SIAM J. CONTROL OPTIM.
Vol. 63, No. 3, pp. 1902–1928

SIGNED-PERTURBED SUMS ESTIMATION OF ARX SYSTEMS: EXACT COVERAGE AND STRONG CONSISTENCY*

ALGO CARÈ[†], ERIK WEYER[‡], BALÁZS CS. CSÁJI[§], AND MARCO C. CAMPI[†]

Abstract. Sign-perturbed sums (SPS) is an identification method that constructs confidence regions for the unknown parameters of a system. In this paper, we consider a new version of SPS for application to autoregressive exogenous systems and establish that the ensuing confidence regions include the true parameters with exact, user-chosen, probability under mild statistical assumptions. This property holds true for any finite number of observed input-output data. Furthermore, the confidence regions are proven to be strongly consistent, that is, they shrink around the true parameters as the number of data points increases and, asymptotically, parameters different from the true ones are almost surely excluded from the regions.

Key words. system identification, confidence regions, finite sample properties, asymptotic properties, least squares

MSC codes. 93B30, 93E24

DOI. 10.1137/24M1646728

1. Introduction.

Estimating parameters of unknown systems based on noisy observations is a classical problem in system identification, as well as signal processing, machine learning, and statistics. Standard solutions such as the least-squares (LS) method or, more generally, prediction error methods provide *point estimates.* In many situations (for example, when the safety, stability, or quality of a process has to be guaranteed), a point estimate needs to be complemented by a *confidence region* that certifies the accuracy of the estimate and serves as a basis for securing robustness. If the noise is known to belong to a given bounded set, set membership approaches can be used to compute the region of the parameters that are consistent with the observed data; see, e.g., [46, 48, 37, 47, 53, 13, 34]. The need for deterministic priors on the noise can be relaxed by working in a probabilistic framework; see, e.g.,

*Received by the editors March 15, 2024; accepted for publication (in revised form) January 24, 2025; published electronically May 28, 2025. A preliminary version of the method presented in this paper was given in a conference paper presented at the 16th IFAC Symposium on System Identication, where a theorem on its finite-sample guarantees was proven under slightly stronger assumptions than those of Theorem 4.1 in this paper. The asymptotic analysis is novel and, in particular, the strong consistency theorem (Theorem 5.5) is stated and proven in this paper for the first time.

https://doi.org/10.1137/24M1646728

Funding: The work of the first author was supported in part by the European Research Consortium for Informatics and Mathematics (ERCIM). The work of the first and second authors was supported in part by the Australian Research Council (ARC) under grant DP130104028. The work of the third author was supported in part by the European Union within the framework of the National Laboratory for Autonomous Systems, grant RRF-2.3.1-21-2022-00002, and by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office (NRDIO), Hungary. The work of the fourth author was supported in part by the Ministry of University and Research (MUR), PRIN 2022 Project N.2022RRNAEX (CUP: D53D23001440006).

[†]Department of Information Engineering, University of Brescia, Brescia, Italy (algo.care@unibs.it, marco.campi@unibs.it).

[‡]Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia (ewey@unimelb.edu.au).

[§]Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), Budapest, Hungary, and Department of Probability Theory and Statistics, Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest, Hungary (balazs.csaji@sztaki.hun-ren.hu).

[32, 17]. However, traditional methods for the construction of confidence regions in a probabilistic setting rely on approximations based on asymptotic results and are valid only if the number of observed data points tends to infinity; see, e.g., [44]. In spite of the well-known fact that evaluating finite-sample estimates based on asymptotic results can lead to misleading conclusions [26], the study of the finite-sample properties of system identification algorithms has remained a niche research topic until recent times.

A notable (and now expanding) literature has been investigating how the quality of finite-sample estimates relates to specific characteristics of the system under consideration. Seminal works in this line, addressing in particular *finite impulse response* (FIR) and *autoregressive exogenous* (ARX) systems, are [74, 28, 75, 72, 66], which have been followed by several more recent studies for various classes of linear [50, 56, 33, 49, 61, 62, 54, 58] and nonlinear [67, 25, 55, 45] systems, also in relation to control frameworks [1, 5, 21, 24]. See [63] for a survey and more references. While confidence regions for the unknown parameters are easily obtained as side products of these investigations, these regions are typically conservative because their validity relies on uniform bounds and, moreover, they have rigid shapes, parametrized by characteristics of the system, which must be known to the user. On the other hand, the goal of producing nonconservative confidence regions by exploiting the observed dataset along more flexible approaches has been pursued by another, complementary research effort, a product of which is the *sign-perturbed sums* (SPS) algorithm, which forms the subject of this paper.

SPS was introduced in [14] with the aim of constructing finite-sample confidence regions for systems that have a known structure, but are otherwise completely unspecified, in a noise quasi distribution-free setup. SPS has connections with the bootstrap literature [20, 27, 22] and, in particular, with wild bootstrap with Rademacher signs [41, 19]. Bootstrap methods, however, either assume that the regressors are independent of the noises, or only provide asymptotic coverage guarantees. The reader is referred to [12] for a discussion on the relation between SPS and other finite-sample methods, such as the bootstrap-style perturbed dataset methods of [39] and the leave-out sign-dominant correlation regions (LSCR) method, a (more conservative) predecessor of SPS that was introduced in [9] and then extended to quite general classes of systems, and applied in a variety of contexts, in [18, 30, 2, 31].

SPS was studied from a computational point of view in [38] and extended to a distributed setup in [76]. Applications of SPS can be found in several domains, ranging from mechanical engineering [69, 70] and technical physics [23, 29, 68] to wireless sensor networks [8] and social sciences [60]. Moreover, the SPS idea constitutes a core technology of several recent algorithms, including techniques for state estimation [51], for the identification of state-space systems [3, 59], for error-in-variables systems [35, 36], and for kernel-based estimation [16, 4].

The theoretical analysis of SPS was conducted in [15] for linear regression models where the regressors are independent of the noise, which, in particular, applies to open-loop FIR systems. In this setting, it was shown that SPS provides exact confidence regions for the parameters and that the region includes the least-squares estimate (LSE). The main assumptions on the noise in [15] are that it forms an independent sequence and that its distribution is symmetric about zero; however, the distribution is otherwise unknown and it can be time-varying, even in each time-step.

In the first part of this paper, we extend the SPS method to cover ARX systems and show that it has the same finite-sample properties as the SPS for FIR systems. In the rest of the paper, we develop an asymptotic analysis of the extended SPS method. Although the characterizing property of SPS remains the finite-sample guarantees, its asymptotic properties are also of interest because they shed light on the capability of SPS to exploit the information carried by a growing amount of data points. Asymptotic properties, for example, play a role in the important problem of detecting model misspecifications; see [10]. The asymptotic analysis of the SPS algorithm presented in [15] was carried out in [73], but that analysis does not apply to the ARX case because of the existing correlation between the regressor vector and the system output. In this paper we show that also SPS for ARX systems is strongly consistent, in the sense that the confidence region shrinks around the true parameters and, asymptotically, all parameters different from the true ones are almost surely excluded from the region.

Structure of the paper. The paper is organized as follows. In the next section we introduce the problem setting and the LS method for ARX models is revisited. Then, in section 3, the SPS method for ARX systems is presented along with its fundamental finite-sample properties (Theorem 4.1). The strong consistency of the method is proved in section 5 (Theorem 5.5). A simple simulation example is given in section 6, and conclusions are drawn in section 7. The proofs of the theorems are postponed to the appendices. A preliminary version of the SPS algorithm for ARX systems was presented in [14], where a theorem on its finite-sample guarantees was proven under slightly stronger assumptions than those of Theorem 4.1 in this paper. The strong consistency theorem (Theorem 5.5) is stated and proven in this paper for the first time.

2. Problem setting.

2.1. Data generating system and problem formulation. The data generating ARX system is given by

$$Y_t + a_1^* Y_{t-1} + \cdots a_{n_a}^* Y_{t-n_a} = b_1^* U_{t-1} + \cdots b_{n_b}^* U_{t-n_b} + N_t, \tag{2.1}$$

where $Y_t \in \mathbb{R}$ is the output, $U_t \in \mathbb{R}$ the input, and $N_t \in \mathbb{R}$ the noise at time t. Equation (2.1) can be written in linear regression form as

$$Y_t = \varphi_t^{\mathrm{T}} \theta^* + N_t, \tag{2.2}$$

$$\varphi_t \triangleq [\, -Y_{t-1}, \ldots, -Y_{t-n_a}, U_{t-1}, \ldots, U_{t-n_b} \,]^{\mathrm{T}}, \tag{2.3}$$

$$\theta^* \triangleq [\, a_1^*, \ldots, a_{n_a}^*, b_1^*, \ldots, b_{n_b}^* \,]^{\mathrm{T}}. \tag{2.4}$$

Aim: construct a confidence region with a user-chosen coverage probability p for the true parameter θ^* from a finite sample of size n, that is, from the regressors $\varphi_1, \ldots, \varphi_n$ and the outputs $Y_1, \ldots, Y_n$.

We make the following two assumptions.

Assumption 2.1. θ^* is a deterministic vector, and the orders n_a and n_b are known.

Assumption 2.2. The initial conditions ($Y_0, \ldots, Y_{1-n_a}$ and $U_0, \ldots, U_{1-n_b}$ in φ_1) and the input sequence $U_1, \ldots, U_n$ are deterministic, and the stochastic noise sequence $N_1, \ldots, N_n$ is symmetrically distributed about zero (that is, for every $s_t \in \{1, -1\}$, $t = 1, 2, \ldots, n$, the joint probability distribution of $(s_1 N_1, \ldots, s_n N_n)$ is the same as that of $(N_1, \ldots, N_n)$) and is otherwise generic.

Assumption 2.2 is relatively "mild" in the following respects: (i) the specific distribution of $\{N_t\}$ is *not* assumed to be known; (ii) the independence of $\{N_t\}$ is *not* assumed, the value of $|N_{t+1}|$ can be a function of the past values of $|N_t|, |N_{t-1}|, \ldots$ (e.g., it could be a GARCH process [6] driven by symmetric innovations); (iii) the marginal distribution of N_t *can* be time-varying (that is, noise is not necessarily identically distributed). Notably, Assumption 2.2 covers the case of a heavy-tailed noise process $\{N_t\}$ (e.g., N_t could be Cauchy distributed, with undefined $\mathbb{E}[N_t]$ and $E[N_t^2]=\infty$). On the other hand, if $\{N_t\}$ has finite second moments, then Assumption 2.2 implies the standard assumptions that $\mathbb{E}[N_t]=0$ and that noises are uncorrelated, i.e., $\mathbb{E}[N_t N_{t-\tau}]=0$, for all $\tau \neq 0$. In fact, Assumption 2.2 implies that $\mathbb{E}[\,N_t \mid |N_t|\,]=\mathbb{E}[\,\mathrm{sign}(N_t) \mid |N_t|\,]\cdot|N_t|=0$ almost surely; hence, $\mathbb{E}[N_t]=0$. Similarly, conditioning on $|N_t|\cdot|N_{t-\tau}|$, it follows that $\mathbb{E}[N_t N_{t-\tau}]=0$. The assumption that the input is deterministic corresponds to an open-loop configuration. We also note that the results in the paper remain valid with some additional generality when $\{U_t\}$ is stochastic and the assumption that $N_1,\dots,N_n$ is symmetrically distributed about zero holds conditionally on $\{U_t\}$.

2.2. Review of the least-squares estimate. The SPS method, which will be introduced in section 3, is LSE-oriented, in the sense that the inclusion of a candidate parameter θ in the SPS confidence region depends on the *normal equation* arising in LS estimation, which we briefly review in this section. Let θ be a generic parameter

$$\theta=[\,a_1,\dots,a_{n_a},b_1,\dots,b_{n_b}\,]^{\mathrm{T}}, \tag{2.5}$$

and let $d=n_a+n_b$ be the number of elements in θ. Let the corresponding *predictor* be given by

$$\hat{Y}_t(\theta)\triangleq\varphi_t^{\mathrm{T}}\theta$$

and the *prediction errors* by

$$\hat{N}_t(\theta)\triangleq Y_t-\hat{Y}_t(\theta)=Y_t-\varphi_t^{\mathrm{T}}\theta. \tag{2.6}$$

The LSE is found by minimizing the sum of the squared prediction errors, that is,

$$\hat{\theta}_n\triangleq\operatorname*{arg\,min}_{\theta\in\mathbb{R}^d}\sum_{t=1}^{n}\hat{N}_t^2(\theta)=\operatorname*{arg\,min}_{\theta\in\mathbb{R}^d}\sum_{t=1}^{n}(Y_t-\varphi_t^{\mathrm{T}}\theta)^2. \tag{2.7}$$

The solution can be found by solving the *normal equation,*

$$\sum_{t=1}^{n}\varphi_t\,\hat{N}_t(\theta)=\sum_{t=1}^{n}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)=0, \tag{2.8}$$

which, when $\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}$ is invertible, has the (unique) solution

$$\hat{\theta}_n=\left(\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}\right)^{-1}\left(\sum_{t=1}^{n}\varphi_t Y_t\right). \tag{2.9}$$

3. Construction of an exact confidence region. The SPS method for ARX systems will be presented in section 3.2. To gradually introduce the main ideas behind it, we first review SPS in the simpler case of FIR systems, where

$$Y_t=b_1^*U_{t-1}+\cdots b_{n_b}^*U_{t-n_b}+N_t. \tag{3.1}$$

This is a simpler case, since the regressors $\varphi_t=[U_{t-1},\dots,U_{t-n_b}]^{\mathrm{T}}$ are independent of the noise sequence, which simplifies the analysis [15].

3.1. Review of SPS for noise-indepedendent regressors. The fundamental step of the original SPS algorithm [15] for (3.1) consists in generating $m-1$ *sign-perturbed sums* by randomly perturbing the signs of the prediction errors in the normal equation (2.8). The sums are given by

$$\begin{aligned} H_i(\theta) &= \sum_{t=1}^{n} \alpha_{i,t}\varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta) \\ &= \sum_{t=1}^{n} \alpha_{i,t}\varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \alpha_{i,t}\varphi_t N_t, \quad i = 1, \dots, m-1, \end{aligned}$$

where $\tilde{\theta} = \theta^* - \theta$, and $\{\alpha_{i,t}\}$ are random signs, i.e., independent and identically distributed (i.i.d.) random variables that take on the values ± 1 with probability $1/2$ each. For a given θ, the *reference sum* is the left-hand side of the *normal equation* (2.8), i.e.,

$$H_0(\theta) = \sum_{t=1}^{n} \varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta) = \sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t.$$

For $\theta = \theta^*$, these sums can be simplified to

$$\begin{aligned} H_0(\theta^*) &= \sum_{t=1}^{n} \varphi_t N_t, \\ H_i(\theta^*) &= \sum_{t=1}^{n} \alpha_{i,t}\varphi_t N_t = \sum_{t=1}^{n} \pm\varphi_t N_t, \end{aligned}$$

where in the last equation we have written $\pm$ instead of $\alpha_{i,t}$ for intuitive understanding. The crucial observation is that, since the regressors are independent of the noise, and the noise is jointly symmetric, it follows that $H_0(\theta^*)$ and $H_i(\theta^*)$ have the *same distribution,* and there is no reason why $\|H_0(\theta^*)\|^2$ ($\triangleq H_0(\theta^*)^{\mathrm{T}} H_0(\theta^*)$) should be bigger or smaller than any other $\|H_i(\theta^*)\|^2$, $i = 1, \dots, m-1$. In fact, in [15] it was proven that the probability that $\|H_0(\theta^*)\|^2$ is the kth largest one in the ordering of the m values $\{\|H_i(\theta^*)\|^2\}_{i=0}^{m-1}$ is exactly $1/m$, and the probability that it is among the q largest ones is $q \cdot \frac{1}{m}$. The SPS region with confidence $1 - \frac{q}{m}$ was then defined in [15] as the set of θ's such that $\|H_0(\theta)\|^2$ is *not* among the qth largest values in the ordering of $\{\|H_i(\theta^*)\|^2\}_{i=0}^{m-1}$. It can be noted that $\|H_0(\theta)\|^2$ is zero when θ is the LSE. Therefore, SPS is LS-driven by design.

Another crucial observation is the following. For "large enough" $\|\tilde{\theta}\|$, we will have

$$\left\| \sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t \right\|^2 > \left\| \sum_{t=1}^{n} \pm\varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \pm\varphi_t N_t \right\|^2,$$

with "high probability" since $\sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta}$ on the left-hand side increases faster than $\sum_{t=1}^{n} \pm\varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta}$ on the right-hand side. Hence, for $\|\tilde{\theta}\|$ large enough, $\|H_0(\theta)\|^2$ dominates in the ordering of $\{\|H_i(\theta)\|^2\}_{i=0}^{m-1}$, and values away from θ^* will therefore be excluded from the confidence region; see [73] for a detailed analysis.

3.2. Main idea behind SPS for ARX systems. In the ARX case, the idea illustrated in the above section cannot be applied directly: in fact, in the ARX case, the distribution of the unperturbed sequence $\{\varphi_t N_t\}$ is different from the distribution of the perturbed one, $\{\alpha_{i,t}\varphi_t N_t\}$, because φ_t depends on the unperturbed noise $\{N_t\}$. Therefore, the distribution of $H_0(\theta^*)$ is different from that of $H_i(\theta^*)$, $i=1,\ldots,m-1$. To get around this difficulty, the key idea in the proposed SPS algorithm for ARX systems is to generate regressors, denoted $\bar{\varphi}_{i,t}(\theta)$, such that $\{\varphi_t N_t\}$ and $\{\alpha_{i,t}\bar{\varphi}_{i,t}(\theta^*)N_t\}$ have the same distribution and use them to define modified versions of $H_0(\theta)$ and $H_i(\theta)$. The elements of $\bar{\varphi}_{i,t}(\theta)$ include, instead of the observed outputs, the outputs of the system corresponding to the parameter θ fed with the perturbed noise $\{\alpha_{i,t}N_t(\theta)\}$. More precisely, the perturbed output sequence $\bar{Y}_{i,1}(\theta),\ldots,\bar{Y}_{i,n}(\theta)$ is generated for every $\theta=[\,a_1,\ldots,a_{n_a},b_1,\ldots,b_{n_b}\,]^{\mathrm{T}}$ according to

$$\bar{Y}_{i,t}(\theta)+a_1\bar{Y}_{i,t-1}(\theta)+\cdots a_{n_a}\bar{Y}_{i,t-n_a}(\theta)\triangleq b_1U_{t-1}+\cdots b_{n_b}U_{t-n_b}+\alpha_{i,t}\hat{N}_t(\theta), \tag{3.2}$$

where $\hat{N}_t(\theta)$ is given by (2.6), and the initial conditions for $\bar{Y}_{i,t}(\theta)$ are $\bar{Y}_{i,t}(\theta)\triangleq Y_t$ for $1-n_a\le t\le 0$. The regenerated regressor is then given as

$$\bar{\varphi}_{i,t}(\theta)\triangleq[-\bar{Y}_{i,t-1}(\theta),\ldots,-\bar{Y}_{i,t-n_a}(\theta),U_{t-1},\ldots,U_{t-n_b}]^{\mathrm{T}}. \tag{3.3}$$

Using this perturbed regressor, the analogue of functions $H_0(\theta)$ and $H_i(\theta)$ defined above can be constructed, and, in what follows, we will denote them $S_0(\theta)$ and $S_i(\theta)$ to avoid confusion. With our definitions, for $\theta=\theta^*$, $S_0(\theta^*)$ and $S_i(\theta^*)$ have the same ordering property as $H_0(\theta^*)$ and $H_i(\theta^*)$ for FIR systems, and therefore the exact coverage probability of the constructed confidence regions can be proven.

3.2.1. SPS for ARX systems. The SPS method for ARX systems is now detailed in two distinct parts. The first, which is called "initialization," sets the main global parameters of SPS and generates the random objects needed for the construction. In the initialization, the user provides the desired confidence probability p. The second part evaluates an indicator function which decides whether or not a particular parameter value θ is included in the confidence region.

The pseudocode for the initialization and the indicator function is given in Tables 1 and 2, respectively. Essentially, the algorithm implements the idea presented in section 3.2, but a few clarifications are in order. In point 4 of Table 2, in the computation of $S_0(\theta)$ and $S_i(\theta)$, the vectors $\frac{1}{n}\sum_{t=1}^{n}\varphi_t\hat{N}_t(\theta)$ and $\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\,\bar{\varphi}_{i,t}(\theta)\hat{N}_t(\theta)$ have been premultiplied by the matrices $R_n^{-\frac{1}{2}}=(\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}})^{-\frac{1}{2}}$ and $R_{i,n}^{-\frac{1}{2}}(\theta)=(\frac{1}{n}\sum_{t=1}^{n}\bar{\varphi}_{i,t}(\theta)\bar{\varphi}_{i,t}^{\mathrm{T}}(\theta))^{-\frac{1}{2}}$. The reason is that this results in well-shaped confidence regions, as discussed in section 6. In point 3 in the initialization (Table 1) a random permutation π is drawn, which is used in the algorithm (precisely, in points 5 and 6 of Table 2) to unambigously decide the ordering among the values $\|S_0(\theta)\|^2,\|S_1(\theta)\|^2,\ldots,\|S_{m-1}(\theta)\|^2$ in the case of ties. Formally, given m real numbers $\{Z_i\}$, $i=0,\ldots,m-1$, the ordering in the algorithm is given by the strict total order $\succ_\pi$ defined as

$$\begin{gathered} Z_k\succ_\pi Z_j \text{ if and only if} \\ (Z_k>Z_j) \text{ or } (Z_k=Z_j \text{ and } \pi(k)>\pi(j)). \end{gathered} \tag{3.4}$$

The p-level *SPS confidence region* with $p=1-\frac{q}{m}$ is given as

$$\widehat{\Theta}_n\triangleq\left\{\,\theta\in\mathbb{R}^d:\text{SPS-INDICATOR}(\,\theta\,)=1\,\right\}.$$

Observe that the LS estimate, $\hat{\theta}_n$, has by definition the property that $S_0(\hat{\theta}_n)=0$. Therefore, the LSE is included in the SPS confidence region, except for the very

TABLE 1
Initialization of the SPS method

PSEUDOCODE: SPS-INITIALIZATION

1. Given a (rational) confidence probability $p \in (0,1)$, set integers $m > q > 0$ such that $p = 1 - q/m$;
2. Generate $n \cdot (m-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with
$$\mathbb{P}r\{\alpha_{i,t} = 1\} = \mathbb{P}r\{\alpha_{i,t} = -1\} = \frac{1}{2},$$
for $i \in \{1, \ldots, m-1\}$ and $t \in \{1, \ldots, n\}$;
3. Generate a random permutation π of the set $\{0, \ldots, m-1\}$, where each of the $m!$ possible permutations has the same probability $1/(m!)$ to be selected.

unlikely situation in which $m - q$ other $S_i(\theta)$ functions (besides $S_0(\theta)$) are null at $\hat{\theta}_n$ and ranked smaller than $S_0(\theta)$ by π.[1]

4. Exact confidence. Like its FIR counterpart, the SPS algorithm for the ARX system generates confidence regions that have *exact* confidence probabilities for any *finite* number of data points. The following theorem holds.

THEOREM 4.1. *Under Assumptions* 2.1 *and* 2.2*, the confidence region constructed by the SPS algorithm in Tables* 1 *and* 2 *has the property that* $\mathbb{P}r\{\theta^* \in \widehat{\Theta}_n\} = 1 - \frac{q}{m}$.

The probability in the statement of Theorem 4.1 is with respect to $\{N_t\}$ and the random elements in the initialization step, Table 1 (i.e., the random signs $\{\alpha_{i,t}\}$ and the random permutation π). The proof of the theorem can be found in Appendix A.1. The simulation examples in section 6 also demonstrate that, when the noise is stationary, the SPS confidence regions compare in size with the heuristic confidence regions obtained by applying the asymptotic system identification theory. However, unlike the asymptotic regions, the SPS regions are theoretically guaranteed for any finite n, and also maintain their guaranteed validity with nonstationary noise patterns.

5. Strong consistency. In addition to the probability of containing the true parameter, another important aspect is the size of the SPS confidence regions. In this section, under some additional mild assumptions, we prove a strong consistency theorem which guarantees that the SPS confidence sets shrink around the true parameter as the sample size increases, and eventually exclude any other parameters $\theta' \neq \theta^*$.

This theorem will be proved under the basic Assumptions 2.1 and 2.2, plus the assumptions discussed below.

5.1. Additional assumptions. Let $A(z^{-1};\theta) = 1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a}$ and $B(z^{-1};\theta) = b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}$, with z^{-1} the delay operator. (2.1) can be compactly rewritten as

$$A(z^{-1};\theta^*)Y_t \triangleq B(z^{-1};\theta)U_t + N_t$$

[1]It is worth mentioning some substantial differences between the construction here proposed and the one of [71]. In [71], extra data (the so-called instrumental variables) are assumed to be available to the user and are required to be correlated with $\{\varphi_t\}$ but independent of the noise $\{N_t\}$. Under this condition, the construction of [71] delivers guaranteed regions around the instrumental-variable estimate. On the other hand, the algorithm proposed in this paper does not require any extra data besides the regressors and the system outputs and is purely LSE-based.

Table 2
Evaluation of the SPS indicator function

Pseudocode: SPS-Indicator (θ)

1. For the given θ, compute the prediction errors for $t \in \{1,\dots,n\}$

$$\hat{N}_t(\theta) \triangleq Y_t - \varphi_t^{\mathrm{T}}\theta.$$

2. Build $m-1$ sequences of sign-perturbed prediction errors $(\alpha_{i,t}\,\hat{N}_t(\theta)), t = 1,\dots,n$.
3. Construct $m-1$ perturbed output trajectories $\bar{Y}_{i,1}(\theta),\dots,\bar{Y}_{i,n}(\theta)$, $i = 1,\dots,m-1$, according to (3.2) with $\bar{Y}_{i,t}(\theta) \triangleq Y_t$ for $1-n_a \le t \le 0$. Form $\bar{\varphi}_{i,t}(\theta)$ according to (3.3).
4. Evaluate

$$S_0(\theta) \triangleq R_n^{-\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}\varphi_t\hat{N}_t(\theta),$$

$$S_i(\theta) \triangleq R_{i,n}^{-\frac{1}{2}}(\theta)\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\,\bar{\varphi}_{i,t}(\theta)\hat{N}_t(\theta),$$

for $i \in \{1,\dots,m-1\}$, where

$$R_n \triangleq \frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}},$$

$$R_{i,n}(\theta) \triangleq \frac{1}{n}\sum_{t=1}^{n}\bar{\varphi}_{i,t}(\theta)\bar{\varphi}_{i,t}^{\mathrm{T}}(\theta),$$

and $^{-\frac{1}{2}}$ denotes the inverse (or pseudoinverse) of the principal square root matrix.
5. Order scalars $\{\|S_i(\theta)\|^2\}$ according to $\succ_\pi$ (see (3.4)).
6. Compute the rank $\mathcal{R}(\theta)$ of $\|S_0(\theta)\|^2$ in the ordering, where $\mathcal{R}(\theta) = 1$ if $\|S_0(\theta)\|^2$ is the smallest in the ordering, $\mathcal{R}(\theta) = 2$ if $\|S_0(\theta)\|^2$ is the second smallest, and so on. Mathematically, $\mathcal{R}(\theta) = 1 + |\{i = 1,\dots,m-1 : \|S_0(\theta)\|^2 \succ_\pi \|S_i(\theta)\|^2\}|$, where $|\cdot|$ denotes cardinality.
7. Return 1 if $\mathcal{R}(\theta) \le m-q$, otherwise return 0.

and (3.2) as

$$A(z^{-1};\theta)\bar{Y}_{i,t}(\theta) \triangleq B(z^{-1};\theta)U_t + \alpha_{i,t}\hat{N}_t(\theta).$$

The following assumption is a standard condition regarding identifiability of the true parameter.

Assumption 5.1 (coprimeness). The polynomials $A(z^{-1};\theta^*)$ and $B(z^{-1};\theta^*)$ are coprime (i.e., only constant polynomials are factors of both of them).

The set of values of θ that are allowed to be included in the confidence region is normally limited by a priori knowledge on the system and, in general, it will be a proper subset of $\mathbb{R}^d$. Although occasionally it can be left implicit, in this paper the subset of values of θ will be denoted by Θ_c and always assumed to be a compact set.

Assumption 5.2 (uniform stability). The families of filters $\{\frac{1}{A(z^{-1};\theta)} : \theta \in \Theta_c\}$ and $\{\frac{B(z^{-1};\theta)}{A(z^{-1};\theta)} : \theta \in \Theta_c\}$ are uniformly stable.

We recall the definition of uniform stability [44]. First, $\frac{1}{A(z^{-1};\theta)}$ and $\frac{B(z^{-1};\theta)}{A(z^{-1};\theta)}$ must be stable for every $\theta \in \Theta_c$. Then, we can define the coefficients $h_0(\theta), h_1(\theta), \dots$ from relation $\sum_{t=0}^{\infty} h_t(\theta) z^{-t} = \frac{1}{A(z^{-1};\theta)}$, and $g_1(\theta), g_2(\theta), \dots$ from relation $\sum_{t=1}^{\infty} g_t(\theta) z^{-t} = \frac{B(z^{-1};\theta)}{A(z^{-1};\theta)}$. Uniform stability means that

$$\sup_{\theta \in \Theta_c} |h_t(\theta)| \leq \bar{h}_t \qquad \text{and} \qquad \sup_{\theta \in \Theta_c} |g_t(\theta)| \leq \bar{g}_t, \ \forall t,$$

for some $\bar{h}_t$ and $\bar{g}_t$ such that

$$\sum_{t=0}^{\infty} \bar{h}_t < \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} \bar{g}_t < \infty.$$

Basically, Assumption 5.2 excludes that the dynamics of the system can be arbitrarily slow and that the static gain can be arbitrarily large.

The following type of conditions are standard for consistency analysis.

Assumption 5.3 (independent noise, moment growth rate). $\{N_t\}$ is a sequence of independent random variables. Moreover, the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[N_t^2] \tag{5.1}$$

exists and

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[N_t^8] < \infty. \tag{5.2}$$

We will assume that the input sequence is persistently exciting. Precisely, following [42], we say that the input sequence is persistently exciting of order $n_a + n_b$ if the limits $m = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} U_t$ (we call m the mean) and $c_{U,k} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} (U_t - m)(U_{t-k} - m)$ exist and are finite for every k, and the matrix

$$\begin{bmatrix} c_{U,0} & \cdots & c_{U,n_a+n_b-1} \\ \vdots & \ddots & \vdots \\ c_{U,n_a+n_b-1} & \cdots & c_{U,0} \end{bmatrix}$$

is positive definite.

Assumption 5.4 (persistent excitation and limited growth rate). The sequence $\{U_t\}$ is persistently exciting of order $n_a + n_b$. Moreover,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} U_t^4 < \infty. \tag{5.3}$$

The reader may be interested in comparing the realizationwise condition (5.3) and the processwise condition (5.2): in this regard, it is worth noting that the processwise condition (5.2) implies that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} N_t^4 < \infty \tag{5.4}$$

holds w.p.1.

In all that follows, we will consider Y_t to be the output of system (2.1) with zero initial conditions and causal input signals (N_t, U_t are zero for $t \leq 0$).

5.2. Result. In the following theorem, $B_\varepsilon(\theta^*)$ denotes the closed Euclidean norm-ball centered at θ^* with radius $\varepsilon > 0$, i.e.,

$$B_\varepsilon(\theta^*) \triangleq \{\,\theta \in \mathbb{R}^d : \|\,\theta - \theta^*\| \leq \varepsilon\,\}.$$

Theorem 5.5 states that the confidence regions $\widehat{\Theta}_n$ will eventually be included in any given norm-ball centered at the true parameter, θ^*.

THEOREM 5.5 (strong consistency). *Under Assumptions* 2.1*,* 2.2*,* 5.1*,* 5.2*,* 5.3*, and* 5.4*, for all* $\varepsilon > 0$*,*

$$\mathbb{P}r\left[\bigcup_{\bar{n}=1}^{\infty}\bigcap_{n=\bar{n}}^{\infty}\left\{\widehat{\Theta}_n \subseteq B_\varepsilon(\theta^*)\right\}\right] = 1.$$

A detailed proof of the theorem is provided in Appendix A.2, preceded by an outline.

6. Numerical example. Consider the following data generating system:

$$Y_t = -a^* Y_{t-1} + b^* U_{t-1} + N_t$$

with zero initial conditions. $a^* = -0.7$ and $b^* = 1$ are the true system parameters and $\{N_t\}$ is a sequence of i.i.d. Laplacian random variables with zero mean and variance 0.1. The input signal was generated according to

$$U_t = 0.75\,U_{t-1} + V_t,$$

where $\{V_t\}$ was a sequence of i.i.d. Gaussian random variables with zero mean and variance 1. The predictor is given by

$$\widehat{Y}_t(\theta) = -aY_{t-1} + bU_{t-1} = \varphi_t^{\mathrm{T}}\theta,$$

where $\theta = [\,a\,b\,]^{\mathrm{T}}$ is the model parameter, and $\varphi_t = [\,-Y_{t-1}\;U_{t-1}\,]^{\mathrm{T}}$ is the regressor at time t.

A 95% confidence region for $\theta^* = [a^*\;\,b^*]^{\mathrm{T}}$ based on $n = 40$ data points, namely (φ_t, Y_t), $t = 1,\dots,40$, was constructed by choosing $m = 100$ and leaving out those values of θ for which $\|S_0(\theta)\|$ was among the five largest values of $\|S_0(\theta)\|, \|S_1(\theta)\|, \dots, \|S_{99}(\theta)\|$.

The SPS confidence region is shown in Figure 1. In the same picture, the confidence ellipsoid based on asymptotic system identification theory is also shown, which is guaranteed to yield a 95% confidence region as $n \to \infty$ [44, Chapters 8–9].[2] It can be observed that the nonasymptotic SPS region is similar in size and shape to the asymptotic confidence region, but it has the advantage that it is guaranteed to contain the true parameter with exact probability 95% for finite values of n ($n = 40$ in this case).

In agreement with Theorem 5.5, the size of the region decreases when n is increased; see Figures 2 and 3. In Figure 3, m is also increased to 4000, and one can observe that there is very little difference between the SPS region and the asymptotic

[2]Precisely, the ellipsoid is computed as $\widetilde{\Theta}_n \triangleq \{\theta : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n\,(\theta - \hat{\theta}_n) \leq \frac{F_{\chi^2}^{-1}(0.95)\hat{\sigma}_n^2}{n}\}$, where $F_{\chi^2}^{-1}(\cdot)$ is the inverse cumulative distribution function of the chi-square distribution with $\dim(\theta^*) = 2$ degrees of freedom, and $\widehat{\sigma}_n^2 \triangleq \frac{1}{n-2}\sum_{t=1}^{n}(Y_t - \varphi_t^{\mathrm{T}}\hat{\theta}_n)^2$ is an estimate of the noise variance.

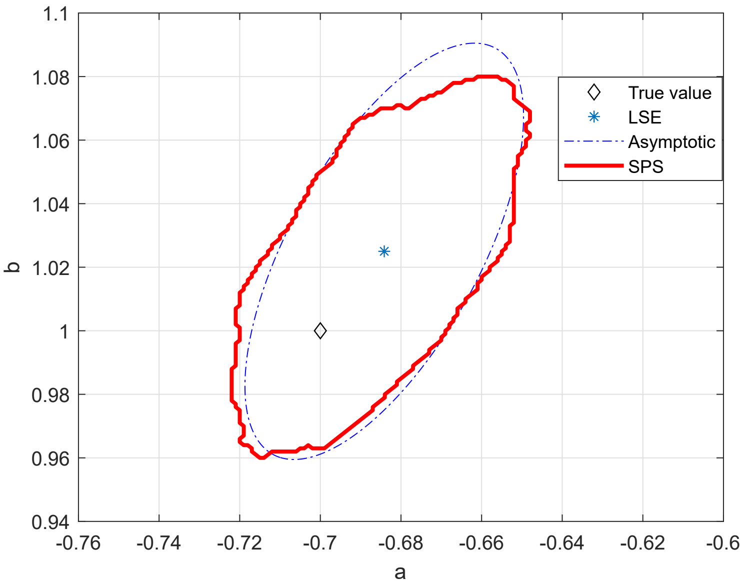

FIG. 1. 95% *confidence regions,* $n = 40$, $m = 100$.

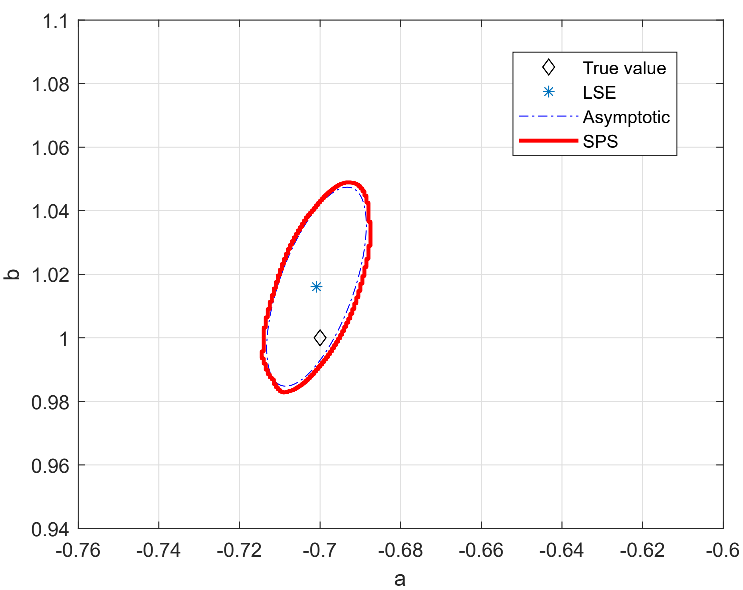

FIG. 2. 95% *confidence regions,* $n = 400$, $m = 100$.

confidence ellipsoid. This exemplifies a quite general behavior that can be informally expressed by saying that when n and m both tend to infinity the SPS confidence region is included in a marginally inflated version of the asymptotic confidence ellipsoids. This circumstance is interesting because it shows that SPS fills smoothly the traditional gap between finite-sample validity and asymptotic optimality. For the FIR case, a formal statement and a proof of this fact can be found in [73, Theorem 3]; an analogue theorem can be stated in our autoregressive setup. The proof,

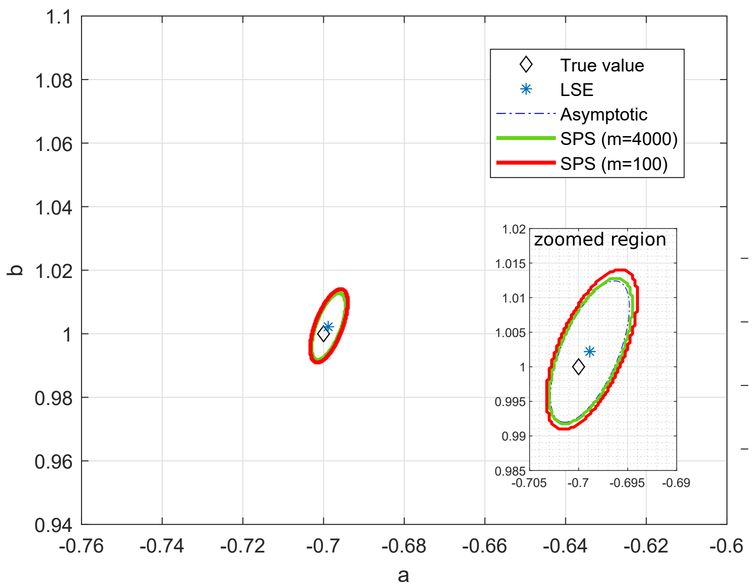

FIG. 3. 95% *confidence regions,* $n = 4000$*,* $m = 4000$*, and* $m = 100$*.*

however, is not trivial and has been omitted from this paper due to space limitations (the interested reader can find more details in the preprint at https://arxiv.org/abs/2402.11528).[3]

6.1. Coverage with asymmetric noise. In this paper, the symmetry of the noise has been assumed. However, it is interesting to evaluate whether the proposed algorithm is sensitive to small asymmetries. While a comprehensive study of the nonsymmetric case is beyond the scope of this work, some numerical evaluations were performed. In particular, the generation mechanism of N_t in the numerical example above was modified as follows: N_t was generated according to an asymmetric Laplace distribution [40] with density

$$f(x) = \begin{cases} \dfrac{\sqrt{2}}{\bar{\sigma}} \dfrac{\kappa}{1+\kappa^2} \exp\left(-\dfrac{\sqrt{2}}{\kappa\bar{\sigma}}(\mu - x)\right) & \text{if } x \leq 0, \\ \dfrac{\sqrt{2}}{\bar{\sigma}} \dfrac{\kappa}{1+\kappa^2} \exp\left(-\dfrac{\kappa\sqrt{2}}{\bar{\sigma}}(x - \mu)\right) & \text{if } x > 0. \end{cases}$$

Then, we evaluated the coverage of the SPS regions over 100,000 Monte Carlo runs for different values of κ. Table 3 reports the results for $\bar{\sigma} = \sqrt{0.1}$ and $\mu = 0$: in this way, when $\kappa = 1$, the distribution is the same as above (symmetric with variance $\sigma^2 = 0.1$) and the coverage is guaranteed to be 95%, while, when κ decreases, the

[3]Similarly as in [73], the result can be proven by using the central limit theorem to show that each $\|S_i(\theta)\|^2$ tends to be Chi-squared distributed as $n \to \infty$ and using the Glivenko–Cantelli theorem to show that the empirical distribution of $\|S_1(\theta)\|^2, \ldots, \|S_{m-1}(\theta)\|^2$ approximates the distribution of a Chi-squared as $m \to \infty$. However, from a technical point of view there is a crucial difference with respect to [73]: while in [73] functions $\|S_0(\theta)\|^2$ and $\|S_i(\theta)\|^2$, $i = 1, \ldots, m-1$, were all quadratic functions, in the present setting functions $\|S_i(\theta)\|^2$, $i = 1, \ldots, m-1$, are ratios of polynomials whose orders increase linearly with n, so that the uniform evaluation of these functions over θ as $n \to \infty$ requires much more attention.

TABLE 3
Coverages (100,000 Monte Carlo runs, $p=95\%$, $m=100$).

	$\kappa=1$	$\kappa=0.9$	$\kappa=0.8$	$\kappa=0.7$	$\kappa=0.5$	$\kappa=0.1$
$n=20$	0.95089	0.94216	0.91277	0.85345	0.67473	0.47360
$n=40$	0.95091	0.93882	0.89617	0.81580	0.64847	0.29411
$n=400$	0.95071	0.93248	0.87574	0.78354	0.58930	0
$n=4000$	0.95010	0.93094	0.84372	0.64998	0.05102	0

distribution becomes skewed toward positive values (the cases $\kappa = 0.5$ and $\kappa = 0.1$ correspond to very skewed distributions) and the coverage is no longer guaranteed.

In these simulations, the noise N_t was nonzero mean and nonzero median when $\kappa \neq 1$ (its mean was $\sigma\frac{1-\kappa^2}{\sqrt{2}\kappa}$ and its median was $-\frac{\sigma}{\kappa\sqrt{2}}\log\frac{1+\kappa^2}{2}$). Then, we modified the location parameter μ so as to generate zero-mean or zero-median noise samples. In these cases, the coverages turned out to be much better, suggesting that correctly estimating the mean or median of the noise is beneficial. For example, with $\kappa = 0.1$, when $n = 40$ we obtained a coverage of 0.94496 for the zero-mean case and 0.91797 for the zero-median case; when $n = 400$ we obtained 0.95003 and 0.80887, respectively. These results, being empirical, are not conclusive. However, they are indicative of the phenomenon that the algorithm exhibits a graceful degradation in the presence of asymmetries (even more so if the mean is correctly accounted for), and, importantly, they are in line with previous studies on the role of asymmetry that were performed for the original SPS [11]. We remark that some of the analyses and robustification techniques in [11] can be carried over to the present setup.

7. Concluding remarks and open problems.

In this paper, we have presented the SPS method for ARX systems. SPS delivers confidence regions around the least-squares estimate that contain with exact, user-chosen, probability the true system parameter under mild assumptions on the data generation system. These regions are built from a finite (and possibly small) sample of input-output data. Besides the exact finite-sample guarantees, we have proven under additional and rather mild assumptions that the method is strongly consistent. Moreover, the shape of the region in relation to the approximate confidence ellipsoids obtained using the asymptotic theory has been briefly discussed.

Finally, we want to mention some further directions of research. While the SPS regions have many desirable features, the exact calculation of the regions is computationally demanding. For FIR systems, an effective ellipsoidal outer approximation of the confidence regions can be practically computed by using convex programming techniques [15] (see also [38]). Obtaining similar results for the ARX system is a challenge of great practical importance. However, when the dimension of the parameter vector θ is small, the SPS region can be computed by checking whether points on a fine grid of the parameter space belong to the confidence set.

Appendix A. Proofs.

A.1. Proof of Theorem 4.1: Exact confidence.

We begin with a definition and two lemmas taken from [15].

DEFINITION A.1. *Let $Z_1, \dots, Z_k$ be a finite collection of random variables and $\succ_{t.o.}$ a strict total order. If for all permutations $i_1, \dots, i_k$ of indices $1, \dots, k$ we have*

$$\mathbb{P}r\{Z_{i_k} \succ_{t.o.} Z_{i_{k-1}} \succ_{t.o.} \dots \succ_{t.o.} Z_{i_1}\} = \frac{1}{k!},$$

then we call $\{Z_i\}$ *uniformly ordered w.r.t. order* $\succ_{t.o.}$.

LEMMA A.2. *Let* $\alpha, \beta_1, \dots, \beta_k$ *be i.i.d. random signs; then the random variables* $\alpha, \alpha \cdot \beta_1, \dots, \alpha \cdot \beta_k$ *are i.i.d. random signs.*

The following lemma highlights an important property of the $\succ_\pi$ relation that was introduced in section 3.

LEMMA A.3. *Let* $Z_1, \dots, Z_k$ *be real-valued, i.i.d. random variables. Then, they are uniformly ordered w.r.t.* $\succ_\pi$.

We are now ready to prove Theorem 4.1.

By construction, the parameter θ^* is in the confidence region if $\|S_0(\theta^*)\|^2$ takes one of the positions $1, \dots, m-q$ in the ascending order (w.r.t. $\succ_\pi$) of the variables $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$. We will prove that $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ are *uniformly ordered,* hence $\|S_0(\theta^*)\|^2$ takes each position in the ordering with probability $1/m$, thus its rank is at most $m-q$ with probability $1-q/m$.

Note that all the functions $S_i(\theta^*)$ depend on the sequence $\{\alpha_{i,t} N_t\}$ via the same function for all i, which we denote as $S(\alpha_{i,1}N_1, \dots, \dots, \alpha_{i,n}N_n) \triangleq S_i(\theta^*)$. This is true also for $S_0(\theta^*)$; in fact, recalling that $\alpha_{0,t} \triangleq 1$, $t \in \{1, \dots, n\}$, it holds that $\alpha_{0,t}\hat{N}_t(\theta^*) = \alpha_{0,t}N_t = N_t$, so $\bar{Y}_{0,t} = Y_t$ and $\bar{\varphi}_{0,t} = \varphi_t$.

Let $b_1, b_2, \dots$ be a sequence of random signs independent of $\{N_t\}$ and $\{\alpha_{i,t}\}$, and define $\sigma_t(N_t)$ as

$$\sigma_t(N_t) = \begin{cases} \operatorname{sign}(N_t) & \text{if } N_t \neq 0, \\ b_t & \text{if } N_t = 0. \end{cases}$$

Clearly, $N_t = \sigma_t(N_t) \cdot |N_t|$ for every value of N_t. By Assumption 2.2, $\{|N_t|\}$ and $\{\sigma_t(N_t)\}$ are independent, and $\{\sigma_t(N_t)\}$ is an i.i.d. sequence. Now, we will work conditioning on $\{|N_t|\}$ by exploiting the independence of $\{|N_t|\}$ from all the other random elements: let us fix a realization of $\{|N_t|\}$ and call it $\{v_t\}$ (all the other random elements are distributed according to their marginal distribution). Then, for all i and t, we introduce $\gamma_{i,t} = \alpha_{i,t}\sigma_t(N_t)$. $\{\alpha_{i,t}\}$, $i = 1, \dots, m-1$, are i.i.d. random signs independent of the other random elements and of $\sigma_t(N_1), \dots, \sigma_t(N_n)$ in particular. Using Lemma A.2, $\gamma_{i,t}$, $i = 0, \dots, m-1$, $t = 1, \dots, n$, are i.i.d. random signs. Thus, $S_i(\theta^*)$ can be equivalently expressed as $S_i(\theta^*) = Z_i$, where $Z_i \triangleq S(\gamma_{i,1}v_1, \dots, \gamma_{i,n}v_n)$. Since Z_i's are obtained by applying the same function to different realizations of an i.i.d. sample, they are also uniformly ordered with respect to $\succ_\pi$ (Lemma A.3). Thus, the uniform ordering property has been proven for a fixed realization of $\{|N_t|\}$. As the realization of $\{|N_t|\}$ was arbitrary, the uniform ordering property of $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ holds unconditionally, and the theorem follows.

A.2. Proof of Theorem 5.5.

A.2.1. Outline of the proof. We define $\hat{\theta}_{i,n}(\theta)$ as the value of $\hat{\theta} \in \mathbb{R}^d$ that minimizes

$$\sum_{i=1}^{n} (\bar{Y}_{i,t}(\theta) - \bar{\varphi}_{i,t}(\theta)^{\mathrm{T}}\hat{\theta})^2, \tag{A.1}$$

i.e., as the LSE if the output sequence were $\bar{Y}_{i,1}(\theta),\dots,\bar{Y}_{i,n}(\theta)$; cf. (2.7).[4] $\hat{\theta}_{i,n}(\theta)$ satisfies

$$\frac{1}{n}\sum_{t=1}^{n}\bar{\varphi}_{i,t}(\theta)\bar{\varphi}_{i,t}(\theta)^{\mathrm{T}}(\hat{\theta}_{i,n}(\theta)-\theta)=\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta). \tag{A.2}$$

Assuming $\hat{\theta}_{i,n}(\theta)$ is unique (we will show that this is the case for n large enough), it is straightforward to check that $\|S_i(\theta)\|^2$ can be written as

$$\|S_i(\theta)\|^2=\|R_{i,n}(\theta)^{\frac{1}{2}}(\theta-\hat{\theta}_{i,n}(\theta))\|^2, \qquad i=1,\dots,m-1, \tag{A.3}$$

where $R_{i,n}(\theta)=\frac{1}{n}\sum_{t=1}^{n}\bar{\varphi}_{i,t}(\theta)\bar{\varphi}_{i,t}^{\mathrm{T}}(\theta)$ (as defined in Table 2). Similarly, $\|S_0(\theta)\|^2$ can be rewritten as

$$\|S_0(\theta)\|^2=\|R_n^{\frac{1}{2}}(\theta-\hat{\theta}_n)\|^2, \tag{A.4}$$

where $\hat{\theta}_n$ is the LSE (2.9), and $R_n=\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}$ (as defined in Table 2). First, we prove that $\hat{\theta}_n\to\theta^*$ w.p.1, and hence that $\|S_0(\theta)\|^2$ eventually stays away from zero outside a ball centered at θ^*. The second step is proving the uniform convergence of $\hat{\theta}_{i,n}(\theta)$ to θ.[5] To do this, we first prove that $R_{i,n}(\theta)$, $i=1,\dots,m-1$, converge uniformly in Θ_c to a matrix function $\bar{R}(\theta)$ that is positive definite, with eigenvalues that are uniformly bounded away from 0 and from ∞. Second, we show that the right-hand side of (A.2) goes to zero uniformly in Θ_c w.p.1.[6] Combining these two facts, we will conclude that $\hat{\theta}_{i,n}(\theta)$ converges to θ, and $\|S_i(\theta)\|^2\to 0$ uniformly. This implies that, for n large enough, $\|S_i(\theta)\|^2$, $i=1,\dots,m-1$, are smaller than $\|S_0(\theta)\|^2$ for all the values of $\theta\in\Theta_c$ outside a small ball centered at θ^*, so that such values of θ are excluded from the confidence region.

A.2.2. Proof. The following two lemmas are the key results to prove the theorem. In the statements of these two lemmas, the assumptions of Theorem 5.5 are left implicit. Their proofs are in Appendix A.2.3.

LEMMA A.4. *The limit matrix*

$$\bar{R}_*\triangleq\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}$$

exists and is finite w.p.1. Moreover, there exists a matrix $\bar{R}(\theta)$, independent of $\{N_t\}$ and $\{\alpha_{i,t}\}$, function of $\theta\in\Theta_c$, such that $\lim_{n\to\infty}\sup_{\theta\in\Theta_c}\|R_{i,n}(\theta)-\bar{R}(\theta)\|=0$, $i=1,\dots,m-1$, w.p.1. $\bar{R}(\theta)$ is continuous in $\theta\in\Theta_c$, $\bar{R}(\theta^*)=\bar{R}_*$, and there exist $\rho_1,\rho_2>0$ such that $I\rho_1\prec\bar{R}(\theta)\prec I\rho_2$ for all $\theta\in\Theta_c$.[7]

[4]In the terminology of [39] this is the minimizer of the cost function corresponding to the "perturbed dataset."

[5]This requires some caution because $\bar{Y}_{i,1}(\theta),\dots,\bar{Y}_{i,n}(\theta)$ is the output of the nonstandard system (3.2), where U_t affects future noise terms through $\hat{N}_{t+1}(\theta),\hat{N}_{t+2}(\theta),\dots$, and the expected value of $\hat{N}_{t+1}^2(\theta)$ given the past is not uniformly bounded. Thus, traditional consistency results such as those in [43], although quite general and inclusive of closed-loop setups, do not apply to this setting.

[6]This step is carried out by using martingale arguments that are inspired by the proof in [43], together with a suitable "conditioning trick."

[7]The symbol "$\prec$" denotes the Loewner partial ordering, i.e., given two matrices A and B, $A\prec B\iff B-A$ is positive definite.

LEMMA A.5. *It holds w.p.*1 *that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} N_t\varphi_t = 0.$$

Moreover, for every $i = 1, \dots, m-1$,

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_c} \left| \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta) \right| = 0$$

*holds true w.p.*1.

We first study the asymptotic behavior of the reference function $\|S_0(\theta)\|^2$.

By definition, the LS estimate $\hat{\theta}_n$ must satisfy the normal equation (see (2.8))

$$\frac{1}{n}\sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}(\hat{\theta}_n - \theta^*) = \frac{1}{n}\sum_{t=1}^{n} N_t\varphi_t. \tag{A.5}$$

The convergence (a.s.) of $\hat{\theta}_n$ to θ^* follows by taking the norm of both the right- and left-hand sides of (A.5) and noting that the right-hand side goes to zero by Lemma A.5. On the other hand, because of $\bar{R}_* = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}} \succ 0$ (Lemma A.4), the left-hand side goes to zero as $n\to\infty$ if and only if $\hat{\theta}_n$ converges to θ^*. Thus,

$$\|\hat{\theta}_n - \theta^*\| \underset{n\to\infty}{\longrightarrow} 0 \text{ w.p.1.} \tag{A.6}$$

Using (A.4), we conclude that

$$\|S_0(\theta)\|^2 \underset{n\to\infty}{\longrightarrow} \|\bar{R}_*^{\frac{1}{2}}(\theta^* - \theta)\|^2 \text{ (uniformly in } \Theta_c) \text{ w.p.1.} \tag{A.7}$$

Now we study the asymptotic behavior of the functions $\|S_i(\theta)\|^2$, $i = 1, \dots, m-1$.

By definition, $\hat{\theta}_{i,n}(\theta)$ satisfies (A.2). By taking the norm of both sides of (A.2) and by using Lemma A.5 we get $\lim_{n\to\infty}\sup_{\theta\in\Theta_c}\|R_{i,n}(\theta)(\hat{\theta}_{i,n}(\theta)-\theta)\|^2 = 0$ w.p.1, while, by Lemma A.4, we have $\sup_{\theta\in\Theta_c}\|R_{i,n}(\theta)(\hat{\theta}_{i,n}(\theta)-\theta)\|^2 \geq \rho_1^2\cdot\|\hat{\theta}_{i,n}(\theta)-\theta\|^2$ for all $\theta\in\Theta_c$, for n large enough. These two facts yield

$$\lim_{n\to\infty}\sup_{\theta\in\Theta_c}\|\hat{\theta}_{i,n}(\theta) - \theta\|^2 = 0 \text{ w.p.1.} \tag{A.8}$$

Using (A.7) and Lemma A.4, we conclude that there exists w.p.1 a (realization dependent) $\bar{n}_0$ such that

$$\|S_0(\theta)\|^2 > \rho_1\epsilon^2 \qquad \forall\theta : ||\theta - \theta^*|| > \epsilon$$

for every $n > \bar{n}_0$. W.p.1, there also exists a (realization dependent) $\bar{n}$ large enough such that, for every $n > \bar{n}$, $R_{i,n}(\theta) \prec I\rho_2$, for all $\theta\in\Theta_c$ and $i = 1, \dots, m-1$ (Lemma A.4), and such that $\|\hat{\theta}_{i,n}(\theta) - \theta\|^2 < \frac{\rho_1\epsilon^2}{\rho_2}$, for all $\theta\in\Theta_c$ and $i = 1, \dots, m-1$ (A.8), which implies

$$\|S_i(\theta)\|^2 < \rho_1\epsilon^2 \qquad \forall\theta\in\Theta_c, \quad i = 1, \dots, m-1.$$

Therefore, for every realization on a set of probability 1, there exist (realization dependent) $\bar{n}_0$ and $\bar{n}$ such that for every $n > \max(\bar{n}_0, \bar{n})$ it holds that $\|S_0(\theta)\| > \|S_i(\theta)\|$, $i = 1, \dots, m-1$, for every $\theta \notin B_\epsilon(\theta^*)$, and this implies the theorem statement.

A.2.3. Proofs of Lemmas A.4 and A.5. Preliminarily, we state some asymptotic results that are useful throughout. In all the lemmas stated in this proof, the assumptions of Theorem 5.5 are left implicit.

Lemma A.6. *W.p.*1 *it holds that*

1.a $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}N_t=0$,

1.b $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}N_tN_{t-k}=\delta_k\cdot\left(\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[N_t^2]\right)<\infty$, *where* $\delta_k=0$ *for every* $k\neq 0$ *and* $\delta_0=1$,

1.c $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}N_tU_{t-k}=0$ *for every* k,

1.d $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}N_tY_{t-k}=0$ *for every* $k\geq 1$.

For every $k\in\mathbb{Z}$, *there exist* $c_{Y,k}<\infty$ *and* $c_{YU,k}<\infty$ *such that, w.p.*1,

2.a $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}Y_tY_{t-k}=c_{Y,k}$ *for every* k,

2.b $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}Y_tU_{t-k}=c_{YU,k}$ *for every* k.

*W.p.*1 *it holds that*

3.a $\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Y_t^4<\infty$,

3.b $\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\|\varphi_t\|^4<\infty$,

3.c $\sup_{\theta\in\Theta_c}(\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\hat{N}_t^4(\theta))<\infty$,

3.d $\sup_{\theta\in\Theta_c}\left(\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\|\bar{\varphi}_{i,t}(\theta)\|^4\right)\leq C<\infty$, *where* C *depends on* $\{N_t\}$ *but not on* $\{\alpha_{i,t}\}$.

For every $\theta\in\Theta_c$ *and* $k\in\mathbb{Z}$, *there exist* $c_{\bar{Y},k}(\theta)<\infty$ *and* $c_{\bar{Y}U,k}(\theta)<\infty$ *such that, w.p.*1,

4.a $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\bar{Y}_{i,t}(\theta)\bar{Y}_{i,t-k}(\theta)=c_{\bar{Y},k}(\theta)$ *for every* k *and every* $i=1,\dots,m-1$,

4.b $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\bar{Y}_{i,t}(\theta)U_{t-k}=c_{\bar{Y}U,k}(\theta)$ *for every* k *and every* $i=1,\dots,m-1$.

Proof. [1.a, 1.b, 1.c] We prove 1.b, since 1.a is easier.

For every $k\neq 0$, $\mathbb{E}[N_tN_{t-k}]=0$. Moreover, by applying twice the Cauchy–Schwarz inequality (once to $\mathbb{E}[\cdot]$ and once to $\sum_{t=1}^{n}\cdot$), we get $\limsup_{n\to\infty}\sum_{t=1}^{n}\frac{\mathbb{E}[(N_tN_{t-k})^2]}{t^2}\leq\limsup_{n\to\infty}\sum_{t=1}^{n}\sqrt{\frac{\mathbb{E}[N_t^4]}{t^2}}\sqrt{\frac{\mathbb{E}[N_{t-k}^4]}{t^2}}\leq\limsup_{n\to\infty}\sqrt{\sum_{t=1}^{n}\frac{\mathbb{E}[N_t^4]}{t^2}}\sqrt{\sum_{t=1}^{n}\frac{\mathbb{E}[N_{t-k}^4]}{t^2}}<\infty$ by Assumption 5.3, (5.2), and the result follows from the Kolmogorov's strong law of large numbers (Theorem B.1 in Appendix B). The case $k=0$ and 1.c can be proven similarly.

[1.d] By using the expression $Y_{t-k}=\sum_{\tau=0}^{\infty}h_\tau N_{t-k-\tau}+\sum_{\tau=1}^{\infty}g_\tau U_{t-k-\tau}$, we write $\frac{1}{n}\sum_{t=1}^{n}N_tY_{t-k}=\frac{1}{n}\sum_{t=1}^{n}N_t(\sum_{\tau=0}^{n}h_\tau N_{t-k-\tau}+\sum_{\tau=1}^{n}g_\tau U_{t-k-\tau})=\sum_{\tau=0}^{n}h_\tau(\frac{1}{n}\sum_{t=1}^{n}N_tN_{t-k-\tau})+\sum_{\tau=1}^{n}g_\tau(\frac{1}{n}\sum_{t=1}^{n}N_tU_{t-k-\tau})$. We focus on the first term; the second one can be dealt with similarly. Using the fact that $N_t=0$ for all $t\leq 0$, the Cauchy–Schwarz inequality yields $\sup_{\tau=1,2,\dots}\frac{1}{n}|\sum_{t=1}^{n}N_tN_{t-k-\tau}|\leq\frac{1}{n}\sum_{t=1}^{n}N_t^2$. Define $C=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}N_t^2$. Fix $\epsilon>0$. By stability, it is possible to choose M such that for each $n\geq M$, $\sum_{\tau=M+1}^{\infty}|h_\tau|<\frac{\epsilon}{2C}$. Thus, $|\sum_{\tau=0}^{n}h_\tau(\frac{1}{n}\sum_{t=1}^{n}N_tN_{t-k-\tau})|\leq|\sum_{\tau=0}^{M}h_\tau(\frac{1}{n}\sum_{t=1}^{n}N_tN_{t-k-\tau})|+\frac{\epsilon}{2}$, which can be made $<\epsilon$ by taking n large enough, because $\max_{\tau=1,\dots,M}|\frac{1}{n}(\sum_{t=1}^{n}N_tN_{t-k-\tau})|$ is the max over a set of finite (M) terms that all go to zero in virtue of 1.b.

[2.a, 2.b] The proof of 2.a is similar to 2.b, so we focus on 2.a. Consider $k\geq 0$, otherwise replace t with $t'=t-k$, and use the same argument. Rewrite $\frac{1}{n}\sum_{t=1}^{n}Y_tY_{t-k}$ as (it is intended that $g_\tau=0$ for $\tau\leq 0$)

$$\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{\tau=0}^{n}(h_\tau N_{t-\tau}+g_\tau U_{t-\tau})\right)\left(\sum_{\ell=0}^{n}(h_\ell N_{t-k-\ell}+g_\ell U_{t-k-\ell})\right)$$
$$=\sum_{\tau=0}^{n}\sum_{\ell=0}^{n}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right)+\sum_{\tau=0}^{n}\sum_{\ell=0}^{n}h_\tau g_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}U_{t-k-\ell}\right)$$
$$+\sum_{\tau=0}^{n}\sum_{\ell=0}^{n}g_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}U_{t-\tau}N_{t-k-\ell}\right)+\sum_{\tau=0}^{n}\sum_{\ell=0}^{n}g_\tau g_\ell\left(\frac{1}{n}\sum_{t=1}^{n}U_{t-\tau}U_{t-k-\ell}\right).$$

All of these terms can be dealt with similarly, so we focus on the first one. $\sum_{\tau=0}^{n}\sum_{\ell=0}^{n} h_\tau h_\ell(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell})$, for $M<n$, can be rewritten as

$$\sum_{\tau=0}^{M}\sum_{\ell=0}^{M}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right)+\sum_{\tau=M+1}^{n}\sum_{\ell=0}^{M}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right)$$
$$+\sum_{\tau=0}^{M}\sum_{\ell=M+1}^{n}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right)+\sum_{\tau=M+1}^{n}\sum_{\ell=M+1}^{n}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right).$$

By virtue of $\sup_{\tau,\ell=0,\dots,n}|\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}|\le\frac{1}{n}\sum_{t=1}^{n}N_t^2$, which converges to a constant as n grows to ∞, and by virtue of the stability of the system, the limit as $n\to\infty$ of all the terms except for the first one can be made arbitrarily close to zero if M is chosen large enough. We are left to deal with the *truncated* sum $\lim_{n\to\infty}\sum_{\tau=0}^{M}\sum_{\ell=0}^{M}h_\tau h_\ell(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell})$, which is Cauchy in M because of the stability of the system, and therefore can be made arbitrarily close to $\lim_{n\to\infty}\sum_{\tau=0}^{n}\sum_{\ell=0}^{n}h_\tau h_\ell(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell})$. More precisely, its argument can be further decomposed as

$$\sum_{\tau=0,\dots,M-k}h_\tau h_{k+\tau}\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}^2\right)+\sum_{\tau=0,\dots,M;\ell=0,\dots,M;\ell\neq k+\tau}h_\tau h_\ell\left(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}N_{t-k-\ell}\right).$$

The limit for $n\to\infty$ of the second term goes to zero because of Lemma A.6 (1.b) applied to a finite number of choices of τ and ℓ, while $\lim_{n\to\infty}\sum_{\tau=0,\dots,M-k}h_\tau h_{k+\tau}(\frac{1}{n}\sum_{t=1}^{n}N_{t-\tau}^2)=c_0\sum_{\tau=0,\dots,M-k}h_\tau h_{k+\tau}$ does not depend on the specific $\{N_t\}$.

[3.a, 3.b, 3.c, 3.d] The sequence $\{Y_t\}$ can be written as the sum of two convolutions, i.e., $\{Y_t\}=(\{N_t\}*\{h_t(\theta^*)\})+(\{U_t\}*\{g_t(\theta^*)\})$, where the t'th sample of the first convolution is $(\{N_t\}*\{h_t(\theta^*)\})_{t'}=\sum_{\tau=0}^{\infty}N_{t'-\tau}h_\tau(\theta^*)$, and the t'th sample of the second convolution is $(\{U_t\}*\{g_t(\theta^*)\})_{t'}=\sum_{\tau=1}^{\infty}U_{t'-\tau}g_\tau(\theta^*)$. Let $\mathbb{1}(P)$ denote the indicator function that is equal to 1 when proposition P is true and is 0 otherwise. For every t and k, define $N_{t|k}\triangleq N_t\cdot\mathbb{1}(t\le k)$, and, similarly, $U_{t|k}\triangleq U_t\cdot\mathbb{1}(t\le k)$, $Y_{t|k}\triangleq Y_t\cdot\mathbb{1}(t\le k)$. Clearly, for every fixed n,

$$\begin{aligned}\|\{Y_{t|n}\}\|_4&=\|(\{N_{t|n}\}*\{h_t(\theta^*)\})+(\{U_{t|n}\}*\{g_t(\theta^*)\})\|_4\\&\le\|(\{N_{t|n}\}*\{h_t(\theta^*)\})\|_4+\|(\{U_{t|n}\}*\{g_t(\theta^*)\})\|_4.\end{aligned}\tag{A.9}$$

Using Young's convolution inequality for sequences (see, e.g., [7, p. 315])

$$\begin{aligned}\|(\{N_{t|n}\}*\{h_t(\theta^*)\})\|_4&\le\|\{N_{t|n}\}\|_4\cdot\|\{h_t(\theta^*)\}\|_1\\&\le\left(\sum_{t=1}^{n}N_t^4\right)^{1/4}\cdot\left(\sum_{t=0}^{\infty}|h_t(\theta^*)|\right),\end{aligned}$$

and similarly for the input term. Due to the stability assumption, $\|\{h_t(\theta^*)\}\|_1 \leq C' < \infty$ and $\|\{g_t(\theta^*)\}\|_1 \leq C'' < \infty$. Hence, we get

$$\frac{1}{n}\sum_{t=1}^{n} Y_t^4 \leq 8C'^4 \frac{1}{n}\sum_{t=1}^{n} N_t^4 + 8C''^4 \frac{1}{n}\sum_{t=1}^{n} U_t^4, \tag{A.10}$$

and, from (5.4) and (5.3), we conclude that $\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Y_t^4 < \infty$ w.p.1.

Inequality

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \|\varphi_t\|_2^4 < \infty \tag{A.11}$$

immediately follows from

$$\|\varphi_t\|_2 \leq \sum_{k=1}^{n_a} |Y_{t-k}| + \sum_{k=1}^{n_b} |U_{t-k}|. \tag{A.12}$$

Moreover, $|\hat{N}_t(\theta)| = |N_t + \varphi_t^{\mathrm{T}}(\theta^* - \theta)| \leq |N_t| + \|\varphi_t\|_2 \cdot \|\theta^* - \theta\|_2 \leq |N_t| + \|\varphi_t\|_2 \cdot \sup_{\theta\in\Theta_c} \|\theta^* - \theta\|_2$. Here, $\sup_{\theta\in\Theta_c} \|\theta^* - \theta\|_2$ is finite because Θ_c is compact and we can conclude that

$$\sup_{\theta\in\Theta_c} \left(\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \hat{N}_t^4(\theta) \right) < \infty. \tag{A.13}$$

The same reasoning that led to (A.10) and (A.11) can be applied to $\{\bar{Y}_t(\theta)\} = (\{\alpha_{i,t}\hat{N}_t(\theta)\} * \{h_t(\theta)\}) + (\{U_t\} * \{g_t(\theta)\})$, and, noting that $\sup_{\theta\in\Theta_c} \sum_{t=1}^{\infty} |h_t(\theta)| \leq K' < \infty$ and $\sup_{\theta\in\Theta_c} \sum_{t=1}^{\infty} |g_t(\theta)| \leq K'' < \infty$ by Assumption 5.2, we immediately get

$$\sup_{\theta\in\Theta_c} \left(\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \|\bar{\varphi}_{i,t}(\theta)\|_2^4 \right) < \infty, \tag{A.14}$$

where the finite bound does not depend on the sequence $\{\alpha_{i,t}\}$.

[4.a, 4.b] Writing $\bar{Y}_{i,t}(\theta) = \sum_{\tau=0}^{n} h_\tau(\theta)\alpha_{i,t}\hat{N}_{t-\tau}(\theta) + \sum_{\tau=1}^{n} g_\tau(\theta)U_{t-\tau} = \sum_{\tau=0}^{n} (h_\tau(\theta)\alpha_{i,t}N_{t-\tau}) + \sum_{\tau=0}^{n} h_\tau(\theta)\alpha_{i,t}\varphi_{t-\tau}^{\mathrm{T}}(\theta^* - \theta) + \sum_{\tau=1}^{n} g_\tau(\theta)U_{t-\tau}$, where $\sum_{\tau=0}^{n}(h_\tau(\theta)\alpha_{i,t}\varphi_{t-\tau}^{\mathrm{T}}(\theta^* - \theta)) = \sum_{\tau=0}^{n} h_\tau(\theta)\alpha_{i,t}(\sum_{\ell=1}^{n_a} Y_{t-\tau-\ell}(a_\ell^* - a_\ell) + \sum_{\ell'=1}^{n_b} U_{t-\tau-\ell'}(b_{\ell'}^* - b_{\ell'}))$, we observe that, modulo the presence of random signs, most of the terms involved in this sum are the same as those encountered in the proofs of results 2.a and 2.b, and they can be dealt with similarly. The term $\sum_{\tau=0}^{n} h_\tau(\theta)\alpha_{i,t}\sum_{\ell=1}^{n_a} Y_{t-\tau-\ell}(a_\ell^* - a_\ell)$ requires some extra care as it gives rise to cross-terms of the kind $(a_\ell^* - a_\ell)^2 \sum_{\tau=0}^{n}\sum_{\lambda=0}^{n} h_\tau(\theta)h_\lambda(\theta)\frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t-\tau-\ell}Y_{t-\tau-\ell}\alpha_{i,t-k-\lambda-\ell}Y_{t-k-\lambda-\ell}$. These terms can be dealt with by conditioning on a fixed sequence $\{N_t\}$; in fact, conditionally on $\{N_t\}$, the sequence $\{\alpha_{i,t-\tau-\ell}Y_{t-\tau-\ell}\alpha_{i,t-k-\lambda-\ell}Y_{t-k-\lambda-\ell}\}_{t=1}^{\infty}$ is independent so that Kolmogorov's strong law of large numbers (Theorem B.1 in Appendix B) applies. In this way, we can conclude that, when $\tau \neq k + \lambda$, $\frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t-\tau-\ell}Y_{t-\tau-\ell}\alpha_{i,t-k-\ell}Y_{t-k-\ell}$ goes to zero w.p.1, while the case $\tau = k + \lambda$ reduces to 2.a. ∎

The following lemma ensures that there is some continuity (on average) in the behavior of $\bar{Y}_{i,1}(\theta), \ldots, \bar{Y}_{i,n}(\theta)$ as θ varies in Θ_c.

LEMMA A.7. *For every $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$\limsup_{n\to\infty} \sup_{\theta_1,\theta_2\in\Theta_c;\theta_1\in B_\delta(\theta_2)} \frac{1}{n}\sum_{t=1}^{n} |\bar{Y}_t(\theta_1) - \bar{Y}_t(\theta_2)|^2 < \epsilon \quad w.p.1.$$

Proof. The proof follows along the same lines as the proof of Lemma A.6, 3.a, 3.b, 3.c, by writing, for each n (and i), $\{\bar{Y}_{i,t|n}(\theta_1)-\bar{Y}_{i,t|n}(\theta_2)\}=\{\bar{Y}_{i,t|n}(\theta_1)\}-\{\bar{Y}_{i,t|n}(\theta_2)\}=\{\alpha_{i,t}\hat{N}_{t|n}(\theta_1)\}*\{h_t(\theta_1)\}-\{\alpha_{i,t}\hat{N}_{t|n}(\theta_2)\}*\{h_t(\theta_2)\}+\{U_{t|n}\}*\{g_t(\theta_1)-g_t(\theta_2)\}=\{\alpha_{i,t}N_{t|n}+\alpha_{i,t}\varphi_{t|n}^{\mathrm{T}}(\theta^*-\theta_2+[\theta_2-\theta_1])\}*\{h_t(\theta_1)\}+\{\alpha_{i,t}N_{t|n}+\alpha_{i,t}\varphi_{t|n}^{\mathrm{T}}(\theta^*-\theta_2)\}*\{h_t(\theta_2)\}+\{U_{t|n}\}*\{g_t(\theta_1)-g_t(\theta_2)\}$. Using the notation $\Delta\theta\triangleq\theta_1-\theta_2$, $\Delta f\triangleq f(\theta_1)-f(\theta_2)$ for a generic function f, we can write

$$
\begin{aligned}
\|\{\Delta\bar{Y}_{i,t|n}\}\|_2 &\le \|\{\alpha_{i,t}N_{t|n}\}*\{\Delta h_t\}\|_2+\|\{\alpha_{i,t}\varphi_{t|n}^{\mathrm{T}}(\theta^*-\theta_2)\}*\{\Delta h_t\}\|_2\\
&\qquad+\|\{\alpha_{i,t}\varphi_{t|n}^{\mathrm{T}}\Delta\theta\}*\{h_t(\theta_1)\}\|_2+\|\{U_{t|n}\}*\Delta g_t\|_2\\
&\le (\text{Young's inequality})\\
&\le \|\{N_{t|n}\}\|_2\cdot\|\{\Delta h_t\}\|_1+(\|\theta^*-\theta_2\|_2\cdot\|\{\|\varphi_{t|n}\|_2\}\|_2)\cdot\|\{\Delta h_t\}\|_1\\
&\qquad+\|\Delta\theta\|_2\cdot\|\{\|\varphi_{t|n}\|_2\}\|_2\cdot\|\{h_t(\theta_1)\}\|_1+\|\{U_{t|n}\}\|_2\cdot\|\{\Delta g_t\}\|_1, \qquad \text{(A.15)}
\end{aligned}
$$

which is a finite quantity in view of Assumption 5.2. Denoting $\sup_{\theta_1,\theta_2\in\Theta_c;\theta_1\in B_\delta(\theta_2)}$ for short as $\sup_{\|\Delta\theta\|<\delta}$, we have $\sup_{\|\Delta\theta\|<\delta}\|\Delta h_t\|_1\le 2\sum_{t=0}^{\infty}\sup_{\theta\in\Theta_c}|h_t(\theta)|<\infty$ and $\sup_{\|\Delta\theta\|<\delta}\|\Delta g_t\|_1\le 2\sum_{t=1}^{\infty}\sup_{\theta\in\Theta_c}|g_t(\theta)|<\infty$. From (A.15), using (5.4) and (5.3), Assumption 5.2, and Lemma A.6 (3.b), it follows that w.p.1 there are (possibly realization dependent) constants C_1, C_2, C_3, C_4 such that

$$
\begin{aligned}
&\limsup_{n\to\infty}\sup_{\|\Delta\theta\|<\delta}\sqrt{\frac{1}{n}}\|\{\Delta\bar{Y}_{i,t|n}\}\|_2\\
&\le C_1\sup_{\|\Delta\theta\|<\delta}\|\{\Delta h_t\}\|_1+C_2\sup_{\|\Delta\theta\|<\delta}\|\{\Delta h_t\}\|_1+\delta\cdot C_3+C_4\sup_{\|\Delta\theta\|<\delta}\|\{\Delta g_t\}\|_1<\infty.
\end{aligned}
$$

Moreover, $\sup_{\|\Delta\theta\|<\delta}\|\{\Delta h_t\}\|_1$ can be made arbitrarily small for δ small enough because $\sup_{\|\Delta\theta\|<\delta}\|\{\Delta h_t\}\|_1\le\sum_{t=0}^{\infty}\sup_{\|\Delta\theta\|<\delta}|\Delta h_t|$ and the following proposition holds.

PROPOSITION A.8. $\sum_{t=0}^{\infty}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t|$ *can be made arbitrarily small for a positive* δ' *small enough.*

Proof. First, write

$$
\begin{aligned}
\sum_{t=0}^{\infty}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t| &= \sum_{t=0}^{M-1}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t|+\sum_{t=M}^{\infty}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t|\\
&\le \sum_{t=0}^{M-1}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t|+2\sum_{t=M}^{\infty}\sup_{\theta\in\Theta_c}|h_t(\theta)|,
\end{aligned}
$$

and note that for any ϵ' we can choose an $M>0$ large enough, such that

$$
\sum_{t=M}^{\infty}\sup_{\theta\in\Theta_c}|h_t(\theta)|<\frac{\epsilon'}{4}.
$$

Now we prove that there exists $\delta'>0$ such that $\sum_{t=0}^{M-1}\sup_{\|\Delta\theta\|<\delta'}|\Delta h_t|<\frac{\epsilon'}{2}$. By Assumption 5.2, the tth coefficient $h_t(\theta)$ of the Laurent series $\sum_{t=0}^{\infty}h_t(\theta)z^{-t}=\frac{1}{A(\theta;z^{-1})}$ can be written as $h_t(\theta)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{A(\theta;e^{-\iota\omega})}e^{\iota\omega t}\mathrm{d}\omega$ (ι denotes the imaginary unit); see, e.g., [52]. This implies that $|h_t(\theta_1)-h_t(\theta_2)|\le\frac{1}{2\pi}\int_{-\pi}^{\pi}|\frac{A(\theta_1-\theta_2;e^{-\iota\omega})-1}{A(\theta_1;e^{-\iota\omega})A(\theta_2;e^{-\iota\omega})}|\mathrm{d}\omega\le\frac{1}{2\pi}K^2\int_{-\pi}^{\pi}|A(\theta_1-\theta_2;e^{-\iota\omega})-1|\mathrm{d}\omega$, where $K\triangleq\sup_{\theta\in\Theta_c,\omega\in[-\pi,\pi]}\frac{1}{|A(\theta;e^{-\iota\omega})|}$. Note that K is finite by Assumption 5.2; in fact, by Assumption 5.2 there exists a finite K' such that for all θ and ω it holds that $K' > \sum_{t=0}^{\infty} |h_t(\theta)| \geq |\sum_{t=0}^{\infty} h_t(\theta) e^{\iota\omega t}| = |\frac{1}{A(\theta; e^{-\iota\omega})}|$. Since $|A(\theta-\theta'; e^{-\iota\omega}) - 1| = |(a_1 - a_1')e^{-\iota\omega} + (a_2 - a_2')e^{-\iota 2\omega} + \cdots + (a_{n_a} - a_{n_a}')e^{-\iota n_a \omega}| \leq n_a \|\theta - \theta'\|_1 \leq n_a^{\frac{3}{2}} \|\theta - \theta'\|_2$, the result follows by choosing $\delta' < n_a^{-\frac{3}{2}} \frac{\pi \epsilon'}{M \cdot K^2}$. ∎

The same argument holds for $\sup_{\|\Delta\theta\| < \delta} \|\Delta g_t\|_1$, and from this the theorem statement follows. ∎

Proof of Lemma A.4. The limit $\bar{R}_* = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$ exists and is finite by Lemma A.6 (2.a, 2.b). The persistent excitation condition on $\{U_t\}$ (Assumption 5.4), together with the fact that polynomials $A(\theta^*; z^{-1})$ and $B(\theta^*; z^{-1})$ are of known orders (Assumption 2.1) and coprime (Assumption 5.1), entails that $\bar{R}_*$ is positive definite; see, e.g., [65, Lemma 10.3], and [42].

From Lemma A.6, 4.a and 4.b, it follows that for each θ the limit matrix $\bar{R}(\theta)$ exists and is independent of i and of the realizations of $\{N_t\}$ and $\{\alpha_{i,t}\}$. When $\theta = \theta^*$, the perturbed output generated by (3.2) is statistically equivalent to the original output, so that $\bar{R}(\theta^*) = \bar{R}_*$.

Let θ_0 be an arbitrary element of Θ_c; we use the notation Δf to denote the difference $f(\theta) - f(\theta_0)$. We first show that

$$\forall \epsilon > 0 \, \exists \delta > 0 \text{ s.t. } \limsup_{n\to\infty} \sup_{\theta,\theta_0 : \theta \in B_\delta(\theta_0)} \|\Delta R_{i,n}\| < \epsilon, \tag{A.16}$$

where the domain Θ_c is implicitly assumed, and it will be omitted in what follows. To prove (A.16), we focus on the matrix $\Delta R_{i,n}$ entry by entry and we study the limiting behavior of entries of the kind Δr_n, where $r_n(\theta) \triangleq \frac{1}{n} \sum_{t=1}^{n} \bar{Y}_{i,t-\ell}(\theta) \bar{Y}_{i,t-\tau}(\theta)$, for some ℓ, τ between 1 and n_a, while other entries in $\Delta R_{i,n}$ that involve U_t can be dealt with similarly. Write $|r_n(\theta) - r_n(\theta_0)| =$

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{t=1}^{n} \bar{Y}_{i,t-\ell}(\theta_0) \Delta \bar{Y}_{i,t-\tau} + \frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\ell} \bar{Y}_{i,t-\tau}(\theta_0) + \frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\ell} \Delta \bar{Y}_{i,t-\tau} \right| \\
&\leq \sqrt{\frac{1}{n} \sum_{t=1}^{n} \bar{Y}_{i,t-\ell}^2(\theta_0)} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\tau}^2} + \sqrt{\frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\ell}^2} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \bar{Y}_{i,t-\tau}^2(\theta_0)} \\
&\quad + \sqrt{\frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\ell}^2} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \Delta \bar{Y}_{i,t-\tau}^2}.
\end{aligned}$$

By taking the $\sup_{\theta,\theta_0:\theta\in B_\delta(\theta_0)}$ on both sides, it is immediate from Lemma A.6, 4.a, and Lemma A.7 that $\sup_{\theta,\theta_0:\theta\in B_\delta(\theta_0)} |r_n(\theta) - r_n(\theta_0)|$ can be made arbitrarily small for every n large enough by choosing δ small enough and (A.16) is established. Since

$$\begin{aligned}
\sup_{\theta,\theta_0:\theta\in B_\delta(\theta_0)} \|\Delta \bar{R}\| &= \sup_{\theta,\theta_0:\theta\in B_\delta(\theta_0)} \limsup_{n\to\infty} \|\Delta R_{i,n}\| \\
&\leq \limsup_{n\to\infty} \sup_{\theta,\theta_0:\theta\in B_\delta(\theta_0)} \|\Delta R_{i,n}\|,
\end{aligned}$$

(A.16) entails uniform continuity of $\bar{R}(\theta)$ over Θ_c, and therefore there exists a finite $\rho_2 > 0$ such that $\bar{R}(\theta) \prec \rho_2 I$ for all $\theta \in \Theta_c$. As for the uniform convergence of $R_{i,n}(\theta)$ to $\bar{R}(\theta)$, we can find a $\delta > 0$ and a finite number M_δ of δ-balls $B_\delta(\theta_0^{(1)}), \ldots, B_\delta(\theta_0^{(M_\delta)})$ that cover Θ_c and are such that, for all n large enough, it holds true that (i) $\max_{j=1,\ldots,M_\delta} \sup_{\theta\in B_\delta(\theta_0^{(j)})} \|R_{i,n}(\theta) - R_{i,n}(\theta_0^{(j)})\| < \frac{\epsilon}{3M_\delta}$ (in view of (A.16)), (ii) $\max_{j=1,\ldots,M_\delta} \|R_{i,n}(\theta_0^{(j)}) - \bar{R}(\theta_0^{(j)})\| < \frac{\epsilon}{3M_\delta}$ (in view of pointwise convergence at the ball centers), and (iii) $\max_{j=1,\dots,M_\delta}\sup_{\theta\in B_\delta(\theta_0^{(j)})}\|\bar R(\theta_0^{(j)})-\bar R(\theta)\|<\frac{\epsilon}{3M_\delta}$ (in view of uniform continuity of $\bar R(\theta)$). Then, for any n large enough, $\sup_\theta\|R_{i,n}(\theta)-\bar R(\theta)\|\le \sum_{j=1}^{M_\delta}\sup_{\theta\in B_\delta(\theta_0^{(j)})}\|R_{i,n}(\theta)-\bar R(\theta)\| \le \sum_{j=1}^{M_\delta}\sup_{\theta\in B_\delta(\theta_0^{(j)})}(\|R_{i,n}(\theta)-R_{i,n}(\theta_0^{(j)})\|+\|R_{i,n}(\theta_0^{(j)})-\bar R(\theta_0^{(j)})\|+\|\bar R(\theta_0^{(j)})-\bar R(\theta)\|)\le\sum_{j=1}^{M_\delta}(\frac{\epsilon}{3M_\delta}+\frac{\epsilon}{3M_\delta}+\frac{\epsilon}{3M_\delta})=\epsilon$.

To see that $\bar R(\theta)\succ I\rho_1$ for all $\theta\in\Theta_c$, recall that $\{\bar Y_{i,t}(\theta)\}=(\{\alpha_{i,t}\hat N_t(\theta)\}*\{h_t(\theta)\})+(\{U_t\}*\{g_t(\theta)\})$, where $\{U_t\}$ is persistently exciting of order n_a+n_b (Assumption 5.4). Any realization of $\{\alpha_{i,t}\hat N_t(\theta)\}$ (in a set of probability 1) is "uncorrelated" with $\{U_t\}$ in the sense that $\lim_{n\to\infty}\frac1n\sum_{t=1}^n\alpha_{i,t}\hat N_t(\theta)U_{t-\tau}=0$ for every τ; moreover, $\{\alpha_{i,t}\hat N_t(\theta)\}$ is persistently exciting of every order in the sense of [42] for every $\theta\ne\theta^*$.[8] Applying standard results on identifiability (e.g., Lemma 10.2 in [65]) it follows immediately that $\bar R(\theta)$ is invertible for every $\theta\in\Theta_c\setminus\{\theta^*\}$. We knew already that $\bar R(\theta^*)=\bar R_*\succ 0$ so that, by continuity of $\bar R(\theta)$, we can conclude that $\bar R(\theta)\succ I\rho_1$ over the whole Θ_c for some $\rho_1>0$.

Proof of Lemma A.5. The first statement follows from Lemma A.6 (1.c, 1.d). As for the second statement, we first prove pointwise convergence, i.e., we prove that for all $\theta\in\Theta_c$,

$$\lim_{n\to\infty}\frac1n\sum_{t=1}^n\alpha_{i,t}\hat N_t(\theta)\bar\varphi_{i,t}(\theta)=0 \text{ w.p.1.}$$

We work conditioning on a sequence $\{N_t\}$, i.e., we fix a realization of the noise, which we recall is independent of the sign-sequences $\{\alpha_{i,t}\}$, $i=1,\dots,m-1$. Therefore, in what follows, all the probabilities and expected values are with respect to the random sign-sequences $\{\alpha_{i,t}\}$, $i=1,\dots,m-1$, only. Since the result that we prove holds conditionally on any realization $\{N_t\}$ in a set of probability 1, then it holds unconditionally w.p.1. For a fixed $\theta\in\Theta$ (and i), define

$$\begin{cases} z_{i,t}(\theta)=z_{i,t-1}(\theta)+\dfrac1t\alpha_{i,t}\hat N_t(\theta)\bar\varphi_{i,t}(\theta),\\ z_{i,0}=0.\end{cases}$$

We aim at showing that each component of $z_{i,n}(\theta)$ is a martingale with bounded variance. From this, convergence of $\frac1n\sum_{t=1}^n\alpha_{i,t}\hat N_t(\theta)\bar\varphi_{i,t}(\theta)$ to zero as $n\to\infty$ will be easily proved.

Clearly, $\mathbb{E}[|z_{i,t}|]<\infty$ for all t. Denote by $\mathcal{A}_t$ the σ-algebra generated by the sequence $\{\alpha_{i,t}\}$ until time t, i.e., by $\alpha_{i,1},\dots,\alpha_{i,t}$. Since $\mathbb{E}[z_{i,t+1}|\mathcal{A}_t]=z_t$, the sequence $\{z_{i,t}^{(j)}\}$ formed by the jth component of the vector $z_{i,t}$ is a martingale. Moreover, $\mathbb{E}[z_{i,t}^{(j)}z_{i,t-1}^{(j)}|\mathcal{A}_{t-1}]=(z_{i,t-1}^{(j)})^2+\mathbb{E}[\alpha_{i,t}|\mathcal{A}_{t-1}]\cdot\frac1t(N_t+\varphi_t^{\mathrm T}(\theta^*-\theta))\bar\varphi_{i,t}(\theta)^{(j)}z_{i,t-1}^{(j)}=(z_{i,t-1}^{(j)})^2$, from which the useful identity

$$\mathbb{E}[(z_{i,t}^{(j)}-z_{i,t-1}^{(j)})^2|\mathcal{A}_{t-1}]=\mathbb{E}[(z_{i,t}^{(j)})^2-(z_{i,t-1}^{(j)})^2|\mathcal{A}_{t-1}] \tag{A.17}$$

follows. Thus,

$$\mathbb{E}[(z_{i,t}^{(j)})^2]=\sum_{k=1}^t\left(\mathbb{E}[(z_{i,k}^{(j)})^2]-\mathbb{E}[(z_{i,k-1}^{(j)})^2]\right) \tag{A.18}$$

[8]Proving these claims is easy if we fix a realization of $\{N_t\}$, and only the signs are left random; then it is just a matter of checking that the conditions for the Kolmogorov's strong law of large numbers (Theorem B.1) are met by the conditionally independent sequences $\{\alpha_t\hat N_t(\theta)\alpha_{t-j}\hat N_{t-j}(\theta)\}$ and $\{\alpha_t\hat N_t(\theta)U_t\}$.

$$= \sum_{k=1}^{t} \mathbb{E}[[\mathbb{E}[(z_{i,k}^{(j)})^2 - (z_{i,k-1}^{(j)})^2 | \mathcal{A}_{k-1}]] \tag{A.19}$$

$$= \sum_{k=1}^{t} \mathbb{E}[(z_{i,k}^{(j)} - z_{i,k-1}^{(j)})^2] \text{ (by (A.17))} \tag{A.20}$$

$$= \sum_{k=1}^{t} \frac{1}{k^2} \hat{N}_k^2(\theta) \mathbb{E}[\bar{\varphi}_{i,k}^{(j)}(\theta)^2] \tag{A.21}$$

$$\leq \sum_{k=1}^{t} \frac{1}{k} \hat{N}_k^2(\theta) \frac{1}{k} \mathbb{E}\left[\|\bar{\varphi}_{i,k}(\theta)\|^2\right] \tag{A.22}$$

$$\leq \sqrt{\sum_{k=1}^{t} \frac{1}{k^2} \hat{N}_k^4(\theta)} \sqrt{\sum_{k=1}^{t} \frac{1}{k^2} \mathbb{E}[\|\bar{\varphi}_{i,k}(\theta)\|^2]^2} \text{ (Cauchy−Schwarz)} \tag{A.23}$$

$$\leq \sqrt{\sum_{k=1}^{t} \frac{1}{k^2} \hat{N}_k^4(\theta)} \sqrt{\mathbb{E}\left[\sum_{k=1}^{t} \frac{1}{k^2} \|\bar{\varphi}_{i,k}(\theta)\|^4\right]} \text{ (Jensen's inequality),} \tag{A.24}$$

and, keeping in mind that the expected value is only w.r.t. $\{\alpha_{i,t}\}$, this is bounded by virtue of Lemma A.6 (3.c, 3.d). Thus, we have proved that $\{z_{i,t}^{(j)}\}$ is a martingale with bounded variance uniformly w.r.t. t, therefore $\sup_t \mathbb{E}[|z_{i,t}^{(j)}|] < \infty$, and we can apply Doob's theorem (Theorem B.2 in Appendix B) to conclude that $\lim_{t\to\infty} z_{i,t}$ is, w.p.1, a limit vector with finite-valued components. Finally, by Kronecker's lemma (Lemma B.3 in Appendix B), $\lim_{t\to\infty} z_{i,t} = \sum_{t=1}^{\infty} \frac{\alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta)}{t} < \infty$ implies $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta) = 0$. As for uniform convergence, using Lemmas A.7 and A.6, one can easily show that there exists a positive δ such that, for n large enough, the values $\{|\frac{1}{n}\sum_{t=0}^{n} \alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta)| : \theta \in B_\delta(\theta')\}$ are ϵ-close to each other, no matter what θ' is. Since Θ_c is compact, a finite number δ-balls cover the whole set Θ_c and therefore $|\frac{1}{n}\sum_{t=0}^{n} \alpha_{i,t}\hat{N}_t(\theta)\bar{\varphi}_{i,t}(\theta)|$ can be made arbitrarily small uniformly on the whole Θ_c for n large enough.

Appendix B. Useful results.

THEOREM B.1 (Kolmogorov's strong law of large numbers [64]). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments, and let $S_n = \sum_{t=1}^{n} \xi_t$. Assume that*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[(\xi_t - \mathbb{E}[\xi_t])^2]}{t^2} < \infty,$$

then

$$\lim_{n\to\infty} \frac{S_n - \mathbb{E}[S_n]}{n} = 0 \ \textit{w.p.1}.$$

The following theorem [57, Chapter VII, section 4, Theorem 1] is fundamental in the study of convergence of the (sub)martingale and can be thought of as a stochastic analogue of the monotone convergence theorem for real sequences.

THEOREM B.2 (Doob). *Let $(\xi_n, \mathcal{F}_n)$ be a submartingale (i.e., $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] \geq \xi_n$ w.p.1) with $\sup_n \mathbb{E}[|\xi_n|] < \infty$. Then, w.p.1, the limit $\lim_{n\to\infty} \xi_n = \xi_\infty$ exists and $\mathbb{E}[|\xi_\infty|] < \infty$.*

LEMMA B.3 (Kronecker [57]). *Let $b_1, b_2, \ldots$ be a sequence of positive increasing numbers such that $\lim_{n\to\infty} b_n = \infty$, and let $c_1, c_2, \ldots$ be a sequence of numbers such that $\sum_{n=1}^{\infty} c_n$ converges. Then, $\lim_{n\to\infty} \frac{1}{b_n}\sum_{j=1}^{n} b_j c_j = 0$. In particular, if $b_n = n$, $c_n = \frac{d_n}{n}$, and $\sum_{n=1}^{\infty} \frac{d_n}{n}$ converges, then $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} d_j = 0$.*

REFERENCES

[1] Y. Abbasi-Yadkori and Cs. Szepesvári, *Regret bounds for the adaptive control of linear quadratic systems*, in Proceedings of the 24th Annual Conference on Learning Theory, 2011, pp. 1–26.

[2] K. Amelin and O. Granichin, *Randomized control strategies under arbitrary external noise*, IEEE Trans. Automat. Control, 61 (2016), pp. 1328–1333, https://doi.org/10.1109/TAC.2015.2463612.

[3] G. Baggio, A. Carè, and G. Pillonetto, *Finite-sample guarantees for state-space system identification under full state measurements*, in Proceedings of the 61st Conference on Decision and Control, IEEE, 2022, pp. 2789–2794.

[4] G. Baggio, A. Carè, A. Scampicchio, and G. Pillonetto, *Bayesian frequentist bounds for machine learning and system identification*, Automatica, 146 (2022), 110599, https://doi.org/10.1016/j.automatica.2022.110599.

[5] R. Boczar, N. Matni, and B. Recht, *Finite-data performance guarantees for the output-feedback control of an unknown system*, in Proceedings of the 57th Conference on Decision and Control, IEEE, 2018, pp. 2994–2999.

[6] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, J. Econometrics, 31 (1986), pp. 307–327, https://doi.org/10.1016/0304-4076(86)90063-1.

[7] P. Bullen, *Dictionary of Inequalities*, 2nd ed., CRC Press, Boca Raton, FL, 2015.

[8] A. Calisti, D. Dardari, G. Pasolini, M. Kieffer, and F. Bassi, *Information diffusion algorithms over WSNs for non-asymptotic confidence region evaluation*, in Proceedings of the International Conference on Communications, IEEE, 2017, pp. 1–7.

[9] M. C. Campi and E. Weyer, *Guaranteed non-asymptotic confidence regions in system identification*, Automatica, 41 (2005), pp. 1751–1764, https://doi.org/10.1016/j.automatica.2005.05.005.

[10] A. Carè, M. C. Campi, B. Cs. Csáji, and E. Weyer, *Facing undermodelling in sign-perturbed-sums system identification*, Systems Control Lett., 153 (2021), 104936, https://doi.org/10.1016/j.sysconle.2021.104936.

[11] A. Carè, B. C. Csáji, and M. C. Campi, *Sign-perturbed sums (SPS) with asymmetric noise: Robustness analysis and robustification techniques*, in Proceedings of the 55th Conference on Decision and Control, IEEE, 2016, pp. 262–267.

[12] A. Carè, B. Cs. Csáji, M. C. Campi, and E. Weyer, *Finite-sample system identification: An overview and a new correlation method*, IEEE Control Syst. Lett., 2 (2018), pp. 61–66, https://doi.org/10.1109/LCSYS.2017.2720969.

[13] D. F. Coutinho, C. E. de Souza, K. A. Barbosa, and A. Trofino, *Robust linear, H_∞ filter design for a class of uncertain nonlinear systems: An LMI approach*, SIAM J. Control Optim., 48 (2009), pp. 1452–1472, https://doi.org/10.1137/060669504.

[14] B. Cs. Csáji, M. C. Campi, and E. Weyer, *Non-asymptotic confidence regions for the least-squares estimate*, in Proceedings of the 16th IFAC Symposium on System Identification, 2012, pp. 227–232.

[15] B. Cs. Csáji, M. C. Campi, and E. Weyer, *Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models*, IEEE Trans. Signal Process., 63 (2015), pp. 169–181, https://doi.org/10.1109/TSP.2014.2369000.

[16] B. Cs. Csáji and K. B. Kis, *Distribution-free uncertainty quantification for kernel methods by gradient perturbations*, Mach. Learn., 108 (2019), pp. 1677–1699, https://doi.org/10.1007/s10994-019-05822-1.

[17] F. Dabbene, M. Sznaier, and R. Tempo, *Probabilistic optimal estimation with uniformly distributed noise*, IEEE Trans. Automat. Control, 59 (2014), pp. 2113–2127, https://doi.org/10.1109/TAC.2014.2318092.

[18] M. Dalai, E. Weyer, and M. C. Campi, *Parameter identification for nonlinear systems: Guaranteed confidence regions through LSCR*, Automatica, 43 (2007), pp. 1418–1425, https://doi.org/10.1016/j.automatica.2007.01.016.

[19] R. Davidson and E. Flachaire, *The wild bootstrap, tamed at last*, J. Econometrics, 146 (2008), pp. 162–169, https://doi.org/10.1016/j.jeconom.2008.08.003.

[20] R. Davidson and J. G. MacKinnon, *Bootstrap methods in econometrics, Chapter* 23, in Palgrave Handbooks of Econometrics: Volume 1 Econometric Theory, T. C. Mills and K. D. Patterson, eds., Palgrave Macmillan, Basingstoke, UK, 2006 pp. 812–838, https://link.springer.com/book/9781403941558.

[21] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, *On the sample complexity of the linear quadratic regulator*, Found. Comput. Math., 20 (2020), pp. 633–679, https://doi.org/10.1007/s10208-019-09426-y.

[22] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

[23] A. D. Evstifeev, G. A. Volkov, A. A. Chevrychkina, and Y. V. Petrov, *Strength performance of* 1230 *aluminum alloy under tension in the quasi-static and dynamic ranges of loading parameters*, Tech. Phys., 64 (2019), pp. 620–624, https://doi.org/10.1134/S1063784219050050.

[24] S. Fattahi, N. Matni, and S. Sojoudi, *Efficient learning of distributed linear-quadratic control policies*, SIAM J. Control Optim., 58 (2020), pp. 2927–2951, https://doi.org/10.1137/19M1291108.

[25] D. Foster, T. Sarkar, and A. Rakhlin, *Learning nonlinear dynamical systems from a single trajectory*, in Learning for Dynamics and Control, Proceedings of Machine Learning Research, 2020, pp. 851–861.

[26] S. Garatti, M. C. Campi, and S. Bittanti, *Assessing the quality of identified models through the asymptotic theory – when is the result reliable?*, Automatica, 40 (2004), pp. 1319–1332, https://doi.org/10.1016/j.automatica.2004.03.005.

[27] L. Godfrey, *Bootstrap Tests for Regression Models*, Palgrave Macmillan, Basingstoke, UK, 2009.

[28] A. Goldenshluger, *Nonparametric estimation of transfer functions: Rates of convergence and adaptation*, IEEE Trans. Inform. Theory, 44 (1998), pp. 644–658, https://doi.org/10.1109/18.661510.

[29] N. Granichin, G. Volkov, Y. Petrov, and M. Volkova, *Randomized approach to determine dynamic strength of ice*, Cybernet. Phys., 10 (2021), pp. 122–126, https://doi.org/10.35470/2226-4116-2021-10-3-122-126.

[30] O. N. Granichin, *The nonasymptotic confidence set for parameters of a linear control object under an arbitrary external disturbance*, Automat. Remote Control, 73 (2012), pp. 20–30, https://doi.org/10.1134/S000511791201002X.

[31] C.-Y. Han, M. Kieffer, and A. Lambert, *Guaranteed confidence region characterization for source localization using RSS measurements*, Signal Process., 152 (2018), pp. 104–117, https://doi.org/10.1016/j.sigpro.2018.03.018.

[32] U. D. Hanebeck, J. Horn, and G. Schmidt, *On combining statistical and set-theoretic estimation*, Automatica, 35 (1999), pp. 1101–1109, https://doi.org/10.1016/S0005-1098(99)00011-4.

[33] M. Hardt, T. Ma, and B. Recht, *Gradient descent learns linear dynamical systems*, J. Mach. Learn. Res., 19 (2018), pp. 1–44, https://jmlr.csail.mit.edu/papers/v19/16-465.html.

[34] M. Karimshoushtari and C. Novara, *Design of experiments for nonlinear system identification: A set membership approach*, Automatica, 119 (2020), 109036, https://doi.org/10.1016/j.automatica.2020.109036.

[35] M. M. Khorasani and E. Weyer, *Non-asymptotic confidence regions for the parameters of EIV systems*, Automatica, 115 (2020), 108873, https://doi.org/10.1016/j.automatica.2020.108873.

[36] M. M. Khorasani and E. Weyer, *Non-asymptotic confidence regions for the transfer functions of errors-in-variables systems*, IEEE Trans. Automat. Control, 67 (2021), pp. 2373–2388, https://doi.org/10.1109/TAC.2021.3080504.

[37] M. Kieffer and E. Walter, *Guaranteed estimation of the parameters of nonlinear continuous-time models: Contributions of interval analysis*, Int. J. Adapt. Control Signal Process., 25 (2011), pp. 191–207, https://doi.org/10.1002/acs.1194.

[38] M. Kieffer and E. Walter, *Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS*, Automatica, 50 (2013), pp. 507–512, https://doi.org/10.1016/j.automatica.2013.11.010.

[39] S. Kolumbán, I. Vajk, and J. Schoukens, *Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets*, Automatica, 51 (2015), pp. 326–331, https://doi.org/10.1016/j.automatica.2014.10.083.

[40] T. Kozubowski and K. Podgórski, *Asymmetric Laplace laws and modeling financial data*, Math. Comput. Model., 34 (2001), pp. 1003–1021, https://doi.org/10.1016/S0895-7177(01)00114-5.

[41] R. Y. Liu, *Bootstrap procedures under some non-iid models*, Ann. Statist., 16 (1988), pp. 1696–1708, https://doi.org/10.1214/aos/1176351062.

[42] L. Ljung, *Characterization of the Concept of 'Persistently Exciting' in the Frequency Domain*, Technical report TFRT-3038, Department of Automatic Control, Lund Institute of Technology, 1971.

[43] L. Ljung, *Consistency of the least squares identification method*, IEEE Trans. Automat. Control, 21 (1976), pp. 779–781, https://doi.org/10.1109/TAC.1976.1101344.

[44] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1999.

[45] H. Mania, M. I. Jordan, and B. Recht, *Active learning for nonlinear system identification with guarantees*, J. Mach. Learn. Res., 23 (2022), pp. 1–30, https://jmlr.org/papers/v23/20-807.html.

[46] M. Milanese, J. Norton, H. Piet-Lahanier, and É. Walter, *Bounding Approaches to System Identification*, Springer, New York, 2013.

[47] M. Milanese and C. Novara, *Unified set membership theory for identification, prediction and filtering of nonlinear systems*, Automatica, 47 (2011), pp. 2141–2151, https://doi.org/10.1016/j.automatica.2011.03.013.

[48] M. Milanese and M. Taragna, *H_∞ set membership identification: A survey*, Automatica, 41 (2005), pp. 2019–2032, https://doi.org/10.1016/j.automatica.2005.07.007.

[49] S. Oymak and N. Ozay, *Non-asymptotic identification of LTI systems from a single trajectory*, in Proceedings of the American Control Conference, IEEE, 2019, pp. 5655–5661.

[50] J. Pereira, M. Ibrahimi, and A. Montanari, *Learning networks of stochastic differential equations*, in Advances in Neural Information Processing Systems 23 (NIPS 2010), 2010, https://papers.nips.cc/paper_files/paper/2010/hash/8df707a948fac1b4a0f97aa554886ec8-Abstract.html.

[51] P. Polterauer, H. Kirchsteiger, and L. del Re, *State observation with guaranteed confidence regions through sign perturbed sums*, in Proceedings of the 54th Conference on Decision and Control, IEEE, 2015, pp. 5660–5665.

[52] J. G. Proakis, *Digital Signal Processing: Principles, Algorithms, and Application-3/E*, Prentice-Hall, Upper Saddle River, NJ, 1996.

[53] M. Quincampoix and V. M. Veliov, *Optimal control of uncertain systems with incomplete information for the disturbances*, SIAM J. Control Optim., 43 (2004), pp. 1373–1399, https://doi.org/10.1137/S0363012903420863.

[54] T. Sarkar, A. Rakhlin, and M. A. Dahleh, *Finite time LTI system identification*, J. Mach. Learn. Res., 22 (2021), pp. 1–61, https://jmlr.org/papers/v22/19-725.html.

[55] Y. Sattar and S. Oymak, *Non-asymptotic and accurate learning of nonlinear dynamical systems*, J. Mach. Learn. Res., 23 (2022), pp. 1–49, https://jmlr.org/papers/v23/20-617.html.

[56] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht, *Linear system identification via atomic norm regularization*, in Proceedings of the 51st Conference on Decision and Control, IEEE, 2012, pp. 6265–6270.

[57] A. N. Shiryaev, *Probability*, 2nd ed., Springer, New York, 1995.

[58] Y. Sun, S. Oymak, and M. Fazel, *Finite sample identification of low-order LTI systems via nuclear norm regularization*, IEEE Open J. Control Syst., 1 (2022), pp. 237–254, https://doi.org/10.1109/OJCSYS.2022.3200015.

[59] S. Szentpéteri and B. Cs. Csáji, *Non-asymptotic state-space identification of closed-loop stochastic linear systems using instrumental variables*, Systems Control Lett., 178 (2023), 105565, https://doi.org/10.1016/j.sysconle.2023.105565.

[60] S. Y. Trapitsin, O. A. Granichina, and O. N. Granichin, *Social capital of professors as a factor of increasing the effectiveness of the university*, in Proceedings of the International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS), IEEE, 2018, pp. 739–742.

[61] A. Tsiamis and G. J. Pappas, *Finite sample analysis of stochastic system identification*, in Proceedings of the 58th Conference on Decision and Control, IEEE, 2019, pp. 3648–3654.

[62] A. Tsiamis and G. J. Pappas, *Linear systems can be hard to learn*, in Proceedings of the 60th Conference on Decision and Control, IEEE, 2021, pp. 2903–2910.

[63] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, *Statistical learning theory for control: A finite-sample perspective*, IEEE Control Syst. Mag., 43 (2023), pp. 67–97, https://doi.org/10.1109/MCS.2023.3310345.

[64] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK, 1998.

[65] M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*, Cambridge University Press, Cambridge, UK, 2007.

[66] M. Vidyasagar and R. L. Karandikar, *System identification: A learning theory approach*, in Control and Modeling of Complex Systems: Cybernetics in the 21st Century Festschrift in Honor of Hidenori Kimura on the Occasion of his 60th Birthday, K. Hashimoto, Y. Oishi, and Y. Yamamoto, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 89–104, https://doi.org/10.1007/978-1-4612-0023-9_6.

[67] M. Vidyasagar and R. L. Karandikar, *A learning theory approach to system identification and stochastic adaptive control*, in Probabilistic and Randomized Methods for Design Under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2006, pp. 265–302, https://doi.org/10.1007/1-84628-095-8_10.

[68] G. A. Volkov, A. A. Gruzdkov, and Y. V. Petrov, *A randomized approach to estimate acoustic strength of water*, in Mechanics and Control of Solids and Structures, V. A. Polyanskiy and A. K. Belyaev, eds., Springer, Berlin, 2022, pp. 633–640, https://doi.org/10.1007/978-3-030-93076-9_30.

[69] M. Volkova, O. Granichin, Y. Petrov, and G. Volkov, *Dynamic fracture tests data analysis based on the randomized approach*, Adv. Syst. Sci. Appl., 17 (2017), pp. 34–41, http://www.scopus.com/inward/record.url?scp=85040729614&partnerID=8YFLogxK.

[70] M. Volkova, O. Granichin, G. Volkov, and Y. V. Petrov, *On the possibility of using the method of sign-perturbed sums for the processing of dynamic test data*, Vestnik St. Petersburg Univ. Math., 51 (2018), pp. 23–30, https://doi.org/10.3103/S1063454118010132.

[71] V. Volpe, B. Cs. Csáji, A. Carè, E. Weyer, and M. C. Campi, *Sign-perturbed sums (SPS) with instrumental variables for the identification of ARX systems*, in Proceedings of the 54th IEEE Conference on Decision and Control, IEEE, 2015, pp. 2115–2120.

[72] E. Weyer, *Finite sample properties of system identification of ARX models under mixing conditions*, Automatica, 36 (2000), pp. 1291–1299, https://doi.org/10.1016/S0005-1098(00)00039-X.

[73] E. Weyer, M. C. Campi, and B. Cs. Csáji, *Asymptotic properties of SPS confidence regions*, Automatica, 82 (2017), pp. 287–294, https://doi.org/10.1016/j.automatica.2017.04.041.

[74] E. Weyer, R. C. Williamson, and I. M. Mareels, *Sample complexity of least squares identification of FIR models*, IFAC Proc. Vol., 29 (1996), pp. 4664–4669, https://doi.org/10.1016/S1474-6670(17)58418-9.

[75] E. Weyer, R. C. Williamson, and I. M. Mareels, *Finite sample properties of linear model identification*, IEEE Trans. Automat. Control, 44 (1999), pp. 1370–1383, https://doi.org/10.1109/9.774109.

[76] V. Zambianchi, F. Bassi, A. Calisti, D. Dardari, M. Kieffer, and G. Pasolini, *Distributed nonasymptotic confidence region computation over sensor networks*, IEEE Trans. Signal Inform. Process. Over Networks, 4 (2018), pp. 308–324, https://doi.org/10.1109/TSIPN.2017.2695403.