Sampling Requirements for Stable Autoregressive Estimation

Abbas Kazemipour, *Student Member, IEEE,* Sina Miran, *Student Member, IEEE,* Piya Pal, *Member, IEEE,* Behtash Babadi, *Member, IEEE,* and Min Wu, *Fellow, IEEE*

Abstract—We consider the problem of estimating the parameters of a linear autoregressive model with sub-Gaussian innovations from a limited sequence of consecutive observations. Assuming that the parameters are compressible, we analyze the performance of the ℓ_1-regularized least squares as well as a greedy estimator of the parameters and characterize the sampling trade-offs required for stable recovery in the non-asymptotic regime. Our results extend those of compressed sensing for linear models where the covariates are i.i.d. and independent of the observation history to autoregressive processes with highly interdependent covariates. We also derive sufficient conditions on the sparsity level that guarantee the minimax optimality of the ℓ_1-regularized least squares estimate. Applying these techniques to simulated data as well as real-world datasets from crude oil prices and traffic speed data confirm our predicted theoretical performance gains in terms of estimation accuracy and model selection.

Index Terms—linear autoregressive processes, sparse estimation, compressive sensing, sampling.

I. INTRODUCTION

Autoregressive (AR) models are among the most fundamental tools in analyzing time series. Applications include financial time series analysis [2] and traffic modeling [3]–[8]. Due to their well-known approximation property, these models are commonly used to represent stationary processes in a parametric fashion and thereby preserve the underlying structure of these processes [9]. In order to leverage the approximation property of AR models, often times parameter sets of very large order are required[10]. For instance, any autoregressive moving average (ARMA) process can be represented by an AR process of infinite order. Statistical inference using these models is usually performed through fitting a long-order AR model to the data, which can be viewed as a truncation of the infinite-order representation [11]–[14]. In general, the ubiquitous long-range dependencies in real-world time series, such as financial data, results in AR model fits with large orders [2].

In various applications of interest, the AR parameters fit to the data exhibit sparsity, that is, only a small number of the parameters are non-zero. Examples include autoregressive communication channel models, quasi-oscillatory data tuned around specific frequencies and financial time series [8], [15],

The authors are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: kaazemi@umd.edu; smiran@umd.edu; ppal@umd.edu; behtash@umd.edu; minwu@umd.edu).
Corresponding author: Behtash Babadi (e-mail: behtash@umd.edu).

[16]. The non-zero AR parameters in these models correspond to significant time lags at which the underlying dynamics operate. Traditional model selection procedures based on criteria such as Final Prediction Error (FPE) [17], Akaike Information Criterion (AIC) [18] and Bayesian Information Criterion (BIC) [19] are not efficient in order selection of sparse AR processes, as the ambient dimension of the parameters can be much larger than the sparsity. Also, these criteria pertain to the asymptotic regimes and their finite sample behavior is not well understood.

In recent years, the theory of compressed (CS) has become the standard framework for measuring and estimating sparse statistical models [20]–[22]. The theoretical guarantees of CS imply that when the number of incoherent measurements are roughly proportional to the sparsity level, then stable recovery of these models is possible. A key underlying assumption in most of the existing theoretical analyses of linear models is the independence and identical distribution (i.i.d.) of the covariates. Exceptions include measurement constructions based on correlated designs, and Toeplitz matrices where the design is extrinsic, fixed in advance and is independent of the underlying sparse signal [23], [24] Such assumptions do not hold for AR processes, as the intrinsic history of the process plays the role of the covariates. Hence the underlying interdependence in the model hinders a straightforward application of existing CS results to AR modeling.

In this paper, we address this issue by extending the analyses of CS theory to the estimation of compressible AR processes, and relaxing the assumptions of i.i.d. covariates. In particular, we will consider an AR process with sparse parameters and will analyze the performance of a LASSO-type estimator (corresponding to ℓ_1-regularized Least Squares (LS) estimation) as well as a greedy solution. We will present theoretical guarantees that extend those of CS theory and characterize fundamental trade-offs between the number of measurements, compressibility, and estimation error of AR processes in the non-asymptotic regime. Our results reveal that when the number of measurements scale *sub-linearly* with the product of the ambient dimension and the sparsity level, then stable recovery of the underlying models is possible, even though the covariates solely depend on the history of the process. This is a significant improvement over existing results for stable AR estimation which require the number of measurements to scale *quadratically* with the ambient dimension [25]. We further establish sufficient conditions on the sparsity level which result in the minimax optimality of the ℓ_1-regularized LS estimator. Finally, we provide simulation results as well as application to

oil price and traffic data which reveal that the sparse estimates significantly outperform traditional techniques such as the Yule-Walker based estimators [26].

The rest of the paper is organized as follows. In Section II, we will introduce the notations and problem formulation. In section III, we will state the main problem of interest and the estimation methods for the parameters of an AR process. We will also present the main theoretical results of this paper on robust estimation of AR parameters. In section III-D we establish the minimax optimality of the ℓ_1-regularized LS estimator. Section IV includes our simulation results on simulated data as well as the real-world financial and traffic data.

II. NOTATIONS AND PROBLEM FORMULATION

Throughout the paper we will use the following notations. We will use the notation x_i^j to denote the vector $[x_i, \cdots, x_j]^T$. We will denote the estimated values by $\widehat{(.)}$ and the biased estimates with the superscript $(.)^b$. Throughout the proofs, c_i's express absolute constant which may change from line to line where there is no ambiguity. By c_η we mean an absolute constant which only depends on a positive constant η.

Consider an AR(p) process defined by

$$x_k = \theta_1 x_{k-1} + \theta_2 x_{k-2} + \cdots + \theta_p x_{k-p} + w_k = \boldsymbol{\theta}^T x_{k-p}^{k-1} + w_k, \tag{1}$$

where $\{w_k\}_{k=-\infty}^{\infty}$ is an i.i.d sub-Gaussian innovation sequence with zero mean and variance σ_w^2. This process can be considered as the output of an LTI system with transfer function

$$H(z) = \frac{\sigma_w^2}{1 - \sum_{\ell=1}^{p} \theta_\ell z^{-\ell}}. \tag{2}$$

Throughout the paper we will assume $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$ to enforce the stability of the filter. We will refer to this assumption as *the sufficient stability assumption*, since an AR process with poles within the unit circle does not necessarily satisfy $\|\boldsymbol{\theta}\|_1 < 1$. However, beyond second-order AR processes, it is not straightforward to state the stability of the process in terms of its parameters in a closed algebraic form, which in turn makes both the analysis and optimization procedures intractable. The AR(p) process given by $\{x_k\}_{k=-\infty}^{\infty}$ in (1) is stationary in the strict sense. Also by (2) the power spectral density of the process equals

$$S(\omega) = \frac{\sigma_w^2}{|1 - \sum_{\ell=1}^{p} \theta_\ell e^{-j\ell\omega}|^2}. \tag{3}$$

The sufficient stability assumption implies boundedness of the spectral spread of the process defined as

$$\rho = \sup_\omega S(\omega) \Big/ \inf_\omega S(\omega).$$

We will discuss how this assumption can be further relaxed in Appendix A-B. The spectral spread of stationary processes in general is a measure of how quickly the process reaches its ergodic state [25]. An important property that we will use later in this paper is that the spectral spread is an upper bound on the eigenvalue spread of the covariance matrix of the process of arbitrary size [27].

We will also assume that the parameter vector $\boldsymbol{\theta}$ is compressible (to be defined more precisely later), and can be well approximated by an s-sparse vector where $s \ll p$. We observe n consecutive snapshots of length p (a total of $n + p - 1$ samples) from this process given by $\{x_k\}_{k=-p+1}^n$ and aim to estimate $\boldsymbol{\theta}$ by exploiting its sparsity; to this end, we aim at addressing the following questions:

- Are the conventional LASSO-type and greedy techniques suitable for estimating $\boldsymbol{\theta}$?
- What are the sufficient conditions on n in terms of p and s, to guarantee stable recovery?
- Given these sufficient conditions, how do these estimators perform compared to conventional AR estimation techniques?

Traditionally, the Yule-Walker (YW) equations or least squares formulations are used to fit AR models. Since these methods do not utilize the sparse structure of the parameters, they usually require $n \gg p$ samples in order to achieve satisfactory performance. The YW equations can be expressed as

$$R\boldsymbol{\theta} = r_{-p}^{-1}, \quad r_0 = \boldsymbol{\theta}^T r_{-p}^{-1} + \sigma_w^2, \tag{4}$$

where $R := R_{p \times p} = \mathbb{E}[x_1^p x_1^{pT}]$ is the $p \times p$ covariance matrix of the process and $r_k = \mathbb{E}[x_i x_{i+k}]$ is the autocorrelation of the process at lag k. The covariance matrix R and autocorrelation vector r_{-p}^{-1} are typically replaced by their sample counterparts. Estimation of the AR(p) parameters from the YW equations can be efficiently carried out using the Burg's method [28]. Other estimation techniques include LS regression and maximum likelihood (ML) estimation. In this paper, we will consider the Burg's method and LS solutions as comparison benchmarks. When n is comparable to p, these two methods are known to exhibit substantial performance differences [29].

When fitted to the real-world data, the parameter vector $\boldsymbol{\theta}$ usually exhibits a degree of sparsity. That is, only certain lags in the history have a significant contribution in determining the statistics of the process. These lags can be thought of as the intrinsic delays in the underlying dynamics. To be more precise, for a sparsity level $s < p$, we denote by $S \subset \{1, 2, \cdots, p\}$ the support of the s largest elements of $\boldsymbol{\theta}$ in absolute value, and by $\boldsymbol{\theta}_S$ the best s-term approximation to $\boldsymbol{\theta}$. We also define

$$\sigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1 \tag{5}$$

and

$$\varsigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_2 \tag{6}$$

which capture the compressibility of the parameter vector $\boldsymbol{\theta}$ in the ℓ_1 and ℓ_2 sense, respectively. Note that by definition $\varsigma_s(\boldsymbol{\theta}) \leq \sigma_s(\boldsymbol{\theta})$. For a fixed $\xi \in (0,1)$, we say that $\boldsymbol{\theta}$ is (s, ξ)-*compressible* if $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$ [30] and $(s, \xi, 2)$-*compressible* if $\varsigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$. Note that $(s, \xi, 2)$-*compressibility* is a weaker condition than (s, ξ)-*compressibility* and when $\xi = 0$, the parameter vector $\boldsymbol{\theta}$ is exactly s-sparse.

Finally, in this paper, we are concerned with the compressed sensing regime where $n \ll p$, i.e., the observed data has

a much smaller length than the ambient dimension of the parameter vector. The main estimation problem of this paper can be summarized as follows: *given observations x^n_{-p+1} from an AR process with sub-Gaussian innovations and bounded spectral spread, the goal is to estimate the unknown p-dimensional $(s, \xi, 2)$-compressible AR parameter vector $\boldsymbol{\theta}$ in a stable fashion (where the estimation error is controlled) when $n \ll p$.*

III. THEORETICAL RESULTS

In this section, we will describe the estimation procedures and present the main theoretical results of this paper.

A. ℓ_1-regularized least squares estimation

Given the sequence of observations x^n_{-p+1} and an estimate $\widehat{\boldsymbol{\theta}}$, the normalized estimation error can be expressed as:

$$\mathfrak{L}\left(\widehat{\boldsymbol{\theta}}\right) := \frac{1}{n}\left\|x^n_1 - X\widehat{\boldsymbol{\theta}}\right\|^2_2, \qquad (7)$$

where

$$X = \begin{bmatrix} x_{n-1} & x_{n-2} & \cdots & x_{n-p} \\ x_{n-2} & x_{n-3} & \cdots & x_{n-p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_0 & x_{-1} & \cdots & x_{-p+1} \end{bmatrix}. \qquad (8)$$

Note that the matrix of covariates $\mathbf{X}$ is Toeplitz with highly interdependent elements. The LS solution is thus given by:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{LS}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathfrak{L}(\boldsymbol{\theta}), \qquad (9)$$

where

$$\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^p | \|\boldsymbol{\theta}\|_1 < 1 - \eta\}$$

is the convex feasible region for which the stability of the process is guaranteed. In the regime of interest when $n \ll p$, the LS estimator is ill-posed and is typically regularized with a smooth norm. In order to capture the compressibility of the parameters, we consider the ℓ_1-regularized LS estimator:

$$\widehat{\boldsymbol{\theta}}_{\ell_1} := \arg\min_{\boldsymbol{\theta} \in \Theta} \mathfrak{L}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1, \qquad (10)$$

where $\gamma_n > 0$ is a regularization parameter. It is easy to verify that the objective function and constraints in Eq. (10) are convex in $\boldsymbol{\theta}$ and hence $\widehat{\boldsymbol{\theta}}_{\ell_1}$ can be obtained using standard numerical solvers. Note that the solution to (10) might not be unique. However, we will provide error bounds that hold for all possible solutions of (10), with high probability.

Consistency of the estimator given by (9) was shown in [2] when $n \to \infty$ for Gaussian innovations. In the case of Gaussian innovations the LS estimates correspond to conditional ML estimation and are asymptotically unbiased under mild conditions, and with p fixed, the solution converges to the true parameter vector as $n \to \infty$. However, when p is allowed to scale with n, the convergence rate is not known. For fixed p, the estimation error is of the order $\mathcal{O}(p/n)$ in general [24]. This makes the analysis of LS estimators, and in general regularized M-estimators, very challenging [31]. Nevertheless, such an analysis has significant practical implications, as it will

reveal sufficient conditions on n with respect to p as well as a criterion to choose γ_n, which result in stable estimation of $\boldsymbol{\theta}$. Throughout our analysis, we fix the ambient dimension p and aim at deriving sufficient conditions on $n \ll p$ resulting in stable estimation. In many applications of interest, the data correlations are exponentially decaying and negligible beyond a certain lag, and hence for large enough p, autoregressive models fit the data very well in the prediction error sense.

Recall that, the Yule-Walker solution is given by

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw}} := \arg\min_{\boldsymbol{\theta} \in \Theta} \mathfrak{J}(\boldsymbol{\theta}) = \widehat{R}^{-1}\widehat{r}^{-1}_{-p}, \qquad (11)$$

where

$$\mathfrak{J}(\boldsymbol{\theta}) := \|\widehat{R}\boldsymbol{\theta} - \widehat{r}^{-1}_{-p}\|_2.$$

We further consider two other sparse estimators for $\boldsymbol{\theta}$ by penalizing the Yule-Walker equations. The ℓ_1-regularized Yule-Walker estimator is defined as:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{2,1}} := \arg\min_{\boldsymbol{\theta} \in \Theta} \mathfrak{J}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1, \qquad (12)$$

where $\gamma_n > 0$ is a regularization parameter. Similarly, using the robust statistics instead of the Gaussian statistics, the estimation error can be re-defined as:

$$\mathfrak{J}_1(\boldsymbol{\theta}) := \|\widehat{R}\boldsymbol{\theta} - \widehat{r}^{-1}_{-p}\|_1,$$

we define the ℓ_1-regularized estimates as

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{1,1}} := \arg\min_{\boldsymbol{\theta} \in \Theta} \mathfrak{J}_1(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1. \qquad (13)$$

B. Greedy estimation

Although there exist fast solvers for the convex problems of the type given by (10), (12) and (13), these algorithms are polynomial time in n and p, and may not scale well with the dimension of data. This motivates us to consider greedy solutions for the estimation of $\boldsymbol{\theta}$. In particular, we will consider and study the performance of a generalized Orthogonal Matching Pursuit (OMP) algorithm [32], [33]. A flowchart of this algorithm is given in Table I for completeness. At each iteration, a new component of $\boldsymbol{\theta}$ for which the gradient of the error metric $\mathfrak{f}(\boldsymbol{\theta})$ is the largest in absolute value is chosen and added to the current support. The algorithm proceeds for a total of $s^\star = \mathcal{O}(s \log s)$ steps, resulting in an

Input: $\mathfrak{f}(\boldsymbol{\theta}), s^\star$
Output: $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}} = \widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{OMP}}$
Initialization: $\begin{cases} \text{Start with the index set } S^{(0)} = \emptyset \\ \text{and the initial estimate } \widehat{\boldsymbol{\theta}}^{(0)}_{\mathsf{OMP}} = 0 \end{cases}$
for $k = 1, 2, \cdots, s^\star$
$\quad j = \arg\max_i \left|\left(\nabla\mathfrak{f}\left(\widehat{\boldsymbol{\theta}}^{(k-1)}_{\mathsf{OMP}}\right)\right)_i\right|$
$\quad S^{(k)} = S^{(k-1)} \cup \{j\}$
$\quad \widehat{\boldsymbol{\theta}}^{(k)}_{\mathsf{OMP}} = \arg\min_{\mathrm{supp}(x) \subset S^{(k)}} \mathfrak{f}(\boldsymbol{\theta})$
end

TABLE I: Generalized Orthogonal Matching Pursuit (OMP)

estimate with $s^\star$ components. When the error metric $\mathfrak{L}(\boldsymbol{\theta})$ is chosen, the generalized OMP corresponds to the original OMP algorithm. For the choice of the YW error metric $\mathfrak{J}(\boldsymbol{\theta})$, we denote the resulting greedy algorithm by ywOMP.

C. Estimation performance guarantees

The main theoretical result regarding the estimation performance of the ℓ_1-regularized LS estimator is given by the following theorem:

Theorem 1. *If $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, there exist positive constants d_1, d_2, d_3 and d_4 such that for $n > d_1 s p^{2/3}$ and a choice of regularization parameter $\gamma_n = d_2 \sqrt{\frac{\log p}{n}}$, any solution $\widehat{\boldsymbol{\theta}}_{\ell_1}$ to (10) satisfies the bound*

$$\left\| \widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta} \right\|_2 \leq d_3 \sqrt{\frac{s \log p}{n}} + \sqrt{d_3 \sigma_s(\boldsymbol{\theta})} \sqrt[4]{\frac{\log p}{n}}, \quad (14)$$

with probability greater than $1 - \mathcal{O}(\frac{1}{n^{d_4}})$. The constants are only functions of σ_{w}^2 and η and are explicitly given in the proof.

Similarly, the following theorem characterizes the estimation performance bounds for the OMP algorithm:

Theorem 2. *If $\boldsymbol{\theta}$ is $(s, \xi, 2)$-compressible for some $\xi < 1/2$, there exist positive constants d_1', d_2', d_3' and d_4' such that for $n > d_1' s p^{2/3} \log s$, the OMP estimate satisfies the bound*

$$\left\| \widehat{\boldsymbol{\theta}}_{\mathsf{OMP}} - \boldsymbol{\theta} \right\|_2 \leq d_2' \sqrt{\frac{s \log s \log p}{n}} + d_3' \frac{\log s}{s^{\frac{1}{\xi} - 2}} \quad (15)$$

after $s^\star = \frac{s}{\eta^2} \log \frac{5s}{\eta^2}$ iterations with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d_4'}}\right)$. The constants are only functions of σ_{w}^2 and η and are explicitly given in the proof.

The results of Theorems 1 and 2 suggest that under suitable compressibility assumptions on the AR parameters, one can estimate the parameters reliably using the ℓ_1-regularized LS and OMP estimators with much fewer measurements compared to those required by the Yule-Walker/LS based methods. To illustrate the significance of these results further, several remarks are in order:

Remark 1. Comparing the sufficient condition $n = \mathcal{O}(sp^{2/3})$ of Theorem 1 with those of [20]–[22] for linear models with i.i.d. measurement matrices given by $n = \mathcal{O}(s \log p)$ reveals that a loss of order $\mathcal{O}(p^{2/3})$ is incurred, although both conditions require $n \ll p$. However, this loss seems to be natural as it stems from two major differences of our setting as compared to traditional CS: first, the sample space characterizing the covariates x_{-p+1}^n is $(n + p - 1)$-dimensional, whereas in traditional CS the sample space of the measurement matrix is np-dimensional. Second, each row of the measurement matrix X highly depends on the entire observation sequence x_1^n, whereas in traditional CS, each row of the measurement matrix is only related to the corresponding measurement. Hence, the aforementioned loss can be viewed as the price of self-averaging of the process accounting for the low-dimensional nature of the covariate sample space and the high inter-dependence of the covariates to the observation

sequence. The dominant loss of $\mathcal{O}(p^{2/3})$ does not seem to be significantly improvable, as AR processes are known for slow convergence to their ergodic state.

Remark 2. On the positive side, compared to existing guarantees for LS estimation techniques for AR processes which typically require $\mathcal{O}(p^2)$ measurements [25], our result predicts a significant saving of the order $p^{4/3} s^{-1}$. It can also be shown that the estimation error for the LS method in general scales as $\sqrt{p/n}$ [24] which is not desirable when $n \ll p$. Our result, however, guarantees a much smaller error rate of the order $\sqrt{s \log p/n}$. Also, the sufficiency conditions of Theorem 2 require high compressibility of the parameter vector $\boldsymbol{\theta}$ ($\xi < 1/2$), whereas Theorem 1 does not impose any extra restrictions on $\xi \in (0, 1)$. Intuitively speaking, these two comparisons reveal the trade-off between computational complexity and measurement/compressibility requirements for convex optimization vs. greedy techniques, which are well-known for linear models [34].

Remark 3. The condition $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$ in Theorem 1 is not restricting for the processes of interest in this paper. This is due to the fact that the boundedness assumption on the spectral spread implies an exponential decay of the parameters (See Lemma 1 of [25]). Finally, the constants d_1, d_1' are increasing with respect to the spectral spread of the process ρ. Intuitively speaking, the closer the roots of the filter given by (2) get to the unit circle (corresponding to larger ρ and smaller η), the slower the convergence of the process will be to its ergodic state, and hence more measurements are required. A similar dependence to the spectral spread has appeared in the results of [25] for ℓ_2-regularized least squares estimation of AR processes.

Remark 4. The main ingredient in the proofs of Theorems 1 and 2 is to establish the restricted eigenvalue (RE) condition introduced in [35] for the covariate matrix X. We will show that if the bounded spectral spread condition holds, then with $n = \mathcal{O}(sp^{2/3})$ the sample covariance matrix is sharply concentrated around the true covariance matrix and hence the RE condition can be guaranteed. All constants appearing in Theorems 1 and 2 are explicitly given in Appendix A-B. As a typical numerical example, for $\eta = 0.5$ and $\sigma_w^2 = 0.01$, the constants of Theorem 1 can be chosen as $d_1 \approx 1000, d_2 \approx 0.5 - 1, d_3 \approx 10$, and $d_4 \approx 4$. The full proofs are given in Appendix A-B.

D. Minimax optimality

In this section, we establish the minimax optimality of the ℓ_1-regularized LS estimator for AR processes with sparse parameters. To this end, we will focus on the class $\mathcal{H}$ of stationary processes which admit an AR(p) representation with s-sparse parameter $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$. The theoretical results of this section are inspired by the enlightening results of [25] on non-asymptotic order selection via ℓ_2-regularized LS estimation in the absence of sparsity, and extend them by studying the ℓ_1-regularized LS estimator of (10).

We define the maximal *estimation* risk over $\mathcal{H}$ to be

$$\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}) := \sup_{\mathcal{H}} \left(\mathbb{E} \left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right] \right)^{1/2}. \quad (16)$$

The minimax estimator is the one minimizing the maximal estimation risk, i.e.,

$$\widehat{\boldsymbol{\theta}}_{\text{minimax}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \quad \mathcal{R}_{\text{est}}(\widehat{\boldsymbol{\theta}}). \tag{17}$$

Minimax estimator $\widehat{\boldsymbol{\theta}}_{\text{minimax}}$, in general, cannot be constructed explicitly [25], and the common practice in non-parametric estimation is to construct an estimator $\widehat{\boldsymbol{\theta}}$ which is *order optimal* as compared to the minimax estimator:

$$\mathcal{R}_{\text{est}}(\widehat{\boldsymbol{\theta}}) \leq L\mathcal{R}_{\text{est}}(\widehat{\boldsymbol{\theta}}_{\text{minimax}}). \tag{18}$$

with $L \geq 1$ being a constant. One can also define the minimax *prediction* risk by the maximal prediction error over all possible realizations of the process:

$$\mathcal{R}_{\text{pre}}^2(\widehat{\boldsymbol{\theta}}) := \sup_{\mathcal{H}} \mathbb{E}\left[\left(x_k - \widehat{\boldsymbol{\theta}}'x_{k-p}^{k-1}\right)^2\right]. \tag{19}$$

In [25], it is shown that an ℓ_2-regularized LS estimator with an order $p^\star = \mathcal{O}(\log n)$ is minimax optimal. This order pertains to the denoising regime where $n \gg p$. Hence, in order to capture long order lags of the process, one requires a sample size exponentially large in p, which may make the estimation problem computationally infeasible. For instance, consider a 2-sparse parameter with only θ_1 and θ_p being non-zero. Then, in order to achieve minimax optimality, $n = \mathcal{O}(2^p)$ measurements are required. In contrast, in the compressive regime where $s, n \ll p$, the goal, instead of selecting p, is to find conditions on the sparsity level s, so that for a given n and large enough p, the ℓ_1-regularized estimator is minimax optimal without explicit knowledge of the value of s (See for example, [36]).

In the following proposition, we establish the minimax optimality of the ℓ_1-regularized estimator over the class of sparse AR processes with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

Proposition 1. *Let x_1^n be samples of an AR process with s-sparse parameters satisfying $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta$ and $s \leq \min\left\{\frac{1-\eta}{\sqrt{8\pi\eta}}\sqrt{\frac{n}{\log p}}, \frac{n}{d_1 p^{2/3}}\right\}$. Then, we have:*

$$\mathcal{R}_{\text{est}}(\widehat{\boldsymbol{\theta}}_{\ell_1}) \leq L\mathcal{R}_{\text{est}}(\widehat{\boldsymbol{\theta}}_{\text{minimax}}).$$

where L is a constant and is only a function of η and σ_{w}^2 and is explicitly given in the proof.

Remark 5. Proposition 1 implies that ℓ_1-regularized LS is minimax optimal in estimating the s-sparse parameter vector $\boldsymbol{\theta}$, for small enough s. The proof of the Proposition 1 is given in Appendix A-D. This result can be extended to compressible $\boldsymbol{\theta}$ in a natural way with a bit more work, but we only present the proof for the case of s-sparse $\boldsymbol{\theta}$ for brevity. We also state the following proposition on the prediction performance of the ℓ_1-regularized LS estimator:

Proposition 2. *Let x_{-p+1}^n be samples of an AR process with s-sparse parameters and Gaussian innovations, then there exists a positive constant d_5 such that for large enough n, p and s satisfying $n > d_1 sp^{2/3}$, we have:*

$$\mathcal{R}_{\text{pre}}^2(\widehat{\boldsymbol{\theta}}_{\ell_1}) \leq d_5 \frac{s\log p}{n} + \sigma_{\text{w}}^2. \tag{20}$$

It can be readily observed that for $n \gg s$ the prediction error variance is very close to the variance of the innovations. The proof is similar to Theorem 3 of [25] and is skipped in this paper for brevity.

IV. APPLICATION TO SIMULATED AND REAL DATA

In this section, we study and compare the performance of Yule-Walker based estimation methods with those of the ℓ_1-regularized and greedy estimators given in section III. These methods are applied to simulated data as well as real data from crude oil price and traffic speed.

A. Simulation studies

In order to simulate an AR process, we filtered a Gaussian white noise process using an IIR filter with sparse parameters. Figure 1 shows a typical sample path of the simulated AR process used in our analysis. For the parameter vector $\boldsymbol{\theta}$, we chose a length of $p = 300$, and employed $n = 1500$ generated samples of the corresponding process for estimation. The parameter vector $\boldsymbol{\theta}$ is of sparsity level $s = 3$ and $\eta = 1 - \|\boldsymbol{\theta}\|_1 = 0.5$. A value of $\gamma_n = 0.1$ is used, which is slightly tuned around the theoretical estimate given by Theorem 1. The order of the process is assumed to be known. We compare the performance of seven estimators: 1) $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ using LS, 2) $\widehat{\boldsymbol{\theta}}_{\text{yw}}$ using the Yule-Walker equations, 3) $\widehat{\boldsymbol{\theta}}_{\ell_1}$ from ℓ_1-regularized LS, 4) $\widehat{\boldsymbol{\theta}}_{\text{OMP}}$ using OMP, 5) $\widehat{\boldsymbol{\theta}}_{\text{yw},\ell_{2,1}}$ using Eq. (12), 6) $\widehat{\boldsymbol{\theta}}_{\text{yw},\ell_{1,1}}$ using Eq. (13), and 7) $\widehat{\boldsymbol{\theta}}_{\text{ywOMP}}$ using the cost function $\mathfrak{J}(\boldsymbol{\theta})$ in the generalized OMP. Figure 2 shows the estimated parameter vectors using these algorithms. It can be visually observed that ℓ_1-regularized and greedy estimators (shown in purple) significantly outperform the Yule-Walker-based estimates (shown in orange).

In order to quantify the latter observation precisely, we repeated the same experiment for $p = 300, s = 3$ and $10 \leq n \leq 10^5$. A comparison of the normalized MSE of the estimators vs. n is shown in Figure 3. As it can be inferred from Figure 3, in the region where n is comparable to p (shaded in light purple), the sparse estimators have a systematic performance gain over the Yule-Walker based estimates, with the ℓ_1-regularized LS and ywOMP estimates outperforming the rest.

The MSE comparison in Figure 3 requires one to know the true parameters. In practice, the true parameters are not available for comparison purposes. In order to quantify the performance gain of these methods, we use statistical tests to assess the goodness-of-fit of the estimates. The common chi-square type statistical tests, such as the F-test, are useful when the hypothesized distribution to be tested against is discrete or categorical. For our problem setup with sub-Gaussian innovations, we will use a number of statistical
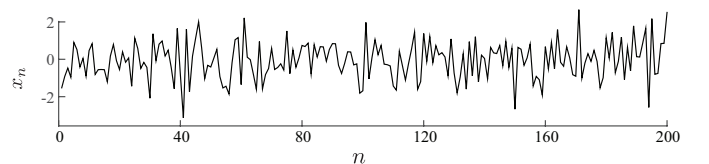

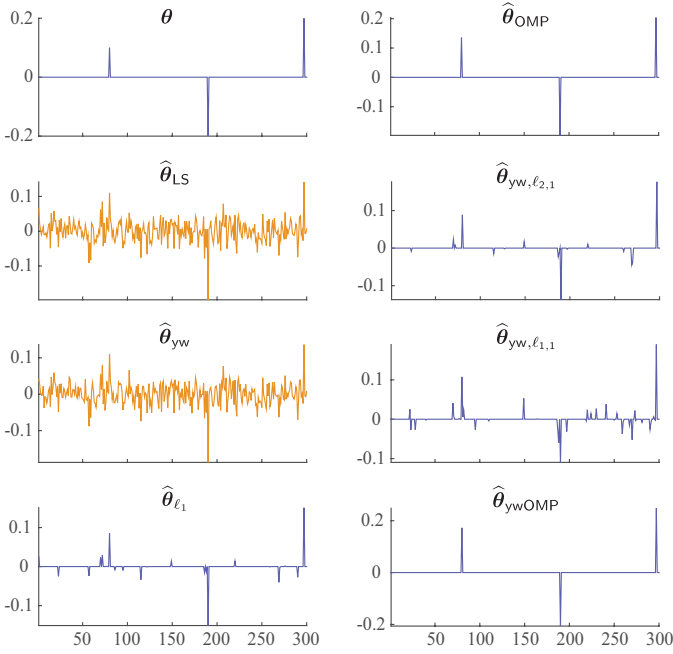Fig. 1: Samples of the simulated AR process.

Fig. 2: Estimates of $\boldsymbol{\theta}$ for $n = 1500$, $p = 300$, and $s = 3$ (These results are best viewed in the color version).
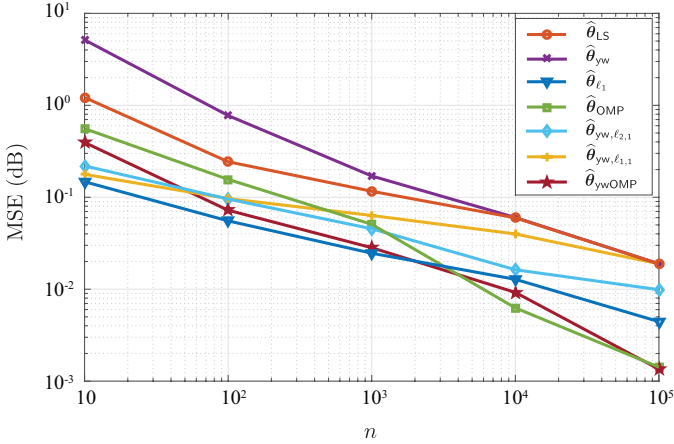

Fig. 3: MSE comparison of the estimators vs. the number of measurements n.

tests appropriate for AR processes, namely, the Kolmogorov-Smirnov (KS) test, the Cramér-von Mises (CvM) criterion, the spectral Cramér-von Mises (SCvM) test and the Anderson-Darling (AD) [37]–[39]. A summary of these tests is given in Appendix B. Table II summarizes the test statistics for different estimation methods. These tests are based on the known results on limiting distributions of error residuals. As noted from Table II, our simulations suggest that the OMP estimate achieves the best test statistics for the CvM, AD and KS tests, whereas the ℓ_1-regularized estimate achieves the best SCvM statistic.

B. Application to the analysis of crude oil prices

In this and the following subsection, we consider applications with real-world data. As for the first application, we apply the sparse AR estimation techniques to analyze the crude oil price of the Cushing, OK WTI Spot Price FOB dataset

TABLE II: Goodness-of-fit tests for the simulated data

Estimate / Test	CvM	AD	KS	SCvM
$\boldsymbol{\theta}$	0.31	1.54	0.031	0.009
$\widehat{\boldsymbol{\theta}}_{LS}$	0.68	5.12	0.037	0.017
$\widehat{\boldsymbol{\theta}}_{yw}$	0.65	4.87	0.034	0.025
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	0.34	1.72	0.030	0.009
$\widehat{\boldsymbol{\theta}}_{OMP}$	0.29	1.45	0.028	0.009
$\widehat{\boldsymbol{\theta}}_{yw,\ell_{2,1}}$	0.35	1.80	0.032	0.009
$\widehat{\boldsymbol{\theta}}_{yw,\ell_{1,1}}$	0.42	2.33	0.040	0.008
$\widehat{\boldsymbol{\theta}}_{ywOMP}$	0.29	1.46	0.030	0.009

[40]. This dataset consists of 7429 daily values of oil prices in dollars per barrel. In order to avoid outliers, usually the dataset is filtered with a moving average filter of high order. We have skipped this procedure by visual inspection of the data and selecting $n = 4000$ samples free of outliers. Such financial data sets are known for their non-stationarity and long order history dependence. In order to remove the deterministic trends in the data, one-step or two-step time differencing is typically used. We refer to [8] for a full discussion of this detrending method. We have used a first-order time differencing which resulted in a sufficient detrending of the data. Figure 4 shows the data used in our analysis. We have chosen $p = 150$ by inspection. The histogram of first-order differences as well the estimates are shown in Figure 5.
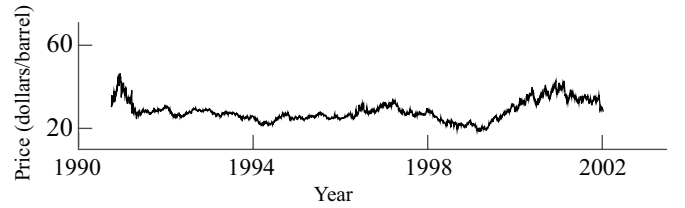

Fig. 4: A sample segment of the Cushing, OK WTI Spot Price FOB data.

A visual inspection of the estimates in Figure 5 shows that the ℓ_1-regularized LS ($\widehat{\boldsymbol{\theta}}_{\ell_1}$) and OMP ($\widehat{\boldsymbol{\theta}}_{OMP}$) estimates consistently select specific time lags in the AR parameters, whereas the Yule-Walker and LS estimates seemingly overfit the data by populating the entire parameter space. In order to perform goodness-of-fit tests, we use an even/odd two-fold cross-validation. Table III shows the corresponding test statistics, which reveal that indeed the ℓ_1-regularized and OMP estimates outperform the traditional estimation techniques.

TABLE III: Goodness-of-fit tests for the crude oil price data

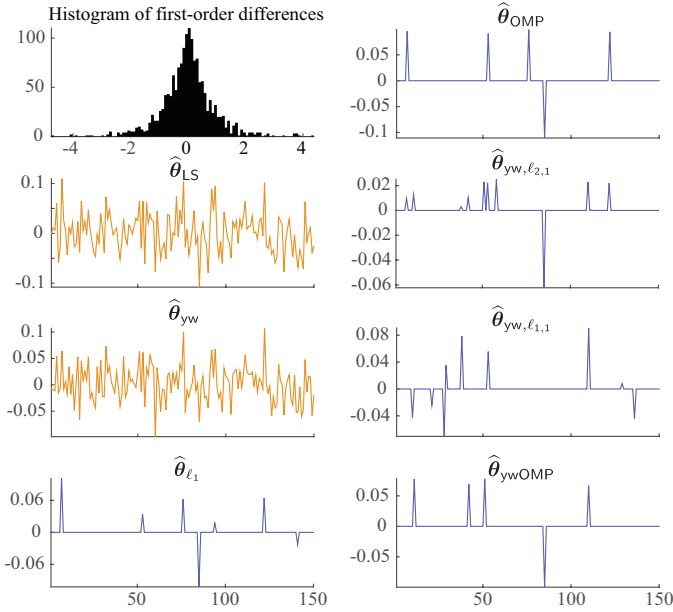| Estimate / Test	CvM	AD	KS	SCvM
$\widehat{\boldsymbol{\theta}}_{LS}$	0.88	5.55	0.055	0.046
$\widehat{\boldsymbol{\theta}}_{yw}$	0.58	3.60	0.043	0.037
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	0.27	1.33	0.031	0.020
$\widehat{\boldsymbol{\theta}}_{OMP}$	0.22	1.18	0.025	0.022
$\widehat{\boldsymbol{\theta}}_{yw,\ell_{2,1}}$	0.28	1.40	0.027	0.021
$\widehat{\boldsymbol{\theta}}_{yw,\ell_{1,1}}$	0.24	1.26	0.027	0.022
$\widehat{\boldsymbol{\theta}}_{ywOMP}$	0.23	1.18	0.026	0.022

Fig. 5: Estimates of $\boldsymbol{\theta}$ for the second-order differences of the oil price data.

C. Application to the analysis of traffic data

Our second real data application concerns traffic speed data. The data used in our simulations is the INRIX ® speed data for I-495 Maryland inner loop freeway (clockwise) between US-1/Baltimore Ave/Exit 25 and Greenbelt Metro Dr/Exit 24 from 1 Jul, 2015 to 31 Oct, 2015 [41], [42]. The reference speed of 65 mph is reported. Given the huge length of the data and its high variability, the following pre-processing was made on the original data:

1) The data was downsampled by a factor of 4 and averaged by the hour in order to reduce its daily variability, that is each lag corresponds to one hour.
2) The logarithm of speed was used for analysis and the mean was subtracted. This reduces the high variability of speed due to rush hours and lower traffic during weekends and holidays.
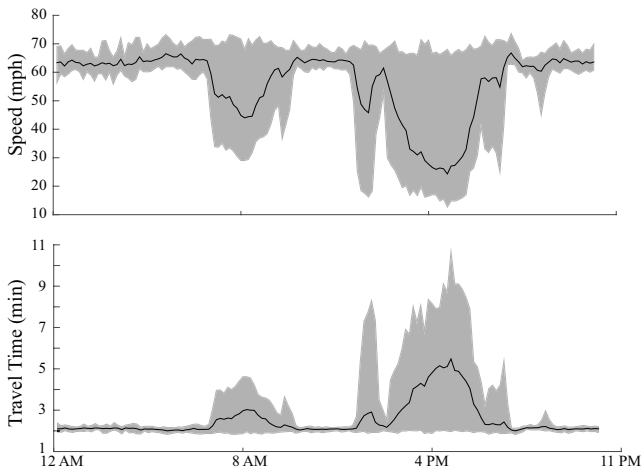

Fig. 6: A sample of the speed and travel time data for I-495.

Figure 6 shows a typical average weekly speed and travel time in this dataset and the corresponding 25-75-th percentiles. As can be seen the data shows high variability around the rush hours of 8am and 4pm. In our analysis, we used the first half of the data ($n = 1500$) for fitting, from which the AR parameters and the distribution and variance of the innovations were estimated. The statistical tests were designed based on the estimated distributions, and the statistics were computed accordingly using the second half of the data. We selected an order of $p = 200$ by inspection and noting that the data seems to have a periodicity of order 170 samples.
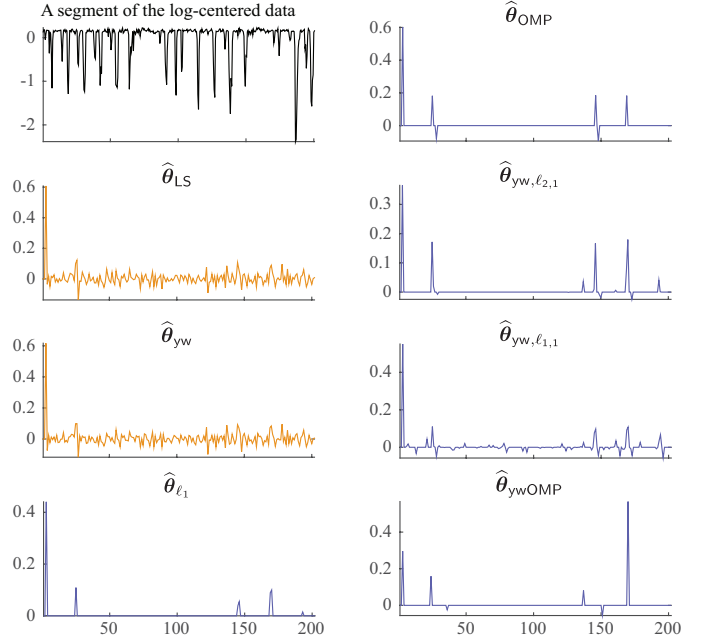

Fig. 7: Estimates of $\boldsymbol{\theta}$ for the traffic speed data.

Figure 7 shows part of the data used in our analysis as well as the estimated parameters. The ℓ_1-regularized LS ($\widehat{\boldsymbol{\theta}}_{\ell_1}$) and OMP ($\widehat{\boldsymbol{\theta}}_{\text{OMP}}$) are consistent in selecting the same components of $\boldsymbol{\theta}$. These estimators pick up two major lags around which $\boldsymbol{\theta}$ has its largest components. The first lag corresponds to about 24 hours which is mainly due to the rush hour periodicity on a daily basis. The second lag is around $150 - 170$ hours which corresponds to weekly changes in the speed due to lower traffic over the weekend. In contrast, the Yule-Walker and LS estimates do not recover these significant time lags.

TABLE IV: Goodness-of-fit tests for the traffic speed data

Test / Estimate	CvM	AD	KS	SCvM
$\widehat{\boldsymbol{\theta}}_{\text{yw}}$	0.012	0.066	0.220	0.05
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	1.4×10^{-7}	2.1×10^{-6}	6.7×10^{-4}	0.25
$\widehat{\boldsymbol{\theta}}_{\text{OMP}}$	0.017	0.082	0.220	1.49
$\widehat{\boldsymbol{\theta}}_{\text{ywOMP}}$	0.025	0.122	0.270	0.14

Statistical tests for a selected subset of the estimators are shown in Table IV. Interestingly, the ℓ_1-regularized LS estimator significantly outperforms the other estimators in three of the tests. The Yule-Walker estimator, however, achieves the best SCvM test statistic.

V. CONCLUSIONS

In this paper, we have investigated sufficient sampling requirements for stable estimation of AR models in the non-asymptotic regime using the ℓ_1-regularized LS and greedy estimation (OMP) techniques. We have further established the minimax optimality of the ℓ_1-regularized LS estimator. Compared to the existing literature, our results provide several major contributions. First, to the best of our knowledge, our sufficient sampling bounds are the first of this kind in assessing the estimation performance of sparse AR models in the non-asymptotic regime. Although ℓ_1-regularized LS and OMP are widely used in practice for AR estimation, the choices of the regularization parameter and the number of greedy iterations, respectively, are often ad-hoc. In contrast, our theoretical results prescribe analytical choices of the aforementioned parameters. In particular, our results suggest an improvement of order $\mathcal{O}(p^{4/3}s^{-1})$ over those of [25] which requires $n \sim \mathcal{O}(p^2)$ for stable AR estimation. Second, our analysis relaxes the widely-assumed hypotheses of i.i.d. covariates, which is clearly violated for AR estimation. To this end, our results extend those of [24] in establishing the RIP for i.i.d. Toeplitz matrices and [43] in establishing the RE condition for correlated Gaussian designs, to the setting or AR estimation where the both the observations and covariates are obtained by the same stochastic process. Moreover, in the linear sparsity regime where s scales linearly with p, our results suggest an improvement of order $\mathcal{O}(p^{1/3})$ to those of [24]. Third, we established the minimax optimality of the ℓ_1-regularized LS estimator for compressible AR parameters.

We further verified the validity of our theoretical results through simulation studies as well as application to real financial and traffic data. These results show that the sparse estimation methods significantly outperform the widely-used Yule-Walker based estimators in fitting AR models to the data. Although we did not theoretically analyze the performance of sparse Yule-Walker based estimators, they seem to perform on par with the ℓ_1-regularized LS and OMP estimators based on our numerical studies. Finally, our results provide a striking connection to our recent work [44], [45] in estimating sparse self-exciting discrete point process models. These models regress an observed binary spike train with respect to its history via Bernoulli or Poisson statistics, and are often used in describing spontaneous activity of sensory neurons. Our results have shown that in order to estimate a sparse history-dependence parameter vector of length p and sparsity s in a stable fashion, a spike train of length $n = \mathcal{O}(s^{2/3}p^{2/3}\log p)$ is required. This leads us to conjecture that our results will hold for a larger class of autoregressive processes, beyond those characterized by linear models.

APPENDIX A
PROOFS OF THEOREMS 1 AND 2

A. The Restricted Strong Convexity of the matrix of covariates

The first element of the proofs of both Theorems 1 and 2 is to establish the Restricted Strong Convexity (RSC) for the matrix X of covariates formed from the observed data. First, we investigate the closely related Restricted Eigenvalue (RE) condition. Let $[\lambda_{\min}(s), \lambda_{\max}(s)]$ be the smallest interval containing the singular values of $\frac{1}{n}(X_S^T X_S)$, where X_S is a sub-matrix X over an index set S of size s.

Definition 1 (Restricted Eigenvalue Condition). *A matrix X is said to satisfy the RE condition of order s if $\lambda_{\min}(s) > 0$.*

Although the RE condition only restricts $\lambda_{\min}(s)$, in the following analysis we also keep track of $\lambda_{\max}(s)$, which appears in some of the bounds. Establishing the RSC for X proceeds in a sequence of lemmas (Lemmas 1–5 culminating in Lemma 6). We first show that the RE condition holds for the true covariance of an AR process:

Lemma 1 (from [46]). *Let $R \in \mathbb{R}^{k \times k}$ be the $k \times k$ covariance matrix of a stationary process with power spectral density $S(\omega)$, and denote its maximum and minimum eigenvalues by $\phi_{\max}(k)$ and $\phi_{\min}(k)$, respectively. Then, $\phi_{\max}(k)$ is increasing in k, $\phi_{\min}(k)$ is decreasing in k, and we have*

$$\phi_{\min}(k) \downarrow \inf_\omega S(\omega), \quad \text{and} \quad \phi_{\max}(k) \uparrow \sup_\omega S(\omega). \tag{21}$$

This result gives us the following corollary:

Corollary 1 (Singular Value Spread of R). *Under the sufficient stability assumption, the singular values of the covariance R of an AR process lie in the interval $\left[\frac{\sigma_w^2}{8\pi}, \frac{\sigma_w^2}{2\pi\eta^2}\right]$.*

Proof: For an AR(p) process

$$S(\omega) = \frac{1}{2\pi}\frac{\sigma_w^2}{|1 - \sum_{\ell=1}^p \theta_\ell e^{-j\ell\omega}|^2}.$$

Combining $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$ with Lemma 1 proves the claim. ∎

Note that by Lemma 1, the result of Corollary 1 not only holds for AR processes, but also for *any* stationary process satisfying $\inf_\omega S(\omega) > 0$ and $\sup_\omega S(\omega) < \infty$, i.e., a process with finite spectral spread.

We next establish conditions for the RE condition to hold for the empirical covariance $\widehat{R}$:

Lemma 2. *If the singular values of R lie in the interval $[\lambda_{\min}, \lambda_{\max}]$, then X satisfies the RE condition of order $s_\star$ with parameters $\lambda_{\min}(s_\star) = \lambda_{\min} - ts_\star$ and $\lambda_{\max}(s_\star) = \lambda_{\max} + ts_\star$, where $t = \max_{i,j}|\widehat{R}_{ij} - R_{ij}|$.*

Proof: Let $\widehat{R} = \frac{1}{n}(X^T X)$. For every $s_\star$-sparse vector $\boldsymbol{\theta}$ we have

$$\boldsymbol{\theta}^T \widehat{R}\boldsymbol{\theta} \geq \boldsymbol{\theta}^T R\boldsymbol{\theta} - t\|\boldsymbol{\theta}\|_1^2 \geq (\lambda_{\min} - ts_\star)\|\boldsymbol{\theta}\|_2^2,$$
$$\boldsymbol{\theta}^T \widehat{R}\boldsymbol{\theta} \leq \boldsymbol{\theta}^T R\boldsymbol{\theta} + t\|\boldsymbol{\theta}\|_1^2 \leq (\lambda_{\max} + ts_\star)\|\boldsymbol{\theta}\|_2^2,$$

which proves the claim. ∎

We will next show that t can be suitably controlled with high probability. Before doing so, we state a key result of Rudzkis [47] regarding the concentration of second-order empirical sums from stationary processes:

Lemma 3. *Let x_{-p+1}^n be samples of a stationary process which satisfies*

$$x_k = \sum_{j=-\infty}^{\infty} b_{j-k}w_j, \tag{22}$$

where w_k's are i.i.d random variables with

$$\left|\mathbb{E}(|w_j|^k)\right|\leq (\tilde{c}\sigma_{\mathsf{w}})^k k!,\ \ k=2,3,\cdots, \tag{23}$$

for some constant $\tilde{c}$ and

$$\sum_{j=-\infty}^{\infty} |b_j| < \infty. \tag{24}$$

Then, the biased sample autocorrelation given by

$$\widehat{r}_k^b = \frac{1}{n+k}\sum_{i,j=1,j-i=k}^{n+k} x_i x_j$$

satisfies

$$\mathbb{P}(|\widehat{r}_k^b - r_k^b| > t) \leq c_1(n+k)\exp\left(-\frac{c_2}{\sigma_{\mathsf{w}}}\frac{t^2(n+k)}{c_3\sigma_{\mathsf{w}}^3 + t^{3/2}\sqrt{n+k}}\right), \tag{25}$$

for positive absolute constants c_1, c_2 and c_3 which are independent of the dimensions of the problem and are only functions of $\tilde{c}$.

Proof: The lemma is a special case of Theorem 4 under Condition 2 of Remark 3 in [47]. ∎

Using the result of Lemma 3, we can control t and establish the RE condition for $\widehat{R}$ as follows:

Lemma 4. *Let m be a positive integer. Then, X satisfies the RE condition of order $(m+1)s$ with a constant $\lambda_{\min}/2$ with probability at least*

$$1 - c_1 p^2(n+p)\exp\left(-\frac{c_3\sqrt{\frac{n}{s}}}{1 + c_4\frac{n+p}{\left(\frac{n}{s}\right)^{3/2}}}\right), \tag{26}$$

where c_1 is the same as in Lemma 3, $c_3 = \frac{c_2}{\sigma_{\mathsf{w}}}\sqrt{\frac{\lambda_{\min}}{2(m+1)}}$ and $c_4 = \frac{c_3\sigma_{\mathsf{w}}^3}{\left(\frac{\lambda_{\min}}{2(m+1)}\right)^{3/2}}$.

Proof: First, note that for the given AR process, condition (22) is verified by the Wold decomposition of the process, condition (23) results from the sub-Gaussian assumption on the innovations, and condition (24) results from the stability of the process. Noting that

$$\widehat{R}_{i,i+k} = \frac{1}{n}\sum_{i=1}^{n} x_i x_{i+k} = \frac{1}{n}\sum_{i,j=1,j-i=k}^{n+k} x_i x_j = \frac{n+k}{n}\widehat{r}_k^b, \tag{27}$$

for $i = 1, \cdots, n$ and $k = 0, \cdots, p-1$, Eq. (25) implies:

$$\mathbb{P}\left(|\widehat{R}_{i,i+k} - R_{i,i+k}| > \tau\right) \leq c_1(n+k)\exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+k)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}}\right). \tag{28}$$

By the union bound and $k \leq p$, we get:

$$\mathbb{P}\left(\max_{i,j}|\widehat{R}_{ij} - R_{ij}| > \tau\right) \leq c_1 p^2(n+p)\exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+p)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}}\right). \tag{29}$$

Choosing $\tau = \frac{\lambda_{\min}}{2(m+1)s}$ and invoking the result of Lemma 2 establishes the result of the lemma. ∎

We next define the closely related notion of the Restricted Strong Convexity (RSC):

Definition 2 (Restricted Strong Convexity [31])**.** *Let*

$$\mathbb{V} := \{\boldsymbol{h} \in \mathbb{R}^p | \|\boldsymbol{h}_{S^c}\|_1 \leq 3\|\boldsymbol{h}_S\|_1 + 4\sigma_s(\boldsymbol{\theta})\}. \tag{30}$$

Then, X is said to satisfy the RSC condition of order s if there exists a positive $\kappa > 0$ such that

$$\frac{1}{n}\boldsymbol{h}^T X^T X \boldsymbol{h} = \frac{1}{n}\|X\boldsymbol{h}\|_2^2 \geq \kappa\|\boldsymbol{h}\|_2^2, \quad \forall \boldsymbol{h} \in \mathbb{V}. \tag{31}$$

The RSC condition can be deduced from the RE condition according to the following result:

Lemma 5 (lemma 4.1 of [35])**.** *If X satisfies the RE condition of order $s_\star = (m+1)s$ with a constant $\lambda_{\min}((m+1)s)$, then the RSC condition of order s holds with*

$$\kappa = \lambda_{\min}((m+1)s)\left(1 - 3\sqrt{\frac{\lambda_{\max}(ms)}{m\lambda_{\min}((m+1)s)}}\right)^2. \tag{32}$$

We can now establish the RSC condition of order s for X:

Lemma 6. *The matrix of covariates X satisfies the RSC condition of order s with a constant $\kappa = \frac{\sigma_{\mathsf{w}}^2}{16\pi}$ with probability at least*

$$1 - c_1 p^2(n+p)\exp\left(-\frac{c_\eta\sqrt{\frac{n}{s}}}{1 + c_\eta'\frac{n+p}{\left(\frac{n}{s}\right)^{3/2}}}\right), \tag{33}$$

where $c_\eta = \frac{c_2\eta}{\sqrt{16\pi(72+\eta^2)}}$ and $c_\eta' = \frac{c_3(16\pi(72+\eta^2))^{3/2}}{\eta^3}$.

Proof: Choosing $m = \lceil\frac{72}{\eta^2}\rceil$, and using Lemmas 2, reflem:re2, and 5 establishes the result. ∎

We are now ready prove Theorems 1 and 2.

B. Proof of Theorem 1

We first establish the so-called vase condition for the error vector $\boldsymbol{h} = \widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}$:

Lemma 7. *For a choice of the regularization parameter $\gamma_n \geq \|\nabla\mathfrak{L}(\boldsymbol{\theta})\|_\infty = \frac{2}{n}\|X^T(x_1^n - X\boldsymbol{\theta})\|_\infty$, the optimal error $\boldsymbol{h} = \widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}$ belongs to the vase*

$$\mathbb{V} := \{\boldsymbol{h} \in \mathbb{R}^p | \|\boldsymbol{h}_{S^c}\|_1 \leq 3\|\boldsymbol{h}_S\|_1 + 4\sigma_s(\boldsymbol{\theta})\}. \tag{34}$$

Proof: Using several instances of the triangle inequality we have:

$$\begin{aligned}
0 &\geq \frac{1}{n}\left(\|x_1^n - X(\boldsymbol{\theta}+\boldsymbol{h})\|_2^2 - \|x_1^n - X\boldsymbol{\theta}\|_2^2\right) + \\
&\quad \gamma_n\left(\|\boldsymbol{\theta}+\boldsymbol{h}\|_1 - \|\boldsymbol{\theta}\|_1\right) \\
&\geq -\frac{1}{n}\|X^T(x_1^n - X\boldsymbol{\theta})\|_\infty\|\boldsymbol{h}\|_1 + \\
&\quad \gamma_n\left(\|\boldsymbol{\theta}_S + \boldsymbol{h}_{S^c} + \boldsymbol{h}_S + \boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}\|_1\right) \\
&\geq -\frac{\gamma_n}{2}(\sigma_s(\boldsymbol{h}) + \|\boldsymbol{h}_S\|_1) + \\
&\quad \gamma_n\left(\|\boldsymbol{\theta}_S + \boldsymbol{h}_{S^c}\|_1 - \|\boldsymbol{h}_S + \boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}\|_1\right) \\
&= -\frac{\gamma_n}{2}(\sigma_s(\boldsymbol{h}) + \|\boldsymbol{h}_S\|_1) + \\
&\quad \gamma_n(\|\boldsymbol{\theta}_S\|_1 + \sigma_s(\boldsymbol{h}) - \|\boldsymbol{h}_S\|_1 - \sigma_s(\boldsymbol{\theta}) - \sigma_s(\boldsymbol{\theta}) - \|\boldsymbol{\theta}_S\|_1) \\
&= \frac{\gamma_n}{2}(\sigma_s(\boldsymbol{h}) - 3\|\boldsymbol{h}_S\|_1 - 4\sigma_s(\boldsymbol{\theta})).
\end{aligned}$$

∎

The following result of Negahban et al. [31] allows us to characterize the desired error bound:

Lemma 8 (Theorem 1 of [31]). *If X satisfies the RSC condition of order s with a constant $\kappa > 0$ and $\gamma_n \geq \|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty$, then any optimal solution $\widehat{\boldsymbol{\theta}}_{\ell_1}$ satisfies*

$$\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2 \leq \frac{2\sqrt{s}\gamma_n}{\kappa} + \sqrt{\frac{2\gamma_n \sigma_s(\boldsymbol{\theta})}{\kappa}}. \qquad (\star)$$

In order to use Lemma 8, we need to control $\gamma_n = \|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty$. We have:

$$\nabla \mathfrak{L}(\boldsymbol{\theta}) = \frac{2}{n} X^T(x_1^n - X\boldsymbol{\theta}), \qquad (35)$$

It is easy to check that by the uncorrelatedness of the innovations w_k's, we have

$$\mathbb{E}\left[\nabla \mathfrak{L}(\boldsymbol{\theta})\right] = \frac{2}{n}\mathbb{E}\left[X^T(x_1^n - X\boldsymbol{\theta})\right] = \frac{2}{n}\mathbb{E}\left[X^T w_1^n\right] = \mathbf{0}. \qquad (36)$$

Eq. (36) is known as the orthogonality principle. We next show that $\nabla \mathfrak{L}(\boldsymbol{\theta})$ is concentrated around its mean. We can write

$$(\nabla \mathfrak{L}(\boldsymbol{\theta}))_i = \frac{2}{n} x_{-i+1}^{n-i}{}^T w_1^n,$$

and observe that the jth element in this expansion is of the form $y_j = x_{n-i-j+1} w_{n-j+1}$. It is easy to check that the sequence y_1^n is a martingale with respect to the filtration given by

$$\mathcal{F}_j = \sigma\left(x_{-p+1}^{n-j+1}\right),$$

where $\sigma(\cdot)$ denote the sigma-field generated by the random variables $x_{-p+1}, x_{-p+2}, \cdots, x_{n-j+1}$. We use the following concentration result for sums of dependent random variables [48]:

Lemma 9. *Fix $n \geq 1$. Let Z_j's be sub-Gaussian $\mathcal{F}_j$-measurable random variables, satisfying for each $j = 1, 2, \cdots, n$,*

$$\mathbb{E}\left[Z_j | \mathcal{F}_{j-1}\right] = 0, \quad \text{almost surely,}$$

then there exists a constant c such that for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n Z_j - \mathbb{E}[Z_j]\right| \geq t\right) \leq \exp\left(-\frac{nt^2}{c^2}\right).$$

Proof: This is a special case of Theorem 3.2 of [48] or Lemma 3.2 of [49], for sub-Gaussian-weighted sums of random variables. The constant c depends on the sub-Gaussian constant of Z_i's. ∎

Since y_j's are a product of two independent sub-Gaussian random variables, they are sub-Gaussian as well. Lemma 9 implies that

$$\mathbb{P}\left(|\nabla \mathfrak{L}(\boldsymbol{\theta})_i| \geq t\right) \leq \exp\left(-\frac{nt^2}{c_0^2 \sigma_w^4}\right). \qquad (37)$$

where $c_0^2 := \frac{c^2}{\sigma_w^4}$ is an absolute constant. By the union bound, we get:

$$\mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq t\right) \leq \exp\left(-\frac{t^2 n}{c_0^2 \sigma_w^4} + \log p\right). \qquad (38)$$

Let d_4 be any positive integer. Choosing $t = c_0 \sigma_w^2 \sqrt{1 + d_4}\sqrt{\frac{\log p}{n}}$, we get:

$$\mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq c_0 \sigma_w^2 \sqrt{1 + d_4}\sqrt{\frac{\log p}{n}}\right) \leq \frac{2}{n^{d_4}}.$$

Hence, a choice of $\gamma_n = d_2 \sqrt{\frac{\log p}{n}}$ with $d_2 := c_0 \sigma_w^2 \sqrt{1 + d_4}$, satisfies $\gamma_n \geq \|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty$ with probability at least $1 - \frac{2}{n^{d_4}}$. Let $d_1 = (2c_\eta')^{2/3}$. Using Lemma 6, the fact that $n > d_1 s p^{2/3}$ by hypothesis, and $p > n$ we have that the RSC of order s hold for $\kappa = \frac{\sigma_w^2}{16\pi}$ with a probability at least $1 - 2c_1 p^3 \exp(-\frac{c_\eta p^{1/3}}{2})$. Combined these two assertions, the claim of Theorem 1 follows for $d_3 = 32\pi c_0 \sqrt{1 + d_4}$.

C. Proof of Theorem 2

The proof is mainly based on the following lemma, adopted from Theorem 2.1 of [33], stating that the greedy procedure is successful in obtaining a reasonable $s^\star$-sparse approximation, if the cost function satisfies the RSC:

Lemma 10. *Let $s^\star$ be a constant such that*

$$s^\star \geq \frac{s}{\eta^2}\log\frac{5s}{\eta^2}, \qquad (39)$$

and suppose that $\mathfrak{L}(\boldsymbol{\theta})$ satisfies RSC of order $s^\star$ with a constant $\kappa > 0$. Then, we have

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}_S\right\|_2 \leq \frac{\sqrt{6}\varepsilon_{s^\star}}{\kappa},$$

where $\eta_{s^\star}$ satisfies

$$\varepsilon_{s^\star} \leq \sqrt{s^\star + s}\|\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty. \qquad (40)$$

Proof: The proof is a specialization of the proof of Theorem 2.1 in [33] to our setting with the spectral spread $1/4\eta^2$. ∎

In order to use Lemma 10, we need to bound $\|\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty$. We have:

$$\mathbb{E}\left[\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\right] = \frac{1}{n}\mathbb{E}\left[X^T(x_1^n - X\boldsymbol{\theta}_S)\right] = \frac{1}{n}\mathbb{E}\left[X^T X(\boldsymbol{\theta} - \boldsymbol{\theta}_S)\right]$$

$$= R(\boldsymbol{\theta} - \boldsymbol{\theta}_S) \leq \frac{\sigma_w^2}{2\pi\eta^2}\varsigma_s(\boldsymbol{\theta})\mathbf{1},$$

where in the second inequality we have used (36), and the last inequality results from Corollary 1. Let d_4' be any positive integer. Using the result of Lemma 9 together with the union bound yields:

$$\mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty \geq c_0 \sigma_w^2 \sqrt{1 + d_4'}\sqrt{\frac{\log p}{n}} + \frac{\sigma_w^2 \varsigma_s(\boldsymbol{\theta})}{2\pi\eta^2}\right) \leq \frac{2}{n^{d_4'}}.$$

Hence, we get the following concentration result for $\varepsilon_{s^\star}$:

$$\mathbb{P}\left(\varepsilon_{s^\star} \geq \sqrt{s^\star + s}\left(c_0 \sigma_w^2 \sqrt{1 + d_4'}\sqrt{\frac{\log p}{n}} + \frac{\sigma_w^2 \varsigma_s(\boldsymbol{\theta})}{2\pi\eta^2}\right)\right)$$

$$\leq \frac{2}{n^{d_4'}}. \qquad (41)$$

Noting that by (39) we have $s^\star + s \leq \frac{4s \log s}{\eta^2}$. Let $d_1' = 4(2c_\eta')^{2/3}$. By the hypothesis of $\varsigma_s(\boldsymbol{\theta}) \leq As^{1-\frac{1}{\xi}}$ for some

constant A, and invoking the results of Lemmas 6 and 10, we get:

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}_S\right\|_2 \leq d_2'\sqrt{\frac{s\log s\log p}{n}} + d_2''\sqrt{s\log s}\varsigma_s(\boldsymbol{\theta})$$

$$\leq d_2'\sqrt{\frac{s\log s\log p}{n}} + d_2''\frac{\sqrt{\log s}}{s^{\frac{1}{\xi}-\frac{3}{2}}},$$

where $d_2' = \frac{16\pi c_0\sqrt{24(1+d_4')}}{\eta}$ and $d_2'' = \frac{A}{\pi\eta^3}$, with probability at least $1 - 2c_1 p^3 \exp(-\frac{c_\eta p^{1/3}}{2}) - \frac{2}{n^{d_4'}} \geq 1 - \frac{3}{n^{d_4'}}$. Finally, we have:

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}\right\|_2 = \left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}_S + \boldsymbol{\theta}_S - \boldsymbol{\theta}\right\|_2$$

$$\leq \left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}_S\right\|_2 + \|\boldsymbol{\theta}_S - \boldsymbol{\theta}\|_2.$$

Choosing $d_3' = 2d_2''$ completes the proof.

D. Proof of Proposition 1

Consider the event defined by

$$\mathcal{A} := \left\{\max_{i,j}|\widehat{R}_{ij} - R_{ij}| \leq \tau\right\}.$$

Eq. (29) in the proof of Lemma 4 implies that:

$$\mathbb{P}(\mathcal{A}^c) \leq c_1 p^2(n+p) \exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+p)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}}\right).$$

By choosing τ as in the proof of Theorem 1, we have

$$\mathcal{R}_{\mathsf{est}}^2(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}) \leq \mathcal{R}_{\mathsf{est}}^2(\widehat{\boldsymbol{\theta}}_{\ell_1}) = \sup_{\mathcal{H}}\left(\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2\right]\right)$$

$$\leq \mathbb{P}(\mathcal{A})d_3^2\frac{s\log p}{n} + \sup_{\mathcal{H}}\mathbb{E}_{\mathcal{A}^c}\left[\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2\right]$$

$$\leq d_3^2\frac{s\log p}{n}$$

$$+ 8(1-\eta)^2 c_1 \exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+p)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}} + 3\log p\right),$$

where the second inequality follows from Theorem 1, and the third inequality follows from the fact that $\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2 \leq 4(1-\eta)^2$ by the sufficient stability assumption. For $n > d_1 sp^{2/3}$, the first term will be the dominant, and thus we get $\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}) \leq 2d_3\sqrt{\frac{s\log p}{n}}$, for large enough n.

As for a lower bound on $\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}})$, we take the approach of [25] by constructing a family of AR processes with sparse parameters $\boldsymbol{\theta}$ for which the minimax risk is optimal modulo constants. In our construction, we assume that the innovations are Gaussian. The key element of the proof is the Fano's inequality:

Lemma 11 (Fano's Inequality). *Let $\mathcal{Z}$ be a class of densities with a subclass $\mathcal{Z}^\star$ of densities $f_{\boldsymbol{\theta}_i}$, parameterized by $\boldsymbol{\theta}_i$, for $i \in \{0, \cdots, 2^M\}$. Suppose that for any two distinct $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}^\star$, $\mathcal{D}_{\mathsf{KL}}(f_{\boldsymbol{\theta}_1}\|f_{\boldsymbol{\theta}_2}) \leq \beta$ for some constant β. Let $\widehat{\boldsymbol{\theta}}$ be an estimate of the parameters. Then*

$$\sup_j \mathbb{P}(\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j|H_j) \geq 1 - \frac{\beta + \log 2}{M}, \tag{42}$$

where H_j denotes the hypothesis that $\boldsymbol{\theta}_j$ is the true parameter, and induces the probability measure $\mathbb{P}(.|H_j)$.

Consider a class $\mathcal{Z}$ of AR processes with s-sparse parameters over any subset $S \subset \{1, 2, \cdots, p\}$ satisfying $|S| = s$, with parameters given by

$$\theta_\ell = \pm e^{-m}\mathbb{1}_S(\ell), \tag{43}$$

where m remains to be chosen. We also add the all zero vector $\boldsymbol{\theta}$ to $\mathcal{Z}$. For a fixed S, we have $2^s + 1$ such parameters forming a subfamily $\mathcal{Z}_S$. Consider the maximal collection of $\binom{p}{s}$ subsets S for which any two subsets differ in at least $s/4$ indices. The size of this collection can be identified by $A(p, \frac{s}{4}, s)$ in coding theory, where $A(n, d, w)$ represents the maximum size of a binary code of length n with minimum distance d and constant weight w [50]. We have

$$A(p, \tfrac{s}{4}, s) \geq \frac{p^{\frac{7}{8}s-1}}{s!},$$

for large enough p (See Theorem 6 in [51]). Also, by the Gilbert-Varshamov bound [50], there exists a subfamily $\mathcal{Z}_S^\star \subset \mathcal{Z}_S$, of cardinality $|\mathcal{Z}_S^\star| \geq 2^{\lfloor s/8 \rfloor} + 1$, such that any two distinct $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}_S^\star$ differ at least in $s/16$ components. Thus for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}^\star := \bigcup_S \mathcal{Z}_S^\star$, we have

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \geq \frac{1}{4}\sqrt{s}e^{-m} =: \alpha, \tag{44}$$

and $|\mathcal{Z}^\star| \geq \frac{p^{\frac{7}{8}s-1}}{s!}2^{\lfloor s/8 \rfloor}$. For an arbitrary estimate $\widehat{\boldsymbol{\theta}}$, consider the testing problem between the $\frac{p^{\frac{7}{8}s-1}}{s!}2^{\lfloor s/8 \rfloor}$ hypotheses $H_j : \boldsymbol{\theta} = \boldsymbol{\theta}_j \in \mathcal{Z}^\star$, using the minimum distance decoding strategy. Using Markov's inequality we have

$$\sup_{\mathcal{Z}}\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right] \geq \sup_{\mathcal{Z}^\star}\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right]$$

$$\geq \frac{\alpha}{2}\sup_{\mathcal{Z}^\star}\mathbb{P}\left(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 \geq \frac{\alpha}{2}\right)$$

$$= \frac{\alpha}{2}\sup_j \mathbb{P}\left(\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j|H_j\right). \tag{45}$$

Let $f_{\boldsymbol{\theta}_j}$ denote joint probability distribution of $\{x_k\}_{k=1}^n$ conditioned on $\{x\}_{-p+1}^0$ under the hypothesis H_j. Using the Gaussian assumption on the innovations, for $i \neq j$, we have

$$\mathcal{D}_{\mathsf{KL}}(f_{\boldsymbol{\theta}_i}\|f_{\boldsymbol{\theta}_j}) \leq \sup_{i\neq j}\mathbb{E}\left[\log\frac{f_{\boldsymbol{\theta}_i}}{f_{\boldsymbol{\theta}_j}}|H_i\right]$$

$$\leq \sup_{i\neq j}\mathbb{E}\left[-\frac{1}{2\sigma_{\mathsf{w}}^2}\sum_{k=1}^n\left(\left(x_k - \boldsymbol{\theta}_i'x_{k-p}^{k-1}\right)^2 - \left(x_k - \boldsymbol{\theta}_j'x_{k-p}^{k-1}\right)^2\right)\Big|H_i\right]$$

$$\leq \sup_{i\neq j}\frac{n}{2\sigma_{\mathsf{w}}^2}\mathbb{E}\left[\left((\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'x_{k-p}^{k-1}\right)^2\Big|H_i\right]$$

$$= \frac{n}{2\sigma_{\mathsf{w}}^2}\sup_{i\neq j}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'R(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)$$

$$\leq \frac{n\lambda_{\mathsf{max}}}{2\sigma_{\mathsf{w}}^2}\sup_{i\neq j}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 \leq \frac{nse^{-2m}}{64\pi\eta^2} =: \beta. \tag{46}$$

Using Lemma 11, (44), (45) and (46) yield:

$$\sup_{\mathcal{Z}}\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right] \geq \frac{\sqrt{s}e^{-m}}{8}\left(1 - \frac{2\left(\frac{nse^{-2m}}{64\pi\eta^2} + \log 2\right)}{s\log p}\right).$$

for p large enough so that $\log p \geq \frac{\log s - \frac{9}{8}}{\frac{3}{8} - \frac{1}{s}}$. Choosing $m = \frac{1}{2} \log \left(\frac{n}{8\pi\eta^2 \log p} \right)$ gives us the claim of Proposition 1 with $L = \frac{d_3}{\eta\sqrt{2\pi}}$ for large enough s and p such that $s \log p \geq \log(256)$. The hypothesis of $s \leq \frac{1-\eta}{\sqrt{8\pi\eta}} \sqrt{\frac{n}{\log p}}$ guarantees that for all $\boldsymbol{\theta} \in \mathcal{Z}^\star$, we have $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta$.

APPENDIX B
STATISTICAL TESTS FOR GOODNESS-OF-FIT

In this appendix, we will give an overview of the statistical goodness-of-fit tests for assessing the accuracy of the AR model estimates. A detailed treatment can be found in [52].

A. Residue-based tests

Let $\widehat{\boldsymbol{\theta}}$ be an estimate of the parameters of the process. The residues (estimated innovations) of the process based on $\widehat{\boldsymbol{\theta}}$ are given by

$$e_k = x_k - \widehat{\boldsymbol{\theta}} x_{k-p}^{k-1}, \qquad i = 1, 2, \cdots, n.$$

The main idea behind most of the available statistical tests is to quantify how close the sequence $\{e_i\}_{i=1}^n$ is to an i.i.d. realization of a known distribution F_0 which is most likely absolutely continuous . Let us denote the empirical distribution of the n-samples by $\widehat{F}_n$. If the samples are generated from F_0 the Glivenko-Cantelli theorem suggests that:

$$\sup_t |\widehat{F}_n(t) - F_0(t)| \xrightarrow{\text{a.s.}} 0.$$

That is, for large n the empirical distribution $\widehat{F}_n$ is uniformly close to F_0. The Kolmogorov-Smirnov (KS) test, Cramér-von Mises (CvM) criterion and the Anderson-Darling (AD) test are three measures of discrepancy between $\widehat{F}_n$ and F_0 which are easy to compute and are sufficiently discriminant against alternative distributions. More specifically, the limiting distribution of the following three random variables are known: The KS test statistic

$$K_n := \sup_t |\widehat{F}_n(t) - F_0(t)|,$$

the CvM statistic

$$C_n := \int \left(\widehat{F}_n(t) - F_0(t) \right)^2 dF_0(t),$$

and the AD statistic

$$A_n := \int \frac{\left(\widehat{F}_n(t) - F_0(t) \right)^2}{F_0(t) \left(1 - F_0(t) \right)} dF_0(t).$$

For large values of n, the Glivenko-Cantelli theorem also suggests that these statistics should be small. A simple calculation leads to the following equivalent for the statistics:

$$K_n = \max_{1 \leq i \leq n} \max \left\{ \left| \frac{i}{n} - F_0(e_i) \right|, \left| \frac{i-1}{n} - F_0(e_i) \right| \right\},$$

$$nC_n = \frac{1}{12n} + \sum_{i=1}^n \left(F_0(e_i) - \frac{2i-1}{2n} \right)^2,$$

and

$$nA_n = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) \left(\log F_0(e_i) + \log \left(1 - F_0(e_i) \right) \right).$$

B. Spectral domain tests for Gaussian AR processes

The aforementioned KS, CvM and AD tests all depend on the distribution of the innovations. For Gaussian AR processes, the spectral versions of these tests are introduced in [39]. These tests are based on the similarities of the periodogram of the data and the estimated power-spectral density of the process. The key idea is summarized in the following lemma:

Lemma 12. *Let $S(\omega)$ be the (normalized) power-spectral density of stationary process with bounded spectral spread, and $\widehat{S}_n(\omega)$ be the periodogram of the n samples of a realization of such a process, then for all ω we have:*

$$\sqrt{n} \left(2 \int_0^\omega \left(\widehat{S}_n(\lambda) - S(\lambda) \right) d\lambda \right) \xrightarrow{\text{d.}} \mathcal{Z}(\omega), \quad (47)$$

where $\mathcal{Z}(\omega)$ is a zero-mean Gaussian process.

The explicit formula for the covariance function of $\mathcal{Z}(.)$ is calculated in [39]. Lemma 12 suggests that for a good estimate $\widehat{\boldsymbol{\theta}}$ which admits a power spectral density $S(\omega; \widehat{\boldsymbol{\theta}})$, one should get a (*close* to) Gaussian process replacing $S(\omega)$ with $S(\omega; \widehat{\boldsymbol{\theta}})$ in (47). The spectral form of the CvM, KS and AD statistics can thus be characterized given an estimate $\widehat{\boldsymbol{\theta}}$.

ACKNOWLEDGMENT

REFERENCES

[1] A. Kazemipour, B. Babadi, and M. Wu, "Sufficient conditions for stable recovery of sparse autoregressive models," in *50th Annual Conference on Information Sciences and Systems (CISS), March 16–18, Princeton, NJ*, 2016.

[2] H. Sang and Y. Sun, "Simultaneous sparse model selection and coefficient estimation for heavy-tailed autoregressive processes," *Statistics*, vol. 49, no. 1, pp. 187–208, 2015.

[3] K. Farokhi Sadabadi, "Vehicular traffic modelling, data assimilation, estimation and short term travel time prediction," Ph.D. dissertation, University of Maryland, College Park, 2014.

[4] S. A. Ahmed and A. R. Cook, *Application of time-series analysis techniques to freeway incident detection*, 1982, no. 841.

[5] M. S. Ahmed and A. R. Cook, *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*, 1979, no. 722.

[6] J. Barceló, L. Montero, L. Marqués, and C. Carmona, "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2175, pp. 19–27, 2010.

[7] S. Clark, "Traffic prediction using multivariate nonparametric regression," *Journal of transportation engineering*, vol. 129, no. 2, pp. 161–168, 2003.

[8] P. M. Robinson, *Time series with long memory*. Oxford University Press, 2003.

[9] H. Akaike, "Fitting autoregressive models for prediction," *Annals of the institute of Statistical Mathematics*, vol. 21, no. 1, pp. 243–247, 1969.

[10] D. S. Poskitt, "Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases," *Annals of the Institute of Statistical Mathematics*, vol. 59, no. 4, pp. 697–725, 2007.

[11] R. Shibata, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *The Annals of Statistics*, pp. 147–164, 1980.

[12] J. W. Galbraith and V. Zinde-Walsh, "On some simple, autoregression-based estimation and identification techniques for arma models," *Biometrika*, vol. 84, no. 3, pp. 685–696, 1997.

[13] J. Galbraith, V. Zinde-Walsh *et al.*, "Autoregression-based estimators for arfima models," CIRANO, Tech. Rep., 2001.

[14] C.-K. Ing, C.-Z. Wei *et al.*, "Order selection for same-realization predictions in autoregressive processes," *The Annals of Statistics*, vol. 33, no. 5, pp. 2423–2474, 2005.

[15] K. E. Baddour and N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1650–1662, 2005.

[16] M. E. Mann and J. Park, "Oscillatory spatiotemporal signal detection in climate studies: A multiple-taper spectral domain approach," *Advances in geophysics*, vol. 41, pp. 1–132, 1999.

[17] H. Akaike, "Maximum likelihood identification of gaussian autoregressive moving average models," *Biometrika*, vol. 60, no. 2, pp. 255–265, 1973.

[18] ——, "Statistical predictor identification," *Annals of the Institute of Statistical Mathematics*, vol. 22, no. 1, pp. 203–217, 1970.

[19] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[20] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[21] E. J. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians Madrid, August 22–30*, 2006, pp. 1433–1452.

[22] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[23] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[24] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5862–5875, 2010.

[25] A. Goldenshluger and A. Zeevi, "Nonasymptotic bounds for autoregressive time series modeling," *Annals of statistics*, pp. 417–444, 2001.

[26] P. Stoica and R. L. Moses, *Introduction to spectral analysis*. Prentice hall Upper Saddle River, 1997, vol. 1.

[27] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[28] J. P. Burg, "Maximum entropy spectral analysis." in *37th Annual International Meeting*. Society of Exploration Geophysics, 1967.

[29] S. L. Marple Jr, "Digital spectral analysis with applications," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1987, 512 p.*, vol. 1, 1987.

[30] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

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

[32] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[33] T. Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6215–6221, 2011.

[34] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[35] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[36] E. J. Candes, "Modern statistical estimation via oracle inequalities," *Acta numerica*, vol. 15, pp. 257–325, 2006.

[37] R. B. D'Agostino, *Goodness-of-fit-techniques*. CRC press, 1986, vol. 68.

[38] S. Johansen, "Likelihood-based inference in cointegrated vector autoregressive models," *OUP Catalogue*, 1995.

[39] T. W. Anderson, "Goodness-of-fit tests for autoregressive processes," *Journal of time series analysis*, vol. 18, no. 4, pp. 321–339, 1997.

[40] "Cushing, ok wti spot price fob dataset," (Date last accessed 14-December-2015). [Online]. Available: http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=D

[41] "Regional integrated transportation information system (ritis)," (Date last accessed 27-December-2015). [Online]. Available: https://ritis.org

[42] "Regional integrated transportation information system (ritis)," (Date last accessed 27-December-2015). [Online]. Available: http://i95coalition.org/projects/regional-integrated-transportation-information-system-ritis

[43] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[44] A. Kazemipour, B. Babadi, and M. Wu, "Sparse estimation of self-exciting point processes with application to LGN neural modeling," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 478–482.

[45] A. Kazemipour, M. Wu, and B. Babadi, "Robust estimation of self-exciting point process models with application to neuronal modeling," *arXiv preprint arXiv:1507.03955*, 2015.

[46] U. Grenander and G. Szegö, *Toeplitz forms and their applications*. Univ of California Press, 1958, vol. 321.

[47] R. Rudzkis, "Large deviations for estimates of spectrum of stationary series," *Lithuanian Mathematical Journal*, vol. 18, no. 2, pp. 214–226, 1978.

[48] S. A. van de Geer, "On Hoeffding's inequality for dependent random variables," in *Empirical Process Techniques for Dependent Data*, H. Dehling and W. Philipp, Eds. Springer, 2001.

[49] ——, *Empirical Processes in M-estimation*. Cambridge university press, 2000.

[50] F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*. Elsevier, 1977, vol. 16.

[51] R. L. Graham and N. Sloane, "Lower bounds for constant weight codes," *Information Theory, IEEE Transactions on*, vol. 26, no. 1, pp. 37–43, 1980.

[52] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*. Wiley New York et al, 1986, vol. 150.