A Finite-Sample Deviation Bound for Stable Autoregressive Processes

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

Abstract
In this paper, we study non-asymptotic deviation bounds of the least squares estimator in Gaussian AR($n$) processes. By relying on martingale concentration inequalities and a tail-bound for $\chi^2$ 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 autoregressive process is usually done by ordinary least squares. This method 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 impelled 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, 2018, 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). Autoregressive processes are also essential for two-stage ARMA estimation algorithms (Stoica and Moses,
A finite-sample deviation bound for stable autoregressive processes

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 present a non-asymptotic analysis of these processes. In summary, the main results of this paper are:

- Via martingale concentration inequalities and bounds on $\chi^2$-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 $\delta$.

The rest of this paper is organized as follows. Our work is put into context in Section 2. We define our notation in Section 3, and explicitly formulate the problem in Section 4. Preliminaries and lemmas are presented in Section 5. After this, 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 6. Discussion of the results and conclusions are presented in Section 7. We also include an Appendix with supplementary material of proofs and auxiliary results.

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 (2019) for stable and unstable AR(1) processes. Unfortunately, the tools used in this 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 Sarkar and Rakhlin (2019); Simchowitz et al. (2018); Faradonbeh et al. (2018); Zheng and Cheng (2018). These contributions consider state-space formulations, with vector autoregressive models (in particular, VAR(1) models). Applying these bounds directly on an AR($n$) process context is not feasible, since in these works it is normally assumed that the noise process perturbs all states. Furthermore, in contrast with state-space descriptions, we are interested in an $n$-th dimensional parameter vector instead of a transition matrix. However, 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 $\infty$-norm of $x$, while $\|x\|_A$ is the weighted 2-norm (i.e. $\|x\|_A := \sqrt{x^T A x}$). 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 $y$. 

---

2
4. Problem formulation

Consider the following AR($n$) process described by

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

where $Y_{t-1}^\top := [y_{t-1} \  y_{t-2} \cdots \ y_{t-n}]$, $e_t$ is a Gaussian white noise of variance $\sigma^2$, and the parameter vector is $\theta^0 := [\theta_1 \ \theta_2 \cdots \ \theta_n]^\top \in \mathbb{R}^n$. Furthermore, assume that $\{y_t\}$ is a stationary process, and that $\theta^0$ 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 $\theta^0$ 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; it gives information about how easily the parameters can be identified through ordinary least squares for a particular system. Thus, 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 $Y := [Y_n \cdots Y_{N-1}]^\top$ and $E := [e_{n+1} \cdots e_N]^\top$, 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

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

where $\varepsilon$ can be expressed as a function of $\delta, N$, and the true parameters. In the following section we cover the main tools we use to obtain the desired bound.

5. Preliminaries: A martingale tail inequality and useful concentration lemmas

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 $\eta_t$ is $\mathcal{F}_t$-measurable and $\eta_t$ 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

$$\nabla_t = V + \sum_{s=1}^t X_s X_s^\top, \quad S_t = \sum_{s=1}^t \eta_s X_s,$$

(3)
Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 1$,

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

Although the result of Proposition 1 is also a deviation bound similar to (2), it relies on the fact that the matrix $\mathbf{V}$ is positive definite, which is valid only for regularized least squares problems. Despite this, in $\mathbf{V}_t$ 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 our 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. We first need a matrix concentration inequality of the form

$$\mathbb{P}(V_{dn} \leq Y^\top Y \leq V_{up}) \geq 1 - \delta. \quad (4)$$

For the derivation of the inequality in (4), we express the AR($n$) process in a state-space formulation

$$x_{t+1} = \begin{bmatrix} \theta_0^\top \\ I_n \end{bmatrix} x_t + \begin{bmatrix} 1 \\ 0_{n\times 1} \end{bmatrix} e_{t+1}, \quad Y_t = \begin{bmatrix} I_n & 0_{n\times 1} \end{bmatrix} x_t, \quad (5)$$

where we denote from now on the transition matrix and input vector in (5) as $A$ and $B$. With this description, we introduce 3 novel lemmas that relate functions of the random variables in (5) with their asymptotic behaviour. Before stating them, we define the following events:

$$\mathcal{E}_1 := \left\{ \rho \left[ \frac{1}{\sigma^2(N - n)} A(x_{n-1} x_{n-1}^\top - x_{N-1} x_{N-1}^\top) A^\top \right] \leq \frac{\epsilon}{3} \right\},$$

$$\mathcal{E}_2 := \left\{ \rho \left[ \frac{1}{\sigma^2(N - n)} \sum_{i=n+1}^{N-2} e_{i+1} B B^\top \right] \leq \frac{\epsilon}{3} \right\},$$

$$\mathcal{E}_3 := \left\{ \rho \left[ \frac{1}{\sigma^2(N - n)} \sum_{i=n+1}^{N-2} \left( e_{i+1} A x_i B^\top + e_{i+1} B x_i A^\top \right) \right] \leq \frac{\epsilon}{3} \right\}.$$

**Lemma 2** Consider the process described in (5), where $\{e_i\}$ is a Gaussian zero-mean i.i.d. of variance $\sigma^2$, and $\{x_t\}$ is a stationary random process. Then,

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

**Lemma 3** Let $\{e_i\}$ be a Gaussian zero-mean i.i.d. sequence of variance $\sigma^2$. 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 4** Consider the same assumptions as in Lemma 2. For any $\epsilon > 0$, we have

$$\mathbb{P}(\mathcal{E}_3) \geq 1 - 2 \exp \left[ -\frac{(N - n)\epsilon}{72(\|\theta_0\|_2 + 1)^2 \beta} \right] - 2 \exp(-\epsilon \sqrt{N}),$$
where

\[ M_\Phi := \max_\omega |e^{j\omega n} - \theta_1 e^{j\omega(n-1)} - \ldots - \theta_n|^2, \]  

\[ \beta := \frac{(n+1)N}{N-n} \left[ \frac{\mathbb{E}\{y_i^2\}}{e\sigma^2} + \frac{2M_\Phi(1+e^{-1/2})}{N^{1/4}} \right]. \]

In the main text, we provide a proof sketch of Lemma 4 only, 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 \( \chi^2 \) random variables found in Laurent and Massart (2000).

**Proof sketch (of Lemma 4).** For any vector \( q := [q_1 \begin{pmatrix} q \end{pmatrix}]^\top \in \mathbb{R}^{n+1} \) of unit 2-norm, we have

\[ q^\top \left( \frac{1}{N-n} \sum_{i=n}^{N-1} e_i (A x_{i-1} B^\top + B x_{i-1} A^\top) \right) q = \frac{2q_1}{N-n} (q_1 \theta_0^\top + \tilde{q}^\top) \sum_{i=n}^{N-1} e_i Y_{i-1}. \]

Since (8) is symmetric around zero, it is sufficient to bound its upper tail. Next, we denote the vector \( z := 2q_1(q_1 \theta_0^\top + \tilde{q}) \). By using Lemma 4.2 of Simchowitz et al. (2018), with \( Z_t = \frac{2}{\sqrt{N-n}} (z, Y_t) \), \( W_t = e_t \), and \( \beta = e\tilde{\beta}(N-n)\sigma^2 \max_{\|q\|_2=1} \|2q_1 \theta_0^\top + \tilde{q}^\top\|_2 \), we obtain the inequality

\[ \mathbb{P} \left[ \left\{ \sum_{i=n}^{N-1} \frac{(z, Y_{i-1}) e_i}{N-n} \geq \frac{\sigma^2 \epsilon}{3} \right\} \cap \left\{ \sum_{i=n}^{N-1} \frac{\|Y_{i-1}\|_2^2}{\sigma^2(N-n)} \leq \epsilon \tilde{\beta} \right\} \right] \leq \exp \left( -\frac{(N-n)\epsilon}{72\beta \max_{\|q\|_2=1} \|q_1(q_1 \theta_0^\top + \tilde{q})\|_2^2} \right). \]

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

\[ \mathbb{P} \left( \sum_{i=n}^{N-1} \frac{(z, Y_{i-1}) e_i}{N-n} \geq \frac{\sigma^2 \epsilon}{3} \right) \leq \exp \left( -\frac{(N-n)\epsilon}{72 \max_{\|q\|_2=1} \|q_1(q_1 \theta_0^\top + \tilde{q})\|_2^2} \right) + \mathbb{P} \left( \sum_{i=1}^{N-2} y_i^2 > \epsilon \tilde{\beta} \right). \]

To tackle the last probability, we note that \( [y_{-2} \cdots y_{N-2}]^\top \sim \mathcal{N}(0, R_N) \), 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 \( \chi^2 \) 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 \( D_N = \text{diag} \{\lambda_i\} \) 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),

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

It is known (see, e.g. (Gray, 2006, Section 4.2)) that the maximum eigenvalue of \( R_N \) is bounded by \( \sigma M_\Phi \), where \( M_\Phi \) is defined as in (6). By letting \( t = \epsilon \sqrt{N} \) in (9), and upper bounding \( \|\lambda\|_2 \) and
we derive a bound for the sample covariance matrix for the least squares estimator for general AR processes.

\[ \|\lambda\|_\infty \text{ by } \sqrt{N}M_\Phi \text{ and } M_\Phi \text{ respectively, we deduce that } \]
\[ \mathbb{P}\left( \frac{n+1}{\sigma^2(N-n)} \sum_{i=1}^{N-2} y_i^2 > \frac{(n+1)N}{N-n} \left[ \frac{\mathbb{E}\{y_i^2\}}{\sigma^2} + \frac{2M_\Phi(\epsilon + \sqrt{\epsilon})}{N^{1/4}} \right] \right) \leq \exp(-\epsilon \sqrt{N}). \]

Finally, note that \( \max_{\|q\|_2=1} \|q_i(q_1\theta^0 + \tilde{q})\|_2 \leq (\|\theta^0\|_2 + 1)^2 \). With this, and considering the complement event, we reach the bound in Lemma 4.

6. Main results

Here we present the main results of this paper. Firstly, in Theorem 5 we derive a \( 1 - \delta \) concentration bound for the sample covariance matrix \( Y^T Y \), 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.

**Theorem 5** Consider the AR(\( n \)) process described in (1), and \( Y = [Y_n \ldots Y_{N-1}]^T \). Given \( \epsilon > 0 \), define the following quantities:
\[
\tilde{V} := \sigma^2 \sum_{i=0}^{\infty} A_i BB^T (A^T)^i, \tag{10}
\]
\[
V_{dn} := (N-n) [I_n \ 0] \left( \tilde{V} - \epsilon \sigma^2 \sum_{i=0}^{\infty} A_i^i (A^T)^i \right) [I_n \ 0],
\]
\[
V_{up} := (N-n) [I_n \ 0] \left( \tilde{V} + \epsilon \sigma^2 \sum_{i=0}^{\infty} A_i^i (A^T)^i \right) [I_n \ 0],
\]
\[
\delta(\epsilon, N) := 2 \left\{ \sqrt{2} \exp \left( \frac{-(N-n)\sigma^2 \epsilon}{24n\mathbb{E}\{y_i^2\}} \right) + \exp \left[ -\frac{N-n}{2} \left( 1 + \frac{\epsilon}{3} - \sqrt{1 + \frac{2\epsilon}{3}} \right) \right] \\
+ \exp \left[ -\frac{(N-n)\epsilon}{72(\|\theta\|_2 + 1)^2} \right] + \exp(-\epsilon \sqrt{N}) \right\}, \tag{11}
\]
where \( M_\Phi \) and \( \tilde{\beta} \) are defined in (6) and (7). Then, for all \( \epsilon > 0 \) such that \( V_{dn} > 0 \), we have
\[ \mathbb{P}(V_{dn} \leq Y^T Y \leq V_{up}) \geq 1 - \delta(\epsilon, N). \tag{12} \]

**Proof.** The AR(\( n \)) process can be rewritten as in (5), where \( x_i \) is equal to \( [Y_i^T \ y_{t-n}]^T \). 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
\[ Y^T Y = [I_n \ 0] \sum_{i=n}^{N-1} x_i x_i^T [I_n \ 0]. \]

The approach consists in first determining a concentration bound for \( \sum_{i=n}^{N-1} x_i x_i^T \), and then relating it to a concentration bound for \( Y^T Y \). In this spirit, we write
\[ x_{i+1} x_{i+1}^T = Ax_i x_i^T A^T + Ax_i e_{i+1} B^T + Be_{i+1} x_i^T A^T + e_{i+1}^2 BB^T, \tag{13} \]
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 (13), we obtain
\[
V_N = AV_N A^\top + \frac{1}{N - n} \left[ A(x_{n-1} x_{n-1}^\top - x_{N-1} x_{N-1}^\top)A^\top + \sum_{i=n-1}^{N-2} (e_{i+1} A x_i B^\top + e_{i+1} B x_i A^\top + e_i^2 B B^\top) \right].
\]

Since the AR\((n)\) system is stable, the Lyapunov equation above has a solution given by \( V_N = \sum_{i=0}^\infty A^i E_N (A^\top)^i \). Note that, by construction, \( V_N \) tends to \( \tilde{V} \) (defined in (10)) 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 its asymptotic expression \( \tilde{V} \). 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 \( \epsilon I \), for all \( N > N(\epsilon) \). Lemmas 2, 3, and 4 are needed for this purpose. With these three lemmas, and by the subadditivity of the spectral radius of Hermitian matrices (Bernstein, 2009, Fact 5.12.2), we have
\[
\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \implies \rho(E_N - \sigma^2 B B^\top) \leq \sigma^2 \epsilon \implies \sigma^2 (B B^\top - \epsilon I) \leq E_N \leq \sigma^2 (B B^\top + \epsilon I),
\]
which occurs with probability not less than \( 1 - \delta(\epsilon, N) \). This also implies (12).

Theorem 5 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 \( \epsilon \), 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 \( \theta^0 \) is assumed to yield a stable process, \( \{e_t\} \) is an i.i.d. Gaussian random process with variance \( \sigma^2 \), and \( \{y_t\} \) is stationary. Then,
\[
\mathbb{P}\left( |w^\top (\hat{\theta}_N - \theta^0)| > 2\sigma \|w\| V_{dn}^{-1/2} \sqrt{\log \left( \frac{\det(V_{up} V_{dn}^{-1} + I_n)^{1/2}}{\delta(\epsilon, N)} \right)} \right) \leq 2 \delta(\epsilon, N), \tag{14}
\]
where \( V_{dn}, V_{up} \) and \( \delta(\epsilon, N) \) are as described in Theorem 5.

**Proof.** We will follow the main ideas in (Sarkar and Rakhl, 2019, Theorem 1). We start by writing an upper bound using the Cauchy-Schwartz inequality
\[
|w^\top (\hat{\theta}_N - \theta^0)| = |w^\top (Y^\top Y)^{-1/2} Y^\top E| \leq \|w^\top (Y^\top Y)^{-1/2}\|_2 \|Y^\top Y\|^{1/2} E_2.
\]

In Theorem 5 we have found determinstic matrices \( V_{dn}, V_{up} \) and a scalar \( \delta(\epsilon, N) \) such that for the event \( \mathcal{E}_{pm} := \{V_{dn} \leq Y^\top Y \leq V_{up}\} \), we have \( \mathbb{P}(\mathcal{E}_{pm}) \geq 1 - \delta(\epsilon, N) \). The next step is to bound the self-normalized norm. This can be done by first defining the event
\[
\mathcal{E}_{sn} := \left\{ \|Y^\top Y\|_{(Y^\top Y + V_{dn})^{-1}} \leq \sqrt{2\sigma^2 \log \left( \frac{\det(Y^\top Y + V_{dn})^{1/2} \det(V_{dn})^{-1/2}}{\delta(\epsilon, N)} \right)} \right\}.
\]

It follows from Proposition 1 that \( \mathbb{P}(\mathcal{E}_{sn}) \geq 1 - \delta(\epsilon, N) \). Also, under \( \mathcal{E}_{pm} \) we have that \( Y^\top Y + V_{dn} \preceq 2Y^\top Y \), which implies \( (Y^\top Y + V_{dn})^{-1} \succeq \frac{1}{2}(Y^\top Y)^{-1} \). So, considering the set \( \mathcal{E}_{pm} \cap \mathcal{E}_{sn} \), we obtain
\[
\mathcal{E}_{pm} \cap \mathcal{E}_{sn} \implies \mathcal{E}_{pm} \cap \left\{ \|Y^\top Y\|_{(Y^\top Y + V_{dn})^{-1}} \leq 2\sigma \sqrt{\log \left( \frac{\det(V_{up} V_{dn}^{-1} + I_n)^{1/2}}{\delta(\epsilon, N)} \right)} \right\}.
\]
Furthermore, observe that $P(\mathcal{E}_{pm} \cap \mathcal{E}_{sn}) \geq 1 - 2\delta(\epsilon, N)$. So, if $\mathcal{E}_{pm} \cap \mathcal{E}_{sn}$ holds, then

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

(15)

which means that the probability of the event in (15) is at least $1 - 2\delta(\epsilon, N)$. By considering the complement event, we obtain the probability bound (14).

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 $\theta_i$ individually, as well as any linear combination of them. Note that $\epsilon$ can be considered a tightness variable, as by setting $\epsilon$ small, more samples are required to guarantee a desired confidence level, but the probability bound will be tighter.

**Remark 7** It is shown in Appendix B that if $\epsilon$ is picked as $\lambda_n - N^{-1/2}$, where $\lambda_n$ is the smallest eigenvalue of

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

then $\delta \sim C e^{-\lambda_n \sqrt{N}}$ for large $N$ and the deviation in (14) is asymptotically a constant in $N$. This suggests that the rate of decay of the probability bound is at least exponential in $\sqrt{N}$.

**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, 2018, 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 (14) 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 5. That is, if tighter bounds for $\mathcal{E}_{pm}$ can be found, then Theorem 6 can be directly improved. Future work concerns proving finite-time variance bounds for the estimated parameters, and extending the analysis for ARX models under sub-Gaussian noise.

**Acknowledgments**

This work was supported by the Swedish Research Council under contract number 2016-06079 (NewLEADS).
A Finite-Sample Deviation Bound for Stable Autoregressive Processes

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. Finite Sample Deviation and Variance Bounds for First Order Autoregressive Processes. *arXiv preprint arXiv:1910.08390*, 2019.

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. *arXiv preprint arXiv:1903.10343*, 2019.

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

A. Krishnamurthy, Z. S. Wu, and V. Syrgkanis. Semiparametric contextual bandits. *arXiv preprint arXiv:1803.04204*, 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*. preprint, 2018.

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 ed. 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.

D. S. Mitrinovic and P. M. Vasic. *Analytic inequalities*, volume 1. Springer, 1970.

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.
Appendix A. Proofs of Lemmas 2 and 3

Proof of Lemma 2. For simplicity we consider \( \bar{\epsilon} := \sigma^2 \epsilon \). We first see that

\[
\mathbb{P} \left( \rho \left[ \frac{1}{N-n} A(x_{n-1}x_{n-1}^\top - x_{N-1}x_{N-1}^\top)A^\top \right] > \frac{\bar{\epsilon}}{3} \right)
\leq \mathbb{P} \left( \rho \left[ \frac{1}{N-n} A(x_{n-1}x_{n-1}^\top + x_{N-1}x_{N-1}^\top)A^\top \right] > \frac{\bar{\epsilon}}{3} \right)
\leq \mathbb{P} \left( \|Ax_{n-1}\|_2^2 + \|Ax_{N-1}\|_2^2 > \frac{\bar{\epsilon}}{3} \right)
\leq \mathbb{P} \left( \|x_{n-1}\|_2^2 + \|x_{N-1}\|_2^2 > \frac{\bar{\epsilon}}{6} \right).
\]

By stationarity, the probabilities of the events

\( \mathcal{E}_n = \left\{ \frac{\|x_{n-1}\|_2^2}{N-n} > \frac{\bar{\epsilon}}{6} \right\} \), \( \mathcal{E}_N = \left\{ \frac{\|x_{N-1}\|_2^2}{N-n} > \frac{\bar{\epsilon}}{6} \right\} \),

are equal. For some fixed \( s > 0 \), we can bound \( \mathbb{P}(\mathcal{E}_n) \) by Chernoff’s inequality (Wainwright, 2019, Chap. 2) and later compute the moment generating function of the generated \( \chi^2 \) distribution:

\[
\mathbb{P}(\mathcal{E}_n) \leq \exp \left( -s \frac{\bar{\epsilon}}{6} \right) \mathbb{E} \left[ \exp \left( \frac{s}{N-n} \sum_{i=1}^{n-1} y_i^2 \right) \right] = \exp \left( -s \frac{\bar{\epsilon}}{6} \right) \prod_{i=1}^{n} \sqrt{1 - \frac{2s}{N-n} \lambda_i(R_n)}
\]

where \( R_n \in \mathcal{S}^n \) is the Toeplitz covariance matrix of \( [y_{-1} \cdots y_{n-1}] \), \( \lambda_i(R_n) \) is the \( i \)-th eigenvalue of the \( R_n \), and \( 0 < s < (N-n)/(2\lambda_{\max}(R_n)) \). By using the Weierstrass product inequality (stated as Proposition 11 in the Auxiliary Results section), we obtain

\[
\mathbb{P}(\mathcal{E}_n) \leq \exp \left( -s \frac{\bar{\epsilon}}{6} \right) \left( 1 - \frac{2sn\mathbb{E}\{y_1^2\}}{N-n} \right)^{-1/2}.
\]

By noting that

\[
\frac{N-n}{2\lambda_{\max}(R_n)} > \frac{N-n}{2\text{tr}(R_n)} > \frac{N-n}{4n\mathbb{E}\{y_1^2\}} > 0,
\]

we set \( s = \frac{(N-n)}{(4n\mathbb{E}\{y_1^2\})} \) to derive the exponential inequality

\[
\mathbb{P} \left( \frac{\|x_{n-1}\|_2^2}{N-n} > \frac{\bar{\epsilon}}{6} \right) \leq \sqrt{2} \exp \left( \frac{-(N-n)\bar{\epsilon}}{24n\mathbb{E}\{y_1^2\}} \right).
\]

So, by the chain of inequalities above, we conclude that

\[
\mathbb{P} \left( \rho \left[ \frac{1}{\sigma^2(N-n)} A(x_{n-1}x_{n-1}^\top - x_{N-1}x_{N-1}^\top)A^\top \right] > \frac{\epsilon}{3} \right) \leq 2\sqrt{2} \exp \left( \frac{-(N-n)\sigma^2\epsilon}{24n\mathbb{E}\{y_1^2\}} \right),
\]

which implies the statement we wanted to prove. \( \blacksquare \)
Proof of Lemma 3. Event $\mathcal{E}_2$ is equivalent to the event

$$\left\{ \left| \frac{1}{N-n} \sum_{i=n}^{N-2} e_i^2 - \sigma^2 \right| \leq \frac{\sigma^2 \epsilon}{3} \right\}.$$

To bound the probability of this event, we shall use a corollary of Lemma 1 of Laurent and Massart (2000) (Corollary 10 in the Auxiliary Results section), which gives high probability bounds on the tails of a $\chi^2$ statistic. Via this result and unnormalizing the $\chi^2$ statistic, we obtain that

$$\mathbb{P} \left( \frac{1}{N-n} \sum_{i=n}^{N-1} e_i^2 - \sigma^2 \geq 2\sigma^2 \sqrt{\frac{x}{N-n} + 2\sigma^2 \frac{x}{N-n}} \right) \leq e^{-x}$$

(16)

and

$$\mathbb{P} \left( \frac{1}{N-n} \sum_{i=n}^{N-1} e_i^2 - \sigma^2 \leq -2\sigma^2 \sqrt{\frac{x}{N-n} + 2\sigma^2 \frac{x}{N-n}} \right) \leq e^{-x}.$$

(17)

Equations (16) and (17) imply that

$$\mathbb{P} \left( -2\sigma^2 \sqrt{\frac{x}{N-n}} \leq \frac{1}{N-n} \sum_{i=n}^{N-1} e_i^2 - \sigma^2 \leq 2\sigma^2 \sqrt{\frac{x}{N-n} + 2\sigma^2 \frac{x}{N-n}} \right) \geq 1 - 2e^{-x}. \quad (18)$$

Next, we are interested in solving the following quadratic equation for positive $x$:

$$2\sigma^2 \sqrt{\frac{x}{N-n} + 2\sigma^2 \frac{x}{N-n}} = \frac{\sigma^2 \epsilon}{3} \implies \sqrt{\frac{x}{N-n}} = \frac{-1 + \sqrt{1 + \frac{2\epsilon}{3}}}{2}$$

$$\implies x = \frac{N-n}{2} \left( 1 + \frac{\epsilon}{3} - \sqrt{1 + \frac{2\epsilon}{3}} \right).$$

By plugging this value of $x$ in (18), we obtain the result.

Appendix B. On the decay rate of the bound

Here we analyze the rate of decay of the deviation bound given in Theorem 6. In particular, we study how fast $\delta$ decays to zero if the confidence bound is held constant.

Consider (14), and denote $\delta' := 2\delta$. Also, denote $\bar{V}_{dn}$ as $\frac{1}{N-n} \bar{V}_{dn}$, and

$$\epsilon := 2\sqrt{\log(2)\sigma \|w\|_2 \bar{V}_{dn}^{-1/2} \|V_{dn}^{-1/2} \|_2 \log \left( \frac{\det(V_{up} V_{dn}^{-1} + I_n)^{1/2}}{\delta'} \right)}.$$ 

First, we analyze $\bar{V}_{dn}$. We write $\bar{V}_{dn}$ as $\bar{V} - \epsilon \Gamma$, where $\bar{V} := [I_n \ 0\ 0] \bar{V} [I_n \ 0 \ 0]^\top$ and $\Gamma := \sigma^2 [I_n \ 0] \sum_{i=0}^{\infty} (A^\top)^i [I_n \ 0]^\top$. Next, let $W W^\top$ be the Cholesky factorization of $\bar{V}$, and let $U D U^\top$ be the eigenvalue decomposition of $W^{-1} \bar{V} W^{-\top}$, where $D = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$, and $\lambda_1 \geq \cdots \geq \lambda_n > 0$. Thus,

$$\bar{V}_{dn}^{-1} = W^{-\top} U \begin{bmatrix} \frac{1}{\lambda_1 - \epsilon} & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \frac{1}{\lambda_n - \epsilon} \end{bmatrix} U^\top W^{-1}.$$
Next, we choose $\epsilon$ such that $\epsilon = \lambda_n - N^{-1/2}$. This choice leads to
\[
\|w^\top \tilde{V}_{dn}^{-1/2}\|_2 = \sqrt{w^\top V_1 w + N^{1/2} w^\top V_2 w},
\]
where $V_1$ and $V_2$ are suitable matrices independent of $N$. In particular,
\[
V_2 = W^{-\top} U \begin{bmatrix} 0_{n-k} & 0 \\ I_k \end{bmatrix} U^\top W^{-1},
\]
with $k$ being the algebraic multiplicity of $\lambda_n$. Also, note that
\[
\det(V_n^{-1}) = \frac{N^{k/2}}{\det(\Gamma) \prod_{i=1}^{n-k}(\lambda_i - \epsilon)}.
\]
Regarding $\delta'$, with the choice of $\epsilon = \lambda_n - N^{-1/2}$ we find that $\delta' \sim C e^{-\lambda_n \sqrt{N}}$ for large $N$. Hence, (19) and (20) lead to the following computations:
\[
\begin{align*}
\epsilon &= 2\sqrt{\log(2)} \sigma \sqrt{\frac{w^\top V_1 w + N^{1/2} w^\top V_2 w}{(N-n)}} \sqrt{\log \left( \frac{\det(V_{up} + V_{dn})^{1/2}}{\det(V_{dn})^{1/2} \delta'} \right)} \\
&= 2\sqrt{\log(2)} \sigma \sqrt{\frac{w^\top V_1 w + N^{1/2} w^\top V_2 w}{(N-n)}} \sqrt{\log \left( \frac{\det \left( \begin{bmatrix} I_n & 0 \\ 0 & V \end{bmatrix} \right)^{1/2}}{\det(V_{dn})^{1/2} \delta'} \right)} \\
&\sim 2\sqrt{\log(2)} \sigma N^{-1/4} \sqrt{w^\top V_2 w} \sqrt{\frac{1}{2} \log \left( \frac{\det \left( \begin{bmatrix} I_n & 0 \\ 0 & V \end{bmatrix} \right)}{C^2 \det(\Gamma) \prod_{i=1}^{n-k}(\lambda_i - \epsilon)} \right) + \frac{k}{4} \log(N) + \lambda_n \sqrt{N}} \\
&= O(1) \text{ as } N \to \infty.
\end{align*}
\]
Hence, the rate of decay is at least exponential in $\sqrt{N}$. That is, for an asymptotically constant confidence bound, the probability of interest decays as $e^{-\lambda_n \sqrt{N}}$. Note that a byproduct of this derivation are the directions of $w$ in which a faster learning can be achieved: Since $V_2$ has rank $k$, a subspace of dimension $n - k$ of directions satisfies $w^\top V_2 w = 0$, for which the dominant term in $\|w^\top V_{dn}^{-1/2}\|_2$ becomes $N^{-1/2}$ instead of $N^{-1/4}$.

Appendix C. Auxiliary results

**Proposition 8 (Lemma 4.2 of Simchowitz et al. (2018))** Let $\{F_t\}_{t \geq 0}$ be a filtration, and $\{Z_t\}_{t \geq 1}$ and $\{W_t\}_{t \geq 1}$ be real-valued processes adapted to $F_t$ and $F_{t+1}$ respectively. Moreover, assume $W_t | F_t$ is mean zero and $\sigma^2$-sub-Gaussian. Then, for any positive real numbers $\alpha$, $\beta$ we have
\[
P \left[ \left\{ \sum_{t=1}^{T} Z_t W_t \geq \alpha \right\} \cap \left\{ \sum_{t=1}^{T} Z_t^2 \leq \beta \right\} \right] \leq \exp \left( -\frac{\alpha^2}{2\sigma^2\beta} \right).
Proposition 9 (Lemma 1 of Laurent and Massart (2000)) Let \((Y_1, \ldots, Y_D)\) be i.i.d. Gaussian variables, with mean 0 and variance 1. Let \(a_1, \ldots, a_D\) be nonnegative. We set

\[
\|a\|_\infty = \sup_{i=1,\ldots,D} |a_i|, \quad \|a\|_2^2 = \sum_{i=1}^D a_i^2.
\]

Let

\[
Z = \sum_{i=1}^D a_i(Y_i^2 - 1).
\]

Then, the following inequalities hold for any positive \(x\):

\[
\mathbb{P}(Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x) \leq \exp(-x),
\]

\[
\mathbb{P}(Z \leq -2\|a\|_2\sqrt{x}) \leq \exp(-x).
\]

Corollary 10 (Corollary of Lemma 1 of Laurent and Massart (2000)) Let \(U\) be a \(\chi^2\) statistic with \(D\) degrees of freedom. For any positive \(x\),

\[
\mathbb{P}(U - D \geq 2\sqrt{Dx} + 2x) \leq \exp(-x),
\]

\[
\mathbb{P}(U - D \leq -2\sqrt{Dx}) \leq \exp(-x).
\]

Proposition 11 (Weierstrass Product Inequality (Mitrinovic and Vasic, 1970, p. 210)) Given real numbers \(0 \leq \lambda_1, \lambda_2, \ldots, \lambda_n \leq 1\), the following inequality holds:

\[
(1 - \lambda_1)(1 - \lambda_2) \cdots (1 - \lambda_n) \geq 1 - \sum_{k=1}^n \lambda_k.
\]

Proof. We proceed by induction. For \(n = 1\), the inequality is obvious. Next, assume that (21) holds for some \(n\). Then,

\[
(1 - \lambda_1)(1 - \lambda_2) \cdots (1 - \lambda_n)(1 - \lambda_{n+1}) \geq \left(1 - \sum_{k=1}^n \lambda_k\right)(1 - \lambda_{n+1})
\]

\[
= 1 - \sum_{k=1}^n \lambda_k - \lambda_{n+1} + \lambda_{n+1}\sum_{k=1}^n \lambda_k
\]

\[
\geq 1 - \sum_{k=1}^{n+1} \lambda_k,
\]

which shows that (21) is valid for \(n + 1\) as well. By induction, the statement follows. 

Proposition 12 (Subadditivity of the spectral radius of Hermitian matrices (Bernstein, 2009, Fact 5.12.2)) Let \(A, B \in \mathbb{R}^{n \times n}\) be Hermitian matrices. Then,

\[
\rho(A + B) \leq \rho(A) + \rho(B).
\]
Proof. For this we will first prove the following chain of inequalities:

\[ \lambda_{\min}(A) + \lambda_{\min}(B) \leq \lambda_{\min}(A + B) \leq \lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B). \]  \hfill (23)

By definition,

\[ \lambda_{\min}(A) + \lambda_{\min}(B) = \min_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^\top Ax}{x^\top x} + \min_{y \in \mathbb{R}^n \setminus \{0\}} \frac{y^\top By}{y^\top y} \leq \min_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^\top (A + B)x}{x^\top x} = \lambda_{\min}(A + B). \]

Similarly,

\[ \lambda_{\max}(A + B) = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^\top (A + B)x}{x^\top x} \leq \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{x^\top Ax}{x^\top x} + \max_{y \in \mathbb{R}^n \setminus \{0\}} \frac{y^\top By}{y^\top y} = \lambda_{\max}(A) + \lambda_{\max}(B). \]

With (23) in mind, and using the fact that for any Hermitian matrix \( H \) we can write \( \rho(H) \) as \( \max\{ -\lambda_{\min}(H), \lambda_{\max}(H) \} \) (Bernstein, 2009, Fact 5.11.5), we see that

\[ \lambda_{\max}(A + B) \leq \max\{ -\lambda_{\min}(A), \lambda_{\max}(A) \} + \max\{ -\lambda_{\min}(B), \lambda_{\max}(B) \} = \rho(A) + \rho(B), \]  \hfill (24)

and

\[ -\lambda_{\min}(A + B) \leq \max\{ -\lambda_{\min}(A), \lambda_{\max}(A) \} + \max\{ -\lambda_{\min}(B), \lambda_{\max}(B) \} = \rho(A) + \rho(B), \]  \hfill (25)

Combining (24) and (25), we reach

\[ \rho(A + B) = \max\{ -\lambda_{\min}(A + B), \lambda_{\max}(A + B) \} \leq \rho(A) + \rho(B), \]

which is what we wanted to prove.