ \Big[y_n(i) - \sum_{m=1}^{N} f_{nm}(\boldsymbol{y}_{L_m}(i)) \Big]^2 \right\} + \Psi\big(\|f_{n1}\|_{\mathcal{H}}, \ldots, \|f_{nN}\|_{\mathcal{H}}\big) \tag{2}$$

where $\mathcal{H}$ is a (normed) function space and $\Psi : \mathbb{R}^N \to [0, \infty[$ is a regularization functional which attempts to promote sparsity in the underlying adjacency matrix by favoring solutions in which many functions f_{nm} are identically zero.

Although several approaches can be considered for nonlinear modeling (such as, e.g., polynomial structural equation models [25]), kernel methods are particularly appealing due to their elegance and efficiency. We consider $\mathcal{H}$ to be an RKHS associated with a positive reproducing kernel $\kappa(\cdot,\cdot)$ [13]. Let us now assume that $f_{nm} \in \mathcal{H}, m \in \mathcal{N} \setminus \{n\}$, and that function Ψ can be decomposed as $\Psi(x_1, \ldots, x_N) = \sum_{m=1}^{N} \psi_m(x_m)$, where each $\psi_m : \mathbb{R} \to [0, \infty[$ is non-decreasing. Then, since (2) employs a convex loss function, the conditions of the Representer Theorem are satisfied [26], which means that the solution to (2) admits a finite-dimensional representation of the form:

$$f^*_{nm}(\cdot) = \sum_{p=1}^{i} \alpha_{nmp} \kappa_m\left(\cdot, \boldsymbol{y}_{L_m}(p)\right)\,, \quad m = 1, \ldots, N\,, \tag{3}$$

where $\alpha_{nmp} \in \mathbb{R}$ are the representation coefficients.

An immediate observation concerning (3) is that the number of coefficients α_{nmp} becomes prohibitively large as i increases. This makes solving (2) online unfeasible. A solution to this problem is the use of kernel dictionaries, which represent f_{nm} as a linear combination of a small number of appropriately selected kernel functions $\kappa_m(\cdot, y_m(\omega_j))$ [18]. Such dictionaries, which we denote by $\mathcal{D}_m = \{\kappa_m(\cdot, y_m(\omega_j)) : \omega_1, \ldots, \omega_{|\mathcal{D}_m|}\}$, can be either constructed online by selecting previous datapoints that satisfy some sparsification criterion (in which case $\omega_j \in \{1, \ldots, i-1\}$) [18], or can be set a priori (in this case, we denote the sample indexes of the dictionary elements as $\omega_j < 0$, with a slight abuse of notation) [20].

Considering dictionaries $\mathcal{D}_m$, an efficient strategy for the online estimation of f_{nm} is the KLMS algorithm. The KLMS aims to minimize the following cost function at every time instant i using the stochastic gradient descent method:

$$J_n(\boldsymbol{\alpha}_n) = \frac{1}{2}\mathbb{E}\Big\{ \big[y_n(i) - \boldsymbol{\alpha}_n^\top \boldsymbol{k}(i)\big]^2 \Big| \{\mathcal{D}_m\}_{m \in \mathcal{N}\setminus\{n\}} \Big\}. \tag{4}$$

where $\boldsymbol{\alpha}_n = \big[\tilde{\boldsymbol{\alpha}}_{n1}^\top, \ldots, \tilde{\boldsymbol{\alpha}}_{nN}^\top\big]^\top$, $\tilde{\boldsymbol{\alpha}}_{nm} = \mathrm{col}\{\alpha_{nmp}\}_{p=1}^{|\mathcal{D}_m|}$, $\boldsymbol{k}(\ell) = \big[\tilde{\boldsymbol{k}}_1^\top(i), \ldots, \tilde{\boldsymbol{k}}_N^\top(i)\big]^\top$, $\tilde{\boldsymbol{k}}_m(i) = \mathrm{col}\{\kappa_m(\boldsymbol{y}_{L_m}(i), \boldsymbol{y}_{L_m}(\omega_j))\}_{j=1}^{|\mathcal{D}_m|}$.

This leads to the following coefficient update rule:

$$\hat{\boldsymbol{\alpha}}_{n(i+1)} = \hat{\boldsymbol{\alpha}}_{n(i)} + \mu \boldsymbol{k}(i) \varepsilon(i)\,, \tag{5}$$

where $\varepsilon(i) \triangleq y_n(i) - \boldsymbol{k}^\top(i)\hat{\boldsymbol{\alpha}}_{n(i)}$ represents the instantaneous error conditioned on the dictionaries $\{\mathcal{D}_m\}_{m=1}^N$, and $\mu > 0$ denotes a small step size. The estimated coefficients $\hat{\boldsymbol{\alpha}}_{n(i)}$ can be related to the adjacency matrix $\boldsymbol{A}$ at instant i by using a threshold τ_n by setting $a_{nm}(i) = 1$ if $\|\widehat{\tilde{\boldsymbol{\alpha}}}_{nm(i)}\|_2 > \tau_n$ and $a_{nm}(i) = 0$ otherwise [27].

3. ALGORITHM ANALYSIS

Let us denote the optimal coefficients which minimize cost function (4) by $\boldsymbol{\alpha}_n^* = \boldsymbol{R}_{kk}^{-1} \boldsymbol{r}_{ky}$, with $\boldsymbol{R}_{kk} \triangleq \mathbb{E}\left\{\boldsymbol{k}(i)\boldsymbol{k}^\top(i)\right\}$ and $\boldsymbol{r}_{ky} \triangleq \mathbb{E}\left\{y_n(i)\boldsymbol{k}(i)\right\}$, and the difference between the current available estimate and the optimal solution by:

$$\boldsymbol{d}_{(i)} \triangleq \hat{\boldsymbol{\alpha}}_{n(i)} - \boldsymbol{\alpha}_n^*\,. \tag{6}$$

The remainder of this analysis concerns the use of the Gaussian kernel, chosen due to its universal approximating capabilities [28]. This kernel is defined as $k_n^{\mathrm{G}}(\boldsymbol{y}_a, \boldsymbol{y}_b) = \exp\left(-\|\boldsymbol{y}_a - \boldsymbol{y}_b\|^2 / 2\sigma^2\right)$, where σ^2 is the kernel bandwidth.

Assumption 1: Dictionaries $\{\mathcal{D}_m\}$ are set beforehand. Inputs $\boldsymbol{y}(i)$ are assumed independent, zero-mean Gaussian random vectors with auto-correlation matrix $\boldsymbol{R}_{yy} \triangleq \mathbb{E}\left\{\boldsymbol{y}(i)\boldsymbol{y}^\top(i)\right\}$.

Assumption 2: Quantity $\boldsymbol{k}(i)\boldsymbol{k}^\top(i)$ is statistically independent of the error vector $\boldsymbol{d}_{(i)}$. A justification for the feasibility of the latter assumption is presented in [29].

Assumption 3: The optimal estimation error $\varepsilon_0(i) \triangleq y_n(i) - \boldsymbol{k}^\top(i)\boldsymbol{\alpha}_n^*$ given by the finite order model is close to the one by the infinite length model, such that $\mathbb{E}\{\varepsilon_0(i)\} \approx 0$.

Error $\varepsilon(i)$ can be expressed in terms of the error vector $\boldsymbol{d}_{(i)}$:

$$\varepsilon(i) = y_n(i) - \boldsymbol{k}^\top(i)\boldsymbol{d}_{(i)} - \boldsymbol{k}^\top(i)\boldsymbol{\alpha}_n^* \,. \tag{7}$$

Replacing (7) into (5) leads to the following recursion:

$$\boldsymbol{d}_{(i+1)} = \boldsymbol{d}_{(i)} - \mu\boldsymbol{k}(i)\boldsymbol{k}^\top(i)\boldsymbol{d}_{(i)} + \mu\boldsymbol{k}(i)\varepsilon_0(i) \,. \tag{8}$$

3.1. Mean error behavior

Before proceeding to the mean behavior analysis, consider the quadratic form ξ of a Gaussian vector $\boldsymbol{\zeta}$, given by $\xi = \boldsymbol{\zeta}\boldsymbol{B}\boldsymbol{\zeta}^\top + \boldsymbol{b}^\top\boldsymbol{\zeta}$, with $\mathbb{E}\{\boldsymbol{\zeta}\} = \boldsymbol{0}, \boldsymbol{R}_{\zeta\zeta} = \mathbb{E}\{\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\}$. For $t \in \mathbb{R}$, the moment-generating function of the random variable ξ is [30, p. 101]:

$$\Psi_\xi(t) \triangleq \mathbb{E}\{\exp(t\xi)\} = \det\{\boldsymbol{I} - 2t\boldsymbol{B}\boldsymbol{R}_{\zeta\zeta}\}^{-\frac{1}{2}} \times \exp\left(\frac{t^2}{2}\boldsymbol{b}^\top\boldsymbol{R}_{\zeta\zeta}\left(\boldsymbol{I} - 2t\boldsymbol{B}\boldsymbol{R}_{\zeta\zeta}\right)^{-1}\boldsymbol{b}\right) . \tag{9}$$

Taking the expectation of relation (8), and employing Assumptions 2 and 3, leads to the mean error (ME) behavior:

$$\mathbb{E}\{\boldsymbol{d}_{(i+1)}\} = (\boldsymbol{I} - \mu\boldsymbol{R}_{kk})\,\mathbb{E}\{\boldsymbol{d}_{(i)}\} \,. \tag{10}$$

Block matrix $\boldsymbol{R}_{kk}$ contains blocks $\boldsymbol{R}_{kk}^{(m_1,m_2)} \triangleq \mathbb{E}\{\boldsymbol{k}_{m_1}(i)\boldsymbol{k}_{m_2}^\top(i)\}$, $\forall m_1, m_2 \in \mathcal{N}_{\backslash n}$. Each entry $(u, v), u = 1, \ldots, |\mathcal{D}_{m_1}|, v = 1, \ldots, |\mathcal{D}_{m_2}|$ of every block $\boldsymbol{R}_{kk}^{(m_1,m_2)}$ is:

$$\left[\boldsymbol{R}_{kk}^{(m_1,m_2)}\right]_{uv} = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}^{(uv)}\right\|^2\right) \times \mathbb{E}\left\{\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}\left\|\boldsymbol{G}\boldsymbol{y}^{(ii)}\right\|^2 - \left(\boldsymbol{y}^{(uv)}\right)^\top\boldsymbol{G}\boldsymbol{y}^{(ii)}\right)\right)\right\} , \tag{11}$$

where $\boldsymbol{y}^{(uv)} = \left[\boldsymbol{y}_{L_{m_1}}^\top(\omega_u), \boldsymbol{y}_{L_{m_2}}^\top(\omega_v)\right]^\top$, and $\boldsymbol{y}^{(ii)}$ and $\boldsymbol{G}$ are given by:

$$\boldsymbol{y}^{(ii)} = \begin{cases} \left[\boldsymbol{y}_{L_{m_1}}^\top(i), \boldsymbol{y}_{L_{m_2}}^\top(i)\right]^\top, & m_1 \neq m_2 \\ \boldsymbol{y}_{L_{m_1}}(i) \,, & m_1 = m_2 \end{cases} , \tag{12}$$

$$\boldsymbol{G} = \begin{cases} \boldsymbol{I} \,, & m_1 \neq m_2 \\ \left[\boldsymbol{I},\, \boldsymbol{I}\right]^\top, & m_1 = m_2 \end{cases} . \tag{13}$$

Making use of (9) with $\boldsymbol{B} = \frac{1}{2}\boldsymbol{G}^\top\boldsymbol{G}$, $\boldsymbol{b} = -\boldsymbol{G}^\top\boldsymbol{y}^{(uv)}$ and $t = -\frac{1}{\sigma^2}$, we obtain:

$$\left[\boldsymbol{R}_{kk}^{(m_1,m_2)}\right]_{uv} = \exp\left(\frac{1}{2\sigma^4}\left(\boldsymbol{y}^{(uv)}\right)^\top\boldsymbol{G}\boldsymbol{H}^{(m_1m_2)}\boldsymbol{G}^\top\boldsymbol{y}^{(uv)}\right) \times \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}^{(uv)}\right\|^2\right)\det\left\{\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{G}^\top\boldsymbol{G}\boldsymbol{R}_{yy}^{(m_1m_2)}\right\}^{-\frac{1}{2}} , \tag{14}$$

where $\boldsymbol{I}$ is of size $(L_{m_1} + L_{m_2}) \times (L_{m_1} + L_{m_2})$, $\boldsymbol{H}^{(m_1m_2)} = \boldsymbol{R}_{yy}^{(m_1m_2)}\left(\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{G}^\top\boldsymbol{G}\boldsymbol{R}_{yy}^{(m_1m_2)}\right)^{-1}$, with $\boldsymbol{R}_{yy}^{(m_1m_2)}$ a block matrix. Each of its blocks $(k, \ell) \in \{1, 2\}^2$ contains $[\boldsymbol{R}_{yy}]_{m_k m_\ell}$ on all entries of its main diagonal, and zeros elsewhere.

From (10), convergence of the coefficients in the mean is assured if the step size satisfies $0 < \mu < 2/\lambda_{\max}(\boldsymbol{R}_{kk})$, where $\lambda_{\max}(\cdot)$ represents the maximum eigenvalue of its matrix argument.

3.2. Mean square error behavior

Let us denote $\boldsymbol{D}_{(i)} \triangleq \mathbb{E}\{\boldsymbol{d}_{(i)}\boldsymbol{d}_{(i)}^\top\}$. Using (8) and the previous assumptions, we obtain the mean square error (MSE) behavior:

$$\boldsymbol{D}_{(i+1)} = \boldsymbol{D}_{(i)} - \mu\left(\boldsymbol{D}_{(i)}\boldsymbol{R}_{kk} + \boldsymbol{R}_{kk}\boldsymbol{D}_{(i)}\right) + \mu^2\boldsymbol{Q} + \mu^2\boldsymbol{R}_{kk}J_{n,\min} \,, \tag{15}$$

with $J_{n,\min}$ being the minimum value of the cost function (4), and:

$$J_{n,\min} = J_n(\boldsymbol{\alpha}_n^*) = \mathbb{E}\{y_n^2(i)\} - \boldsymbol{r}_{ky}^\top\boldsymbol{R}_{kk}^{-1}\boldsymbol{r}_{ky} \,, \tag{16}$$

$$\boldsymbol{Q} = \mathbb{E}\left\{\boldsymbol{k}(i)\boldsymbol{k}^\top(i)\boldsymbol{d}_{(i)}\boldsymbol{v}_{(i)}^\top\boldsymbol{k}(i)\boldsymbol{k}^\top(i)\right\} . \tag{17}$$

These second order moments relate to the MSE via [31]:

$$J_{n,\text{MSE}}(i) \triangleq \mathbb{E}\{\varepsilon^2(i)\} = J_{n,\min} + \text{Tr}\{\boldsymbol{R}_{kk}\boldsymbol{D}_{(i)}\} \,, \tag{18}$$

and to the MSD through:

$$\text{MSD}(i) \triangleq \mathbb{E}\{\|\boldsymbol{d}_{(i)}\|^2\} = \text{Tr}\{\boldsymbol{D}_{(i)}\} \,. \tag{19}$$

Let us note $k_D = \sum_{m\in\mathcal{N}_{\backslash n}} |\mathcal{D}_m|$, the total number of dictionary entries. We make use of Assumption 2, leading to the writing of the (u, v)-th entry of $\boldsymbol{Q}$ as:

$$[\boldsymbol{Q}]_{uv} = \sum_{a=1}^{k_D}\sum_{b=1}^{k_D}\mathbb{E}\left\{[\boldsymbol{k}(i)]_u[\boldsymbol{k}(i)]_v[\boldsymbol{k}(i)]_a[\boldsymbol{k}(i)]_b\right\}\left[\boldsymbol{D}_{(i)}\right]_{ab} . \tag{20}$$

We introduce matrix $\boldsymbol{K}^{(u,v)}$, whose (a, b)-th entry is:

$$\left[\boldsymbol{K}^{(u,v)}\right]_{ab} = \mathbb{E}\left\{[\boldsymbol{k}(i)]_u[\boldsymbol{k}(i)]_v[\boldsymbol{k}(i)]_a[\boldsymbol{k}(i)]_b\right\} . \tag{21}$$

Now we can write relation (20) as:

$$[\boldsymbol{Q}(i)]_{uv} = \text{Tr}\{\boldsymbol{K}^{(u,v)}\boldsymbol{D}_{(i)}\} \,. \tag{22}$$

To compute the expectation in (21), we need to find which block of $\boldsymbol{k}(i)$ (and which element therein) is being indexed by values u, v, a and b. Let us define functions $\varsigma, \varrho : \mathbb{N}_* \to \mathbb{N}_*$, which relate an index u in block vector $\boldsymbol{k}(i)$ to its corresponding constituent block $\varsigma(u)$ and entry $\varrho(u)$: $[\boldsymbol{k}(i)]_u = [\tilde{\boldsymbol{k}}_{L_{\varsigma(u)}}(i)]_{\varrho(u)}$ (i.e., $[\boldsymbol{k}(i)]_u$ is the $\varrho(u)$-th entry of the $\varsigma(u)$-th block). Then, by denoting $m_1 = \varsigma(u)$, $m_2 = \varsigma(v)$, $m_3 = \varsigma(a)$, $m_4 = \varsigma(b)$ and $\omega_p = \omega_{\varrho(u)}$, $\omega_q = \omega_{\varrho(v)}$, $\omega_r = \omega_{\varrho(a)}$, $\omega_s = \omega_{\varrho(b)}$, we can write the following:

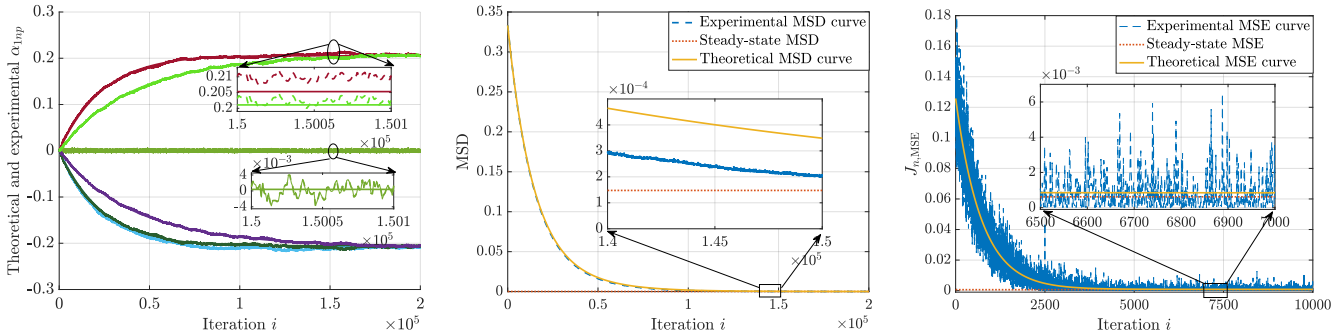
$$\left[\boldsymbol{K}^{(u,v)}\right]_{ab} = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}^{\text{d}}\right\|^2\right) \times \mathbb{E}\left\{\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|\widetilde{\boldsymbol{G}}\boldsymbol{y}^{\text{i}}\right\|^2 - \left(\boldsymbol{y}^{\text{d}}\right)^\top\widetilde{\boldsymbol{G}}\boldsymbol{y}^{\text{i}}\right)\right)\right\},$$

with $\boldsymbol{y}^{\text{d}} = \left[\boldsymbol{y}_{L_{m_1}}^\top(\omega_p), \boldsymbol{y}_{L_{m_2}}^\top(\omega_q), \boldsymbol{y}_{L_{m_3}}^\top(\omega_r), \boldsymbol{y}_{L_{m_4}}^\top(\omega_s)\right]^\top$ collecting the dictionary entries and $\boldsymbol{y}^{\text{i}} = \text{col}\{\boldsymbol{y}_{L_m}(i)\}_{m\in\bigcup_{\ell=1}^4\{m_\ell\}}$ collecting the unique instantaneous measurements (i.e., without the repeated elements when there is $m_r = m_s$ for $r \neq s$), respectively. Matrix $\widetilde{\boldsymbol{G}}$, which depends on $m_1, \ldots, m_4$, duplicates any repeated inputs on its image, such that

$$\widetilde{\boldsymbol{G}}\boldsymbol{y}^{\text{i}} = \left[\boldsymbol{y}_{L_{m_1}}^\top(i), \boldsymbol{y}_{L_{m_2}}^\top(i), \boldsymbol{y}_{L_{m_3}}^\top(i), \boldsymbol{y}_{L_{m_4}}^\top(i)\right]^\top.$$

We now use relation (9), with $\boldsymbol{B} = \frac{1}{4}\widetilde{\boldsymbol{G}}^\top\widetilde{\boldsymbol{G}}$, $\boldsymbol{b} = -\widetilde{\boldsymbol{G}}^\top\boldsymbol{y}^{\text{d}}$ and $t = -\frac{1}{\sigma^2}$. Thus, we obtain:

$$\left[\boldsymbol{K}^{(u,v)}\right]_{ab} = \det\left\{\boldsymbol{I} + \frac{1}{2\sigma^2}\widetilde{\boldsymbol{G}}^\top\widetilde{\boldsymbol{G}}\boldsymbol{R}_{yy}^{(m_1\to4)}\right\}^{-\frac{1}{2}} \times \exp\left(\frac{1}{2\sigma^4}\left(\boldsymbol{y}^{\text{d}}\right)^\top\widetilde{\boldsymbol{G}}\boldsymbol{H}^{(m_1\to4)}\widetilde{\boldsymbol{G}}^\top\boldsymbol{y}^{\text{d}} - \frac{1}{2\sigma^2}\left\|\boldsymbol{y}^{\text{d}}\right\|^2\right) , \tag{23}$$



(a) Theoretical and experimental entries of $\boldsymbol{\alpha}$. Continuous lines are the theoretical curves, while dashed line are experimental ones.

(b) Experimental, steady-state, and theoretical MSD curves.

(c) Experimental, steady-state, and theoretical MSE curves. For reasons of better visibility, a different initialization $\hat{\boldsymbol{\alpha}}_n(0)$ was used.

Fig. 1: Analysis validation in the mean and mean square sense.

where $\boldsymbol{H}^{(m_1\to 4)} = \boldsymbol{R}_{yy}^{(m_1\to 4)}\left(\boldsymbol{I}+\frac{1}{2\sigma^2}\widetilde{\boldsymbol{G}}^\top\widetilde{\boldsymbol{G}}\boldsymbol{R}_{yy}^{(m_1\to 4)}\right)^{-1}$, identity $\boldsymbol{I}$ is of size $\sum_{\ell=1}^4 L_{m_\ell}\times\sum_{\ell=1}^4 L_{m_\ell}$, and $\boldsymbol{R}_{yy}^{(m_1\to 4)}$ is a block matrix formed similarly to $\boldsymbol{R}_{yy}^{(m_1 m_2)}$: each of its blocks $(k,\ell)\in\{1,2,3,4\}^2$ contains $[\boldsymbol{R}_{yy}]_{m_k m_\ell}$ on all entries of its main diagonal, and zeros elsewhere.

By stacking the columns of $\boldsymbol{D}_{(i)}$ on top of each other, i.e., $\bar{\boldsymbol{d}}_{(i)} = \mathrm{vec}\left\{\boldsymbol{D}_{(i)}\right\}$ and making use of the properties of the vectorization operator, recursion (15) can now be computed as:

$$\bar{\boldsymbol{d}}_{(i+1)} = \boldsymbol{F}_0\bar{\boldsymbol{d}}_{(i)} + \mu^2 J_{n,\min}\bar{\boldsymbol{r}}_{kk}\,, \tag{24}$$

where $\bar{\boldsymbol{r}}_{kk} = \mathrm{vec}\{\boldsymbol{R}_{kk}\}$, and:

$$\boldsymbol{F}_0 = \boldsymbol{I}_2 - \mu(\boldsymbol{I}\otimes\boldsymbol{R}_{kk} + \boldsymbol{R}_{kk}\otimes\boldsymbol{I}) + \mu^2\boldsymbol{F}_1\,. \tag{25}$$

We remark upon the fact that the identity matrix $\boldsymbol{I}_2$ is of size $k_D^2\times k_D^2$, while $\boldsymbol{I}$ is if size $k_D\times k_D$. Also, entries of the matrix $\boldsymbol{F}_1$ are $[\boldsymbol{F}_1]_{u+(v-1)k_D,a+(b-1)k_D} = \left[\boldsymbol{K}^{(u,v)}\right]_{ab}$.

Assuming a small enough step size μ, the algorithm is mean-square stable as $i\to\infty$, and converges towards:

$$\lim_{i\to\infty}\bar{\boldsymbol{d}}_{(i)} = \mu^2 J_{n,\min}\left(\boldsymbol{I}-\boldsymbol{F}_0\right)^{-1}\bar{\boldsymbol{r}}_{kk} = \bar{\boldsymbol{d}}_{(\infty)}\,. \tag{26}$$

4. EXPERIMENTAL VALIDATION

We consider a simulation scenario with i.i.d. Gaussian data $\boldsymbol{y}(i)$ generated using a correlation matrix $\boldsymbol{R}_{yy}$, depicted in (27). We note that this particular correlation matrix also corresponds to the data correlation of the linear model $\boldsymbol{y}(i) = \boldsymbol{A}\boldsymbol{y}(i) + \boldsymbol{v}(i)$, with $\boldsymbol{v}(i)$ zero-mean Gaussian noise with covariance $\sigma_v^2\boldsymbol{I}$, where $\sigma_v = 0.05$. This model, although not inherently nonlinear, allows direct knowledge of the 5-node ground truth matrix $\boldsymbol{A}$ given in (28). Moreover, it offers exact knowledge of the statistical properties of the input $\boldsymbol{y}$, necessary in the analysis. See [17] for the methods' behavior in nonlinear settings. For the algorithm we selected $L_m = 1$, for all m, a step size $\mu = 5\cdot 10^{-2}$, kernel bandwidths of $\sigma = 1$, and each node stored a dictionary $\mathcal{D}_m$ with 3 entries, chosen in a uniform grid on $[-1,1]$. We compare the theoretical model for the ME given by (10) and for the MSE given by (15) with the empirical performance of the algorithm, averaged over 100 Monte Carlo runs. Fig.1a shows both the theoretical and experimental values for a subset of non-zero coefficients α_{nmp}, for $n=1$. Fig. 1b and 1c show both the theoretical and experimental MSD and MSE curves, as well as their steady-state values, computed using relations (26), (19), (18) and (15). It can be seen that the theoretical model was able to predict the behavior of the algorithm very accurately, both in the mean and mean-square sense. Moreover, the resulting theoretical curves serve to show that an adequate step-size, as well as other parameters, can be selected in order to attain a desirable application-dependent performance. In particular, with a small dictionary of only 3 entries for each node, it was possible to obtain a very small MSE at steady-state, which means that the estimated $\hat{f}_{nm}$ were able to predict the observations y_n accurately. A similar behavior can also be observed for the MSD. In the recovery of the network topology $\boldsymbol{A}$, we obtained an average error of around 10%, corresponding to only two links estimated incorrectly. This indicates that the considered algorithm can achieve a good trade-off between performance and computational complexity.

$$\boldsymbol{R}_{yy} = 10^{-4}\cdot\begin{pmatrix} 8.52 & 1.70 & -2.84 & -2.84 & 1.70\\ 1.70 & 8.52 & 1.70 & -2.84 & -2.84\\ -2.84 & 1.70 & 8.52 & 1.70 & -2.84\\ -2.84 & -2.84 & 1.70 & 8.52 & 1.70\\ 1.70 & -2.84 & -2.84 & 1.70 & 8.52\end{pmatrix} \tag{27}$$

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1\\ 1 & 0 & 1 & 0 & 1\\ 1 & 0 & 0 & 1 & 0\\ 0 & 1 & 1 & 0 & 1\\ 1 & 0 & 1 & 1 & 0\end{pmatrix} \tag{28}$$

5. CONCLUSION

Designing online, adaptive network topology identification strategies which account for nonlinear interactions between network agents is an important topic in network analysis. In this paper, we presented a statistical convergence analysis of the kernel LMS algorithm applied to graph topology identification. The derived model characterizes the performance of the algorithm as a function of parameters such as the kernel bandwidth and the dictionary. In turn, this allows for precise tuning of such parameters in order to obtain a desired performance in transient or in steady state. Moreover, the theoretical model quantifies the typical statistical behavior of the algorithm both in the presence and in the absence of a link between a given node pair. Future work can harness this fact in order to design an automatic thresholding rule by comparing the behavior of the empirical algorithm coefficients to the one dictated by the model.

6. REFERENCES

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

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

[3] P. M. Djurić and C. Richard, *Cooperative and Graph Signal Processing: Principles and Applications*, Academic Press, 2018.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] W. Holmes 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.

[12] 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.

[13] V. I. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*, vol. 152, Cambridge University Press, 2016.

[14] 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.

[15] 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.

[16] Y. Shen and G. B. Giannakis, "Online identification of directional graph topologies capturing dynamic and nonlinear dependencies," in *Proc. IEEE Data Science Workshop (DSW)*, Lausanne, Switzerland, 2018, pp. 195–199.

[17] M. Moscu, R. Borsoi, and C. Richard, "Online graph topology inference with kernels for brain connectivity estimation," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 1200–1204.

[18] 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.

[19] M. Moscu, R. Borsoi, and C. Richard, "Online kernel-based graph topology identification with partial-derivative-imposed sparsity," in *Proc. 28th European Conference on Signal Processing (EUSIPCO)*, Amsterdam, The Netherlands, 2020, pp. 2190–2194.

[20] 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)*, Florence, Italy, 2014, pp. 7243–7247.

[21] W. Gao, J. Chen, C. Richard, J. Huang, and R. Flamary, "Kernel LMS algorithm with forward-backward splitting for dictionary learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[22] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized online learning with kernels," *Signal Processing, IEEE Transactions on*, vol. 66, no. 12, pp. 3240–3255, 2018.

[23] S. Petkoski and V. K. Jirsa, "Transmission time delays organize the brain network synchronization," *A Philosophical Transactions of the Royal Society*, vol. 377, no. 2153, 2019.

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

[25] 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*, pp. 57–88, 1996.

[26] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *International conference on computational learning theory*. Springer, 2001, pp. 416–426.

[27] 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.

[28] W. Liu, J. C. Príncipe, and S. Haykin, *Kernel adaptive filtering: a comprehensive introduction*, chapter 1, pp. 1–26, John Wiley & Sons, Ltd, 2010.

[29] J. Minkoff, "Comment on the "Unnecessary assumption of statistical independence between reference signal and filter weights in feedforward adaptive systems"," *IEEE Transactions on Signal Processing*, vol. 49, no. 5, pp. 1109–, 2001.

[30] J. Omura and Thomas Kailath, *Useful Probability Distributions*, vol. 2, Stanford Electronics, September 1965.

[31] W. D. Parreira, J. C. M. Bermudez, C. Richard, and J.-Y. Tourneret, "Stochastic behavior analysis of the Gaussian kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2208–2222, 2012.