This paper has been accepted for publication in Systems & Control letters
The final version is available at: https://doi.org/10.1016/j.sysconle.2021.104936

Facing Undermodelling in Sign-Perturbed-Sums System Identification*

A. Carè[a], M. C. Campi[a], B. Cs. Csáji[b], E. Weyer[c]

[a] *University of Brescia, Brescia, IT*
[b] *SZTAKI: Institute for Computer Science and Control, Budapest, HU*
[c] *The University of Melbourne, Melbourne, AU*

Abstract

Sign-Perturbed Sums (SPS) is a finite sample system identification method that constructs exact, non-asymptotic confidence regions for the unknown parameters of linear systems without using any knowledge about the disturbances except that they are symmetrically distributed. In the available literature, the theoretical properties of SPS have been investigated under the assumption that the order of the system model is known to the user. In this paper, we analyse the behaviour of SPS when the model assumed by the user does not match the data generation mechanism, and we propose a new SPS algorithm able to detect the circumstance that the model order is incorrect.

Keywords: system identification, confidence regions, finite sample results, least squares, parameter estimation, distribution-free results, linear systems, uncertainty evaluation

1. Introduction

Estimating parameters of partially unknown systems based on observations corrupted by noise is a fundamental problem in system identification, signal processing and statistics at large, with an impact on control and prediction methods in machine learning, [1, 2, 3]. Several standard approaches, such as the Least Squares (LS) method or, more generally, prediction error methods, can be employed to obtain *point estimates* of the unknown parameters. In many situations, for example when the safety, stability or quality of a process has to be guaranteed, a point estimate should be accompanied with an *uncertainty region* that certifies the accuracy of the estimate. In standard *statistical* system identification approaches, confidence regions are constructed based on theoretical analyses that, typically, have only asymptotic validity, while guarantees that are valid for a finite sample of observations require strong assumptions on the data generation mechanism. Along a different route, finite sample results can be obtained by resorting to *worst-case* identification approaches where the noise is guaranteed to belong to a bounded set, see e.g. [4, 5, 6, 7]. In this approach, a region for the unknown system parameters can be constructed by including in the region those parameter vectors that are consistent with the observed data given all the possible realizations of the noise in the noise bounding set. Such methodologies construct uncertainty

*Corresponding author Marco C. Campi.
Email addresses: algo.care@unibs.it (A. Carè), marco.campi@unibs.it (M. C. Campi), csaji@sztaki.hu (B. Cs. Csáji), ewey@unimelb.edu.au (E. Weyer)

regions that are robust, but often conservative in practice. These issues have recently led to studies, see e.g. [8] and the references therein, where the worst-case approach is complemented with statistical knowledge to reduce conservatism, and, on a different line of research, to a class of statistical finite sample identification algorithms (see, e.g., [9, 10, 11, 12, 13, 14, 15, 16, 17, 18] and [19] for a recent overview) that (i) do not rely on noise bounding sets and (ii) are robust to statistical uncertainties.

This paper focuses on one of these statistical, finite sample algorithm: the *Sign-Perturbed Sums* (SPS) algorithm, [12, 20]. SPS constructs confidence regions around the Least Squares Estimate (LSE). In the case of Finite Impulse Response (FIR) systems, several important properties of SPS have been proven rigorously. In particular, we recall here that the regions constructed by SPS have an *exact coverage probability* (i.e., they include the true parameter vector with an exact and user-chosen probability) independently of the specific, unknown, distribution of the noise, which is only assumed to be symmetric and forms an independent sequence. Moreover, the SPS regions are *strongly consistent* [20].

All the known properties of SPS, however, have been derived under the assumption that the true data generating system belongs to the model class considered by the user, with known model order (at least, an upper-bound to the model order should be known). In system identification practice, this information typically comes from domain knowledge or a previous stage of the identification workflow, which is known as *model order selection*. Model order selection is a standard topic in the system identification literature and various supporting tools are available to the user, see e.g., [3, 1, 21], and [22, 23, 24, 25] for more recent contributions. However, this is still a difficult and theoretically challenging problem, which might end up with an underestimate of the model order, [26, 27, 28, 29].

1.1. Aim and structure of the paper

In this paper, we study the behaviour of SPS in the presence of undermodelling and argue that, with undermodelling, the SPS regions, which are no more exact, may induce a false sense of confidence in the incautious user. Thus, we introduce a modified version of SPS with the following properties:

- if the system is not undermodelled, the algorithm builds exact, non-asymptotic confidence regions for the true model parameter vector;
- if the system is undermodelled, the algorithm has a propensity to warn the user that there is a mismatch between the data generation mechanism and the postulated model class (see the simulation section for practical examples) and, moreover, it certainly detects undermodelling when the number of data points becomes large (see the asymptotic Theorem 5).

After a brief discussion of the related literature in the following Section 1.2, we review the standard SPS method in Section 2. Standard SPS in the presence of undermodelling is analysed in Section 3 and this theoretical analysis will provide a motivation for the new algorithm UD-SPS (SPS with Undermodelling Detection), which we introduce and study in Section 4. Computational aspects of UD-SPS are discussed in Section 5. An illustration on a numerical example is offered in Section 6. Section 7 presents our conclusions and outlines some future directions of research.

1.2. Related literature

Although a simulation experiment on the effect of undermodelling on SPS was carried out in [12], the available literature on the theory of SPS does not consider this possibility. If we look at the broader class of statistical finite sample identification methods, the algorithm in [30] allows the user to estimate a subset of the unknown system parameters, which makes the algorithm applicable also when the true model order is higher than the selected one. However, differently from SPS, the algorithm in [30] does not build regions around the LS estimate, and, most importantly, it can be applied only if the measurable input is known to satisfy (or can be chosen so as to satisfy) precise statistical properties (e.g., the input has to be a symmetric white process, or a filtered version of it through a known filter). In this paper, instead, no assumptions will be made on the input except for standard excitation conditions. The conference paper [31] contains a preliminary exposition of the ideas of this paper, which are here revised, developed and accompanied with rigorous theoretical derivations that were not given in [31]; the discussion on the practical and computational aspects and the numerical examples are also new.

2. The SPS algorithm

In this section we summarize the basic ideas behind the SPS algorithm, and recall the fundamental theorem about its exact confidence and its asymptotic properties. The concepts and results of this section are presented more comprehensively and made precise in [12] and [20]. The reader is referred to these papers for more details and comparisons with related methodologies, such as bootstrap methods. For an overview of SPS in the context of statistical finite sample methods, see also [19].

2.1. The SPS set-up

Consider the following scalar linear regression system

$$Y_t = \varphi_t^\top \theta^* + N_t, \tag{1}$$

where t is the discrete time index, $\{Y_t\}$ is the output, $\{N_t\}$ the noise, $\{\varphi_t\}$ is a measured d-dimensional input regressor and $\theta^* = [\, b_1^*, \ldots, b_d^* \,]^\top$ is the d-dimensional parameter vector to be estimated. A notable case is when (1) describes the input-output equation of a Finite Impulse Response (FIR) system fed with an input $\{U_t\}$, in which case $\varphi_t = [\, U_{t-1}, \ldots, U_{t-d} \,]^\top$, while more more general models such as those based on Laguerre polynomials are also included, see Section II.B in [12].

Following [12], in this paper we will always treat the measured input $\{\varphi_t\}$ as *deterministic*, but we remark that all the results here presented can be immediately generalized to random inputs when they are independent of the noise $\{N_t\}$. The noise sequence $\{N_t\}$ is made up of random variables that are independent (not necessarily identically distributed) and symmetric about zero[1].

Given n observations, i.e., $Y_1, \ldots, Y_n$ and $\varphi_1, \ldots, \varphi_n$, the Least Squares Estimate (LSE) $\hat{\theta}_n$ is the value of θ that minimizes the sum of the squared prediction errors $\{Y_t - \varphi_t^\top \theta\}$, i.e.,

$$\hat{\theta}_n \triangleq \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{n} (Y_t - \varphi_t^\top \theta)^2. \tag{2}$$

SPS constructs exact confidence regions around $\hat{\theta}_n$ for the unknown true parameter θ^*.

[1] For a discussion of the robustness of SPS with respect to violations of the symmetry assumptions, see [32].

2.2. Review of the SPS algorithm and its core ideas

It is well-known that the LSE, given by (2), satisfies the normal equation

$$\sum_{t=1}^{n} \varphi_t (Y_t - \varphi_t^\top \theta) = 0, \tag{3}$$

whose solution can be written as

$$\hat{\theta}_n \;=\; \Big(\sum_{t=1}^{n} \varphi_t \varphi_t^\top \Big)^{-1} \Big(\sum_{t=1}^{n} \varphi_t Y_t \Big) \tag{4}$$

provided that $\sum_{t=1}^{n} \varphi_t \varphi_t^\top$ is invertible. The fundamental step in SPS consists of generating $m-1$ *sign-perturbed sums* by randomly perturbing the sign of the prediction error in the normal equations (3); namely, for $i = 1, \ldots, m-1$, define

$$H_i(\theta) \;\triangleq\; \sum_{t=1}^{n} \varphi_t \alpha_{i,t} (Y_t - \varphi_t^\top \theta)$$

where $\{\alpha_{i,t}\}$ are random signs, i.e., i.i.d. random variables that take on the values ± 1 with probability 1/2 each. These perturbed sums are suitably compared with a *reference sum*, which is simply the unperturbed sum

$$H_0(\theta) \;\triangleq\; \sum_{t=1}^{n} \varphi_t (Y_t - \varphi_t^\top \theta).$$

For reasons discussed in [12, 20], it is convenient to "normalize" the sums $H_0(\theta), \ldots, H_{m-1}(\theta)$ as follows

$$S_i(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} H_i(\theta), \;\; i = 0, \ldots, m-1, \tag{5}$$

where $R_n \;=\; \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^\top$ and "$^{-\frac{1}{2}}$" denotes the inverse of the principal square root. Denote by $\mathcal{R}(\theta)$ the *rank* of the element $\|S_0(\theta)\|^2$ ($= S_0(\theta)^\top S_0(\theta)$) in the ordering of the m elements $\|S_0(\theta)\|^2, \|S_1(\theta)\|^2, \ldots, \|S_{m-1}(\theta)\|^2$, e.g., $\mathcal{R}(\theta) = 1$ means that $\|S_0(\theta)\|^2$ is the smallest one, $\mathcal{R}(\theta) = 2$ means that $\|S_0(\theta)\|^2$ is the second smallest, and so on. In case of ties, the tie is broken by randomization, see [12] for details.

A *crucial* fact is that, when $\theta = \theta^*$ (which, of course, we are not aware of, as θ^* is unknown), we have

$$S_0(\theta^*) \;=\; R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t N_t,$$

$$S_i(\theta^*) \;=\; R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \pm \varphi_t N_t$$

where in the last equation we have written $\pm$ instead of $\{\alpha_{i,t}\}$ for a more intuitive visualization. Since $\{\pm N_t\}$ is distributed exactly as $\{N_t\}$ (use the symmetry assumption about zero), $S_0(\theta^*)$ and $S_i(\theta^*)$ have the same distribution and there is no reason for $\|S_0(\theta^*)\|^2$ to be bigger or smaller than any other $\|S_i(\theta^*)\|^2$; indeed, it can be proven rigorously that $\mathcal{R}(\theta^*)$ is uniformly distributed between 1 and m (i.e., $\|S_0(\theta^*)\|^2$ takes any position in the ranking with equal probability) and this

is the key for the construction of the SPS region. Letting q be a number between 1 and m, the SPS region (at level $1 - \frac{q}{m}$) is formally defined as

$$\widehat{\Theta}_n \triangleq \left\{ \theta : \mathcal{R}(\theta) \leq m\left(1 - \frac{q}{m}\right) \right\}, \tag{6}$$

and the following theorem, proven as Theorem 1 in [12], follows from the arguments above.

Theorem 1 (Exact Confidence)**.** If $N_1, \ldots, N_n$ is a sequence of independent random variables distributed symmetrically about zero, then it holds that

$$\mathbb{P}r\{\, \theta^* \in \widehat{\Theta}_n \,\} \;=\; 1 - \frac{q}{m}. \qquad *$$

It is also of special interest for the present study that, under mild excitation assumptions, SPS is *strongly consistent*, i.e., for every $\varepsilon > 0$, $\widehat{\Theta}_n$ is almost surely contained in the ε-ball centred at the true parameter value θ^* for every n large enough. The main idea behind the consistency of SPS is the following. If we compare $\|S_0(\theta)\|^2$ and $\|S_i(\theta)\|^2$ ($i \neq 0$) after rewriting them as functions of $\Delta\theta \triangleq \theta^* - \theta$, we get that, when $\|\Delta\theta\| \neq 0$ it holds that $\|S_0(\theta)\|^2 > \|S_i(\theta)\|^2$ for n large enough (in fact, the term $\sum_{t=1}^{n} \varphi_t\varphi_t^\top \Delta\theta$ on the left-hand side grows faster with n than the term $\sum_{t=1}^{n} \pm\varphi_t\varphi_t^\top \Delta\theta$ on the right-hand side, because the latter is tamed by the effect of the random signs). Therefore, any $\theta \neq \theta^*$ is ruled out in the long run and removed from $\widehat{\Theta}_n$ because of the high value of $\mathcal{R}(\theta)$ (clearly, proving that this happens uniformly for all the θ's outside any small ball requires some additional and nontrivial underpinning).

3. SPS in the presence of undermodelling

From the previous section we know that a *crucial* role in the SPS idea is played by the condition that $\{Y_t - \varphi_t^\top \theta^*\} = \{N_t\}$, i.e., the actual noise sequence is reconstructed from the measured data when the true parameter is correctly guessed. This is obviously true when the true system structure is known, which is the standard assumption in the SPS literature.

In this section, we study the behaviour of the SPS algorithm in a more general setting which includes the possibility that the *true data generation mechanism* does not match the model that has been postulated by the user (which we call the *user-chosen model*).

3.1. Definition of the user-chosen model and the true data-generation mechanism

3.1.1. The user-chosen model

We assume that the user has chosen the following class of predictive models in line with the system described in equation (1):

$$\widehat{Y}_t(\theta) = b_1 U_{t-1} + b_2 U_{t-2} + \cdots + b_{\hat{d}} U_{t-\hat{d}}, \tag{7}$$

where $\{U_t\}$ is the deterministic system input, $\theta = [b_1, \ldots, b_{\hat{d}}]^\top$ is the vector of impulse response coefficients that the user wants to estimate, $\hat{d}$ is the postulated model order.[2]

[2]The reader can easily check that the theoretical results and ideas of this paper do not depend on this specific FIR structure, and can be immediately applied to more general set-ups. A particularly useful generalization is to consider $\{L_1(z)U_t\}, \ldots, \{L_{\hat{d}}(z)U_t\}$ in place of $\{U_t\}, \ldots, \{U_{t-\hat{d}}\}$ where $\{L_k(z)U_t\}$ is $\{U_t\}$ filtered through the transfer function $L_k(z)$. A celebrated choice of the transfer functions $L_k(z)$ is the Laguerre polynomials, [33, 34, 35, 36, 2]. Clearly, the FIR case is recovered when $L_k(z) = z^{-k}$.

3.1.2. The true data generation mechanism

In this paper we will consider two possible situations. The first situation is when the data generation mechanism matches the user-chosen model. In this case, we say that there is a "matching" data generation mechanism. Precisely, the data generation mechanism is matching when data $\{Y_t\}$ are generated according to relation

$$Y_t = \widehat{Y}_t(\theta^*) + N_t \tag{8}$$

for some value θ^* of the parameter θ and the noise $\{N_t\}$ is independent, symmetric about zero.

In the matching case, the standard SPS is guaranteed to deliver a confidence region in $\mathbb{R}^{\hat{d}}$ that contains the true parameter θ^* with user-chosen probability. Moreover, under mild assumptions, the SPS region shrinks around θ^*, see Section 2 and the references therein.

In this paper we also consider the situation of a "non-matching" data generation mechanism: we say that the data generation mechanism is non-matching if, contrary to what is postulated by the user, there is no θ^* such that (8) is valid for an independent and symmetric $\{N_t\}$.[3]

For the sake of generality, we avoid to assume any specific structure for the non-matching data generation mechanism. Our analysis will just assume the existence of $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \varphi_t Y_t$ and require mild restrictions on the growth rate of $\{Y_t\}$. In the specific case where the data generation mechanism is linear (e.g., FIR system of order $d > \hat{d}$, ARX and ARMAX systems), our general assumptions are easily seen to be implied by standard sets of conditions.

3.2. The effect of undermodelling on standard SPS

In the case of a non-matching data generation mechanism one could hope that the SPS region behaves so as to "warn" the user that the algorithm is being applied outside its domain of applicability. Unfortunately, this is not the case. The SPS region is constructed so as to be nonempty and to include the LSE $\hat{\theta}_n$ (see [12]); in addition, Theorem 2 below shows that, under mild assumptions, the SPS region shrinks around $\hat{\theta}_\infty$, defined as $\lim_{n\to\infty} \hat{\theta}_n$, as n increases. This behaviour can easily convey to the user the wrong idea that the estimate becomes more and more precise as more data are collected, which can have deleterious effects if, for example, the SPS region is employed to tune a robust controller against plant parametric uncertainty.

Let $B_\varepsilon(\hat{\theta}_\infty)$ denote the Euclidean norm-ball centred at $\hat{\theta}_\infty$ with radius $\varepsilon > 0$, i.e., $B_\varepsilon(\hat{\theta}_\infty) = \{\theta : \|\hat{\theta}_\infty - \theta\| \le \varepsilon\}$. Theorem 2 below states that the event that $\widehat{\Theta}_n \subseteq B_\varepsilon(\hat{\theta}_\infty)$ for all n larger than some (realization-dependent) value $\bar{n}$ is a probability 1 event.

Theorem 2 (Asymptotic behaviour of SPS)**.** *Let $R_n = \frac{1}{n}\sum_{t=1}^{n} \varphi_t\varphi_t^\top$, and assume that there exists a finite limit matrix $\bar{R} \succ 0$ such that*

$$\lim_{n\to\infty} R_n = \bar{R}. \tag{9}$$

Assume also that there exists a finite real vector $\bar{Y}_\varphi$ such that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \varphi_t Y_t = \bar{Y}_\varphi \ w.p.1, \tag{10}$$

and, moreover, that

[3] Clearly, undermodelling is a particular case of non-matching.

$$\sum_{t=1}^{\infty} \frac{\|\varphi_t\|^4}{t^2} < \infty \tag{11}$$

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[Y_t^2]^2}{t^2} < \infty. \tag{12}$$

Then,

$$\hat{\theta}_\infty \triangleq \lim_{n\to\infty} \hat{\theta}_n = \bar{R}^{-1}\bar{Y}_\varphi \text{ (w.p.1)}, \tag{13}$$

and, for all $\varepsilon > 0$*,*

$$\mathbb{P}r\left[\bigcup_{\bar{n}=1}^{\infty}\bigcap_{n=\bar{n}}^{\infty}\left\{\widehat{\Theta}_n \subseteq B_\varepsilon(\hat{\theta}_\infty)\right\}\right] = 1, \tag{14}$$

where $B_\varepsilon(\hat{\theta}_\infty) \triangleq \{\theta \in \mathbb{R}^{\hat{d}} : \|\hat{\theta}_\infty - \theta\| \le \varepsilon\}$. *

Proof. Using (4), (9) and (10) we immediately get

$$\hat{\theta}_n = R_n^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\varphi_t Y_t\right) \to \bar{R}^{-1}\bar{Y}_\varphi \text{ (w.p.1)}.$$

From the definition (5) of $S_i(\theta)$, for every $i = 1, \ldots, m-1$, we get $\|S_i(\theta)\| \le \|{R_n}^{-\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t Y_t\| + \|{R_n}^{-\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t{\varphi_t}^\top\| \cdot \|\theta\|$. We want to show that for every $\delta > 0$, there exists with probability 1 an n_0 such that for all $n > n_0$ we get $\|S_i(\theta)\| \le \delta + \delta\|\theta\|$ (for all θ), from which it follows that

$$\|S_i(\theta)\| \le \delta + \delta\|\theta - \hat{\theta}_\infty\| + \delta\|\hat{\theta}_\infty\|. \tag{15}$$

Note that $\{\alpha_{i,t}\varphi_t\varphi_t^\top\}$ is a sequence of independent random matrices (we recall that $\{\varphi_t\}$ is a deterministic sequence). Any element of $\alpha_{i,t}\varphi_t\varphi_t^\top$ is zero mean, with variance upper bounded by $\|\varphi_t\|^4$. Condition (11) allows us to apply the Kolmogorov's Strong Law of Large Numbers (Theorem 6 in the Appendix) element-wise and conclude that $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t{\varphi_t}^\top = 0$ w.p.1. By (9), ${R_n}^{-\frac{1}{2}} \to \bar{R}^{-\frac{1}{2}}$ so that, w.p.1, $\|{R_n}^{-\frac{1}{2}}\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t{\varphi_t}^\top\| < \delta$ for any n large enough.

A similar reasoning applies to the term including the averaged sum $\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t Y_t$, but this requires extra attention because, in general, $\{\alpha_{i,t}\varphi_t Y_t\}$ is not a sequence of independent random vectors: to prove that this averaged sum converges to 0 w.p.1 we resort to martingale theory, as follows. Define $M_n = \sum_{t=1}^{n}\frac{1}{t}\alpha_{i,t}\varphi_t Y_t$ for $n = 1, 2, \ldots$. Each component of the vector M_n is a random variable with zero mean whose variance is bounded by $\sigma^2_{M_n} = \sum_{t=1}^{n}\frac{1}{t}\|\varphi_t\|^2\frac{1}{t}\mathbb{E}[Y_t^2]$. Note that $\sigma^2_{M_n}$ can be bounded uniformly for all n: in fact, by the Cauchy-Schwarz inequality, we get

$$\sigma^2_{M_n} \le \sqrt{\sum_{t=1}^{n}\frac{\|\varphi_t\|^4}{t^2}}\sqrt{\sum_{t=1}^{n}\frac{\mathbb{E}[Y_t^2]^2}{t^2}} \le \sqrt{\sum_{t=1}^{\infty}\frac{\|\varphi_t\|^4}{t^2}}\sqrt{\sum_{t=1}^{\infty}\frac{\mathbb{E}[Y_t^2]^2}{t^2}},$$

whose finiteness is ensured by (11) and (12). Denoting by $\mathcal{F}_n$ the σ-algebra generated by the random variables $Y_n, Y_{n-1}, \ldots$ and $\alpha_{i,n}, \ldots, \alpha_{i,1}$, we have that

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = \mathbb{E}[M_n + \alpha_{i,n+1}\frac{\varphi_{n+1}Y_{n+1}}{n+1}|\mathcal{F}_n] = M_n \text{ (w.p.1)},$$

i.e., each component of M_n is a martingale adapted to $\mathcal{F}_n$. Therefore, we can appeal to Doob's theorem (Theorem 7 in the Appendix) to conclude that $\lim_{n\to\infty} M_n = \sum_{t=1}^{\infty} \frac{1}{t}\alpha_{i,t}\varphi_t Y_\ell < \infty$ w.p.1., which, by Kronecker's Lemma (Theorem 8 in the Appendix), entails that $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\varphi_t Y_t = 0$ w.p.1., and this concludes the proof that the function $\|S_i(\theta)\|$ can be upper-bounded according to (15) for all n large enough. Since m is finite and does not depend on n, bound (15) can be ensured for all $i = 1, \dots, m-1$.

On the other hand, using (4), $\|S_0(\theta)\|^2$ can be rewritten as $(\theta - \hat{\theta}_n)^\top R_n(\theta - \hat{\theta}_n)$, from which $\|S_0(\theta)\| = \|R_n^{\frac{1}{2}}(\theta - \hat{\theta}_n)\| \geq \|R_n^{\frac{1}{2}}(\theta - \hat{\theta}_\infty)\| - \|R_n^{\frac{1}{2}}(\hat{\theta}_n - \hat{\theta}_\infty)\|$. Note now that (9) implies that there exists a $\varrho > 0$ such that, for all n large enough, $\|R_n^{\frac{1}{2}}(\theta - \hat{\theta}_\infty)\| \geq \varrho\|\theta - \hat{\theta}_\infty\|$. Moreover, using (13), we also get that for all $\delta > 0$ there exists with probability 1 an n_1 such that for all $n > n_1$ we have $\|R_n^{\frac{1}{2}}(\hat{\theta}_n - \hat{\theta}_\infty)\| \leq \delta$. From these two facts, we conclude that, w.p.1, for all n large enough,

$$\|S_0(\theta)\| \geq \varrho\|\theta - \hat{\theta}_\infty\| - \delta. \tag{16}$$

Since the values of ϱ and $\hat{\theta}_\infty$ are realization-independent, for every $\varepsilon > 0$, we can a-priori choose $\delta > 0$ such that $(\varrho - \delta)\varepsilon > 2\delta + \delta\|\hat{\theta}_\infty\|$. Then, putting (15) and (16) together, we conclude that, with probability 1, for all n large enough, the relationships $\|S_0(\theta)\| > \|S_i(\theta)\|$, $i = 1, \dots, m-1$, are valid for all the values of θ such that $\|\theta - \hat{\theta}_\infty\| > \varepsilon$: hence, no points outside the ball $B_\varepsilon(\hat{\theta}_\infty)$ are included in the SPS region (6), and statement (14) is proven.

□

Remark 1. *Strictly speaking, Theorem 2 applies both to the matching and the non-matching cases. In the case of a non-matching data generation mechanism, the SPS region deceivingly shrinks around the parameter vector $\hat{\theta}_\infty$, which may not be related to the identification goals. On the other hand, in the case of a matching data generation mechanism, Theorem 2 recovers under a slightly different set of conditions a convergence result previously proven in Theorem 2 of [14].*

In summary, the standard SPS algorithm does not provide any mechanism to help the user recognize that in a given situation the assumptions for the validity of the method are not satisfied. This is the motivation for the UD-SPS algorithm that is presented and studied in detail in the next section.

Remark 2. *It is worth noticing that the SPS algorithm delivers guaranteed regions also when $\hat{d}$ is overestimated. In fact, if the data generation mechanism is matching for a certain $\hat{d}$, then it is matching also for larger values of $\hat{d}$ (it is enough to set the additional parameters to zero and see that* (8) *remains true). For this reason, in the presence of uncertainty on the correct model order, one might want to increase the number of parameters $\hat{d}$ in an attempt to increase the flexibility of the adopted model class and reduce the mismatch between the identified system and the data generating mechanism. This strategy, however, comes at a price. As the information that is carried by the data gets diluted in a higher-dimensional identification procedure, the variance of the least squares estimate gets larger and the SPS region tends to include more parameter values for the same confidence level. In contrast, the capability to detect undermodelling enables one to safely reduce $\hat{d}$ to small values, thus making an efficient use of the available data.*

4. UD-SPS: A modified SPS method

We here define the UD-SPS algorithm, and discuss the main idea behind it. Clarifying the connection between UD-SPS and the standard SPS makes it easy to prove that UD-SPS inherits

the most important properties of standard SPS when the system is correctly specified (matching case). Then, we show that UD-SPS can be used to detect undermodelling.

4.1. Formal definition of UD-SPS

UD-SPS constructs a region $\widehat{\Theta}_n^{\mathrm{UD}}$ in which a candidate θ is or is not included depending on the ranking of the following functions

$$\begin{aligned} Q_0(\theta) &\triangleq R_n'^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} (Y_t - \varphi_t^\top \theta), \\ Q_i(\theta) &\triangleq R_n'^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} (Y_t - \varphi_t^\top \theta), \\ &\quad \text{for} \qquad i = 1, \dots, m-1 \end{aligned} \tag{17}$$

where ψ_t is a vector that includes s extra input values, precisely[4]

$$\psi_t \triangleq [\, U_{t-\hat{d}-1}, \cdots, U_{t-\hat{d}-s} \,]^\top, \tag{18}$$

and $R_n' = \frac{1}{n} \sum_{t=1}^{n} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} \begin{bmatrix} \varphi_t^\top & \psi_t^\top \end{bmatrix}$.

In (18), s is a user-chosen parameter. Clearly, s need *not* be equal to the difference between the true order of the system (which is unknown) and $\hat{d}$, and s is usually chosen to be a small number, e.g., 1,2 or 3, see Remark 3. The invertibility of R_n' is assumed here and throughout the remainder of the paper.[5]

We now provide a formal construction of $\widehat{\Theta}_n^{\mathrm{UD}}$ by resorting to a pseudocode. The main aim of this pseudocode is to define $\widehat{\Theta}_n^{\mathrm{UD}}$ in a precise and formal way, while a discussion on the practical computation needed for the implementation of $\widehat{\Theta}_n^{\mathrm{UD}}$ is postponed to Session 5. The intuition behind UD-SPS and its connection with SPS is presented in the next Section 4.2.

The UD-SPS pseudocode is in two parts. The initialization (Table 1) sets the main global parameters and generates the random objects needed for the construction. Moreover, the user provides the desired confidence probability p and the parameter s. Like in the initialization of the standard SPS, a *random* permutation π of the set $\{0, \dots, m-1\}$ is generated, whose only role is to break ties when there are two $\|Q_i(\theta)\|^2$ functions with the same value for a given θ. In fact, based on π, one can define the strict total order $Z_j >_\pi Z_k \Leftrightarrow [\big(Z_j > Z_k\big)$ or $\big(Z_j = Z_k$ and $\pi(j) > \pi(k)\big)]$, (so that π breaks the tie $Z_j = Z_k$) for the variables $Z_i = \|Q_i(\theta)\|^2$, $i = 0, \dots, m-1$. The second part (Table 2) defines the indicator function UD-SPS-Indicator(θ) that determines if a particular parameter vector θ is included in the region $\widehat{\Theta}_n^{\mathrm{UD}}$. In principle, the region $\widehat{\Theta}_n^{\mathrm{UD}}$ is obtained by applying to each and every candidate θ the UD-SPS-Indicator(θ) algorithm.

[4]While this choice of ψ_t is the most natural one in case of the FIR structure (7), the theory developed applies to a general ψ_t, as it can be verified from the derivations in this and in the following sections.

[5]This assumption simplifies the definition and the study of UD-SPS, and is analogous to the assumption of the invertibility of R_n for standard SPS. Generalizations are possible that consider pseudoinverses rather than inverses. We also remark that the only motivation for having R_n' in (17) is to obtain a better shape of the UD-SPS region, while other deterministic choices of R_n' or a downright elimination of the R_n' matrix in the algorithm (i.e., substitution of R_n' with the identity matrix in (17)) do not invalidate the statement of Theorems 3, 4, 5.

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. Set the integer s that determines the length of $$\psi_t \triangleq [\, U_{t-\hat{d}-1}, \cdots, U_{t-\hat{d}-s} \,]^\top;$$ 3. Calculate the outer product $$R'_n \triangleq \frac{1}{n} \sum_{t=1}^{n} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} \begin{bmatrix} \varphi_t^\top & \psi_t^\top \end{bmatrix},$$ and let ${R'_n}^{1/2}$ be the principal square root of R'_n (it holds that ${R'_n}^{1/2} {R'_n}^{1/2} = R'_n$); 4. Generate $n \cdot (m-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with $$\mathbb{P}r\{\alpha_{i,t} = 1\} = \mathbb{P}r\{\alpha_{i,t} = -1\} = \tfrac{1}{2},$$ for $i \in \{1, \ldots, m-1\}$ and $t \in \{1, \ldots, n\}$; 5. 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.

Table 1

Pseudocode: UD-SPS-Indicator (θ)
1. For the given θ, for $t \in \{1, \ldots, n\}$, compute the prediction errors $$\varepsilon_t(\theta) \triangleq Y_t - \varphi_t^{\mathrm{T}} \theta;$$ 2. Let $$Q_0(\theta) \triangleq {R'_n}^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} \varepsilon_t(\theta),$$ $$Q_i(\theta) \triangleq {R'_n}^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t} \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} \varepsilon_t(\theta),$$ for $i \in \{1, \ldots, m-1\}$; 3. Order scalars $\{\|Q_i(\theta)\|^2\}$ according to $>_\pi$; 4. Compute the rank $\mathcal{R}(\theta)$ of $\|Q_0(\theta)\|^2$ in the ordering where $\mathcal{R}(\theta) = 1$ if $\|Q_0(\theta)\|^2$ is the smallest in the ordering, $\mathcal{R}(\theta) = 2$ if $\|Q_0(\theta)\|^2$ is the second smallest, and so on; 6. Return 1 if $\mathcal{R}(\theta) \leq m - q$, otherwise return 0.

Table 2

The p-level UD-SPS *confidence region* is as follows

$$\widehat{\Theta}_n^{\mathrm{UD}} \triangleq \left\{\theta \in \mathbb{R}^{\hat{d}} : \text{UD-SPS-Indicator}(\theta) = 1\right\}. \tag{19}$$

4.2. The idea behind UD-SPS

The following Fact 1 is key to understand and prove the main features of UD-SPS.

Fact 1. *The UD-SPS region $\widehat{\Theta}_n^{UD}$ can be interpreted as the intersection between a $\hat{d}$-dimensional space and a full-fledged standard SPS region, say $\widehat{\Theta}'_n$, that is associated with the regressor $[\varphi_t^\top \ \psi_t^\top]^\top$ and that lives in the domain $\{\theta' \in \mathbb{R}^{\hat{d}+s}\}$. Precisely, $\widehat{\Theta}_n^{UD} = \{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top \ \mathbf{0}^\top]^\top \in \widehat{\Theta}'_n\}$.* $*$

In order to see that Fact 1 is true, consider the functions $S'_0(\theta'), S'_1(\theta'), \ldots, S'_{m-1}(\theta')$ of $\theta' \in \mathbb{R}^{\hat{d}+s}$ defined as

$$\begin{aligned} S'_0(\theta') &= R'^{-\frac{1}{2}}_n \frac{1}{n}\sum_{t=1}^{n} \varphi'_t(Y_t - \varphi'^\top_t \theta'), \\ S'_i(\theta') &= R'^{-\frac{1}{2}}_n \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\varphi'_t(Y_t - \varphi'^\top_t \theta'), \\ &\quad \text{for} \qquad i = 1, \ldots, m-1, \end{aligned} \tag{20}$$

where

$$R'_n = \frac{1}{n}\sum_{t=1}^{n} \varphi'_t \varphi'^\top_t \tag{21}$$

and

$$\varphi'^\top_t = \begin{bmatrix} \varphi_t^\top & \psi_t^\top \end{bmatrix}^\top = [U_{t-1}, \ldots, U_{t-\hat{d}}, U_{t-\hat{d}-1}, U_{t-\hat{d}-s}]. \tag{22}$$

These are the functions based on which a standard SPS region, which we denote $\widehat{\Theta}'_n$, can be built in the augmented space $\mathbb{R}^{\hat{d}+s}$ for the model $\widehat{Y}_t = \varphi'^\top_t \theta'$. Comparing (20) with (17), it can be observed that functions (20) take the same values of functions (17) whenever θ' is restricted to $\mathbb{R}^{\hat{d}} \times \{0\}^s$, i.e.,

$$S'_i(\theta')_{|\theta'=[\theta^\top \mathbf{0}^\top]^\top} = Q_i(\theta), \tag{23}$$

from which the statement in Fact 1 follows.

4.3. UD-SPS with correct system specification

Let us first consider the case of a matching data generation mechanism. We prove that the coverage property of SPS carries over to UD-SPS.

Theorem 3. *For a* matching *data generation mechanism, it holds that*

$$\mathbb{P}r\{\, \theta^* \in \widehat{\Theta}_n^{UD} \,\} \ = \ 1 - \frac{q}{m}. \qquad *$$

Proof. Let $\theta'^* \triangleq [\, \theta^{*\top}, \mathbf{0}^\top]^\top$. Recalling that $\widehat{\Theta}_n^{\mathrm{UD}} = \{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top \mathbf{0}^\top]^\top \in \widehat{\Theta}'_n\}$ (see Fact 1) one concludes that $\theta^* \in \widehat{\Theta}_n^{\mathrm{UD}}$ if and only if $\theta'^* \in \widehat{\Theta}'_n$. Then, the conclusion of the theorem follows by observing that $\theta'^* \in \widehat{\Theta}'_n$ with probability $1 - \frac{q}{m}$ by Theorem 1 (see Section 2).

□

In addition, UD-SPS is strongly consistent.

Theorem 4 (Strong consistency of UD-SPS)**.** *Assume that the data generating system is* matching*, that there exists a finite limit matrix* $\bar{R}' > 0$ *such that*

$$\lim_{n\to\infty} R'_n = \bar{R}', \tag{24}$$

and, moreover, that

$$\sum_{t=1}^{\infty} \frac{\|\varphi'_t\|^4}{t^2} < \infty, \tag{25}$$

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[N_t^2]^2}{t^2} < \infty. \tag{26}$$

Then, for all $\varepsilon > 0$*,*

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

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

Proof. The equation of the matching data generation mechanism, $Y_t = \varphi_t^\top \theta^* + N_t$, can be rewritten as $Y_t = \varphi'^\top_t \theta^{*\prime} + N_t$ where $\varphi'^\top_t = [\varphi_t{}^\top \psi_t{}^\top]$ is the augmented regressor and $\theta^{*\prime\top} = [\theta^{*\top}\ \mathbf{0}^\top]$ is the "true" parameter vector in the augmented $(\hat{d}+s)$-dimensional space. The LSE in the augmented space is $\hat{\theta}'_n = (\frac{1}{n}\sum_{t=1}^n \varphi'_t\varphi'^\top_t)^{-1}\frac{1}{n}\sum_{t=1}^n \varphi'_t(\varphi'^T_t\theta^{*\prime}+N_t) = \theta^{*\prime}+(\frac{1}{n}\sum_{t=1}^n \varphi'_t\varphi'^\top_t)^{-1}\frac{1}{n}\sum_{t=1}^n \varphi'_t N_t$. By (24), the matrix $(\frac{1}{n}\sum_{t=1}^n \varphi'_t\varphi'^\top_t)^{-1}$ converges to a positive definite matrix as $n \to \infty$. Each component of the vector sequence $\{\varphi'_t N_t\}$ is independent zero mean with a variance bounded by $\bar{\sigma}_t^2 = \|\varphi'_t\|^2\mathbb{E}[N_t^2]$, and we have that $\sum_{t=1}^\infty \frac{\bar{\sigma}_t^2}{t^2} < \sqrt{\sum_{t=1}^\infty \frac{\|\varphi'_t\|^4}{t^2}}\sqrt{\sum_{t=1}^\infty \frac{(\mathbb{E}[N_t^2])^2}{t^2}} < \infty$ by the Cauchy-Schwarz inequality and (25), (26). Thus, the Kolmogorov's Strong Law of Large Numbers (Theorem 6 in the Appendix) yields that $\lim_{n\to\infty}(\frac{1}{n}\sum_{t=1}^n \varphi'_t\varphi'^\top_t)^{-1}\frac{1}{n}\sum_{t=1}^n \varphi'_t N_t = 0$ w.p.1 so that $\hat{\theta}'_\infty \triangleq \lim_{n\to\infty}\hat{\theta}'_n = \theta'^*$. Let $B'_\varepsilon(\theta^{*\prime}) = \{\vartheta' \in \mathbb{R}^{\hat{d}+s} : \|\theta^{*\prime} - \vartheta'\| \leq \varepsilon\}$. Theorem 2 ensures that $\widehat{\Theta}'_n \subseteq B'_\varepsilon(\theta^{*\prime})$ for all n large enough, which implies that

$$\{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top\ \mathbf{0}^\top]^\top \in \widehat{\Theta}'_n\} \subseteq \{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top\ \mathbf{0}^\top]^\top \in B'_\varepsilon(\theta^{*\prime})\}$$

for n large enough. The theorem claim follows by recalling that $\widehat{\Theta}_n^{\mathrm{UD}} = \{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top\ \mathbf{0}^\top]^\top \in \widehat{\Theta}'_n\}$ (Fact 1) and noting that $\{\theta \in \mathbb{R}^{\hat{d}} : [\theta^\top\ \mathbf{0}^\top]^\top \in B'_\varepsilon(\theta^{*\prime})\} = B_\varepsilon(\theta^*)$.

□

4.4. UD-SPS in the presence of undermodelling

Consider now the case of a non-matching data generation mechanism.

In this case, we know that the LSE estimate will converge to a deceiving value $\hat{\theta}_\infty$, and the $\widehat{\Theta}_n$ region shrinks around this point. This behaviour can be explained by noting that Y_t can be expressed according to equation

$$Y_t = \widehat{Y}_t(\hat{\theta}_\infty) + E_t,$$

where $E_t \triangleq Y_t - \widehat{Y}_t(\hat{\theta}_\infty)$ is a residual that, in the unmaching case, *does not* satisfy the assumption of independence and symmetry about zero. By design, SPS is myopically bound to the postulated

structure (7) and it has no mechanism to discriminate between a residual like E_t that includes the effect of undermodelled dynamics and a residual that is independent of the input signal (as is in the matching case). UD-SPS, instead, by constructing a region that can be interpreted as a restriction of a virtual region $\widehat{\Theta}'_n$ living in an augmented parameter space, is capable of discovering if the residual can be better described by also using the extra components of ψ_t, thus revealing undermodelling.

Building on the above reasoning, the following theorem asserts that the UD-SPS algorithm issues a "warning" in the non-matching case by generating, for n large enough, an empty region.

The statement of the theorem, equation (29), is formulated using the same notation as in Theorems 2 and 4, and means that, with probability 1, there is a (realization-dependent) value of $\bar{n}$ (which depends on the information content carried by data) such that the region $\widehat{\Theta}_n^{\mathrm{UD}}$ is empty for every $n \geq \bar{n}$. Hence, unlike SPS, UD-SPS has a propensity to warn the user when the method is employed beyond its domain of applicability, and undermodelling manifests itself through the constructions of empty UD-SPS regions.

In the theorem, condition (28) is a technical detectability condition, which formalizes the idea that ψ_t is able to (partly) describe E_t. This condition is mild because full description of E_t is not required and ψ_t needs only to capture some content that is present in E_t (see the remark after the proof of the theorem for more explanation).

Theorem 5 (Undermodelling detection)**.** *Assume that the conditions* (24) (25)*,* (12)*,* (10) *hold true, and that there exists a finite vector* $\bar{Y}_\psi$ *such that* $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \psi_t Y_t = \bar{Y}_\psi$ *w.p.1. With the definition*

$$\bar{Y}_{\varphi'} \triangleq \begin{bmatrix} \bar{Y}_\varphi \\ \bar{Y}_\psi \end{bmatrix}, \tag{27}$$

($\bar{Y}_\varphi$ *is defined in* (10)*), if*

$$(\bar{R}')^{-1} \bar{Y}_{\varphi'} \notin \mathbb{R}^{\hat{d}} \times \{0\}^s, \tag{28}$$

then

$$\mathbb{P}r\left[\bigcup_{\bar{n}=1}^{\infty} \bigcap_{n=\bar{n}}^{\infty} \left\{\widehat{\Theta}_n^{UD} = \emptyset\right\}\right] = 1. \tag{29}$$

*

Proof. The conditions of Theorem 2 are satisfied in the augmented space where φ_t is replaced by ${\varphi'_t}^\top = [\varphi_t^\top \ \psi_t^\top]$. Then, by applying Theorem 2 in the augmented space, we conclude that, with probability 1, the SPS region in the augmented space shrinks uniformly around the limit point $\hat{\theta}'_\infty = \bar{R}'^{-1} \bar{Y}_{\varphi'} \notin \mathbb{R}^{\hat{d}} \times \{0\}^s$. Thus, with probability 1, for n large enough, the region $\widehat{\Theta}'_n$ has empty intersection with the $\hat{d}$-dimensional subspace in which the UD-SPS region is defined (Fact 1) and the theorem statement follows. □

Remark 3 (On the detectability condition (28))**.** *Condition* (28) *can be intuitively explained as follows. Assume that we want to express the residual* E_t *as a linear combination of the input variables in the augmented regressor* $\varphi'_t = [\varphi_t^\top \ \psi_t^\top]^\top$*, that is, we look for the* $\hat{\nu}_n$ *that minimises the sum of the squared residuals* $(E_t - {\varphi'_t}^\top \nu)^2$*. This gives* $\hat{\nu}_n = (R'_n)^{-1} \frac{1}{n} \sum_{t=1}^{n} \varphi'_t E_t$ *and, since* $E_t = Y_t - \varphi_t^\top \hat{\theta}_\infty = Y_t - {\varphi'_t}^\top \begin{bmatrix} \hat{\theta}_\infty \\ \mathbf{0} \end{bmatrix}$*, we get* $\hat{\nu}_n \to \left((\bar{R}')^{-1} \bar{Y}_{\varphi'} - \begin{bmatrix} \hat{\theta}_\infty \\ \mathbf{0} \end{bmatrix}\right)$*. The detectability condition requires that at least one of the last* s *components of this vector be nonzero, and hence the residual* E_t *can be better explained by* ψ_t *and* φ_t *than* φ_t *alone.*

*

5. Computational aspects of UD-SPS and practical usage

The UD-SPS method of Section 4 allows one to easily verify whether a certain value of θ is inside or outside the UD-SPS region. On the other hand, in system identification practice, one often desires to construct the whole region to which the system parameter vector has to belong, which may take a long time if each single point in a grid has to be examined. Moreover, with UD-SPS, it is important to check whether the region is empty, which might be a challenging task to be accomplished via a gridding procedure. In this section, we study the shape of the UD-SPS region, by describing its ellipsoidal "building blocks", which will prove useful to construct suitable approximants of the UD-SPS region.

5.1. Ellipsoids as building blocks

Let $\mathcal{E}_i = \{\theta : \|Q_i(\theta)\|^2 >_\pi \|Q_0(\theta)\|^2\}$, $i = 1, \dots, m-1$. Then, the region $\widehat{\Theta}_n^{\mathrm{UD}}$, (19), can be expressed as

$$\widehat{\Theta}_n^{\mathrm{UD}} = \bigcup_{I \subseteq \{1,\dots,m-1\}, |I|=q} \bigcap_{i \in I} \mathcal{E}_i. \tag{30}$$

5.2. Explicit expression for $\mathcal{E}_i$

Preliminarily, we define the useful shorthands

$$P_n = \frac{1}{n}\sum_{t=1}^{n} \varphi_t Y_t,\ P'_n = \frac{1}{n}\sum_{t=1}^{n} \varphi'_t Y_t,\ \widetilde{P}'_{i,n} = \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\varphi'_t Y_t;$$

$$G_n = \frac{1}{n}\sum_{t=1}^{n} \varphi'_t \varphi_t^\top,\ \widetilde{G}_{i,n} = \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,n}\varphi'_t \varphi_t^\top,$$

$$\widetilde{R}'_{i,n} = \frac{1}{n}\sum_{t=1}^{n} \alpha_{i,t}\varphi'_t {\varphi'_t}^\top,$$

and recall that $R_n = \frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^\top$, $R'_n = \frac{1}{n}\sum_{t=1}^{n} \varphi'_t {\varphi'_t}^\top$. Note that, when $\theta' = \begin{bmatrix} \theta \\ \mathbf{0} \end{bmatrix}$, the following identities are satisfied

$$\theta^\top G_n^\top = {\theta'}^\top R'_n, \qquad {\theta'}^\top R'_n \theta' = \theta^\top R_n \theta, \qquad {\theta'}^\top P'_n = \theta^\top P_n; \tag{31}$$

$$\theta^\top \widetilde{G}_{i,n}^\top = {\theta'}^\top \widetilde{R}'_{i,n}. \tag{32}$$

By definition, $\|Q_0(\theta)\|^2 = (P'_n - G_n\theta)^\top (R'_n)^{-1}(P'_n - G_n\theta) = \theta^\top (G_n^\top (R'_n)^{-1} G_n)\theta - 2\theta^\top G_n^\top (R'_n)^{-1} P'_n + {P'_n}^\top {R'_n}^{-1} P'_n$, which, using the identities (31), simplifies to

$$\begin{aligned} &\|Q_0(\theta)\|^2 \\ &\quad = \theta^\top R_n \theta - 2\theta^\top P_n + {P'_n}^\top (R'_n)^{-1} P'_n \\ &\quad = \theta^\top A_0 \theta - 2\theta^\top b_0 + c_0, \end{aligned} \tag{33}$$

with the obvious definitions of A_0, b_0 and c_0. Similarly,

$$\begin{aligned}
&\|Q_i(\theta)\|^2 \\
&\quad = \theta^\top(\widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n})\theta - 2\theta^\top\widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{P}'_{i,n} + \widetilde{P}'_{i,n}{}^\top(R'_n)^{-1}\widetilde{P}'_{i,n} \\
&\quad = \theta^\top A_i\theta - 2\theta^\top b_i + c_i,
\end{aligned} \tag{34}$$

with the obvious definitions of A_i, b_i and c_i. Except for the case when $A_0 - A_i = 0$,[6] $\|Q_0(\theta)\|^2 - \|Q_i(\theta)\|^2$ is a quadratic function, as a consequence of the fact, which we prove next, that $(A_0 - A_i) \succeq 0$. To show that $(A_0 - A_i) \succeq 0$, we first notice that $(R'_n - \widetilde{R}'_{i,n}{}^T(R'_n)^{-1}\widetilde{R}'_{i,n}) \succeq 0$ (in fact, Lemma 4 of [12] applies verbatim to the augmented matrices R'_n and $\widetilde{R}'_{i,n}$). Then, defining $\theta' = \begin{bmatrix} \theta \\ \mathbf{0} \end{bmatrix}$ and using (32) we have that $\theta^\top(R_n - \widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n})\theta = \theta'^\top(R'_n - \widetilde{R}'_{i,n}{}^T(R'_n)^{-1}\widetilde{R}'_{i,n})\theta' \geq 0$ for every θ, so that $(R_n - \widetilde{G}_{i,n}^T R'^{-1}_n \widetilde{G}_{i,n}) \succeq 0$. Clearly, when the $A_0 - A_i$ matrix is positive definite (as it is expected in practical cases), there is a unique minimizer θ_c of the quadratic function. However, to deal with the possibility that this matrix is only positive semi-definite, we write the (possibly non-unique) mininum point θ_c of $\|Q_0(\theta)\|^2 - \|Q_i(\theta)\|^2$ by resorting to the Moore-Penrose pseudo-inverse:

$$\theta_c = (A_0 - A_i)^\dagger(b_0 - b_i). \tag{35}$$

The corresponding minimum value $v_i \triangleq \|Q_0(\theta_c)\|^2 - \|Q_i(\theta_c)\|^2$ can be directly computed by substitution. Thus, the ellipsoid can be written as

$$\mathcal{E}_i = \{\theta : (\theta - \theta_c)^\top(R_n - \widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n})(\theta - \theta_c) <_\pi -v_i\}.^7$$

This expression of $\mathcal{E}_i$ is handy as it allows to recognize quickly whether the set is empty ($v_i > 0$ implies empty set). Moreover, it allows one to compute useful quantities, such as the volume of the ellipsoid[8], tight axes-aligned bounding boxes[9], etc.

According to (30), when $q = 1$, the UD-SPS region $\widehat{\Theta}_n^{\mathrm{UD}}$ is a simple union of sets $\mathcal{E}_i$. Therefore, in the $q = 1$ case, the analysis above can be directly exploited for a satisfactory computation and evaluation of $\widehat{\Theta}_n^{\mathrm{UD}}$. Choosing $q > 1$ yields a more complex construction that, however, could be beneficial in some respects (e.g., optimality results proven in [20] assume that both q and m increase when $n \to \infty$). For this latter case, we derive in what follows a handy ellipsoidal outer approximation of $\widehat{\Theta}_n^{\mathrm{UD}}$ that builds on the results of this section. It is worth mentioning that other methods that have been explored for the approximation of SPS regions could be applied *mutatis mutandis* to UD-SPS; in particular, the reader is referred to [37] for interval analysis methods and to [38] for sampling schemes.

[6] When $A_0 - A_i = 0$, the function $\|Q_0(\theta)\|^2 - \|Q_i(\theta)\|^2$ degenerates into a simple affine (or constant) function. This eventuality, which is very rare under reasonable input excitation conditions, can be easily detected and dealt with, and is not discussed further here.

[7] In practice, $<_\pi$ can be replaced by the handier $\leq$ at the price of possibly including some points that had been excluded by the tie-break rule π. Moreover, the reader will have noticed that, in the case where the matrix $A_0 - A_i$ is only semidefinite, $\mathcal{E}_i$ extends indefinitely in the directions of eigenvectors corresponding to null eigenvalues.

[8] The volume can be obtained by the formula $\sqrt{\det^{-1}\left(\frac{R_n - \widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n}}{-v_i}\right)}$.

[9] The vector whose components are the lower bounds can be computed as $\theta_c - \sqrt{v_i \mathrm{diag}[(\widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n} - R_n)^{-1}]}$, where $\sqrt{\cdot}$ is component-wise, while the vector with the upper bounds is $\theta_c + \sqrt{v_i \mathrm{diag}[(\widetilde{G}_{i,n}^T(R'_n)^{-1}\widetilde{G}_{i,n} - R_n)^{-1}]}$.

5.3. Ellipsoidal outer approximation for $\widehat{\Theta}_n^{UD}$

To begin with, note that whenever $\mathcal{E}_i$ is nonempty we have

$$\begin{aligned}\mathcal{E}_i \subseteq&\{\theta \in \mathbb{R}^{\hat{d}} : \|Q_0(\theta)\|^2 \le \|Q_i(\theta)\|^2\}\\ \subseteq&\{\theta : \|Q_0(\theta)\|^2 \le \sup_{\bar{\theta}:\|Q_0(\bar{\theta})\|^2\le\|Q_i(\bar{\theta})\|^2} \|Q_i(\bar{\theta})\|^2\}\\ =&\{\theta : \|Q_0(\theta)\|^2 \le \sup_{\|Q_0(\bar{\theta})\|^2\le\|Q_i(\bar{\theta})\|^2} \|Q_0(\bar{\theta})\|^2\}.\end{aligned}$$

Using (33) and (34), the value $\sup_{\|Q_0(\bar{\theta})\|^2\le\|Q_i(\bar{\theta})\|^2} \|Q_0(\bar{\theta})\|^2$ can be written as the solution to the optimization problem

$$\begin{aligned}\text{minimize}_{\theta\in\mathbb{R}^{\hat{d}}} \quad & -\theta^\top A_0\theta + 2\theta^\top b_0 - c_0\\ \text{subject to} \quad & \theta^\top(A_0 - A_i)\theta - 2\theta^\top(b_0 - b_i) + c_0 - c_i \le 0.\end{aligned} \tag{36}$$

The optimal value of this non-convex program can be obtained by solving the following convex semidefinite program (SDP), see Appendix B in [39].

$$\begin{aligned}\text{maximize}_{\gamma,\lambda} \quad & \gamma\\ \text{subject to} \quad & \lambda \ge 0\\ & \begin{bmatrix} \lambda(A_0 - A_i) - A_0 & \lambda(b_i - b_0) + b_0\\ (\lambda(b_i - b_0) + b_0)^\top & \lambda(c_0 - c_i) - \gamma - c_0 \end{bmatrix} \ge 0.\end{aligned} \tag{37}$$

Denoting by γ_i^* the optimal value, and repeating the procedure for all $i = 1, \dots, m-1$ for which $\mathcal{E}_i$ is non-empty (set $\gamma_i = -\infty$ otherwise), we define $\gamma_*^{(q)}$ as the q-th largest of γ_i, $i = 1, \dots, m-1$. Then we have that any point θ that is contained in at least q of the sets $\mathcal{E}_i$ cannot be outside the ellipsoid

$$\widetilde{\Theta}_n^{\mathrm{UD}} \;=\; \{\,\theta \in \mathbb{R}^{\hat{d}} : \|Q_0(\theta)\|^2 \le \gamma_*^{(q)}\,\}, \tag{38}$$

and therefore $\widetilde{\Theta}_n^{\mathrm{UD}}$ certainly contains $\widehat{\Theta}_n^{\mathrm{UD}}$ by (30).

6. A numerical experiment

The true data-generation mechanism is the following FIR(2) system

$$Y_t \;=\; 0.75U_{t-1} - 0.3U_{t-2} + N_t, \tag{39}$$

where $\{N_t\}$ is a sequence of i.i.d. Laplacian random variables with zero mean and variance 0.1. The input signal is generated as

$$U_t \;=\; -0.5\,U_{t-1} + V_t,$$

where $\{V_t\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance 1, independent of N_t.

The simple identification problem that we provide here to illustrate some basic facts of UD-SPS is described as follows: the sequence $Y_1, \dots, Y_n$, $n = 50$, is available, together with the measured input $U_1, \dots, U_{50}$, and we want to build a confidence region for the model parameters at probabilistic level 80%. In what follows, we will compare the output of SPS and UD-SPS on

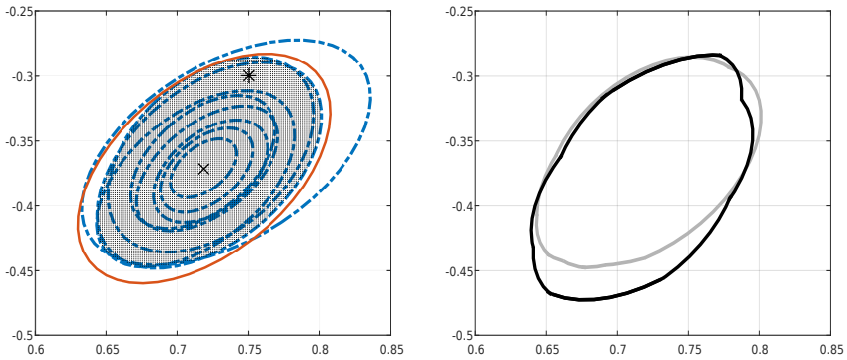

Figure 1: Left – the shaded area is the UD-SPS region at level 80%, $\widehat{\Theta}_n^{\mathrm{UD}}$, with $n = 50$ data, in the (b_1, b_2) space. The star $*$ denotes the true parameter value; the cross $\times$ is the LSE. The dashed ellipses are the contours of the sets $\mathcal{E}_i$. The solid red ellipsoid is the ellipsoidal outer approximation of $\widehat{\Theta}_n^{\mathrm{UD}}$. Right – the contour of $\widehat{\Theta}_n^{\mathrm{UD}}$ (lighter line) is compared with the the contour of the SPS region (darker line) for the same set of data.

repeated experiments. In all simulations, m and q are set to 10 and 2 respectively, with the aim of achieving an 80% confidence level.

First, we simulated on 10,000 independent experiments the behaviour of a user who has postulated a FIR(1) model and uses SPS and UD-SPS to estimate the unknown parameter. The SPS region was obviously always nonempty as it had to contain the LS estimate. On the other hand, UD-SPS with $s = 1$ detected undermodelling by yielding an empty region in all but 14 cases out of 10,000. Moreover, the number of cases where the region was not empty went down to zero as the number of data points was increased to $n = 70$ data points. This demonstrates the effectiveness of the new UD-SPS in detecting undermodelling when it occurs.

Then, again with $n = 50$, we applied SPS and UD-SPS under the correct system specification, i.e., FIR(2). UD-SPS was applied with $s = 1$ (therefore, $\varphi_t = [U_{t-1}, U_{t-2}]^\top$ and $\psi_t = [U_{t-3}]$). The UD-SPS region included the true parameters $b_1^* = 0.75$, $b_2^* = -0.3$ in 8,012 cases out of 10,000, and SPS in 8,057 out of 10,000 cases. These results are in agreement with the theoretical guarantee that the constructed region is an 80% coverage region. The region $\widehat{\Theta}_n^{\mathrm{UD}}$ at level 80% obtained in one of these 10,000 experiments is represented by the shaded area in Figure 1, on the left. The dashed lines are the contours of the 9 ellipsoidal sets $\mathcal{E}_i$, $i = 1, \ldots, m-1$, which are the building blocks to construct all the regions at probability levels $(1 - \frac{j}{9})$, $j = 1, \ldots, 9$: for example, the region of interest, at level 80%, is the region containing the points at the intersection of at least two sets $\mathcal{E}_i$; if instead one wants to visualize the region at level 90%, this is the union of all the $\mathcal{E}_i$'s, while the region at level 10% (the smallest UD-SPS region, and with the less confidence, that can be constructed for the present choice of the parameter m) is the small region at the intersection of all the $\mathcal{E}_i$'s. The ellipsoidal outer approximation of $\widehat{\Theta}_n^{\mathrm{UD}}$ (constructed using the procedure in Section 5) is represented in the figure as a solid-line ellipse in red. Note that the UD-SPS region does not intersect the horizontal axis ($b_2 = 0$): this means that in this case UD-SPS would detect undermodelling when applied with $\hat{d} = 1$ and $s = 1$. Figure 1, on the right, shows together the contour of the UD-SPS region (lighter line) and the contour of the SPS region (darker line) constructed from the same set of data.

7. Conclusions and future work

In this paper we have extended the study of SPS, a statistical, guaranteed, finite sample system identification method, to the case of undermodelled dynamics. While undermodelled dynamics is not covered by the standard SPS algorithm, we have also shown that the SPS algorithm does not contain any mechanism to warn the user that the algorithm is working outside of its domain of applicability.

We have then proposed an extension of SPS, called UD-SPS, which is able to detect that the algorithm is working outside of its domain of applicability as soon as a sufficient amount of data is collected. We have shown that UD-SPS provides guaranteed confidence regions if the model order is correctly specified, otherwise it almost surely detects undermodelling in the long run.

The findings presented in this paper should not be seen as setting a final word, rather the authors conceive this work as a stepping stone towards future investigations. Driven by the observation that the standard SPS algorithm has no internal mechanisms to invalidate an unsuitable model class, our effort has been geared towards manufacturing a new scheme that warns the user in case an inappropriate model class is being used, while also preserving the original ability to provide guaranteed regions when the model class is correct. The next step we envisage is to exploit these new ideas as a means for complexity control in adaptive model selection schemes with probabilistic guarantees: as soon as data are informative enough for UD-SPS to detect that a model class is inadequate, the model class is replaced with a more suitable one. Setting up this iterative procedure is at the present time still an open research plan. One fundamental aspect that needs to be better characterized is whether a model class that has yet not been invalidated by UD-SPS remains appropriate for the intended use of the model. This is fundamental to allow for a seamless transition from one model class to the next while getting the system to suffer moderate distress. We hope that the present paper will offer a solid platform to foster activities in these important directions.

8. Acknowledgement

The authors would like to thank Gianluigi Pillonetto for discussions on the computational aspects of the UD-SPS region, in particular, for drawing our attention to the notable computational advantages of the $q = 1$ case.

References

[1] L. Ljung, System Identification: Theory for the User, 2nd Edition, Prentice-Hall, Upper Saddle River, 1999.
[2] L. Ljung, Perspectives on system identification, Annual Reviews in Control 34 (1) (2010) 1–12.
[3] T. Söderström, P. Stoica, System Identification, Prentice Hall International, Hertfordshire, UK, 1989.
[4] M. Milanese, J. Norton, H. Piet-Lahanier, E. Walter, Bounding approaches to system identification, Springer Science & Business Media, 2013.
[5] M. Milanese, M. Taragna, H_∞ set membership identification: A survey, Automatica 41 (12) (2005) 2019 – 2032.
[6] M. Milanese, C. Novara, Unified set membership theory for identification, prediction and filtering of nonlinear systems, Automatica 47 (10) (2011) 2141 – 2151.
[7] M. Karimshoushtari, C. Novara, Design of experiments for nonlinear system identification: A set membership approach, Automatica 119 (2020) 109036.
[8] F. Dabbene, M. Sznaier, R. Tempo, Probabilistic optimal estimation with uniformly distributed noise, IEEE Transactions on Automatic Control 59 (8) (2014) 2113–2127.
[9] M. C. Campi, E. Weyer, Guaranteed non-asymptotic confidence regions in system identification, Automatica 41 (2005) 1751–1764.

[10] M. Dalai, E. Weyer, M. C. Campi, Parameter identification for non-linear systems: guaranteed confidence regions through LSCR, Automatica 43 (2007) 1418–1425.
[11] O. N. Granichin, The nonasymptotic confidence set for parameters of a linear control object under an arbitrary external disturbance, Automation and Remote Control 73 (1) (2012) 20–30.
[12] B. Cs. Csáji, M. C. Campi, E. Weyer, Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models, IEEE Transactions on Signal Processing 63 (1) (2015) 169–181.
[13] S. Kolumbán, I. Vajk, J. Schoukens, Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets, Automatica 51 (2015) 326–331.
[14] K. Amelin, O. Granichin, Randomized control strategies under arbitrary external noise, IEEE Transactions on Automatic Control 61 (5) (2016) 1328–1333.
[15] M. Volkova, O. Granichin, Y. Petrov, G. Volkov, Dynamic fracture tests data analysis based on the randomized approach, Advances in Systems Science and Applications 17 (3) (2017) 34–41.
[16] C.-Y. Han, M. Kieffer, A. Lambert, Guaranteed confidence region characterization for source localization using rss measurements, Signal Processing 152 (2018) 104–117.
[17] B. Cs. Csáji, K. B. Kis, Distribution-free uncertainty quantification for kernel methods by gradient perturbations, Machine Learning 108 (8) (2019) 1677–1699.
[18] M. M. Khorasani, E. Weyer, Non-asymptotic confidence regions for the parameters of EIV systems, Automatica 115 (2020) 108873.
[19] A. Carè, B. Cs. Csáji, M. C. Campi, E. Weyer, Finite-sample system identification: An overview and a new correlation method, IEEE Control Systems Letters 2 (1) (2018) 61–66.
[20] E. Weyer, M. C. Campi, B. Cs. Csáji, Asymptotic properties of SPS confidence regions, Automatica 82 (2017) 287 – 294.
[21] P. Stoica, Y. Selen, Model-order selection: a review of information criterion rules, IEEE Signal Processing Magazine 21 (4) (2004) 36–47.
[22] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, L. Ljung, Kernel methods in system identification, machine learning and function estimation: A survey, Automatica 50 (3) (2014) 657–682.
[23] L. Ljung, T. Chen, B. Mu, A shift in paradigm for system identification, International Journal of Control 93 (2) (2020) 173–180.
[24] D. Piga, M. Forgione, S. Formentin, A. Bemporad, Performance-oriented model learning for data-driven MPC design, IEEE Control Systems Letters 3 (3) (2019) 577–582.
[25] M. Scandella, M. Mazzoleni, S. Formentin, F. Previdi, Kernel-based identification of asymptotically stable continuous-time linear dynamical systems, International Journal of Control.
[26] R. J. Brooks, A. M. Tobias, Choosing the best model: Level of detail, complexity, and model performance, Mathematical and Computer Modelling 24 (4) (1996) 1 – 14.
[27] C. Cobelli, F. Bettini, A. Caumo, M. J. Quon, Overestimation of minimal model glucose effectiveness in presence of insulin response is due to undermodeling, American Journal of Physiology-Endocrinology and Metabolism 275 (6) (1998) E1031–E1036.
[28] A. E. Nordsjö, T. Wigren, On estimation of errors caused by non-linear undermodelling in system identification, International Journal of Control 75 (14) (2002) 1100–1113.
[29] K. A. McLean, K. B. McAuley, Mathematical modelling of chemical processes –obtaining the best model predictions and parameter estimates using identifiability and estimability procedures, The Canadian Journal of Chemical Engineering 90 (2) (2012) 351–366.
[30] M. C. Campi, S. Ko, E. Weyer, Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics, Automatica 45 (2009) 2175–2186.
[31] A. Carè, M. C. Campi, B. Cs. Csáji, E. Weyer, Undermodelling detection with sign-perturbed sums, IFAC-PapersOnLine 50 (1) (2017) 2744–2749.
[32] A. Carè, B. Cs. Csáji, M. C. Campi, Sign-Perturbed Sums (SPS) with asymmetric noise: Robustness analysis and robustification techniques, in: Procs. of the 55th IEEE Conference on Decision and Control, 2016, pp. 262–267.
[33] B. Wahlberg, System identification using laguerre models, IEEE Transactions on Automatic Control 36 (5) (1991) 551–562.
[34] B. Wahlberg, P. Mäkilä, On approximation of stable linear dynamical systems using Laguerre and Kautz functions, Automatica 32 (5) (1996) 693 – 708.
[35] B. Ninness, F. Gustafsson, A unifying construction of orthonormal bases for system identification, IEEE Transactions on automatic control 42 (4) (1997) 515–521.
[36] P. S. Heuberger, P. M. Van den Hof, B. Wahlberg, Modelling and identification with rational orthogonal basis functions, Springer Science & Business Media, 2005.
[37] M. Kieffer, E. Walter, Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS, Automatica 49 (2013) 507–512.

[38] A. Carè, G. Pillonetto, M. C. Campi, Uncertainty bounds for kernel-based regression: A Bayesian SPS approach, in: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), 2018, pp. 1–6.
[39] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2009.
[40] A. N. Shiryaev, Probability, 2nd Edition, Springer, 1995.

Appendix A. Useful results

Theorem 6 (Kolmogorov's Strong Law of Large Numbers, [40]). *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

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

Theorem 7 (Doob, [40]). *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 with probability 1, the limit $\lim_{n\to\infty} \xi_n = \xi_\infty$ exists and $\mathbb{E}[|\xi_\infty|] < \infty$.*

Theorem 8 (Kronecker's lemma, [40]). *Let $\{b_t\}$ a sequence of positive increasing numbers such that $\lim_{t\to\infty} b_t = \infty$, and let x_t be a sequence of numbers such that $\sum_{t=1}^{\infty} x_t$ converges. Then, $\lim_{n\to\infty} \frac{1}{b_n} \sum_{t=1}^{n} b_t x_t = 0$. In particular, if $b_t = t$, $x_t = \frac{y_t}{t}$ and $\sum_{t=1}^{\infty} \frac{y_t}{t}$ converges, then $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} y_t = 0$.*