Nuclear Norm Regularization for Overparametrized Hammerstein Systems

Tillmann Falck, Johan A.K. Suykens, Johan Schoukens, Bart De Moor

***Abstract*—In this paper we study the overparametrization scheme for Hammerstein systems [1] in the presence of regularization. The quality of the convex approximation is analysed, that is obtained by relaxing the implicit rank one constraint. To obtain an improved convex relaxation we propose the use of nuclear norms [2], instead of using ridge regression. On several simple examples we illustrate that this yields a solution close to the best possible convex approximation. Furthermore the experiments suggest that ridge regression in combination with a projection step yield a generalization performance close to the one obtained by nuclear norms.**

I. Introduction

The identification of block structured systems like the one shown in Figure 1 is an important problem in system identification [3]. They have several advantages over pure black box identification schemes. On the one hand algorithms that exploit prior knowledge about the structure of a model yield better performances than black box identification techniques [4]. On the other hand a block structure model is much better in terms of interpretability as the blocks can be for example visualized or related to physical properties. Popular classes of block structured systems are Hammerstein, Wiener, Hammerstein-Wiener and Wiener-Hammerstein systems. Several different state of the art identification approaches for Wiener-Hammerstein systems have been compared in a special session at SYSID2009 [5]. Most of those methods can be applied to some or all of the afore mentioned classes. If the linear dynamics of a system are known a priori a more general class of block structured nonlinear systems based on linear fractional transformations (LFTs) [6], [7] can be estimated.

This paper is concerned with the identification of Hammerstein systems (Fig. 1) but the results can be extended to Hammerstein-Wiener and Wiener-Hammerstein systems with minor modifications. The only assumptions are that the nonlinear block f is smooth and the model order for H is known. The results are based on the overparametrization technique studied in [1], [8]. This method has an implicit rank constraint which is neglected by most publications. In this paper, we analyze how regularization affects the rank constraint and propose the use of nuclear norms [2] to obtain low rank solutions.

T. Falck, J. Suykens and B. De Moor are with SCD research group of the Department of Electrical Engineering (ESAT) at the Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium. Email: {tillmann.falck,johan.suykens,bart.demoor}@esat.kuleuven.be

J. Schoukens is with the Faculty of Engineering - ELEC at the Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussel, Belgium. E-mail: johan.schoukens@vub.ac.be

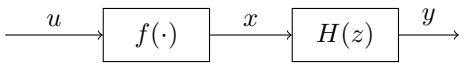


Fig. 1: General structure of a Hammerstein system

The notation used in this paper is as follows. Vectors are denoted by lowercase boldface letters and matrices by boldface capitals. Elements of vectors and matrices are denoted by subscripts on the non boldface letter and the transpose is indicated by a superscript ^{T}. The estimate of a variable $\boldsymbol{x}$ is denoted by $\hat{\boldsymbol{x}}$. The backshift operator in time is denoted by z and defined as $x_{t-1} = z^{-1}x_t$. The trace of a square matrix $\boldsymbol{X}$ is denoted by $\mathrm{tr}(\boldsymbol{X})$, $\|\boldsymbol{X}\|_F$ is the Frobenius norm of $\boldsymbol{X}$ and $\|\boldsymbol{X}\|_2$ is the largest singular value of $\boldsymbol{X}$ and called spectral norm. Finally '$\succeq$' is used for conic inequalities.

After giving a brief introduction in the present section, overparametrization for Hammerstein systems is described in Section II. The focus is on its implicit rank constraint and regularization problem formulations. In particular we give examples for which the rank constraint is not automatically satisfied. Section IV briefly reviews nuclear norms as tool for rank minimization and states an improved problem formulation. It takes the rank constraint into account and is able to obtain low rank solutions more readily than the unmodified method. Numerical simulations are performed in Section V. Finally the conclusions are drawn in VI.

II. Identification of Hammerstein Systems

Definition 1 (Hammerstein System): The system shown in Figure 1 consisting of a cascade of a static nonlinearity and a dynamic linear system is called Hammerstein system. It is described by

$$y_t = H(z)x_t, \quad x_t = f(u_t),$$

where f typically possesses a smoothness property.
Suppose that $H(z)$ is parametrized as an ARX model $B(z)/A(z)$ with orders $P, Q \in \mathbb{N}$. Thus it can be written as

$$y_t = \sum_{q=0}^{Q} b_q x_{t-q} + \sum_{p=1}^{P} a_p y_{t-p} = \boldsymbol{b}^T \boldsymbol{x}_t + \boldsymbol{a}^T \boldsymbol{y}_{t-1} \qquad (1)$$

with parameters $\boldsymbol{a} = [a_1, \ldots, a_P]^T \in \mathbb{R}^P$ and $\boldsymbol{b} = [b_0, \ldots, b_Q]^T \in \mathbb{R}^{Q+1}$ and lagged outputs $\boldsymbol{y}_{t-1} = [y_{t-1}, \ldots, y_{t-P}]^T$ and inputs $\boldsymbol{x}_t = [x_t, \ldots, x_{t-Q}]^T$. Furthermore we represent f as the linear combination of M basis

functions

$$x_t = f(u_t) = \sum_{m=1}^{M} c_m \psi_m(u_t) = \boldsymbol{c}^T \boldsymbol{\psi}(u_t). \tag{2}$$

For convenience the known basis functions are collected in a vector $\boldsymbol{\psi}(x) = [\psi_1(x), \ldots, \psi_M(x)]^T$ and the corresponding weights in $\boldsymbol{c} = [c_1, \ldots, c_M]^T$. We restrict x to the closed domain $\mathcal{D}$ and assume that the basis functions ψ_m are square integrable on $\mathcal{D}$.

Given a set of measurements $\{(u_t, y_t)\}_{t=1}^N$ of inputs and outputs a nonlinear least squares problem that fits the measurements to the model parameters $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ can be stated. The regularized optimization problem that solves the estimation task is

$$\begin{aligned} &\min_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},e_t,x_t} \frac{1}{2}\lambda\nu(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + \frac{1}{2}\sum_{t=Q+1}^{N} e_t^2 \\ &\text{subject to} \\ &\quad y_t = \boldsymbol{b}^T\boldsymbol{x}_t + \boldsymbol{a}^T\boldsymbol{y}_{t-1} + e_t, \qquad t = Q+1, \ldots, N \\ &\quad x_t = \boldsymbol{c}^T\boldsymbol{\psi}(u_t), \qquad\qquad\quad t = 1, \ldots, N. \end{aligned} \tag{3}$$

The regularization parameter λ is non negative. Note that the objective is always convex as the regularization term $\nu(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is convex by assumption. The constraints are non convex due to the multiplication of $\boldsymbol{b}$ and $\boldsymbol{c}$. Therefore the complete problem is non convex as well.

If only little information on the static nonlinearity is present it is usually necessary to employ many basis functions. This can lead to numerical conditioning problems as well as overfitting of the model. Both problems can be tackled using regularization. This reduces the number of effective parameters and performs a bias variance tradeoff. In the scope of this paper we consider regularization terms $\nu(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ that are convex functions of $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$. A frequently used regularization scheme is ridge regression [9]. It can be written as $\nu(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \|\boldsymbol{c}\|_2^2 = \boldsymbol{c}^T\boldsymbol{c}$ when applied only to the coefficients of f.

The problem (3) needs to be transformed into an equivalent problem, such that a convex relaxation can be found. To state the equivalent problem consider a slightly different regularization term

$$\nu(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \|\boldsymbol{b}\|_2^2 \|\boldsymbol{c}\|_2^2. \tag{4}$$

Remark 1: By taking the logarithm this can be transformed into a convex problem

$$\bar{\nu}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \begin{cases} \min\limits_{\beta,\gamma} 2\log(\beta) + 2\log(\gamma) \\ \text{subject to} \\ \quad \exp(\|\boldsymbol{b}\|_2) \leq \beta, \exp(\|\boldsymbol{c}\|_2) \leq \gamma. \end{cases}$$

Consider that the optimal solution to problem (3) for ν given by (4) and λ is obtained by $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$. Then solving (3) for $\bar{\nu}$ with $\bar{\lambda}$ given by

$$\bar{\lambda} = \frac{\|\boldsymbol{b}^*\|_2^2 \|\boldsymbol{c}^*\|_2^2}{\|\boldsymbol{b}^*\|_2^2 + \|\boldsymbol{c}^*\|_2^2}\lambda$$

will have the same optimal solution $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$.

III. Overparametrization

A. Overview

The technique of overparametrization [1], [8] can be summarized in three steps.

1) Introduction of new variables $\boldsymbol{c}_q \in \mathbb{R}^M$ for $q = 0, \ldots, Q$ such that a new model equation can be formulated as $y_t = \sum_{q=0}^{Q} \boldsymbol{c}_q^T\boldsymbol{\psi}(u_{t-q}) + \boldsymbol{a}^T\boldsymbol{y}_{t-1} + e_t$.
2) Stating and solving a linear least squares problem in $\boldsymbol{a}$ and the new variables $\boldsymbol{C} = [\boldsymbol{c}_0, \ldots, \boldsymbol{c}_Q]$.
3) Projecting the solution $\boldsymbol{C}$ onto the original model class parametrized by $\boldsymbol{b}$ and $\boldsymbol{c}$.

In the first phase, new variables $\boldsymbol{C}$ are introduced and the equivalent problem to (3) in the new variables is stated. This can be achieved for the regularizer in (4).

Proposition 1 (Equivalent problem): The nonlinear least squares problem in $\boldsymbol{a}$, $\boldsymbol{C}$ and e_t

$$\begin{aligned} &\min_{\boldsymbol{a},\boldsymbol{C},e_t} \frac{1}{2}\lambda\nu(\boldsymbol{a},\boldsymbol{C}) + \frac{1}{2}\sum_{t=Q+1}^{N} e_t^2 \\ &\text{subject to} \\ &\quad y_t = \operatorname{tr}(\boldsymbol{C}^T\boldsymbol{\Psi}_t) + \boldsymbol{a}^T\boldsymbol{y}_{t-1} + e_t, \qquad t = Q+1, \ldots, N \\ &\quad \operatorname{rank}(\boldsymbol{C}) = 1 \end{aligned} \tag{5}$$

with $\nu(\boldsymbol{a}, \boldsymbol{C}) = \|\boldsymbol{C}\|_F^2$ is equivalent to (3) with (4) in $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and e_t. The matrix $\boldsymbol{C}$ is of dimension $M \times (Q+1)$ and $\boldsymbol{\Psi}_t = [\boldsymbol{\psi}(u_t), \ldots, \boldsymbol{\psi}(u_{t-Q})]$ for $t = Q+1, \ldots, N$.

Proof: Consider new variables $\boldsymbol{c}_q \in \mathbb{R}^M$ to rewrite $\boldsymbol{b}^T\boldsymbol{x}_t = \sum_{q=0}^{Q} b_q \boldsymbol{c}^T\boldsymbol{\psi}(u_{t-q})$ in (3) as

$$\sum_{q=0}^{Q} \boldsymbol{c}_q^T\boldsymbol{\psi}(u_{t-q}) = \operatorname{tr}(\boldsymbol{C}^T\boldsymbol{\Psi}_t)$$

with $\boldsymbol{C} = [\boldsymbol{c}_0, \ldots, \boldsymbol{c}_Q]$ and subject to $\operatorname{rank}(\boldsymbol{C}) = 1$. Imposing $\operatorname{rank}(\boldsymbol{C})$ is equal to one is then equivalent to requiring $\boldsymbol{c}_q = b_q\boldsymbol{c}$ or equivalently $\boldsymbol{C} = \boldsymbol{c}\boldsymbol{b}^T$.

For the regularization term it holds

$$\|\boldsymbol{C}\|_F^2 = \|\boldsymbol{c}\boldsymbol{b}^T\|_F^2 = \|\boldsymbol{b}\|_2^2 \|\boldsymbol{c}\|_2^2.$$

■

For the second stage of the overparametrization technique, the nonconvex problem has to be relaxed to a convex one. Afterwards the resulting linear least squares problem can be solved to obtain estimates for $\boldsymbol{a}$ and $\boldsymbol{C}$.

Proposition 2 (Convex relaxation): A well known convex relaxation for (5) to a linear least squares problem can be obtained by dropping the rank one constraint.

$$\begin{aligned} &\min_{\boldsymbol{a},\boldsymbol{C},e_t} \frac{1}{2}\lambda\nu(\boldsymbol{a},\boldsymbol{C}) + \frac{1}{2}\sum_{t=Q+1}^{N} e_t^2 \\ &\text{subject to} \\ &\quad y_t = \operatorname{tr}(\boldsymbol{C}^T\boldsymbol{\Psi}_t) + \boldsymbol{a}^T\boldsymbol{y}_{t-1} + e_t, \qquad t = Q+1, \ldots, N \end{aligned} \tag{6}$$

with $\nu(\boldsymbol{a}, \boldsymbol{C}) = \|\boldsymbol{C}\|_F^2$.

Rewriting the regularization term as

$$\nu(\boldsymbol{a}, \boldsymbol{C}) = \|\boldsymbol{C}\|_F^2 = \|[\boldsymbol{c}_0, \ldots, \boldsymbol{c}_Q]\|_F^2 = \sum_{q=0}^{Q} \boldsymbol{c}_q^T \boldsymbol{c}_q = \sum_{q=1}^{Q} \|\boldsymbol{c}_q\|_2^2 \quad (7)$$

provides an interpretation in absence of the rank constraint. Instead of regularizing the coefficients of a single function f the new regularizer is applied to the coefficients $\boldsymbol{c}_q$ of implicit functions $f_q(u_t) = \boldsymbol{c}_q^T \boldsymbol{\psi}(u_t)$.

The last stage of the overparametrization technique finally projects the estimated parameters $\widehat{\boldsymbol{C}}$ onto the original rank one model class parametrized by $\boldsymbol{b}$ and $\boldsymbol{c}$. The projection step is carried out by solving $\hat{\boldsymbol{b}}, \hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{b},\boldsymbol{c}} \|\widehat{\boldsymbol{C}} - \boldsymbol{c}\boldsymbol{b}^T\|_F$. This can be done by computing the singular value decomposition (SVD) of $\widehat{\boldsymbol{C}}$.

B. Rank one constraint violation

The bias introduced by the projection step is analyzed in [10]. In the presence of the regularization term ν it is hard to quantify this bias. Thus it cannot be compensated easily. Therefore we are interested in regularization schemes that yield low rank solutions for $\boldsymbol{C}$. In this way no additional bias is introduced by the projection step. Therefore the bias does not have to be removed in an additional processing step.

A solution to a relaxed version of (5) is a good approximation to the original problem if the estimate $\widehat{\boldsymbol{C}}$ is low rank. This provides another justification to investigate regularization schemes that promote low rank solutions. In that case the constraint violation of $\operatorname{rank}(\boldsymbol{C}) = 1$ will be small and the obtained solution close to the global optimal solution that could be obtained by solving the non convex problem.

A sufficient condition for $\widehat{\boldsymbol{C}}$ to be asymptotically rank one is established by Bai [8]. It holds for the unregularized problem $\nu(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = 0$ under the following assumptions

1) the nonlinearity can be expressed in terms of the chosen basis $f^{\text{true}}(x) = \sum_{m=1}^{M} c_m^{\text{true}} \psi_m(x)$ and
2) the matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}(u_1), \ldots, \boldsymbol{\psi}(u_N)]$ has full row rank (a persistency of excitation condition).

The results are established in the presence of an additive and zero mean noise term with finite second moments. Without using a very rich set of basis functions assumption 1 is hard to satisfy in practice. Assumption 2 is relatively easy to satisfy in theory but in the presence of many basis functions much data is needed to also achieve a good numerical conditioning. In many practical applications either assumption one or two might not be satisfied. This is especially true for colored input signals u_t, like present in Wiener-Hammerstein systems.

In the following section we summarize some recent advances in convex optimization for heuristics for rank minimization and propose a new regularization scheme for overparametrized problems that helps finding low rank solutions for (5).

IV. Convex Relaxation for Rank Minimization

A. Brief Overview

Constraining the rank of a matrix is in general non-convex and a NP-hard problem. Therefore convex heuristics have been developed to obtain approximate solutions in polynomial time. For a rectangular matrix $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ the convex envelope of the rank function $\operatorname{rank}(\boldsymbol{X})$ for $\|\boldsymbol{X}\|_2 \leq 1$ is the nuclear norm [2], [11]. It is defined as

$$\|\boldsymbol{X}\|_* = \sum_{n=1}^{\min(M,N)} \sigma_n(\boldsymbol{X})$$

where $\sigma_n(\boldsymbol{X})$ is the n-th singular value of the matrix $\boldsymbol{X}$. The nuclear norm can be seen as a generalization of the l_1 vector norm. In fact it is the l_1 norm of the singular values of a matrix. As the l_1-norm induces sparsity of a vector, the nuclear norm promotes low rank solutions. Conditions for sparse solutions are studied in [12].

The nuclear norm has already been applied to several problems in system identification [2], [13], [14]. References to applications of nuclear norms in compressed sensing and of l_1-norms in many fields ranging from statistics to signal processing, including system identification, can be found in the references cited in this section.

To solve nuclear norm problems in polynomial time using general purpose convex optimization solvers it can be formulated as semidefinite programming (SDP) problem. $\|\boldsymbol{X}\|_*$ can be computed as [2]

$$\begin{aligned} &\min_{\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}} \frac{1}{2}\operatorname{tr}(\boldsymbol{U}) + \frac{1}{2}\operatorname{tr}(\boldsymbol{V}) \\ &\text{subject to} \\ &\begin{bmatrix} \boldsymbol{U} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{V} \end{bmatrix} \succeq 0 \end{aligned}$$

with $\boldsymbol{U} = \boldsymbol{U}^T$ and $\boldsymbol{V} = \boldsymbol{V}^T$ symmetric. The number of variables in this SDP embedding is large. Therefore only small scale problems can be solved using general purpose solvers. A brief overview of ongoing research on solvers is given in [11, Section 5]. Some examples are interior point solvers that exploit special problem structure [14] and first order methods based on subgradients [15], [16].

The scope of this paper is the modeling aspect, therefore we employ a general purpose solver [17] and restrict ourselves to small problem sizes.

B. Application to overparametrized problems

The nuclear norm heuristic can be used to obtain an improved approximation of the nonlinear least squares problem in (3). This is outlined in the following proposition.

Proposition 3 (Nuclear norm based convex relaxation): Replacing the regularizer $\nu(\boldsymbol{a}, \boldsymbol{C}) = \|\boldsymbol{C}\|_F^2$ in (6) by $\|\boldsymbol{C}\|_*$ takes the rank constraint on $\boldsymbol{C}$ into account. Thus a tighter convex relaxation to (5) is obtained when dropping the rank

one constraint

$$\min_{\boldsymbol{a},\boldsymbol{C},e_t} \lambda\|\boldsymbol{C}\|_* + \frac{1}{2}\sum_{t=Q+1}^{N} e_t^2$$

subject to

$$y_t = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Psi}_t) + \boldsymbol{a}^T\boldsymbol{y}_{t-1} + e_t, \quad t = Q+R+2,\ldots,N.$$

Remark 2: Most literature on sparse recovery with l_1 and nuclear norms does not square the norm in the objective. We follow this convention here as it is usually required by solvers. Note though the solution for the squared norm can be obtained for $\lambda = \lambda_S\|\boldsymbol{C}\|_*$.

V. Numerical Experiments

In the following we compare different choices for the regularization term ν. On the one hand the effect it has on the rank of the estimate $\widehat{\boldsymbol{C}}$ is analyzed. On the other hand the generalization performance is studied. We consider the following regularizers

1) ridge regression: $\nu(\boldsymbol{a},\boldsymbol{C}) = \|\boldsymbol{C}\|_F^2$ (cf. (7)) and
2) nuclear norm: $\nu(\boldsymbol{a},\boldsymbol{C}) = \|\boldsymbol{C}\|_*^2$.

To illustrate possible dependence on the choice of basis functions, two differrent bases are used. First Hinge functions which we define as

$$\psi_1^H(x) = 1,\ \psi_2^H(x) = x \text{ and}$$
$$\psi_m^H(x) = \begin{cases} x - b_m, & x \geq b_m \\ 0, & \text{otherwise} \end{cases}$$

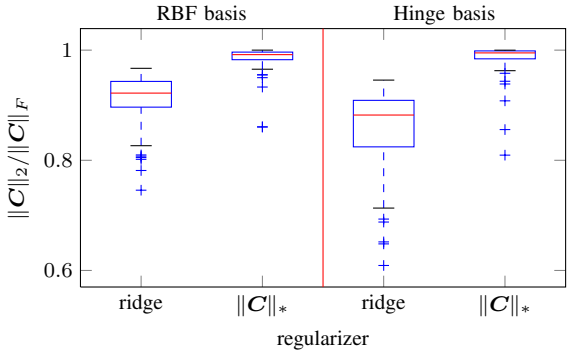
for $m = 3,\ldots,M$. The location of the kinks is given by the parameters b_m which we assume given. For our experiments we use a uniform distribution. The other considered basis functions are Gaussian Radial Basis Functions (RBFs) defined by

$$\psi_m^{\mathrm{RBF}}(\boldsymbol{x}) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}_m\|_2^2/\sigma^2).$$

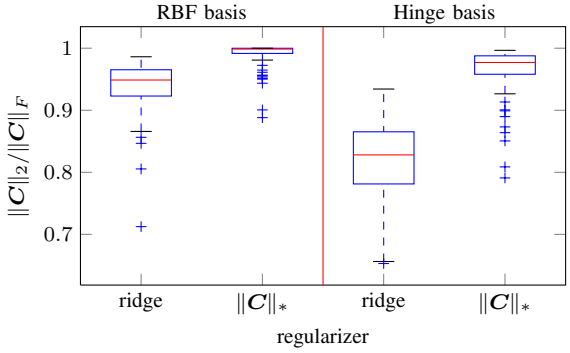
The supporting points $\boldsymbol{z}_m$ and the bandwidth σ are assumed to be known. During the following experiments the bandwidth $\sigma = 1$ is used along with a uniform distribution of basis functions.

A. Setup

We consider the identification of Hammerstein systems with the nonlinearities $f_1(u_t) = \mathrm{sinc}(u_t)$ and $f_2(u_t) = \tanh(u_t)$. The linear block $H(z)$ is minimum phase with 5 randomly generated poles and zeros. To obtain a more challenging estimation problem we use filtered white Gaussian noise $v_t \sim \mathcal{N}(0,1)$ as input signal $u_t = 0.9u_{t-1} + v_t$. For each example we generate 300 samples corrupted by additive white Gaussian noise with variance $\sigma^2 = 0.2^2$. The data is split into three equal parts, the first is used for estimation, the second to select the regularization parameter λ and the last to assess the generalization performance before projection. The root mean squared error (RMSE) is used to asses model performance. All examples use 30 basis functions and use the true model orders for P and Q.



(a) $f_1(u_t) = \mathrm{sinc}(u_t)$



(b) $f_2(u_t) = \tanh(u_t)$

Fig. 2: Analysis of rank one constraint violation for the different regularization schemes. The statistics are generated from 100 consecutive runs. The setup is the one described in V-A.

B. Rank constraint violation

The ratio

$$\frac{\|\boldsymbol{C}\|_2}{\|\boldsymbol{C}\|_F} = \frac{\sigma_1}{\sqrt{\sum_{k=1}^{K}\sigma_k^2}}$$

is a measure for the closeness of $\boldsymbol{C}$ to rank one. Here $K = \min(Q+1, M)$ and $\sigma_1 \geq \cdots \geq \sigma_K$ is the ordered sequence of singular values of $\boldsymbol{C}$. Thus a value close to one means that most energy is concentrated in the largest singular value. Therefore such a matrix is close to rank one.

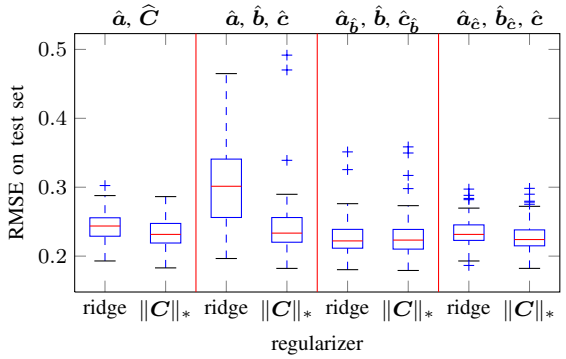
From the results shown in Figure 2 it can be concluded that the nuclear norm is superior to ridge regression in terms of low rank solutions. This trend is more pronounced for the Hinge basis. The approximation of the saturation like tanh is slightly better than that of sinc for both regularizers.

C. Projection schemes

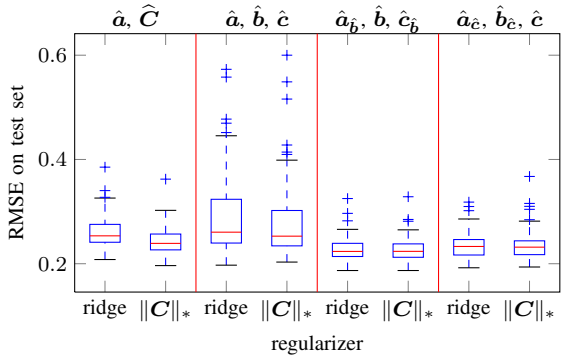
Let $\widehat{\boldsymbol{C}}$ have the singular value decomposition

$$\widehat{\boldsymbol{C}} = \sum_{k=1}^{K}\sigma_k\boldsymbol{u}_k\boldsymbol{v}_k^T$$

with $K = \min(Q+1, M)$. Then $\hat{\boldsymbol{c}} = \sqrt{\sigma_1}\boldsymbol{u}_1$ and $\hat{\boldsymbol{b}} = \sqrt{\sigma_1}\boldsymbol{v}_1$ are a solution to the projection problem $\hat{\boldsymbol{b}},\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{b},\boldsymbol{c}}\|\widehat{\boldsymbol{C}} - \boldsymbol{c}\boldsymbol{b}^T\|_F$. The projection of $\widehat{\boldsymbol{C}}$ onto the original parameters can be used for predictions in several ways.



(a) with RBF basis



(b) with Hinge basis

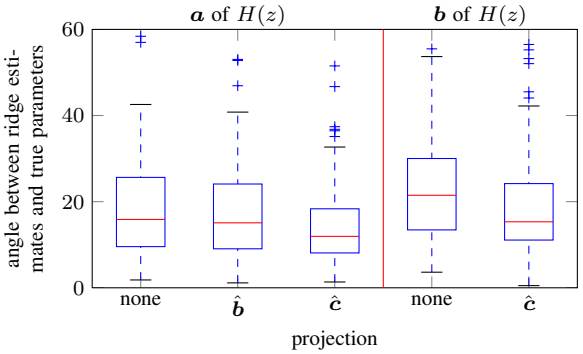
Fig. 3: Comparison of different projection schemes applied to the solutions of the two regularized problems. The plots depict generalization performaces for the unprojected model, as well as for the different projections listed in Sec. V-C. The statistics are generated from 100 consecutive runs. The setup is the one described in V-A with sinc as nonlinearity.

1) Use $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{c}}$ in a model of given by (1) and (2).
2) Fix $\hat{\boldsymbol{b}}$ and solve (3) for $\hat{\boldsymbol{a}}_{\hat{\boldsymbol{b}}}$ and $\hat{\boldsymbol{c}}_{\hat{\boldsymbol{b}}}$ and use those estimates.
3) Fix $\hat{\boldsymbol{c}}$ and solve (3) for $\hat{\boldsymbol{a}}_{\hat{\boldsymbol{c}}}$ and $\hat{\boldsymbol{b}}_{\hat{\boldsymbol{c}}}$ and use those estimates.

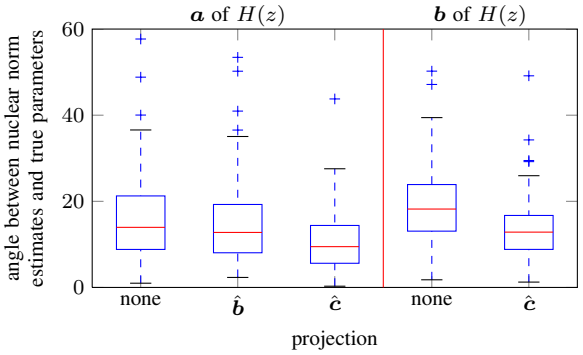
From Figure 3 we can follow while 2) and 3) usually improve the generalization performance, 1) leads to a degradation. Furthermore using the estimate for $\boldsymbol{b}$ to reestimate the remaining parameters is slightly better than using $\hat{\boldsymbol{c}}$ for that purpose. The projection levels the performance of the regularization schemes. Even though the differences are small the nuclear norm usually has a small advantage.

D. Parameter estimates

Figure 4 shows the angle between the true coefficients for $H(z)$ and their estimated values. We compare the parameter estimates for the unprojected estimates and their values after projection and reestimation. The best correlation is obtained if the model is reestimated using the estimate $\hat{\boldsymbol{c}}$. Reestimation using $\hat{\boldsymbol{b}}$ yields only small improvements. This is in contrast to the result for generalization performance in the previous section where the refined model using $\hat{\boldsymbol{b}}$ was best. Comparing ridge regression and nuclear norm in the upper and lower panel of Fig. 4 respectively, shows that the nuclear norm is slightly better in recovering the parameters. Finally in Fig.



(a) for ridge regression



(b) for nuclear norm

Fig. 4: Comparison of parameter estimates for coefficients of the linear block $H(z)$. The estimates are compared for the unprojected as well as the projected models. The statistics are generated from 100 consecutive runs. The setup is the one described in V-A with sinc as nonlinearity.

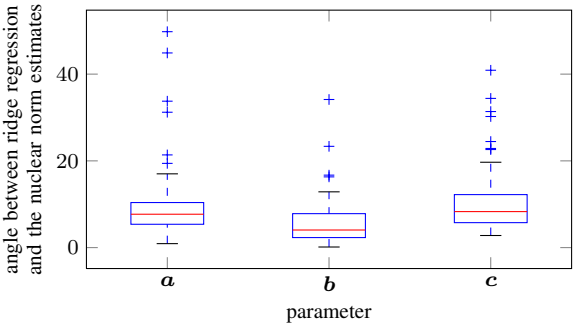


Fig. 5: Similarity of parameter estimates compared for the unprojected ridge regression and nuclear norm estimates. models. The statistics are generated from 100 consecutive runs. The setup is the one described in V-A with sinc as nonlinearity.

TABLE I: Generalization performance and rank one constraint violation for the Wiener-Hammerstein system described in [5]. Comparison of ridge regression and nuclear norm.

regularizer	$\|\boldsymbol{C}\|_2/\|\boldsymbol{C}\|_F$	$10^3 \cdot$ RMSE, using $\widehat{\boldsymbol{C}}$	$\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{c}}$	$\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{c}}_{\hat{\boldsymbol{b}}}$	$\hat{\boldsymbol{b}}_{\hat{\boldsymbol{c}}}$, $\hat{\boldsymbol{c}}$
ridge regression	0.62	2.41	43.2	3.33	5.09
nuclear norm	0.77	2.38	19.0	2.75	4.72

5 the correlation between the parameter estimates for ridge regression and the nuclear norm is shown. The correlation between the different estimates is much higher than between the true parameters and their estimates. Yet on average there is still an angle of $5^\circ - 10^\circ$ between them.

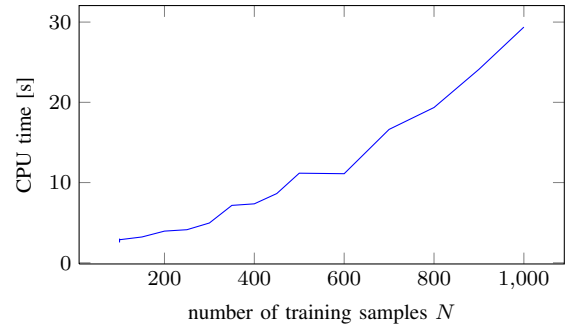
E. Real work example

In the following we compare the regularization schemes on the Wiener-Hammerstein benchmark [5] data set. We extend the modeling approach from purely Hammerstein systems to Wiener-Hammerstein systems. Therefore replace u_t by $\boldsymbol{u}_t = [u_t, u_{t-1}, \ldots, u_{t-R}]$ and extend $f : \mathbb{R} \to \mathbb{R}$ such that $f : \mathbb{R}^{R+1} \to \mathbb{R}$. This means that a multi-dimensional static nonlinearity is estimated instead of a static one-dimensional nonlinearity. The objective of this example is only to evaluate the two regularization schemes on a real world data set. Without special solvers the nuclear norm problem can only be solved for small numbers of basis functions. Therefore the absolute performances are lower than those in the special session at SYSID2009 devoted to the identification of [5].

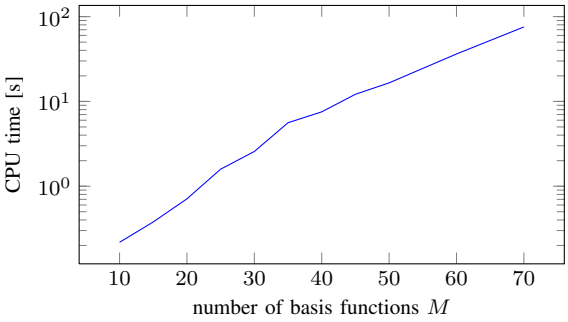
As in the previous sections we use a validation set for model selection and an independent test set for the final evaluations. Training, validation and test set are 1000 samples long consecutive parts of the complete time series starting from sample 10,000. We restrict the simulations to 30 RBF basis functions, which are distributed according to a normal distribution. The bandwidth of the RBF functions, the standard deviation of the normal distribution as well as the model orders P, Q, R and the regularization parameter λ are selected based on performance on the validation set. The selection procedure is run for the model with ridge regression. The selected model orders are $\hat{P} = 13$, $\hat{Q} = 20$ and $\hat{R} = 2$.

Table I summarizes the results. It can be seen that the results for artificial problems can be transferred to a real data set. The nuclear norm regularization results in a much better approximation of the rank-1 constraint. Yet in terms of generalization performance the model obtained by ridge regression yields very similar values.

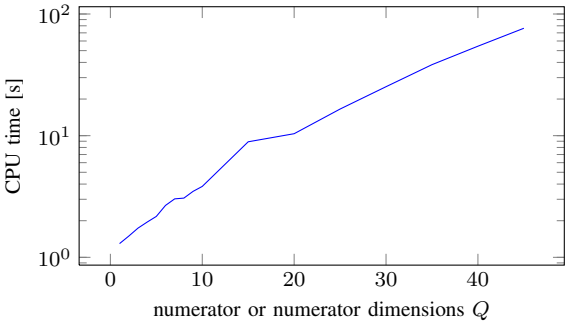
The overall poor performance and the decrease in performance after projection is due to the low number of basis functions. As the model order R grows, the fixed number of basis functions have to cover a larger space. Therefore the selected model orders are chosen suboptimally. The estimate for R is selected small enough to be described by 30 basis functions. Then the orders P and Q are chosen high such that they can compensate for the loss of expressive power.



(a) basis functions $M = 30$, numerator dimensionality $Q = 5$



(b) training samples $N = 100$, numerator dimensionality $Q = 5$



(c) training samples $N = 100$, basis functions $M = 30$

Fig. 6: Time to estimate a single instance of (3) as a function of training samples N, of basis functions M and of numerator coefficients Q while fixing the other quantities. The plots show average CPU times for 20 executions of the same problem.

F. Numerical complexity

In Figure 6 we show the runtime of an estimation problem including the nuclear norm. The setup is identical to the one used in V-A. $\mathrm{sinc}(u_t)$ is used as nonlinearity. The measurements were taken on a single node of the VIC3 supercomputer[1] at the K.U.Leuven. Only a single core of a Xeon 5420 with 2.5 GHz is used for the simulation. It can be seen that the computation time increases on a linear scale in the number of training samples N and approximately exponential in the number of basis functions M and numerator coefficients Q. Note that the computational burden can be decreased by exploiting structure [14] or using first order schemes [15], [16].

[1] https://vscentrum.be/

VI. Conclusions

We considered the identification of Hammerstein systems using many basis functions. This approach is powerfull in practice as only little prior knowledge on the nonlinearity is needed. The drawback is that effective regularization is needed to cope with the large number of parameters, especially if overparametrization is applied to convexify the problem. We discussed two different regularization schemes, ridge regression and nuclear norms. The nuclear norm is a better convex approximation to the original non convex problem (5) as the rank one constraint of the overparametrization is taken into account. For artificial examples the nuclear norm yields almost rank one solutions. Therefore its solution is almost feasible for the non convex problem and globally optimal for the convex approximation. This means that it is the best convex approximation that can be obtained for (3) with the regularizer given by (4).

The experiments have shown that ridge regression and nuclear norm yield very similar prediction performance after projection onto the true model class. Therefore, for our examples, ridge regression in combination with projection are a good approximation for the non convex initial problem (3).

Acknowledgments

Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC); IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011; EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger.

Johan Suykens is a professor and Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Johan Schoukens is a full professor at the Vrije Universiteit Brussel, Belgium.

References

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

[2] M. Fazel, "Matrix Rank Minimization with Applications," PhD, Stanford, 2002.

[3] L. Ljung, *System identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.

[4] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, pp. 1691–1724, December 1995.

[5] J. Schoukens, J. A. K. Suykens, and L. Ljung, "Wiener-Hammerstein benchmark," in *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009.

[6] K. Hsu, K. Poolla, and T. L. Vincent, "Identification of Structured Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2497–2513, 2008.

[7] K. Hsu, T. L. Vincent, G. Wolodkin, S. Rangan, and K. Poolla, "An LFT approach to parameter estimation," *Automatica*, vol. 44, no. 12, pp. 3087–3092, 2008.

[8] E.-W. Bai, "An optimal two stage identification algorithm for Hammerstein-Wiener nonlinear systems," in *Proceedings of the 1998 American Control Conference*, vol. 5. American Autom. Control Council, 1998, pp. 2756–2760.

[9] R. W. Hoerl, "Ridge Analysis 25 Years Later," *The American Statistician*, vol. 39, no. 3, pp. 186–192, 1985.

[10] H. Hjalmarsson and J. Schoukens, "On Direct Identification of Physical Parameters in Non-Linear Models," in *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS)*, Stuttgart, Germany, 2004.

[11] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," *accepted for SIAM Review*. [Online]. Available: http://faculty.washington.edu/mfazel/lowrank.html

[12] B. Recht, W. Xu, and B. Hassabi, "Null Space Conditions and Thresholds for Rank Minimization," Tech. Rep., 2009. [Online]. Available: http://www.ist.caltech.edu/~brecht/papers/08.RecXuHas.Thresholds.pdf

[13] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Sparse and Low-Rank Matrix Decompositions System Identification," in *15th IFAC Symposium on System Identification*, Saint-Malo, France, 2009.

[14] Z. Liu and L. Vandenberghe, "Interior-Point Method for Nuclear Norm Approximation with Application to System Identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, January 2009.

[15] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," Tech. Rep., 2009. [Online]. Available: http://www.optimization-online.org/DB_HTML/2009/03/2268.html

[16] D. Goldfarb and S. Ma, "Convergence of fixed point continuation algorithms for matrix rank minimization," Tech. Rep., June 2009. [Online]. Available: http://arxiv.org/abs/0906.3499

[17] J. Dahl and L. Vandenberghe, "Python Software for Convex Optimization (CVXOPT)." [Online]. Available: http://abel.ee.ucla.edu/cvxopt