A Finite-Sample Deviation Bound for Stable Autoregressive Processes

Rodrigo A. González GRODRIGO@KTH.SE
Cristian R. Rojas CRRO@KTH.SE
Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M. Zeilinger

Abstract

In this paper, we study non-asymptotic deviation bounds of the least squares estimator for Gaussian AR(n) processes. By relying on martingale concentration inequalities and a tail-bound for χ^2 distributed variables, we provide a concentration bound for the sample covariance matrix of the process output. With this, we present a problem-dependent finite-time bound on the deviation probability of any fixed linear combination of the estimated parameters of the AR(n) process. We discuss extensions and limitations of our approach.

Keywords: Autoregressive Processes, Non-Asymptotic Estimation, Least Squares, Finite Sample Analysis.

1. Introduction

Autoregressive (AR) processes are ubiquitous in engineering sciences, as they are applied in econometrics, time series analysis (Box et al., 2015), system identification (Ljung, 1999), signal processing (Kay, 1993), machine learning and control.

Given sampled data, the identification of the parameters of an AR process is usually done by ordinary least squares, which is known to have asymptotically optimal statistical performance (Mann and Wald, 1943; Durbin, 1960) and is related to Maximum Likelihood in a Gaussian framework. Despite its success in practical applications, most analyses of the least squares method are asymptotic. Finite-time analyses of this method are still rare in the literature, despite being important for computing the number of samples needed for achieving a specified accuracy, deriving finite-time confidence sets, and designing robust control schemes. Non-asymptotic performance bounds have been historically difficult to derive since most of the classical statistical methods are better suited for asymptotic results.

In recent years, new statistical tools from the theory of self-normalizing processes (De la Peña et al., 2008) and high dimensional probability (Wainwright, 2019) have shown to be useful for analyzing a wide range of regression models. These tools have impulsed research on finite-time properties of the least squares estimator, with unifying efforts from the system identification, control and machine learning communities. Among topics of interest, we can find sample complexity bounds (Jedra and Proutiere, 2019), $1-\delta$ probability bounds on parameter errors (Sarkar and Rakhlin, 2019), and confidence bounds (Lattimore and Szepesvári, 2020, Chap. 20).

Even though autoregression is a key aspect in dynamical systems and regression models, finite-time properties of AR(n) processes have not yet been studied deeply. AR(n) processes are of particular interest, as they build the foundations for studying general regression models such as ARX and ARMAX models, which are widely used in linear system identification (Ljung, 1999). Autoregres-

sive processes are also essential for two-stage ARMA estimation algorithms (Stoica and Moses, 2005), and for speech production models (Makhoul, 1975). For a greater understanding on how least squares performs on different autoregressive processes for finite-sample data, here we perform a non-asymptotic analysis of the least squares estimator of the coefficients of these processes. In summary, the main results of this paper are:

- via martingale concentration inequalities and bounds on χ^2-distribution tails, we derive a finite-time problem-dependent concentration bound for the sample data covariance matrix of an n-th order autoregressive process;
- using the previous result, we provide a bound on the deviation of any fixed linear combination of the parameters of an AR(n) process around its true value, such that larger deviations occur only with probability at most δ.

The rest of this paper is organized as follows. Our work is put into context in Section 2, and we define our notation in Section 3. In Section 4 the problem is explicitly formulated, and a preliminary result is given. We state and prove our concentration inequalities for the sample covariance matrix and deviation bound of the parameters of an AR(n) process in Section 5.[1] Section 6 provides the proof for a key result used in the previous section, and a discussion of the results is presented in Section 7.

2. Relation to prior work

In a time-series context, Bercu et al. (1997) and Bercu (2001) studied large deviation rates of the least squares estimator in an AR(1) process. These contributions provide problem independent bounds, and do not generalize to AR(n) processes. A problem dependent finite-time deviation and variance bound was provided by González and Rojas (2020) for stable and unstable AR(1) processes. Unfortunately, the tools used in that work cannot be extended to a multivariate setting. Asymptotic properties of AR(n) models were obtained by Lai and Wei (1983).

In a broader context, one of the first non-asymptotic results in system identification was presented in Campi and Weyer (2002), where a uniform bound for the difference between empirical and theoretical identification costs was obtained. More recently, among works that have analyzed finite-time identification for stochastic processes are Jedra and Proutiere (2020); Sarkar and Rakhlin (2019); Simchowitz et al. (2018); Faradonbeh et al. (2018) and Zheng and Cheng (2018). These contributions consider state-space formulations with first order vector autoregressive models that, contrary to the description of an AR(n) process in state-space, normally assume that the noise process perturbs all states instead of only one. In particular, the performance bounds in Simchowitz et al. (2018) consider the estimation of the full transition matrix instead of the parameters of interest for AR(n) modeling. By leveraging the direct relationship between the coefficients of interest and the transition matrix of the underlying state-space in controller form, bounds for AR(n) processes could possibly be obtained by finding the (finite-time) optimal projection of the A matrix whose error in operator norm is bounded in Simchowitz et al. (2018), such that the resulting matrix is exactly the one provided by the LS estimate of the underlying AR(n) process. After this, a concentration inequality that bounds the error over $\hat{A}(T)$ and the transition matrix in controller form of the AR(n)

1. Supplementary material regarding proofs of Lemmas 3 and 4, and auxiliary results, is available in the preprint version of this paper (González and Rojas, 2019).

process would be needed. Although plausible, the results we seek do not seem direct from (Simchowitz et al., 2018). Instead, our analysis resembles that of (Sarkar and Rakhlin, 2019, Theorem 1) in the derivation of a matrix concentration bound, and similarly to the works cited above, a Gramian matrix associated with the real process dictates the learning rate.

3. Notation

Given a matrix $A \in \mathbb{R}^{n\times n}$, $\rho(A)$ denotes its spectral radius. If x is a vector and A is a fixed positive definite matrix, then $\|x\|_2$ and $\|x\|_\infty$ denote the 2 and ∞-norm of x, while $\|x\|_A$ is the weighted 2-norm (i.e., $\|x\|_A := \sqrt{x^\top A x}$). If B is an event, B^c denotes its complement, and $\mathbb{P}(B)$ refers to its probability of occurrence. $\mathbb{E}\{y\}$ denotes the expected value of the random variable y.

4. Problem formulation and preliminary result

Consider the following AR(n) process described by

$$y_t = Y_{t-1}^\top \theta^0 + e_t, \tag{1}$$

where $Y_{t-1}^\top := [y_{t-1} \;\; y_{t-2} \;\; \dots \;\; y_{t-n}]$, e_t is a Gaussian white noise of variance σ^2, and the parameter vector is $\theta^0 := [\theta_1 \;\; \theta_2 \;\; \dots \;\; \theta_n]^\top \in \mathbb{R}^n$. Furthermore, assume that $\{y_t\}$ is a stationary process, and that θ^0 is such that the AR(n) process is asymptotically stationary, which implies that $p(x) = x^n - \theta_1 x^{n-1} - \cdots - \theta_{n-1}x - \theta_n$ is a Schur polynomial. In this work, we are interested in how $w^\top \hat{\theta}_N$ concentrates around its true value $w^\top \theta^0$, where $w \in \mathbb{R}^n$ with $\|w\|_2 = 1$ is fixed and $\hat{\theta}_N$ is the least squares estimator of θ^0 given the data $\{y_t\}_{t=1}^N$. By allowing w to be chosen freely, we study deviation probabilities for single parameters, or linear combinations of them. Note that this probability depends on the true parameters, an thus it gives information about how easily the parameters can be identified through ordinary least squares for a particular system. In other words, our interest is in interpretability; in particular, we are concerned on how the least squares estimator performs under different AR processes.

Unfortunately, an explicit expression of the deviation probability (or equivalently, the confidence region) of interest is elusive in the literature. Therefore, it is of our interest to find an upper bound of it instead. If we define $Y := [Y_n \;\; \dots \;\; Y_{N-1}]^\top$ and $E := [e_{n+1} \;\; \dots \;\; e_N]^\top$, we can write $w^\top(\hat{\theta}_N - \theta^0)$ as $w^\top (Y^\top Y)^{-1} Y^\top E$, and hence we pursue a bound of the form

$$\mathbb{P}(|w^\top (Y^\top Y)^{-1} Y^\top E| > \varepsilon) \leq \delta, \tag{2}$$

where ε can be expressed as a function of δ, N, and the true parameters. Note that the stochastic quantity $w^\top(\hat{\theta}_N - \theta^0)$ is a self-normalized process. That is, it is unit free and therefore not affected by scale changes (De la Peña et al., 2008). These processes are now ubiquitous in the machine learning community, as they arise naturally in, e.g., finite-time analysis of linear systems (Simchowitz et al., 2018) and stochastic bandit problems (Krishnamurthy et al., 2018).

To derive a bound like (2), we make use of a martingale tail inequality introduced in Abbasi-Yadkori et al. (2011), which is valid for sub-Gaussian stochastic processes.

Proposition 1 (Abbasi-Yadkori et al. (2011)) *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for*

some $R > 0$, i.e.

$$\forall \lambda \in \mathbb{R} \quad \mathbb{E}[e^{\lambda \eta_t} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 1$, define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s. \tag{3}$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 1$,

$$\|S_t\|^2_{\overline{V}_t{}^{-1}} \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

Although the result of Proposition 1 is also a deviation bound similar to (2), it relies on the fact that the matrix V is positive definite, which is valid only for regularized least squares problems. Despite this, in $\overline{V}_t$ we recognize the sample covariance matrix $\sum_{s=1}^{t} X_s X_s^\top$, which plays an important role in our main result. The key idea behind the proposed approach is to first obtain a finite-sample probability bound on the matrix $Y^\top Y = \sum_{s=n}^{N-1} Y_s Y_s^\top$, and use this result together with Proposition 1 to derive the novel deviation bound.

5. Main results

In this section, we present our finite-time bounds for AR processes. Firstly, in Theorem 2 we derive a $1-\delta$ concentration bound for the sample covariance matrix $Y^\top Y$, and then use this result to obtain a deviation bound for the least squares estimator for general AR(n) processes, which is presented in Theorem 6. For the following, we express the AR(n) process in a state-space formulation

$$x_{t+1} = \begin{bmatrix} \theta^{0\top} & 0 \\ I_n & 0_{n\times 1} \end{bmatrix} x_t + \begin{bmatrix} 1 \\ 0_{n\times 1} \end{bmatrix} e_{t+1}, \quad Y_t = \begin{bmatrix} I_n & 0_{n\times 1} \end{bmatrix} x_t, \tag{4}$$

where we denote from now on the transition matrix and input vector in (4) as A and B respectively.

Theorem 2 *Consider the AR(n) process described in (1), and $Y = [Y_n \ \dots \ Y_{N-1}]^\top$. Given $\epsilon > 0$, define the following quantities:*

$$\overline{V} := \sigma^2 \sum_{i=0}^{\infty} A^i B B^\top (A^\top)^i, \tag{5}$$

$$V_{\text{dn}} := (N-n) \begin{bmatrix} I_n & 0 \end{bmatrix} \left(\overline{V} - \epsilon\sigma^2 \sum_{i=0}^{\infty} A^i (A^\top)^i\right) \begin{bmatrix} I_n \\ 0 \end{bmatrix},$$

$$V_{\text{up}} := (N-n) \begin{bmatrix} I_n & 0 \end{bmatrix} \left(\overline{V} + \epsilon\sigma^2 \sum_{i=0}^{\infty} A^i (A^\top)^i\right) \begin{bmatrix} I_n \\ 0 \end{bmatrix},$$

$$\delta(\epsilon, N) := 2\Bigg\{\sqrt{2}\exp\left(\frac{-(N-n)\sigma^2\epsilon}{24n\mathbb{E}\{y_1^2\}}\right) + \exp\left[-\frac{N-n}{2}\left(1 + \frac{\epsilon}{3} - \sqrt{1 + \frac{2\epsilon}{3}}\right)\right]$$
$$+ \exp\left[-\frac{(N-n)\epsilon}{72(\|\theta\|_2 + 1)^2 \tilde{\beta}}\right] + \exp(-\epsilon\sqrt{N})\Bigg\}, \tag{6}$$

where

$$M_\Phi := \max_\omega |e^{j\omega n} - \theta_1 e^{j\omega(n-1)} - \cdots - \theta_n|^{-2}, \tag{7}$$

$$\tilde{\beta} := \frac{(n+1)N}{N-n}\left[\frac{\mathbb{E}\{y_1^2\}}{\epsilon\sigma^2} + \frac{2M_\Phi(1+\epsilon^{-1/2})}{N^{1/4}}\right]. \tag{8}$$

Then, for all $\epsilon > 0$ such that $V_{\text{dn}} \succ 0$, we have

$$\mathbb{P}(V_{\text{dn}} \preceq Y^\top Y \preceq V_{\text{up}}) \geq 1 - \delta(\epsilon, N). \tag{9}$$

Proof The AR(n) process can be rewritten as in (4), where x_t is equal to $\begin{bmatrix} Y_t^\top & y_{t-n} \end{bmatrix}^\top$. Note that the eigenvalues of A are precisely the poles of the autoregressive process, with an extra eigenvalue at 0. We are interested in bounding

$$Y^\top Y = \begin{bmatrix} I_n & 0 \end{bmatrix} \sum_{i=n}^{N-1} x_i x_i^\top \begin{bmatrix} I_n \\ 0 \end{bmatrix}.$$

The approach consists in first determining a concentration bound for $\sum_{i=n}^{N-1} x_i x_i^\top$, and then relating it to a concentration bound for $Y^\top Y$. In this spirit, we write

$$x_{i+1}x_{i+1}^\top = Ax_i x_i^\top A^\top + Ax_i e_{i+1} B^\top + Be_{i+1}x_i^\top A^\top + e_{i+1}^2 BB^\top, \tag{10}$$

and denote V_N as $(N-n)^{-1}\sum_{i=n}^{N-1} x_i x_i^\top$. If we sum over $i = n-1, \ldots, N-2$ in (10), we obtain

$$V_N = AV_N A^\top + \underbrace{\frac{1}{N-n}\left[A(x_{n-1}x_{n-1}^\top - x_{N-1}x_{N-1}^\top)A^\top + \sum_{i=n-1}^{N-2}\left(e_{i+1}Ax_iB^\top + e_{i+1}Bx_i^\top A^\top + e_{i+1}^2 BB^\top\right)\right]}_{E_N}.$$

Since the AR(n) process is asymptotically stationary, the Lyapunov equation above has as solution $V_N = \sum_{i=0}^\infty A^i E_N (A^\top)^i$. By construction, V_N tends to $\overline{V}$ (defined in (5)) with probability 1 as N tends to infinity (see, e.g. (Söderström, 2002, p. 64)). Thus, our goal is to obtain a finite-sample concentration bound that relates V_N with $\overline{V}$. For this, we bound $\sum_{i=n-1}^{N-2} e_i^2$ by its variance, and bound the other terms of E_N by a small matrix quantity ϵI, for all $N > N(\epsilon)$. Lemmas 3, 4 and 5 are needed for this purpose, which bound the probability of the following events:

$$\mathcal{E}_1 := \left\{\rho\left[A(x_{n-1}x_{n-1}^\top - x_{N-1}x_{N-1}^\top)A^\top\right] \leq \epsilon\sigma^2(N-n)/3\right\},$$
$$\mathcal{E}_2 := \left\{\rho\left[\textstyle\sum_{i=n-1}^{N-2} e_{i+1}^2 BB^\top - BB^\top\right] \leq \epsilon\sigma^2(N-n)/3\right\},$$
$$\mathcal{E}_3 := \left\{\rho\left[\textstyle\sum_{i=n-1}^{N-2}\left(e_{i+1}Ax_iB^\top + e_{i+1}Bx_i^\top A^\top\right)\right] \leq \epsilon\sigma^2(N-n)/3\right\}.$$

Lemma 3 *Consider the process described in (4), where $\{e_i\}$ is a Gaussian zero-mean i.i.d. of variance σ^2, and $\{x_t\}$ is a stationary random process. Then,*

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - 2\sqrt{2}\exp\left(\frac{-(N-n)\sigma^2\epsilon}{24n\mathbb{E}\{y_1^2\}}\right).$$

Lemma 4 *Let $\{e_i\}$ be a Gaussian zero-mean i.i.d. sequence of variance σ^2. Then,*

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - 2\exp\left[-\frac{N-n}{2}\left(1 + \frac{\epsilon}{3} - \sqrt{1 + \frac{2\epsilon}{3}}\right)\right].$$

Lemma 5 *Consider the same assumptions as in Lemma 3. For any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\mathcal{E}_3\right) \geq 1 - 2\exp\left[-\frac{(N-n)\epsilon}{72(\|\theta^0\|_2+1)^2\tilde{\beta}}\right] - 2\exp(-\epsilon\sqrt{N}),$$

where M_Φ and $\tilde{\beta}$ are defined in (7) and (8) respectively.

Due to space constraints, we provide proof of Lemma 5 only[2], which can be found in Section 6. With these three lemmas, and by the subadditivity of the spectral radius of Hermitian matrices (Bernstein, 2009, Fact 5.12.2), we have

$$\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \implies \rho(E_N - \sigma^2 BB^\top) \leq \sigma^2\epsilon \implies \sigma^2(BB^\top - \epsilon I) \preceq E_N \preceq \sigma^2(BB^\top + \epsilon I),$$

which occurs with probability not less than $1-\delta(\epsilon, N)$. This also implies (9). ■

Theorem 2 delivers a finite-sample bound on the sample covariance matrix. Naturally, this matrix will deviate from its expected value by a small amount for large sample sizes. Note that this bound depends on a fixed value ϵ, which can be chosen arbitrarily small. As most self-normalized process bounds, $\delta(\epsilon, N)$ does not depend on the variance of the process noise.

With this result, we are ready to state the desired deviation bound in Theorem 6.

Theorem 6 *Consider the AR(n) process described in (1), where θ^0 is assumed to yield an asymptotically stationary process, $\{e_t\}$ is an i.i.d. Gaussian random process with variance σ^2, and $\{y_t\}$ is stationary. Then,*

$$\mathbb{P}\left(|w^\top(\hat{\theta}_N - \theta^0)| > 2\sigma\|w^\top V_{\text{dn}}^{-1/2}\|_2\sqrt{\log\left(\frac{\det(V_{\text{up}}V_{\text{dn}}^{-1}+I_n)^{1/2}}{\delta(\epsilon,N)}\right)}\right) \leq 2\delta(\epsilon,N), \qquad (11)$$

where V_{dn}, V_{up} and $\delta(\epsilon, N)$ are as described in Theorem 2.

Proof We will follow the main ideas in (Sarkar and Rakhlin, 2019, Theorem 1). We start by writing an upper bound using the Cauchy-Schwartz inequality

$$|w^\top(\hat{\theta}_N - \theta^0)| = |w^\top(Y^\top Y)^{-1}Y^\top E| \leq \|w^\top(Y^\top Y)^{-1/2}\|_2\|(Y^\top Y)^{-1/2}Y^\top E\|_2.$$

In Theorem 2 we have found deterministic matrices $V_{\text{dn}}, V_{\text{up}}$ and a scalar $\delta(\epsilon, N)$ such that, for the event $\mathcal{E}_{\text{pm}} := \{V_{\text{dn}} \preceq Y^\top Y \preceq V_{\text{up}}\}$, we have $\mathbb{P}(\mathcal{E}_{\text{pm}}) \geq 1-\delta(\epsilon, N)$. The next step is to bound the self-normalized norm. This can be done by first defining the event

$$\mathcal{E}_{\text{sn}} := \left\{\|Y^\top E\|_{(Y^\top Y+V_{\text{dn}})^{-1}} \leq \sqrt{2\sigma^2\log\left(\frac{\det(Y^\top Y+V_{\text{dn}})^{1/2}\det(V_{\text{dn}})^{-1/2}}{\delta(\epsilon,N)}\right)}\right\}.$$

It follows from Proposition 1 that $\mathbb{P}(\mathcal{E}_{\text{sn}}) \geq 1-\delta(\epsilon, N)$. Also, under $\mathcal{E}_{\text{pm}}$ we have that $Y^\top Y + V_{\text{dn}} \preceq 2Y^\top Y$, which implies $(Y^\top Y + V_{\text{dn}})^{-1} \succeq \frac{1}{2}(Y^\top Y)^{-1}$. So, considering the set $\mathcal{E}_{\text{pm}} \cap \mathcal{E}_{\text{sn}}$, we obtain

$$\mathcal{E}_{\text{pm}} \cap \mathcal{E}_{\text{sn}} \implies \mathcal{E}_{\text{pm}} \cap \left\{\|(Y^\top Y)^{-1/2}Y^\top E\|_2 \leq 2\sigma\sqrt{\log\left(\frac{\det(V_{\text{up}}V_{\text{dn}}^{-1}+I_n)^{1/2}}{\delta(\epsilon,N)}\right)}\right\}.$$

2. Proofs of Lemmas 3 and 4 can be found in the preprint version of this paper (González and Rojas, 2019).

Furthermore, observe that $\mathbb{P}(\mathcal{E}_{\text{pm}} \cap \mathcal{E}_{\text{sn}}) \geq 1 - 2\delta(\epsilon, N)$. So, if $\mathcal{E}_{\text{pm}} \cap \mathcal{E}_{\text{sn}}$ holds, then

$$|w^\top (Y^\top Y)^{-1} Y^\top E| \leq 2\sigma \|w^\top V_{\text{dn}}^{-1/2}\|_2 \sqrt{\log\left(\frac{\det(V_{\text{up}} V_{\text{dn}}^{-1} + I_n)^{1/2}}{\delta(\epsilon, N)}\right)}, \tag{12}$$

which means that the probability of the event in (12) is at least $1 - 2\delta(\epsilon, N)$. By considering the complement event, we obtain the probability bound (11). ■

Theorem 6 provides a finite-sample confidence bound on the deviation of the weighted parameter vector $w^\top \hat{\theta}_N$ with respect to its asymptotic value $w^\top \theta^0$. This result delivers probability bounds on the deviation each parameter θ_i individually, as well as any linear combination of them. Note that ϵ can be considered a *tightness variable*, as by setting ϵ small, more samples are required to guarantee a desired confidence level, but the probability bound will be tighter.

To end this analysis, we derive the decay rate of our probability bound in Corollary 7.

Corollary 7 *If ϵ is picked as $\lambda_n - N^{-1/2}$, where λ_n is the smallest eigenvalue of*

$$\left(\sigma^2 \begin{bmatrix} I_n \\ 0 \end{bmatrix}^\top \sum_{i=0}^{\infty} A^i (A^\top)^i \begin{bmatrix} I_n \\ 0 \end{bmatrix}\right)^{-1/2} \begin{bmatrix} I_n \\ 0 \end{bmatrix}^\top \overline{V} \begin{bmatrix} I_n \\ 0 \end{bmatrix} \left(\sigma^2 \begin{bmatrix} I_n \\ 0 \end{bmatrix}^\top \sum_{i=0}^{\infty} A^i (A^\top)^i \begin{bmatrix} I_n \\ 0 \end{bmatrix}\right)^{-\top/2},$$

then $\delta \sim Ce^{-\lambda_n \sqrt{N}}$ for large N and the deviation in (11) is asymptotically a constant in N. This shows that the rate of decay of the probability bound is at least exponential in $\sqrt{N}$.

6. Proof of Lemma 5

Here we present a sketch of the proof of Lemma 5, in which we use a martingale concentration inequality from Simchowitz et al. (2018) and exploit the Gaussianity of $\{e_t\}$ by applying a concentration inequality for χ^2 random variables found in Laurent and Massart (2000).

Proof of Lemma 5 For any vector $q := \begin{bmatrix} q_1 & \tilde{q}^\top \end{bmatrix}^\top \in \mathbb{R}^{n+1}$ of unit 2-norm, we have

$$q^\top \left(\frac{1}{N-n} \sum_{i=n}^{N-1} e_i (A x_{i-1} B^\top + B x_{i-1}^\top A^\top)\right) q = \frac{2q_1}{N-n} (q_1 \theta^{0\top} + \tilde{q}^\top) \sum_{i=n}^{N-1} e_i Y_{i-1}. \tag{13}$$

Since (13) is symmetric around zero, it is sufficient to bound its upper tail. Next, we denote the vector $z := 2q_1(q_1\theta^0 + \tilde{q})$. By using Lemma 4.2 of Simchowitz et al. (2018), with $Z_t = \frac{1}{\sqrt{N-n}} \langle z, Y_{t-1} \rangle$, $W_t = e_t$, and $\beta = \epsilon \tilde{\beta} (N-n) \sigma^2 \max_{\|q\|_2 = 1} \|2q_1 \theta^{0\top} + \tilde{q}^\top\|_2^2$, we obtain the inequality

$$\mathbb{P}\left[\left\{\sum_{i=n}^{N-1} \frac{\langle z, Y_{i-1} \rangle e_i}{N-n} \geq \frac{\sigma^2 \epsilon}{3}\right\} \cap \left\{\sum_{i=n}^{N-1} \frac{\|Y_{i-1}\|_2^2}{\sigma^2 (N-n)} \leq \epsilon \tilde{\beta}\right\}\right] \leq \exp\left(-\frac{(N-n)\epsilon}{72 \tilde{\beta} \max\limits_{\|q\|_2 = 1} \|q_1(q_1\theta^0 + \tilde{q})\|_2^2}\right).$$

Using the well-known inequality $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$, and the fact that $\sum_{i=n}^{N-1} \|Y_{i-1}\|_2^2 \leq (n+1) \sum_{i=-1}^{N-2} y_i^2$, we obtain

$$\mathbb{P}\left(\sum_{i=n}^{N-1} \frac{\langle z, Y_{i-1} \rangle e_i}{N-n} \geq \frac{\sigma^2 \epsilon}{3}\right) \leq \exp\left(-\frac{(N-n)\epsilon}{72 \max\limits_{\|q\|_2 = 1} \|q_1(q_1\theta^0 + \tilde{q})\|_2^2 \tilde{\beta}}\right) + \mathbb{P}\left(\frac{n+1}{\sigma^2 (N-n)} \sum_{i=-1}^{N-2} y_i^2 > \epsilon \tilde{\beta}\right).$$

To tackle the last probability, we note that $\begin{bmatrix} y_{-1} & \cdots & y_{N-2} \end{bmatrix}^\top \sim \mathcal{N}(0, R_N)$, where R_N is a symmetric Toeplitz covariance matrix of eigenvalues $\{\lambda_i\}_{i=1}^N$. Hence, $Z = \sum_{i=-1}^{N-2} y_i^2$ is a generalized χ^2 random variable, whose distribution is equal to the distribution of $v^\top R_N v$, where $v \sim \mathcal{N}(0, I_N)$. By the singular value decomposition $R_N = U_N D_N U_N^\top$ where $D_N = \operatorname{diag}\{\lambda_i\}$ and U_N is a unitary matrix, and the rotation invariance of v (Vershynin, 2018, Chap. 3), we see that

$$\mathbb{P}\left(\sum_{i=-1}^{N-2} y_i^2 - \mathbb{E}\left\{\sum_{i=-1}^{N-2} y_i^2\right\} > t\right) = \mathbb{P}\left(\sum_{i=1}^{N} \lambda_i(\tilde{v}_i^2 - 1) > t\right),$$

where $\tilde{v} \sim \mathcal{N}(0, I_N)$. Then, by (Laurent and Massart, 2000, Lemma 1),

$$\mathbb{P}\left(\sum_{i=-1}^{N-2} y_i^2 - \mathbb{E}\left\{\sum_{i=-1}^{N-2} y_i^2\right\} > 2\|\lambda\|_2\sqrt{t} + 2\|\lambda\|_\infty t\right) \leq \exp(-t). \tag{14}$$

It is known (see, e.g. (Gray, 2006, Section 4.2)) that the maximum eigenvalue of R_N is bounded by σM_Φ, where M_Φ is defined as in (7). By letting $t = \epsilon\sqrt{N}$ in (14), and upper bounding $\|\lambda\|_2$ and $\|\lambda\|_\infty$ by $\sqrt{N}M_\Phi$ and M_Φ respectively, we deduce that

$$\mathbb{P}\left(\frac{n+1}{\sigma^2(N-n)}\sum_{i=-1}^{N-2} y_i^2 > \underbrace{\frac{(n+1)N}{N-n}\left[\frac{\mathbb{E}\{y_1^2\}}{\sigma^2} + \frac{2M_\Phi(\epsilon+\sqrt{\epsilon})}{N^{1/4}}\right]}_{\epsilon\tilde{\beta}}\right) \leq \exp(-\epsilon\sqrt{N}).$$

Finally, note that $\max_{\|q\|_2=1} \|q_1(q_1\theta^0 + \tilde{q})\|_2^2 \leq (\|\theta^0\|_2 + 1)^2$. With this, and considering the complement event, we reach the bound in Lemma 5. ■

7. Discussion and conclusions

In this paper, we have provided finite-sample guarantees for the least squares estimates of the coefficients of general AR(n) processes. For this, a concentration bound for the sample covariance matrix was derived. In this bound, the Gramian matrix $\sum_{i=0}^{\infty} A^i(A^\top)^i$ in V_{dn} and V_{up} shows that faster processes need less samples to guarantee concentration of the covariance matrix, which is a natural result. Regarding Theorem 6, we find that the fixed vector w impacts the confidence bound through the inverse of V_{dn}, which resembles the results obtained in (Lattimore and Szepesvári, 2020, Eq. 20.2) for least squares estimates of linear bandit algorithms with deterministic actions. The $\log\det$ term is also unsurprising, as it also appears in finite-sample analysis of LTI systems (see, e.g. (Sarkar and Rakhlin, 2019, Eq. 12)). The deterministic matrices V_{dn} and V_{up} in (11) capture the correct behavior of the confidence bound, since it is large when the uncertainty on the sample covariance matrix is also large. Also, note that the proof of Theorem 6 heavily relies on bounding the probability of the normal matrix $Y^\top Y$, but it is easily decoupled from Theorem 2. That is, if tighter bounds for $\mathcal{E}_{\text{pm}}$ can be found, then Theorem 6 can be directly improved. Future work concerns proving finite-time variance bounds for the estimated parameters, extending the analysis for ARX models under sub-Gaussian noise, and deriving sharp lower bounds for AR(n) processes.

Acknowledgments

This work was supported by the Swedish Research Council under contract number 2016-06079 (NewLEADS).

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

B. Bercu. On large deviations in the gaussian autoregressive process: stable, unstable and explosive cases. *Bernoulli*, 7(2):299–316, 2001.

B. Bercu, F. Gamboa, and A. Rouault. Large deviations for quadratic forms of stationary gaussian processes. *Stochastic Processes and their Applications*, 71(1):75–90, 1997.

D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control,* 5th ed. John Wiley & Sons, 2015.

M. C. Campi and E. Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

V. H. De la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2008.

J. Durbin. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1):139–153, 1960.

M. Faradonbeh, A. Tewari, and G. Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

R. A. González and C. R. Rojas. A Finite-sample Deviation Bound for Stable Autoregressive Processes. *arXiv preprint arXiv:1912.08103*, 2019.

R. A. González and C. R. Rojas. Finite sample deviation and variance bounds for first order autoregressive processes. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* accepted for publication, 2020.

R. M. Gray. Toeplitz and Circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

Y. Jedra and A. Proutiere. Sample Complexity Lower Bounds for Linear System Identification. In *58th IEEE Conference on Decision and Control (CDC), Nice, France*, pages 2676–2681, 2019.

Y. Jedra and A. Proutiere. Finite-time identification of stable linear systems: Optimality of the least-squares estimator. *arXiv preprint arXiv:2003.07937*, 2020.

S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

A. Krishnamurthy, Z. S. Wu, and V. Syrgkanis. Semiparametric Contextual Bandits. In *International Conference on Machine Learning (ICML)*, pages 2776–2785, 2018.

T. L. Lai and C. Z. Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of multivariate analysis*, 13(1): 1–23, 1983.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

L Ljung. *System Identification: Theory for the User,* 2nd Edition. Prentice-Hall, 1999.

J. Makhoul. Linear Prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

H. B. Mann and A. Wald. On the statistical treatment of linear stochastic difference equations. *Econometrica, Journal of the Econometric Society*, pages 173–220, 1943.

T. Sarkar and A. Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473, 2018.

T. Söderström. *Discrete-time Stochastic Systems: Estimation and Control*. Springer, 2002.

P. Stoica and R. L. Moses. *Spectral Analysis of Signals*. Prentice Hall, 2005.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Y. Zheng and G. Cheng. Finite time analysis of vector autoregressive models under linear restrictions. *arXiv preprint arXiv:1811.10197*, 2018.