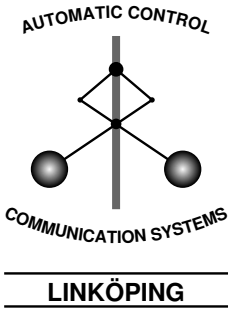
Non-linear System Identification Via Direct Weight Optimization

Jacob Roll, Alexander Nazin, Lennart Ljung

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: roll@isy.liu.se, nazine@ipu.rssi.ru, ljung@isy.liu.se

13th September 2005



Report no.: LiTH-ISY-R-2696
Submitted to Automatica

Technical reports from the Control & Communication group in Linköping are available at http://www.control.isy.liu.se/publications.

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 (or an upper bound) 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.

Keywords: Non-parametric identification, Function approximation, Minimax techniques, Quadratic programming, Nonlinear systems, Mean-square error, Local structures

Non-linear System Identification Via Direct Weight Optimization

Jacob Roll*, Alexander Nazin†, and Lennart Ljung*
* Div. of Automatic Control, Linköping University
SE-58183 Linköping, Sweden
Email: roll, ljung@isy.liu.se
† Institute of Control Sciences, Profsoyuznaya str., 65
117997 Moscow, Russia
Email: nazine@ipu.rssi.ru

2005-09-13

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 (or an upper bound) 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., (Chen and Billings, 1992; Harris et al., 2002; Roll et al., 2002; Sjöberg et al., 1995; Suykens et al., 2002; 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.

One may ask if it is meaningful to 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 under certain conditions 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 basic problem setting considered is as follows: Given data $\{\varphi(t), y(t)\}_{t=1}^{N}$ from the system

$$y(t) = f_0(\varphi(t)) + e(t) \tag{1}$$

where $f_0(\cdot)$ is unknown, $\varphi(t)$ is the regression vector, and $e(t)$ is noise, we would like to find a good linear (affine) estimator of the function f_0

$$\hat{f}(\varphi^*) = w_0 + \sum_{t=1}^{N} w_t y(t) \tag{2}$$

at a given point φ^*. The performance of the estimator will depend on how the weights w_0 and w_t are selected, and we can thus view the problem as an optimization problem in the weights; it just remains to specify a criterion to minimize. An interesting alternative would be the *mean-square error (MSE)*

$$W(\varphi^*, f_0, w^N) = E[(f_0(\varphi^*) - \hat{f}(\varphi^*))^2] \tag{3}$$

where $w^N = [w_0 \ w_1 \ \dots \ w_N]^T$ and the expectation is taken with respect to the noise terms $e(t)$. Unfortunately, the MSE itself is not computable (it depends on the unknown function f_0). If, however, we know that f_0 belongs to a certain function class $\mathcal{F}$, we can use the *worst-case MSE*

$$\sup_{f \in \mathcal{F}} W(\varphi^*, f, w^N) \tag{4}$$

as a criterion function, getting a minimax approach to the estimation problem. As we will see, though, the worst-case MSE is not easily computable for all function classes, and we might have to resort to upper bounds.

1.1 Related approaches

The affine estimator (2) is in fact very common in the literature on non-linear estimation, and many methods have been suggested to determine the weights w_t. A very wide-spread technique is formed by so called *kernel methods* (see e.g., Härdle, 1990). Then the weights w_t depend on the distance between the given point φ^* and the observation points $\varphi(t)$ via the *kernel function* K_H:

$$w_t = K_H(\varphi(t) - \varphi^*) = K_H(\widetilde{\varphi}(t))$$

where we define $\widetilde{\varphi}(t) = \varphi(t) - \varphi^*$. The index H indicates the *bandwidth* of the estimator, typically

$$K_H(\widetilde{\varphi}) = 0 \qquad \text{if} \quad \|H^{-1}\widetilde{\varphi}\| > 1$$

where H is a positive definite, symmetric matrix. Natural normalization (with $w_0 = 0$) is to let $\sum w_t = 1$. Then this kernel estimator is known as the Nadaraya-Watson estimator, (Nadaraya, 1964; Watson, 1964).

Often the kernel functions are formed from just one basic function K, which is scaled by the bandwidth matrix H, so that

$$K_H(\widetilde{\varphi}) = K(H^{-1}\widetilde{\varphi})$$

It is common that the kernel function is spherically symmetric and H is a scaled identity matrix, $H = hI$. Asymptotic analysis (as $N \to \infty$) shows that an optimal choice in many cases is obtained for the *spherical Epanechnikov kernel* (Epanechnikov, 1969)

$$K(\widetilde{\varphi}) = C(1 - \|\widetilde{\varphi}\|^2)_+ \tag{5}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ and C is a normalization constant.

Another popular approach is the *local polynomial modelling* approach, (Fan and Gijbels, 1996), where the estimator is determined by locally fitting a polynomial to the given data (the Nadaraya-Watson estimator is obtained as a special case, by fitting local constant models to the data). For this, we need to solve a weighted least-squares problem, which for a first-order polynomial takes the form

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{t=1}^{N} K_H(\widetilde{\varphi}(t)) \left(y(t) - (\beta_0 + \beta_1^T \widetilde{\varphi}(t))\right)^2 \tag{6}$$

Here the estimate $\hat{f}(\varphi^*) = \hat{\beta}_0$. It is easy to show that $\hat{\beta}_0$ is linear in y, and the weights w_t in (2) are thus implicitly determined. K_H is a kernel function as above, that focuses the polynomial fit of the function to observations in the vicinity of φ^* (hence the name *local polynomial*). The bandwidth h can be determined to asymptotically minimize the worst-case MSE over all linear estimators (see Fan and Gijbels, 1996, for details).

In fact, the affine estimator (2) includes several other approaches to function estimation, such as *kriging*, (Cressie, 1993), *Gaussian processes*, (Gibbs, 1997; Rasmussen, 1996) *least squares support vector machines (LS-SVM)*, (Suykens et al., 2002) and several others. We refer to (Suykens et al., 2002), Section 3.6 for a discussion of these connections.

The different methods mentioned above for choosing the weights w_t in the linear estimator (2) are typically justified using asymptotic arguments, as $N \to \infty$. However, in reality only a finite number of data is given. Furthermore, these data may be sparsely and non-uniformly distributed, in particular when the dimension n of φ is high. This might deteriorate the performance of the estimation methods. To a certain extent, this problem can be compensated for by choosing the bandwidth in an adaptive way (see, e.g., Lepski and Spokoiny, 1997). However, the shape of the kernels is still fixed and is not adjusted to the actual data.

In contrast, the DWO approach considered in this paper is a non-asymptotic approach, which takes the positions of the actual data observations into account

and finds the optimal weights for the estimator (2). It is an extension of what has previously been presented in (Roll et al., 2002, 2003a,b). 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.

For concreteness and clarity reasons, we will mainly concentrate on giving the details of the main algorithms for a special class of functions with Lipschitz continuous gradient. These algorithms are derived in Section 5. 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

$$y(t) = f_0(Z^{t-1}) + e(t) \tag{7}$$

where

$$Z^{t-1} = [y(1), u(1), y(2), u(2), \ldots, y(t-1), u(t-1)] \tag{8}$$

and $e(t)$ is a white noise term. In the notation we here assume that the system is single-input-single-output. It is 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 $\varphi(t)$:

$$\varphi(t) = g(Z^{t-1}) \tag{9a}$$

$$y(t) = f_0(\varphi(t)) + e(t) \tag{9b}$$

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

$$\varphi(t) = [u(t-1) \ \ldots \ u(t-n_b)]^T \tag{10a}$$

for NFIR (nonlinear finite impulse response) models, or

$$\varphi(t) = [y(t-1) \ldots y(t-n_a) \ u(t-1) \ldots u(t-n_b)]^T \tag{10b}$$

for NARX (nonlinear autoregressive with exogenous input) models. See Leontaritis and Billings (1985); Sjöberg et al. (1995) for definitions of different nonlinear model classes.

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 by a linear (affine) combination of observed outputs:

$$\hat{f}(\varphi^*) = w_0 + \sum_{t=1}^{N} w_t y(t) \tag{11}$$

The coefficients will in general depend on the function argument:

$$w_t = w_t(\varphi^*) \tag{12}$$

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 independent of $y^N = [y(1) \ \dots \ y(N)]^T$. 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 optimal performance for linear estimators in terms of minimax risk is not much worse than the overall theoretical optimal performance (Fan and Gijbels, 1996, Theorem 3.11).
- Quite often, a parameterized linear regression model structure with a number of fixed basis functions is used to approximate f_0 in (9):

$$f(\varphi(t), \theta) = \sum_{k=1}^{d} \theta_k f_k(\varphi(t)) \tag{13}$$

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

Estimating the parameter θ in (13) by linear least squares gives an expression

$$\hat{\theta}_N = \left(\sum_{k=1}^{N} F(\varphi(k)) F^T(\varphi(k)) \right)^{-1} \sum_{t=1}^{N} F(\varphi(t)) y(t) \tag{14}$$

where

$$F(\varphi) = \begin{bmatrix} f_1(\varphi) \\ \vdots \\ f_d(\varphi) \end{bmatrix} \tag{15}$$

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

$$\hat{f}_N(\varphi^*) = f(\varphi^*, \hat{\theta}_N) = F^T(\varphi^*)\hat{\theta}_N = \sum_{t=1}^{N} w_t y(t) \tag{16}$$

where

$$\begin{aligned} w_t &= w_t(Z^N, \varphi^*) \\ &= F^T(\varphi^*) \left(\sum_{k=1}^{N} F(\varphi(k)) F^T(\varphi(k)) \right)^{-1} F(\varphi(t)) \end{aligned} \tag{17}$$

We see that this is an expression linear in $y(t)$ just as in (11) (with $w_0 = 0$). 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 (11). 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 get a non-random quality measure by taking the expectation of the square of η:

$$W(\varphi^*, f_0, w^N) = E\eta^2(\varphi^*) \tag{18}$$

to form the mean-square error (MSE) of the estimate.

Remark 1. In the computations, we will assume that $\{\varphi(t)\}_{t=1}^N$ are deterministic. The case of random regression vectors can be treated simply by replacing the expectation in (18) (and subsequent expressions) by the conditional expectation given $\{\varphi(t)\}_{t=1}^N$. Furthermore, the noise $e(t)$ and $\varphi(\tau)$ should be independent for all t, τ. This assumption is violated, e.g., if $\varphi(t)$ depends on $y(\tau)$ for some τ, as in NARX models. However, in practice this has only minor implications (see Remark 3 and Section 6.3).

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:

$$f_0 \in \mathcal{F} \tag{19}$$

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 with respect to w^N:

$$w^* = \underset{w^N}{\operatorname{argmin}} \sup_{f_0 \in \mathcal{F}} W(\varphi^*, f_0, w^N) \tag{20}$$

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. Consequently, it is then convex in w^N. Since the maximum over a set of convex functions is also convex, it means that

$$\tilde{W}(\varphi^*, \mathcal{F}, w^N) = \sup_{f_0 \in \mathcal{F}} W(\varphi^*, f_0, w^N) \tag{21a}$$

is convex and that the problem

$$w^* = \operatorname*{argmin}_{w^N} \tilde{W}(\varphi^*, \mathcal{F}, w^N) \tag{21b}$$

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 (20) that they will depend on

1. The function class $\mathcal{F}$. We shall discuss different such classes shortly.
2. The given regression vectors $\varphi(1), \ldots, \varphi(N)$.
3. 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 (Braun et al., 2001; Stenman, 1999) and also *Just in Time*-models (Cybenko, 1989) 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. The model on demand approach can be especially advantageous when working with complex systems, for which a global parametric model would be difficult to compute, with the risk of getting stuck in local minima. In such cases, the model on demand approach is easier to handle, and we know the approximate computation times in advance.

4 Examples of Some Function Classes

Let us discuss the minimization of (21) 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 $f_k(\varphi)$:

$$\begin{aligned} \mathcal{F}_{\text{par}} = \{f | f(\varphi) = \sum_{k=1}^{d} \theta_k f_k(\varphi) = F(\varphi)^T \theta \\ \text{for some } \theta \in \mathbb{R}^d\} \\ F(\varphi) = \begin{bmatrix} f_1(\varphi) & \dots & f_d(\varphi) \end{bmatrix}^T \end{aligned} \tag{22}$$

In this case we can easily show the following proposition:

Proposition 1. *Consider the problem (21) for the function class (22) when the noise terms $e(t)$ in (9) 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. The minimizing weights w^* are then given by (17).*

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 (22). The MSE (18) can be written

$$\begin{aligned} W(\varphi^*, f_0, w^N) &= E\left[\left(\sum_{t=1}^{N} w_t y(t) - f_0(\varphi^*)\right)^2\right] \\ &= E\left[\left(\sum_{t=1}^{N} w_t (f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2\right] \\ &= \left(\sum_{t=1}^{N} w_t F(\varphi(t))^T \theta_0 - F(\varphi^*)^T \theta_0\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ &= \left(\left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*)\right)^T \theta_0\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \end{aligned} \tag{23}$$

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

$$\sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \tag{24}$$

Under this requirement, the bias term completely disappears from (23). To find

the solution of (21), we hence need to solve the optimization problem

$$\begin{aligned} \min_{w^N} \quad & \sum_{t=1}^{N} w_t^2 \\ \text{subj. to} \quad & \sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \end{aligned} \tag{25}$$

But this is nothing less than finding the least-norm solution to (24), which is given exactly by (17), and the proposition is proved. □

So in the case of the function class (22) 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 (21) seems to be reasonable.

4.2 Functions with Local Smoothness

The real advantage of considering (21), however, is that we can use less knowledge about the true function f_0.

It may seem a very specific type of prior knowledge about the system to assume that it belongs to a specified family like (22). It could be more natural to have some idea about the local smoothness of the predictor function. In Section 5 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 (11) depends primarily on the observations close to the target regressor φ^*.

To be more specific, the function class we will mainly consider in Section 5 is defined by

$$\begin{aligned} \mathcal{F}_2(Q) = \{ f \in \mathcal{C}^1 \,|\, & \|\nabla f(\varphi + h) - \nabla f(\varphi)\|_{Q^{-1}} \leq \|h\|_Q \\ & \forall \varphi, h \in \mathbb{R}^n \} \end{aligned} \tag{26}$$

where ∇ denotes gradient, $\|h\|_Q \triangleq \sqrt{h^T Q h}$ and Q is a symmetric, positive definite matrix. For twice differentiable functions f, the inequality can be interpreted as an upper bound on the Hessian of f. However, we also allow functions that are not twice differentiable. A special case of (26) is given by $Q = LI$, where L is a scalar and I is the identity matrix. In this case, (26) becomes a standard Lipschitz condition on the gradient, with L as the Lipschitz constant.

4.3 A Realistic Noise Model

In some contexts the simple description of the term e in (7) as white noise or even random variables is rejected. Indeed, there could be many reasons why this is not a realistic description. Another alternative 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 known constant. See, among many references, e.g., (Deller, 1989; Milanese and Belforte, 1982; Milanese et al., 1996; Schweppe, 1968). 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 $e_u(t)$:

$$y(t) = f_0(Z^{t-1}) + e_u(t) + e_s(t), \quad |e_u(t)| \leq C_e \tag{27}$$

In this case it is easy to define function classes $\mathcal{F}_{\text{UBB}}$ that include the component e_u. For example, for the function class (22) for f_0 we would have the version

$$y(t) = f_0(\varphi(t)) + e_u(t) + e_s(t), \quad |e_u(t)| \leq C_e \tag{28a}$$

$$\Rightarrow y(t) = \tilde{f}_0(\varphi(t)) + e_s(t) \tag{28b}$$

$$\tilde{f}_0 \in \mathcal{F}_{\text{UBB}} = \{f| \, |f(\varphi) - F(\varphi)^T\theta| \leq C_e \text{ for some } \theta \in \mathbb{R}^d\} \tag{28c}$$

Note that the functions in this class in general are non-smooth. DWO solutions for this function class are investigated in (Nazin et al., 2003).

5 The Direct Weight Optimization Algorithm

Let us now turn to the main topic of this paper. In this section, the DWO approach will be described for the function class introduced in Section 4.2. In Section 5.3 we will also briefly outline the approach for the realistic noise models described by (28). For more general expressions, see (Roll et al., 2005).

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

$$y(t) = f_0(\varphi(t)) + e(t) \tag{29}$$

where f_0 is an unknown function, $f_0 : \mathbb{R}^n \to \mathbb{R}$, $\varphi(t) \in \mathbb{R}^n$, 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 $\mathcal{F}_2(Q)$ described by (26).

Now, the problem to solve is to find an estimator (11) to estimate $f_0(\varphi^*)$ at a certain point φ^*, such that the worst-case MSE (21) 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.

Remark 3. When estimating NARX 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 in Section 7, the method often works well in practice anyway. See Section 6.3 for a discussion about this.

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 (11) can be written

$$\begin{aligned} W(\varphi^*, f_0, w^N) &= E\left(w_0 + \sum_{t=1}^N w_t y(t) - f_0(\varphi^*)\right)^2 \\ &= E\left(w_0 + \sum_{t=1}^N w_t (f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2 \\ &= \left(w_0 + \sum_{t=1}^N w_t f_0(\varphi(t)) - f_0(\varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^N w_t^2 \\ &= \Bigg(w_0 + \sum_{t=1}^N w_t \left(f_0(\varphi(t)) - f_0(\varphi^*) - \nabla^T f_0(\varphi^*)\widetilde{\varphi}(t)\right) \\ &\quad + f_0(\varphi^*)\left(\sum_{t=1}^N w_t - 1\right) + \nabla^T f_0(\varphi^*) \sum_{t=1}^N w_t \widetilde{\varphi}(t)\Bigg)^2 \\ &\quad + \sigma^2 \sum_{t=1}^N w_t^2 \end{aligned} \tag{30}$$

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 (30) can get arbitrarily large unless we impose the following constraints on the weights:

$$\sum_{t=1}^N w_t = 1 \tag{31a}$$

$$\sum_{t=1}^N w_t \widetilde{\varphi}(t) = 0 \tag{31b}$$

In other words, for the worst-case MSE to be finite, (31) 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 (31), any linear function is estimated with zero bias.

Under the restrictions (31) and by using the definition (26) of $\mathcal{F}_2$, we get the following upper bound on the MSE:

$$\begin{aligned} &W(\varphi^*, f_0, w^N) \\ &\quad \le \left(\frac{1}{2}\sum_{t=1}^N |w_t| \|\widetilde{\varphi}(t)\|_Q^2 + |w_0|\right)^2 + \sigma^2 \sum_{t=1}^N w_t^2 \end{aligned} \tag{32}$$

Note that the upper bound is tight whenever the weights w_t and w_0 are non-negative. 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:

$$\begin{aligned} \min_{w^N} \quad & \frac{1}{4}\left(\sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ \text{subj. to} \quad & \sum_{t=1}^{N} w_t = 1 \\ & \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{33}$$

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

$$\begin{aligned} \min_{w^N, s} \quad & \frac{1}{4}\left(\sum_{t=1}^{N} s_t \|\widetilde{\varphi}(t)\|_Q^2\right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq w_t \\ & s_t \geq -w_t \\ & \sum_{t=1}^{N} w_t = 1 \\ & \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{34}$$

and can be solved efficiently to get the optimal w^N.

Remark 4. We may note that for the case $Q = 0$, $\mathcal{F}_2(Q)$ is nothing but the class of linear functions. Hence, in this case we are back to the situation in Section 4.1, i.e., the solution to (34) will be the classical least-squares solution for an ARX system. In the other extreme case, when $\sigma^2 = 0$, we get a linear interpolation between the data points. In that case, if $\varphi^* = \varphi(t)$ for some t, the corresponding estimate $\hat{f}_N(\varphi^*)$ will, quite naturally, equal $y(t)$.

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

$$\begin{aligned} \mathcal{F}_2(Q, \delta, \Delta, R) = \{ f \in \mathcal{F}_2(Q) \,|\, & |f(\varphi^*) - a| \leq \delta, \\ & \|\nabla f(\varphi^*) - b\|_{R^{-1}} \leq \Delta \} \end{aligned} \tag{35}$$

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

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

MSE:

$$W(\varphi^*, f_0, w^N) \leq \left(\frac{1}{2} \sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 \right. \\ + \left| w_0 + a \left(\sum_{t=1}^{N} w_t - 1 \right) + b^T \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right| \\ \left. + \delta \left| \sum_{t=1}^{N} w_t - 1 \right| + \Delta \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{36}$$

This upper bound can for any given $w = [w_1 \ \ldots \ w_N]^T$ be minimized with respect to w_0, giving

$$w_0 = -a \left(\sum_{t=1}^{N} w_t - 1 \right) - b^T \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \tag{37}$$

By inserting this into (36), the upper bound on the MSE is reduced to

$$W(\varphi^*, f_0, w^N) \leq \left(\frac{1}{2} \sum_{t=1}^{N} |w_t| \|\widetilde{\varphi}(t)\|_Q^2 \right. \\ \left. + \delta \left| \sum_{t=1}^{N} w_t - 1 \right| + \Delta \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{38}$$

We can now minimize (38) 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)*

$$\begin{aligned} &\min_{w^N, s, r} \ r_c \\ &\text{subj. to} \ \left| \sum_{t=1}^{N} w_t - 1 \right| \leq r_a \\ &\qquad \left\| \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) \right\|_R \leq r_b \\ &\qquad |w_t| \leq s_t, \qquad t = 1, \ldots, N \\ &\left\| \begin{bmatrix} 2 \left(\delta \cdot r_a + \Delta \cdot r_b + \frac{1}{2} \sum_{t=1}^{N} \|\widetilde{\varphi}(t)\|_Q^2 s_t \right) \\ 2\sigma s \\ 1 - r_c \end{bmatrix} \right\| \leq 1 + r_c \end{aligned} \tag{39}$$

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

Note that, since we have incorporated more information about the true function than in Section 5.1, the optimal upper bound on the MSE obtained from (39) will never be worse than what we get from (34). On the other hand, if the prior information about $f(\varphi^*)$ and $\nabla f(\varphi^*)$ is too imprecise (i.e., if δ and Δ are large enough), it is not necessarily better either. In fact, under some relatively general conditions it can be shown that for large enough values of δ and Δ, (39) and (34) will give exactly the same solutions. See (Roll et al., 2003b) for more details.

5.3 Dealing with the Realistic Noise Model

The computations in the previous sections can easily be adjusted to the models introduced in Section 4.3. Given a true system described by (28) and an affine estimator (11), we can write the MSE as follows (where θ_0 is the true, unknown parameter vector corresponding to f_0):

$$\begin{aligned} W_{\mathrm{UBB}}(\varphi^*, f_0, w^N) &= E\left(w_0 + \sum_{t=1}^{N} w_t y(t) - f_0(\varphi^*)\right)^2 \\ &= E\left(w_0 + \sum_{t=1}^{N} w_t(\tilde{f}_0(\varphi(t)) + e_s(t)) - F(\varphi^*)^T\theta_0\right)^2 \\ &= \left(w_0 + \sum_{t=1}^{N} w_t \tilde{f}_0(\varphi(t)) - F(\varphi^*)^T\theta_0\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ &= \left(w_0 + \sum_{t=1}^{N} w_t \left(\tilde{f}_0(\varphi(t)) - F(\varphi(t))^T\theta_0\right)\right. \\ &\quad \left. + \theta_0^T\left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*)\right)\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \end{aligned} \tag{40}$$

This quantity can get arbitrarily large unless we impose the restriction

$$\sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \tag{41}$$

Under this restriction, however, we obtain the following upper bound on the MSE:

$$W_{\mathrm{UBB}}(\varphi^*, f_0, w^N) \leq \left(|w_0| + C_e \sum_{t=1}^{N} |w_t|\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{42}$$

Choosing $w_0 = 0$ (which is clearly optimal), the QP to minimize becomes

$$\begin{aligned} \min_{w^N, s} \quad & C_e^2 \left(\sum_{t=1}^{N} s_t\right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq w_t \\ & s_t \geq -w_t \\ & \sum_{t=1}^{N} w_t F(\varphi(t)) = F(\varphi^*) \end{aligned} \tag{43}$$

which, again, can be solved efficiently.

6 Properties of the Solutions

In this section, some interesting properties of the solutions to (34) for the local smoothness class of Section 5.1 are investigated.

6.1 Finite Bandwidth

Since only local smoothness of the predictor function is assumed, very few conclusions about function values can be drawn 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 (or its upper bound) *for any finite data record* N.

In particular, for the problem (34), 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* (34) *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* (w^*, s^*), *it holds that*

$$w_t^* = \begin{cases} P(\widetilde{\varphi}(t)) - \mu_3 \|\widetilde{\varphi}(t)\|_Q^2, & \mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \leq P(\widetilde{\varphi}(t)) \\ 0, & |P(\widetilde{\varphi}(t))| \leq \mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \\ P(\widetilde{\varphi}(t)) + \mu_3 \|\widetilde{\varphi}(t)\|_Q^2, & P(\widetilde{\varphi}(t)) \leq -\mu_3 \|\widetilde{\varphi}(t)\|_Q^2 \end{cases} \tag{44}$$

with $P(\widetilde{\varphi}(t))$ *given by*

$$P(\widetilde{\varphi}(t)) = \mu_1 + \mu_2^T \widetilde{\varphi}(t) \tag{45}$$

Remark 5. 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 (44) is illustrated for the univariate case in Figure 1.

Remark 6. When data are symmetrically spread (i.e., if the nonzero $\widetilde{\varphi}(k)$ can be paired so that for each pair $(\widetilde{\varphi}(i), \widetilde{\varphi}(j))$ we have $\widetilde{\varphi}(i) = -\widetilde{\varphi}(j)$), it can be shown that $\mu_2 = 0$ (see Roll (2003, Theorem 3.3) for the univariate case). This means that the weights will be exactly the same as for the Epanechnikov kernel (5) with appropriately chosen bandwidth.

Proof. The proof uses the *Karush-Kuhn-Tucker (KKT) conditions*. Since the QP (34) 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).

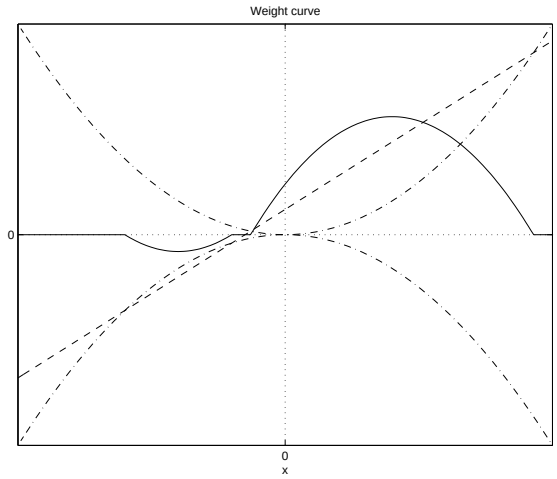


Figure 1: Principal shape of the weight curve (44) for the univariate case (solid curve). The dash-dotted parabolas are $\pm\mu_3\widetilde{\varphi}^2$, and the dashed line is $\mu_1+\mu_2\widetilde{\varphi}$. (The weight curve is scaled by a factor 4 to make the figure more clear.)

The Lagrangian function of (34) can be written

$$\begin{aligned}\mathcal{L}(w,s;\mu,\lambda) = \frac{1}{4}\left(\sum_{t=1}^{N} s_t\|\widetilde{\varphi}(t)\|_Q^2\right)^2 + \sigma^2\sum_{t=1}^{N} s_t^2 \\ - 2\sigma^2\mu_1\left(\sum_{t=1}^{N} w_t - 1\right) - 2\sigma^2\mu_2\sum_{t=1}^{N} w_t\widetilde{\varphi}(t) \\ - 2\sigma^2\sum_{t=1}^{N}(\lambda_t^+(s_t - w_t) + \lambda_t^-(s_t + w_t))\end{aligned} \tag{46}$$

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:

$$\mu_1 + \mu_2 \widetilde{\varphi}(t) = \lambda_t^+ - \lambda_t^- \tag{47a}$$

$$\frac{1}{4\sigma^2}\left(\sum_{k=1}^{N} |w_k^*| \|\widetilde{\varphi}(k)\|_Q^2\right) \|\widetilde{\varphi}(t)\|_Q^2 + |w_t^*| = \lambda_t^+ + \lambda_t^- \tag{47b}$$

$$\sum_{t=1}^{N} w_t^* = 1 \tag{47c}$$

$$\sum_{t=1}^{N} w_t^* \widetilde{\varphi}(t) = 0 \tag{47d}$$

$$s_t^* = |w_t^*| \tag{47e}$$

$$\lambda_t^+(|w_t^*| - w_t^*) = 0 \tag{47f}$$

$$\lambda_t^-(|w_t^*| + w_t^*) = 0 \tag{47g}$$

$$\lambda_t^{\pm} \geq 0, \quad t = 1, \ldots, N \tag{47h}$$

Let

$$\mu_3 = \frac{1}{4\sigma^2}\left(\sum_{k=1}^{N} |w_k^*| \|\widetilde{\varphi}(k)\|_Q^2\right) \tag{48}$$

From (47f) and (47g), 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

$$w_t^* = \mu_1 + \mu_2 \widetilde{\varphi}(t) - \operatorname{sgn}(w_t^*) \mu_3 \|\widetilde{\varphi}(t)\|_Q^2, \quad w_t^* \neq 0 \tag{49}$$

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 (47a), (47b), and (47h) 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, (44) 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 (34) considerably smaller. Having solved (34), one can easily check whether or not the excluded weights really should be zero, by checking if the excluded data points satisfy $|\mu_1 + \mu_2^T \widetilde{\varphi}(t)| \leq \mu_3 \|\widetilde{\varphi}(t)\|_Q^2$ (the middle case of (44)).

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 (for the univariate case) that using the Epanechnikov kernel would yield an asymptotically optimal (continuous) kernel estimator with respect to the worst-case MSE if the upper bound (32) was tight. Since DWO minimizes (32), one would therefore 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 $\varphi^* \in (-1/2, 1/2)$ under an equally spaced fixed design model*

$$\varphi(k) = \frac{k-1}{N-1} - \frac{1}{2}, \quad k = 1, \ldots, N \tag{50}$$

and with $\sigma > 0$. Let w^* be the minimizer of (33). Then asymptotically, as $N \to \infty$,

$$w_k^* \approx \frac{3}{4} C_N \max\{1 - \left(\frac{\widetilde{\varphi}(k)}{h_N}\right)^2, 0\}, \ k = 1, \ldots, N \tag{51}$$

where

$$C_N \asymp \frac{1}{N h_N}, \quad h_N \asymp \left(\frac{15\sigma^2}{L^2 N}\right)^{1/5} \quad \text{as } N \to \infty \tag{52}$$

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 7. Theorem 2 implies that the optimal weights (51) 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).

Remark 8. When the data inside the bandwidth are lying symmetrically around φ^*, e.g., when $\varphi^* = 0$, it follows that the relation (51) will hold exactly also for finite N, i.e.,

$$w_k^* = \frac{3}{4} C_N \max\{1 - \left(\frac{\widetilde{\varphi}(k)}{h_N}\right)^2, 0\}, \ k = 1, \ldots, N \tag{53}$$

where C_N and h_N obey (52) (see (Roll, 2003, Remark 3.2 and Theorem 3.3) for details).

Proof. For this proof, a special version of Theorem 1 is needed (see Roll, 2003, Theorem 3.2), from which it follows that there are three numbers $\mu_1 > 0$, μ_2, and $\mu_3 > 0$, such that

$$w_k^* = \max\{\mu_1 + \mu_2\widetilde{\varphi}(k) - \mu_3\widetilde{\varphi}^2(k), 0\}, \quad k = 1, \ldots, N \tag{54}$$

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

$$\mu_2^2 \leq 4\mu_3\mu_1 \tag{55}$$

Also recall that the KKT conditions (47) 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

$$\lim_{N\to\infty} \frac{\mu_2^2}{\mu_3\mu_1} = 0 \tag{56}$$

for the three parameters μ_1, μ_2, and μ_3 satisfying (47c), (47d), and (48), with the weights w_k^* given by (54). 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

$$\mu_1 + \mu_2 a - \mu_3 a^2 = 0, \quad \mu_1 + \mu_2 b - \mu_3 b^2 = 0, \quad a < b \tag{57}$$

and suppose that $[a, b] \in [-0.5 - \varphi^*, 0.5 - \varphi^*]$. If we find a solution to the system of the three equations (47c), (47d), and (48) with respect to $\mu_1 > 0$, μ_2, and $\mu_3 > 0$, and (55) is satisfied, then we have proved (54). The following asymptotic relation for nonnegative weights (54) holds true as $N \to \infty$:

$$\frac{1}{N}\sum_{k=1}^{N} w_k\widetilde{\varphi}^m(k) = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right)\widetilde{\varphi}^m d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{58}$$

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

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

Thus, the equations (47c), (47d), and (48) may be written as follows:

$$\frac{1}{N} = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right) d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{59}$$

$$0 = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right)\widetilde{\varphi} d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{60}$$

$$\frac{4\sigma^2}{L^2}\frac{\mu_3}{N} = \int_a^b \left(\mu_1 + \mu_2\widetilde{\varphi} - \mu_3\widetilde{\varphi}^2\right)\widetilde{\varphi}^2 d\widetilde{\varphi} + O(h/N)\left(\mu_1 + |\mu_2| + \mu_3\right) \tag{61}$$

with

$$a = \frac{\mu_2 - \sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3}, \quad b = \frac{\mu_2 + \sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3}$$
$$h = \frac{\sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3} \tag{62}$$

Note that the terms $O(h/N)$ in (59)–(61) 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, Section 3.4) for a detailed proof) that the solution to (59)–(61) has the following asymptotics:

$$\mu_1 \asymp \frac{3}{4Nh_N}, \quad \mu_2 = O(N^{-1}), \quad \mu_3 \asymp \frac{\mu_1}{h_N^2} \tag{63}$$

with

$$h = \frac{\sqrt{\mu_2^2 + 4\mu_3\mu_1}}{2\mu_3} \asymp h_N = \left(\frac{15\sigma^2}{L^2 N}\right)^{1/5} \tag{64}$$

Thus, we obtain

$$\lim_{N\to\infty} \frac{\mu_2^2}{\mu_3\mu_1} = \lim_{N\to\infty} \frac{\mu_1}{\mu_3}\left(\frac{\mu_2}{\mu_1}\right)^2 = 0 \tag{65}$$

and relation (56) is proved.

Since $\mu_2 = o(\mu_1)$, the relation (51) follows directly from (54) and (63). This proves the theorem. □

6.3 Using the DWO Approach for Dynamic Systems

What happens when the assumption about $e(t)$ and $\varphi(\tau)$ being independent for all t, τ is violated? Let us have a closer look at the problem. For simplicity we consider the basic case of Section 5.1 with the function class $\mathcal{F}_2(Q)$ and $w_0 = 0$. Suppose that the regression vector $\varphi(t)$ contains $y(t-1)$. This would mean that if $\varphi(t)$, $t = 1, \ldots, N$ are given, then also the corresponding $y(t-1)$ are given. Hence, the MSE can be rewritten as

$$\begin{aligned}
W(\varphi^*, f_0, w^N) &= E[(\hat{f}(\varphi^*) - f_0(\varphi^*))^2 | \{\varphi(t)\}_{t=1}^N] \qquad (66)\\
&= \left(\sum_{t=1}^N w_t f_0(\varphi(t)) - f_0(\varphi^*)\right)^2 \\
&+ 2\left(\sum_{t=1}^N w_t f_0(\varphi(t)) - f_0(\varphi^*)\right) \\
&\quad \cdot \left(\sum_{t=1}^{N-1} w_t \big(y(t) - f_0(\varphi(t))\big)\right) \\
&+ 2\sum_{t=1}^{N-2}\sum_{j=t+1}^{N-1} w_t w_j \big(y(t) - f_0(\varphi(t))\big)\big(y(j) - f_0(\varphi(j))\big) \\
&+ \sum_{t=1}^{N-1} w_t^2 \big(y(t) - f_0(\varphi(t))\big)^2 + \sigma^2 w_N^2
\end{aligned}$$

Since f_0 is unknown, there is generally no way to evaluate this expression, and we cannot get an upper bound either. However, for large N, the second and third terms of the last expression should generally be much smaller than the squared terms, since $y(t) - f_0(\varphi(t)) = e(t)$ is the noise contribution, which is averaged in these sums. Furthermore, the fourth and fifth terms should be well approximated by

$$\sigma^2 \sum_{t=1}^{N} w_t^2 \tag{67}$$

Hence, it seems reasonable to approximate the second and third terms in this expression by 0, and the fourth and fifth terms by (67), and we are back to the MSE expression (30). The only difference is that this is not the true MSE anymore, but an approximation. As we will see in Section 7, the approach works well in practice also for NARX systems.

One should also note that, as $N \to \infty$, the weights from the DWO approach will be nonzero only in a very small neighborhood of φ^*. For most reasonable dynamic systems, unless φ^* is an equilibrium point, this means that the regression vectors corresponding to the nonzero weights will have very different indices t, and so they will in general only be weakly correlated. This means that the approximation of the worst-case MSE used by the DWO approach will be asymptotically correct.

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 (69) below. When the optimal weights are determined for these target points, they are however calculated only using the data in Z_e^N. This will make comparisons to other methods more fair.

7.1 A nonlinear ARX (NARX) system

We begin by considering a model of NARX type, where Q and σ^2 are known.

Example 1. Consider the following NARX system:

$$\begin{aligned} y(t) = &\begin{bmatrix} 0.1 & -0.1 & 0.25 & 0.5 \end{bmatrix} \cdot \varphi(t) \\ &+ \frac{L}{2}\Big(\|\varphi(t)\|^2 - 2\left(\max\{\|\varphi(t)\|^2, 1\} - 1\right) \\ &+ 2\left(\max\{\|\varphi(t)\|^2, 2\} - 2\right) \\ &- \left(\max\{\|\varphi(t)\|^2, 3\} - 3\right)\Big) + e(t) \end{aligned} \tag{68}$$

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 (26) with $Q = LI = 0.1I$. $N = 300$ data samples were collected by simulation with an input $u(t) \in N(0, 1)$.

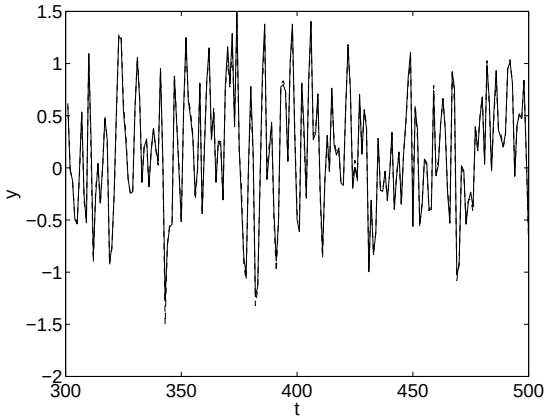


Figure 2: Simulated (solid) and true (dashed) output (validation data) for system (68), modeled using the DWO approach with $Q = 0.1I$. The fit is 90.8%.

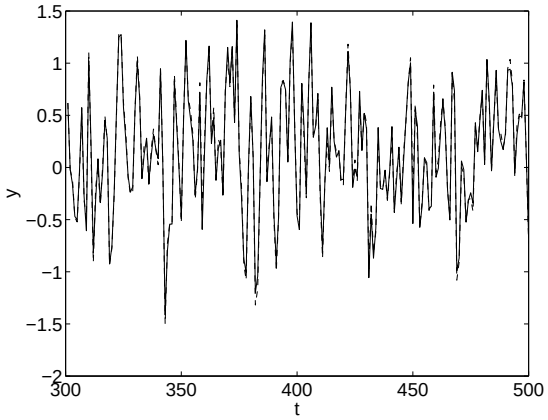


Figure 3: Simulated (solid) and true (dashed) output for system (68), modeled using an artificial neural network. The fit is 89.8%.

To test the quality of the proposed approach, another set of 200 data samples with $u(t) \in N(0, 1)$ was collected. The DWO technique was then used to simulate the system output. It is worth commenting on how this is done: The predictor function is defined by

$$\hat{y}(t|t-1) = \hat{f}(\varphi(t))$$
$$\varphi(t) = [y(t-1) \dots y(t-n_a)\ u(t-1) \dots u(t-n_b)]$$

(where in this particular example $n_a = n_b = 2$). A simulation of the model uses

only the input, so it is accomplished recursively as

$$\begin{aligned} y_s(t) &= \hat{f}(\varphi_s(t)) \\ \varphi_s(t) &= [y_s(t-1) \dots y_s(t-n_a)\ u(t-1) \dots u(t-n_b)] \end{aligned} \tag{69}$$

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

$$\left(1 - \sqrt{\frac{\sum_t (y(t) - y_s(t))^2}{\sum_t (y(t) - \bar{y})^2}}\right) \cdot 100\% \tag{70}$$

where $\bar{y}$ is the arithmetic mean of y.

In Figure 2, the resulting output is compared to the true, noiseless output. As can be seen, the simulation gives a good result (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 NARX 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

$$\begin{aligned} x_1(t+1) &= \left(\frac{x_1(t)}{1+x_1^2(t)} + 1\right) \sin x_2(t) \\ x_2(t+1) &= x_2(t) \cos x_2(t) + x_1(t) e^{-\frac{x_1^2(t)+x_2^2(t)}{8}} \\ &\quad + \frac{u^3(t)}{1+u^2(t)+0.5\cos(x_1(t)+x_2(t))} \\ y(t) &= \frac{x_1(t)}{1+0.5\sin x_2(t)} + \frac{x_2(t)}{1+0.5\sin x_1(t)} + e(t) \end{aligned} \tag{71}$$

The noise term $e(t)$ is added in accordance with (Stenman, 1999, Section 5.7.2) and has a variance of 0.1. The states are assumed not to be measurable, and following the discussion in (Stenman, 1999), an NARX331 structure is used to model the system, i.e., $n_a = n_b = 3$. As estimation data, $N = 50000$ samples were generated by simulation using a uniformly distributed random input $u(t) \in [-2.5, 2.5]$. To validate the model, the input signal

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

was used. Figure 4 shows the simulated output when Q was chosen to be $0.1I$. (Note that there is no true value of Q in this case, since the true system is not an NARX system.) 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), or with the results reported in (Narendra and Li, 1996) which are of the same quality order (no explicit numbers are given).

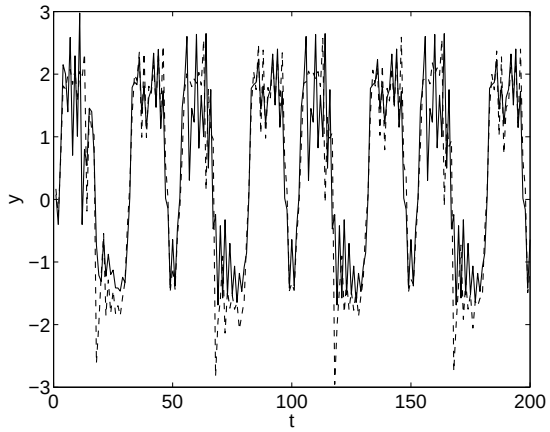


Figure 4: Simulated (solid) and true (dashed) output for system (71), modeled using the DWO approach with $Q = 0.1I$. The fit is 49.7%.

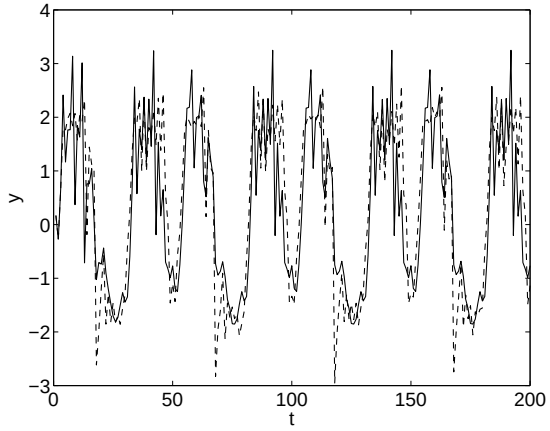


Figure 5: Simulated (solid) and true (dashed) output for system (71), modeled using an artificial neural network. The fit is 47.1%.

7.3 Choice of Q and σ

In the previous example, the matrix Q was not known *a priori* but was regarded as a design parameter, and was chosen to be constant over the entire state-space. An alternative would be to estimate a local Q for each point φ^*. 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

$$\hat{H}(\varphi^*) = T(\varphi^*)D(\varphi^*)T^T(\varphi^*) \tag{72}$$

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

$$\hat{Q}(\varphi^*) = T(\varphi^*)\bar{D}(\varphi^*)T^T(\varphi^*) \tag{73}$$

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

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

If the noise variance σ^2 is also unknown, we can estimate it as well. This can be done using, e.g., the C_p criterion (Cleveland and Devlin, 1988; Mallows, 1973), modified as described in (Stenman, 1999, Section 4.4.5) and (Roll, 2003, Section 7.2). One should observe that using $\alpha\hat{Q}(\varphi^*)$ and $\alpha\hat{\sigma}(\varphi^*)$ for an arbitrary $\alpha > 0$ does not influence the resulting weights w_t; that is, only the ratio between $\|Q\|$ and σ is relevant for the resulting weights as long as the "shape" of Q is fixed.

7.4 Cell Dynamics

The following example deals with data simulated from equations of the same character as the glucose metabolism in cell dynamics. Here, Q and σ are estimated using the procedure described in the previous subsection.

Example 3. A set of 200 data samples has been collected from the system

$$\begin{aligned} \dot{x}_1 &= -\frac{x_1 - x_2}{1 + x_1 + x_2} + \frac{u - x - 1}{1 + u + x_1 + x_1 u} \\ \dot{x}_2 &= \frac{x_1 - x_2}{1 + x_1 + x_2} - \frac{x_2 - 1}{1 + x_2 + 1} \\ y &= x_2 \end{aligned} \tag{74}$$

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 $\varphi(t) = \begin{bmatrix} y(t-1) & y(t-2) & u(t-1) & u(t-2) \end{bmatrix}^T$. 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 (70). The DWO approach gave a fit of 72.9% and the neural network model a fit of 66.4% in this case.

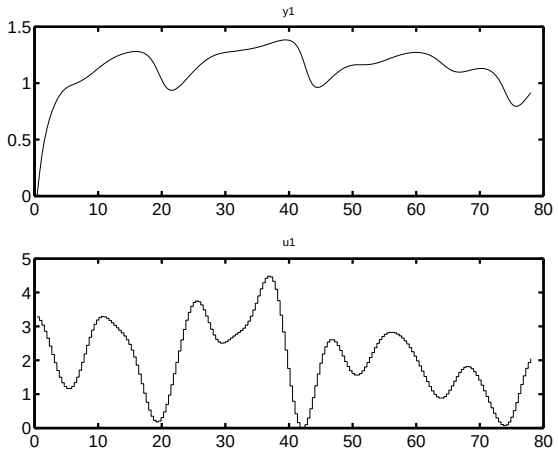


Figure 6: Estimation data from system (74) (Below: input; Above: output).

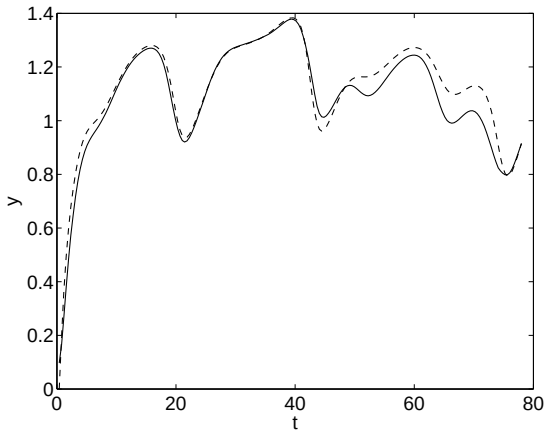


Figure 7: Simulated (solid) and true (dashed) output for system (74), modeled using the DWO approach with $Q = 0.1I$. The fit is 72.9%.

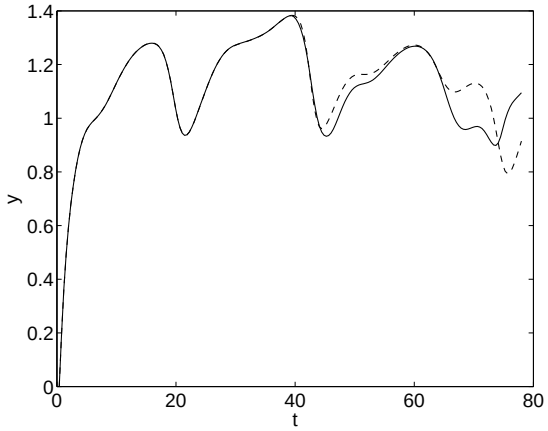


Figure 8: Simulated (solid) and true (dashed) output for system (74), modeled using a sigmoidal neural network with 10 neurons. The fit is 66.4%.

8 Conclusions

There are two main conclusions from this paper:

- The nonlinear identification/estimation problem can be formulated as a direct optimization of a minmax criterion with respect to weights in a 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 (21) is always a convex optimization problem, which gives many useful advantages: Potentially efficient algorithms (Boyd and Vandenberghe, 2004) and unique minima. However, the problem in general is to compute the supremum over $\mathcal{F}$ for fixed w^N. This is often a nontrivial problem (depending on the nature of $\mathcal{F}$), and we might have to resort to upper bounds as in (32) in this paper. In some cases, though, the worst-case MSE is actually computable. This is the case, e.g., for the local smoothness function class $\mathcal{F}_2(Q)$ from Section 4.2 when f_0 is univariate. However, preliminary experiments indicate that only a minor improvement (typically in the order of maximally a few percents decrease in the criterion function value) of the estimates are obtained by using the corresponding optimal estimator, compared to the standard DWO estimator described in Section 5.1. See (Roll, 2003, Section 6.2) for details. Corresponding theoretical conclusions for the asymptotic case 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 worthwhile 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, (26). The local estimation algo-

rithms 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 of 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) for some ideas to estimate the Hessian, and (Juditsky et al., 2004) for implicit estimation of the Lipschitz constant in order to obtain an adaptive estimator, 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, also when the assumptions on independence between noise and regressors are violated. 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, as is often the case for non-convex methods; for instance, this is a well known hassle with 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, (ILOG, Inc., 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, which would be of great value for using the method within a model predictive control (MPC) framework, along the same lines as in (Stenman, 1999, Chapter 7).

References

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. *International Journal of Control*, 74(18):1708–1717, December 2001.

S. Chen and S. A. Billings. Neural networks for nonlinear dynamic system modeling and identification. *International Journal of Control*, 56(2):319–346, August 1992.

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.

Noel A. C. Cressie. *Statistics for spatial data*. Wiley, New York, 1993.

G. Cybenko. Approximation by superpositions of a sigmoidal functions. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

J. R. Deller. Set membership identification in digital signal processing. *IEEE ASSP Magazine*, 6(4):4–20, October 1989.

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14:153–158, 1969.

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

Mark N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997.

W. Härdle. *Applied Nonparametric Regression*. Number 19 in Econometric Society Monographs. Cambridge University Press, 1990.

Chris Harris, Xia Hong, and Qiang Gan. *Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach*. Springer-Verlag, 2002.

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

Anatoli Juditsky, Alexander Nazin, Jacob Roll, and Lennart Ljung. Adaptive DWO estimator of a regression function. In *NOLCOS´04*, Stuttgart, September 2004.

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.

Sergei L. Leonov. Remarks on extremal problems in nonparametric curve estimation. *Statistics & Probability Letters*, 43(2):169–178, 1999.

I. J. Leontaritis and S. A. Billings. Input-output parametric models for non-linear systems - part ii: sthocastic non-linear systems. *International Journal of Control*, 41(2):329–344, 1985.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, December 1997.

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

M. Milanese and G. Belforte. Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors: Linear families of models and estimators. *IEEE Transactions on Automatic Control*, 27(2):408–414, April 1982.

M. Milanese, J. Norton, H. Piet-Lahanier, and É. Walter, editors. *Bounding Approaches to System Identification*. Kluwer Academic/Plenum Publishers, New York, May 1996.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10:186–190, 1964.

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.

Alexander Nazin, Jacob Roll, and Lennart 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, December 2003.

Carl Edward Rasmussen. *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*. PhD thesis, University of Toronto, 1996.

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 of nonlinear dynamic systems using direct weight optimization. In *13th IFAC Symposium on System Identification*, pages 1554–1559, Rotterdam, August 2003a.

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

Jacob Roll, Alexander Nazin, and Lennart Ljung. A general direct weight optimization framework for nonlinear system identification. To be presented at the 16th IFAC World Congress on Automatic Control, July 2005.

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 Transactions on Automatic Control*, 13(1):22–28, February 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.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā, Series A*, 26:359–372, 1964.