Online Joint Topology Identification and Signal Estimation with Inexact Proximal Online Gradient Descent

Bakht Zaman, Student Member, IEEE, Luis Miguel Lopez Ramos, Member, IEEE, and Baltasar Beferull-Lozano, Senior Member, IEEE
WISENET Center, Department of ICT, University of Agder, Grimstad, Norway

Abstract—Identifying the topology that underlies a set of time series is useful for tasks such as prediction, denoising, and data completion. Vector autoregressive (VAR) model based topologies capture dependencies among time series, and are often inferred from observed spatio-temporal data. When the data are affected by noise and/or missing samples, the tasks of topology identification and signal recovery (reconstruction) have to be performed jointly. Additional challenges arise when i) the underlying topology is time-varying, ii) data become available sequentially, and iii) no delay is tolerated. To overcome these challenges, this paper proposes two online algorithms to estimate the VAR model-based topologies. The proposed algorithms have constant complexity per iteration, which makes them interesting for big data scenarios. They also enjoy complementary merits in terms of complexity and performance. A performance guarantee is derived for one of the algorithms in the form of a dynamic regret bound. Numerical tests are also presented, showcasing the ability of the proposed algorithms to track the time-varying topologies with missing data in an online fashion.

I. INTRODUCTION

In many applications involving complex systems, causal relations among time series are computed. These relations form a causality graph, where each node corresponds to a time series, and oftentimes reveal the topology of e.g. an underlying social, biological, or brain network [1]. A causality graph provides insights about the complex system under analysis, and enables certain tasks such as forecasting [2], signal reconstruction [3], anomaly detection [4], and dimensionality reduction [5]. While most prior works assume that the data are fully observable at every node and time-instant, this is not the case in certain real-world scenarios [6], [7], due to diverse reasons: For instance, in sensor networks, the data at a node may be partially observed due to faulty equipments/sensors, dropped data packages due to network congestion, or under-observation of certain signals with the purpose of saving energy (e.g. sporadic observations based on the variations of the measured signal). In social networks, user data may be partially available due to security or privacy reasons. In ecological networks, uncontrollable factors such as weather conditions limit the ability to have reliable counts of a certain species. This paper considers the problem of online topology identification with streaming noisy data where some values are missing, simultaneously with the reconstruction of the input signals.

Identifying graphs capturing spatio-temporal “interactions” among time series has attracted great attention in the literature [1], [8]. Among the popular approaches for undirected topologies, correlation and partial correlation graphs [1], Markov random fields [9]–[12], and more recent approaches based on graph signal processing [13]–[15] are adopted in the literature. For directed interactions, one may employ structural equation models (SEM) [16], [17] or Bayesian networks [12, Sec. 8.1]. However, the methods above account only for memoryless interactions, meaning that they cannot accommodate delayed (causal) interactions where the value of a time series at a given time instant is related to the past values of other time series.

An important notion of causality among time series is due to Granger [18] based on the optimal prediction error, which is generally difficult to determine optimally [19, p. 33], [20]. Thus, alternative causality definitions based on vector autoregressive (VAR) models are typically preferred [21]–[23]. VAR topologies are estimated assuming Gaussianity and stationarity in [24], [25] and assuming sparsity in [26]–[29]. All these approaches assume that the graph does not change over time. Since this is not the case in many applications, approaches have been devised to identify time-varying topologies, both undirected [30]–[32] and directed [33].

All previously discussed approaches process the entire data set at once, and cannot accommodate data arriving sequentially. Hence, their complexity becomes prohibitive for long observation windows. The modern approach to tackle these issues is online optimization, where an estimate is refined with every new data instance. Existing online topology identification algorithms include [17], [34]–[38] for memoryless interactions, and [39] for nonlinear memory-based dependencies.

The topology identification becomes more challenging when the input data is noisy. In [40], joint signal estimation and topology identification is pursued based on a spatio-temporal smoothness-based graph learning algorithm. This setting can become even more challenging when the data are incomplete. Several batch approaches to identify topologies in the presence of noisy data with missing values are available for undirected [41] and (directed) VAR-based [42], [43] topologies.

The problem of online time series prediction with missing

The work in this paper was supported by the SFI Offshore Mechatronics grant 237896/E30, the PETROMAKS Smart-Rig grant 244205 and the IKTPLUSS Indurb grant 270730/O70 from the Research Council of Norway.

The authors are with the WISENET Center, Dept. of ICT, University of Agder, Jon Lilletunsvei 3, Grimstad, 4879 Norway. E-mails: {bakht.zaman, luismiguel.lopez, baltsasar.beferull}@uia.no.
data is considered in [44], where the goal is to predict the future values; and [45], where the missing values are imputed by their estimates. Theoretical guarantees in the form of static regret bounds are presented; however, those works adopt a univariate autoregressive (AR) process model, and thus do not extract information about the relations among multiple time series. Moreover, these works consider a static (stationary) model and analyze the static regret. An approach to jointly estimate the signal and topology is presented in [46] for a structural VAR model (SVARM) when the observations contain noisy and missing values; in that work, different batch and online algorithms are proposed, and an identifiability result is stated. However, no performance guarantees showing the tracking capabilities of the proposed online algorithm are presented.

The present work proposes online algorithms to estimate time-varying, memory-aware causality graphs from a collection of time series affected by noise and missing data, while reconstructing the input signals (denoising and imputation of missing values). The contributions include two algorithms that track VAR-causality graphs, with fixed computational complexity per sample, which renders them suitable for sequential and big-data scenarios. The proposed algorithms have complementary merits: the first one has very low computational complexity and is supported by dynamic regret bounds. However, no performance guarantees showing the tracking capabilities of the proposed online algorithm are presented.

After introducing the basic batch formulation of the jointly optimal estimation and reconstruction, the contributions are:

- **C1)** The problem of online estimation and reconstruction is formulated as a sequential decision problem (where decisions have an impact in future states), with a cost function inspired from the formulation in [46]. This cost function involves signal reconstruction mismatch (how far the estimated signal is from the signal predicted by the past values and the topology), deviation from the noisy samples, time-variation of the estimated topology (distance between parameter estimates that are adjacent in time), and a sparsity-promoting term. Three hyperparameters (one of them becoming the learning rate) allow to trade off between the aforementioned metrics.

- **C2)** A first algorithm, termed Joint Signal and Topology Identification via Sparse Online learning (JSTISO), which at each iteration simultaneously estimates the signal and the topology by minimizing a simple loss function. Such a loss function has been derived by applying some simplifying assumptions to the cost function in C1, adapting it to the (much more tractable) framework of online convex optimization. The resulting algorithm has very low complexity and is similar to the algorithm TISO discussed in [47]. At each iteration, the proposed algorithm (from the noisy observations with missing values) and the topology

- **C3)** A second algorithm, named Joint Signal and Topology Identification via Recursive Sparse Online learning (JSTIRSO) to improve the tracking capability of JSTISO at the price of a moderate increase in computation. The difference with respect to JSTISO is that the loss function that JSTIRSO optimizes is augmented with an additional term based on the cost function proposed in [47], which is in turn inspired by the celebrated recursive least squares (RLS) algorithm and facilitates the analysis.

- **C4)** To characterize the performance of JSTIRSO when the topology is time-varying, a dynamic regret bound is derived. The derived dynamic regret bound depend on the properties of the data, the error due to missing values, and the parameters of the algorithm.

- **C5)** Finally, the performance of the proposed algorithms is empirically validated through numerical tests.

The rest of the paper is organized as follows: Sec. II presents the model and a batch formulation for tracking of VAR causality graphs. Sec. III introduces the sequential joint tracking and signal estimation, and reviews the online convex optimization approach. To solve the sequential problem in an online fashion, an approximate loss function is derived in Sec. IV, and an algorithm (JSTISO) is derived. An alternative loss function is presented in Sec. V and it is argued why it is expected to yield better tracking performance. The proposed algorithm (JSTIRSO) is characterized analytically (dynamic regret analysis) in Sec. VI. Numerical results are presented in Sec. VII, and Sec. VIII concludes the paper.

**Notation.** Bold lowercase (uppercase) letters denote column vectors (matrices). Operators $\mathbb{E}[\cdot], \partial, (\cdot)^\top, \text{vec}(\cdot)$, and $\lambda_{\max}(\cdot)$, respectively denote expectation, sub-differential, matrix transpose, vectorization, and maximum eigenvalue of a matrix. The operator $\nabla$ denotes gradient and $\nabla^\top$ represents a subgradient. Symbols $0_N$ and $0_{N \times N}$ represent all-zero vector and matrix, $1_N$ all-ones vector, and $I_N$ identity matrix, all of the given size. Finally, $[\cdot]_+ = \max(\cdot, 0)$, and $I$ is the indicator satisfying $I\{x\} = 1$ if $x$ is true and $I\{x\} = 0$ otherwise.

## II. Model and Problem Formulation

Consider a collection of $N$ time series, where $y_n[t], t = 0, 1, \ldots, T - 1$, denotes the value of the $n$-th time series at time $t$. A causality graph $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$ is a graph where the $n$-th vertex in $\mathcal{V} = \{1, \ldots, N\}$ is identified with the $n$-th time series $y_n[t]$ and there is an edge (or arc) from $n'$ to $n$ $(\langle n, n' \rangle \in \mathcal{E})$ if and only if (iff) $y_{n'}[t]$ causes $y_n[t]$ according to a certain causality notion. A prominent notion of causality can be defined using VAR models. To this end, consider the order-$P$ VAR model [48]:

$$y[t] = \sum_{p=1}^{P} A_p y[t-p] + u[t], \quad (1)$$

where $y[t] \equiv [y_1[t], \ldots, y_N[t]]^\top$, $A_p \in \mathbb{R}^{N \times N}$, $p = 1, \ldots, P$, are the matrices of VAR parameters and $u[t] \equiv [u_1[t], \ldots, u_N[t]]^\top$ is the innovation process, generally assumed to be a temporally white, zero-mean stochastic process, i.e., $\mathbb{E}[u[t]] = 0_N$ and $\mathbb{E}[u[t]u^\top[\tau]] = 0_{N \times N}$ for $t \neq \tau$.

With $a_{n,n'}^{(p)}$, the $n$, $n'$-th entry of $A_p$, (1) takes the following
form
\[
y_n[t] = \sum_{n'=1}^{N} \sum_{p=1}^{P} a_{n,n'}^{(p)} y_{n'}[t - p] + u_n[t]
\]
\[
= \sum_{n' \in \mathcal{N}(n)} \sum_{p=1}^{P} a_{n,n'}^{(p)} y_{n'}[t - p] + u_n[t], \tag{2}
\]
for \(n = 1, \ldots, N\), where \(\mathcal{N}(n) \triangleq \{n' : a_{n,n'} \neq 0_P\}\) and
\[
a_{n,n'}^{(p)} \triangleq \begin{bmatrix} a_{n,n',1}^{(p)} & \cdots & a_{n,n',n}^{(p)} \end{bmatrix}^\top. \tag{3}
\]

With this model we can introduce the concept of VAR causality \([49]\), which embodies a similar spirit to that of Granger causality, but is much less challenging to compute. Given a process order \(P\), it is said that time series \(y_i[t]\) \(VAR\)-causes time series \(y_j[t]\) iff the \(P\) most recent values of \(y_j[t]\) carry information that allows to reduce the prediction MSE of \(y_j[t]\) (compared to the optimal prediction based on all other time series in the set under consideration). While in the definition of Granger causality the notion of optimal prediction is not clearly specified, the VAR model allows a clear definition of an optimal predictor.

When \(u[t]\) is a zero-mean and temporally white stochastic process, the term \(\hat{y}_n[t] \triangleq \sum_{n' \in \mathcal{N}(n)} \sum_{p=1}^{P} a_{n,n'}^{(p)} y_{n'}[t - p]\) in (2) is the minimum mean square error estimator of \(y_n[t]\) given the previous values of all time series \(\{y_{n'}[\tau], n' = 1, \ldots, N, \tau < t\}\); see e.g. \([20, \text{Sec. 12.7}]\). The set \(\mathcal{N}(n)\) therefore collects the indices of those time series that participate in this optimal predictor of \(y_n[t]\); in other words, the information provided by time series \(y_{n'}[t]\) with \(n' \notin \mathcal{N}(n)\) is not informative to predict \(y_n[t]\). This allows us to express the definition of VAR-causality in a clearer and more compact way: \(y_{n'}[t] VAR\)-causes \(y_n[t]\) whenever \(n' \in \mathcal{N}(n)\). Equivalently, \(y_n[t] VAR\)-causes \(y_{n'}[t]\) if \(a_{n,n'} \neq 0_P\). VAR causality relations among the \(N\) time series can be represented using a causality graph where \(E \triangleq \{(n,n') : a_{n,n'} \neq 0_P\}\). Clearly, in such a graph, \(\mathcal{N}(n)\) is the in-neighborhood of node \(n\). To quantify the strength of these causality relations, a weighted graph can be constructed by assigning e.g. the weight \(\|a_{n,n'}\|_2^2\) to the edge \((n,n')\).

So far we have considered a static VAR model. However, to analyze dynamic systems in practical applications, a time-varying VAR model (often associated with a dynamic topology) is introduced:
\[
y[t] = \sum_{p=1}^{P} A_p^{(t)} y[t - p] + u[t], \tag{4}
\]
where the parameters \(\{A_p^{(t)}\}_{p=1}^{P}\) follow a certain law of motion such as introduced in \([50, \text{Ch. 18}]\). With these definitions, the problem can be formally stated as: given the observations \(\{y[t]\}_{t=0}^{T-1}\) (in batch form) and the VAR process order, \(P\), find the time-varying VAR coefficients \(\{A_p^{(t)}\}_{p=1}^{P}\) such that it yields sparse topology at each time instant. Without assumptions on the variations of the topologies, the problem involves more unknown variables than data and is ill-posed. In this case, we assume that the variations in the topology are constrained, so that the cumulative norm difference between consecutive sets of parameters does not exceed a given budget \(B\). The formulation in \([26]\) can be extended to a time-varying model as follows:
\[
\begin{aligned}
\arg \min_{\{A_p^{(t)}\}_{p=1}^{P}} & \frac{1}{2(T-P)} \sum_{t=P}^{T-1} \|y[t] - \sum_{p=1}^{P} A_p^{(t)} y[t - p]\|_2^2 \\
+ & \sum_{p=1}^{P} \Omega \left(\{A_p^{(t)}\}_{p=1}^{P}\right) \\
\text{s. t.} & \sum_{t=P+1}^{T-1} \|\text{vec} \left(\{A_p^{(t)}\}_{p=1}^{P}\right) - \text{vec} \left(\{A_p^{(t-1)}\}_{p=1}^{P}\right)\|_2^2 \leq B,
\end{aligned} \tag{5a}
\]
where the first term in the cost function is the least-squares loss, and the second term is a group sparsity-promoting regularization function defined as
\[
\Omega \left(\{A_p^{(t)}\}_{p=1}^{P}\right) \triangleq \lambda \sum_{n=1}^{N} \sum_{n'=1}^{N} \mathbb{1}_{\{n' \neq n\}} \|a_{n,n'}^{(t)}\|_2^2, \tag{6}
\]
where \(a_{n,n'}^{(t)}\) has the same structure as (3) with time-varying VAR parameters. The regularization function \(\Omega\) promotes sparse edges in the causality graphs. The parameter \(\lambda\) is a user-defined constant that controls the sparsity in the edges of the graph. The constraint (5b) restricts the amount of variation in the VAR parameters, and is necessary for the problem to have a meaningful solution (otherwise it would be very ill-posed). In this work, we consider that some data values will be missing (for reasons already stated in the introduction), and the observed values will be affected by measurement noise.

To formulate the problem of estimating the causality graphs with missing values and noise in the observation vector, consider a subset of \(\mathcal{V}\) where the signal is observed, given by \(\mathcal{M}_t \subseteq \mathcal{V}\). The (random) pattern of missing values is collected in the masking vector \(m[t] \in \mathbb{R}^N\) where \(m_n[t], n = 1, \ldots, N\), are i.i.d. Bernoulli random variables taking value \(1\) with probability \(\rho\) and zero with probability \(1 - \rho\). Let \(\hat{y}[t]\) be the observation obtained at time \(t\), given by
\[
\hat{y}[t] = m[t] \odot (y[t] + \epsilon[t]), \tag{7}
\]
where \(\odot\) denotes element-wise product, and \(\epsilon[t]\) is the observation noise vector.

In batch setting, the problem of estimating time-varying topologies with missing values is stated as: given the noisy observations \(\{\hat{y}[t]\}_{t=0}^{T-1}\) with missing values, and the VAR process order \(P\), find the coefficients \(\{A_p^{(t)}\}_{p=1}^{P}\) such that it yields a sparse topology. However, thanks to VAR model, it is easier to estimate the topology from the observation vector directly if the missing values are reconstructed (imputed), and the topology (VAR parameters) helps in such reconstruction. Thus, a natural approach is to jointly estimate the signal and the topology.

In batch setting, the approach advocated in \([46]\) is to solve the following problem, which includes joint estimation of the
signal and the VAR coefficients:

\[
\left\{ \tilde{y}[t], \{ \hat{A}_p^{(t)} \}_{p=1}^P \right\}_{t=p}^{T-1} = \arg\min_{\{ y[t], \{ A_p^{(t)} \}_{p=1}^P \}_{t=p}^{T-1}} \frac{1}{2} \sum_{t=p}^{T-1} \left\| y[t] - \sum_{p=1}^P A_p^{(t)} y[t-p] \right\|_2^2 + \frac{1}{2} \sum_{t=p}^{T-1} \Omega \left( \{ A_p^{(t)} \}_{p=1}^P \right) + \frac{1}{2} \sum_{t=p}^{T-1} \left\| A_p^{(t)} - A_p^{(t-1)} \right\|_F^2, \tag{8}
\]

where the first term is a least-squares (LS) fitting error for all time instants (where the \( t \)-th term in the summation fits the signal based on the \( P \) previous observations and the VAR coefficients at time \( t \)), the second term penalizes the mismatch between the observation vector and the reconstructed signal (recall that \( |M_t| \) is the number of nodes where the signal is observed\(^1\)), the third term is a regularization function that promotes sparsity in the edges, and the fourth term limits the variations in the coefficients (it comes from the dualization of the constraint in (5)). The parameter \( \nu > 0 \) is a constant to control the trade-off between the prediction error based on the VAR coefficients and the mismatch between the measured samples and the signal reported after the reconstruction. The parameter \( \lambda \) controls the sparsity in the edges while \( \beta \) controls the magnitude of the cumulative norm of the difference between consecutive coefficients. The resulting problem in (8) is (separately) convex in \( \{ y[t] \}_{t=p}^{T-1} \) and in \( \{ A_p^{(t)} \}_{p=1}^P \) \( \{ A_p^{(t)} \}_{p=1}^P \), but not jointly convex. A stationary point of (8) can be found via alternating minimization [46, Corollary 1]. Each subproblem in alternating minimization can be solved via proximal gradient descent.

In the next section, we describe how to solve this problem in an online fashion where the data are coming sequentially.

### III. Online Signal Reconstruction and Topology Inference

The batch formulation in (8) uses information from all time instants to produce a sequence of reconstructed signal values and VAR parameter (topology) estimates. On the other hand, an online formulation should allow us to produce such a sequence with minimum delay and with fixed complexity (at the price of lower accuracy). Specifically, here we are interested in an algorithm that, at each time instant \( t \), produces an estimate of \( \tilde{y}[t] \) and \( \{ \hat{A}_p^{(t)} \}_{p=1}^P \) as soon as the partial observation \( y[t] \) is received.

To this end, we design an online criterion such that its sum over time matches the batch objective in (8). As a preliminary step, define

\[
\ell_t \left( \{ y[\tau] \}_{\tau=t-P}^t, \{ \hat{A}_p^{(t)} \}_{p=1}^P \right) \triangleq \\
\frac{1}{2} \left\| y[t] - \sum_{p=1}^P \hat{A}_p^{(t)} y[t-p] \right\|_2^2 + \frac{\nu}{2 |M_t|} \left\| y[t] - m[t] \odot y[t] \right\|_2^2.
\tag{9}
\]

Now we can use the expression above\(^2\), and the definition of \( \Omega(\cdot) \) from (6), to define the dynamic cost function:

\[
c_t \left( \{ y[\tau] \}_{\tau=t-P}^t, \{ A_p^{(t)} \}_{p=1}^P, \{ A_p^{(t-1)} \}_{p=1}^P \right) \triangleq \\
c_t \left( \{ y[\tau] \}_{\tau=t-P}^t, \hat{A}_p^{(t)} \right) + \Omega \left( \{ A_p^{(t)} \}_{p=1}^P \right) + \beta \sum_{t=P}^{T-1} \left\| A_p^{(t)} - A_p^{(t-1)} \right\|_F^2.
\tag{10}
\]

The objective function in (8) can be rewritten as \( \sum_t c_t(\cdot, \cdot, \cdot) \). It becomes clear that producing an estimate of \( y[t] \) and \( \{ A_p^{(t)} \}_{p=1}^P \) does not only have an impact on \( c_t(\cdot, \cdot, \cdot) \), but also on \( c_t(\cdot, \cdot, \cdot, \cdot) \). Such a coupling in time is taken into account in the framework of dynamic programming (or reinforcement learning), where the goal is to find a policy \( \pi \) of the form

\[
\pi : \mathbb{R}^{PN} \times \mathbb{R}^{PN^2} \times \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^{N} \times \mathbb{R}^{N^2 P}
\pi \left( \{ y[\tau] \}_{\tau=t-P}^t, \{ A_p^{(t-1)} \}_{p=1}^P, \tilde{y}[t], m[t] \right) \mapsto \hat{y}[t], \{ A_p^{(t)} \}_{p=1}^P
\tag{11}
\]

such that the cumulative cost is minimized in expectation. Learning such a policy (via e.g., deep reinforcement learning) would be computationally intensive and require a high amount of data, and it is left out of the scope of the present paper. Instead, we propose to approximate such a policy using the much more tractable framework of online convex optimization (reviewed next). Fortunately enough, the structure of (10) resembles that of the composite problems that can be efficiently dealt with via proximal online gradient descent (OGD). In the next section, an approximation of the cost function discussed above will be taken in a way such that we can derive a proximal OGD update over \( \{ A_p^{(t-1)} \}_{p=1}^P \).

In the remainder of this section, the theoretical background of proximal OGD and inexact proximal OGD (IP-OGD) will be introduced. In Sec. IV, we will explain the approximations we take in order to be able to apply the IP-OGD framework [51] to the online problem at hand.

#### A. Theoretical background: composite problems

In the sequel, we present a framework to solve composite-objective optimization problems in an online fashion.

Consider a sequence of functions such that each of them can be split into two parts (a loss function and a regularization function). Generally, each function in the sequence is given by

\[
h_t(a) \triangleq f_t(a) + \Omega_t(a),
\tag{12}
\]

\(^1\)For those time instants where \( |M_t| = 0 \), the term affected by the fraction will not be considered in the optimization, so division-by-zero error is avoided.

\(^2\)The splitting of the arguments of \( \ell_t \) into present and past samples will become useful in subsequent sections.
where \( f_1 : \mathcal{X} \rightarrow \mathbb{R} \) is a general convex loss function, \( \Omega : \mathcal{X} \rightarrow \mathbb{R} \) is the convex regularization function where \( \mathcal{X} \) is a convex set. Note that the function \( \Omega(\cdot) \) can vary with time, however, in this work, it will remain constant.

Given such a sequence of functions, the online learning setting requests to generate, at each time \( t \), a hypothesis or estimate \( a[t] \), given the previous functions \( \{h_t\}_{t=0}^{\tau-1} \). The quality of the proposed estimate \( a[t] \) will be assessed by \( h_t(a[t]) \). Since the estimate must be delivered before \( h_t \) is made available, the possibility of generating good estimates is subject to certain assumptions on how much the sequence of optimal estimates (which is only known in hindsight) changes over time. In the context of this work, \( a[t] \) corresponds to the topology, and the online learning task corresponds to the tracking of the time-varying topologies, subject to the assumption that the topology changes slowly over time.

The performance metric usually considered in online learning algorithms for static problems is the static regret, which compares the algorithm’s performance with a static (constant in time) hindsight solution. Although online algorithms with sublinear regret [52] can be applied to track slowly time-varying solutions, the static regret is not an adequate metric for quantifying how well an algorithm infers time-varying models. To characterize the performance of online algorithms in time-varying scenarios, dynamic regret has become popular for problems where the hindsight solution is also time-varying [53]. Mathematically, the dynamic regret is defined as

\[
R_d(T) \triangleq \sum_{t=1}^{T} \left[ h_t(a^*[t]) - h_t(a[t]) \right],
\]

where \( a[t] \) is the estimate of the online algorithm and \( a^*[t] \) is the optimal solution\(^3\) at time \( t \), given by \( a^*[t] \triangleq \text{arg min}_a h_t(a) \). Note that optimal solutions are time-varying.

Next, we present an online algorithm to solve the composite problem given in (12). It is well known that composite problems can be efficiently solved via proximal methods [54], [55], which are based on the so-called proximity operator, which we briefly discuss next. The proximity (prox) operator of a scaled function \( \eta \Psi \) at point \( v \) is defined by [54]:

\[
\text{prox}_\eta \Psi(v) \triangleq \text{arg min}_{x \in \text{dom } \Psi} \left\{ \Psi(x) + \frac{1}{2\eta} \| x - v \|_2^2 \right\},
\]

where \( \Psi(\cdot) \) is minimized together with a quadratic proximal term, which makes the minimization objective strongly convex. The prox operator of a function at point \( v \) can be interpreted as minimizing the function while being close to \( v \), and the parameter \( \eta \) controls the trade-off between the two objectives. Based on this prox operator, there are various algorithms which work under very general conditions. Usually, the proximal algorithms are used to solve composite problems (differentiable plus non-differentiable term) and they exhibit good convergence guarantees. Some of the existing algorithms such as gradient descent, projected gradient descent, etc. can be shown to be special cases of proximal algorithms.

An extremely popular algorithm for solving composite problems is proximal gradient descent (PGD) [54]. At each iteration, a gradient descent step is performed on the differentiable component of the objective and then the prox operator of the non-differentiable function at the resultant vector is performed. This process is repeated until convergence. In its online version, namely proximal OGD, only one iteration of the proximal gradient is performed at each time instant based on the available data sample, instead of running until convergence. In many cases, the full information about the cost function is not available to the algorithm. To deal with this issue, IP-OGD [51] assumes that an inexact gradient is available and the analysis of the algorithm includes the error between the true gradient and the available inexact gradient. The IP-OGD algorithm enjoys solid theoretical guarantees regarding its performance in tracking time-varying parameters.

### IV. Deriving an approximate loss function

The expressions in the previous section [cf. (10)] represent the problem of joint estimation and reconstruction from a rather ideal point of view because, even though the optimal policy would allow the best possible tracking, finding such a policy is nearly intractable. Fortunately, adding a few simple assumptions can give rise to a composite objective problem that can be solved using the approach described in Sec. III-A.

Our approach consists in treating, at time \( t \), the \( P \) previous reconstructed samples, \( \{\hat{y}[\tau]\}_{\tau=t-P} \), as random variables. Although those variables are dependent of the estimated \( V \) AR parameters, we adopt the simplifying approximation of assuming that they are independent. After doing so, the deterministic function \( c_h(\cdot) \) can be replaced with a random function

\[
C_t \left( \{\hat{y}[\tau]\}, \{A_p^{(t)}\}_{p=1}^P \right) = \\
\ell_t \left( \{\hat{y}[\tau]\}_{\tau=t-P}^t, \{\hat{A}_p^{(t)}\}_{p=1}^P \right) + \Omega \left( \{A_p^{(t)}\}_{p=1}^P \right) \\
+ \beta \sum_{p=1}^P \|A_p^{(t)} - \hat{A}_p^{(t-1)}\|_F^2,
\]

which is jointly convex in its arguments. Notice that, if \( \{\hat{y}[\tau]\}_{\tau=t-P}^t \) and \( \hat{y}[t] \) were equal to the true (yet unobservable) signals \( \{y[\tau]\}_{\tau=t-P}^t \), this setting would be the same that is dealt with in [47], by direct application of proximal OGD. Since the aforementioned signal estimates are inexact versions of the true signals, in the present work we will use the IP-OGD framework discussed in [51] to analyze the regret of the resulting algorithm.

Before proceeding to the formulation of the online algorithm, two remarks are in order.

**Remark 1:** The cost function has as inputs the signal estimate and the \( V \) AR parameters. It is assumed that the \( V \) AR parameters change smoothly with time, but we cannot assume that the signals vary smoothly with time. Recall that in each proximal OGD iteration, a minimization is solved involving a first-order approximation of the loss \( \ell_t \), the regularizer \( \Omega \) (not linearized), and a proximal term that ensures that the variable estimated at time \( t \) is close in norm to its previous estimate.
but not \( y[t] \).

**Remark 2:** As a consequence of the simplifying assumption of random independent reconstructed samples, the function \( C_t(\cdot) \) becomes separable across nodes.

As a consequence of the approximations above, the joint optimization over \( \{ A_p^t \}_{p=1}^P \) and \( y[t] \) can be reformulated into an optimization only over \( \{ A_p^t \}_{p=1}^P \) as follows. Note first that minimizing it can be split into first minimizing over \( y[t] \) and then over \( \{ A_p^t \}_{p=1}^P \). The first minimization admits a closed form, which in turn is convex in \( \{ A_p^t \}_{p=1}^P \). Specifically, we can write

\[
\min_{y[t], \{ A_p^t \}_{p=1}^P} \sum_{t \in \mathcal{T}} C_t \left( y[t], \{ A_p^t \}_{p=1}^P \right) = \min_{y[t], \{ A_p^t \}_{p=1}^P} \left( \sum_{t \in \mathcal{T}} L_t \left( \{ A_p^t \}_{p=1}^P \right) \right),
\]

where

\[
L_t \left( \{ A_p^t \}_{p=1}^P \right) \triangleq \min_{y[t]} \ell_t \left( \{ y[\tau] \}_{\tau = t - P}^{t-1}, y[t], \{ A_p^t \}_{p=1}^P \right),
\]

and the analytical minimization in (17) is shown in Sec. IV-A. Once a closed form is available for \( L_t \), IP-OGD can be applied. The inexactness comes from the previously estimated (reconstructed) \( \{ y[\tau] \}_{\tau = t - P}^{t-1} \). That is what makes \( L_t \) a random function, more specifically an inexact version of the “true” loss function which would be given by

\[
L_t^{\text{true}} \left( \{ A_p^t \}_{p=1}^P \right) \triangleq \min_{y[t]} \ell_t \left( \{ y[\tau] \}_{\tau = t - P}^{t-1}, y[t], \{ A_p^t \}_{p=1}^P \right),
\]

but is unavailable because the true signal values \( \{ y[\tau] \}_{\tau = t - P}^{t-1} \) would be needed to evaluate it.

Observe that the loss function in (17) is separable across nodes, i.e.,

\[
L_t \left( \{ A_p^t \}_{p=1}^P \right) = \sum_{n=1}^N L_t^{(n)}(a_n[t]) = \sum_{n=1}^N \min_{y_n[t]} \ell_t^{(n)}(\hat{y}[t], y_n[t], a_n[t]),
\]

where

\[
\ell_t^{(n)}(\hat{y}[t], y_n[t], a_n[t]) \triangleq \frac{1}{2} \left( (y_n[t] - \hat{g}[t]^\top a_n[t])^2 + \frac{\nu m_n[t]}{|M_n|}(y_n[t] - \hat{y}_n[t])^2 \right),
\]

with

\[
\hat{g}[t] = \text{vec} \left( [\hat{y}[t-1], \ldots, \hat{y}[t-P]]^\top \right),
\]

and

\[
L_t^{(n)}(a_n[t]) \triangleq \min_{y_n[t]} \ell_t^{(n)}(\hat{y}[t], y_n[t], a_n[t]).
\]

To arrive at the loss function, the minimizer (signal reconstruction) will be derived; then, a closed-form expression for \( L_t^{(n)} \) will be obtained.

### A. Signal reconstruction and loss function in closed form

We discuss here the (sub)problem of estimating the signal from a noisy observation vector with missing values given a (fixed) topology. The resulting estimator is a convex combination of the signal prediction via the VAR process and the values present in the observation vector. More formally, the reconstruction subproblem consists in estimating \( \hat{y}[t] \) given \( \hat{y}[t], m[t], g[t], \) and \( \{ A_p^t \}_{p=1}^P \). Notice from (9) that \( \hat{y}[t] \) and \( m[t] \) are implicit in the definition of \( \ell_t(\cdot) \):

\[
\hat{y}[t] = \min_{y[t]} \ell_t(\hat{g}[t], y[t], a_n[t]).
\]

The solution for the \( n \)-th entry of \( \hat{y}[t] \) is

\[
\hat{y}_n[t] = \arg \min_{y_n[t]} \ell_t^{(n)}(\hat{g}[t], y_n[t], a_n[t]),
\]

which has a closed form given by

\[
\hat{y}_n[t] = \left( 1 - U_n[t] \right) \hat{g}[t]^\top a_n[t] + U_n[t] \hat{y}_n[t],
\]

where

\[
U_n[t] \triangleq \frac{\nu}{|M_n| + \nu} m_n[t].
\]

Observe that \( U_n[t] \in [0, \nu/(1 + \nu)) \) holds \( \forall t, n \).

After substituting (23) into (21) and simplifying, the loss function can be expressed as

\[
L_t^{(n)}(a_n[t]) = \frac{1}{2} U_n[t] (\hat{y}_n[t] - \hat{g}^\top[t] a_n[t])^2
\]

and it will be used in Sec. IV-B to derive the IP-OGD iterates. Proximal OGD involves linearizing part of the objective, in this case \( L_t^{(n)}(\cdot) \), which requires computing the gradient. The latter is given by

\[
\nabla L_t^{(n)}(a_n[t]) = U_n[t] (\hat{g}[t] \hat{g}^\top[t] a_n[t] - \hat{y}_n[t] \hat{g}[t]).
\]

### B. Application of IP-OGD to Joint Signal and Topology Estimation

The gradient defined in (26) depends on \( \hat{g}[t] \), which is conformed using the estimates \( \hat{y}[t-P] \) for \( \hat{g}[t] \), which in turn will generally differ from the true signals. This is translated into an error in the gradient and this is why IP-OGD is advocated here.

Let \( f_t^{(n)} \) be a general loss function, and let \( \mathcal{F}_t^{(n)} \) be a random function that is an inexact version of \( f_t^{(n)} \). Using \( \mathcal{F}_t^{(n)} \) and \( \Omega^{(n)}(a_n) \) in (12), with a constant step size \( \alpha \), the IP-OGD iteration is:

\[
a_n[t] = \text{prox}_{\Omega^{(n)}} \left( a_n[t-1] - \alpha \nabla \mathcal{F}_t^{(n)}(a_n[t-1]) \right).
\]

Let \( a_t^{(n)}[t] \triangleq a_n[t-1] - \alpha \nabla \mathcal{F}_t^{(n)}(a_n[t-1]) \), and

\[
a_t^{(n)}[t] = \left( a_t^{(1,n)}[t], \ldots, a_t^{(N,n)}[t] \right)^\top,
\]

which enables us to write the above update expression as

\[
a_n[t] = \text{prox}_{\Omega^{(n)}} \left( a_t^{(n)}[t] \right)
\]

\[
= \arg \min_{z_n} \left( \Omega^{(n)}(z_n) + \frac{1}{2\alpha} \| z_n - a_t^{(n)}[t] \|_2^2 \right).
\]
Using the regularizing function $\Omega^{(n)}(a_n) \triangleq \lambda \sum_{n'=1}^{N} \mathbb{I}\{n \neq n'\} \|a_{n,n'}\|_2^2$, [cf. (6)] that $\Omega = \sum_{n=1}^{N} \Omega^{(n)}$, the update yields

$$a_n[t] = \arg\min_{(z_{n,n'})_{n'=1}^{N}} \left( \lambda \sum_{n'=1}^{N} \mathbb{I}\{n \neq n'\} \|z_{n,n'}\|_2^2 + \frac{1}{2\alpha} \sum_{n'=1}^{N} \|z_{n,n'} - a_{n,n'}^f[t]\|_2^2 \right),$$

which is separable across $n'$ and the solution to the $n'$-th subproblem is given by the group soft-thresholding:

$$a_{n,n'}[t] = \arg\min_{z_{n,n'}} \left( \mathbb{I}\{n \neq n'\} \|z_{n,n'}\|_2^2 + \frac{1}{2\alpha} \|z_{n,n'} - a_{n,n'}^f[t]\|_2^2 \right) = a_{n,n'}^f[t],$$

(recall that $a_{n,n'}^f[t]$ is a subvector of $a_n[t]$ as defined in (28)). The algorithm JSTISO, which is intended at minimizing $C_l(\cdot,\cdot)$ in (15), is obtained when $F_n(t)$ is set to be $L_n^{(n)}$. All required steps are summarized in Procedure 1.

**Remark 3:** For those time instants and nodes where an entry is missing, $\nabla_{a_n[t]} L_n^{(n)} = 0_N$, but Procedure 1 applies the soft-thresholding operator (29) to the corresponding coefficients. While this may seem counterintuitive, the shrinking is justified by the model at hand. The time-varying parameters are modeled as a random walk whose innovations are compound by a) a Gaussian distributed term, plus b) a term that attracts the VAR parameters towards 0 (which is the source of sparsity). The term a) justifies the Frobenius norm in and the term b) justifies the presence of $\Omega$ in (8).

**Remark 4:** The proposed algorithm only differs from TISO [47] in the estimate refinement (line 11). Actually, if line 3 is modified to obtain $\hat{g}[t]$ directly from the observations, the resulting parameter estimates $\{a_{n}[t]\}_{n=1}^{N}$ will coincide with those of TISO.

V. AN ALTERNATIVE LOSS FUNCTION FOR IMPROVED TRACKING

The loss function in the previous approach is an instantaneous loss, which only depends on the current sample. While this keeps the complexity of the iterations very low, and may be sufficient for online estimation of a static VAR model, it is sensitive to noise and input variability, and thus it is expected to perform poorly when attempting at tracking a time-varying model. In [47], a running average loss function is designed drawing inspiration from the relation between least mean squares (LMS) and recursive least squares (RLS) to improve the tracking capabilities of TISO. In this paper, similar steps will lead to a second approach, where a running average loss function is adopted, which depends on the past reconstructed signal values. In this second approach, the loss function is set as

$$\tilde{\ell}_t \left( \{\hat{y}[\tau]\}_{\tau=0}^{t-1}, y[t], \{A_p^{(t)}\}_{p=1}^{P} \right) = \frac{1}{2} \left\| y[t] - \sum_{p=1}^{P} A_p^{(t)} \hat{y}[t-p] \right\|_2^2 + \frac{\nu}{2|\mathcal{M}|} \left\| y[t] - m[t] \odot y[t] \right\|_2^2 + \frac{1}{2} \sum_{\tau=p}^{t-1} \left\| \hat{y}[\tau] - \sum_{p=1}^{P} A_p^{(t)} \hat{y}[\tau-p] \right\|_2^2,$$

(30)

where $\gamma$ is a user-selected forgetting factor which controls the weight of past (reconstructed) samples of $y[t]$. The modeling principles in the previous section (treating the previously reconstructed samples as a random variable, and minimizing over $y[t]$) are applied to the alternative deterministic loss $\tilde{\ell}_t$, enabling to define the random loss function $\hat{L}_t$ as

$$\hat{L}_t \left( \{A_p^{(t)}\}_{p=1}^{P} \right) \triangleq \min_{y[t]} \tilde{\ell}_t \left( \{\hat{y}[\tau]\}_{\tau=0}^{t-1}, y[t], \{A_p^{(t)}\}_{p=1}^{P} \right),$$

(31)

which can be rewritten in terms of $\ell_t$ as

$$\hat{L}_t \left( \{A_p^{(t)}\}_{p=1}^{P} \right) = \min_{y[t]} \ell_t \left( \{\hat{y}[\tau]\}_{\tau=0}^{t-1}, y[t], \{A_p^{(t)}\}_{p=1}^{P} \right) + \frac{1}{2} \sum_{\tau=p}^{t-1} \left\| \hat{y}[\tau] - \sum_{p=1}^{P} A_p^{(t)} \hat{y}[\tau-p] \right\|_2^2.$$

(32)

Regarding the signal reconstruction, the minimizer of (31) is:

$$\hat{y}[t] = \arg\min_{y[t]} \ell_t \left( \{\hat{y}[\tau]\}_{\tau=0}^{t-1}, y[t], \{A_p^{(t)}\}_{p=1}^{P} \right) = \arg\min_{y[t]} \frac{1}{2} \left\| y[t] - \sum_{p=1}^{P} A_p^{(t)} \hat{y}[t-p] \right\|_2^2 + \frac{\nu}{2|\mathcal{M}|} \left\| y[t] - m[t] \odot y[t] \right\|_2^2.$$

(33)

Observe that (33) coincides with the reconstruction problem in (22) and, therefore, its solution is given by (23).
Next, to derive the closed-form solution for $\tilde{L}_t$ in this approach, we substitute the closed-form expression of $\tilde{y}[t]$ from (23) into (31):

$$
\tilde{L}_t \left( \left\{ A_t^{(p)} \right\}^P_{p=1} \right) =
\frac{1}{2} \sum_{n=1}^{N} \left[ U_n[t](\tilde{y}_n[t] - \hat{g}^T[t]a_n[t])^2 \right] + \frac{1}{2} \sum_{n=1}^{N} \left( \sum_{\tau = T}^{t-1} \gamma^{t-\tau} \tilde{y}_n[\tau] \right) + \gamma a_n[t] \Phi[t-1]a_n[t] - 2\gamma \tilde{r}_n[t-1]a_n[t],
$$
(34)

where

$$\tilde{\Phi}[t] \triangleq \sum_{\tau = T}^{t} \gamma^{t-\tau} \hat{g}[\tau] \hat{g}^T[\tau],$$
(35a)

$$\tilde{r}_n[t] \triangleq \sum_{\tau = T}^{t} \gamma^{t-\tau} \tilde{y}_n[\tau] \hat{g}[\tau].$$
(35b)

The variables above can be efficiently computed via recursive expressions\(^4\). Note that $\tilde{L}_t$ is also separable across nodes, i.e.,

$$\tilde{L}_t(\cdot) = \sum_{n=1}^{N} \tilde{L}_t^{(n)}(\cdot),$$
(36)

where

$$\tilde{L}_t^{(n)}(a_n) \triangleq \tilde{L}_t^{(n)}(a_n) + \sum_{\tau = T}^{t-1} \gamma^{t-\tau} \tilde{y}_n^2[\tau] + \gamma a_n^T \Phi[t-1]a_n - 2\gamma \tilde{r}_n[t-1]a_n.$$

The algorithm JSTIRSO is obtained when $F_t^{(n)}$ is set to be $\tilde{L}_t^{(n)}$, following similar steps to those in Sec. IV-B. The gradient of $\tilde{L}_t^{(n)}$ w.r.t. $a_n[t]$ is given by

$$\nabla \tilde{L}^{(n)}_t(a_n[t]) = U_n[t] \left( \hat{g}[t] \hat{g}^T[t] a_n[t] - \tilde{y}_n[t] \hat{g}[t] \right) + \gamma \Phi[t-1]a_n[t] - \gamma \tilde{r}_n[t-1].$$
(38)

All required steps are summarized in Procedure 2. The initial values for $\Phi[P-1]$ and $\tilde{r}[P-1]$ can be set depending on available prior information; if no such information is available, one can choose a small $\sigma$ and set $\Phi[P-1] = \sigma^2 I$, and $\tilde{r}[P-1] = 0 \forall n$.

**Remark 5:** An observation similar to Remark 4 applies for JSTIRSO and TIRSO [47].

VI. PERFORMANCE ANALYSIS

To analyze the performance of JSTIRSO, we present analytical results in this section. First, the assumptions considered in the analysis are stated and then, two lemmas followed by the main theorem about the dynamic regret bound of JSTIRSO are presented. Moreover, a third lemma stating a bound on the error in the gradient is presented and discussed. Finally, a corollary with a simpler dynamic regret bound is presented.

\(^4\)The recursive expressions are explicit in lines 4 and 6 in Procedure 2.

**Procedure 2** Tracking time-varying topologies with missing data via JSTIRSO

**Input:** $P, \lambda, \alpha, \nu, \gamma, \sigma^2$

**Initialization:** $\{ \tilde{y}[\tau] \}_{\tau=0}^{P-1}$, $\{ a_n[P-1], \tilde{r}_n[P-1] \}_{n=1}^{N}$, $\Phi[P-1]$

1: for $t = P, P+1, \ldots$
2: Receive observation $\tilde{y}[t]$ and masking vector $m[t]$
3: Obtain $\hat{g}[t]$ from $\{ \tilde{y}[t - \tau] \}_{\tau=1}^{P}$ via (20)
4: $\Phi[t] = \Phi[t-1] + \hat{g}[t] \hat{g}^T[t]$
5: for $n = 1, \ldots, N$
6: $\tilde{r}_n[t] = \gamma \tilde{r}_n[t-1] + \tilde{y}_n[t] \hat{g}[t]$
7: Obtain $U_n[t]$ via (24)
8: $a_n[t] = a_n[t-1] - \alpha \nabla \tilde{L}_t^{(n)}(\tilde{a}_n[t-1])$ [cf. (38)]
9: for $n' = 1, 2, \ldots, N$
10: $\tilde{a}_n[n', t] = \tilde{a}_t^{(n')}[t] \left[ 1 - \frac{\alpha}{\| \tilde{a}_n[n', t] \|_2} \right] + \tilde{a}_t^{(n')}[t]$
11: end for
12: $\tilde{a}_n[t] = [\tilde{a}_n[1, t], \ldots, \tilde{a}_n[N, t]]^T$
13: Compute $\tilde{y}_n[t]$ via (23)
14: end for
15: Output $\{ \tilde{a}_n[t] \}_{n=1}^{N}, \tilde{y}[t]$
16: end for

To quantify the inexactness in our algorithm, we need to define the following quantities:

$$g[t] \triangleq \text{vec} \left( \begin{bmatrix} \tilde{y}[t-1] & \ldots & \tilde{y}[t-P] \end{bmatrix}^T \right),$$
(39a)

$$\Phi[t] \triangleq \sum_{\tau = T}^{t} \gamma^{t-\tau} g[\tau] g^T[\tau],$$
(39b)

$$r_n[t] \triangleq \sum_{\tau = T}^{t} \gamma^{t-\tau} \tilde{y}_n[\tau] g[\tau],$$
(39c)

which can be thought as the true versions of $\Phi[t]$ and $\tilde{r}_n[t]$. The following assumptions will be considered for the characterization of JSTIRSO:

A1. **Bounded samples:** There exists $B_g > 0$ such that $|y_n[t]|^2 \leq B_g$, $|\tilde{y}_n[t]|^2 \leq B_g$, and $|\tilde{y}_n[t]|^2 \leq B_y \forall n, t$.

A2. **Bounded minimum eigenvalue of $\Phi$** and $\tilde{\Phi}$: There exists $\beta_{\text{min}} > 0$ such that $\lambda_{\text{min}}(\Phi[t]) \geq \beta_{\text{min}}$ and $\lambda_{\text{min}}(\tilde{\Phi}[t]) \geq \beta_{\text{min}} \forall t \geq P$.

A3. **Bounded maximum eigenvalue of $\Phi$** and $\tilde{\Phi}$: There exists $L > 0$ such that $\lambda_{\text{max}}(\Phi[t]) \leq L$ and $\lambda_{\text{max}}(\tilde{\Phi}[t]) \leq L, \forall t \geq P$.

A4. **Bounded errors in $g$, $\Phi$, $r_n$ due to noise and missing values:**

$$\| g[t] - \hat{g}[t] \|_2 \leq B_g \forall t,$$
(40a)

$$\lambda_{\text{max}}(\hat{\Phi}[t] - \Phi[t]) \leq B_\Phi \forall t,$$
(40b)

$$\| \tilde{r}_n[t] - r_n[t] \|_2 \leq B_r \forall n, t.$$  
(40c)

The forthcoming results depend on the error in the gradient, given by

$$e^{(n)}[t] \triangleq \nabla \tilde{L}_t^{(n)}(a_n[t]) - \nabla \tilde{L}_t^{(n)}(\tilde{a}_n[t])$$
(41)
The dynamic regret for JSTIRSO corresponding to the $n$-th node is defined as
\[
\tilde{L}_t^{(n)\text{true}}(a_n[t]) \equiv \min_{y_n[t]} \tilde{L}_t^{(n)} \left( \{y[t]\}_{t=0}^{T-1}, y_n[t], a_n[t] \right)
\]
is the true (exact) gradient (where $\{y[t]\}_{t=0}^{T-1}$ are the (unobservable) true signal values), and $\nabla \tilde{L}_t^{(n)}(a_n)$ is the inexact gradient defined in (38). The latter is inexact due to the error in the reconstructed entries of $\tilde{g}$, and the error in $\tilde{g}$ comes in turn from the missing values and noisy observations.

Dynamic regret analysis is generally expressed in terms of metrics that express how challenging tracking becomes, e.g., how fast the optimal parameters vary. Specifically in our case, the dynamic regret will be expressed in terms of the variation in consecutive optimal solutions (often referred to as path metrics that express how challenging tracking becomes, e.g., turn from the missing values and noisy observations).

Lemma 2: Under assumptions A1 and A3, we have
\[
\left\| \nabla \tilde{L}_t^{(n)}(a_n[t]) \right\|_2 \leq \frac{\nu}{1 + \nu} \left( PB_y + 2 \sqrt{P}NB_y B_g + B_g^2 + \gamma L \frac{1}{1 - \gamma} \right) \times \frac{1}{\beta \gamma} \left( \frac{1}{1 + \nu} \right) \sqrt{P}NB_y + \sqrt{P^2NB_y^2 - 1} \gamma \sqrt{P}NB_y.
\]
Proof: See Appendix A in the supplementary material.

Lemma 2: All the subgradients of the regularization function $\Omega^{(n)}$ are bounded by $\lambda \sqrt{N}$, i.e., $\|u_t\|_2 \leq \lambda \sqrt{N}$, where $u_t \in \partial \Omega^{(n)}(a_n[t])$.

Proof: See the proof of Theorem 5 in [47].

Next, we present a bound on the dynamic regret of JSTIRSO. Theorem 1: Under assumptions A1, A2, A3, and A4, let $\{a_n[t]\}_{t=P}^T$ be generated by JSTIRSO (Procedure 2) with a constant step size $\alpha \in (0, 1/L)$. If there exists $\sigma$ such that
\[
\|a_n[t] - a_n[t - 1]\|_2 \leq \sigma, \quad \forall t \geq P + 1,
\]
then the dynamic regret of JSTIRSO satisfies:
\[
\hat{R}_{n}^{(n)}[T] \leq \frac{1}{\alpha \beta \nu} \left[ B_v + \lambda \sqrt{N} \right] \left( \|a_n[P] - a_n[P]\|_2 + W^{(n)}[T] \right) + \alpha E^{(n)}[T],
\]
where $B_v$ is defined in (45).

Proof: In order to derive the dynamic regret of JSTIRSO, since $\tilde{h}_t$ is convex, we have by definition
\[
\hat{h}_t^{(n)}(a_n[t]) \geq \tilde{h}_t^{(n)}(a_n[t]) + \left( \nabla^s \tilde{h}_t^{(n)}(a_n[t]) \right)^T (a_n[t] - a_n[t]),
\]
\forall a_n[t], a_n[t]$ where a subgradient of $\tilde{h}_t^{(n)}$ is given by $\nabla^s \tilde{h}_t^{(n)}(a_n[t]) = \nabla \tilde{L}_t^{(n)}(a_n[t]) + u_t$ with $u_t \in \partial \Omega^{(n)}(a_n[t])$. Rearranging (48) and summing both sides of the inequality from $t = P$ to $T$ results in:
\[
\sum_{t=P}^{T} \left( \tilde{h}_t^{(n)}(a_n[t]) - \tilde{h}_t^{(n)}(a_n[t]) \right) \leq \sum_{t=P}^{T} \left( \nabla^s \tilde{h}_t^{(n)}(a_n[t]) \right)^T (a_n[t] - a_n[t]).
\]
(49)

By applying the Cauchy-Schwarz inequality on each term of the summation in the r.h.s. of the above inequality, we obtain
\[
\sum_{t=P}^{T} \left( \tilde{h}_t^{(n)}(a_n[t]) - \tilde{h}_t^{(n)}(a_n[t]) \right) \leq \sum_{t=P}^{T} \left( \nabla^s \tilde{h}_t^{(n)}(a_n[t]) \right)^2 \|a_n[t] - a_n[t]\|_2.
\]
(50)

Next, we apply Lemma 2 in [51] in order to bound $\|\nabla^2 \tilde{h}_t^{(n)}(a_n[t])\|_2$. From the definition of $\nabla^s \tilde{h}_t^{(n)}(a_n[t])$ and by the triangular inequality, we have
\[
\|\nabla^s \tilde{h}_t^{(n)}(a_n[t])\|_2 \leq \|\nabla \tilde{L}_t^{(n)}(a_n[t])\|_2 + \|u_t\|_2.
\]
(51)

From Lemma 1 and Lemma 2, we have $\|\nabla^2 \tilde{h}_t^{(n)}(a_n[t])\|_2 \leq B_v + \lambda \sqrt{N}$. Substituting this bound into (50) leads to:
\[
\sum_{t=P}^{T} \left( \tilde{h}_t^{(n)}(a_n[t]) - \tilde{h}_t^{(n)}(a_n[t]) \right) \leq \sum_{t=P}^{T} \left[ B_v + \lambda \sqrt{N} \right] \|a_n[t] - a_n[t]\|_2.
\]
(52)

Next, we apply Lemma 2 in [51] in order to bound $\sum_{t=P}^{T} \|a_n[t] - a_n[t]\|_2$ in (52). The hypotheses of Lemma 2 are Lipschitz smoothness of $\tilde{L}_t^{(n)}$, Lipschitz continuity of $\Omega^{(n)}$, and strong convexity of $\tilde{L}_t^{(n)}$. Lipschitz continuity of $\Omega^{(n)}$ is proved in Lemma 2 whereas strong convexity of $\tilde{L}_t^{(n)}$ is implied by the assumption A2. To verify that $\tilde{L}_t^{(n)}$ is Lipschitz-smooth, it suffices to realize that $\tilde{L}_t^{(n)}$ is twice-differentiable, and thus assumption A3 is equivalent to saying that $\tilde{L}_t^{(n)}$ is $L$-Lipschitz smooth.
it to bound \( \| \tilde{a}_n[t] - \tilde{a}^\circ_n[t] \|_2 \) in (52) yields:

\[
\sum_{t=P}^T \left[ \tilde{h}_t^{(n)}( \tilde{a}_n[t] ) - \tilde{h}_t^{(n)}( \tilde{a}^\circ_n[t] ) \right] \leq \\
\frac{1}{\alpha \beta \ell} \left[ B_v + \lambda \sqrt{N} \right] \left( \| a_n[P] - \tilde{a}^\circ_n[P] \|_2 + W^{(n)}[T] + \alpha E^{(n)}[T] \right).
\] (53)

This concludes the proof (note that initializing \( \tilde{a}_n[P] = 0_{N,P} \) can lead to further simplification).

The bound on the dynamic regret for JSTIRSO depends on \( W^{(n)}[T] \) and \( E^{(n)}[T] \), which formalizes how much the variability and uncertainty affect the parameter estimation. If both the path length \( W^{(n)}[T] \) and the cumulative error \( E^{(n)}[T] \) are sublinear, then the dynamic regret bound becomes sublinear. This will happen in a setting where the variations in the model, the probability of missing data, and the noise level are all vanishing, e.g., when the estimation starts during a transient period after which the conditions become milder. In many practical settings however, these assumptions will not hold, yielding a dynamic regret that grows linearly with time. Requiring sublinear regret in these settings is arguably unpractical, and the rate of growth of the regret can be used as a benchmark to compare different approaches.

The cumulative error \( E^{(n)}[T] \) can be bounded as a function of the quantities introduced in A4 (related to the inexactness of the reconstructed samples). The following lemma establishes that under such assumptions, the error on the gradient (i.e., \( \| e^{(n)}[t] \|_2 \) is always bounded.

**Lemma 3:** Under assumptions A1 and A4, let \( \{ \tilde{a}_n[t] \}_{t=P}^T \) be generated by JSTIRSO (Procedure 2) with a constant step size \( \alpha \in (0, 1/L] \). Then, the error associated with the inexact gradient [cf. (41)] is bounded as \( \| e^{(n)}[t] \|_2 \leq B_e \), where

\[
B_e \triangleq \left( \gamma B_y + \left( \frac{\nu}{1 + \nu} \right) \left( 2 \sqrt{P N B_y B_g + B_y^2} \right) \right) \times \frac{\sqrt{P N B_y}}{\beta \ell} \left( \frac{\nu}{1 + \nu} + \frac{1}{1 - \gamma} \right) + \gamma B_r + \left( \frac{\nu}{1 + \nu} \right) B_g \sqrt{B_y}.
\] (54)

**Proof:** See Appendix B in the supplementary material.

This bound depends on three kinds of quantities: a) bounds related to the inexactness of the reconstructed signal, b) simple properties of the data time series, and c) the hyperparameters \( \nu \) and \( \gamma \). Note that \( \| e^{(n)}[t] \|_2 \) and \( E^{(n)}[T] \) are related via (43). In those cases where the sources of uncertainty are such that \( \| e^{(n)}[t] \|_2 \) does not vanish, the above bound can be used to replace \( E^{(n)}[T] \) in the regret bound in (47) with an expression that depends on the quantities expressed in A4.

**Corollary 1:** Under the hypotheses in Theorem 1, the dynamic regret of JSTIRSO satisfies:

\[
\tilde{R}_d^{(n)}[T] \leq \frac{1}{\alpha \beta \ell} \left[ B_v + \lambda \sqrt{N} \right] \left( \| a_n[P] - \tilde{a}^\circ_n[P] \|_2 + W^{(n)}[T] \right) + \alpha T B_e.
\] (55)

Observe that the above regret bound is linear in \( T \), and this case was commented after Theorem 1. The growth rate of the dynamic regret is \( \alpha B_e \).

Intuitively, the dynamic regret characterizes the ability to predict the next signal observation from the estimated parameters and reconstructed signals. A remaining challenge is to determine under which conditions the algorithms are able to identify parameters and signals. This is important because, under identifiability conditions, one could claim that the lower the regret bound is, the closer the reconstructed signals will be to the true signals. Consequently, apart from obtaining a smaller value of \( B_g \), also \( \{ \Phi[t], \tilde{r}_n[t] \} \) will become closer to the (not observable) \( \{ \Phi[t], r_n[t] \} \), which will be associated with smaller values of the quantities \( B_g \), \( B_r \). The dependency of these bounds on the regret and the interaction between such bounds are topics that lie out of the scope of the present work, and could give rise to improved regret bounds.

**VII. Numerical Tests**

To analyze the performance of the proposed algorithms, the metrics that have been evaluated are the mean squared deviation (NMSD) for both the signal and the VAR parameters (topology). The NMSD for the signal is given by

\[
\text{NMSD}_s[t] = \frac{\mathbb{E}[\| y[t] - \hat{y}[t] \|_2^2]}{\mathbb{E}[\| y[t] \|_2^2]},
\] (56)

where \( y[t] \) is the true signal while \( \hat{y}[t] \) is the reconstructed signal. The NMSD for the graph (topology) is defined as:

\[
\text{NMSD}_g[t] = \frac{\mathbb{E}[\sum_{n=1}^N \| a_n[t] - a_n^{true}(t) \|_2^2]}{\mathbb{E}[\sum_{n=1}^N \| a_n^{true}(t) \|_2^2]},
\] (57)

which measures the difference between the estimates \( \{ a_n[t] \}_t \) and the time-varying true VAR coefficients \( \{ a_n^{true}(t) \}_t \).

1) Data generation: We consider a dynamic VAR model where the coefficients change abruptly in two specific points in time. To generate the synthetic data, an Erdős-Rényi random graph is generated with edge probability \( p_e \) and self-loop probability 1. This random graph underlies the data generation, and its binary adjacency matrix determines which entries of the V AR coefficients change abruptly in two specific points. This is performed by generating each time a new set of V AR coefficients while keeping the binary adjacency matrix fixed. The innovation process samples are drawn independently as \({{\{ \{ \phi[t], \tilde{r}_n[t] \} \}}_t} \) i.i.d. from a standard normal distribution. Each of the matrices \( \{ \{ a_n^{true}[t] \}_{p=0}^P \}_{t=1}^T \) is then scaled down by a constant that ensures that the V AR process is stable [48]. The innovation process samples are drawn independently as \( u[t] \sim \mathcal{N}(0_N, \sigma_u^2 I_N) \). At \( t = T/3 \) and \( t = 2T/3 \), the model changes abruptly from one model to another model. This is performed by generating each time a new set of V AR coefficients while keeping the binary adjacency matrix fixed.

2) Competing alternatives: The performance of JSTISO (Procedure 1) and JSTIRSO (Procedure 2) is evaluated and compared with that for two competing alternatives. The first alternative is a simple procedure based on TIRSO [47] where the missing values are imputed directly as their predicted values via the V AR model (1), and the noisy samples are not refined: this procedure is referred to as ‘NaiveTIRSO’. The second alternative is an adaptation of JISGOT [46, Algorithm 4], which, is to the best of our knowledge, the state of the art in joint signal and topology estimation. The JISGOT algorithm refines the previous \( P \) signal estimates and runs
several iterations at each time instant, incurring higher computational complexity. The values for the parameter $\nu$ in JSTISO, JSTIRSO, and JISGoT are selected via grid search to minimize the squared deviation for a validation signal. The values of the regularization parameter $\lambda$ were selected by hand to yield a moderate level of sparsity.

3) Discussion of results: In Fig. 1 and Fig. 2, the NMSD for the signal estimation [cf. (56)] and the NMSD for the topology estimation [cf. (57)] are respectively presented for the four algorithms under test.

It can be observed that JSTISO tracks the topology more slowly than the other three algorithms in Fig. 2 as expected since this procedure disregards the past completely. Both JSTISO and JISGoT outperform NaiveTIRSO in terms of signal estimation in Fig. 1 with no significantly different margins, suggesting that the proposed denoising strategy has an effect similar to that implemented in the current adaptation of JISGoT. Note, however, that JSTISO requires less computation than JISGoT. As a side note, NaiveTIRSO exhibits a decent adaptivity after abrupt changes, but its NMSD saturates earlier than that of the other algorithms.

JSTIRSO achieves a lower NMSD for both the signal and graph after a sufficient number of time instants after an abrupt change. Expectedly, due to the choice of the loss function in JSTIRSO, it attains lower levels of NMSD and NMSD than that of JISGoT, despite JSTIRSO does not do any refining of the previous signal estimates. These results suggest that an algorithm combining the strategies developed here with those in [46] could give even better results at the cost of a moderate increase in complexity, which is left as a possible future research avenue.

VIII. CONCLUSIONS

To track time-varying topologies from noisy observations in the presence of missing data, two online algorithms, i.e., JSTISO and JSTIRSO, have been proposed by minimizing an online joint optimization criterion. Thanks to a carefully formulated loss function, joint signal and topology estimation can be carried out efficiently (especially in the case of the low-complexity JSTISO); moreover, the performance of JSTIRSO has been characterized theoretically. To this end, a dynamic regret bound has been derived as a function of the path length (which quantifies the variation in the topologies) and cumulative error on the gradient (which quantifies the effect of noise and missing values). The error on the gradient is in turn bounded [cf. Lemma 3] as a function of the maximum deviation of the inexact variables $\hat{\mathbf{g}}, \hat{\Phi}[t], \hat{r}[t]$ from their associated true values. The bound on the dynamic regret becomes sublinear in scenarios where the variation in the time-varying topologies, the probability of missing data, and the observation noise level are vanishing with time. Numerical results have shown that JSTISO and JSTIRSO can track the time-varying topologies from noisy observations with missing values. Future research avenues include the combination of the proposed strategy with tools related to Kalman filtering and smoothing [46] to ascertain its improvement in terms of performance.

REFERENCES

[1] E. D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, Springer, New York, 2009.
[2] E. Isufi, A. Loukas, N. Perraudin, and G. Leus, “Forecasting time series with arma recursions on graphs,” arXiv preprint arXiv:1810.08581, 2018.
[3] P. D. Lorenzo, S. Barbarossa, P. Banelli, and S. Sardellitti, “Adaptive least mean squares estimation of graph signals,” IEEE Trans. Signal Info. Process. Netw., vol. 2, no. 4, pp. 555–568, Dec. 2016.
[4] C. Liu, S. Ghosal, Z. Jiang, and S. Sarkar, “An unsupervised spatiotemporal graphical modeling approach to anomaly detection in distributed CPS,” in ACM/IEEE Int. Conf. Cyber-Physical Syst., Apr. 2016, pp. 1–10.
[5] Y. Shen, P. A. Traganitis, and G. B. Giannakis, “Nonlinear dimensionality reduction on graphs,” in Proc. IEEE Int. Workshop Comput. Advan. Multi-Sensor Adapt. Process., Curacao, Netherlands Antilles, Dec. 2017.
[6] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, Inc., USA, 2014.
[7] E. Pavez and A. Ortega, “Covariance matrix estimation with non uniform and data dependent missing observations,” arXiv preprint arXiv:1910.06679, 2019.

[8] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, “Connecting the dots: Identifying network structure via graph signal processing,” arXiv preprint arXiv:1810.13066, 2018.

[9] D. Angelosante and G. B. Giannakis, “Sparse graphical modeling of piecewise-stationary time series,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Prague, Czech Republic, 2011, pp. 1960–1963.

[10] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” Biostatistics, vol. 9, no. 3, pp. 432–441, 2008.

[11] S. L. Lauritzen, Graphical models, Oxford University Press, 1996.

[12] C. W. J. Granger, “Some recent development in a concept of causality,” J. Econometrics, vol. 13, no. 2, pp. 228, 2020.

[13] Y. Shen, B. Baingana, and G. B. Giannakis, “Tensor decompositions for learning Laplacian matrix in smooth graph signal representations,” IEEE Trans. Signal Process., vol. 64, no. 23, pp. 6160–6173, Dec. 2016.

[14] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, “Network topology inference from spectral templates,” IEEE Trans. Signal Info. Process. Netw., vol. 3, no. 3, pp. 467–483, Sep. 2017.

[15] R. Shafipour and G. Mateos, “Online topology inference from streaming stationary graph signals with partial connectivity information,” Algorithms, vol. 13, no. 9, pp. 228, 2020.

[16] R. B. Kline, Principles and Practice of Structural Equation Modeling, Guilford Publications, 2015.

[17] Y. Shen, B. Baingana, and G. B. Giannakis, “Tensor decompositions for identifying directed graph topologies and tracking dynamic networks,” IEEE Trans. Signal Process., vol. 65, no. 14, pp. 3675–3687, Jul. 2017.

[18] C. W. J. Granger, “Some recent development in a concept of causality,” J. Econometrics, vol. 39, no. 1-2, pp. 199–211, Sep. 1988.

[19] A. Zellner, “Causality and econometrics,” in Carnegie-Rochester Conference series on Public Policy, Elsevier, 1979, vol. 10, pp. 9–54.

[20] S. M. Kay, Fundamentals of Statistical Signal Processing, Vol. 1: Estimation Theory, Prentice-Hall, 1993.

[21] R. Goebel, A. Roesbroeck, D.S. Kim, and E. Fornisano, “Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and Granger causality mapping,” Magn. Reson. Imag., vol. 21, no. 10, pp. 1251–1261, 2003.

[22] S. Basu, A. Shojaie, and G. Michailidis, “Network Granger causality with inherent grouping structure,” J. Mach. Learn. Res., vol. 16, no. 2, pp. 417–453, Mar. 2015.

[23] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox, “Neural Granger causality for nonlinear time series,” arXiv preprint arXiv:1802.05842, 2018.

[24] F. R. Bach and M. I. Jordan, “Learning graphical models for stationary time series,” IEEE Trans. Signal Process., vol. 52, no. 8, pp. 2189–2199, Aug. 2004.

[25] J. Songsiri and L. Vandenberghe, “Topology selection in graphical models of autoregressive processes,” J. Mach. Learn. Res., vol. 11, pp. 2671–2705, Oct. 2010.

[26] A. Bolstad, B. D. Van Veen, and R. Nowak, “Causal network inference via group sparse regularization,” IEEE Trans. Signal Process., vol. 65, no. 8, pp. 2077–2092, Apr. 2017.

[27] J. Songsiri, “Sparse autoregressive model estimation for learning Granger causality in time series,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Vancouver, BC, May 2013, pp. 3198–3202.

[28] J. Mei and J. M. F. Moura, “Signal processing on graphs: Causal modeling of unstructured data,” IEEE Trans. Signal Process., vol. 65, no. 8, pp. 2077–2092, Apr. 2017.

[29] M. Ayazoglu, M. Sznajer, and N. Ozay, “Blind identification of sparse dynamic networks and applications,” in IEEE Conf. Decision Control Eur. Control Conf., 2011, pp. 2944–2950.

[30] M. Kolak, L. Song, A. Ahmed, and E. P. Xing, “Estimating time-varying networks,” Ann. Appl. Statist, pp. 94–123, 2010.

[31] J. Lee, G. Li, and J. D. Wilson, “Varying-coefficient models for dynamic networks,” arXiv preprint arXiv:1702.03672, 2017.

[32] K. Yamada, Y. Tanaka, and A. Ortega, “Time-varying graph learning with constraints on graph temporal variation,” 2020.

[33] L. M. Lopez-Ramos, D. Romero, B. Zaman, and B. Beferull-Lozano, “Dynamic network identification from non-stationary vector autoregressive time series,” in Proc. IEEE Global Conf. Signal Inf. Process., 2018, pp. 775–777.

[34] D. Hallac, Y. Park, S. Boyd, and J. Leskovec, “Network inference via the time-varying graphical lasso,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2017, pp. 205–213.
Supplementary Material

APPENDIX A

PROOF OF LEMMA 1

To bound $\|\nabla \tilde{L}_1^{(n)}(a_n[t])\|_2$, taking the norm on both sides of (38) and applying the triangular inequality yields

$$
\left\| \nabla \tilde{L}_1^{(n)}(a_n[t]) \right\|_2 \\
\leq \left\| U_n[t]\tilde{g}[t]\tilde{g}^T[t]a_n[t] \right\|_2 \\
+ \left\| U_n[t]\tilde{y}_n[t]\tilde{g}[t] \right\|_2 \\
+ \left\| \gamma \Phi[t-1]a_n[t] \right\|_2 + \| \gamma \tilde{r}_n[t-1] \|_2
$$

Next, using assumptions A1 and A4, it can be easily shown that

$$
\lambda_{\max}(\tilde{g}[t]\tilde{g}^T[t]) \leq PNB_y + 2\sqrt{PNB_yB_g} + B_g^2. 
$$

Substituting this bound in the above expression and using assumption A3 yields

$$
\left\| \nabla \tilde{L}_1^{(n)}(a_n[t]) \right\|_2 \\
\leq \left\| U_n[t] \left( PNB_y + 2\sqrt{PNB_yB_g} + B_g^2 \right) a_n[t] \right\|_2 \\
+ U_n[t] \sqrt{PNB_y} + \gamma L \| a_n[t] \|_2 + \| \gamma \tilde{r}_n[t-1] \|_2. \quad (59)
$$

Next, an upper bound of $\tilde{r}_n[t-1]$ is derived. By the definition of $\tilde{r}_n[t]$ and assumption A1, we have

$$
\| \tilde{r}_n[t-1] \|_2 = \left\| \sum_{\tau=p}^{t-1} \tau^{t-1-\tau} \tilde{y}_n[\tau] \tilde{g}^T[\tau] \right\|_2 \\
\leq \frac{1}{\gamma} \left\| \sum_{\tau=p}^{t-1} \tau^{t-1-\tau} \sqrt{B_y} \sqrt{B_y} 1_{NP} \right\|_2 \\
= \frac{1}{\gamma} B_y \sqrt{PN} y \left( \frac{1}{\gamma} \right)^{t-1} \\
= \frac{1}{\gamma} B_y \sqrt{PN} \gamma (1 - \gamma^{t-p}) \leq \sqrt{PNB_y} \frac{1}{1 - \gamma}. \quad (60a)
$$

Using the above bound in (59), it follows that

$$
\left\| \nabla \tilde{L}_1^{(n)}(a_n[t]) \right\|_2 \\
\leq \left\| U_n[t] \left( PNB_y + 2\sqrt{PNB_yB_g} + B_g^2 \right) a_n[t] \right\|_2 \\
+ U_n[t] \sqrt{PNB_y} + \gamma L \| a_n[t] \|_2 + \gamma \sqrt{PNB_y} \frac{1}{1 - \gamma}. \quad (61a)
$$

The next step is to derive a bound on $\|a_n[t]\|_2$. To this end, from (29) and (38), it follows that

$$
\|a_n[t+1]\|_2 \\
\leq \|a_n[t]\|_2 \\
= \|a_n[t] - \alpha_t \tilde{v}_n[t]\|_2 \\
= \|a_n[t] - \alpha_t \left( U_n[t] \tilde{g}[t] \tilde{g}^T[t]a_n[t] - U_n[t] \tilde{y}_n[t] \tilde{g}[t] \\
+ \gamma \Phi[t-1]a_n[t] - \gamma \tilde{r}_n[t-1] \right) \|_2 \\
= \left\| \left( I - \alpha_t \gamma \Phi[t-1] - \alpha_t U_n[t] \tilde{g}[t] \tilde{g}^T[t] \right) a_n[t] + \alpha_t U_n[t] \tilde{y}_n[t] \tilde{g}[t] + \alpha_t \gamma \tilde{r}_n[t-1] \right\|_2. \quad (62)
$$

Applying triangular inequality and by assumption A2, we have

$$
\|a_n[t+1]\|_2 \\
\leq \lambda_{\max} \left( I - \alpha_t \gamma \Phi[t-1] - \alpha_t U_n[t] \tilde{g}[t] \tilde{g}^T[t] \right) \|a_n[t]\|_2 \\
+ \alpha_t \| U_n[t] \tilde{y}_n[t] \tilde{g}[t] \|_2 + \alpha_t \gamma \| \tilde{r}_n[t-1] \|_2 \quad (63a) \\
= 1 - \alpha_t \gamma \lambda_{\min} \left( \Phi[t-1] + \alpha_t U_n[t] \tilde{g}[t] \tilde{g}^T[t] \right) \|a_n[t]\|_2 \\
+ \alpha_t \| U_n[t] \tilde{y}_n[t] \tilde{g}[t] \|_2 + \alpha_t \gamma \| \tilde{r}_n[t-1] \|_2 \quad (63b) \\
\leq 1 - \alpha_t \gamma \lambda_{\min} \left( \Phi[t-1] \right) \|a_n[t]\|_2 + \alpha_t \| U_n[t] \tilde{y}_n[t] \tilde{g}[t] \|_2 \\
+ \alpha_t \gamma \| \tilde{r}_n[t-1] \|_2 \quad (63c) \\
\leq (1 - \alpha_t \gamma \beta_y) \|a_n[t]\|_2 + \alpha_t \| U_n[t] \tilde{y}_n[t] \tilde{g}[t] \|_2 \\
+ \alpha_t \gamma \| \tilde{r}_n[t-1] \|_2. \quad (63d)
$$

Substituting the bound on $\|\tilde{r}_n[t-1]\|_2$ from (60b) into the above expression, we have

$$
\|a_n[t+1]\|_2 \\
\leq (1 - \alpha_t \beta_y \gamma) \|a_n[t]\|_2 + \alpha_t \left( U_n[t] \sqrt{PNB_y} + \frac{\sqrt{PNB_y}}{1 - \gamma} \right). \quad (64)
$$

Setting $\alpha_t = \alpha$ and for $0 < \alpha \leq 1/L$, it can be proven by recursively substituting into (63d) (similar steps to those in the proof of [47, Theorem 5]), that

$$
\|a_n[t+1]\|_2 \leq \frac{1}{\beta_t \gamma} \left( \frac{\nu}{1 + \nu} \sqrt{PNB_y} + \frac{\sqrt{PNB_y}}{1 - \gamma} \right) \forall t. \quad (65)
$$

Substituting the above bound into (61b) completes the proof.

APPENDIX B

PROOF OF LEMMA 3

The error in the gradient for ISTIRSO is given by (41) and can be rewritten as:

$$
e^{(n)}[t] = U_n[t] \tilde{g}[t] \tilde{g}^T[t] - g[t] g^T[t] a_n[t] \\
+ U_n[t] \tilde{y}_n[t] (g[t] - \tilde{g}[t]) + \gamma (\tilde{r}_n[t-1] - r[t-1]) \\
+ \gamma (\Phi[t-1] - \Phi[t-1]) a_n[t]. \quad (66)
$$
Next, we take the norm on both sides of the above equation
\[
\left\| e^{(n)}[t] \right\|_2 \leq \left\| U_n[t](\hat{g}[t]g^T[t] - g[t]g^T[t])a_n[t] \right\|_2 \\
+ \left\| \gamma (\Phi[t - 1] - \Phi[t - 1])a_n[t] \right\|_2 \\
+ \left\| \gamma (r_n[t - 1] - \hat{r}_n[t - 1]) \right\|_2 + \left\| U_n[t]y_n[t](g[t] - \hat{g}[t]) \right\|_2 \\
\leq \gamma \lambda_{\max} (\Phi[t - 1] - \Phi[t - 1]) \| a_n[t] \|_2 + U_n[t] \lambda_{\max} (\hat{g}[t]g^T[t] - g[t]g^T[t]) \| a_n[t] \|_2 + \gamma B_r + U_n[t] \| y_n[t] \| g[t] - \hat{g}[t] \|_2,
\]  
(67)
where the first inequality holds because of the triangular inequality and the second inequality holds because of the Cauchy-Schwarz inequality.

Besides, combining A1 and (40a) it can be proven that
\[
\lambda_{\max} (\hat{g}[t]g^T[t] - g[t]g^T[t]) \leq 2 \sqrt{PNB_B y_B g + B_g^2}.  
\]  
(68)
By substituting (40) and (68) into (67), we obtain
\[
\left\| e^{(n)}[t] \right\|_2 \\
\leq \gamma B \| a_n[t] \|_2 + U_n[t] \left(2 \sqrt{PNB_B y_B g + B_g^2}\right) \| a_n[t] \|_2 \\
+ \gamma B_r + U_n[t] B_g \sqrt{B_y} 
\]  
(69a)
\[
\leq \gamma B \| a_n[t] \|_2 + \left(\frac{\nu}{1 + \nu}\right) \left(2 \sqrt{PNB_B y_B g + B_g^2}\right) \| a_n[t] \|_2 \\
+ \gamma B_r + \left(\frac{\nu}{1 + \nu}\right) B_g \sqrt{B_y} 
\]  
(69b)
\[
= \left(\gamma B \| a_n[t] \|_2 + \left(\frac{\nu}{1 + \nu}\right) \left(2 \sqrt{PNB_B y_B g + B_g^2}\right)\right) \| a_n[t] \|_2 \\
+ \gamma B_r + \left(\frac{\nu}{1 + \nu}\right) B_g \sqrt{B_y}, 
\]  
(69c)
where the final result comes from substituting an upper bound on \( U_n[t] \) and rearranging terms. We can use here the same bound on \( \| a_n[t] \|_2 \) that was derived in the proof of Lemma 1 [cf. (65)]:
\[
\| a_n[t + 1] \|_2 \leq \frac{\sqrt{PNB_y}}{3 \beta_\ell} \left(\frac{\nu}{1 + \nu} + \frac{1}{1 - \gamma}\right) \quad \forall t;  
\]  
(70)
substituting the above bound into (69c) completes the proof.