Generalized Autoregressive Linear Models for Discrete High-Dimensional Data

Parthe Pandit, *Student Member, IEEE*, Mojtaba Sahraee-Ardakan, *Student Member, IEEE*, Arash A. Amini, Sundeep Rangan, *Fellow, IEEE*, and Alyson K. Fletcher, *Member, IEEE*

***Abstract*—Fitting multivariate autoregressive (AR) models is fundamental for time-series data analysis in a wide range of applications. This article considers the problem of learning a *p*-lag multivariate AR model where each time step involves a linear combination of the past *p* states followed by a probabilistic, possibly nonlinear, mapping to the next state. The problem is to learn the linear connectivity tensor from observations of the states. We focus on the sparse setting, which arises in applications with a limited number of direct connections between variables. For such problems, ℓ_1-regularized maximum likelihood estimation (or M-estimation more generally) is often straightforward to apply and works well in practice. However, the analysis of such methods is difficult due to the feedback in the state dynamic and the presence of nonlinearities, especially when the underlying process is non-Gaussian. Our main result provides a bound on the mean-squared error of the estimated connectivity tensor as a function of the sparsity and the number of samples, for a class of discrete multivariate AR models, in the high-dimensional regime. Importantly, the bound indicates that, with sufficient sparsity, consistent estimation is possible in cases where the number of samples is significantly less than the total number of parameters.**

***Index Terms*—Autoregressive processes, compressed sensing, high-dimensional time series analysis, long-termmemory, nonlinear dynamical systems, maximum likelihood estimation.**

I. Introduction

We CONSIDER the problem of learning a *p*-lag autoregressive (AR) generalized linear model (GLM) for a multivariate time series involving *N*-variables: $\boldsymbol{x}^t = (x_i^t) \in \mathbb{R}^N$, where $x_i^t \in \mathcal{X}_i \subseteq \mathbb{R}$ for all $i \in [N]$, $t \in \mathbb{Z}$. A particular case of the model we consider is of the form,

$$x_i^t \mid z_i^t \sim \mathbb{Q}_i(\cdot \mid z_i^t), \quad z_i^t = f_i\Big(\big\langle \Theta_i^*, \mathbf{X}^{t-1} \big\rangle\Big), \tag{1}$$

where the inner product corresponds to $\mathbb{R}^{N\times p}$, for $t = 1, 2, \ldots$ and $i = 1, 2, \ldots, N$ where $\mathbf{X}^{t-1} = [\boldsymbol{x}^{t-1}\ \boldsymbol{x}^{t-2}\ \ldots\ \boldsymbol{x}^{t-p}] \in \mathbb{R}^{N\times p}$ is the *p*-lag history of the process up to time $t-1$, and $\mathbb{Q}_i(\cdot \mid z_i^t)$ is a probabilistic link function. The problem is to estimate the unknown parameters $\Theta_i^* \in \mathbb{R}^{N\times p}$ for $i = 1, 2, \ldots, N$, given observations of n time samples $\boldsymbol{x}^t$, $t = 1, \ldots, n$. The conditional distributions $\mathbb{Q}_i(\cdot \mid z_i^t)$ and link functions f_i are assumed to be known.

Modeling problems of this form appear in a wide-range of applications with time-series data. For example, in neural modeling, $\boldsymbol{x}^t$ can represent a vector of spike counts or some other measure of activity from N neurons or brain regions. In this case, estimation of the tensor Θ^* in (1) can provide insight into the neural connectivity. Other applications include genomics, econometrics [3], data science, sociology, business management, financial markets [4], [5] and natural language processing.

A key challenge in estimating the multivariate AR(*p*) models is the large number of unknown parameters to estimate, particularly as the dimension of the process, N, and number of time lags, p, grows. However, in many cases, one can assume some sparsity constraint in the connectivity tensor Θ^*. For example, in neural modeling, there are physically limited numbers of direct connections between brain regions. Under a sparsity assumption, it is common to estimate Θ^* via an ℓ_1-regularized M-estimator of the form,

$$\begin{aligned}\widehat{\Theta} := \underset{\Theta\in\mathbb{R}^{N\times N\times p}}{\operatorname{argmin}}\ & \frac{1}{n}\sum_{i=1}^{N}\sum_{t=1}^{n} \mathcal{L}_{it}\Big(x_i^t; \big\langle \Theta_i, \mathbf{X}^{t-1} \big\rangle\Big) \\ & + \lambda_n \|\Theta\|_{1,1,1},\end{aligned} \tag{2}$$

where $\mathcal{L}_{it} : \mathcal{X}_i \times \mathbb{R} \to \mathbb{R}$ are loss functions and $\lambda_n\|\Theta\|_{1,1,1}$ is an ℓ_1 regularizer (precise definitions will be given in the Section II below). The broad goal of this article is to analyze the sample complexity of such ℓ_1-regularized M-estimators. That is, given a sparsity constraint on Θ^*, and the number of measurements, n, how well can we estimate Θ^*?

A. Key Contributions

We consider the case where $\{\mathcal{X}_i\}_{i=1}^N$ are bounded countable subsets of $\mathbb{R}$. We analyze the ℓ_1-regularized M-estimator (2) when the loss functions $v \mapsto \mathcal{L}_{it}(u; v)$ are strongly convex,

Manuscript received May 15, 2020; revised November 23, 2020; accepted November 24, 2020. Date of publication December 2, 2020; date of current version January 7, 2021. The work of Parthe Pandit, Mojtaba Sahraee-Ardakan, and Alyson K. Fletcher was supported in part by the National Science Foundation under Grant 1254204 and Grant 1738286, and in part by the Office of Naval Research under Grant N00014-15-1-2677. The work of Sundeep Rangan was supported in part by the National Science Foundation under Grant 1116589, Grant 1302336, and Grant 1547332; in part by the Semiconductor Research Corporation; in part by National Institute of Standards and Technology; and in part by the industrial affiliates of NYU WIRELESS. Portions of this article were presented at AISTATS 2019 [1]. Code available at [2]. *(Corresponding author: Parthe Pandit.)*

Parthe Pandit and Mojtaba Sahraee-Ardakan are with the Department of Electrical and Computer Engineering, University of California Los Angeles, Los Angeles, CA 90095 USA (e-mail: parthepandit@ucla.edu; msahraee@ucla.edu).

Arash A. Amini and Alyson K. Fletcher are with the Department of Statistics, University of California Los Angeles, Los Angeles, CA 90095 USA (e-mail: aaamini@ucla.edu; akfletcher@ucla.edu).

Sundeep Rangan is with the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY 11201 USA (e-mail: srangan@nyu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2020.3041714, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2020.3041714

for all $u \in \mathcal{X}_i$. We assume that the connectivity tensor can be approximated by a sparse tensor with at most $s_{\max}$ non-zero values in each slice Θ_i^*. Under these assumptions, our main result in Theorem 1 establishes the consistency of the regularized M-estimator (2) in the high-dimensional regime of $n = \mathsf{poly}(s_{\max} \log(N^2 p))$ under some regularity conditions.

In proving our main result, we establish the so-called restricted strong convexity (RSC) [6] for a large class of loss functions, for a dependent non-Gaussian discrete-valued multivariate process. Our proof of the RSC property requires showing a restricted eigenvalue condition, which is nontrivial due to the non-Gaussian and highly-correlated nature of the design matrix. What makes the problem more challenging is the existence of feedback from more than just the immediate past (the case $p > 1$).

We establish the RSC for general $p \geq 1$ using the novel approach of viewing the p-block version of the process as a Markov chain. The problem becomes significantly more challenging when going from $p = 1$ to even $p = 2$. The difficulty with this *higher-order* Markov chain is that its *Dobrushin contraction coefficient* is trivially 1. We develop techniques to get around this issue which could be of independent interest (see Section VII). Our techniques hold for all $p \geq 1$.

Much of the previous work towards proving the RSC condition has either focused on the independent sub-Gaussian case [7], [8] or the dependent Gaussian case [9], [10] for which powerful Gaussian concentration results such as the Hanson–Wright inequality [11] are still available. Our approach is to use concentration results for Lipschitz functions of Markov chains over countable spaces, and strengthen them to uniform results using metric entropy arguments. In doing so, we circumvent the use of empirical processes which require additional assumptions for estimation [12]. Moreover, our approach allows us to identify key properties of the model that allow for sample-efficient estimation.

Although discrete time series are often modeled using the specific link functions such as `logit` or `softmax`, our result allows more flexibility to choose the link functions. For example in the Bernoulli AR(p) and Truncated-Poisson AR(p) cases discussed in Section III-B, any Lipschitz continuous, log-convex link function can be used. The analysis also brings out crucial properties of the link function, and the role it plays in determining the estimation error and sample complexity.

Our model also allows for each individual time series x_i^t to lie in distinct spaces $\mathcal{X}_i$ which is desirable in practical applications with heterogeneous types of data.

B. Previous Work

There is a vast literature on recovering sparse vectors in under-sampled settings [13], [14], [15], [16]. The generic results show that if a vector θ is s-sparse in a p-dimensions, it can be estimated in $n = \Omega(s \log(p))$ measurements. However, these results typically do not have feedback as in the AR process considered here.

The estimation of sparse Gaussian VAR(p) processes with linear feedback has been considered only more recently [9], [17], [18], [19], [20]. For these models, a restricted eigenvalue condition can be established fairly easily, by reducing the problem, even in the time-correlated setting, to the concentration of quadratic functionals of Gaussian vectors for which powerful inequalities exist [11]. These techniques do not extend to non-Gaussian setups.

In the non-Gaussian setting, Hall *et al.* [21] and Zhou and Raskutti [22] recently considered a multivariate time series evolving as a GLM driven by the history of the process similar to our model. The Bernoulli AR(1) and Poisson AR(1) with $p = 1$ lags were considered as special cases of this model. They provide statistical guarantees on the error rate for the ℓ_1 regularized estimator. More importantly, their results are restricted to the case $p = 1$ which does not allow the explicit encoding of long-term dependencies. More recently, Mark *et al.* [23], [24] considered a model closer to ours for multivariate AR(p) processes with lags $p = 1$ or $p = 2$.

A key contribution of ours is to bring out the explicit dependence on p in the AR(p) models, allowing for a general $p \geq 1$. In the special cases we consider: the Bernoulli AR(p) and the Truncated-Poisson AR(p), we show how the scaling of the sample complexity and the error rate with p can be controlled by the properties of the link function f_i and a certain norm of the parameter tensor.

Our results improve upon those in [21], [23] when applied to the Bernoulli AR(p) and Truncated-Poisson AR(p). Due to the key observation that an AR(p) over a countable space can be viewed as a higher order Markov chain, our analysis relaxes several assumptions made by [21], [23]. In doing so, we achieve better sample complexities with explicit dependence on p. Our analysis borrows from martingale-based concentration inequalities for Lipschitz functions of Markov chains [25].

The univariate Bernoulli AR(p) process for $p \geq 1$ was considered by Kazemipour *et al.* [26], [27] where they analyzed a multilag Bernoulli process for a single neuron. Their analysis does not extend to the $N > 1$ case. Even for $N = 1$, their analysis is restricted to the biased process with $\mathbb{P}(x_1^t = 1 | \mathbf{X}^{t-1}) < \frac{1}{2}$ for all t. Mixing times of the Bernoulli AR(1) have been considered in [28]. However, their discussion is again limited to $p = 1$.

The rest of the paper is organized as follows. In Section II, we introduce the generalized discrete VAR(p) model and the associated class of regularized M-estimators. Section III presents our main result, Theorem 1, on the consistency of the regularized M-estimator and discusses its assumptions and implications. Applications of Theorem 1 to the special cases of Binomial and Truncated-Poisson processes are detailed in Section III-B. In Section IV, we provide simulation results corroborating our theoretical predictions. Section V provides an overview of the proof of Theorem 1. In Section VII, we present new techniques for deriving concentration inequalities for dependent multivariate processes. We conclude with a discussion and point to some open problems and directions for solving them in Section VIII.

Notation: For two sequence $\{a_n\}$ and $\{b_n\}$, we write either of $a_n \gtrsim b_n$ or $b_n \lesssim a_n$ or $b_n = O(a_n)$ or $a_n = \Omega(a_n)$ to mean that there is a constant $C > 0$ such that $a_n \geq C b_n$ for all n. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$. We write $a_n \gg b_n$

or $b_n \ll a_n$ or $b_n = o(a_n)$ if $b_n/a_n \to 0$ as $n \to \infty$. We use $[N]$ to denote the set $\{1, 2, \dots, N\}$. For a subset $\mathcal{X}$ of a vector space, we write $\mathcal{X}^{\times p}$ for the set of matrices with p columns from $\mathcal{X}$. Formally $\mathcal{X}^{\times p} := \{(x_1, x_2, \dots, x_p) \mid x_i \in \mathcal{X},\ i \in [p]\}$. For example, $(\mathbb{R}^N)^{\times p}$ is the same as the set of real-valued $N \times p$ matrices. In addition, Table I in the supplementary material provides a list of all notations used in the paper.

II. Models and Methods

To state our results in their full generality, we consider a slightly more general model than (1). We assume that the multivariate time series $\boldsymbol{x}^t = (x_i^t) \in \mathcal{X} \subset \mathbb{R}^N$ evolves as,

$$x_i^t \mid z_i^t \sim \mathbb{Q}_i(\,\cdot \mid z_i^t) \tag{3a}$$

$$z_i^t = f_i\Big(\big\langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \big\rangle_{\mathbb{R}^{N \times L}}\Big) \tag{3b}$$

$$x_i^t \perp\!\!\!\perp x_j^t \mid \boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \dots \tag{3c}$$

for $t = 1, 2, \dots$ and $i = 1, 2, \dots, N$. The key difference here is that we have added a matrix $\mathbf{D} = [\boldsymbol{d}_1\ \boldsymbol{d}_2\ \dots\ \boldsymbol{d}_L] \in \mathbb{R}^{p \times L}$, a known dictionary of filters $\{\boldsymbol{d}_\ell\}_{\ell=1}^L$. When $\mathbf{D} = I_{p \times p}$, we obtain the special case (1). The role of this dictionary will be explained below. To model the discrete-valued nature of the states, we assume that $\boldsymbol{x}^t \in \mathcal{X} := \prod_{i=1}^N \mathcal{X}_i$ where each $\mathcal{X}_i$ is a bounded countable subset of $\mathbb{R}$. The matrix $\mathbf{X}^{t-1} = [\boldsymbol{x}^{t-1}\ \boldsymbol{x}^{t-2}\ \dots\ \boldsymbol{x}^{t-p}] \in \mathbb{R}^{N \times p}$ is the p-lag history of the process up to time $t-1$, and $\mathbb{Q}_i(\,\cdot \mid z)$ is a distribution on $\mathcal{X}_i$ parameterized by z. For example an exponential family distribution with mean parameter z. The matrices $\Theta_i^* \in \mathbb{R}^{N \times L}$, $i \in [N]$ are the (unknown) model parameters and $\langle \cdot, \cdot \rangle_{\mathbb{R}^{N \times L}}$ is the inner product. A process of this form will be denoted GVAR(p).

The distribution $\mathbb{Q}_i(\,\cdot \mid z_i^t)$ represents the conditional distribution of x_i^t given the past $\boldsymbol{x}^{t-1}, \boldsymbol{x}^{t-2}, \dots$. Functions $f_i : \mathbb{R} \to \mathbb{R}$ are similar to the inverse-link functions in GLMs, and can be nonlinear in general. It is worth noting that $\mathcal{X}_i$ and $\mathbb{Q}_i$ can vary for every variable $i \in [N]$ making the model extremely flexible to include heterogeneous types of discrete data.

The inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^{N \times L}}$ in (3) is the Hilbert-Schmidt inner product on $\mathbb{R}^{N \times L}$, and can be expanded as:

$$\big\langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \big\rangle_{\mathbb{R}^{N \times L}} = \sum_{j=1}^{N} \sum_{\ell=1}^{L} \Theta_{ij\ell}^* \big\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \big\rangle_{\mathbb{R}^p} \tag{4}$$

where $\mathbf{x}_j^{t-*} := [x_j^{t-1}\ x_j^{t-2}\ \cdots\ x_j^{t-p}]$ is the p-lag history of variable j up to time $t-1$, i.e., the j^{th} row of $\mathbf{X}^{t-1}$. Note that $(\mathbf{X}^{t-1}\mathbf{D})_{j\ell} = \big\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \big\rangle_{\mathbb{R}^p}$. The parameter $(\Theta_i^*)_{j\ell} = \Theta_{ij\ell}^* \in \mathbb{R}$ captures the dependence of variable x_i^t on the past activity of variable j, via $\mathbf{x}_j^{t-*}$. The vectors $\boldsymbol{d}_\ell \in \mathbb{R}^p$ act as filters that modulate the mean of variable x_i^t based on the past activity of all the variables, that is, x_j^k for $j \in [N]$, and $t - p \le k < t$.

A. Dictionary and Network Interpretations

The filters $\{\boldsymbol{d}_\ell\}$ serve two main purposes: (i) interpretability and (ii) dimension reduction. For example, in neuroscience applications where the types of spiking behaviors are limited, the presence of a dictionary causes the model to favor specific forms of interactions between the spiking activities of two neurons. We refer to [29] which explores these filters for various interactive behaviors among neurons such as bursting, tonic spiking, phasic spiking, etc. The dictionary increases the interpretability of the parameter Θ_i^*—one interprets $(\Theta_i^*)_{j\ell}$ as measuring the effect of the activity of neuron i on neuron j, as explained by interaction type ℓ. Thus, the sparsity of Θ_i^* is more meaningful in the presence of a dictionary. An earlier version of this article [1] considered modeling the interaction with the past as $\langle \Theta_i^*, \mathbf{X}^{t-1} \rangle$ where Θ_i^* lies in $\mathbb{R}^{N \times p}$, corresponding to taking $\mathbf{D} = I_{p \times p}$, the identity matrix, in (4c). The formulation with a general dictionary $\mathbf{D}$ has the added advantage of potentially reducing the number of free parameters from Np to NL. When $L \ll p$, this leads to a massive dimension reduction. The bilinear term $\langle \Theta_i^*, \mathbf{X}^{t-1}\mathbf{D} \rangle_{\mathbb{R}^{N \times L}} = \langle \Theta_i^* \mathbf{D}^\top, \mathbf{X}^{t-1} \rangle_{\mathbb{R}^{N \times p}}$ can also be thought of as a low-rank approximation to the parameter, forcing one factor to be fixed by $\mathbf{D}$. By adding pre-existing knowledge of temporal interactions between variables, the dictionary allows for a rich model with fewer parameters, leading to more (sample) efficient estimators for Θ^*.

The parameter Θ^* can be interpreted as representing a network among variables x_i^t, $i \in [N]$. A slice $\Theta_{**\ell}$ can be thought of as an adjacency matrix for the *influence network* explained by coupling behavior ℓ. If neurons i and j are not connected, then $\Theta_{ij\ell} = 0$ for all $\ell \in [L]$. For example, in the neural spike train application, one can reveal a latent network among the neurons (i.e., who influences whose firing) just from the observations of patterns of neural activity, a task which is of significant interest in neuroscience [30], [31], [32]. Similarly, in the context of social networks, one might be interested in who is influencing whom [33].

B. Examples

The GVAR(p) process of the form (3) can be applied in a wide range of applications. For example, letting $\mathbb{Q}_i(\,\cdot \mid z) = \text{Ber}(z)$ and $f_i(u) = (1 + e^{-u})^{-1}$ recovers the Bernoulli autoregressive process in [1]. Similarly, $\mathbb{Q}_i(\,\cdot \mid z) = \text{Binomial}(K_i, z)$ and $f_i(u) = (1 + e^{-u})^{-1}$ models a Binomial process with K_i trials (for coordinate i) and success probability z. Such a model can be suitable for modeling count data. Another common model for point processes in neuroscience [31] is the Truncated-Poisson autoregressive process given by $\mathbb{Q}_i(\,\cdot \mid z) = \mathbb{P}(\min(M_i, Z) \in \cdot\,)$ where $Z \sim \text{Poi}(z)$, and $f_i(u) = \exp(u)$ or $f_i(u) = \log(1 + e^u)$ for some integer M_i [21], [23]. Although we focus on single-parameter discrete distributions in this article, the ideas can be easily extended to distributions with multiple parameters. For example, one can construct a categorical or multinomial process, by allowing z_i^t to be vector-valued and taking f_i to be the `softmax` function.

C. Regularized M-Estimation

We are primarily interested in parameter estimation in the high-dimensional regime where $n \ll N$. To make the estimation feasible, we assume that the activity of each variable i depends on the past activity of only a few number of variables,

$s_i \ll N$. We refer to s_i as the *in-degree* of variable i. Our main result provides sufficient conditions under which parameter Θ^* can be estimated in the high-dimensional setting where $n = \mathsf{poly}(\{s_i\}_{i=1}^N, \log(NLp))$.

Given a collection of loss functions $\mathcal{L}_{it}:\mathcal{X}_i \times \mathbb{R} \to \mathbb{R}$, for $i \in [N]$ and $t \in \mathbb{Z}$, we consider the following ℓ_1-regularized M-estimator

$$\widehat{\Theta} := \underset{\Theta\in\mathbb{R}^{N\times N\times L}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}_i(\Theta_i) + \lambda_n\|\Theta\|_{1,1,1}.$$
$$\mathcal{L}_i(\Theta_i) := \tfrac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{it}\Big(x_i^t\,;\,\big\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\big\rangle\Big) \qquad (5)$$

where we use the notation

$$\|\boldsymbol{M}\|_{p,q,r} := \left(\sum_{i=1}^{a}\left\{\sum_{j=1}^{b}\left(\sum_{k=1}^{c}|M_{ijk}|^r\right)^{\frac{q}{r}}\right\}^{\frac{p}{q}}\right)^{\frac{1}{p}} \qquad (6)$$

to denote a norm of a $a \times b \times c$ tensor $\boldsymbol{M}$ (when $p, q, r > 1$). We also use a similar norm notation for matrices $\|\boldsymbol{M}\|_{p,q} := \sum_{i=1}^{a}(\sum_{j=1}^{b}|M_{ij}|^q)^{\frac{p}{q}}$. For $p = q = r = 2$, we denote the norm subscript by F.

Since both the loss function and the ℓ_1 penalty are decomposable, we can solve each of the N problems in (5) indexed by i separately,

$$\widehat{\Theta}_i := \underset{\Theta_i\in\mathbb{R}^{N\times L}}{\operatorname{argmin}}\ \mathcal{L}_i(\Theta_i) + \lambda_n\|\Theta_i\|_{1,1} \quad \forall i \in [N]. \qquad (7)$$

The possible dependence of $\mathcal{L}_{it}$ on t in the M-estimator (5) allows for the incorporation of time-discounting factors such as γ^t for some $\gamma < 1$. We consider a large class of loss functions later stated explicitly in Assumptions (A2) and (A3). This class always includes the negative-log likelihood function for exponential family distributions $\mathbb{Q}_i(\cdot\,|f_i(v))$ with log-concave link f_i, and pseudo-likelihood functions in some cases. When $\mathcal{L}_{it}$ are chosen to be convex, the whole problem (5) is unconstrained, convex, with a coercive objective function, whereby the solution $\widehat{\Theta}$ is unique. Furthermore, the estimator (5) can be solved efficiently using any non-smooth convex optimization solver, such as the subgradient methods or proximal gradient descent methods [34]. An implementation for the general problem in (5) is available at [2] which implements both the subgradient method as well as the proximal gradient method.

Each iteration of both of these methods involve computation of the gradient of the loss function followed by finding the sub-gradient or proximal mapping for the regularization. Computing the gradient of the loss is the most expensive step. The gradient of the loss is

$$\nabla\mathcal{L}(\Theta_i) = \frac{1}{n}\sum_{t=1}^{n}\mathcal{L}'_{it}\Big(x_i^t;\big\langle\Theta_i, \mathbf{X}^{t-1}\mathbf{D}\big\rangle\Big)\mathbf{X}^{t-1}\mathbf{D}, \qquad (8)$$

where in $\mathcal{L}'_{it}(\cdot\,;\,\cdot)$ the derivative is with respect to the second argument. To compute the gradient, $\mathbf{X}^{t-1}\mathbf{D}$ can be precomputed once by multiplying $\mathbb{X} := \{\boldsymbol{x}^t\}_{t=-p+1}^{n}$ and $\mathbf{D}$. Hence, the complexity of obtaining the gradient $\nabla\mathcal{L}(\Theta_i)$ at each iteration is dominated by that of computing $\big\langle\Theta_i, \mathbf{X}^{t-1}\mathbf{D}\big\rangle$ for all i, that is, $O(nNL)$. To solve the optimization problem, one can then use the subgradient method with a provable convergence rate of $1/\sqrt{k}$ after k steps. This relatively slow rate is due to the nonsmoothness of the objective function. Alternatively, we can use the proximal gradient method that converges at a rate of $1/k$. Then, the overall computational complexity of obtaining an ε-optimal solution is $O(nNL/\varepsilon)$. The parallel implementation in (7) allows for massive speed-ups in computation when using GPUs. The main result of this article concerns the statistical complexity of the estimator and is agnostic to the choice of the optimization solver.

Our main result establishes the statistical properties of estimator (5) such as consistency, sample complexity and error rate. Our analysis also highlights desirable properties of the loss functions $\mathcal{L}_{it}$ and the nonlinearities f_i for achieving consistency. The result also shows the effect of the dictionary $\mathbf{D}$ in increasing the sample-efficiency of the estimator.

III. Main Results

Our main result concerns the estimation error of the parameters $\{\widehat{\Theta}_i\}_{i=1}^N$, obtained by solving (7). We implicitly assume Θ_i^* to be approximately s_i-sparse. This assumption is encoded via the ℓ_1-approximation errors

$$\omega_i := \min_{\boldsymbol{\beta}\in\mathbb{R}^{N\times L}}\big\{\|\boldsymbol{\beta} - \Theta_i^*\|_1 \mid \|\boldsymbol{\beta}\|_{0,0} \le s_i\big\}. \qquad (9)$$

We also impose the following assumptions:

(A1) The process is wide-sense stationary and stable, i.e., the power spectral density matrix exists:

$$\mathcal{X}(\omega) := \sum_{\ell=-\infty}^{\infty}\operatorname{Cov}\Big(\boldsymbol{x}^t, \boldsymbol{x}^{t-\ell}\Big)e^{-j\omega\ell} \in \mathbb{C}^{N\times N},$$
$$\min_{\omega\in[-\pi,\pi)} \lambda_{\min}(\mathcal{X}(\omega)) \ge C_{\mathcal{X}}^2 > 0.$$

(A2) The loss function $v \mapsto \mathcal{L}_{it}(u, v)$ is twice differentiable and strongly convex for all u, with curvature $\kappa_i > 0$, i.e., $\partial_v^2\mathcal{L}_{it}(u; v) \ge \kappa_i$ for all $u \in \mathcal{X}_i$, $v \in \mathbb{R}$, $i \in [N]$, $t \in \mathbb{N}_+$.

(A3) $|\partial_v\mathcal{L}_{it}(u, v)| \le C_{\mathcal{L}}$, and for all $v \in \mathbb{R}$, $i \in [N]$, $t \in \mathbb{N}_+$ we have

$$U \sim \mathbb{Q}_i(\cdot\,|f_i(v)) \implies \mathbb{E}[\partial_v\mathcal{L}_{it}(U; v)] = 0.$$

Assumption (A3) guarantees that Θ^* is the minimizer of the population loss, and is necessary for the consistency of the M-estimator. The second half of the assumption is generally satisfied if the loss is taken to be the log-likelihood function. The next example verifies this for single-parameter exponential families.

Example 1: Assume that $Q_i(\cdot \mid z)$ is an exponential family with density $x \mapsto \exp(xz - \phi(z))$, for all i. Here, z is the so-called natural parameter of the family and ϕ is the log-partion function. Let $U \sim Q(\cdot \mid f_i(v))$ and take $\mathcal{L}_{it}(x, v)$ to be the log-likelihood of this model, that is,

$$\mathcal{L}_{it}(x; v) = -xf_i(v) + \phi(f_i(v)).$$

This class includes Bernoulli, Poisson, and Gaussian (with known variance) AR processes among others. We have

$$\partial_v\mathcal{L}_{it}(U; v) = -Uf_i'(v) + \phi'(f_i(v))f_i'(v).$$

By a standard property of the exponential family $\mathbb{E}[U] = \phi'(f_i(v))$, hence $\mathbb{E}[\partial_v \mathcal{L}_{it}(U; v)] = 0$ verifying the second half of (A3). If, in addition, the family has bounded support and both ϕ and f_i are Lipschitz, then the entire (A3) holds. Distributions such as Poisson and Gaussian violate the boundedness assumption. However, the truncated version of these distributions belong to the exponential family and satisfy the boundedness condition.

Example 2: Under the same exponential family distribution as in Example 1, the second half of (A3) also holds for the squared error loss

$$\mathcal{L}_{it}(x; v) = \big[x - \phi'(f_i(v))\big]^2.$$

To verify this, it is enough to observe that

$$\partial \mathcal{L}_{it}(U; v) = 2\big[U - \phi'(f_i(v))\big] \cdot \phi''(f_i(v)) f_i'(v),$$

and use $\mathbb{E}[U] = \phi'(f_i(v))$.

These two examples show that (A3) is satisfied for commonly used loss functions. As for (A2), we recall that in an exponential family with the natural parameterization, the log-partition function $\phi(\cdot)$ is convex. Assumption (A2), however, requires the map $v \mapsto \mathcal{L}_{it}(u, v)$ to be strongly convex. Extra care should be taken in choosing the loss and $f_i(\cdot)$ to ensure that this assumption is satisfied. The stability Assumption (A1) is further discussed in the remarks following the main result.

Let us now define a few constants necessary to state our main result. Let

$$\begin{aligned} C_{\mathbf{D}} &:= \max_{\ell} \|\boldsymbol{d}_\ell\|_1, \\ G = G_f(\Theta^*) &:= 64 C_{\mathbf{D}}^4 B^4 \Big(1 + p^2 \psi\big(\tau_1(\Theta^*)\big)\Big), \end{aligned} \tag{10}$$

where $\psi(x) = (1 - x^{-1})^{-2}$ and

$$\begin{aligned} \tau_1(\Theta^*) &:= \sup_{z, y \in \mathcal{X}^{\times p}} \|\mathbb{P}_z - \mathbb{P}_y\|_{\mathrm{TV}} < 1, \\ \mathbb{P}_z &:= \mathbb{P}\big(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = z\big), \quad z \in \mathcal{X}^{\times p}. \end{aligned} \tag{11}$$

Here, $\mathcal{X}^{\times p} \subset \mathbb{R}^{N \times p}$ denotes the set of matrices consisting of p columns, each from $\mathcal{X}$. Note that $\mathbb{P}_z$ is t-invariant. Fix $\mathcal{U} \subset [N]$ and let us write

$$\begin{aligned} &s_{\max} := \max_{i \in \mathcal{U}} s_i, \quad s_+ := \sum_{i \in \mathcal{U}} s_i, \quad \overline{\kappa} := \max_{i \in \mathcal{U}} \kappa_i \\ &\underline{\kappa} := \frac{C_{\mathfrak{X}}^2}{8} \min_{i \in \mathcal{U}} \kappa_i, \quad \text{and} \quad \widetilde{\omega}_+ := \sum_{i \in \mathcal{U}} \overline{\kappa} \frac{\omega_i^2}{s_i} + 4\omega_i, \end{aligned} \tag{12}$$

where κ_i and $C_{\mathfrak{X}}$ are specified in (A2) and (A1). We are now ready to state the main result:

Theorem 1: Suppose that $\{\boldsymbol{x}^t\}_{t=-p+1}^n$ are samples from process (3), with each $\mathcal{X}_i$ being a countable subsets of $[-B, B]$ for some $B > 0$, and satisfying (A1). Fix a subset $\mathcal{U} \subseteq [N]$ and let $\{\widehat{\Theta}_i\}_{i \in \mathcal{U}}$ be the solutions of (7) with loss functions $\mathcal{L}_{it}$ satisfying (A2)-(A3). Fix $c_1 > 2$ and let $c = c_1/2 - 1$. If

$$\begin{aligned} &\lambda_n = 2BC_{\mathcal{L}} C_{\mathbf{D}} \sqrt{c_1 \log(|\mathcal{U}| NL)/n}, \quad \text{and} \\ &n \gtrsim \frac{G}{C_{\mathfrak{X}}^6} s_{\max}^3 \log(NL), \end{aligned}$$

then, with probability at least $1 - (NL)^{-Cs_{\max}} - (|\mathcal{U}| NL)^{-c}$,

$$\sum_{i \in \mathcal{U}} \|\widehat{\Theta}_i - \Theta_i^*\|_F^2 \le \frac{9}{\underline{\kappa}^2} s_+ \lambda_n^2 + \frac{\widetilde{\omega}_+}{\underline{\kappa}} \lambda_n. \tag{13}$$

where $C = \mathcal{O}\Big(C_{\mathfrak{X}}^{-2}\Big)$ only depends on $C_{\mathfrak{X}}$.

The error bound in (13) can be written, up to constants, as:

$$\sum_{i \in \mathcal{U}} \|\widehat{\Theta}_i - \Theta_i^*\|_F^2 \lesssim \frac{s_+ \log(NL)}{n} + \widetilde{\omega}_+ \sqrt{\frac{\log(NL)}{n}}. \tag{14}$$

The two terms in the bound correspond to the estimation and approximation errors, respectively. The estimation error scales at the so-called *fast rate* $\log(NL)/n$, while the approximation error scales at the slower rate $\sqrt{\log(NL)/n}$. For the exact sparsity model, where $\omega_i = 0$ for all i, the approximation error vanishes and the estimator achieves the fast rate. For simplicity, assume that $C_{\mathcal{L}}, C_{\mathbf{D}} \lesssim 1 \lesssim C_{\mathfrak{X}}$. Then, the overall (excess) sample complexity for consistent estimation is

$$n \gg \max\Big\{G s_{\max}^3,\ s_+,\ (\widetilde{\omega}_+)^2\Big\} \log(NL). \tag{15}$$

By consistency, we mean that the estimator converges to the true parameter when n grows to infinity, as long as the above condition holds, even when the rest of the parameters s, p, L and N grow to infinity alongside n. We discuss the meaning of the "excess" qualification for the sample complexity in the remarks below.

Bound (14) has a logarithmic dependence on N, the number of variables in the process, which is a notable feature of our work. Compared to some of the previous work [27], we overcome the $N > 1$ barrier for the BAR model while allowing for $p > 1$ dependence on the past. The bound also depends logarithmically on L. This means that dictionary $\mathbf{D}$ can be overcomplete, allowing for Θ^* to be sparse, for nearly no additional cost.

A. Remarks on Theorem 1

Let us make a few comments on the various choices in Theorem 1:

a) *Choice of the Loss $\mathcal{L}$:* Theorem 1 holds for any loss function satisfying conditions (A2) and (A3). For the Bernoulli AR process, the negative log-likelihood $\mathcal{L}_{i,t}(u, v) = -u \log f_i(v) - (1 - u) \log(1 - f_i(v))$ satisfies these assumptions for any log-concave f_i; see [1]. For the Truncated-Poisson AR process, the negative log-likelihood takes the form $\mathcal{L}_{it}(u, v) = f_i(v) - u \log f_i(v) + \log(u!)$ and satisfies the assumptions for $f_i(v) = \exp(v)$ or $f_i(v) = \log(1 + e^v)$.

b) *Choice of $\mathcal{U}$:* The result in Theorem 1 has been stated for a general $\mathcal{U} \subseteq [N]$. Taking $\mathcal{U} = [N]$, gives a bound on the Frobenius norm of the entire tensor $\|\widehat{\Theta} - \Theta^*\|_F^2$. On the other extreme, we can take $\mathcal{U} = \{i\}$ to obtain bounds on each slice of the tensor with better scaling with sparsity. For example, in the exact sparsity setting, we obtain $\|\widehat{\Theta}_i - \Theta_i^*\|^2 \lesssim s_i \log(NL)/n$, avoiding the extra price of $(\sum_{j \neq i} s_j) \log(NL)/n$ that we pay for the entire tensor.

c) *Scaling With Sparsity:* Considering the exact sparsity setting, the scaling of the sample complexity (15) with sparsity is $n = \Omega(s_+ \vee s_{\max}^3)$. In the worst case, $s_+ = s_{\max}$ and we get a cubic dependence on sparsity which is not ideal. However, when $s_+ \gtrsim s_{\max}^3$, Theorem 1 requires $n = \Omega(s_+)$ which is the optimal scaling with sparsity. (This can be seen by noting that in the linear independent setting, one cannot do better than $n = \Omega(s_+)$.) Our result also holds for the more general case of $\omega_i \neq 0$. For example, for the ℓ_q ball sparsity with $q \in (0, 1)$, we have $\omega_i = \mathcal{O}(s_i^{1-1/q})$ hence $\omega_i^2/s_i + w_i = \mathcal{O}(\omega_i) = \mathcal{O}(s_i)$ and $\widetilde{\omega}_+ = \mathcal{O}(s_+)$ and the same sample complexity as the exact sparsity case holds.

It is not clear if the worst-case cubic dependence on the sparsity can be improved without imposing restrictive assumptions. It is worth noting that in our proof, the additional s_i^2 factor comes from concentration inequality (33) in Lemma 5. This additional factor can be removed if one were able to show sub-Gaussian concentration for deviations of the order of $\|\boldsymbol{\beta}\|_F^2$ instead of $\|\boldsymbol{\beta}\|_{1,1}^2$, in Lemma 5. It remains open whether such concentration is possible and under what additional assumptions. Section VII provides a more detailed discussion on this concentration inequality. Figure 4(a) in Section IV suggests a superlinear dependence on s, hinting that the situation may not be as simple as the i.i.d. case.

For $p = 1$, a sample complexity of $\rho^3 \log(N)$ was reported in [21, Corollary 1]. One can verify that ρ in their model is equal to $s_{\max}$ in ours, hence they obtain the same $s_{\max}^3$ dependence on sparsity. Similarly for $p = 2$, the result in [23, Th. 4.4] requires $(s/r_\rho^2)\log(N)$ samples where s and r_ρ are sparsity parameters defined therein and r_ρ is inversely related to $s_{\max}$ in the worst case, yielding a similar cubic dependence on sparsity as ours. Furthermore, it appears that their analysis only holds for $s_{\max} = \mathcal{O}(1)$, whereas we make no such assumption. In short, to our knowledge, no prior work has broken the $s_{\max}^3$ barrier in the non-Gaussian AR setting.

d) *Scaling With Lag p:* Our result is the first to provide sufficient conditions for a sample complexity logarithmic in p in the case of the identity dictionary, for any value of N. As will be discussed in Section III-B, the dependence of the (excess) sample size n on p could be as good as $O(\log L)$ for a general dictionary, under certain tail and normalization conditions. In these cases, one could obtain an $O(1)$ growth of n as function of p in the best case (when $L = O(1)$) and an $O(\log p)$ growth in the worse case (the identity dictionary). In contrast, [27, Th. 1] requires $s^{2/3}p^{2/3}\log(p)$ samples, for the identity dictionary, and their proof relies heavily on $N = 1$.

Our bound scales with p through G which is defined in terms of the contraction coefficient $\tau_1(\Theta^*)$ in (11). The contraction coefficient only depends on Θ^* and is always less than 1. Intuitively, if Θ^* is too large, then for two different initializations $\boldsymbol{z}$ and $\boldsymbol{y}$, the distributions $\mathbb{P}(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = \boldsymbol{y})$ and $\mathbb{P}(\mathbf{X}^{t+p} = \cdot \mid \mathbf{X}^t = z)$ may significantly differ. A clear sufficient condition for $G = \mathcal{O}(1)$ is to have $\tau_1(\Theta^*) = \mathcal{O}(p^{-1})$ as well as $C_{\mathbf{D}} \lesssim 1$. The challenge is to control $\tau_1(\Theta^*)$ in terms of the size of Θ^*. Section III-B further discusses sufficient conditions under which $G = \mathcal{O}(1)$. There, we show that for certain exponential families, the scaling depends on the behavior of the tail of $k \mapsto |(\boldsymbol{d}_\ell)_k|$, that is, how fast the *influence from the past* dies down in the filters $\{\boldsymbol{d}_\ell\}$.

A subtle point worth noting here, which does not arise in ordinary M-estimation with i.i.d. measurements, is that n is in fact the *excess* sample-size one needs beyond the p initial samples. It is clear that at least p initial samples are needed for estimating a p-lag process. Examples discussed in Section III-B provide conditions that guarantee that the excess sample size, n, needed for consistent estimation is $O(\log L)$ as p grows, the smallest order one could hope for.

e) *Stability Assumption (A1):* We use Assumption (A1) to guarantee that the strong convexity holds for the population loss $\Theta \mapsto \mathbb{E}\mathcal{L}(\Theta)$. This is key in guaranteeing that any parameter tensor $\widehat{\Theta}$ that maximizes the regularized loss function in (5) does not deviate far from the true parameter Θ^*.

Assumption (A1) is by now standard in time-series estimation literature [9], [10], [35]. The quantity $C_{\mathcal{X}}$ is fundamental to multivariate time-series analysis, however, its behavior as a function of the parameters of the model is not yet fully understood. Intuitively, $C_{\mathcal{X}}$ is related to the *flatness* of the power spectral density (PSD) $\mathcal{X}$, and the stability of the process. For the $N = 1$ case, $C_{\mathcal{X}} > 0$ implies that the process does not have zeros on the unit circle in the spectral domain.

In general, $C_{\mathcal{X}}$ could potentially depend on N, indirectly via Θ^*. In subsequent discussions of Theorem 1, we have assumed that $C_{\mathcal{X}}$ stays uniformly bounded away from zero as N grows. This assumption is explicitly stated as $C_{\mathcal{X}} \gtrsim 1$. Our main result (Theorem 1), however, holds for all positive values of $C_{\mathcal{X}}$, regardless of its growth rate. Even if $C_{\mathcal{X}} = o(1)$ with respect to N, Theorem 1 still gives a consistency result, albeit with a worse dependence on N.

The dependence of $C_{\mathcal{X}}$ on N occurs through the scaling of the true parameter Θ^*. That $C_{\mathcal{X}}$ is in general bounded below by a constant (or has a slow decay as a function of N) is part of the folklore of the time series literature. It is reasonable to assume that this holds for certain structured Θ^*. However, obtaining exact conditions on Θ^* for $C_{\mathcal{X}} \gtrsim 1$ to hold is, in general, a non-trivial open problem, even for univariate Gaussian AR(p) processes. The main difficulty is that the relation between the power spectral density of the process and its parameter is indirect and via the Z-transform. Nevertheless, conditions are known in special cases. See for example the discussion surrounding in [9, Proposition 2.2], where explicit conditions are given on the parameter matrix of a VAR(1) Gaussian process, for $C_{\mathcal{X}}$ to stay bounded away from zero.

B. Special Cases

Let us now look at the applications of Theorem 1 to two special cases often considered in discrete-valued time series modeling — Binomial and Poisson AR processes. We take $\mathcal{U} = [N]$ throughout this section. To apply the theorem, we need to upper-bound $G_f(\Theta^*)$ in each case. Since the ψ function in (10) is non-decreasing on $[0, 1)$, it is enough to control $\tau_1(\Theta^*)$. In fact, a sufficient condition for $G_f(\Theta^*) = \mathcal{O}(1)$ is to have $\tau_1(\Theta^*) = \mathcal{O}(\frac{1}{p})$ and $C_{\mathbf{D}} = O(1)$.

The quantity $\tau_1(\Theta^*)$ is the maximum total variation distance between the p-step conditional distributions of the process, starting from two initial states $\boldsymbol{y}$ and $\boldsymbol{z}$. The Pinsker's inequality [36, p. 44] can be used to further control the total variation distance by the KL divergence, which is the natural choice for comparing two exponential family distributions with independent coordinates.

Recall $\mathcal{X} = \prod_{i=1}^{N} \mathcal{X}_i \subset [-B, B]^N$ and the notation $\mathbb{P}_z$ from (11). Pinsker's inequality yields

$$\tau_1^2(\Theta^*) \le \sup_{z,y \in \mathcal{X}^{\times p}} \tfrac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_z \| \mathbb{P}_y), \tag{16}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the KL-divergence. We now state upper bounds on $D_{\mathrm{KL}}(\mathbb{P}_z\|\mathbb{P}_y)$ for the two cases of the Binomial and Poisson processes. A quantity of interest is the tail decay of the dictionary elements $\{\boldsymbol{d}_\ell\}_{\ell=1}^{L}$, measured by

$$\gamma_{t\ell} := \sum_{m=t}^{p} \big|(\boldsymbol{d}_\ell)_m\big|. \tag{17}$$

Let us define the following norm on Θ,

$$\|\Theta\|_\star := \left(\sum_{i,t} L_i^2 \left[\sum_{j,\ell} \gamma_{t\ell} |\Theta_{ij\ell}| \right]^2 \right)^{1/2}$$

where L_i is the Lipschitz constant of the link function f_i, and the summations run over $(i, t, j, \ell) \in [N] \times [p] \times [N] \times [L]$. One can often establish a bound of the form

$$D_{\mathrm{KL}}(\mathbb{P}_z\|\mathbb{P}_y) \le C_f B^2 \|\Theta^*\|_\star^2 \tag{18}$$

where C_f depends on $\{f_i\}$ and Θ^* is the true parameter generating the samples.

Lemma 1: Consider a Binomial AR process given by (3) with $\mathcal{X}_i = \{0, 1, \dots, K_i\}$, where $K_i \le B$, and $\mathbb{Q}_i(\cdot\,|\,z) = \mathrm{Bin}(K_i, z)$. Assume that f_i is L_i-Lipschitz, and for some $\varepsilon \in (0, \frac{1}{2})$, $f_i : \mathbb{R} \to [\varepsilon, 1-\varepsilon]$ for all i. Then, (18) holds with $C_f = 6/\varepsilon$.

The case of $B = 1$ recovers the result for the Bernoulli Autoregressive Process in [1].

Lemma 2: Consider a Truncated Poisson AR process given by (3) with $\mathcal{X}_i = \{0, 1, \dots, K_i\}$ and $\mathbb{Q}_i(\cdot\,|\,z) = \mathbb{P}(\min(K_i, Z) \in \cdot)$ where $Z \sim \mathrm{Poi}(z)$ and $K_i \le B$. Assume that f_i is L_i-Lipschitz, and for some $\varepsilon > 0$, $f_i : \mathbb{R} \to [\varepsilon, \infty)$ for all i. Then, (18) holds with $C_f = 4/\varepsilon$.

Combining with (16), we have the following corollary.

Corollary 1: Under the assumptions of Lemma 1 or 2,

$$\tau_1(\Theta^*) \lesssim \frac{B}{\sqrt{\varepsilon}} \|\Theta^*\|_\star.$$

In particular, if $C_{\mathcal{L}}, C_{\mathbf{D}} \lesssim 1 \lesssim C_{\mathcal{X}}$ and $\|\Theta^*\|_\star = O(1/p)$, then $G = O(1)$ and the following is sufficient for consistency:

$$n \gg \max\Big\{ s_{\max}^3,\ s_+,\ (\widetilde{\omega}_+)^2 \Big\} \log(NL).$$

In other words, Corollary 1 provides conditions under which consistent estimation is possible with (excess) sample complexity that grows at most logarithmically in L.

Let us consider some examples for which $\|\Theta^*\|_\star = O(1/p)$. For the purpose of illustration, let us separate the tail decay of Θ^*, along the lag dimension, by assuming that

$$\left|\Theta^*_{ij\ell}\right| \le R_{ij} h_\ell, \quad \forall\, (i, j, \ell) \in [N] \times [N] \times [L].$$

for some sequence $\{h_\ell\}_{\ell=1}^{\infty}$ such that $\sum_{\ell=1}^{\infty} h_\ell < \infty$ and a matrix $R = (R_{ij})$. Assume that $\Theta^*_{ij\ell}$ is normalized so that $\|R\|_{2,1} = O(1)$. Moreover, assume that $\max_i L_i = O(1/p)$. Since in model (3), the input to each f_i involves terms $\langle \mathbf{x}_j^{t-*}, \boldsymbol{d}_\ell \rangle_{\mathbb{R}^p}$, each of which is essentially a sum of p terms (see (4)), the aforementioned assumption on the Lipschitz constant is a natural normalization that prevents the saturation of the nonlinearities f_i as p grows. Equivalently, we can make this condition more explicit by replacing $f_i(\cdot)$ in the definition of model (3) with $\tilde{f}_i(\frac{1}{p}\cdot)$ and assuming that $\tilde{f}_i$ have Lipschitz constants uniformly bounded by a constant.

Under the above modeling assumptions, consider the following two dictionaries:

Case (a): The identity dictionary, where $L = p$ and $(\boldsymbol{d}_\ell)_m = 1\{m = \ell\}$. In this case, $\gamma_{t\ell} = 1\{t \le \ell\}$. Then,

$$\|\Theta\|_\star \lesssim \frac{1}{p} \|R\|_{2,1} \left[\sum_{t=1}^{p} \left(\sum_{\ell=t}^{p} h_\ell \right)^2 \right]^{1/2} = O\Big(\frac{1}{p}\Big)$$

assuming that $\sum_{t=1}^{\infty} (\sum_{\ell=t}^{\infty} h_\ell)^2 < \infty$ which holds, for example, if h_ℓ decays at least as fast as $\ell^{-1-\alpha/2}$ for some $\alpha > 1$. Note that in this case $C_{\mathbf{D}} \asymp 1$ is trivially satisfied.

Case (b): A general dictionary, with filters satisfying the decay rate $\max_\ell |(\boldsymbol{d}_\ell)_m| \lesssim m^{-\alpha-1}$ for some $\alpha > 1$. Then, $\max_\ell \gamma_{t\ell} \lesssim t^{-\alpha}$ and

$$\|\Theta\|_\star \lesssim \frac{1}{p} \|R\|_{2,1} \left(\sum_{t=1}^{p} t^{-2\alpha} \right)^{1/2} \sum_{\ell=1}^{p} h_\ell = O\Big(\frac{1}{p}\Big)$$

using $\sum_{t=1}^{\infty} t^{-2\alpha} < \infty$ and $\sum_{\ell=1}^{\infty} h_\ell < \infty$. Moreover, since we have $C_{\mathbf{D}} \lesssim \sum_{m=1}^{p} m^{-\alpha-1}$, it follows that $C_{\mathbf{D}} = O(1)$ as p grows.

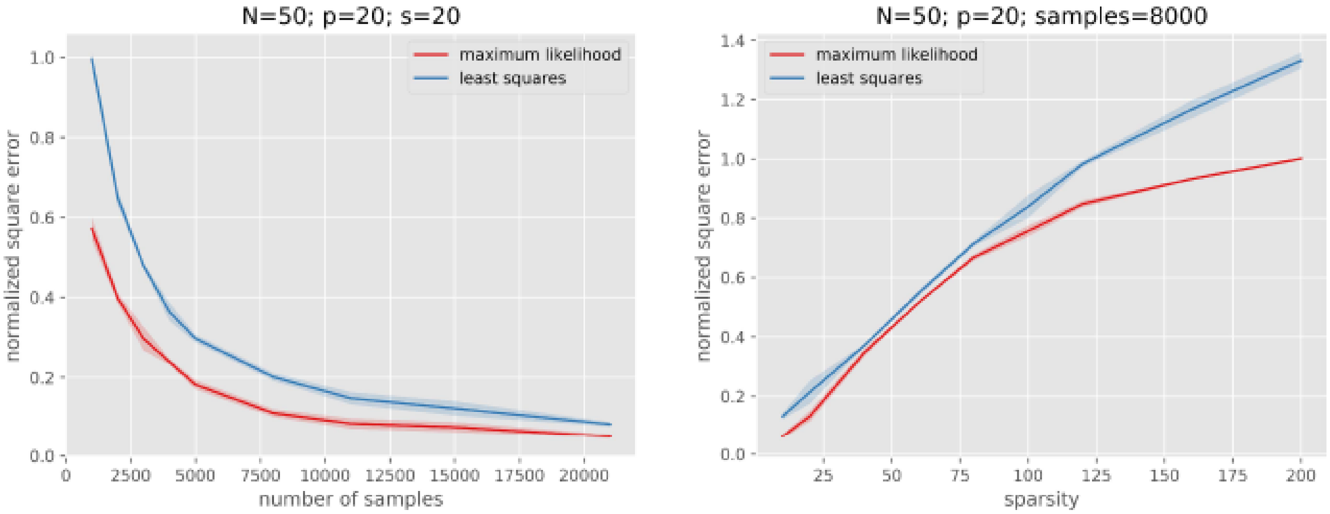
Thus in both cases, Corollary 1 guarantees that the excess sample size n needed for consistency grows at most logarithmically in L. This translates to an $O(\log p)$ growth in the case the identity dictionary but could be as low as $O(1)$ for a dictionary with the number of filters L not growing with p. Note that the summability condition on h_ℓ in case (b) is milder than that in case (a), showing the trade-off between the tail decay of Θ (along the lag dimension) and the tail decay of the dictionary filters. Having fast decaying filters relaxes the decay requirement on the tails of Θ.

IV. Simulations

In this section, we evaluate the performance of the estimator in (5) using simulated data. We generate the data using the model in (3). In all the examples, we first randomly generate Θ^* and $\mathbf{D}$. To generate Θ^*, we select the support of Θ_i^* for each i uniformly at random based on the sparsity s_i. We then fill the support with i.i.d. draws of the normal distribution, and finally normalize such that $\|\Theta_i^*\|_{1,1}$ is a constant.

To report the performance of (5), we use the metric normalized squared error (NSE) defined as:

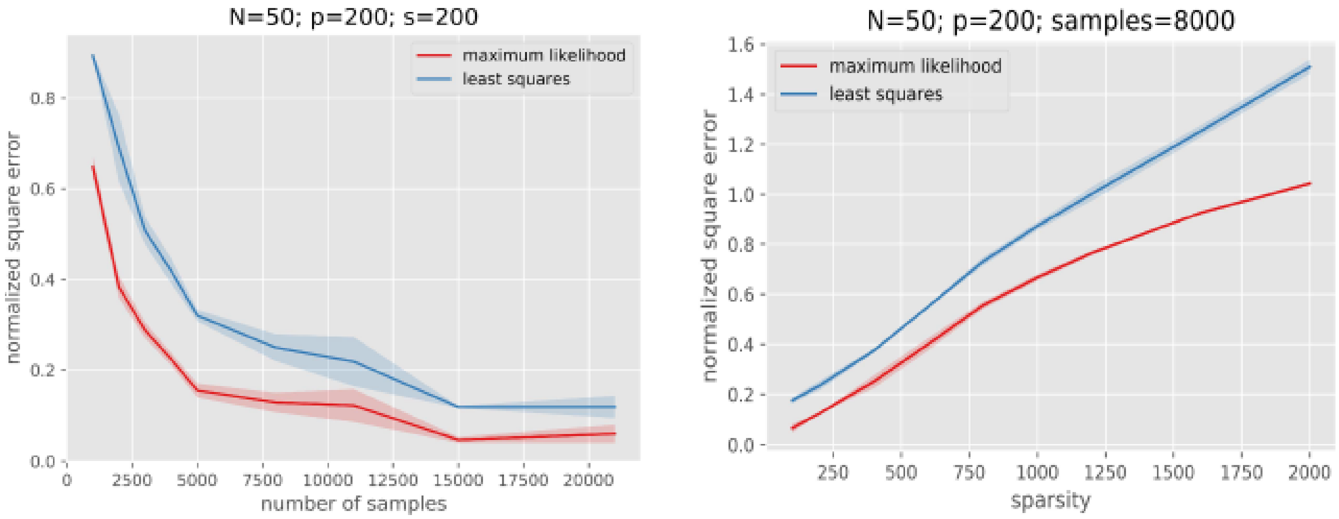
$$\mathrm{NSE}(\Theta^*, \widehat{\Theta}) = \frac{\|\Theta^* - \widehat{\Theta}\|_F^2}{\|\Theta^*\|_F^2}. \tag{19}$$



(a) NSE vs. sample size for a Poisson process without dictionary.

(b) NSE vs. sparsity for a Poisson process without dictionary.

Fig. 1. Poisson AR(p) process without a dictionary (i.e., $\mathbf{D} = \boldsymbol{I}_p$).



(a) NSE vs. sample size for a Poisson process with dictionary.

(b) NSE vs. sparsity for a Poisson process with dictionary.

Fig. 2. Poisson AR(p) process with dictionary of size $L = 20$.

to normalize variations in the size of the parameter across independent instances of Θ^*. An implementation is provided at [2]. We consider the following 3 processes.

A. Poisson AR(p) Process Without Dictionary

We evaluate the performance of the regularized maximum likelihood and the regularized least-squares estimators on a Poisson process with no dictionary, i.e., $\mathbf{D} = \boldsymbol{I}_p$. For the Poisson process, we use the inverse link function $f_i(z) = \log(1 + e^z)$. Then, these estimators have the form of (5) with

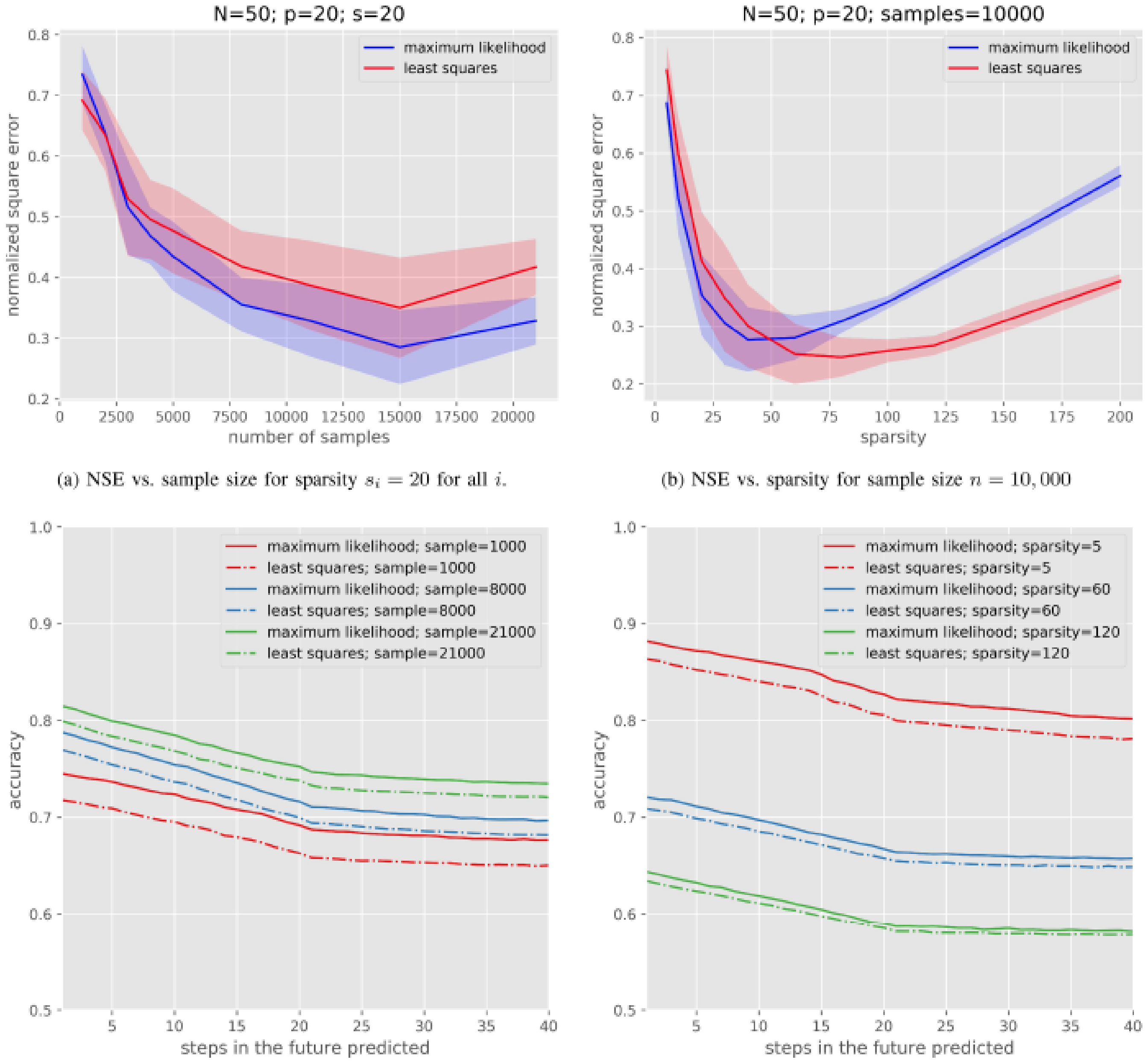
$$\mathcal{L}_{it}^{\mathrm{ML}}(x_i^t; z_i^t) = z_i^t - x_i^t \log(z_i^t), \qquad (20a)$$

$$\mathcal{L}_{it}^{\mathrm{LS}}(x_i^t; z_i^t) = (x_i^t - z_i^t)^2, \qquad (20b)$$

where $z_i^t = f(\langle \Theta_i^*, \mathbf{X}^{t-1} \rangle)$, since $\mathbf{D} = \boldsymbol{I}_p$. Note that the M-estimation problem in (5) corresponding to (20a) is convex, whereas it is non-convex for (20b) (we report a local minimum). Here, we generate the ground truth parameters as mentioned before with $N = 50$ and $p = 20$ and we use $\lambda_n = 0.05/\sqrt{n}$. When comparing *NSE* v/s n, each Θ_i has sparsity 20. The results are shown in Figure 1. The error shades correspond to one standard deviation over 5 independent instances of $(\Theta^*, \widehat{\Theta})$. With the NSE metric, the regularized maximum likelihood estimator appears to perform better for the Poisson AR(p) process, for the random ensemble of problems generated in these examples.

B. Poisson AR(p) Process With Dictionary

We choose $\mathbf{D}$ to be entrywise i.i.d. Gaussian with standard deviation σ/p for a constant σ, so that the ℓ_1-norm of all columns of $\mathbf{D}$ are close to a constant for large p (the constant being the mean of a folded normal distribution). The process is generated as in the previous example using (3). We take $N = 50, p = 200$, and $L = 20$ such that the process has very long range dependencies. We again consider the two



(a) NSE vs. sample size for sparsity $s_i = 20$ for all i.

(b) NSE vs. sparsity for sample size $n = 10,000$

(c) Accuracy vs. steps predicted in the future for different n.

(d) Accuracy vs. steps predicted in the future for different s.

Fig. 3. Bernoulli AR(p) process without dictionary.

regularized M-estimators: the regularized maximum likelihood and the regularized least-squares with the inverse link function $f(z) = \log(1 + e^z)$. These estimators are identical to the ones in (20a) and (20b), except that $z_i^t = f(\langle \Theta_i, \mathbf{X}^{t-1}\mathbf{D}\rangle)$ with $\mathbf{D} \neq \boldsymbol{I}_p$.

The results are shown in Figure 2. They are very similar to Figure 1. In accordance with our theoretical results, these figures suggest that for an AR processes with very long range dependencies, estimating the parameter is easier in the presence of a dictionary.

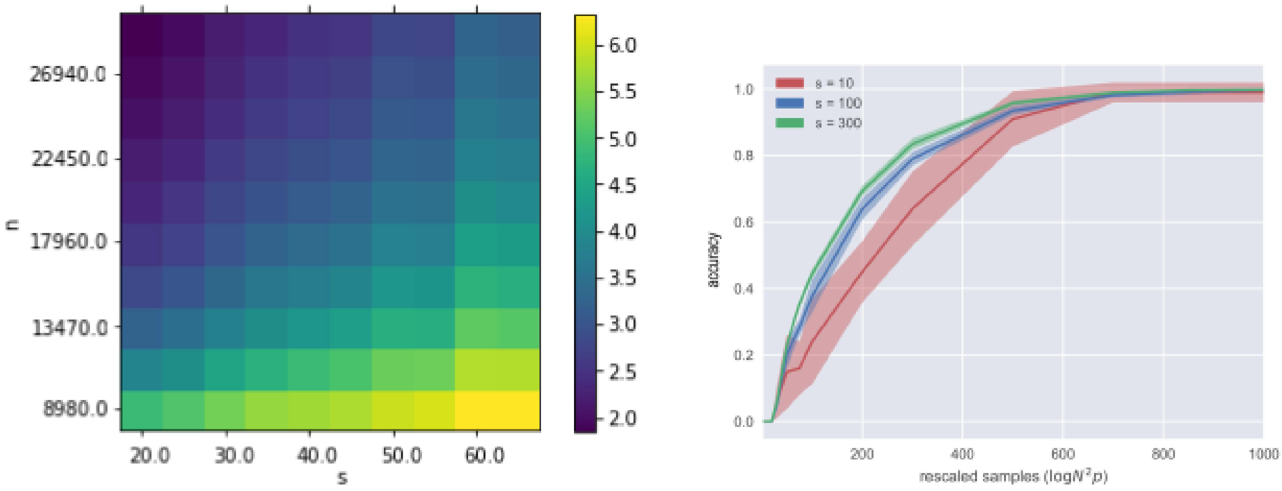
C. Bernoulli AR(p) Process Without Dictionary

Finally, we look at a Bernoulli autoregressive process. We use the sigmoid function, $f(z) = 1/(1 + e^{-z})$, as the inverse link function. We compare the performance of regularized maximum likelihood estimator to regularized least-squares estimator. Both of these estimators have the form of (5) with

$$\mathcal{L}_{it}^{\mathrm{ML}}\left(x_i^t; z_i^t\right) = -z_i^t \log\left(x_i^t\right) - \left(1 - x_i^t\right) \log\left(1 - z_i^t\right) \quad (21a)$$

$$\mathcal{L}_{it}^{\mathrm{LS}}\left(x_i^t; z_i^t\right) = \left(x_i^t - z_i^t\right)^2, \quad (21b)$$

where $z_i^t = f(\langle \Theta_i, \mathbf{X}^{t-1}\rangle)$ is the mean parameter of the dimension i of the Bernoulli process at time t. Note that due to inverse link function, despite convexity of square loss with respect to z_i^t, the optimization problem corresponding to least square estimator is non-convex and our results do not apply to it. Nevertheless, we observe that its performance is similar to maximum likelihood estimator.



(a) Average Frobenius norm of the error over 20 runs with $N = 20$, $p = 20$. Each pixel corresponds to a pair (s, n) for Θ^*.

(b) Fraction of support recovered by taking the largest s entries of $\hat{\Theta}$ as the estimator of support. Here $N = 100$, $p = 1$.

Fig. 4. Simulation results for Bernoulli AR(p) process.

Figure 3 shows different measures of performance of the regularized maximum likelihood estimator. We have set $N = 50$, $p = 20$ and $\lambda_n = 0.05/\sqrt{n}$ as recommended by Theorem 1, in these examples. Figure 3(a) shows how the normalized estimation error changes with respect to the number of training samples.

The sparsity is 20 for each Θ_i. Note that we are using the same regularization parameter for both estimators and not the optimal λ_n, i.e., without any cross-validation. The error shades correspond to one standard deviation. Figure 3(b) shows the normalized square error for different sparsity levels. For small values of sparsity, the denominator Θ^* has a small norm which causes high normalized error, however for higher values of sparsity, we see the linear dependence on sparsity as predicted by Theorem 1.

The next two figures correspond to generalization error as opposed to estimation error in the first two figures. Here, we use the estimated parameters $\widehat{\Theta}$ to predict the process in the future and calculate the accuracy of prediction. We use 5 MCMC runs of the process to estimate the accuracy. The plot shows average accuracy over all N variables of the process. Figure 3(c) shows the accuracy vs. steps in the future for different training sample sizes and Figure 3(d) shows it for different levels of sparsity. There is a prominent change in the accuracy plots at 21 steps. This corresponds to $p = 20$ where the future of the process is being estimated purely based on simulated samples using the estimated parameter. Prior to this point, parts of the samples being used to make the predictions are True values and not estimated ones. As expected, the accuracies improve as the number of training samples increase with sparsity fixed, and they decrease as sparsity level increases with number of training samples fixed. Figure 4(a) shows the estimation error for different sample sizes and sparsity levels.

Finally, we also use the regularized maximum likelihood estimator to perform support recovery, i.e., assuming that the true parameter tensor is exactly s-sparse, how does the support estimated from $\widehat{\Theta}$ compare to the support of Θ^*? To do so, we need to estimate the support from $\widehat{\Theta}$. If we know the sparsity s, we can estimate the support by taking the indices corresponding to the s largest entries of $\widehat{\Theta}$ in magnitude. If we do not know the sparsity in advance, we can estimate the support based on a threshold chosen by cross-validation. Given a threshold γ, the estimated support would be

$$\widehat{\text{supp}}(\Theta) := \left\{(j, k, \ell) : \left|\widehat{\Theta}_{jk\ell}\right| \geq \gamma\right\}.$$

Note that our theoretical results do not give any guarantees for support recovery. In order to guarantee support recovery, a stronger result bounding the error uniformly for each entry of $\widehat{\Theta}$ is required, i.e., we need to control $\|\widehat{\Theta} - \Theta^*\|_{\infty,\infty,\infty}$ with high probability. Therefore, more work is needed to obtain theoretical guarantees for support recovery. Nevertheless, our simulations show that the estimator is able to recover the support very well. Figure 4(b) shows the results for a process with $p = 1, N = 100$ and three different sparsities. For recovering the support, we assumed that the sparsity s is known, and took the indices corresponding to the s largest entries of $\widehat{\Theta}$ as the recovered support. The fraction of the correctly recovered indices is plotted against the sample size. Figure 4(b) shows that if the sample size is below some threshold, no entries of the support are recovered, while above the threshold, the recovered fraction gradually increases to 1.

V. Proof Sketch for Theorem 1

We now outline the proof of Theorem 1. Our analysis applies the framework of Negahban *et al.* [6]. Let

$$\mathcal{L}_i(\boldsymbol{\beta}) := \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{it,}\left(x_i^t; \left\langle \boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\right\rangle\right), \quad \boldsymbol{\beta} \in \mathbb{R}^{N \times L}.$$

Fix $\mathcal{U} \subseteq [N]$ and set $\Theta_{\mathcal{U}} := (\Theta_i)_{i\in\mathcal{U}}$ and similarly $\Theta^*_{\mathcal{U}} := (\Theta^*_i)_{i\in\mathcal{U}}$ and $\widehat{\Theta}_{\mathcal{U}} := (\widehat{\Theta}_i)_{i\in\mathcal{U}}$, all tensors in $\mathbb{R}^{|\mathcal{U}|\times N\times L}$. We also write $\mathcal{L}_{\mathcal{U}}(\Theta_{\mathcal{U}}) = \sum_{i\in\mathcal{U}} \mathcal{L}_i(\Theta_i)$. We have

$$\widehat{\Theta}_{\mathcal{U}} = \underset{\Theta_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times N\times L}}{\operatorname{argmin}} \mathcal{L}_{\mathcal{U}}(\Theta_{\mathcal{U}}) + \|\Theta_{\mathcal{U}}\|_{1,1,1}. \tag{22}$$

In the sequel, $\nabla\mathcal{L}_{\mathcal{U}}$ and $\nabla^2\mathcal{L}_{\mathcal{U}}$ are the gradient and Hessian of $\mathcal{L}_{\mathcal{U}}$ with respect to variable $\Theta_{\mathcal{U}}$. When $n \ll |\mathcal{U}|NL$, the empirical Hessian, $\nabla^2\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})$, is rank-deficient, hence the loss function is flat in many directions around $\Theta^*_{\mathcal{U}}$. The approach of Negahban *et al.* [6] is to guarantee that $\mathcal{L}_{\mathcal{U}}$ is positively curved in certain directions, including $\widehat{\Delta}_{\mathcal{U}} := \widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}}$.

In particular, if the regularization parameter λ_n is large enough, specifically

$$\lambda_n \geq 2\|\nabla\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})\|_{\infty,\infty,\infty}, \tag{23}$$

then, the error tensor $\widehat{\Delta}_{\mathcal{U}}$ lies in a small *cone-like* subset $\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$—to be defined below—and on this set, $\mathcal{L}_{\mathcal{U}}$ is "nearly" strongly convex, i.e., $\nabla^2\mathcal{L}_{\mathcal{U}}(\Theta^*_{\mathcal{U}})$ is uniformly quadratically bounded below.

For a set $S \subseteq [N]\times[L]$, let $\boldsymbol{\beta}_S$ denote the projection of $\boldsymbol{\beta}$ on the subspace of matrices with support S. For $\boldsymbol{\beta}^*$ define:

$$\mathbb{C}(S;\boldsymbol{\beta}^*) := \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_{1,1} \leq 3\|\boldsymbol{\beta}_S\|_{1,1} + 4\left\|\boldsymbol{\beta}^*_{S^c}\right\|_{1,1}\right\}. \tag{24}$$

Note that this is a *cone-like* subset of $\mathbb{R}^{N\times L}$ around $\boldsymbol{\beta}^*$. See [6] for a visualization. Let $\mathcal{S} := \bigcup_{i\in\mathcal{U}}\{i\}\times S_i$ where $S_i \subseteq [N]\times[L]$ for $i \in \mathcal{U}$. Equivalently, $\mathcal{S} = \bigsqcup_{i\in\mathcal{U}} S_i$ using the notation of *disjoint union*. With some abuse of notation, we write $\mathcal{S}^c := \bigcup_{i\in\mathcal{U}}\{i\}\times S_i^c$. The cone-like set $\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$ is defined as follows:

$$\mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}}) := \left\{(\Delta_i)_{i\in\mathcal{U}} : \Delta_i \in \mathbb{C}(S_i;\Theta^*_i),\ \forall i \in \mathcal{U}\right\}. \tag{25}$$

For loss functions $\mathcal{L}_i$, $i \in \mathcal{U}$, and for $\boldsymbol{\delta}, \boldsymbol{\beta}^* \in \mathbb{R}^{N\times L}$, let

$$R\mathcal{L}_i(\boldsymbol{\delta};\boldsymbol{\beta}^*) := \mathcal{L}_i(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \mathcal{L}_i(\boldsymbol{\beta}^*) - \left\langle\nabla\mathcal{L}_i(\boldsymbol{\beta}^*), \boldsymbol{\delta}\right\rangle, \tag{26}$$

be the remainder of the first-order Taylor expansion of $\mathcal{L}_i$ around $\boldsymbol{\beta}^*$. Following [6], we say that $\mathcal{L}_{\mathcal{U}}$ satisfies restricted strong convexity (RSC) at $\Theta^*_{\mathcal{U}}$ with curvature $\kappa > 0$ and tolerance τ^2 if for all $\Delta \in \mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$, we have,

$$\sum_{i\in\mathcal{U}} R\mathcal{L}_i(\Delta_i;\Theta^*_i) \geq \kappa \sum_{i\in\mathcal{U}} \|\Delta_i\|_F^2 - \tau^2. \tag{27}$$

The left-hand side is the remainder of the first-order Taylor expansion of $\mathcal{L}_{\mathcal{U}}$ around $\Theta^*_{\mathcal{U}}$, that is, $R\mathcal{L}_{\mathcal{U}}(\Delta_{\mathcal{U}};\Theta^*_{\mathcal{U}})$—defined similar to (26).

Now, assume that (23) and (27) hold. Then, [6, Th. 1] implies that $\widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}} \in \mathcal{C}(\mathcal{S};\Theta^*_{\mathcal{U}})$, and that

$$\|\widehat{\Theta}_{\mathcal{U}} - \Theta^*_{\mathcal{U}}\|_F^2 \leq \tfrac{9\lambda_n^2}{\kappa^2}|\mathcal{S}| + \tfrac{\lambda_n}{\kappa}\left(2\tau^2 + 4\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^c}\|_{1,1,1}\right). \tag{28}$$

The above inequality provides a family of bounds, one for each choice of $\mathcal{S} = \bigsqcup_{i\in\mathcal{U}} S_i$. Decreasing $|\mathcal{S}|$ reduces the first term, but potentially increases $\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^c}\|_{1,1,1}$. We choose $\mathcal{S}$ to balance the two. Let $S_i^* \subset [N]\times[L]$ be the support of the minimizer in (9), so that $|S_i^*| = s_i$. We take $\mathcal{S} = \mathcal{S}^* = \bigsqcup_{i\in\mathcal{U}} S_i^*$. Consequently, $|\mathcal{S}^*| = \sum_{i\in\mathcal{U}} s_i$ and $\|(\Theta^*_{\mathcal{U}})_{\mathcal{S}^{*c}}\|_{1,1,1} = \sum_{i\in\mathcal{U}} \omega_i$. For this choice of $\mathcal{S}$, Proposition 1 below shows that (27) holds, with high probability. To state the concentration inequality, recall the definitions (12).

Proposition 1: Under Assumptions (A1) and (A2), if we have,

$$n \gtrsim \frac{G}{C_{\mathcal{X}}^6} s_{\max}^3 \log(NL) \tag{29}$$

then, the RSC property (27) for $\mathcal{S} = \mathcal{S}^*$ holds with curvature $\kappa = \underline{\kappa}$ and tolerance $\tau^2 = \frac{\overline{\kappa}}{2}\sum_{i\in\mathcal{U}} \omega_i^2/s_i$, with probability at least $1 - (NL)^{-Cs_{\max}}$ where $C = \mathcal{O}(C_{\mathcal{X}}^{-2})$.

Lemma 12 in Appendix A in the supplementary material shows that $\Theta^*_{\mathcal{U}}$ is in fact the minimizer of the expected loss $\mathbb{E}\mathcal{L}_{\mathcal{U}}(\cdot)$. Lemma 13 in Appendix A in the supplementary material shows that taking $\lambda_n = \mathcal{O}(\sqrt{\log(|\mathcal{U}|NL)/n})$ is enough for (23) to hold with high probability. Putting the pieces together proves Theorem 1. The next section sketches a proof of Proposition 1.

VI. Restricted Strong Convexity: Proof of Proposition 1

Showing the RSC property (27) for a particular choice of $\mathcal{S}$ is a major contribution of our work. This is a nontrivial task since it involves uniformly controlling a dependent non-Gaussian empirical process. Even for i.i.d. samples, the task is challenging since the quantity to be controlled, $\Delta \mapsto R\mathcal{L}(\Delta;\Theta^*)$, is a *random function* that needs to be uniformly bounded below. Controlling the behavior of this function becomes significantly harder without the independence assumption.

We proceed by a establishing a series of intermediate lemmas which are proved in Appendix A in the supplementary material. First, we show that $\boldsymbol{\beta} \mapsto R\mathcal{L}_i(\boldsymbol{\beta};\Theta^*_i)$ is lower-bounded by the following quadratic form:

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) := \frac{1}{n}\sum_{t=1}^{n}\left\langle\boldsymbol{\beta}, \mathbf{X}^{t-1}\mathbf{D}\right\rangle^2, \tag{30}$$

where $\mathbb{X} := \{\boldsymbol{x}^t\}_{t=-p+1}^n$.

Lemma 3 (Quadratic Lower Bound): Under Assumption (A2),

$$R\mathcal{L}_i(\boldsymbol{\beta};\Theta^*_i) \geq \frac{\kappa_i}{2}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) \tag{31}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^{N\times L}$ and $i \in [N]$.

Notice that $\boldsymbol{\beta} \mapsto \mathcal{E}(\boldsymbol{\beta};\mathbb{X})$ is a random function due to the randomness in $\mathbb{X}$. Importantly, $\mathcal{E}(\cdot\,;\mathbb{X})$ does not depend on the choice of i. The following set of results establish some important properties of the random function $\mathcal{E}(\cdot\,;\mathbb{X})$.

Lemma 4 (Strong Convexity at the Population Level): Under Assumption (A1),

$$\mathbb{E}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) \geq C_{\mathcal{X}}^2\,\|\boldsymbol{\beta}\|_F^2, \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{N\times L}. \tag{32}$$

Next, we show that for a fixed $\boldsymbol{\beta}$, the quantity $\mathcal{E}(\boldsymbol{\beta};\mathbb{X})$ concentrates around its mean. Section VII provides a sketch of the proof of the following concentration inequality:

Lemma 5 (Concentration Inequality): For any $\boldsymbol{\beta} \in \mathbb{R}^{N\times L}$, if $\mathbb{X}$ is generated as (3), then with probability at least $1 - 2\exp(-nt^2/G)$, we have

$$\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) > \mathbb{E}\mathcal{E}(\boldsymbol{\beta};\mathbb{X}) - t\|\boldsymbol{\beta}\|_{1,1}^2. \tag{33}$$

Finally, for a fixed $i \in [N]$ we use the structural properties of set $\mathbb{C}(S_i^*;\Theta^*_i)$ along with Lemmas 4 and 5 to give a uniform quadratic lower bound on $\mathcal{E}(\boldsymbol{\beta};\mathbb{X})$, which holds with high probability:

Lemma 6: Fix $i \in \mathcal{U}$. For constants $C_1, C_2 > 0$, if $s_i \geq \frac{C_{\mathbb{X}}^2}{C_1}$, then with probability $\geq 1 - \exp(\frac{C_2}{C_{\mathbb{X}}^2} s_i \log(NL) - \frac{nC_{\mathbb{X}}^4}{16Gs_i^2})$,

$$\mathcal{E}(\boldsymbol{\beta}; \mathbb{X}) \geq \frac{C_{\mathbb{X}}^2}{4}\|\boldsymbol{\beta}\|_F^2 - \omega_i^2/s_i, \quad \forall \boldsymbol{\beta} \in \mathbb{C}(S_i^*; \Theta_i^*).$$

The proof of Lemma 9 (see Appendix B) in the supplementary material makes use of a discretization argument. Proving uniform laws are challenging when the parameter space is not finite. The discretization of the set $\mathbb{C}(S^*; \Theta^*)$ uses estimates of the *entropy numbers* for absolute convex hulls of collections of points (Lemma 14 in the supplementary material). These estimates are well-known in approximation theory and have been previously adapted to the analysis of regression problems in [7]. The following technical lemma allows us to put the above results together:

Lemma 7: For all $i \in \mathcal{U}$, let a_i, b_i, d_i, p_i be positive constants, and consider random variables $X_i, Y_i \in \mathbb{R}$ which satisfy $Y_i \geq a_i X_i$, and $\mathbb{P}(X_i < b_i - d_i) \leq p_i$ for all $i \in \mathcal{U}$. Then with probability at least $1 - |\mathcal{U}| \max_{i \in \mathcal{U}} p_i$, we have,

$$\sum_{i \in \mathcal{U}} Y_i > \left(\min_{i \in \mathcal{U}} a_i\right) \sum_{i \in \mathcal{U}} b_i - \left(\max_{i \in \mathcal{U}} a_i\right) \sum_{i \in \mathcal{U}} d_i$$

Proposition 1 follows by taking $Y_i = R\mathcal{L}_i(\Delta_i; \Theta_i^*)$, $X_i = \mathcal{E}(\Delta_i, \mathbb{X})$, $a_i = \frac{\kappa_i}{2}$, $b_i = \frac{C_{\mathbb{X}}^2}{4}\|\Delta_i\|_F^2$, and $d_i = \omega_i^2/s_i$.

VII. Concentration Under Dependence: Proof of Lemma 5

In this section, we sketch the proof of Lemma 5 which is a concentration inequality for $\boldsymbol{\beta} \mapsto \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$, a quadratic empirical process based on dependent non-Gaussian variables with long-term dependence. For independent sub-Gaussian variables $\{\mathbf{X}^{t-1}\}$, such a concentration result is often called the Hanson–Wright inequality [11, Th. 1]. Providing similar inequalities for dependent random variables is significantly more challenging. For dependent Gaussian variables, the machinery of the Hanson–Wright inequality can still be adapted to derive the desired result [9, Proposition 2.4]. However, these arguments do not extend easily to non-Gaussian dependent variables and hence other techniques are needed to provide such concentration inequalities.

Recent results [37], [38] on the concentration of empirical processes derived from Markov chains could provide improvements on the results we present here. However, since we are dealing with a non-Markovian process (when $p > 1$), such results are not directly applicable. A key observation, discussed in Section I-B, is that process (3) can be represented as a discrete-space *p-Markov chain*. This allows us to use concentration results for dependent processes in countable metric spaces. There are several results for such processes; see [25], [39], [40], and [41] for a review. Here, we apply that of Kontorovich and Ramanan [25]. These concentration inequalities are stated in terms of various mixing and contraction coefficients of the underlying process. The challenge is to control the contraction coefficients in terms of the process parameter Θ^*, which in our case is done using quantities $\tau_1(\Theta^*)$ and $G_f(\Theta^*)$. Some results developed in this section hold more generally for any $p-$Markov process, even those outside the current autoregressive framework.

We start by stating the result of Kontorovich and Ramanan [25] for a process $\{X^t\}_{t \in [n]}$ consisting of (possibly dependent) random variables taking values in a countable space $\mathcal{X}$. For any $\ell \geq k \geq 1$, define the *mixing coefficient*

$$\eta_{k\ell} \triangleq \sup_{w,w',y} \Big\|\mathbb{P}\Big(X_\ell^n = \cdot \mid X_k = w', X_1^{k-1} = y\Big) - \mathbb{P}\Big(X_\ell^n = \cdot \mid X_k = w, X_1^{k-1} = y\Big)\Big\|_{\mathrm{TV}}, \quad (34)$$

where the supremum is over $w, w' \in \mathcal{X}$ and $y \in \mathcal{X}^{k-1}$. Here, $X_u^v := (X^t, u \leq t \leq v)$ is viewed either as a member of $\mathcal{X}^{\times(v-u+1)}$ (the set of a matrices with $v - u + 1$ columns from $\mathcal{X}$) or simply as a vector in $\mathcal{X}^{v-u+1}$. Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be an upper triangular matrix with entries $\eta_{k\ell}$ for $\ell \geq k$ and zero otherwise. Let $\|\mathbf{H}\|_\infty := \max_k \sum_{\ell \geq k} \eta_{k\ell}$ be the ℓ_∞ operator norm of $\mathbf{H}$.

Proposition 2 [25, Th. 1.1]: Let $\phi : \mathcal{X}^n \to \mathbb{R}$ be an L_ϕ-Lipschitz function of $\{X^t\}_{t=1}^n$ with respect to the Hamming norm, then for all $\varepsilon > 0$, with probability at least $1 - 2\exp(-\frac{\varepsilon^2}{2nL_\phi^2\|\mathbf{H}\|_\infty^2})$, we have

$$\left|\phi(\{X^t\}_{t=1}^n) - \mathbb{E}\phi(\{X^t\}_{t=1}^n)\right| < \varepsilon. \quad (35)$$

We apply the above result to $\phi = \mathcal{E}(\boldsymbol{\beta}; \mathbb{X})$ by finding an upper bound for the Lipschitz constant L_ϕ of the map $\mathbb{X} \mapsto \mathcal{E}(\boldsymbol{\beta}, \mathbb{X})$ with respect to the Hamming distance over $\mathcal{X}^{\times(n+p-1)} = (\prod_{i=1}^N \mathcal{X}_i)^{\times(n+p-1)}$. Lemma 16 in Appendix C shows that $L_\phi \leq (4B^2C_{\mathbf{D}}^2/n)\|\boldsymbol{\beta}\|_{1,1}^2$, whereas Lemma 17 in Appendix C shows that $\|\mathbf{H}\|_\infty^2 \leq 2(1 + p^2\psi_1(\Theta^*))$, where the quantity $\psi_1(\Theta^*)$ is defined below equation (10). Lemma 17 is a general result that applies to any p-lag Markov chain, including the GVAR(p) processes considered in this article. In Appendix C we also develop some tools for controlling $\|\mathbf{H}\|_\infty$ in terms of the contraction coefficient of another related Markov chain obtained via a non-standard state augmentation.

Applying Proposition 2 with $\varepsilon = t\|\boldsymbol{\beta}\|_{1,1}^2$, and using the upper bounds for L and $\|\mathbf{H}\|_\infty^2$ concludes the proof.

VIII. Discussion

Fitting autoregressive AR(p) models with multiple lags is of broad interest in multivariate time series analysis. We consider a large class of multivariate discrete-valued AR(p) processes with nonlinear feedback. We study statistical properties of a general ℓ_1 regularized M-estimator for this model, and provide sufficient conditions on the model hyperparameters under which consistent estimation is possible. Under assumptions of approximate sparsity, our result shows that a sample complexity $\Omega(\mathsf{poly}(s), \log(Np))$ is achievable. Our experiments validate the theoretical results on simulated data. Commonly occurring special cases of discrete-valued processes such as Bernoulli AR(p) and Truncated-Poisson AR(p) are explored in detail. The proof technique develops concentration inequalities and identifies mixing properties of higher order Markov chains which may be of independent interest. These techniques were previously unknown to the best of our knowledge.

Several open questions remain to be uncovered for the general AR(p) model. For example, the current model explores the case of bounded, discrete valued data. Getting around this assumption requires finding concentration inequalities for random averages of the form in Lemma 5 for real-valued random processes. Also, it remains unknown whether the dependence on the sparsity hyperparameter s is optimal, since there is a small gap between our upper bound and the naive lower bound. Finally, it would be interesting to study parameter estimation, and potentially even controls, for the case of partial observability, i.e., when we observe $g(\boldsymbol{x}^t)$ and not $\boldsymbol{x}^t$ fully, akin to partially-observed Markov decision processes (POMDPs).

References

[1] P. Pandit, M. Sahraee-Ardakan, A. Amini, S. Rangan, and A. K. Fletcher, "Sparse multivariate bernoulli processes in high dimensions," in *Proc. 22nd Int. Conf. Artif. Intell. Stat.*, 2019, pp. 457–466.

[2] M. Sahraee-Ardakan, P. Pandit, A. Amini, S. Rangan, and A. K. Fletcher. (2020). *Multivariate Autoregressive Generalized Linear Model Regression in PyTorch*. [Online]. Available: https://github.com/mojtabasah/AR_process

[3] C. De Mol, D. Giannone, and L. Reichlin, "Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components?" *J. Econ.*, vol. 146, no. 2, pp. 318–328, 2008.

[4] A. Timmermann, "Excess volatility and predictability of stock prices in autoregressive dividend models with learning," *Rev. Econ. Stud.*, vol. 63, no. 4, pp. 523–557, 1996.

[5] D. N. DeJong and C. H. Whiteman, "The temporal stability of dividends and stock prices: Evidence from the likelihood function," *Amer. Econ. Rev.*, vol. 81, no. 3, pp. 600–617, 1991.

[6] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[7] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[8] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Ann. Stat.*, vol. 36, no. 4, pp. 1567–1594, 2008.

[9] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Stat.*, vol. 43, no. 4, pp. 1535–1567, 2015.

[10] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multiresponse tensor regression," *Ann. Stat.*, vol. 47, no. 3, pp. 1554–1584, 2019.

[11] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, no. 32, p. 9, 2013.

[12] A. Rakhlin, K. Sridharan, and A. Tewari, "Sequential complexities and uniform martingale laws of large numbers," *Probab. Theory Related Fields*, vol. 161, nos. 1–2, pp. 111–153, 2015.

[13] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[14] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math. J. Courant Inst. Math. Sci.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[15] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[16] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[17] T. T. Cai, Z. Ren, and H. H. Zhou, "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation," *Electron. J. Stat.*, vol. 10, no. 1, pp. 1–59, 2016.

[18] T. L. McMurry and D. N. Politis, "High-dimensional autocovariance matrices and optimal linear prediction," *Electron. J. Stat.*, vol. 9, no. 1, pp. 753–788, 2015.

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

[20] D. F. Ahelegbey, M. Billio, and R. Casarin, "Sparse graphical vector autoregression: A bayesian approach," *Ann. Econ. Stat.*, vols. 123–124, pp. 333–361, Dec. 2016.

[21] E. C. Hall, G. Raskutti, and R. M. Willett, "Learning high-dimensional generalized linear autoregressive models," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2401–2422, Apr. 2019.

[22] H. H. Zhou and G. Raskutti, "Non-parametric sparse additive autoregressive network models," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1473–1492, Mar. 2019.

[23] B. Mark, G. Raskutti, and R. Willett, "Network estimation from point process data," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2953–2975, May 2019.

[24] B. Mark, G. Raskutti, and R. Willett, "Network estimation via poisson autoregressive models," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi Sensor Adapt. Process. (CAMSAP)*, 2017, pp. 1–5.

[25] L. A. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *Ann. Probab.*, vol. 36, no. 6, pp. 2126–2158, 2008.

[26] A. Kazemipour, "Compressed sensing beyond the IID and static domains: Theory, algorithms and applications," 2018. [Online]. Available: arXiv:1806.11194.

[27] A. Kazemipour, M. Wu, and B. Babadi, "Robust estimation of self-exciting generalized linear models with application to neuronal modeling," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3733–3748, Jul. 2017.

[28] D. Katselis, C. Beck, and R. Srikant, "Mixing times and structural inference for bernoulli autoregressive processes," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 3, pp. 364–378, Sep. 2019.

[29] A. I. Weber and J. W. Pillow, "Capturing the dynamical repertoire of single neurons with generalized linear models," *Neural Comput.*, vol. 29, no. 12, pp. 3260–3289, 2017.

[30] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural Comput.*, vol. 17, no. 9, pp. 1927–1961, 2005.

[31] A. C. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Comput.*, vol. 15, no. 5, pp. 965–991, 2003.

[32] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: State-of-the-art and future challenges," *Nat. Neurosci.*, vol. 7, no. 5, p. 456, 2004.

[33] M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia, "Sequential anomaly detection in the presence of noise and limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5544–5562, Aug. 2012.

[34] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," in *Optimization for Machine Learning*. Cambridge, MA, USA: MIT Press, 2011.

[35] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer, 2005.

[36] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[37] J. Fan, B. Jiang, and Q. Sun, "Hoeffding's lemma for markov chains and its applications to statistical learning," 2018. [Online]. Available: arXiv:1802.00211.

[38] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff–Hoeffding bounds for Markov chains: Generalized and simplified," 2012. [Online]. Available: arXiv:1201.0559.

[39] K. Marton *et al.*, "Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration," *Ann. Probab.*, vol. 24, no. 2, pp. 857–866, 1996.

[40] P.-M. Samson *et al.*, "Concentration of measure inequalities for markov chains and Φ-mixing processes," *Ann. Probab.*, vol. 28, no. 1, pp. 416–461, 2000.

[41] A. Kontorovich, "Obtaining measure concentration from Markov contraction," *Markov Processes Related Fields*, vol. 18, no. 4, pp. 613–638, 2012.

[42] S. A. van de Geer, "On Hoeffding's inequality for dependent random variables," in *Empirical Process Techniques for Dependent Data*. Boston, MA, USA: Birkhäuser, 2002, pp. 161–169.

[43] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Found. Trends Commun. Inf. Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[44] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[45] A. Rhodius, "On the maximum of ergodicity coefficients, the dobrushin ergodicity coefficient, and products of stochastic matrices," *Linear Algebra Appl.*, vol. 253, nos. 1–3, pp. 141–154, 1997.

[46] M. Krein and V. Smulian, "On regularly convex sets in the space conjugate to a banach space," *Ann. Math.*, vol. 41, no. 3, pp. 556–583, 1940.