Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Imposing Symmetry in Least Squares Support Vector Machines Regression

Marcelo Espinoza, Johan A.K. Suykens, Bart De Moor
K.U. Leuven, ESAT-SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.
{marcelo.espinoza,johan.suykens}@esat.kuleuven.ac.be

Abstract— **In this paper we show how to use relevant prior information by imposing symmetry conditions (odd or even) to the Least Squares Support Vector Machines regression formulation. This is done by adding a simple constraint to the LS-SVM model, which finally translates into a new kernel. This equivalent kernel embodies the prior information about symmetry, and therefore the dimension of the final dual system is the same as the unrestricted case. We show that using a regularization term and a soft constraint provides a general framework which contains the unrestricted LS-SVM and the symmetry-constrained LS-SVM as extreme cases. Imposing symmetry improves substantially the performance of the models, which can be seen in terms of generalization ability and in the reduction of model complexity. Practical examples of NARX models and time series prediction show satisfactory results.**

I. Introduction

In applied nonlinear system identification, the estimation of a nonlinear black-box model in order to produce accurate forecasts starting from a set of observations is a common practice. Kernel based estimation techniques, such as Support Vector Machines (SVMs) and Least Squares Support Vector Machines (LS-SVMs) have shown to be powerful nonlinear black-box regression methods [9], [16]. Both techniques build a linear model in the so-called feature space where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping φ. This is converted to the dual space by means of the Mercer's theorem and the use of a positive definite kernel, without computing explicitly the mapping φ. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution [2]. The LS-SVM formulation, on the other hand, solves a linear system in dual space under a least-squares cost function [14], where the sparseness property can be obtained by e.g. sequentially pruning the support value spectrum [12] or via a fixed-size subset selection approach [13]. The LS-SVM training procedure involves the selection of a kernel parameter and the regularization parameter of the cost function, which can be done e.g. by cross-validation, Bayesian techniques [8] or others.

Particularly when there is no a priori information about the model structure, a full nonlinear black-box model can give satisfactory results [11] for prediction or control. However, in applied work it is often the case that there exist some a priori information about the nonlinear behavior of the system under identification [7]. According to the principle that "do not estimate what you already know", the use of prior knowledge in the modelling stage can lead to important improvements. For the case when there is prior knowledge about the model structure in such a way that it is known that the nonlinearity only affects some of the inputs (and other inputs enter the model in a linear parametric way), the use of a Partially Linear LS-SVM model has been shown to improve the practical performance against a full black-box model [4], [5]. Moreover, there are cases where the simple knowledge of a general property of the symmetry of the nonlinearity can be used to improve the final modelling results [1].

In this paper we focus on the case where there exists prior knowledge on the symmetry of the unknown nonlinearity. The simple knowledge that a nonlinear function may show an even or odd symmetry can be imposed into the LS-SVM formulation in a straightforward way, reducing the model complexity and improving the generalization ability. The use of this type of prior knowledge is particularly helpful when the data available for modelling does not cover all the range on which we would like to use the model for further simulation or prediction, as it is usually the case for nonlinear time series identification [10]. In addition, we show the difference between imposing the prior knowledge as a hard or a soft constraint, providing a framework to include prior information that may not be entirely exact. This paper is structured as follows. Section II describes the derivation of the LS-SVM with symmetry constrains. Section III shows the case where the prior information is imposed via a regularization parameter and a soft constraint. Numerical applications are described on Section IV.

II. LS-SVM with Symmetry Constraints

The inclusion of a symmetry constraint (odd or even) to the nonlinearity within the LS-SVM regression framework can be formulated as follows. Given the dataset $\{\boldsymbol{x}_k, y_k\}_{k=1}^N$, with $\boldsymbol{x}_k \in \mathbb{R}^p$ and $y_k \in \mathbb{R}$, the goal is to estimate a model of the form

$$y_k = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_k) + b + e_k, k = 1, \ldots, N, \qquad (1)$$

where $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}^{n_h}$ is the mapping to a high dimensional (and possibly infinite dimensional) feature space, and the error terms e_k are assumed to be i.i.d. with zero mean and constant (and finite) variance. The following

0-7803-9568-9/05/$20.00 ©2005 IEEE

optimization problem with a regularized cost function is formulated:

$$\min_{\boldsymbol{w},b,e_k} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2$$
$$\text{s.t.} \quad \begin{cases} y_k = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + e_k, & k = 1,\ldots,N, \\ \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) = a\boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_k), & k = 1,\ldots,N, \end{cases} \tag{2}$$

with $a \in \{-1, 1\}$ a given constant. The first restriction is the standard model formulation in the LS-SVM framework. The second restriction is a shorthand for the cases where we want to impose the nonlinear function $\boldsymbol{w}^T\boldsymbol{\varphi}(x_k)$ to be even (resp. odd) by using $a = 1$ (resp. $a = -1$). The solution is formalized in the following lemma.

Lemma 1: Given the problem (2) and a positive definite kernel function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ satisfying the assumptions $K(\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(-\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $K(-\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ $\forall k, l = 1,\ldots,N$, the solution to (2) is given by the system

$$\left[\begin{array}{c|c} \frac{1}{2}(\boldsymbol{\Omega} + a\boldsymbol{\Omega}^*) + \frac{1}{\gamma}\mathbf{I} & \mathbf{1} \\ \hline \mathbf{1}^T & 0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ \hline b \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ \hline 0 \end{array}\right], \tag{3}$$

with $\boldsymbol{\Omega}_{k,l} = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $\boldsymbol{\Omega}^*_{k,l} = K(-\boldsymbol{x}_k, \boldsymbol{x}_l)$ $\forall k, l = 1,\ldots,N$.

Proof: Building the Lagrangian of the regularized cost function,

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, b, e_k, \alpha_k, \beta_k) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \\ - \sum_{k=1}^{N}(\boldsymbol{w}^T\boldsymbol{\varphi}(x_k) + b + e_k - y_k) - \\ - \sum_{k=1}^{N}(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) - a\boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_k)), \end{aligned} \tag{4}$$

with $\alpha_k, \beta_k \in \mathbb{R}$ the Lagrange multipliers, and taking the optimality conditions $\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = 0$, $\frac{\partial\mathcal{L}}{\partial b} = 0$, $\frac{\partial\mathcal{L}}{\partial e_k} = 0$, $\frac{\partial\mathcal{L}}{\partial\beta_k} = 0$ $\frac{\partial\mathcal{L}}{\partial\alpha_k} = 0$, the following system of equations is obtained:

$$\begin{aligned} \boldsymbol{w} &= \sum_{l=1}^{N}(\alpha_l + \beta_l)\boldsymbol{\varphi}(\boldsymbol{x}_l) - a\sum_{l=1}^{N}\beta_l\boldsymbol{\varphi}(-\boldsymbol{x}_l) \\ &\sum_{i=1}^{N}\alpha_i = 0, \\ \gamma e_k &= \alpha_k, \qquad k = 1,\ldots,N \\ y_k &= \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + e_k \quad k = 1,\ldots,N \\ \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) &= a\boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_k), \quad k = 1,\ldots,N \end{aligned}$$

Using Mercer's theorem, $\boldsymbol{\varphi}(\boldsymbol{x}_k)^T\boldsymbol{\varphi}(\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ for a positive definite kernel function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$[13]. Under the assumptions that $K(\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(-\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $K(-\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ $\forall k, l = 1,\ldots,N$, the elimination of $\boldsymbol{w}, e_k$ and β_k yields

$$y_k = \frac{1}{2}\sum_{l=1}^{N}\alpha_l[K(\boldsymbol{x}_l, \boldsymbol{x}_k) + aK(-\boldsymbol{x}_l, \boldsymbol{x}_k)] + b + \frac{1}{\gamma}\alpha_k \tag{5}$$

and the final dual system can be written as

$$\left[\begin{array}{c|c} \frac{1}{2}(\boldsymbol{\Omega} + a\boldsymbol{\Omega}^*) + \frac{1}{\gamma}\mathbf{I} & \mathbf{1} \\ \hline \mathbf{1}^T & 0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ \hline b \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ \hline 0 \end{array}\right], \tag{6}$$

with $\boldsymbol{\Omega}_{k,l} = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $\boldsymbol{\Omega}^*_{k,l} = K(-\boldsymbol{x}_k, \boldsymbol{x}_l)$ $\forall k, l = 1,\ldots,N$. ■

Remark 1: Kernel functions. For a positive definite kernel function $K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ some common choices are: $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \boldsymbol{x}_k^T\boldsymbol{x}_l$ (linear kernel); $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = (\boldsymbol{x}_k^T\boldsymbol{x}_l + c)^d$ (polynomial of degree d, with $c > 0$ a tuning parameter); $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \exp(-||\boldsymbol{x}_k - \boldsymbol{x}_l||_2^2/\sigma^2)$ (RBF kernel), where σ is a tuning parameter.

Remark 2: Equivalent Kernel. The final model becomes

$$\hat{y}(\boldsymbol{x}) = \sum_{l=1}^{N}\alpha_l K_{eq}(\boldsymbol{x}_l, \boldsymbol{x}) + b. \tag{7}$$

where

$$K_{eq}(\boldsymbol{x}_l, \boldsymbol{x}) = \frac{1}{2}[(K(\boldsymbol{x}_l, \boldsymbol{x}) + aK(-\boldsymbol{x}_l, \boldsymbol{x})] \tag{8}$$

is the equivalent symmetric kernel that embodies the restriction about the nonlinearity. It is important to note that the final dual system (11) has the same dimensions as the one obtained with the traditional unrestricted LS-SVM. Therefore, imposing the second constraint does not increase the dimension of the system to be solved, as the new information is translated into the kernel level.

Remark 3: Validity of the Assumptions. The assumptions $K(\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(-\boldsymbol{x}_k, \boldsymbol{x}_l)$ and $K(-\boldsymbol{x}_k, -\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$ $\forall k, l = 1,\ldots,N$ are easily verified for all kernel functions that can be expressed in terms of the distance between vectors, $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = K(\|\boldsymbol{x}_k - \boldsymbol{x}_l\|)$ (stationary kernels, e.g. RBF kernel) and those expressed in terms of the dot product $K(\boldsymbol{x}_k, \boldsymbol{x}_l) = K(\boldsymbol{x}_k^T\boldsymbol{x}_l)$ (nonstationary kernels, e.g. polynomial kernel), which are the most common kernel functions used in practical work. From a theoretical point of view, in general the kernel function can be described by its spectral representation. For the general class of kernels for which the polynomial and RBF kernels are particular cases, the spectral representation can be written as [6]:

$$K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\cos(\boldsymbol{\theta}_1^T\boldsymbol{x}_k - \boldsymbol{\theta}_2^T\boldsymbol{x}_l)F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \tag{9}$$

(where F is a bounded symmetric measure). Under this representation, noting that $\cos(z) = \cos(-z)$, it is easy to verify the required assumptions:

$$\begin{aligned} K(\boldsymbol{x}_k, -\boldsymbol{x}_l) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\cos(\boldsymbol{\theta}_1^T\boldsymbol{x}_k + \boldsymbol{\theta}_2^T\boldsymbol{x}_l)F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\cos(-[-\boldsymbol{\theta}_1^T\boldsymbol{x}_k - \boldsymbol{\theta}_2^T\boldsymbol{x}_l])F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\cos(-\boldsymbol{\theta}_1^T\boldsymbol{x}_k - \boldsymbol{\theta}_2^T\boldsymbol{x}_l)F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= K(-\boldsymbol{x}_k, \boldsymbol{x}_l) \end{aligned}$$

$$
\begin{aligned}
K(-\boldsymbol{x}_k,-\boldsymbol{x}_l) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \cos(-\boldsymbol{\theta}_1^T\boldsymbol{x}_k+\boldsymbol{\theta}_2^T\boldsymbol{x}_l)F(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \cos(-[\boldsymbol{\theta}_1^T\boldsymbol{x}_k-\boldsymbol{\theta}_2^T\boldsymbol{x}_l])F(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \cos(\boldsymbol{\theta}_1^T\boldsymbol{x}_k-\boldsymbol{\theta}_2^T\boldsymbol{x}_l)F(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\\
&= K(\boldsymbol{x}_k,\boldsymbol{x}_l)
\end{aligned}
$$

Therefore, for a large class of kernels, mostly those used in practical nonlinear system identification, the required assumptions hold. However, this may not be a general property for all possible kernels, especially those defined in new applications fields (e.g. text, chemical molecules, etc.).

III. Imposing Symmetry via a Regularization Term

In this section the symmetry is imposed as a soft constraint, which can be interpreted as a weak prior knowledge. Under the same definitions for the initial dataset $\{\boldsymbol{x}_i,y_i\}_{i=1}^N$ and the model formulation, now the following optimization problem with a regularized cost function is formulated:

$$
\begin{aligned}
\min_{\boldsymbol{w},b,e_k} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\gamma_1\frac{1}{2}\sum_{k=1}^N e_k^2+\gamma_2\frac{1}{2}\sum_{k=1}^N r_k^2\\
\text{s.t.} \quad & \begin{cases} y_k=\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k)+b+e_k, & k\in\mathcal{K},\\ \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k)=a\boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_k)+r_k, & k\in\mathcal{K}, \end{cases}
\end{aligned}
\tag{10}
$$

with $a\in\{-1,1\}$ a given constant and $\mathcal{K}=1,\ldots,N$. The second restriction, imposing $\boldsymbol{w}^T\boldsymbol{\varphi}(x_k)$ to be even (resp. odd) by using $a=1$ (resp. $a=-1$), contains now a residual term r_k thus allowing the restriction not to be exact. The "fitting" of this second restriction is included on the cost function via a new regularization term γ_2. The solution is formalized in the following lemma.

Lemma 2: Given the problem (10) and a positive definite kernel function $K:\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}$ satisfying $K(\boldsymbol{x}_k,-\boldsymbol{x}_l)=K(-\boldsymbol{x}_k,\boldsymbol{x}_l)$ and $K(-\boldsymbol{x}_k,-\boldsymbol{x}_l)=K(\boldsymbol{x}_k,\boldsymbol{x}_l)\ \forall k,l=1,\ldots,N$, the solution to (10) is given by the system

$$
\left[\begin{array}{c|c} \boldsymbol{\Omega}_{eq}+\frac{1}{\gamma_1}\mathbf{I} & \mathbf{1}\\ \hline \mathbf{1}^T & 0 \end{array}\right]\left[\begin{array}{c}\boldsymbol{\alpha}\\ \hline b\end{array}\right]=\left[\begin{array}{c}\boldsymbol{y}\\ \hline 0\end{array}\right],
\tag{11}
$$

where

$$
\boldsymbol{\Omega}_{eq}=\frac{1}{2}(\boldsymbol{\Omega}+a\boldsymbol{\Omega}^*)+\frac{1}{2\gamma_2}(a\boldsymbol{\Omega}^*-\boldsymbol{\Omega}+\frac{1}{2\gamma_2}\boldsymbol{I})^{-1}
\tag{12}
$$

and $\boldsymbol{\Omega}_{k,l}=K(\boldsymbol{x}_k,\boldsymbol{x}_l)$ and $\boldsymbol{\Omega}^*_{k,l}=K(-\boldsymbol{x}_k,\boldsymbol{x}_l)\ \forall k,l=1,\ldots,N$.

Proof: Building the Lagrangian as in (4) and taking the optimality conditions $\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}}=0$, $\frac{\partial\mathcal{L}}{\partial b}=0$, $\frac{\partial\mathcal{L}}{\partial e_k}=0$, $\frac{\partial\mathcal{L}}{\partial\beta_k}=0$ $\frac{\partial\mathcal{L}}{\partial\alpha_k}=0$ and $\frac{\partial\mathcal{L}}{\partial r_k}=0$, we obtain the system

$$
\begin{aligned}
\boldsymbol{w}=\sum_{i=1}^N(\alpha_i+\beta_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)-a\sum_{i=1}^N\beta_i\boldsymbol{\varphi}(-\boldsymbol{x}_i)&\\
\sum_{i=1}^N\alpha_i=0,&\\
\gamma_1 e_k=\alpha_k, \qquad k=1,\ldots,N&\\
\gamma_2 r_k=-\beta_k \qquad k=1,\ldots,N&\\
y_k=\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k)+b+e_k \quad k=1,\ldots,N&\\
\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k)=a\boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_k)+r_k, k=1,\ldots,N.&
\end{aligned}
$$

From this system, one can express a relation between the vectors of lagrange multipliers $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ as

$$
(\boldsymbol{\Omega}-a\boldsymbol{\Omega}^*)\boldsymbol{\alpha}=(2a\boldsymbol{\Omega}^*-2\boldsymbol{\Omega}+\frac{1}{\gamma_2}\boldsymbol{I})\boldsymbol{\beta}
\tag{13}
$$

On the other hand, the elimination of $\boldsymbol{w}$ and e_k using the optimality conditions gives (in matrix notation),

$$
\boldsymbol{y}=\boldsymbol{\Omega}\boldsymbol{\alpha}+\boldsymbol{\Omega}\boldsymbol{\beta}-a\boldsymbol{\Omega}^*\boldsymbol{\beta}+\mathbf{1}b+\frac{1}{\gamma_1}\boldsymbol{\alpha}
\tag{14}
$$

Expressing $\boldsymbol{\beta}$ in terms of $\boldsymbol{\alpha}$ from (13) into (14) gives the final system (11). ■

Remark 4: Role of second regularization term. Imposing symmetry as a soft constraint gives rise to a new equivalent kernel

$$
\boldsymbol{\Omega}_{eq}=\frac{1}{2}(\boldsymbol{\Omega}+a\boldsymbol{\Omega}^*)+\frac{1}{2\gamma_2}(a\boldsymbol{\Omega}^*-\boldsymbol{\Omega}+\frac{1}{2\gamma_2}\boldsymbol{I})^{-1}
\tag{15}
$$

which is equal to the equivalent kernel of Section II when $\gamma_2\to\infty$. This means that the hard constrained case of Section II is a particular case of the soft constrained derivation. In addition, we see that when $\gamma_2\to 0$ the regularized cost function from (10) becomes the cost function of the standard LS-SVM. When $\gamma_2\to 0$ working with the soft constraint, the optimality condition related to r_k gives $\beta_k=0$ thus killing the effect of the second constraint. Therefore, imposing symmetry via a regularization parameter and a soft constraint covers a continuum of cases: from the standard unconstrained LS-SVM ($\gamma_2\to 0$, no prior knowledge) to the hard constrained case of Section II ($\gamma_2\to\infty$, absolute prior knowledge). From this perspective, the regularization term γ_2 can measure the degree upon which the symmetry constraint can be imposed. This is also related to the Bayesian framework where prior information can be imposed via a regularization term [13], [8].

IV. Illustrative Examples

In this section, some examples of the effects of imposing symmetry to the LS-SVM are presented. On all cases, an RBF kernel is used and the parameters σ and γ are found by 10-fold cross validation over the corresponding training sample. On each example, the results using the standard LS-SVM (i.e. full black-box model) are compared to those obtained with the symmetry-constrained LS-SVM (S-LS-SVM) from (2). The examples are defined in such a way

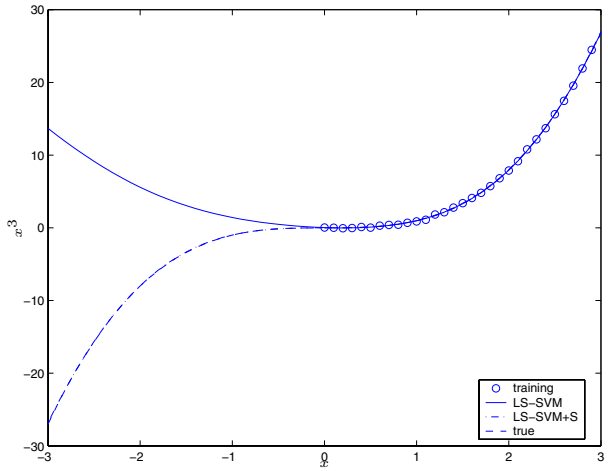

Fig. 1. Training points and predictions with LS-SVM (thin line), S-LS-SVM (dot-dashed) and the actual values (dashed line).

that there is not enough training datapoints on every region of the relevant space; thus, it is very difficult for a black-box model to "learn" the symmetry just by using the available information. The examples are compared in terms of their complexity (effective number of parameters [17]), the performance in the training sample (cross-validation mean squared error, MSE-CV) and the generalization performance (MSE out of sample, MSE-OUT). The results are shown on Table I.

A. Cubic function

The model to be identified is $y_k = x_k^3 + \varepsilon_k$, where ε_k is drawn from a Normal distribution with zero mean and variance 0.2. The training data for this example consists of $x_k \in [0, 3]$ in increments of 0.1, thus containing only positive values. The goal is to observe how well does the model generalize to the negative values of x_k. The model is formulated simply as $y_k = \varphi(x_k) + e_k$ to be identified by standard LS-SVM and by S-LS-SVM, where the symmetric condition is implemented by using $a = -1$ in (2) (odd function). Figure 1 shows the performance of the estimated models. Clearly the S-LS-SVM can generalize better by making use of the symmetry information. The effective number of parameters is reduced from 4.4 (LS-SVM) to 3 (S-LS-SVM).

B. Sinc function in 2-D

The model to be identified is $y_k = 0.5[\text{sinc}(x_k) + \text{sinc}(z_k)] + \varepsilon_k$, where ε_k is drawn from a Normal distribution with zero mean and variance 0.1. Training values for x_k range from -2.9 to 2.9, whereas the training values for z_k only take positive values in the range 0 to 2.9. The black box model is formulated as $y_k = \varphi(x_k, z_k) + e_k$ and it is estimated by LS-SVM and S-LS-SVM. The final models are then used to generalize to the other half of the space, where the input z_k has negative values. Clearly the result

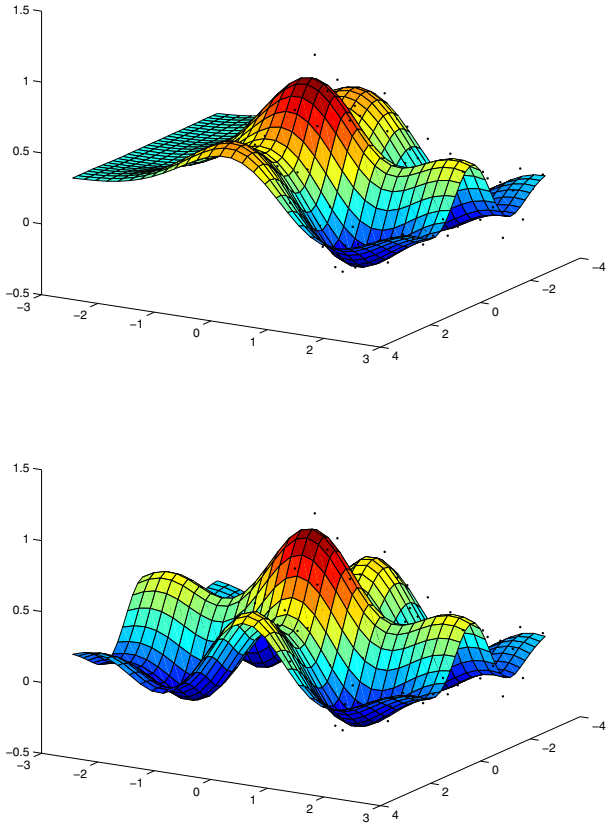

Fig. 2. Training points and predicted surface with LS-SVM (Top) and S-LS-SVM (Bottom) for the sinc function example.

from LS-SVM provides a good generalization in the range of the training data, but it fails in the region where there are no training points. The top panel of Figure 2 shows that the generalization produced by the LS-SVM is flat in the region of interest. The inclusion of a symmetric constraint ($a = 1$) corrects the problem and improves the generalization ability of the S-LS-SVM model, as shown on the bottom panel of Figure 2. In this case, the effective number of parameters is reduced from 29 to 25.

C. Lorenz attractor

This example is taken from [1]. The $x-$coordinate of the Lorenz attractor is used as an example of a time series generated by a dynamical system. Usually chaotic time series are used to produce assessments about the generalization performance of a particular black-box methodology, as in time series competitions [18], [15]. A Nonlinear AutoRegressive (NAR) black-box model is formulated:

$$y(t) = \varphi(y(t-1), y(t-2), \ldots, y(t-p)) + e(t)$$

to be identified by LS-SVM and S-LS-SVM. The order p is selected during the cross-validation process as an extra parameter. A sample of 1000 datapoints is used

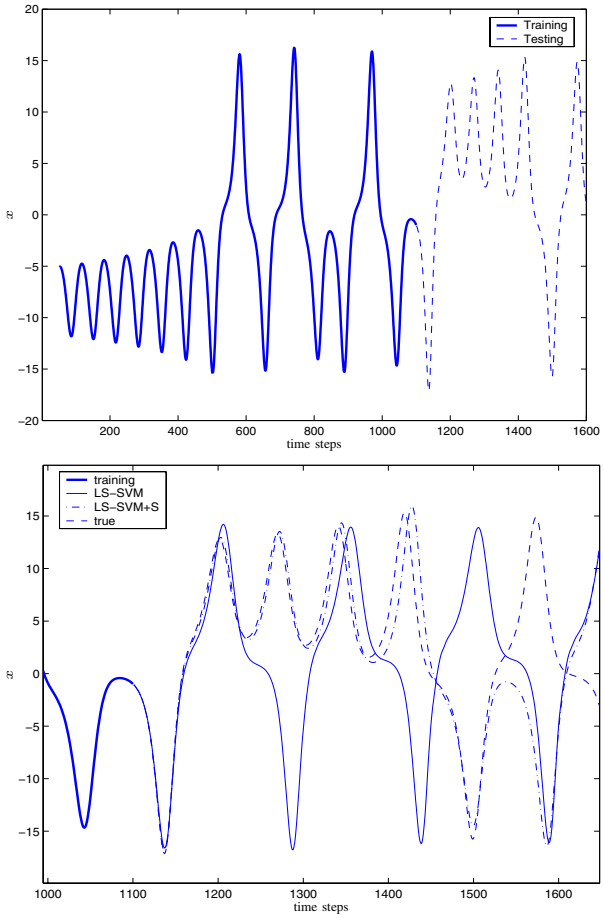

Fig. 3. (Top) The series from the $x-$coordinate of the Lorenz attractor, part of which is used for training. (Bottom) Simulations with LS-SVM (thin line), S-LS-SVM (dot-dashed) compared to the actual values (dashed line).

for training, which corresponds to an unbalanced sample over the evolution of the system. After each model is estimated, they are used in simulation mode, where the future predictions are computed with the estimated model $\hat{\varphi}$ using past predictions:

$$\hat{y}(t) = \hat{\varphi}(\hat{y}(t-1), \hat{y}(t-2), \ldots, \hat{y}(t-p)).$$

Figure 3 (top) shows the training sequence (thick line) and the future evolution that the models should be able to simulate up to a certain timestep. Figure 3 (bottom) shows the generalization zone, with the simulations obtained with LS-SVM (thin line) and S-LS-SVM (dot-dashed line). Clearly the S-LS-SVM can simulate the system for the next 500 timesteps, far beyond the 100 points that can be simulated by the LS-SVM. The effective number of parameters is reduced from 237 (LS-SVM) to 137 (S-LS-SVM).

D. The SilverBox Data

The real-life nonlinear dynamical system that was used in the NOLCOS 2004 Special Session benchmark [10] consists of a sequence of 130,000 datapoints for the input u and the

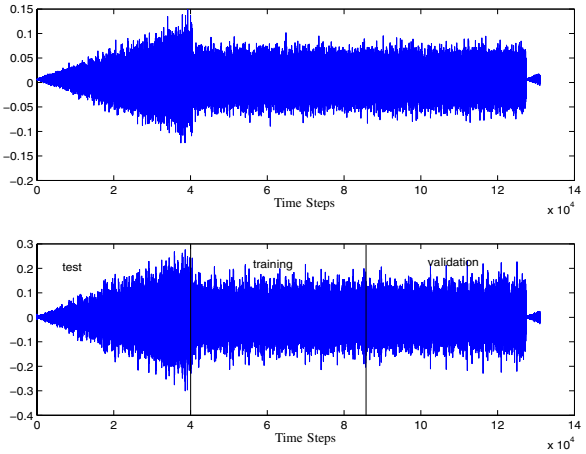

Fig. 4. Input (Top) and Output (Bottom) sequences for the SilverBox dataset. The data used for training, validation and testing is indicated.

output y measured from a real physical system. Figure 4 shows the output time series, along with the definition of the data that was used for training and final testing. The final test consists on produce a simulation for the first 40,000 datapoints (the "head of the arrow"), which requires the models to generalize on a zone of wider amplitude than the one used for training. A full black-box LS-SVM model reached excellent levels of performance [3] and now we want to check if the knowledge of the existence of an odd nonlinearity can improve further on. A NARX black-box model is formulated,

$$y(t) = \varphi(y(t-1), y(t-2), \ldots, y(t-p), \ldots \\ \ldots, u(t-1), u(t-2), \ldots, u(t-p)) + e(t)$$

which is estimated with LS-SVM and S-LS-SVM[1]. Figure 5 shows the residuals obtained with LS-SVM (top) and S-LS-SVM (Bottom) on the simulation exercise. In spite of the good performance of the LS-SVM, achieving a root mean squared error (RMSE) of 3.24×10^{-4} on this simulation, there are still some larger residuals to the end of the sequence. This is the zone of wider amplitude of the dataset. Imposing symmetry with the S-LS-SVM improves the generalization performance on the simulation by reducing the RMSE to 2.84×10^{-4}. Fewer peaks are visible in the residuals obtained with S-LS-SVM.

V. Conclusions

We have shown how to impose simple constraints with prior information about the symmetry of the unknown nonlinear function to be identified using LS-SVM. The constraint with the symmetry condition (odd or even) translates into an equivalent kernel. This makes the dimension of the

[1]Due to the large size of this sample, a fixed-size version on primal space of the LS-SVM is required. The interested reader is referred to [13], [3] for details on the equivalence between the dual and primal space formulations.

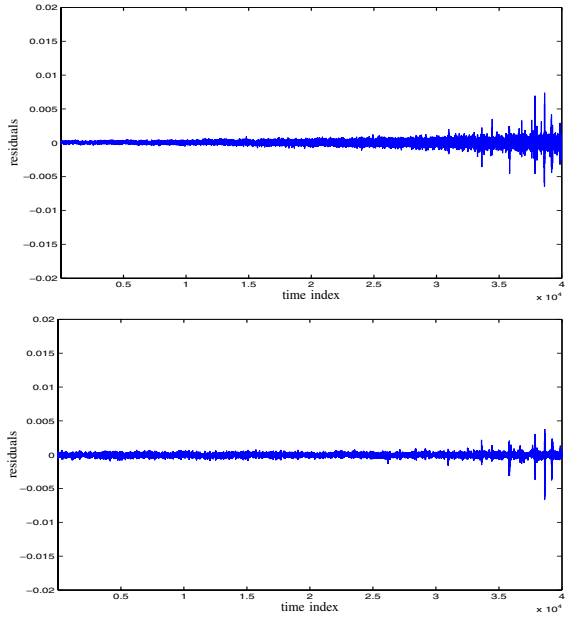

Fig. 5. Residuals of the SilverBox simulations on the test set. LS-SVM (Top) and S-LS-SVM (Bottom)

	1-D Cubic	2-D Sinc	Lorenz	SilverBox
LS-SVM				
N_{eff}	4.4	29	237	490
MSE-CV	0.011	0.010	3.41×10^{-4}	1.75×10^{-4}
MSE-OUT	156.2	0.027	52.057	3.24×10^{-4}
LS-SVM with Symmetry Constraint				
N_{eff}	3.0	25	137	490
MSE-CV	0.009	0.008	1.62×10^{-6}	0.54×10^{-4}
MSE-OUT	0.006	0.001	0.085	2.84×10^{-4}

TABLE I

PERFORMANCE COMPARISON BETWEEN LS-SVM AND S-LS-SVM.

constrained dual system to remain equal to the unrestricted case. Imposing prior knowledge as a hard constraint is a straightforward extension of the LS-SVM, where the new kernel embodies the prior information. When the symmetry is imposed as a soft constraint, the associated regularization term can be interpreted as the indicator up to which extent the prior knowledge can be imposed. When this regularization term goes to infinity, the hard constraint case is recovered. When it goes to zero, the standard LS-SVM is recovered. Practical examples of imposing symmetry show satisfactory results, in the context of NARX models and time series prediction. The generalization ability of the models is improved, and the complexity is reduced.

ACKNOWLEDGMENTS

This work is supported by grants and projects for the Research Council K.U.Leuven (GOA- Mefisto 666, GOA- Ambiorics, several PhD/ Postdocs & fellow grants), the Flemish Government (FWO: PhD/ Postdocs grants, projects G.0211.05, G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, ICCoS, ANMMM; AWI; IWT: PhD grants, GBOU (McKnow) Soft4s), the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/ 01/40), the EU (FP5- Quprodis; ERNSI, Eureka 2063- Impact; Eureka 2419- FLiTE) and Contracts Research / Agreements (ISMC /IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). J. Suykens and B. De Moor are an associated professor and a full professor at the K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

REFERENCES

[1] L.A. Aguirre, R. Lopes, G. Amaral, and C. Letellier. Constraining the topology of neural networks to ensure dynamics with symmetry properties. *Physical Review E*, 69, 2004.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[3] M. Espinoza, K. Pelckmans, L. Hoegaerts, J.A.K. Suykens, and B De Moor. A comparative study of LS-SVMs applied to the silverbox identification problem. In *Proceedings of the 6th IFAC Conference on Nonlinear Control Systems (NOLCOS)*, 2004.

[4] M. Espinoza, J. Suykens, and B De Moor. Partially linear models and least squares support vector machines. In *Proc. of the 43rd IEEE Conference on Decision and Control*, 2004.

[5] M. Espinoza, J.A.K. Suykens, and B. De Moor. Model structure determination and identification with kernel based partially linear models. Technical Report 04-110, ESAT-SCD-SISTA, K.U.Leuven, Belgium, 2004.

[6] M. Genton. Classes of kernel for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[7] T. Johansen. Identification of non-linear systems using empirical data and prior knowledge-an optimization approach. *Automatica*, 32(3):337–356, 1996.

[8] D.J.C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11:1035–1068, 1999.

[9] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[10] J. Schoukens, G. Nemeth, Y. Crama, P. abd Rolain, and R. Pintelon. Fast approximate identification of nonlinear systems. *Automatica*, 39(7), 2003.

[11] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear Black-box Modelling in System Identification: a Unified Overview. *Automatica*, 31:1691–1724, 1995.

[12] J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1-4):85–105, 2002.

[13] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[14] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[15] J.A.K. Suykens and J Vandewalle. The k.u.leuven competition data : a challenge for advanced neural network techniques. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN'2000)*, pages 299–304, Brugges,Belgium, 2000.

[16] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New-York, 1995.

[17] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

[18] A.S. Weigend and N.A. Gershenfeld, editors. *Time Series Prediction. Forecasting the Future and Understanding the past*. Addison-Wesley, Reading, MA, 1993.