Instrumental Varaibles Based Least Squares Support Vector Machine for Identification of Nonlinear Systems [⋆]

Vincent Laurain [a], Roland Tóth [b], Dario Piga [b], Wei Xing Zheng [c],

[a] *Centre de Recherche en Automatique de Nancy, Université de Lorraine, CNRS, 2 rue Jean Lamour, 54519 Vandoeuvre-lès-Nancy Cedex, France.*

[b] *Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands.*

[c] *School of Computing and Mathematics, University of Western Sydney, Penrith NSW 2751, Australia.*

Abstract

Least-Squares Support Vector Machines (LS-SVM's), originating from statistical learning theory, represent a promising approach to identify nonlinear systems via nonparametric estimation of nonlinearities in a computationally and stochastically attractive way. However, application of LS-SVM's in the identification context is formulated as a linear regression aiming at the minimization of the ℓ_2 loss in terms of the prediction error. This formulation corresponds to the assumption of an auto-regressive noise structure, which, especially in the nonlinear context, is often found to be too restrictive in practical applications. In this paper, Instrumental Variable (IV) based estimation is integrated into the LS-SVM approach providing, under minor conditions, a consistent identification of nonlinear systems in case of a noise modeling error. It is shown how the cost function of the LS-SVM is modified to achieve an IV-based solution. Although, a practically well applicable choice of the instrumental variable is proposed for the derived approach, optimal choice of this instrument in terms of the estimates associated variance still remains to be an open problem. The effectiveness of the proposed IV based LS-SVM scheme is also demonstrated by a representative example based on a Monte Carlo study.

Key words: support vector machines; instrumental variables; nonlinear identification; machine learning; non-parametric estimation.

1 Introduction

Support vector machines (SVM's) have been originally developed as a class of *supervised learning* methods in stochastic learning theory. Their original purpose was to provide efficient tools for data analysis and pattern recognition in classification problems and regression analysis [1,2]. SVMs have had a paramount impact on the *machine learning* field since their extension as a theoretical framework in that setting [3]. These methods also offer an attractive, so-called non-parametric way of data-driven dynamic modeling, *i.e.*, *system identification*, especially in the nonlinear context. In that context, these approaches are part of the data-driven model learning avenue [4,5], focusing on the paradigm of estimation of the targeted system without posing prior assumptions on their dynamical nature or the non-linearities involved. Most of the research interest regarding identification with SVM's has been dedicated to *nonlinear block models* so far, using various *least-square* SVM (LS-SVM) approaches where the original nonlinear estimation problem is posed as a linear regression [6–8]. In general, LS-SVM's are particular variations of the original support vector machine approach using an ℓ_2 loss function on the prediction error of the model. They also have lot of similarities to *Gaussian Processes* where identification of nonlinear models have been also considered recently in a linear regression setting. Particular advantages of these approaches are the convex

[⋆] This paper was not presented at any IFAC meeting. Corresponding author Roland Tóth. Tel. +31-40-247-2655. Fax +31-40-243-4582.

[⋆⋆]This work was supported by the Netherlands Organization for Scientific Research (NWO, grant no.: 639.021.127) and by the French ministries of Foreign Affairs, Education and Research and the French-Dutch Academy (PHC Van Gogh project, n. 29342QL).

Email addresses: vincent.laurain@univ-lorraine.fr (Vincent Laurain), r.toth@tue.nl (Roland Tóth), d.piga@tue.nl (Dario Piga), W.Zheng@uws.edu.au (Wei Xing Zheng).

nature of the corresponding optimization problem and an attractive trade-off between regularization bias and variance of the estimates [7].

Given the large number of parameters typically involved in LS-SVM's, these approaches can also be seen as so-called *over-parametrization methods* in the nonlinear framework [9,10]. However, due to the existence of powerful regularization methods for SVM's [1,2], the variance of the estimated nonlinear functions is significantly lower than in the classical over-parametrization approaches [7]. On the other hand, LS-SVM's also offer the possibility of incorporating a model structure and prior knowledge on the nonlinearities unlike other nonparametric methods (*e.g.*, [11]). The latter is an important property, providing perspectives of joining structural information handling and the power of prior-free modeling – a much need improvement in system identification as pointed out in [12] –.

A particular handicap of the variants of LS-SVM's (and also GP's) is that the used linear regression form under the ℓ_2 loss function corresponds to the assumption of an auto-regressive noise structure, which, especially in the nonlinear context, is often found to be too restrictive in practical applications. In the classical identification literature, significant research efforts have been devoted to achieve consistent estimation in case of rather general noise assumptions corresponding to the situations commonly encountered in practice [13]. To introduce the same generality of noise structures, some steps have been taken in the LS-SVM context such as the recurrent LS-SVM developed in [14] and the linear parametric noise model equipped SVM derived in [15]. However, the classical results in identification highlight that the chosen noise model, *i.e.*, the prejudice on the assumed noise, plays an important role in the consistency of the estimates. Therefore, in the light of a non-parametric prior-free modeling objective, the question rises why we should bound ourself to a priori specified noise assumption, especially in the general nonlinear context. By turning to the classical results, we can find that variants of linear regression based methods, *e.g.*, *instrumental variable* (IV) approaches, have been developed to cope with realistic assumptions on the noise without specifying a direct parametrization or structure [13,16]. The strength of IV methods in the LTI case has been found in delivering consistent estimates independently on the chosen noise model assumption in a computationally attractive way [17]. Consequently, in order to take the next step on the data-driven model learning avenue, it is required to mitigate our prior assumptions on the noise.

To achieve this objective, in this paper we consider the idea of introducing the IV scheme into the LS-SVM regression structure, which was first proposed in[18]. As a significant improvement of the initial scheme described in [18], in this paper, we provide a rigorous treatment of instrumental variables based LS-SVM's, showing that an *instrumental* LS-SVM (IV-SVM) method can be derived via the dual solution of the IV optimization problem [19,13]. Furthermore, this contribution not only preserves the computationally attractive feature of the original approach by satisfying the Mercer conditions, but also provides unbiased estimates for general noise model structures/conditions; opening a large set of application areas for data-driven model learning.

The paper is organized as follows: the considered identification problem setting is introduced in Section 2. This is followed in Section 3 by the derivation of the primal and dual solutions for the ℓ_2 optimization problem associated with LS-SVM methods, pointing out the stochastic shortcoming of this scheme under general noise conditions. In Section 4, the IV-based estimation associated optimization problem is introduced and solutions are derived both in the primal and dual forms leading to the core result of the IV-SVM approach. This is followed by integrating the dual IV solution into the LS-SVM estimation scheme for nonlinear dynamic systems resulting in the IV-SVM method. In Section 5, implementation of the proposed estimation scheme is discussed together with the selection of kernel functions and tuning of the hyper parameters. To demonstrate the advantages of the IV-SVM, a Monte Carlo study in Section 6 is provided in which the identification of a nonlinear system with a colored noise is analyzed. Finally, conclusions and some future directions of research are given in Section 7.

2 Problem description

In order to set the stage for the upcoming discussion, the considered identification problem is defined in this section, showing what advantages are enjoyed by non-parametric approaches in contrast to over-parametrization based methods.

2.1 The data-generating system

As an objective of the identification scenario, the data-driven estimation of a rather general nonlinear discrete-time system $\mathcal{S}_\mathrm{o}$ with affine nonlinearities is considered. For the sake of simplicity regarding the upcoming derivations, the system $\mathcal{S}_\mathrm{o}$ is assumed to be *single-input single-output* (SISO). The behavior of $\mathcal{S}_\mathrm{o}$ is described by the following difference equation

$$y(k) = \sum_{i=1}^{n_\mathrm{a}} f_i^\mathrm{o}\big(y(k-i)\big) + \sum_{j=0}^{n_\mathrm{b}} g_j^\mathrm{o}\big(u(k-j)\big) + c^\mathrm{o} + v_\mathrm{o}(k), \quad (1)$$

where u and y are the input and output signals of $\mathcal{S}_\mathrm{o}$ corresponding to a valid *input-output* (I/O) partition, $k \in \mathbb{Z}$ denotes the discrete time, $c^\mathrm{o} \in \mathbb{R}$ is a constant, $f_i^\mathrm{o}, g_j^\mathrm{o} : \mathbb{R} \to \mathbb{R}$ are a set of possibly nonlinear functions being bounded and sufficiently smooth on $\mathbb{R}$ with $f_i^\mathrm{o}(0) = 0$, $g_j^\mathrm{o}(0) = 0$ centered and $v_\mathrm{o}(k)$ is a zero-mean stochastic noise process (not necessarily white). Note that the system defined by (1) is general enough to describe usual block structures such as *Hammerstein* or *Wiener* systems. Formulation of (1) in the *multi-input multi-output* (MIMO) case is also available as shown in [7]. Note that in case $v_{\mathrm{o}} = e_{\mathrm{o}}$, where e_{o} is white, (1) can be seen as a specific *nonlinear auto regressive with exogenous input* (NARX) model. However, it is important to note that this case corresponds to a more restrictive system class than the NARX class defined in [20]:

$$y(k) = f_{\mathrm{o}}\big(x(k)\big) + c^{\mathrm{o}} + v_{\mathrm{o}}(k), \tag{2}$$

where

$$x(k)=[y(k-1)\dots y(k-n_{\mathrm{a}})\ u(k)\dots\ u(k-n_{\mathrm{b}})]^\top,$$

$x(k) \in \mathbb{R}^{n_{\mathrm{g}}}$, $n_{\mathrm{g}} = n_{\mathrm{a}} + n_{\mathrm{b}} + 1$ and $f_{\mathrm{o}} : \mathbb{R}^{n_{\mathrm{g}}} \to \mathbb{R}$ is bounded, nonlinear and zero at the origin. We are also going to analyze the applicability of the proposed SVM approach w.r.t. such a general system, but to preserve tractability of the notation, the main line of discussion is based on (1).

2.2 The modeling paradigm

In general, the main difficulty in parametric identification of nonlinear systems such as (1) is that the involved nonlinearities are *a priori* unknown – a natural expectation for modeling, but a rather difficult problem to be handled using only measurements –. To shed light on the nature of the problem, consider the first step in setting up an estimation problem by which we are able to find an "adequate" model of the data-generating system. The first step is the definition of a parametric model set $\mathcal{M} = \{\mathcal{M}_\theta \mid \theta \in \Theta\}$ with parameters $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ in which this most "adequate" model $\mathcal{M}_\theta$ is searched for. The adequateness of the estimated model $\mathcal{M}_\theta \in \mathcal{M}$ is assessed using a cost a function or error measure $\mathcal{V}(\mathcal{M}_\theta, \mathcal{D}_N)$ w.r.t. a given data set $\mathcal{D}_N = \{y(k), u(k)\}_{k=1}^N$ generated by $\mathcal{S}_{\mathrm{o}}$. However, in order to choose a model set $\mathcal{M}$ able to describe the dynamics of (1) accurately, the parametrization involved must represent a wide class of nonlinearities. As the nonlinearities and the associated dynamics can be arbitrary, such an objective is hopeless to be achieved without adequate prior assumptions. Furthermore, the more degrees of freedom, *i.e.*, over-parametrization, are present in $\mathcal{M}$, the more sensitive the estimation problem is to the corruption of noise, which increases the variance of the estimates, and the more demanding the conditions are on the excitation signals. On the other hand, if $\mathcal{M}$, *i.e.*, the associated parametrization is not complex enough, then the dynamics of $\mathcal{S}_{\mathrm{o}}$ cannot be captured and the estimation process ultimately leads to a structural bias. This reflection on the well-known bias/variance trade-off problem sets the primary objective to finding a suitable nonlinear parametric model set $\mathcal{M}$ (*e.g.*, parametrized in terms of suitable basis functions) which can precisely describe the nonlinearities while keeping the number of parameters as low as possible [20]. This translates to considering the model structure selection and the choice of the parametrization to be the part of the estimation problem itself.

Besides of many sparse estimator and regularization based approaches, *e.g.*, [21–25], the so-called class of "non-parametric" identification methods, like the LS-SVM's, aim to achieve this objective via an implicit parametrization of the data relations. The assumption is made that each nonlinearity $f(\centerdot)$ can be modeled as the projection $\phi^\top(\centerdot)\theta$ by using an n_{H} dimensional mapping $\phi : \mathbb{R}^{n_{\mathrm{g}}} \to \mathbb{R}^{n_{\mathrm{H}}}$ (where n_{H} is potentially infinite) from the space of input-output samples to the so called feature space of the output samples. This so called feature map ϕ and the parameters θ are estimated together using the concepts of the *reproducing kernel theory* [26] without requiring from the user to define a parametrization of f explicitly.

Before properly addressing the LS-SVM problem and in order to clearly develop the motivations for the proposed approach, it is assumed that each nonlinearity in (1) can be written as a function expansion:

$$f_i^{\mathrm{o}}\big(y(k-i)\big) = \sum_{l=1}^{n_{\mathrm{H}}} \theta_{i,l}^{\mathrm{o}} \phi_{i,l}\big(y(k-i)\big), \tag{3a}$$

$$g_j^{\mathrm{o}}\big(u(k-j)\big) = \sum_{l=1}^{n_{\mathrm{H}}} \theta_{j+n_{\mathrm{a}}+1,l}^{\mathrm{o}} \phi_{j+n_{\mathrm{a}}+1,l}\big(u(k-j)\big), \tag{3b}$$

where $\{\phi_{i,l} : \mathbb{R} \to \mathbb{R}\}_{i=1,l=1}^{n_{\mathrm{g}},n_{\mathrm{H}}}$ are *a priori* unknown zero centered functional basis over a function space $\mathcal{Q}(\mathbb{R}) \subset \mathbb{R}^{\mathbb{R}}$, for example, the set of real continuous functions $\mathcal{C}(\mathbb{R})$, and $\{\theta_{i,l}^{\mathrm{o}} \in \mathbb{R}\}_{i=1,l=1}^{n_{\mathrm{g}},n_{\mathrm{H}}}$ are constant parameters. This assumption, which is usually taken in over-parameterization methods altogether with the a priori selection of each $\phi_{i,l}$, leads to the parametrized model $\mathcal{M}_\theta$ described as

$$y(k) = \varphi^\top(k)\ \theta + e(k), \tag{4}$$

where $e(k)$ qualifies as the *prediction error*. The regressor $\varphi(k)$ and the parameter vector θ are n_θ-dimensional vectors, with $n_\theta = (n_{\mathrm{a}} + n_{\mathrm{b}} + 1)n_{\mathrm{H}} + 1$, defined as

$$\varphi(k) = \Big[\ 1\ \ \phi_1^\top\big(y(k-1)\big)\ \dots\ \phi_{n_{\mathrm{a}}}^\top\big(y(k-n_{\mathrm{a}})\big)$$
$$\phi_{n_{\mathrm{a}}+1}^\top\big(u(k)\big)\ \dots\ \phi_{n_{\mathrm{g}}}^\top\big(u(k-n_{\mathrm{b}})\big)\ \Big]^\top, \tag{5a}$$

$$\theta = \Big[\ c\ \ \theta_1^\top\ \dots\ \theta_{n_{\mathrm{g}}}^\top\ \Big]^\top, \tag{5b}$$

where $\phi_i(\centerdot) = [\ \phi_{i,1}(\centerdot)\ \dots\ \phi_{i,n_{\mathrm{H}}}(\centerdot)\ ]^\top$, $c \in \mathbb{R}$ and $\theta_i = [\theta_{i,1}\ \dots\ \theta_{i,n_{\mathrm{H}}}]^\top$. Under this setting, $\mathcal{S}_{\mathrm{o}}$ belongs to the model set $\mathcal{M} = \{\mathcal{M}_\theta \mid \theta \in \mathbb{R}^{n_\theta}\}$, *i.e.*, the collection of all models in the form of (4). Therefore, there exists a $\theta_\mathrm{o} \in \mathbb{R}^{n_\theta}$ such that

$$y(k) = \varphi^\top(k)\ \theta_\mathrm{o} + v_\mathrm{o}(k). \tag{6}$$

Note that such a regression form can also be generalized for (2). As discussed in Section 1, it is important to find the feature maps ϕ_i, based on the given data set $\mathcal{D}_N$, which can achieve a good trade-off between the following objectives:

- minimize n_H, *i.e.*, the number of estimated parameters (minimizing the variance of θ);
- represent each function f_i^o and g_j^o with minimal error (minimizing the structural bias).

Note that in case the feature maps ϕ_i would be known, a regularization based estimation of θ could be applied to achieve such a trade-off. Next, a particular regularized estimator is considered which gives the possibility to estimate θ and ϕ_i together via the so-called *kernel trick*.

3 Duals and the support vector machine

To characterize an estimate for (4) based on data, the quality of the model fit is formulated in terms of the cost function (error measure) $\mathcal{V}(\mathcal{M}_\theta, \mathcal{D}_N)$, which in case of LS-SVM's is a regularized *least-squares* (LS) criterion denoted in a compact form as:

$$\mathcal{V}(\theta, e) = \frac{1}{2}\theta^\top\theta + \frac{\gamma}{2N}\sum_{k=1}^{N} e^2(k) = \frac{1}{2}\|\theta\|_{\ell_2}^2 + \frac{\gamma}{2N}\|e(k)\|_{\ell_2}^2, \tag{7}$$

where $e(k) = y(k) - \varphi^\top(k)\ \theta$ is the *prediction error* w.r.t. $\mathcal{D}_N$ and the scalar $\gamma > 0$ is the *regularization parameter*. In case ϕ_i's are known, the estimate of the parameters θ (based on a data set $\mathcal{D}_N$) is the solution of the following minimization problem:

$$\min_{\theta, e}\ \mathcal{V}(\theta, e), \tag{8a}$$

$$\text{s.t.}\ \ e(k) = y(k) - \varphi^\top(k)\ \theta, \quad k = 1, \dots, N, \tag{8b}$$

As the dimension n_H of the regressor φ is usually large (and potentially infinite), hence the use of the regularization term $\|\theta\|_{\ell_2}^2$, whose importance is characterized by γ in (7), is essential to achieve an efficient solution in terms of the bias/variance trade-off. Hence (7) is a so-called *sum-of-norms* criterion. In order to construct an estimate of both the feature maps and the parameters together, it is necessary to develop the solution of problem (8) both in a primal and in a dual form.

3.1 Solution in the primal form

The primal solution of problem (8) implicitly assumes that the regressor terms w.r.t. ϕ_i are given and well defined. This allows to obtain the primal solution by substituting (8b) into the objective function $\mathcal{V}(\theta, e)$ and then deriving the analytical solution of

$$\frac{\partial \mathcal{V}(\theta, e)}{\partial \theta} = 0. \tag{9}$$

This minimum for $\mathcal{V}(\theta, e)$ is achieved at:

$$\hat{\theta}_\mathrm{P} = \left(\frac{1}{\gamma}I_{n_\theta} + \frac{1}{N}\sum_{k=1}^{N}\varphi(k)\varphi^\top(k)\right)^{-1} \cdot \left(\frac{1}{N}\sum_{k=1}^{N}\varphi(k)y(k)\right), \tag{10}$$

where I_{n_θ} denotes the identity matrix of size n_θ. By using the notation

$$Y = [\, y(1)\ \dots\ y(N)\,]^\top \in \mathbb{R}^N, \tag{11a}$$

$$\Phi = [\, \varphi(1)\ \dots\ \varphi(N)\,]^\top \in \mathbb{R}^{N \times n_\theta}, \tag{11b}$$

the primal solution in (10) can be written as:

$$\hat{\theta}_\mathrm{P} = \underbrace{\left(\frac{1}{N}\Phi^\top\Phi + \frac{1}{\gamma}I_{n_\theta}\right)^{-1}}_{R_\mathrm{P}(\gamma, N)} \frac{1}{N}\Phi^\top Y. \tag{12}$$

3.2 Solution in the dual form

The solution of (8) can also be obtained in a dual form which allows to avoid the implicit assumption that the regressor terms are defined and all equality constraints are satisfied. The dual solution of (8) is obtained by constructing the *Lagrangian*:

$$\mathcal{L}(\theta, e, \alpha) = \mathcal{V}(\theta, e) - \sum_{k=1}^{N}\alpha_k\left(\varphi^\top(k)\ \theta + e(k) - y(k)\right), \tag{13}$$

with $\alpha_k \in \mathbb{R}$ being the Lagrangian multipliers. The global optimum is obtained when the *Karush-Kuhn-Tucker* (KKT) conditions:

$$\frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \qquad \alpha_k = \frac{\gamma}{N}e(k), \tag{14a}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \rightarrow \qquad y(k) = \varphi^\top(k)\ \theta + e(k), \tag{14b}$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \quad \rightarrow \qquad \theta = \sum_{k=1}^{N}\alpha_k\varphi(k). \tag{14c}$$

are fulfilled for all $k = 1, \dots, N$. Substitution of (14a) and (14c) into (14b) leads to

$$y(k) = \underbrace{\varphi^\top(k)\left(\sum_{k=1}^{N}\alpha_k\varphi(k)\right)}_{\theta} + \underbrace{\gamma^{-1}N\alpha_k}_{e(k)}, \tag{15}$$

for $k \in \{1, \ldots, N\}$. The set of linear equations in (15) can be written as

$$Y = \left[\Phi\Phi^\top + \frac{N}{\gamma} I_N\right] \alpha \tag{16}$$

where $\alpha = [\alpha_1 \ \ldots \ \alpha_N]^\top \in \mathbb{R}^N$ and I_N is the identity matrix of size N. The Lagrangian multipliers α are then given by

$$\alpha = \underbrace{\left(\frac{1}{N}\Phi\Phi^\top + \gamma^{-1} I_N\right)^{-1}}_{R_\mathrm{D}(\gamma,N)} \frac{1}{N} Y \tag{17}$$

Once the Lagrangian multipliers α are computed through (17), the estimate $\hat{\theta}_\mathrm{D}$ of the model parameters θ is obtained from (14c), i.e.,

$$\hat{\theta}_\mathrm{D} = \Phi^\top \underbrace{\left(\frac{1}{N}\Phi\Phi^\top + \gamma^{-1} I_N\right)^{-1} \frac{Y}{N}}_{\alpha}. \tag{18}$$

Remark 1 (Zero duality gap [27]) *Since the primal problem* (8) *is a convex quadratic problem with linear equality constraints, strong duality holds for* (8)*. Therefore, the dual and primal solutions are equivalent, i.e.,* $\hat{\theta}_\mathrm{D} = \hat{\theta}_\mathrm{P}$.

3.3 Consistency analysis

The importance of Remark 1 lays in allowing to analyze the asymptotic properties of the dual solution via the primal solution. The bias $B_{\hat{\theta}}$ of the estimate $\hat{\theta}_\mathrm{P}$ and hence the bias of $\hat{\theta}_\mathrm{D}$ can be studied through

$$B_{\hat{\theta}} = \bar{\mathbb{E}}\{\hat{\theta}_\mathrm{P} - \theta_\mathrm{o}\}, \tag{19}$$

where $\bar{\mathbb{E}}\{\centerdot\} = \lim_{N\to\infty} \mathbb{E}\{\centerdot\}$. Using the notation $R_\mathrm{P}(\gamma) = \bar{\mathbb{E}}\{R_\mathrm{P}(\gamma, N)\}$, the bias can be expressed as:

$$B_{\hat{\theta}} = \underbrace{\left(\gamma R_\mathrm{P}(\gamma)\right)^{-1}\theta_\mathrm{o}}_{B^\mathrm{r}_{\hat{\theta}}} + \underbrace{R_\mathrm{P}^{-1}(\gamma)\bar{\mathbb{E}}\left\{\frac{1}{N}\sum_{k=1}^{N}\varphi(k)v_\mathrm{o}(k)\right\}}_{B^\mathrm{n}_{\hat{\theta}}}. \tag{20}$$

For the derivation, see the Appendix. The term $B^\mathrm{r}_{\hat{\theta}}$ can be seen as the regularization bias which is determined by the regularization term in (7). On the other hand, $B^\mathrm{n}_{\hat{\theta}}$ is determined by the residual and hence can be seen as the direct effect of the noise. The interplay between the two terms corresponds to a trade-off between bias and variance of the estimates.

3.3.1 On the γ parameter bias/variance trade-off

Most often in over-parametrization context, the information matrix $(\Phi^\top\Phi)$ is rank deficient and it can be expressed as $(\Phi^\top\Phi) = U\Lambda U^\top$ with $UU^\top = I_{n_\theta}$, $\Lambda \in \mathbb{R}^{n_\theta \times n_\theta}$:

$$\Lambda = \mathrm{Diag}(\lambda_1 \ldots \lambda_r, 0, \ldots 0). \tag{21}$$

and hence, $\Phi^\top = U\sqrt{\Lambda}V^\top$, with $V \in R^{N\times n_\theta}$. Under this notation, the bias terms become:

$$B^\mathrm{r}_{\hat{\theta}} {=} \bar{\mathbb{E}}\left[U\mathrm{Diag}\left(\frac{\gamma^{-1}}{\gamma^{-1}{+}\frac{\lambda_1}{N}} \cdots \frac{\gamma^{-1}}{\gamma^{-1}{+}\frac{\lambda_r}{N}}, 1 \ldots 1\right) U^\top\right]\theta_\mathrm{o},$$
$$B^\mathrm{n}_{\hat{\theta}} {=} \bar{\mathbb{E}}\left[\frac{1}{N}U\mathrm{Diag}\left(\frac{\sqrt{\lambda_1}}{\gamma_1^{-1}{+}\frac{\lambda_1}{N}} \cdots \frac{\sqrt{\lambda_r}}{\gamma_1^{-1}{+}\frac{\lambda_r}{N}}, 0 \ldots 0\right) V^\top W_\mathrm{o}\right],$$

with $W_\mathrm{o} = [v_\mathrm{o}(1) \ldots v_\mathrm{o}(N)]^\top$. Moreover, let $\mathrm{Ker}\{\Phi^\top\Phi\}$ be the null space of $\Phi^\top\Phi$ and $\mathrm{Im}\{\Phi^\top\Phi\}$ its image. It is first interesting to notice that the θ components in $\mathrm{Ker}\{\Phi^\top\Phi\}$ are always set to zero independently from γ. Consider the behavior of θ in $\mathrm{Im}\{\Phi^\top\Phi\}$, for both the extreme cases $\gamma \to 0$ and $\gamma \to \infty$.

- For $\gamma \to 0$, $\hat{\theta} \to 0$ and it can straightforwardly concluded that $B^n_{\hat{\theta}} \to 0$, $B^r_{\hat{\theta}} \to -\theta_\mathrm{o}$ while the variance of $\hat{\theta}$ tends to exactly 0. Therefore, this case can be seen as a maximal regularization bias with no variance.
- For $\gamma \to \infty$, $B^r_{\hat{\theta}} \to 0$ and $B^n_{\hat{\theta}} \to B_\mathrm{LS}$ where B_LS is the bias of the unregularized LS estimate of $\hat{\theta}$. Moreover, using a similar reasoning, it is possible to show that the variance of $\hat{\theta}$ in this case is equal to the variance of the LS estimate in $\mathrm{Im}\{\Phi^\top\Phi\}$. This gives the other extremum case for which the noise has the maximal effect both in terms of bias and variance.

This indicates that the choice of γ defines an expected bias/variance trade-off for $\hat{\theta}$ except for disturbing term $B^\mathrm{n}_{\hat{\theta}}$ which is an unwanted artifact of the noise. The term $B^\mathrm{n}_{\hat{\theta}}$, characterizing (19), is a bias directly linked to the noise and can become important depending on the noise conditions. In the primal setting, it is well known that $B^\mathrm{n}_{\hat{\theta}} = 0$ under the condition that the regressor $\varphi(k)$ is not correlated with the noise $v_\mathrm{o}(k)$, i.e.,

C1 $\quad \mathbb{E}\{\varphi(k)v_\mathrm{o}(k)\} = 0, \quad \forall k \in \mathbb{Z}.$

This implies that C1 must also hold for the dual estimate in order to eliminate the bias due to the noise as $\gamma \to \infty$. Unfortunately, C1 only holds if v_o is white as $\varphi(k)$ is constructed from past samples of u and y. In fact, while $u(k-j)$ for any $j \in \mathbb{Z}$ is uncorrelated to the noise, $y(k-i)$ for $i > 0$ is uncorrelated to $v_\mathrm{o}(k)$ only in the case when this additive noise is white. However, in real-world applications, assuming that the noise in the data is white and exhibits an autoregressive form such as in (6) is highly unlikely to happen. Furthermore, such a noise model to be valid also requires accurate knowledge/estimation of the nonlinearities involved in the regressor. Consequently, in most practical applications, the minimization of criterion (7) will lead to a biased estimate, most likely to be dominated by $B^{\mathrm{n}}_{\hat{\theta}}$. Therefore, it would be highly advantageous to unbalance the bias/variance trade-off, in order to keep the variance of the LS estimate with a null bias when $\gamma \to \infty$. This is the aim of the next section, which introduces an IV method in order to cope with this issue, guaranteeing $B^{\mathrm{n}}_{\hat{\theta}} = 0$ for the dual estimate. Hence, the proposed IV approach significantly improves the applicability of the SVM scheme.

3.4 The support vector machine

For deriving the primal and dual solutions for (8), we started with the assumption that the functional basis $\{\phi_{i,l}\}_{i=1,l=1}^{n_{\mathrm{g}},n_{\mathrm{H}}}$ are *a priori* unknown and the dimension of the parametrization is possibly infinite. This means that, unlike to over-parametrization approaches where each $\{\phi_{i,l}\}_{i=1,l=1}^{n_{\mathrm{g}},n_{\mathrm{H}}}$ is a-priori fixed and n_{H} is finite and known, in the considered context $\varphi(k)$ is composed of unknown feature maps with $n_{\mathrm{H}} \to \infty$ implying that $n_\theta \to \infty$. Hence, $\hat{\theta}_{\mathrm{P}}$ cannot be explicitly computed via (12). The importance of the LS-SVM approach lies in the fact that the vector $\alpha \in \mathbb{R}^N$ in the dual solution can be analytically obtained without the proper knowledge of the feature maps Φ. In fact, what becomes possible via the so called *kernel trick* is to estimate the required feature maps in a considered function class using the dual solution.

To properly introduce the kernel trick based LS-SVM approach, consider the following preliminaries: Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a similarity measure, which is positive definite, like an *inner product*, w.r.t. (the finite measure space) $\mathcal{X} \subseteq \mathbb{R}^n$ with $0 < n < \infty$. Then K is called the kernel function of the operator T_K defined as

$$(T_K f)(x) = \int_{\mathcal{X}} K(x, x') f(x')\, dx', \tag{22}$$

if (22) exists for all $f \in \mathcal{L}_2(\mathcal{X})$ (square integrable maps $f : \mathcal{X} \to \mathbb{R}$).

Theorem 1 (Mercer's theorem, [28,29]) *Let $K \in \mathcal{L}_\infty(\mathcal{X}^2)$ be a symmetric real-valued function. Then, the integral operator (22) is positive definite in the sense that $\langle f, T_K f\rangle_{\mathcal{L}_2} > 0$ for all $f \in \mathcal{L}_2(\mathcal{X})$, if*

$$\int_{\mathcal{X}^2} K(x, x') f(x) f(x')\, dx\, dx' \geq 0, \tag{23}$$

for all $f \in \mathcal{L}_2(\mathcal{X})$.

Now, we would like to represent the positive definite inner products $\Phi\Phi^\top$ appearing in (17) *via* such symmetric kernel functions with $\mathcal{X}$ being the regression space. Suppose that K is a continuous symmetric kernel function on the closed interval $\mathcal{X} = [a, b]$. Then there is an orthonormal basis $\{\phi_l\}_{l=1}^\infty \in \mathcal{L}_2([a, b])$ such that K has the representation [29]

$$K(x_i, x_j) = \sum_{l=1}^{\infty} \lambda_l\, \phi_l(x_i)\, \phi_l(x_j), \tag{24}$$

where the convergence is absolute and uniform w.r.t. $\lambda_l \geq 0$. However, this also allows to represent any $f \in \mathcal{L}_2([a, b])$ in the reproducing kernel Hilbert space of K as a series expansion $f(\centerdot) = \sum_{j=1}^m \alpha_j K(\centerdot, x_j)$, with $\mathcal{X} = \{x_1, \ldots, x_m\}$ and $\alpha_j \in \mathbb{R}$ [29]. Furthermore, it allows to represent inner products of unknown functions as series expansions in terms of K.

Define the so-called *Grammian matrix* as $G = \Phi\Phi^\top$. According to the Mercer's theorem, the Grammian matrix G can be defined in terms of kernel functions without the explicit knowledge of Φ. Notice that G can be decomposed as

$$[G]_{j,k} = \sum_{i=1}^{n} [G^{(i)}]_{j,k}, \tag{25}$$

where each $G^{(i)}$ represents the inner product

$$[G^{(i)}]_{j,k} = \big\langle \phi_i(x_i(j)), \phi_i(x_i(k)) \big\rangle = K^{(i)}\big(x_i(j), x_i(k)\big). \tag{26}$$

Here, $K^{(i)}$ qualifies as a positive definite kernel function on the sample set $\mathcal{X} = \mathcal{D}_N$ and

$$x_i(k) = y(k - i), \qquad i = 1, \ldots, n_{\mathrm{a}}, \tag{27a}$$

$$x_{n_{\mathrm{a}}+1+j}(k) = u(k - j), \qquad j = 0, \ldots, n_{\mathrm{b}}. \tag{27b}$$

Consequently, a given set of kernel functions $K^{(i)}$ defines G and hence characterizes implicitly Φ, providing the solution for (17) in terms of

$$\alpha = \left(\frac{1}{N} G + \gamma^{-1} I_N\right)^{-1} \frac{Y}{N}. \tag{28}$$

This is called the *kernel trick* [1], [2], which allows the identification of the nonlinear functions f_i^{o}, g_j^{o} in (3a) without explicitly defining the feature maps involved. The resulting dual approach is called the LS-SVM. A typical choice of kernel, which provides uniformly effective representation of a large class of smooth functions, is the *Radial Basis Function* (RBF) kernel:

$$K^{(i)}\big(x_i(j), x_i(k)\big) = \exp\left(\tfrac{-\|x_i(j) - x_i(k)\|_2^2}{\sigma_i^2}\right). \tag{29}$$

However, other positive definite kernels, like *linear*, *polynomial*, *rational*, *spline* or *wavelet* kernels, can also be used [2]. Choosing the most appropriate kernel highly depends on the problem at hand. Automatic kernel selection for general SVM problems is possible and is discussed in [30].

Another remark is that the parameter vector $\hat{\theta}_{\mathrm{D}}$ is never accessible in the LS-SVM framework, and only the combined estimation $f_i(\centerdot) = \phi_i^\top(\centerdot)\theta_i$ (or $g_j(\centerdot) = \phi_{n_a+1+j}^\top(\centerdot)\theta_{n_a+1+j}$) is computable in an expansion form using the kernel functions defined:

$$f_i(\centerdot) = \phi_i^\top(\centerdot)\theta_i = \sum_{k=1}^{N} \alpha_k K^{(i)}(x_i(k), \centerdot). \qquad (30)$$

However, the resulting function estimator is directly linked to the dual solution of (6). This means that if the kernel functions defined feature map (and therefore the associated feature space $\mathcal{X}$) is correlated with the noise v_o (C1 is violated), then the function estimates are biased, i.e., $B_{\hat{\theta}}^{n} \neq 0$ implies that (30) is biased. In order to eliminate this bias, i.e., $B_{\hat{\theta}}^{n}$, an IV based modification of the LS-SVM is proposed in the next section.

4 Instruments in the primal and the dual form

Among the available identification approaches used in the regression framework, the principle idea of the *Instrumental Variable* (IV) approach has been successfully applied in many contexts to elegantly resolve the inconsistency problem of LS regression under correlated noise v_o [31,19,13,16]. In the sequel, our objective is to develop an IV extension of the LS-SVM, allowing a much wider applicability of this identification approach in practice.

We have seen previously that the most difficult condition required for the consistency of the LS-SVM is C1. In most practical problems, the regressor is correlated (implicitly or explicitly) to the noise and hence C1 does not hold. Thus, in the parametric context, an IV identification criterion has been introduced which relaxes C1 to a less restrictive condition and prevents the deterioration of the estimation performance [19]. The idea is to introduce a *so-called* instrument signal $\zeta : \mathbb{Z} \to \mathbb{R}^{n_\theta}$ such that the consistency condition w.r.t. the noise bias becomes:

X1 $\quad \mathbb{E}\{\zeta(k)v_o(k)\} = 0, \quad \forall k \in \mathbb{Z}.$

While condition C1 depends on $\varphi(k)$ and therefore on the model assumed, X1 depends on $\zeta(k)$ which can be chosen by the user. This idea grants a wide range of possible solutions for achieving consistency by picking instruments uncorrelated to the noise.

To respect the consistency conditions, the IV estimate can be seen as the minimizer of the IV criterion:

$$\begin{aligned}\mathcal{W}(\theta, e) &= \frac{1}{2}\theta^\top\theta + \frac{\gamma}{2N^2}\|\sum_{k=1}^{N}\zeta(k)e(k)\|_{\ell_2}^2 \\ &= \frac{1}{2}\|\theta\|_{\ell_2}^2 + \frac{\gamma}{2N^2}\|\Gamma^\top E\|_{\ell_2}^2, \end{aligned} \qquad (31)$$

based on the data set $\mathcal{D}_N$ and with Γ and E defined as

$$\Gamma = \left[\, \zeta(1) \ \ldots \ \zeta(N) \,\right]^\top, \qquad (32a)$$

$$E = \left[\, e(1) \ \ldots \ e(N) \,\right]^\top, \qquad (32b)$$

and $\zeta(k) \in \mathbb{R}^{n_\theta}$ being the instrument:

$$\zeta(k) = \left[\, 1 \ \phi_1^\top(\xi_1(k)) \ \ldots \ \phi_{n_g}^\top(\xi_{n_g}(k)) \,\right]^\top, \qquad (33)$$

chosen by the user so that condition X1 is satisfied. The specific choice of the instrument signals $\{\xi_i\}_{i=1}^{n_g}$ regarding the considered identification setting is discussed later. It is important to note that (31) introduces a different sum of norms criterion than (7). In this respect, the bias-variance trade off and even the introduced regularization bias via (31) does not necessary scales as in (7). On one hand, this makes comparison of the two estimation problems difficult in the considered nonlinear context, while on the other hand makes possible to achieve better reduction of the bias with a smaller sacrifice on the side of the variance.

The motivations to pursue an IV-scheme based solution for bias reduction are the following:

- In general, recent IV approaches offer similar performance as the optimal (minimum variance and unbiased estimate) prediction error methods in case of correct assumptions on the system and noise models.
- As it will be shown later, the IV-based LS-SVM problem can be solved in a very similar way to the LS-SVM problem, implying approximately the same computational load as well as the same complexity.
- Most importantly, the IV-schemes provide consistent estimates in case of incorrect noise assumptions.

While the IV methods are now widely used under the primal form of the optimization problem, they have never been introduced in a dual setting to the best of the authors' knowledge. Thus, the question arises: can the parallelism between the primal and dual solutions, explored in Section 3, be used to introduce an IV scheme for the dual form without any performance degradation?

4.1 IV in the primal form

First, the minimizer of (31) is derived based on the classical results. Let us define

$$\Psi = \sum_{k=1}^{N} \zeta(k)\varphi^\top(k) = \Gamma^\top\Phi, \qquad (34a)$$

$$D = \sum_{k=1}^{N} \zeta(k)y(k) = \Gamma^\top Y. \qquad (34b)$$

Then, the minimizer of $\mathcal{W}(\theta, e)$ (with the implicit assumptions that (8b) is satisfied and each ϕ_i is given) is

$$\hat{\theta}_{\mathrm{P}}^{\mathrm{IV}} = \arg \min_{\theta \in \mathbb{R}^{n_\theta}} \frac{1}{2}\theta^\top \theta + \frac{\gamma}{2N^2} \left\| \Psi\theta - D \right\|_{\ell_2}^2, \qquad (35)$$

which is equivalent with the following algebraic problem

$$\hat{\theta}_{\mathrm{P}}^{\mathrm{IV}} = \mathrm{sol}\left\{ \theta + \frac{\gamma}{N^2}\left(\Psi^\top \Psi \theta - \Psi^\top D \right) = 0 \right\}. \qquad (36)$$

This problem has an analytic solution given by

$$\hat{\theta}_{\mathrm{P}}^{\mathrm{IV}} = \underbrace{\left(\frac{1}{\gamma} I_{n_\theta} + \frac{1}{N^2}\Psi^\top \Psi \right)^{-1}}_{R_{\mathrm{P}}^{\mathrm{IV}}(\gamma, N)} \frac{1}{N^2}\Psi^\top D. \qquad (37)$$

Remark 2 (Bias of the IV estimate) *If the instrument satisfies condition X1, then, under minor assumptions, the bias $B_{\hat{\theta}}^{\mathrm{IV}}$ of the estimate $\hat{\theta}_{\mathrm{P}}^{\mathrm{IV}}$ is given by*

$$B_{\hat{\theta}}^{\mathrm{IV}} = \underbrace{\left(\gamma R_{\mathrm{P}}^{\mathrm{IV}}(\gamma) \right)^{-1} \theta_{\mathrm{o}}}_{B_{\hat{\theta}}^{\mathrm{IV,r}}}, \qquad (38)$$

where $R_{\mathrm{P}}^{\mathrm{IV}}(\gamma) = \bar{\mathbb{E}}\{R_{\mathrm{P}}^{\mathrm{IV}}(\gamma, N)\}$.

For a proof see the Appendix.

Consequently, in the IV-SVM scheme, the bias is solely conditioned by the regularization term. Again, by defining $\Psi^\top \Psi = U'^\top \Lambda' U'^\top$ where

$$\Lambda' = \mathrm{Diag}(\lambda'_1 \dots \lambda'_{r'}, 0, \dots 0) \qquad (39)$$

the argument provided in Section 3.3.1, implies that $B_{\hat{\theta}}^{\mathrm{IV}} \to 0$ as $\gamma \to \infty$ for the θ components in $\mathrm{Im}\{\Psi^\top \Psi\}$. This is however not true for the components in $\mathrm{Ker}\{\Psi^\top \Psi\}$. Consequently, in order to ensure a bias improvement with respect to the LS approach, the chosen instrument must fulfill the following condition:

X2 $\quad \mathrm{Im}\{\Phi^\top \Phi\} = \mathrm{Im}\{\Psi^\top \Psi\}$.

In other words $\Phi^\top \Phi = U \Lambda U^\top$ with Λ as in (21) and $\Psi^\top \Psi = U \tilde{\Lambda} U^\top$ with

$$\tilde{\Lambda} = \mathrm{Diag}(\tilde{\lambda}_1 \dots \tilde{\lambda}_r, 0, \dots 0). \qquad (40)$$

It can be noticed that this condition is similar to full ranking condition in the linear regression framework [19]. In practice, verifying this condition might be a tedious task. Nevertheless, choosing ξ_i correlated to x_i, $\forall i = 1 \dots n_\theta$ can ensure the correlation between Φ and Γ and consequently, condition X2.

Under this conditions, it can be concluded that the proposed estimate fulfills the aimed features: the bias/variance trade-off has been dramatically changed with respect to the LS optimization scheme. Hence, for data sets where the bias is large with respect to the variance, the proposed estimate can seriously increase the quality of the estimates. Again, the γ parameter has to be optimized based on validation data for the exact same reasons as exposed in section 3.3.1.

4.2 IV in the dual form

Next, the IV solution is derived in a dual form which will be used to define the IV-SVM. Let us rewrite the primal minimization problem of (31) as

$$\min_{\theta \in \mathbb{R}^{n_\theta}} \quad \frac{1}{2}\theta^\top \theta + \frac{\gamma}{2N^2} \left\| \Gamma^\top E \right\|_{\ell_2}^2, \qquad (41a)$$

$$\text{s.t.} \quad e(k) = y(k) - \varphi^\top(k)\theta, \qquad (41b)$$

based on the data set $\mathcal{D}_N$. Introduce the *Lagrangian*

$$\mathcal{L}(\theta, e, \alpha) = \mathcal{W}(\theta, e) - \sum_{k=1}^{N} \alpha_k \left(\varphi^\top(k)\theta + e(k) - y(k) \right), \quad (42)$$

with $\alpha_k \in \mathbb{R}$. The global optimum is obtained when the KKT conditions (necessary and sufficient) are fulfilled:

$$\frac{\partial \mathcal{L}}{\partial e} = 0 \quad \to \quad \alpha_k = \frac{\gamma}{N^2}\Gamma\Gamma^\top e(k), \qquad (43a)$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \to \quad y(k) = \varphi^\top(k)\theta + e(k), \qquad (43b)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \quad \to \quad \theta = \sum_{k=1}^{N} \alpha_k \varphi(k). \qquad (43c)$$

By substituting (43c) into (43b), we get

$$y(k) = \left(\sum_{\tau=1}^{N} \alpha_\tau \varphi^\top(\tau) \right) \varphi(k) + e(k), \qquad (44)$$

which can be written in the matrix form:

$$Y = \Phi\Phi^\top \alpha + E. \qquad (45)$$

Then, substitution of (45) into (43a) leads to

$$\alpha = \frac{\gamma}{N^2}\Gamma\Gamma^\top \left(Y - \Phi\Phi^\top \alpha \right), \qquad (46)$$

which has the solution

$$\alpha = \left(\frac{1}{N^2} HG + \gamma^{-1} I_N \right)^{-1} \frac{1}{N^2} HY, \qquad (47)$$

where $H = \Gamma\Gamma^\top$. Hence, H is defined similar to G in terms of (26) via the chosen kernels evaluated on ζ. According to (43c), $\theta = \Phi^\top \alpha$ and therefore

$$\hat{\theta}_{\mathrm{D}}^{\mathrm{IV}} = \Phi^\top \underbrace{\left[\frac{1}{N^2} HG + \gamma^{-1} I_N\right]^{-1}}_{R_{\mathrm{D}}^{\mathrm{IV}}(\gamma,N)} \frac{1}{N^2} HY. \tag{48}$$

As for the LS-SVM case, once α are computed from (47), the dual estimate $\hat{\theta}_{\mathrm{D}}^{\mathrm{IV}}$ of the parameters θ is then given by (43c) and, again, the primal and dual solutions are equivalent. Consequently, under condition X1, $\hat{\theta}_{\mathrm{D}}^{\mathrm{IV}}$ is consistent w.r.t. the residual bias independently on the correlation properties of $v_{\mathrm{o}}(k)$ (as long as it is zero-mean). Similarly, to the standard LS-SVM, the estimate of the nonlinear functions f_i^{o} and g_j^{o} is given by

$$f_i(\boldsymbol{.}) = \sum_{k=1}^{N} \alpha_k K^{(i)}(x_i(k), \boldsymbol{.}), \tag{49a}$$

$$g_j(\boldsymbol{.}) = \sum_{k=1}^{N} \alpha_k K^{(n_{\mathrm{a}}+1+j)}(x_{n_{\mathrm{a}}+1+j}(k), \boldsymbol{.}). \tag{49b}$$

4.3 Identification of more general systems

If we consider a more general class of systems described in the form of (2), the IV-SVM and LS-SVM methods still remain applicable. Although, one major difference is that in this case, due to the lack of strong priors on the separation of dynamics and nonlinearities, a single but multidimensional kernel function is needed , as it must be based on the whole set of signals $x = [\, x_1 \, \dots \, x_{n_{\mathrm{g}}} \,]^\top$ which contains the dynamics already (time operators).

5 Implementation of the IV-SVM

In this section, the properties and practical implementation of the IV-SVM method are discussed together with the selection of the instruments, the kernel functions and the hyper-parameters.

5.1 The choice of the instrument

So far, we have not shed light on the specific choice of the variables $\xi(k)$ which eventually generate the instrument $\zeta(k)$ used in the derivation of the IV-SVM. In the previous section, we have shown that, in order to guarantee consistency of the dual parameters $\hat{\theta}_{\mathrm{D}}^{\mathrm{IV}}$ w.r.t. the noise bias, the instrument $\zeta(k)$ has to satisfy condition X1, which means that the variables $\xi(k)$ have to be chosen by the user so that they are independent of the noise realization $v_{\mathrm{o}}(k)$. It is worth pointing out that, in the linear identification framework, the optimal instruments minimizing the asymptotic covariance matrix of the estimated parameters are given by the noise-free input and output samples [19]. On the other hand, in a nonlinear context, the choice of an optimal instrument depends highly on the system structure and the noise model assumed, and is mostly an open problem. Nevertheless, the bias results exposed in Sections 4.1 and 3.3.1 can be derived in the same way for the variance leading to the conclusion that the variance properties is strongly linked to the linear regression theory for $\mathrm{Im}\{\Psi^\top\Psi\}$ as $\gamma \to \infty$. Hence, the linear regression theory is here invoked regarding the choice of an instrument. More precisely, the variable $\xi_i(k)$ is chosen to be maximally correlated with the noise-free part of the sample $x_i(k)$ in order to both satisfy X2 and X1. This leads to the following choice of instruments

$$\xi_i(k) = \breve{y}(k-i), \quad i = 1, \dots, n_{\mathrm{a}}, \tag{50a}$$

$$\xi_{n_{\mathrm{a}}+1+j}(k) = u(k-j), \quad j = 0, \dots, n_{\mathrm{b}}, \tag{50b}$$

where $\breve{y}$ is the noise free output signal of the data-generating system $\mathcal{S}_{\mathrm{o}}$. As such signals are not available in practice, one needs to restrict himself to $\breve{y}$ being approximated by the simulated output of an estimated model of the system, e.g., a model obtained via the LS-SVM approach. This choice of the instrument resembles to the widely used IV solution for linear regression [19,13].

The following iterative IV-SVM scheme can be implemented in order to mitigate the effect of the estimated noiseless signals on the IV scheme and hence “maximize” the accuracy of the IV-SVM solution by iteratively refining the instruments.

Algorithm 1 Refined IV-SVM

```
Require: model structure (4) in terms of model orders
    $n_{\mathrm{a}}$ and $n_{\mathrm{b}}$, data set $\mathcal{D}_N = \{y(k), u(k)\}_{k=1}^N$, regularization parameter $\gamma$, kernel functions $\{K^{(i)}\}_{i=1}^{n_{\mathrm{g}}}$ with
    $n_{\mathrm{g}} := n_{\mathrm{a}} + n_{\mathrm{b}} + 1$.
1: set $\tau \leftarrow 0$.
2: compute the matrices $G^{(i)}$ via the kernels $K^{(i)}$ applied on $\mathcal{D}_N$.
3: estimate $\alpha^{(0)}$ via (17) resulting in the model estimate $\mathcal{M}^{(0)}$.
4: repeat
5:    set $\tau \leftarrow \tau + 1$
6:    use $\mathcal{M}^{(\tau-1)}$ to generate, by simulation, $\{\breve{y}^{(\tau)}(k)\}_{k=1}^N$.
7:    calculate $\{\xi_i(k)\}_{i=1,k=1}^{n_{\mathrm{g}},N}$ via (50) using $\{\breve{y}^{(\tau)}(k), u(k)\}_{k=1}^N$.
8:    compute $H$ in terms of the kernel functions $K^{(i)}$ applied on $\{\xi_i(k)\}_{i=1,k=1}^{n_{\mathrm{g}},N}$.
9:    estimate $\alpha^{(\tau)}$ via (47) resulting in the model estimate $\mathcal{M}^{\tau}$.
10: until $\alpha^{(\tau)}$ has converged.
11: return Model structure $\mathcal{M}^{(\tau)}$ with estimates of the nonlinear functions $\hat{f}_i$ and $\hat{g}_j$ obtained via (49).
```

5.2 Enforcing zero-centering of the nonlinear functions

Note that, to ensure identifiability, it is essential to impose normalization of f_i and g_j as otherwise, any pair of functions $\tilde{\phi}_1^\top(\cdot)\tilde{\theta}_1 = \phi_1^\top(\cdot)\theta_1 - \bar{c}$ and $\tilde{\phi}_2^\top(\cdot)\tilde{\theta}_2 = \phi_2^\top(\cdot)\theta_2 + \bar{c}$ would provide the same input-output realization for any constant $\bar{c}$. Consequently, in order to respect the system description (1) and the assumption that f_i^{o} and g_j^{o} are equal to zero at the origin, it is essential to impose that

$$\phi_i^\top(0)\theta_i = 0, \qquad i = 1, \ldots, n_{\mathrm{g}}. \tag{51}$$

The primal problem (41) can be then modified by adding the constraints (51), i.e.,

$$\min_{\theta \in \mathbb{R}^{n_\theta}} \quad \frac{1}{2}\theta^\top\theta + \frac{\gamma}{2N^2}\left\|\Gamma^\top E\right\|_{\ell_2}^2, \tag{52a}$$

$$\text{s.t.} \quad e(k) = y(k) - \varphi^\top(k)\theta, \tag{52b}$$

$$\phi_i^\top(0)\theta_i = 0. \tag{52c}$$

By introducing the *Lagrangian*

$$\mathcal{L}(\theta, e, \alpha, \beta) = \mathcal{W}(\theta, e) - \sum_{k=1}^{N} \alpha_k \big(\varphi^\top(k)\theta + e(k) - y(k)\big) - \sum_{i=1}^{n_{\mathrm{g}}} \beta_i \phi_i^\top(0)\theta_i, \tag{53}$$

the following expression of the dual parameters $\alpha \in \mathbb{R}^N$ and $\beta \in \mathbb{R}^{n_{\mathrm{g}}}$ can be derived from the KKT conditions:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{N^2}HG + \frac{1}{\gamma}I_N & \frac{1}{N^2}HD_\Phi \\ \frac{1}{N}D_\Phi^\top & \frac{1}{N}D_0 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{N^2}HY \\ 0_{n_{\mathrm{g}}} \end{bmatrix}, \tag{54}$$

with $1_N = [1 \ \ldots \ 1]^\top \in \mathbb{R}^N$, $0_{n_{\mathrm{g}}} = [0 \ \ldots \ 0]^\top \in \mathbb{R}^{n_{\mathrm{g}}}$ and

$$\Phi_i = \big[\, \phi_i(x_i(1)) \ \ldots \ \phi_i(x_i(N)) \,\big]^\top, \tag{55a}$$

$$D_\Phi = \big[\, \Phi_1\phi_1(0) \ \ldots \ \Phi_{n_{\mathrm{g}}}\phi_{n_{\mathrm{g}}}(0) \,\big]^\top, \tag{55b}$$

$$D_0 = \mathrm{diag}\Big(\phi_1^\top(0)\phi_1(0), \ \ldots, \ \phi_{n_{\mathrm{g}}}^\top(0)\phi_{n_{\mathrm{g}}}(0)\Big). \tag{55c}$$

The kernel trick can be then used to characterize the entries of H, G, D_Φ and D_0 in terms of the kernel functions $K^{(i)}$. This leads to the following estimate of the constant term c^{o} and of the nonlinear functions f_i^{o}:

$$c = 1_N^\top \alpha, \tag{56a}$$

$$f_i(\centerdot) = \sum_{k=1}^{N} \alpha_k K^{(i)}(x_i(k), \centerdot) + \beta_i K^{(i)}(0, \centerdot), \tag{56b}$$

$$g_j(\centerdot) = \sum_{k=1}^{N} \alpha_k K^{(n_{\mathrm{a}}+1+j)}(x_{n_{\mathrm{a}}+1+j}(k), \centerdot) + \beta_{n_{\mathrm{a}}+1+j} K^{(n_{\mathrm{a}}+1+j)}(0, \centerdot). \tag{56c}$$

A detailed derivation of eqs. (56) can be found in the technical report [32].

5.3 The choice of γ and the kernels

In this subsection, choice/optimization of the hyper-parameter γ (defining the cost function $\mathcal{W}(\theta, e)$ in (31)) and the choice/tuning of the kernel functions $K^{(i)}$ is discussed.

As it has been briefly explained in Section 3, choice of the most appropriate kernel for the modeling problem at hand highly depends on the structure of the system to be identified and on the available data. Besides of the discussed *radial basis function* kernels which are adequate to represent a large class of smooth functions in terms of the expansion $f(\centerdot) = \sum_{j=1}^{m} \alpha_j K(\centerdot, x_j)$ (see Section 3.4, other positive definite kernels, like *linear*, *polynomial*, *rational*, *spline* or *wavelet* kernels, can also be used [2]. However, these choices have an impact on the function class in which the expansion is made rather than the actual decay rate of the expansion error. So it becomes a question, how the particular parameters of these kernel functions, like σ_i in (29) should be chosen to maximize the decay rate of the expansion w.r.t. the estimated unknown functional terms and hence the accuracy of the obtained model. Furthermore, the optimal choice of the regularization parameter γ is dependent on the choice of kernel functions, hence the overall optimization of all such hyper-parameters can not be independently accomplished.

If we restrict our attention to the RBF case, a simple methodology can be used to optimize the kernel functions $K^{(i)}$ and γ for the system to be estimated. As a first step, we can reduce the number of hyper-parameters by requiring that

$$\sigma_i = \sigma_{\mathrm{y}} \quad \text{for all } i = 1, \ldots, n_{\mathrm{a}} \tag{57a}$$

$$\sigma_{n_{\mathrm{a}}+1+j} = \sigma_{\mathrm{u}} \quad \text{for all } j = 0, \ldots, n_{\mathrm{b}} \tag{57b}$$

In this way, all kernels $K^{(i)}$ (with $i = 1, \ldots, n_{\mathrm{g}}$) used in the IV-SVM scheme are characterized by only two hyper-parameters, i.e., σ_{y} and σ_{u}. Then, the parameters σ_{y}, σ_{u} and γ are tuned via cross-validation based optimization. For instance, the values of σ_{y}, σ_{u} and γ providing the most accurate model w.r.t. an independent "validation" data set can be computed through a three-dimensional grid-search procedure over the space of hyper-parameters. Other numerically efficient techniques for the computation of the optimal hyper-parameters by means of genetic algorithms and particle swarm optimization are discussed in [33–35].

6 Simulation example

In this section, performance of the IV-SVM and of the standard LS-SVM approaches are compared using an extensive Monte-Carlo study based on a simulation example.

6.1 The data-generating system

The data-generating system $\mathcal{S}_\mathrm{o}$, considered in this study, is described by the difference equation

$$\begin{aligned} y(k) = -0.7y(k-1) + f^\mathrm{o}\big(y(k-2)\big) + g^\mathrm{o}\big(u(k)\big) + \\ - 0.4u(k-1) + c^\mathrm{o} + v_\mathrm{o}(k), \quad (58) \end{aligned}$$

where the constant parameter c^o is 0, while $f(\cdot)$ and $g(\cdot)$ are defined by the following nonlinear functions

$$f^\mathrm{o}(x) = \tfrac{1}{8}x^2, \; g^\mathrm{o}(x) = \begin{cases} 0.5 & \text{if} \quad x \geq 0.5, \\ x & \text{if} \quad -0.5 < x < 0.5, \\ -0.5 & \text{if} \quad x \leq -0.5. \end{cases}$$

Furthermore, $v_\mathrm{o}(k)$ is a zero-mean colored noise generated by filtering a white noise sequence $e_\mathrm{o}(k) \sim \mathcal{N}(0, \sigma_e^2)$ with standard deviation $\sigma_e = 0.07$ with the filter

$$v_\mathrm{o}(k) = a_1 v_\mathrm{o}(k) + b_0 e_\mathrm{o}(k) + b_1 e_\mathrm{o}(k-1), \qquad (59)$$

where $a_1 = 0.95$, $b_0 = 1.5$ and $b_1 = -0.3$. To generate data sets $\mathcal{D}_N = \{u(k), y(k)\}_{k=1}^N$ by $\mathcal{S}_\mathrm{o}$, the system is excited with a $N = 1000$ long white input sequence with uniform distribution $\mathcal{U}(-1, 1)$ starting with zero initial conditions. In order to provide representative results, 100 of such data sets are generated with independent realizations of the noise and the input sequences, setting up a Monte Carlo study with $N_\mathrm{MC} = 100$ runs. To asses the level of the noise, the *signal-to-noise ratio* (SNR), defined as

$$\mathrm{SNR} = 10 \log \left(\frac{\sum_{k=1}^N y_\mathrm{o}^2(k)}{\sum_{k=1}^N v_\mathrm{o}^2(k)} \right), \qquad (60)$$

with $y_\mathrm{o}(k)$ denoting the noise-free output of $\mathcal{S}_\mathrm{o}$, has been computed. The resulting 11 dB average SNR corresponds to a significant noise present in the data.

6.2 Model structures

In order to provide a fair comparison between the LS-SVM method and the IV-SVM method, the same model structure is used by the two approaches. In particular, the following ARX model structure $\mathcal{M}_\theta$ is considered:

$$\begin{aligned} y(k) = y(k-1)\theta_1 + \phi_2^\top\big(y(k-2)\big)\theta_2 + \\ + \phi_3^\top\big(u(k)\big)\theta_3 + u(k-1)\theta_4 + c + e(k), \quad (61) \end{aligned}$$

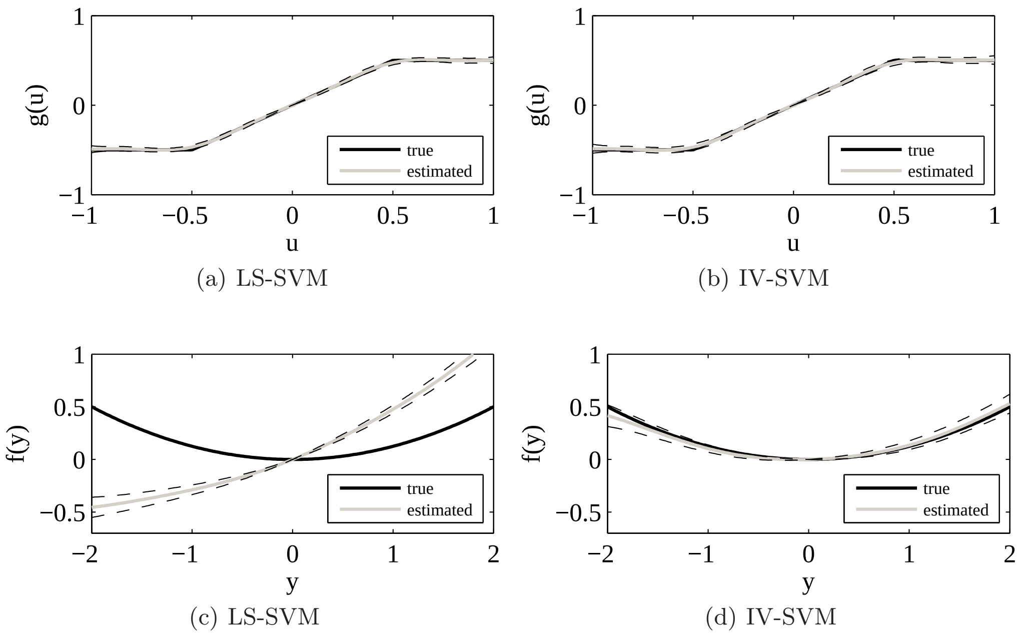
with $e(k)$ denoting the residual error. Note that the functions $\phi_1(y(k-1))$ and $\phi_4(u(k-1))$ are explicitly defined as $y(k-1)$ and $u(k-1)$, respectively. This a-priori information can be easily exploited in the SVM identification context. On the other hand, the feature maps $\phi_2(\cdot), \phi_3(\cdot) : \mathbb{R} \to \mathbb{R}^{n_\mathrm{H}}$ are not a-priori imposed and their dimension n_H is potentially infinite. RBF kernels are used to characterize these nonlinearities. Since the aim of the example is to compare the estimation properties of the LS-SVM and the IV-SVM approaches, the same modeling setting is applied in both approaches by using the same kernel functions and hyper-parameters. The IV-SVM approach is implemented with the refined scheme of Algorithm 1 defining the choice of instruments. The parameters σ_2 and σ_3 associated with the RBF kernel function $K^{(2)}$ and $K^{(3)}$ used to estimate ϕ_2 and ϕ_3 are chosen via cross validation, in particular, by maximizing the *best fit rate* (BFR), computed on a validation data set, of the model estimates. The BFR is defined as

$$\mathrm{BFR} = \max \left\{ 1 - \frac{\|y(k) - \hat{y}(k)\|_2}{\|y(k) - \overline{y}(k)\|_2}, 0 \right\} \cdot 100\%, \qquad (62)$$

with $\overline{y}(k)$ denoting the mean value of the output sequence $y(k)$, while $\hat{y}(k)$ is the simulated output of the estimated model. Based on an exhaustive grid search, $\sigma_2 = 1.7$ and $\sigma_3 = 0.5$ have been obtained for the LS-SVM. This choice of hyper-parameters have been also applied in the IV-SVM in order to guarantee fair comparison w.r.t. the original LS-SVM approach. On the other hand, the regularization-parameters γ_LS and γ_IV have been optimized separately, since they are linked to different criteria, i.e., $\mathcal{V}(\theta, e)$ (in eq. (7)) and $\mathcal{W}(\theta, e)$ (in eq. (31)), respectively. An exhaustive search aiming at the maximization of the cross-validation based BFR has lead to the choice of $\gamma_\mathrm{LS} : \frac{\gamma_\mathrm{LS}}{N} = 12.5$ and $\gamma_\mathrm{IV} : \frac{\gamma_\mathrm{IV}}{N^2} = 16.5$.

6.3 Obtained results

First, note that since the $\phi_1(\cdot)$ and $\phi_4(\cdot)$ are explicitly defined, the parameters θ_1 and θ_4 are directly identified. On the other hand, the parameters θ_2 and θ_3 are not directly accessible and only a combined estimation of the functions $f(\cdot) = \phi_2^\top(\cdot)\theta_2$ or $g(\cdot) = \phi_3^\top(\cdot)\theta_3$ can be computed. The mean value and standard deviation (over the 100 Monte Carlo runs) of the estimates $\hat{\theta}_1$ and $\hat{\theta}_4$ obtained through the LS-SVM and the IV-SVM algorithms are reported in Table 1, which shows that, in line with the theory, the standard LS-SVM approach provides a biased estimate of θ_1 and θ_4, while the parameters identified through the IV-SVM are unbiased. The estimation results of the nonlinear function $g(\cdot)$ obtained by the LS-SVM and the IV-SVM methods are shown in Fig. 1a-b, where the mean estimated function together with the standard deviation interval over the 100 Monte Carlo runs are plotted. Both the algorithms provide an accurate estimate of the function $g(\cdot)$. The estimation results of the nonlinearity $f(\cdot)$ are reported in Fig. 1c-d, which shows that the mean estimate of $f(\cdot)$ obtained through the IV-SVM algorithm is centered on the true one, while the LS-SVM approach provides a biased estimate of the nonlinearity $f(\cdot)$.



(a) LS-SVM

(b) IV-SVM

(c) LS-SVM

(d) IV-SVM

Fig. 1. Comparison of the estimation results: true nonlinearity $g^{\rm o}(u)$ and $f^{\rm o}(y)$ (solid black lines), mean estimate (solid gray line) and standard deviation intervals (dashed black line) over the 100 Monte Carlo runs.

Table 1
Mean and standard deviation of the estimated parameters $\hat{c}$, $\hat{\theta}_1$ and $\hat{\theta}_4$ over the 100 Monte Carlo runs.

	True Value	LS-SVM	IV-SVM
mean $\hat{c}$	0	-0.0044	-0.0035
std $\hat{c}$	–	0.0290	0.0468
mean $\hat{\theta}_1$	-0.7	-0.2684	-0.6920
std $\hat{\theta}_1$	–	0.0451	0.0359
mean $\hat{\theta}_4$	-0.4	-0.6983	-0.4062
std $\hat{\theta}_4$	–	0.0320	0.0371

The performance of the LS-SVM and the IV-SVM algorithms is also tested on a noiseless validation data set $\mathcal{D}_N^{\rm V}$ with $N = 600$ input-output pairs, where the input is a triangle wave with range from -0.5 to 0.5 and with a period of 50 samples. Using this validation data, the simulated output sequences $\hat{y}$ of the models estimated by the LS-SVM and the IV-SVM algorithms are plotted in Fig. 2a-b, while the error between the true output $y_{\rm o}(k)$ in $\mathcal{D}_N^{\rm V}$ and $\hat{y}(k)$ is plotted in Fig. 2c-d. The average BFR and the *mean squared error* (MSE), defined as

$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}\left(y_{\rm o}(k) - \hat{y}(k)\right)^2, \tag{63}$$

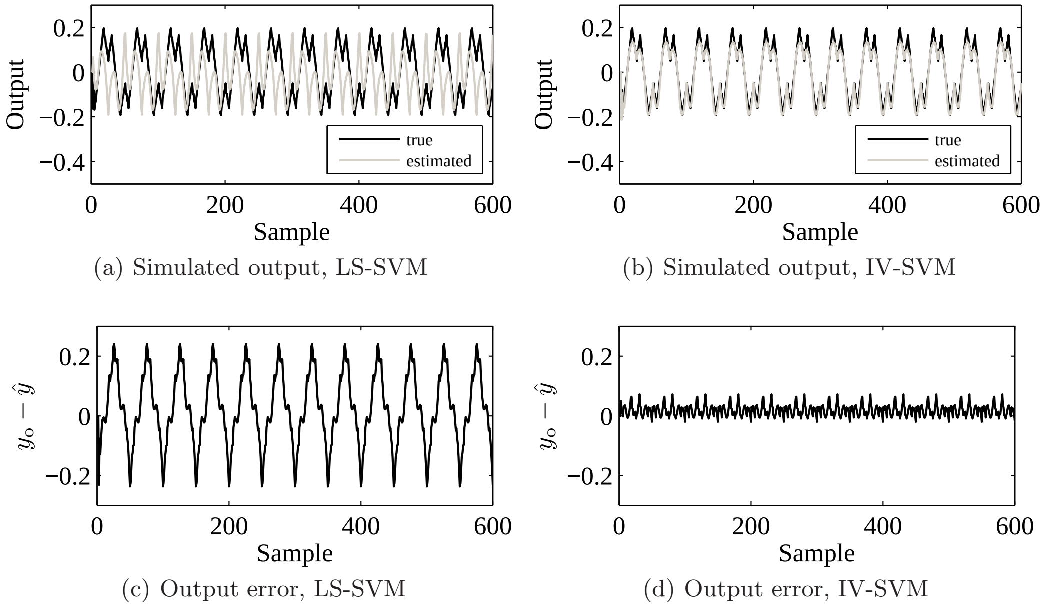
computed on the simulated response $\hat{y}$ of the estimated models are reported in Table 2. The obtained results show that IV-SVM algorithm significantly outperforms the standard LS-SVM approach.

Table 2
Validation results of the estimated models. Average and standard deviation of the best fit rate (BFR) and of the mean squared error (MSE).

	mean(BFR)	std(BFR)	mean(MSE)	std(MSE)
LS-SVM	15%	14%	0.0127	0.0030
IV-SVM	95%	2%	0.0009	0.0003

7 Conclusions

In this paper, identification of nonlinear models *via* the LS-SVM identification scheme has been analyzed, resulting in a consistency analysis at the parameter level for this method issued from the machine learning community. The analysis has shown that due to the minimization of the ℓ_2-loss over the prediction using an NARX structure, the choice of the regularization parameter under general noise conditions corresponds to a trade-off between a regularization bias deriving from the optimization scheme and a noise bias deriving from the measured data. Consequently, in case of a system structure other than NARX, the user is unable to suppress the bias on the estimates, which can lead to poor structural learning capabilities. Consequently, an IV-SVM optimization criterion was introduced in order to cope with this limitation while preserving the attractive computational properties of the LS-SVM. It has been shown that the IV-SVM scheme allows the elimination of the bias in case the noise process can be written as a zero mean additive process. Hence, the proposed scheme considerably widens the applicability of LS-SVM based methods. A suitable choice for the required instrument has been



(a) Simulated output, LS-SVM

(b) Simulated output, IV-SVM

(c) Output error, LS-SVM

(d) Output error, IV-SVM

Fig. 2. Validation data based comparison of the true output $y_{\mathrm{o}}(k)$ (black line) with the simulated output sequence $\hat{y}(k)$ (gray line) of one model estimated from the Monte Carlo simulation by the standard LS-SVM and the proposed IV-SVM approaches.

discussed and an iterative solution inspired from linear regression IV methods has been proposed. The performance of the resulting IV-SVM algorithm with respect to the regular LS-SVM method has been demonstrated on a challenging example. Finally, while generalization in terms of the nonlinear dynamics follows relatively easily, generalization for non-zero mean and/or nonlinearly distorted noise is technically more demanding and remains the objective of future research in the LS-SVM context.

8 Appendix

PROOF. (Bias of the LS estimate) Consider $B_{\hat{\theta}} = \bar{\mathbb{E}}\{\hat{\theta}_{\mathrm{P}} - \theta_{\mathrm{o}}\}$. Then, based on (10),

$$B_{\hat{\theta}} = \bar{\mathbb{E}}\left\{ R_{\mathrm{P}}^{-1}(\gamma,N)\left(\frac{1}{N}\sum_{k=1}^{N}\varphi(k)y(k)\right) - \theta_{\mathrm{o}}\right\} \tag{64}$$

By using that $y(k) = \varphi^{\top}(k)\theta_{\mathrm{o}} + v_{\mathrm{o}}(k)$, (64) can be written as

$$\begin{aligned} B_{\hat{\theta}} &= \bar{\mathbb{E}}\left\{ R_{\mathrm{P}}^{-1}(\gamma,N)\frac{1}{N}\left(\Phi^{\top}\Phi\theta_{\mathrm{o}} + \sum_{k=1}^{N}\varphi(k)v_{\mathrm{o}}(k)\right) - \theta_{\mathrm{o}}\right\} \\ &= \bar{\mathbb{E}}\Big\{ \underbrace{R_{\mathrm{P}}^{-1}(\gamma,N)\frac{1}{N}\left(\Phi^{\top}\Phi - NR_{\mathrm{P}}(\gamma,N)\right)\theta_{\mathrm{o}}}_{(\gamma R_{\mathrm{P}}(\gamma,N))^{-1}\theta_{\mathrm{o}}}\Big\} \\ &\quad + \bar{\mathbb{E}}\left\{ R_{\mathrm{P}}^{-1}(\gamma,N)\frac{1}{N}\sum_{k=1}^{N}\varphi(k)v_{\mathrm{o}}(k)\right\} \end{aligned}$$

Then, by using $R_{\mathrm{P}}(\gamma) = \bar{\mathbb{E}}\{R_{\mathrm{P}}(\gamma,N)\}$, (20) follows directly.

PROOF. (Bias of the IV estimate) Consider $B_{\hat{\theta}}^{\mathrm{IV}} = \bar{\mathbb{E}}\{\hat{\theta}_{\mathrm{P}}^{\mathrm{IV}} - \theta_{\mathrm{o}}\}$. Then

$$B_{\hat{\theta}}^{\mathrm{IV}} = \bar{\mathbb{E}}\left\{ \left(R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\right)^{-1}\frac{1}{N^2}\Psi^{\top}\sum_{k=1}^{N}\zeta(k)y(k) - \theta_{\mathrm{o}}\right\}. \tag{65}$$

Using the same derivation as in the LS case, it follows that

$$\begin{aligned} B_{\hat{\theta}}^{\mathrm{IV}} &= \bar{\mathbb{E}}\Big\{ \underbrace{\left(R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\right)^{-1}\frac{1}{N^2}\left(\Psi^{\top}\Psi - N^2R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\right)\theta_{\mathrm{o}}}_{\left(\gamma R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\right)^{-1}\theta_{\mathrm{o}}}\Big\} \\ &\quad + \bar{\mathbb{E}}\Big\{ \left(R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\right)^{-1}\frac{1}{N^2}\Psi^{\top}\sum_{k=1}^{N}\zeta(k)v_{\mathrm{o}}(k)\Big\} \end{aligned}$$

Now under minor assumptions (stationary data and that ζ is uncorrelated with v_{o}) it holds that $R_{\mathrm{P}}^{\mathrm{IV}}(\gamma) = \bar{\mathbb{E}}\{R_{\mathrm{P}}^{\mathrm{IV}}(\gamma,N)\}$ and $Q_*(\gamma) = \bar{\mathbb{E}}\{\left(R_{\mathrm{P}}^{\mathrm{IV}}(\gamma)\right)^{-1}\frac{1}{N}\Psi^{\top}\}$ exist and

$$B_{\hat{\theta}}^{\mathrm{IV}} = \left(\gamma R_{\mathrm{P}}^{\mathrm{IV}}(\gamma)\right)^{-1}\theta_{\mathrm{o}} + Q_*(\gamma)\bar{\mathbb{E}}\left\{\frac{1}{N}\sum_{k=1}^{N}\zeta(k)v_{\mathrm{o}}(k)\right\}. \tag{66}$$

If condition X1 is satisfied, then it holds that

$$\bar{\mathbb{E}}\left\{\frac{1}{N}\sum_{k=1}^{N}\zeta(k)v_{\mathrm{o}}(k)\right\} = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\underbrace{\mathbb{E}\left\{\zeta(k)v_{\mathrm{o}}(k)\right\}}_{=0} = 0.$$

References

[1] V. Vapnik, *Statistical Learning Theory.* Wiley-Interscience, 1998.

[2] B. Schölkopf and A. Smola, *Learning with kernels.* Cambridge MA: MIT Press, 2002.

[3] N. Cristianini and J. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge, UK: Cambridge University Press, 2000.

[4] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, pp. 1–12, 2010.

[5] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinearsystem identification," *IEEE Trans. on Automatic Control*, vol. 56, no. 12, pp. 2825–2840, 2011.

[6] T. Falck, K. Pelckmans, J. Suykens, and B. De Moor, "Identification of wiener-hammerstein systems using LS-SVMs," in *15th IFAC symposium on System Identification*, Saint Malo, France, July 2009.

[7] I. Goethals, K. Pelckmans, J. Suykens, and B. De Moor, "Identification of MIMO Hammerstein models using least squares support vector machines," *Automatica*, vol. 41, no. 7, pp. 1263–1272, 2005.

[8] F. Giri and E.-W. Bai, *Lecture Notes in Control and Information Sciences*, ser. Block-oriented Nonlinear System Identification. Springer-Germany, 2010.

[9] E.-W. Bai, "An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems," *Automatica*, vol. 34, no. 3, pp. 333–338, 1998.

[10] F. H. I. Chang and R. Luus, "A noniterative method for identification using the Hammerstein model," *IEEE Trans. Automatic Control*, vol. 16, no. 5, pp. 464–468, 1971.

[11] W. Greblicki and M. Pawlak, *Non-Parametric System Identification.* Cambridge, UK: Cambridge University Press, 2008.

[12] B. Wahlberg, H. Hjalmarsson, and J. Mårtensson, "On identification of cascade systems," in *Proc. of the 17th IFAC World Congress*, Seoul, South Korea, July 2008.

[13] L. Ljung, *System Identification, theory for the user*, 2nd ed. Prentice-Hall, 1999.

[14] J. Suykens and J. Vandewalle, "Recurrent least squares support vector machines," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 7, pp. 1109–1114, 2000.

[15] T. Falck, J. Suykens, and B. De Moor, "Linear parametric noise models for least squares support vector machines," in *Proc. of the 49th IEEE Conf. on Decision and Control*, Atlanta, USA, Dec. 2010, pp. 6389–6394.

[16] H. Garnier and L. Wang, editors, *Identification of Continuous-time Models from Sampled Data.* Springer-Verlag, 2008.

[17] P. C. Young, *Recursive estimation and time-series analysis. An introduction to the student and the practionner.* Berlin: Springer-Verlag, 2011.

[18] V. Laurain, W. Zheng, and R. Tóth, "Introducing instrumental variables in the LS-SVM based identification framework," in *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011, pp. 3198–3203.

[19] T. Söderström and P. Stoica, *Instrumental Variable Methods for System Identification.* Springer-Verlag, New York, 1983.

[20] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[21] L. Breiman, "Better subset regression using the nonnegative garotte," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[22] R. Tibshirani, "Regression shrinkage and selection with the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[23] M. Yuan and Y. Lin, "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 143–161, 2007.

[24] C. R. Rojas and H. Hjalmarsson, "SPARSEVA: Sparse estimation based on a validation criterion," in *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011, pp. 2825–2830.

[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[26] S. Saitoh, *Theory of Reproducing Kernels and its Applications.* Harlow, England: Longman Scientific & Technical, 1988.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[28] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 209, pp. 415–446, 1909.

[29] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2001.

[30] T. Howley and M. G. Madden, "The genetic kernel support vector machine: Description and evaluation," *Artificial Intelligence Review*, vol. 24, no. 3-4, pp. 379–395, 2005.

[31] P. C. Young, *Recursive Estimation and Time-Series Analysis.* Berlin: Springer-Verlag, 1984.

[32] V. Laurain, R. Tóth, and D. Piga, "An instrumental least squares support vector machine for nonlinear system identification: enforcing zero-centering constraints," Eindhoven University of Technology, Tech. Rep. TUE-CS-2013-001, 2013.

[33] M. W. Chang and C. J. Lin, "Leave-one-out bounds for support vector regression model selection," *Neural Computation*, vol. 17, no. 5, pp. 1188–1222, 2005.

[34] S. An, W. Liu, and S. Venkatesh, "Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression," *Pattern Recognition*, vol. 40, no. 8, pp. 2154–2162, 2007.

[35] X. C. Guo, J. H. Yang, C. G. Wu, C. Y. Wang, and Y. C. Liang, "A novel LS-SVMs hyper-parameter selection based on particle swarm optimization," *Neurocomputing*, vol. 71, no. 16, pp. 3211–3215, 2008.