Non-linear System Identification Via Direct Weight Optimization

Jacob Roll* Alexander Nazin† Lennart Ljung*

January 27, 2004

Abstract

A general framework for estimating nonlinear functions and systems is described and analyzed in this paper. Identification of a system is seen as estimation of a predictor function. The considered predictor function estimate at a particular point is defined to be affine in the observed outputs, and the estimate is defined by the weights in this expression. For each given point, the maximal mean square error of the function estimate over a class of possible true functions is minimized with respect to the weights, which is a convex optimization problem. This gives different types of algorithms depending on the chosen function class. It is shown how the classical linear least squares is obtained as a special case and how unknown-but-bounded disturbances can be handled.

Most of the paper deals with the method applied to locally smooth predictor functions. It is shown how this leads to local estimators with a finite bandwidth, meaning that only observations in a neighborhood of the target point will be used in the estimate. The size of this neighborhood (the bandwidth) is automatically computed and reflects the noise level in the data and the smoothness priors.

The approach is applied to a number of dynamical systems to illustrate its potential.

1 Introduction

Identification of non-linear systems is a very broad and diverse field. Very many approaches have been suggested, attempted and tested. See among many references, e.g. Sjöberg et al. (1995), Harris et al. (2002), Suykens et al. (2002), Roll et al. (2004), Vidyasagar (1997).

In this paper we suggest a new perspective on non-linear system identification, which we call *Direct Weight Optimization, DWO*. It is based on postulating an estimator that is linear in the observed outputs and then determining the weights in this estimator by direct optimization of a suitably chosen (min-max) criterion.

*Div. of Automatic Control, Linköping University, SE-58183 Linköping, Sweden, e-mail: roll, ljung@isy.liu.se

†Institute of Control Sciences, Profsoyuznaya str., 65, 117997 Moscow, Russia, e-mail: nazine@ipu.rssi.ru

One may ask if it is meaningful add one more approach to the already rich flora of methods. However, our suggested approach has some interesting features:

- We will obtain estimates from linear regression models as a special case
- We will obtain a framework for dealing with a class of "realistic" noise descriptions, including so called unknown-but-bounded noises
- We will obtain classical local kernel methods as a special case, equipped with a technique to determine the optimal finite so called *bandwidth* of such methods.

The direct weight optimization approach considered in this paper is an extension of what has previously been presented in (Roll et al., 2002, 2003). See also (Roll, 2003) for a more detailed presentation.

The paper is organized as follows: In the next section predictor models will be defined, which shows the relationship between identification of dynamic systems and (predictor) function estimation. In Section 3 the DWO approach to function estimation is described, and in Section 4 it is shown how function classes defined by basis function expansions can be dealt with, also in the case when unknown-but-bounded disturbances may affect the outputs.

The main algorithms are derived in Section 5 where locally smooth predictor functions are studied. The basic properties of the resulting algorithms are studied in Section 6 and some numerical examples are given in Section 7.

2 Predictor Models

We denote by y and u the output and the input of the system, and we shall assume that the input-output data are sampled with a unit sampling interval. There are many ways to describe a nonlinear system: Input-output form, state-space equations, or predictor forms. We shall here use the predictor (or innovations) form. That means that the output at time t, $y(t)$ is written as

$$y(t) = f_0(Z^{t-1}) + e(t) \tag{1}$$

where

$$Z^{t-1} = \{y(1), u(1), y(2), u(2), \dots, y(t-1), u(t-1)\} \tag{2}$$

In the notation we here assume that the system is single-input-single-output. It immediate to extend to several inputs. For the multi-output case, one would consider the predictor functions for each of the outputs separately, at the same time as allowing Z^{t-1} to contain all past inputs and outputs.

It is a common special case that the predictor function f_0 depends on past data only via a finite and fixed dimensional vector $\varphi(t)$:

$$\varphi(t) = h(Z^{t-1}) \tag{3a}$$
$$y(t) = f_0(\varphi(t)) + e(t) \tag{3b}$$

This vector will be called the *regression vector*. The identification problem is then to determine the two functions h and f_0 from observed data. Often the function h is postulated to be of a simple form, e.g.

$$\varphi(t) = \big[u(t-1), \dots, u(t-nb)\big] \quad \text{(NLFIR) or} \tag{4a}$$

$$\varphi(t) = \big[y(t-1), \dots, y(t-na), u(t-1) \dots, u(t-nb)\big] \quad \text{(NLARX)} \tag{4b}$$

3 The Problem Formulation

We shall consider the situation that the regression vector representation has been selected and the predictor function at a particular argument φ^* is estimated a linear (affine) combination of observed outputs:

$$\hat{f}(\varphi^*) = w_0 + \sum_{t=1}^{N} w_t y(t) \tag{5}$$

The coefficients will in general depend on the function argument:

$$w_t = w_t(\varphi^*) \tag{6}$$

The problem we will discuss in this paper is *how to select the weights w_t in this expression*. This approach we call *Direct Weight Optimization, DWO.*

3.1 Is it restrictive to consider only estimates linear in y?

It may seem restrictive to postulate an estimate that is linear in the observed data. (In fact as long as we do not impose any conditions on the w_t this is no restriction, but we will later assume that the w_t are essentially independent of y^N.) This means that certain non-linear estimators will be ruled out. However. there are two main arguments that this limitation is not so severe:

- For function estimation, it is known from general results that the theoretical lower accuracy bounds for linear estimators are not significantly larger than the overall theoretical (Cramér-Rao) lower bound, see e.g. Fan and Gijbels (1996).
- Quite often, a linear regression model structure for (3) is postulated:

$$\hat{y}(t|\theta) = f(\varphi(t), \theta) = \sum_{k=1}^{d} \theta_k f_k(\varphi(t)) \tag{7}$$

  (This is the case, e.g. for wavelet expansions, for the neuro-fuzzy models treated e.g. in Harris et al. (2002), etc.)

Estimating the parameter θ in (7) by linear least squares gives an expression

$$\hat{\theta}_N = \left(\sum_{k=1}^{N} F(\varphi(k)) F^T(\varphi(k)) \right)^{-1} \sum_{t=1}^{N} F(\varphi(t)) y(t) \tag{8}$$

where

$$F(\varphi) = \begin{bmatrix} f_1(\varphi) \\ \vdots \\ f_d(\varphi) \end{bmatrix} \tag{9}$$

This parameter estimate inserted into the function value at φ^* gives

$$\hat{f}_N(\varphi^*) = f(\varphi^*, \hat{\theta}_N) = F^T(\varphi^*)\hat{\theta}_N = \sum_{t=1}^{N} w_t y(t) \tag{10}$$

where

$$w_t = w_t(Z^N, \varphi^*) = F^T(\varphi^*) \left(\sum_{k=1}^{N} F(\varphi(k)) F^T(\varphi(k)) \right)^{-1} F(\varphi(t)) \tag{11}$$

We see that this is an expression linear in $y(t)$ just as in (5). This indicates that confining ourselves to this type of estimator is not so restrictive.

3.2 How to formulate criteria for choice of weights?

So, let us focus on the estimator structure (5). We can evaluate the quality of the estimator by forming the error at regressor φ^*:

$$\eta(\varphi^*) = f_0(\varphi^*) - \hat{f}(\varphi^*)$$

This error depends on the regression point, φ^*, the true predictor function f_0, the weights w_t, and the random observations $y(t), t = 1, \ldots, N$. We can relieve the dependence on the random data by taking the expectation of the square of η as the quality measure:

$$W(\varphi^*, f_0, w^N) = E\eta^2(\varphi^*) \tag{12}$$

to form the Mean Square Error (MSE) of the estimate.

Remark 1. A technical point: When forming this expectation, we will normally treat the weights w as deterministic constants. Depending on the choice of regressors, it may be that as the optimal weights are determined later on, they may turn out to depend also on the output sequence $y(t)$. In the calculations we will simply disregard this complication and treat the derived algorithms as if the weights were deterministic. One may also note that as the model is applied to a fresh data set (validation data) this complication will be less pronounced.

It would thus be desirable to select the weights w_t to minimize W. Clearly these best weights would depend on the true — unknown — predictor function f_0. Although this predictor function is unknown, we could assume that we know it to belong to a certain class of functions:

$$f_0 \in \mathcal{F} \tag{13}$$

We shall discuss such classes later. A reasonable estimator would be to select the weights so that the maximum of $W(\varphi^*, f_0, w^N)$ over $f_0 \in \mathcal{F}$ is minimized wrt w^N:

$$w^{\text{opt}} = \arg\min_{w^N} \sup_{f_0 \in \mathcal{F}} W(\varphi^*, f_0, w^N) \tag{14}$$

This is the criterion we will adopt.

3.3 Convexity

Note that $\eta(\varphi^*)$ is linear in w^N, which means that η^2 and its expected value $W(\varphi^*, f_0, w^N)$ is quadratic in w^N for any fixed φ^* and f_0. In particular it is then convex in w^N. Since the maximum over a set of convex functions is also convex, it means that

$$\tilde{W}(\varphi^*, \mathcal{F}, w^N) = \sup_{f_0 \in \mathcal{F}} W(\varphi^*, f_0, w^N) \tag{15a}$$

is convex and that the problem

$$w^{\text{opt}} = \arg\min_{w^N} \tilde{W}(\varphi^*, \mathcal{F}, w^N) \tag{15b}$$

is a *convex optimization problem*. This allows for potentially efficient algorithms, and in particular, there will be no local minima that are not global.

3.4 Model on demand

What will the optimal weights depend on? We see from (14) that they will depend on

1. The function class $\mathcal{F}$. We shall discuss different such classes shortly.
2. The regression point φ^* ("the target value").

The latter fact means that the determination of optimal weights will depend on the target value, and that the estimation procedure must be repeated for each new such value of interest. The term *Model on Demand* has been used for this approach, Stenman (1999), Braun et al. (2001), and also *Just in Time*-models, Cybenko (1996), since the model is constructed and delivered (using a data base of observed data) only when needed at a certain point φ^*. In the artificial intelligence community, the approach is known under the name *Lazy Learning*, Atkeson et al. (1997).

One should realize the fact that the model is computed "on demand" means that the estimation data Z^N is never condensed into a model. The estimation data must be kept along and is used every time the predictor function $\hat{f}$ is evaluated at some point. This may seem to defy the idea of a model as a compact summary of observed data, but it should be stressed that with today's cheap memory and very fast retrieval from large data bases, this does not pose any practical problem. It is true that there will be no analytical expression for $\hat{f}$, but just an algorithm to compute this function for any chosen argument. However, other non-linear black box models, like neural networks or trees are also essentially only mechanisms for function value computation, due to their complex internal structure.

4 Examples of Some Function Classes

Let us discuss the minimization of (15) for some different function classes $\mathcal{F}$.

4.1 $\mathcal{F}$ is a linear hull of basis functions

Consider the case that the function class consists of functions that are obtained as linear combinations of a finite number of basis functions $f_k(\varphi)$:

$$\begin{aligned} \mathcal{F} &= \{f | f(\varphi) = \sum_{k=1}^{d} \theta_k f_k(\varphi) = F(\varphi)^T \theta \text{ for some } \theta\} \\ F(\varphi)^T &= \begin{bmatrix} f_1(\varphi) & \dots & f_d(\varphi) \end{bmatrix} \end{aligned} \tag{16}$$

In this case we can easily show the following proposition:

Proposition 1. *Consider the problem (15) for the function class (16) The minimizing weights w^N are then given by* (11)*.*

Remark 2. Note that this is the same solution as obtained by estimating θ by linear least squares and evaluating the resulting model in φ^*.

Proof. Let θ_0 be the (unknown) parameters of f_0 in the set (16). The MSE (12) can be written

$$\begin{aligned} W(\varphi^*, f_0, w^N) &= E[\left(\sum_{t=1}^{N} w_t y(t) - f_0(\varphi^*)\right)^2] \\ &= E[\left(\sum_{t=1}^{N} w_t (f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2] \\ &= \left(\sum_{t=1}^{N} w_t F(\varphi(t))^T \theta_0 - F(\varphi^*)^T \theta_0\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ &= \left(\left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*)\right)^T \theta_0\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \end{aligned} \tag{17}$$

In the last expression, we can see that the bias term may be arbitrarily large, unless we choose our weights such that

$$\sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \tag{18}$$

Under this requirement, the bias term completely disappears from (17). To find the solution of (15), we hence need to solve the optimization problem

$$\begin{aligned} \min_{w} \quad & \sum_{t=1}^{N} w_t^2 \\ \text{subj. to} \quad & \sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \end{aligned} \tag{19}$$

But this is nothing less than finding the least-norm solution to (18), which is given exactly by (11), and the proposition is proved. □

So in the case of the function class (16) the DWO approach does not give any new method, but just the classical least squares. This is in a sense reassuring, indicating that the problem formulation (15) seems to be reasonable.

4.2 A Realistic Noise Model

The real advantage of considering (15), however, is that we can use less knowledge about the true function f_0.

In some contexts the simple description of the term e in (1) as white noise or even random variables is rejected. Indeed, there could be many reasons why this is not a realistic description. Other models that have been used is the so-called *unknown-but-bounded* assumption, where all that is assumed known is that $|e(t)| \leq C_e, \forall t$, with C_e being a constant. See, among many references, e.g.Schweppe (1968), Milanese and Belforte (1982), Deller (1989). This description may lead to conservative estimates, since one must be prepared for "malicious" disturbances. A quite realistic and attractive noise description is to assume that e has a stochastic (white noise) component $e_s(t)$ and an unknown-but-bounded component $e_u(t)$:

$$y(t) = f_0(Z^{t-1}) + e_u(t) + e_s(t), \quad |e_u(t)| \leq C_e \tag{20}$$

In this case it is easy to define function classes $\mathcal{F}$ that include the component e_u. For example, for the function class (16) for f_0 we would have the version

$$y(t) = f_0(\varphi(t)) + e_u(t) + e_s(t), \quad |e_u(t)| \leq C_e \tag{21a}$$

$$\Rightarrow y(t) = \tilde{f}_0(\varphi(t)) + e_s(t) \tag{21b}$$

$$\tilde{f}_0 \in \mathcal{F} = \{f| \, |f - F(\varphi)^T\theta| \leq C_e \text{ for some } \theta\} \tag{21c}$$

DWO solutions for this function class are investigated in Nazin et al. (2003).

5 Functions with Local Smoothness

We now turn to the main topic of this paper.

It may seem a very specific type of prior knowledge about the system to assume that it belongs to a specified family like (16). It could be more natural to have some idea about the local smoothness of the predictor function. In this and the following sections we shall work with such classes of $\mathcal{F}$. As it may be expected, these classes lead to local estimation methods, that is the function estimate (5) depends primarily on the observations close to the target regressor φ^*.

So to repeat the framework, assume that we are given data $\{\varphi(t), y(t)\}_{t=1}^N$ from a system described by

$$y(t) = f_0(\varphi(t)) + e(t) \tag{22}$$

where f_0 is an unknown function, $f_0 : \mathbb{R}^n \to \mathbb{R}$, and $e(t)$ are zero-mean, i.i.d. random variables with known variance σ^2, and where $e(t)$ and $\varphi(\tau)$ are independent for all t, τ. Also assume that f_0 belongs to the function class

$$\mathcal{F}_2(Q) = \{f \in \mathcal{C}^1 \mid \|\nabla f(\varphi + h) - \nabla f(\varphi)\|_{Q^{-1}} \leq \|h\|_Q \ \forall \varphi, h \in \mathbb{R}^n\} \tag{23}$$
$$(\|h\|_Q = h^T Q h)$$

where ∇ denotes gradient and Q is a symmetric, positive definite matrix. Intuitively, the inequality can be interpreted as an upper bound on the Hessian

of f. A special case is given by $Q = LI$, where L is a scalar and I is the identity matrix. In this case, (23) becomes a standard Lipschitz condition on the gradient, with L as the Lipschitz constant.

Now, the problem to solve is to find an estimator (5) to estimate $f_0(\varphi^*)$ in a certain point φ^*, such that the worst-case MSE (15) is minimized. However, in general, the worst-case MSE is very difficult to compute. Instead, we will give an upper bound on the worst-case MSE, which will be minimized with respect to the weights w_t of the estimator.

5.1 Minimizing an upper bound on the worst-case MSE

For convenience, let us introduce the notation $\widetilde{\varphi}(t) = \varphi(t) - \varphi^*$. Under the above assumptions, the MSE for an affine estimator (5) can be written

$$\begin{aligned}
W(\varphi^*, f_0, w^N) &= E\left(w_0 + \sum_{t=1}^{N} w_t y(t) - f_0(\varphi^*)\right)^2 \\
&= E\left(w_0 + \sum_{t=1}^{N} w_t (f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2 \\
&= \left(w_0 + \sum_{t=1}^{N} w_t f_0(\varphi(t)) - f_0(\varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \qquad (24) \\
&= \Bigg(w_0 + \sum_{t=1}^{N} w_t \left(f_0(\varphi(t)) - f_0(\varphi^*) - \nabla^T f_0(\varphi^*)\widetilde{\varphi}(t)\right) \\
&\quad + f_0(\varphi^*)\left(\sum_{t=1}^{N} w_t - 1\right) + \nabla^T f_0(\varphi^*) \sum_{t=1}^{N} w_t \widetilde{\varphi}(t)\Bigg)^2 \\
&\quad + \sigma^2 \sum_{t=1}^{N} w_t^2
\end{aligned}$$

where the first squared term of the last expression is the squared bias, and the last term is the variance of the estimate.

Since there are no bounds on $f_0(\varphi^*)$ and $\nabla^T f_0(\varphi^*)$ in $\mathcal{F}_2(Q)$, it is easy to see that the bias term of the MSE (24) can get arbitrarily large unless we impose the following constraints on the weights:

$$\sum_{t=1}^{N} w_t = 1 \qquad (25a)$$

$$\sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \qquad (25b)$$

In other words, for the worst-case MSE to be finite, (25) has to hold. Moreover, as we will see soon, a natural choice of w_0 should be zero, i.e., we get a linear estimator. With $w_0 = 0$ and under the restrictions (25), any linear function is estimated with zero bias.

Under the restrictions (25), we get the following upper bound on the MSE:

$$W(\varphi^*, f_0, w^N) \leq \left(\frac{1}{2} \sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 + |w_0| \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{26}$$

This upper bound can now be minimized with respect to the weights w_t. As already hinted, the minimization with respect to w_0 is obtained by choosing $w_0 = 0$. Hence, the optimization problem to solve is the following:

$$\begin{aligned} \min_{w} \quad & \frac{1}{4} \left(\sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ \text{subj. to} \quad & \sum_{t=1}^{N} w_t = 1 \\ & \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{27}$$

By using slack variables, this problem can easily be formulated as a convex *quadratic program (QP)*

$$\begin{aligned} \min_{w,s} \quad & \frac{1}{4} \left(\sum_{t=1}^{N} s_t \|\widetilde{\varphi}(t)\|_Q^2 \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq w_t \\ & s_t \geq -w_t \\ & \sum_{t=1}^{N} w_t = 1 \\ & \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{28}$$

and can be solved efficiently to get the optimal w_t.

5.2 Using knowledge about the function and gradient values

Sometimes we might know some bounds on the function value and/or its gradient in φ^*. To incorporate this information, let us consider the function class

$$\mathcal{F}_2(Q, \delta, \Delta, R) = \{ f \in \mathcal{F}_2(Q) \,|\, |f(\varphi^*) - a| \leq \delta,\ \|\nabla f(\varphi^*) - b\|_{R^{-1}} \leq \Delta \} \tag{29}$$

where R is a positive definite matrix, $a, \delta, \Delta \in \mathbb{R}$, and $b \in \mathbb{R}^n$.

Assuming that $f_0 \in \mathcal{F}_2(Q, \delta, \Delta, R)$, we get the following upper bound on the

MSE:

$$W(\varphi^*, f_0, w^N) \leq \left(\frac{1}{2} \sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 \right. \\ + \left| w_0 + a \left(\sum_{t=1}^{N} w_t - 1 \right) + b^T \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right| \\ \left. + \delta \left| \sum_{t=1}^{N} w_t - 1 \right| + \Delta \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{30}$$

This upper bound can for any given $w = \{w_1, \ldots, w_N\}$ be minimized with respect to w_0, giving

$$w_0 = -a \left(\sum_{t=1}^{N} w_t - 1 \right) - b^T \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \tag{31}$$

By inserting this into (30), the upper bound on the MSE is reduced to

$$W(\varphi^*, f_0, w^N) \leq \left(\frac{1}{2} \sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 \right. \\ \left. + \delta \left| \sum_{t=1}^{N} w_t - 1 \right| + \Delta \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{32}$$

We can now minimize (32) with respect to the weights w_t. Simple but tedious reformulations show that the optimization problem to solve is equivalent to a *second order cone program (SOCP)*

$$\begin{aligned} \min_{w,s,r} \quad & r_c \\ \text{subj. to} \quad & \left\| \begin{bmatrix} 2\left(\delta r_a + \Delta r_b + \frac{1}{2}\sum_{t=1}^{N} \|\widetilde{\varphi}(t)\|_Q^2 s_t\right) \\ 2\sigma s \\ 1 - r_c \end{bmatrix} \right\| \leq 1 + r_c \\ & \left| \sum_{t=1}^{N} w_t - 1 \right| \leq r_a \\ & \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \leq r_b \\ & |w_t| \leq s_t, \qquad t = 1, \ldots, N \end{aligned} \tag{33}$$

This is a standard convex optimization problem (see, e.g., Boyd and Vandenberghe (2004)) and can be solved efficiently.

6 Properties of the Solutions

6.1 Finite Bandwidth

Since only local smoothness of the predictor function is assumed, no inference about function values can be made from data far away from the target point. It

is therefore to be expected that the weights w_t will decrease with the distance $|\varphi(t) - \varphi^*|$. An interesting property of the DWO approach is that in many cases, most of the weights will not only decrease but become exactly zero. This can be thought of as an automatic finite bandwidth, i.e., the estimates will automatically become local: The estimate of f at φ^* will only depend on those observations $y(t), \varphi(t)$ that are in the vicinity of φ^*, $|\varphi(t) - \varphi^*| < h$, where h would be the bandwidth. This is a typical feature of so called *kernel methods* for function estimation, see e.g. Härdle (1990). In those cases the bandwidth is typically chosen *ad hoc* or using asymptotic (in N) arguments. In our case, as we shall see, the bandwidth is automatically determined and minimizes the worst case MSE *for any finite data record* N.

In particular, for the problem (28), we can show the following theorem (see also Sacks and Ylvisaker (1978) for a similar theorem in a slightly different setting).

Theorem 1. *Suppose that the problem* (28) *is feasible, and that* $\sigma > 0$. *Then there exist* $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}^n$, *and* $\mu_3 \in \mathbb{R}$, $\mu_3 \geq 0$, *such that for an optimal solution* (s^*, w^*), *it holds that*

$$w_t^* = \begin{cases} \mu_1 + \mu_2^T \widetilde{\varphi}(t) - \mu_3 \|\widetilde{\varphi}(t)\|_Q^2, & \mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \leq \mu_1 + \mu_2^T \widetilde{\varphi}(t) \\ 0, & |\mu_1 + \mu_2^T \widetilde{\varphi}(t)| \leq \mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \\ \mu_1 + \mu_2^T \widetilde{\varphi}(t) + \mu_3 \|\widetilde{\varphi}(t)\|_Q^2, & \mu_1 + \mu_2^T \widetilde{\varphi}(t) \leq -\mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \end{cases} \tag{34}$$

Remark 3. In words, some of the weights will lie along at most two paraboloid segments, one positive and one negative, and the rest will be zero. The expression (34) is illustrated for the univariate case in Figure 1.

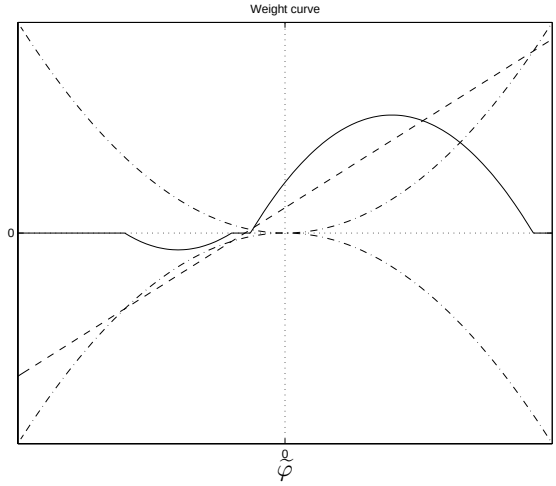

Figure 1: Principal shape of the weight curve (solid curve). The dash-dotted parabolas are $\pm\mu_3\widetilde{\varphi}^2$, and the dashed line is $\mu_1 + \mu_2\widetilde{\varphi}$. (The weight curve is scaled by a factor 4 to make the figure more clear.)

Proof. The proof uses the *Karush-Kuhn-Tucker (KKT) conditions* (see, e.g., (Nocedal and Wright, 1999)). Since the QP (28) is a convex optimization problem with linear constraints, the KKT conditions are necessary and sufficient conditions for optimality of a solution (see, e.g., (Boyd and Vandenberghe, 2004) for details).

The Lagrangian function of (28) can be written

$$\begin{aligned}\mathcal{L}(w,s;\mu,\lambda) = \frac{1}{4}\left(\sum_{t=1}^{N} s_t \|\widetilde{\varphi}(t)\|_Q^2\right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 - 2\sigma^2\mu_1\left(\sum_{t=1}^{N} w_t - 1\right) \\ - 2\sigma^2\mu_2 \sum_{t=1}^{N} w_t\widetilde{\varphi}(t) - 2\sigma^2 \sum_{t=1}^{N}(\lambda_t^+(s_t - w_t) + \lambda_t^-(s_t + w_t))\end{aligned} \tag{35}$$

where $\lambda_t^{\pm} \geq 0$, $t = 1, \ldots, N$, and μ are the Lagrangian multipliers, scaled by a factor $1/2\sigma^2$. Since $s_t^* = |w_t^*|$ (trivially) for an optimal solution (w^*, s^*), the KKT conditions are equivalent to the following relations:

$$\mu_1 + \mu_2\widetilde{\varphi}(t) = \lambda_t^+ - \lambda_t^- \tag{36a}$$

$$\frac{1}{4\sigma^2}\left(\sum_{k=1}^{N} |w_k^*| \|\widetilde{\varphi}(k)\|_Q^2\right)\|\widetilde{\varphi}(t)\|_Q^2 + |w_t^*| = \lambda_t^+ + \lambda_t^- \tag{36b}$$

$$\sum_{t=1}^{N} w_t^* = 1 \tag{36c}$$

$$\sum_{t=1}^{N} w_t^*\widetilde{\varphi}(t) = 0 \tag{36d}$$

$$s_t^* = |w_t^*| \tag{36e}$$

$$\lambda_t^+(|w_t^*| - w_t^*) = 0 \tag{36f}$$

$$\lambda_t^-(|w_t^*| + w_t^*) = 0 \tag{36g}$$

$$\lambda_t^{\pm} \geq 0, \quad t = 1, \ldots, N \tag{36h}$$

Let

$$\mu_3 = \frac{1}{4\sigma^2}\left(\sum_{k=1}^{N} |w_k^*| \|\widetilde{\varphi}(k)\|_Q^2\right) \tag{37}$$

From (36f) and (36g), we can see that $w_t^* > 0$ implies $\lambda_t^- = 0$, and that $w_t^* < 0$ implies $\lambda_t^+ = 0$. Hence, we can eliminate $\lambda_t^{\pm}$ from the KKT conditions in these cases, getting

$$w_t^* = \mu_1 + \mu_2\widetilde{\varphi}(t) - \operatorname{sgn}(w_t^*)\mu_3\|\widetilde{\varphi}(t)\|_Q^2, \quad w_t^* \neq 0 \tag{38}$$

We can see that

$$\begin{aligned} w_t^* > 0 \quad &\Rightarrow \quad \mu_1 + \mu_2\widetilde{\varphi}(t) > \mu_3\|\widetilde{\varphi}(t)\|_Q^2 \\ w_t^* < 0 \quad &\Rightarrow \quad \mu_1 + \mu_2\widetilde{\varphi}(t) < -\mu_3\|\widetilde{\varphi}(t)\|_Q^2 \end{aligned}$$

Finally, if $w_t^* = 0$, we get from (36a), (36b), and (36h) that

$$\begin{aligned} 2\lambda_t^+ &= \mu_1 + \mu_2\widetilde{\varphi}(t) + \mu_3\|\widetilde{\varphi}(t)\|_Q^2 \geq 0 \\ 2\lambda_t^- &= -\mu_1 - \mu_2\widetilde{\varphi}(t) + \mu_3\|\widetilde{\varphi}(t)\|_Q^2 \geq 0 \end{aligned}$$

which implies

$$-\mu_3\|\widetilde{\varphi}(t)\|_Q^2 \leq \mu_1 + \mu_2\widetilde{\varphi}(t) \leq \mu_3\|\widetilde{\varphi}(t)\|_Q^2$$

From these expressions, (34) is readily obtained. □

One advantage with the described property is that, instead of having to explicitly prescribe a bandwidth for the estimator, we can give the noise variance σ^2 and the upper bound Q on the Hessian, which can also be thought of as giving an upper bound for the approximation error we would make by locally approximating the system by a linear model. This might in many cases be a more natural choice of design parameters.

Theorem 1 also opens up for a possible reduction of the computational complexity: Since many of the weights w_t will be zero, we can already beforehand exclude data that will most likely correspond to zero weights, thus making the QP (28) considerably smaller. Having solved (28), one can easily check whether or not the excluded weights really should be zero, by checking if the excluded data points satisfy $|\mu_1 + \mu_2\widetilde{\varphi}(t)| \leq \mu_3\|\widetilde{\varphi}(t)\|_Q^2$ (the middle case of (34)).

Another appealing property is that the weights automatically adapt to how the actual data samples are spread, and can easily handle sparse data sets or data lying asymmetrically. This should be particularly desirable when the dimension of the regression vectors is high.

6.2 Asymptotic Behavior

In (Legostaeva and Shiryaev, 1971), it was shown that using the Epanechnikov kernel would yield an asymptotically optimal (continuous) kernel estimator with respect to the worst-case MSE if the upper bound (26) was tight. Therefore, one would expect that the weights w_k of the DWO approach would asymptotically converge to the weights using the Epanechnikov kernel with an asymptotically optimal bandwidth (see Fan and Gijbels, 1996). In the following theorem, we show this for a special univariate case.

Theorem 2. *Consider the problem of estimating an unknown function $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{F}_2(L)$, where $L > 0$ is the Lipschitz constant, at a given internal point $\varphi_0 \in (-1/2, 1/2)$ under an* equally spaced fixed design model

$$\varphi(k) = \frac{k-1}{N-1} - \frac{1}{2}, \quad k = 1, \ldots, N \tag{39}$$

and with $\sigma > 0$. Let w^* be the minimizer of (27)*. Then asymptotically, as $N \to \infty$,*

$$w_k^* \approx \frac{3}{4} C_N \max\{1 - \left(\frac{\widetilde{\varphi}(k)}{h_N}\right)^2, 0\}, \quad k = 1, \ldots, N \tag{40}$$

where

$$C_N \asymp \frac{1}{Nh_N}, \quad h_N \asymp \left(\frac{15\sigma^2}{L^2 N}\right)^{1/5} \quad \text{as } N \to \infty \tag{41}$$

Hence, the optimal weights (40) *approximately coincide with related asymptotically optimal weights and bandwidth of the local polynomial estimator for the worst-case function in $\mathcal{F}_2(L)$, as given in (Fan and Gijbels, 1996).*

Here $a_N \asymp b_N$ means asymptotic equivalence of two real sequences (a_N) and (b_N), that is $a_N/b_N \to 1$ as $N \to \infty$.

Remark 4. When the data are lying symmetrically around φ_0, e.g., when $\varphi_0 = 0$, it follows that the relation (40) will hold exactly also for finite N, i.e.,

$$w_k^* = \frac{3}{4} C_N \max\{1 - \left(\frac{\widetilde{\varphi}(k)}{h_N}\right)^2, 0\}, \quad k = 1, \ldots, N \tag{42}$$

where C_N and h_N are given by (41) (see Roll, 2003, for details).

Proof. For this proof, a special version of Theorem 1 is needed (see Roll, 2003), from which it follows that there are three numbers $\mu_1 > 0$, μ_2, and $\mu_3 > 0$, such that

$$w_k^* = \max\{\mu_1 + \mu_2 \widetilde{\varphi}(k) - \mu_3 \widetilde{\varphi}^2(k), 0\}, \quad k = 1, \ldots, N \tag{43}$$

if and only if $\mu_1 + \mu_2 \widetilde{\varphi}(k) + \mu_3 \widetilde{\varphi}^2(k) \geq 0$ for all $k = 1, \ldots, N$, which is the case if

$$\mu_2^2 \leq 4 \mu_3 \mu_1 \tag{44}$$

Also recall that the KKT conditions (36) applied in the proof of Theorem 1 represent necessary and sufficient conditions for optimality of the solution to the considered QP problem. Thus, in order to prove the first part of the theorem, it suffices to demonstrate that

$$\lim_{N \to \infty} \frac{\mu_2^2}{\mu_3 \mu_1} = 0 \tag{45}$$

for the three parameters μ_1, μ_2, and μ_3 satisfying (36c), (36d), and (37), with the weights w_k^* given by (43). Denote the support of the function $w(\widetilde{\varphi}) = \max\{\mu_1 + \mu_2 \widetilde{\varphi} - \mu_3 \widetilde{\varphi}^2, 0\}$ by $[a, b]$, that is

$$\mu_1 + \mu_2 a - \mu_3 a^2 = 0, \quad \mu_1 + \mu_2 b - \mu_3 b^2 = 0, \quad a < b \tag{46}$$

and suppose that $[a, b] \in [-0.5 - \varphi_0, 0.5 - \varphi_0]$. If we find a solution to the system of the three equations (36c), (36d), and (37) with respect to $\mu_1 > 0$, μ_2, and $\mu_3 > 0$, and (44) is satisfied, then we have proved (43). The following asymptotic relation for nonnegative weights (43) holds true as $N \to \infty$:

$$\frac{1}{N} \sum_{k=1}^{N} w_k \widetilde{\varphi}^m(k) = \int_a^b \left(\mu_1 + \mu_2 \widetilde{\varphi} - \mu_3 \widetilde{\varphi}^2\right) \widetilde{\varphi}^m d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{47}$$

for any $m = 0, 1, 2$, where

$$h = \frac{b - a}{2}$$

Thus, the equations (36c), (36d), and (37) may be written as follows:

$$\frac{1}{N} = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right) d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{48}$$

$$0 = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right) \widetilde{\varphi} d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{49}$$

$$\frac{4\sigma^2}{L^2}\frac{\mu_3}{N} = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right) \widetilde{\varphi}^2 d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{50}$$

with

$$a = \frac{\mu_2 - \sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3}, \quad b = \frac{\mu_2 + \sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3}, \quad h = \frac{\sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3} \tag{51}$$

Note that the terms $O(h/N)$ in (48)–(50) do not depend on (μ_1, μ_2, μ_3). Consequently, $O(h/N)|\mu_2|$ is uniformly bounded over μ_2 as $N \to \infty$.

Now, one might verify by direct substitution (see Roll, 2003, for a detailed proof) that the solution to (48)–(50) has the following asymptotics:

$$\mu_1 \asymp \frac{3}{4Nh_N}, \quad \mu_2 = O(N^{-1}), \quad \mu_3 \asymp \frac{\mu_1}{h_N^2} \tag{52}$$

with

$$h = \frac{\sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3} \asymp h_N \asymp \left(\frac{15\sigma^2}{L^2 N}\right)^{1/5} \tag{53}$$

Thus, we obtain

$$\lim_{N\to\infty} \frac{\mu_2^2}{\mu_3\mu_1} = \lim_{N\to\infty} \frac{\mu_1}{\mu_3}\left(\frac{\mu_2}{\mu_1}\right)^2 = 0 \tag{54}$$

and relation (45) is proved.

Since $\mu_2 = o(\mu_1)$, the relation (40) follows directly from (42) and 52). This proves the theorem. □

7 Examples of Applications to Dynamical Systems

In this section we shall apply the DWO technique for locally smooth predictors to a number of simulated examples. Generally speaking we shall build the models using a certain estimation data set Z_e^N and then test the model on another validation data set Z_v^M. This means that the "target points" φ^* will be generated in simulations using the validation set as $\varphi_s(t)$ in (56) below. When the optimal weights are determined for these target points, they are however calulated only using the data in Z_e^N. This will make comparisons to other methods more fair.

7.1 A nonlinear ARX (NLARX) system

In this section, the direct weight optimization approach is applied to data from different systems. We begin by considering a model of NLARX type Sjöberg et al. (1995), where Q and σ^2 are known.

When estimating NLARX models, one should realize that our assumptions about $e(t)$ and $\varphi(\tau)$ being independent for all t, τ are violated (as opposed to the NFIR case, where φ only depends on the input u, not on the output y). However, as we will see, the method often works well in practice anyway. See remark 1 in Section 3 and (Roll, 2003) for a discussion about this.

Example 1. Consider the following NLARX system:

$$\begin{aligned} y(t) = &\begin{bmatrix} 0.1 & -0.1 & 0.25 & 0.5 \end{bmatrix} \cdot \varphi(t) \\ &+ \frac{L}{2}\Big(\|\varphi(t)\|^2 - 2\left(\max\{\|\varphi(t)\|^2, 1\} - 1\right) + 2\left(\max\{\|\varphi(t)\|^2, 2\} - 2\right) \\ &- \left(\max\{\|\varphi(t)\|^2, 3\} - 3\right)\Big) + e(t) \end{aligned} \tag{55}$$

where

$$\varphi(t) = \begin{bmatrix} y(t-1) & y(t-2) & u(t-1) & u(t-2) \end{bmatrix}^T$$

$L = 0.1$ and $e(t) \in N(0, 0.01)$, i.e., $\sigma = 0.1$. Note that this system satisfies (23) with $Q = LI = 0.1I$. $N = 300$ data samples were collected for an input $u(t) \in N(0, 1)$.

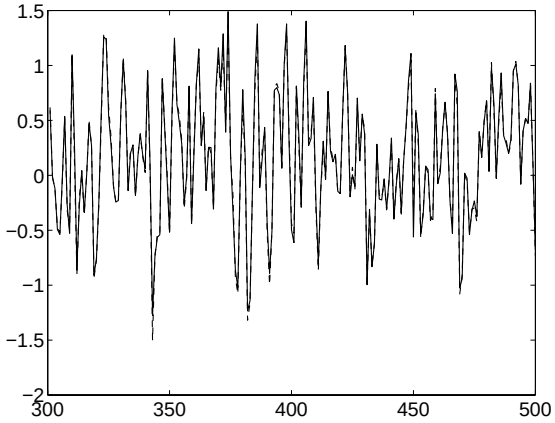

Figure 2: Simulated (solid) and true (dashed) output for system (55), modeled using the DWO approach with $Q = 0.1I$. The fit is 90.8%.

The system was then simulated using the DWO approach for another set of 200 data samples with $u(t) \in N(0, 1)$.

It is worth commenting on how this is done: The predictor function is defined by

$$\begin{aligned} \hat{y}(t|t-1) &= \hat{f}(\varphi(t)) \\ \varphi(t) &= \begin{bmatrix} y(t-1), \ldots, y(t-n_a), u(t-1), \ldots, u(t-n_b) \end{bmatrix} \end{aligned}$$

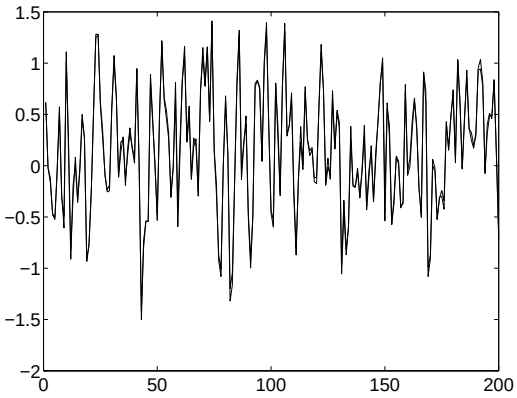

Figure 3: Simulated and true output for system (55), modeled using an artificial neural network. The fit is 89.8%.

A simulation of the model uses only the input, so it is accomplished recursively as

$$y_s(t) = \hat{f}(\varphi_s(t)) \tag{56}$$

$$\varphi_s(t) = \big[y_s(t-1), \dots, y_s(t-n_a), u(t-1), \dots, u(t-n_b)\big] \tag{57}$$

For the DWO approach the estimate of the function f when evaluated at $\varphi_s(t)$ was using data from the estimation set only, and not the validation set. To evaluate the fit between the simulated output y_s and the measured output y we use the percentage

$$[1 - \text{norm}(y - y_s)/\text{norm}(y - \text{mean}(y))] * 100 \tag{58}$$

In Figure 2, the resulting output is compared to the true, noiseless output. As can be seen, the result is very good (90.8% fit). An artificial neural network with 10 sigmoidal units in the hidden layer achieved 89.8% fit (see Figure 3).

7.2 The Narendra-Li system

It can also be interesting to see how the DWO approach can perform when the true system is not of NLARX type, since this is often the case in real applications. The following is an example of this.

Example 2. Let us consider a nonlinear benchmark system proposed by Narendra and Li (1996). The system is defined in state-space form by

$$\begin{aligned} x_1(t+1) &= \left(\frac{x_1(t)}{1+x_1^2(t)} + 1\right)\sin x_2(t) \\ x_2(t+1) &= x_2(t)\cos x_2(t) + x_1(t)e^{-\frac{x_1^2(t)+x_2^2(t)}{8}} \\ &\quad + \frac{u^3(t)}{1+u^2(t)+0.5\cos(x_1(t)+x_2(t))} \\ y(t) &= \frac{x_1(t)}{1+0.5\sin x_2(t)} + \frac{x_2(t)}{1+0.5\sin x_1(t)} + e(t) \end{aligned} \tag{59}$$

The noise term $e(t)$ is added in accordance with (Stenman, 1999) and has a variance of 0.1. The states are assumed not to be measurable, and following the discussion in (Stenman, 1999), a NLARX331 structure is used to model the system, i.e.,

$$\varphi(t) = \begin{bmatrix} y(t-1) & y(t-2) & y(t-3) & u(t-1) & u(t-2) & u(t-3) \end{bmatrix}^T$$

As estimation data, $N = 50000$ samples were generated using a uniformly distributed random input $u(t) \in [-2.5, 2.5]$. To validate the model, the input signal

$$u(t) = \sin\frac{2\pi t}{10} + \sin\frac{2\pi t}{25}, \quad t = 1, \ldots, 200$$

was used. Figure 4 shows the simulated output when Q was chosen to be $0.1I$. The results are reasonable (49.7% fit), and can be compared with the results using a neural network with 20 hidden sigmoidal units, which achieved 47.1% fit (see Figure 5).

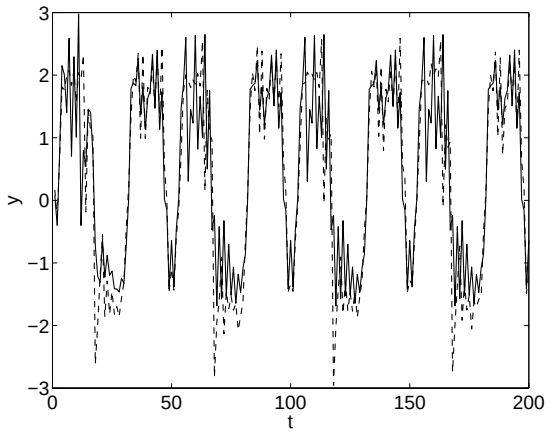

Figure 4: Simulated (solid) and true (dashed) output for system (59), modeled using the DWO approach with $Q = 0.1I$. The fit is 49.7 %.

In the previous example, the matrix Q was not known *a priori*, and was chosen to be constant over the entire state-space. An alternative would be to estimate Q. A (somewhat ad hoc) way of doing this is to estimate the Hessian $H(\varphi^*)$ of f (by locally fitting a cubic model to the data). Then the estimate $\hat{H}(\varphi^*)$ can be factorized according to

$$\hat{H}(\varphi^*) = T(\varphi^*)D(\varphi^*)T^T(\varphi^*) \tag{60}$$

where $T(\varphi^*)$ is orthogonal and $D(\varphi^*) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal. Finally, choose

$$\hat{Q}(\varphi^*) = T(\varphi^*)\bar{D}(\varphi^*)T^T(\varphi^*) \tag{61}$$

where $\bar{D}(\varphi^*) = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_n|)$.

Some adaptive techniques to estimate the Lipschitz constant directly from data (at each target value) are suggested in Juditsky et al. (2004).

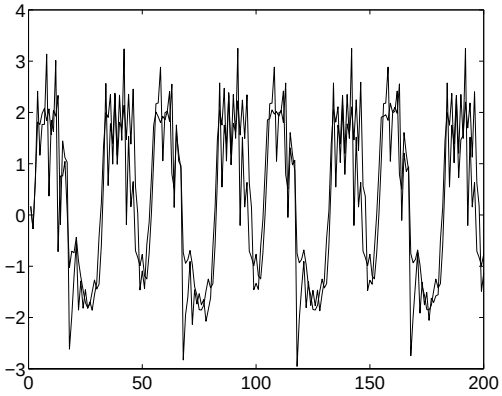

Figure 5: Simulated and true output for system (59), modeled using an artificial neural network. The fit is 47.1 %.

7.3 Cell Dynamics

In the following example, Q is estimated using the procedure described in the previous subsection. Furthermore, σ is treated as unknown and is estimated using the C_p criterion (Cleveland and Devlin, 1988; Mallows, 1973).

Example 3. This example deals with data simulated from equations of the same character as the glucose metabolism in cell dynamics.

A set of 200 data samples has been collected from the system

$$\begin{aligned} \dot{x}_1 &= -\frac{x_1 - x_2}{1 + x_1 + x_2} + \frac{u - x - 1}{1 + u + x_1 + x_1 u} \\ \dot{x}_2 &= \frac{x_1 - x_2}{1 + x_1 + x_2} - \frac{x_2 - 1}{1 + x_2 + 1} \\ y &= x_2 \end{aligned} \tag{62}$$

The given data set (input u and output y) is shown in Figure 6. The data were applied to the DWO estimation procedure with regression vector $\varphi(t)^T = \big[y(t-1), y(t-2), u(t-1), u(t-2)\big]$. The first 100 data were used as estimation data. Then the system was simulated for all 200 data samples, using the DWO approach with Q and σ estimated as described above. (This means that only the data set up to time 40 (= sample 100) was used when the regressors φ^* in the set from 101 to 200 were estimated.) The result can be seen in Figure 7. It can be compared to the result from a sigmoidal neural network with the same regressors and 10 neurons, shown in Figure 8.

The fit is determined as in (58). The DWO approach gave a fit of 72.95 % and the neural network model a fit of 66.41 % in this case.

8 Conclusions

There are two main conclusions from this paper

- The nonlinear identification/estimation problem can be formulated as a direct optimization of a min/max criterion with respect to weights in a

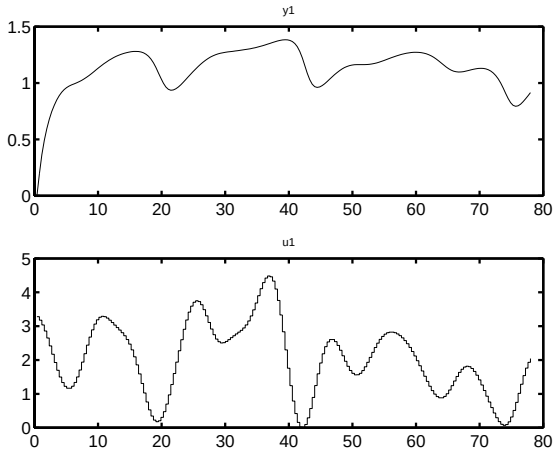

Figure 6: Estimation data from system (62) (Below: input; Above: output).

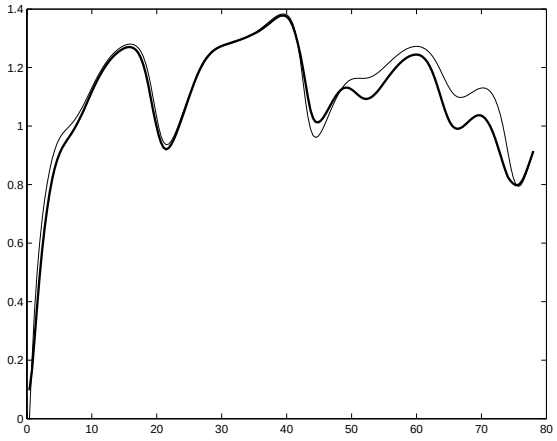

Figure 7: Simulated (thick) and true (thin) output for system (62), modeled using the DWO approach with $Q = 0.1I$. The fit is 72.95 %

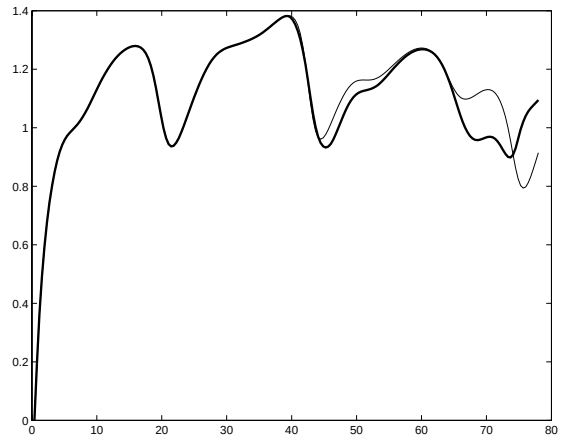

Figure 8: Simulated (thick) and true (thin) output for system (62), modeled using a sigmoidal neural network with 10 neurons. The fit is 66.41 %

linear estimator. This formulation has a potential to serve as quite a general guideline for dealing with problems with various prior information.

- When applied to locally smooth predictor functions, algorithms are obtained that are competitive alternatives to more traditional black-box identification methods, such as artificial neural networks.

In the general formulation we have noted that the DWO-approach (15) is always a convex optimization problem, which gives many useful advantages: Potentially efficient algorithms, Boyd and Vandenberghe (2004), and unique minima. However, the problem is to compute the supremum over $\mathcal{F}$ for fixed w_0 and w. This is often a nontrivial problem, and we might have to resort to upper bounds as in (26) in this paper. In some cases, though, the worst-case MSE is actually computable. This is the case, e.g., when f_0 is univariate. However, experiments show that only a minor improvement of the estimates are obtained by using the corresponding optimal estimator, compared to the standard DWO estimator. See (Roll, 2003) for details. Corresponding theoretical conclusions have been obtained by Leonov (1999).

The potential to treat less exact prior information about the predictor function, such as it "being close" to a linear hull of basis functions, is worth while to consider further. It may give insights and alternative algorithms for *unknown-but-bounded* disturbances (or "set membership" identification methods).

The main part of this paper has however dealt with the DWO algorithm applied to locally smooth predictor functions, (23). The local estimation algorithms obtained in this way have several features in common with classical kernel methods and local polynomial approaches. An interesting feature is that the DWO approach automatically gives the optimal bandwidth for such methods, even for finite data records. Of particular value is that the actual distribution on observations is properly taken care of, be it sparse and/or unevenly spread.

The local smoothness approach depends on prior information of the noise level and of the size of (an upper bound of) the Hessian of the predictor function. We estimated those from data in a fairly *ad hoc* way in Example 3. It would be

of interest to develop efficient and robust methods for this task. See Stenman (1999) and Juditsky et al. (2004) for some ideas to estimate the Hessian and Lipschitz constant, respectively, and e.g. Fan and Gijbels (1996) for estimation on the noise level.

The numerical examples show that the suggested approach could be a viable alternative to more conventional black-box methods. Actually, the fits obtained for the DWO approach were slightly better than for neural networks in all three cases. It should be remarked that the DWO approach gives an exact minimization of the chosen criterion, and is therefore not depending on iterative search and initial parameter estimates that may lead to non-global, local minima. This is a well known hassle with e.g. artificial neural networks. On the other hand, the DWO approach gives "models-on-demand", and the estimation has to be repeated for each given argument φ^*. See the discussion in Section 3.4.

Moreover, our current implementation of the DWO method applied to locally smooth functions is quite slow: it is based on MATLAB code calling a Quadratic Programming solver from CPLEX, Anonymous (2000). As mentioned in Section 6, it is possible to reduce the computational complexity for the calculation of the weights. This should be investigated further. An interesting goal is to push the estimation time to the order of magnitude of evaluating the function value of a complex neural network or to the sampling times of even fast sampled control systems.

References

Anonymous. *CPLEX 7.0 User's Manual*. Gentilly, France, 2000.

C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1-5):11–73, February 1997.

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

M. W. Braun, D. Rivera, and Anders Stenman. A 'model-on-demand' identification methodology for non-linear process systems. *Int. J. Control*, 74(18): 1708–1717, Dec 2001.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, September 1988.

G. Cybenko. Just-in-time learninging and estimation. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, NATO ASI Seried, pages 423 –434, Berlin, 1996. Springer.

J. R. Deller. Set membership identification in digital signal processing. *IEEE ASSP Magazine*, 4:4–20, 1989.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, 1996.

C. Harris, X. Hong, and Q. Gan. *Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach*. Springer-Verlag, 2002.

W. Härdle. *Applied Nonparametric Regression*. Cambridge Univeristy Press, Cambridge, UK, 1990.

A. Juditsky, A. Nazin, J. Roll, and L. Ljung. Adaptive DWO estimator of a regression function. In *Proc. NOLCOS'04*, Stuttgart, 2004. Submitted.

I. L. Legostaeva and A. N. Shiryaev. Minimax weights in a trend detection problem of a random process. *Theory of Probability and its Applications*, 16 (2):344–349, 1971.

S. L. Leonov. Remarks on extremal problems in nonparametric curve estimation. *Statistics & Probability Letters*, 43(2):160–178, 1999.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–676, 1973.

M. Milanese and G. Belforte. Estimations theory and uncertainty intervals evaluation in the presence of unkown but bounded errors: Linear families of models and estimators. *IEEE Trans. on Automatic Control*, AC-27:408–414, 1982.

K. S. Narendra and S.-M. Li. Neural networks in control systems. In P. Smolensky, M. C. Mozer, and D. E. Rumelhart, editors, *Mathematical Perspectives on Neural Networks*, chapter 11, pages 347–394. Lawrence Erlbaum Associates, 1996.

A. Nazin, J. Roll, and L. Ljung. A study of the DWO approach to function estimation at a given point: Approximately constant and approximately linear function classes. Technical Report LiTH-ISY-R-2578, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, Dec 2003.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

J. Roll. *Local and Piecewise Affine Approaches to System Identification*. PhD thesis, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, April 2003.

J. Roll, A. Nazin, and L. Ljung. A non-asymptotic approach to local modelling. In *The 41st IEEE Conference on Decision and Control*, pages 638–643, December 2002.

J. Roll, A. Nazin, and L. Ljung. Local modelling with a priori known bounds using direct weight optimization. In *European Control Conference*, Cambridge, September 2003.

J. Roll, A. Nazin, and L. Ljung. Direct weight optimization for nonparametric estimation of a regression function at a given point. *Scand. J. Statist.*, 2004. Submitted.

J. Sacks and D. Ylvisaker. Linear estimation for approximately linear models. *The Annals of Statistics*, 6(5):1122–1137, 1978.

F.C. Schweppe. Recursive state estimation - unknown but bounded errors and system inputs. *IEEE Trans. on Automatic Control*, 13(37):22–28, 1968.

J. Sjöberg, 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*, 31(12):1691–1724, 1995.

A. Stenman. *Model on Demand: Algorithms, Analysis and Applications*. PhD thesis, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, 1999.

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.

M. Vidyasagar. *A Theory of Learning and Generalization*. Springer Verlag, London, 1997.