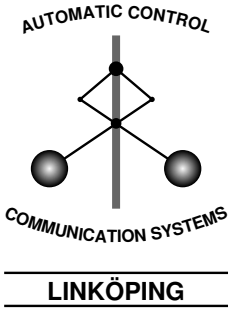
Consistent Nonparametric Estimation of NARX Systems Using Convex Optimization

Jacob Roll, Martin Enqvist, 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, maren@isy.liu.se,
ljung@isy.liu.se

22nd December 2005



Report no.: LiTH-ISY-R-2721
Submitted to the 44th IEEE Conference on Decision and Control and European Control Conference ECC'05, Seville, Spain

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

Abstract

In this paper, a nonparametric method based on quadratic programming (QP) for identification of nonlinear autoregressive systems with exogenous inputs (NARX systems) is presented. We consider a mixed parametric/nonparametric model structure. The output is assumed to be the sum of a parametric linear part and a nonparametric Lipschitz continuous part. The consistency of the estimator is shown assuming only that an upper bound on the true Lipschitz constant is given. In addition, different types of prior knowledge about the system can easily be incorporated. Examples show that the method can give accurate estimates also for small data sets and that the estimate of the linear part sometimes can be improved compared to the linear least squares estimate.

Keywords: System identification, NARX systems, nonparametric models

Consistent Nonparametric Estimation of NARX Systems Using Convex Optimization

Jacob Roll, Martin Enqvist, Lennart Ljung

2005-12-22

Abstract

In this paper, a nonparametric method based on quadratic programming (QP) for identification of nonlinear autoregressive systems with exogenous inputs (NARX systems) is presented. We consider a mixed parametric/nonparametric model structure. The output is assumed to be the sum of a parametric linear part and a nonparametric Lipschitz continuous part. The consistency of the estimator is shown assuming only that an upper bound on the true Lipschitz constant is given. In addition, different types of prior knowledge about the system can easily be incorporated. Examples show that the method can give accurate estimates also for small data sets and that the estimate of the linear part sometimes can be improved compared to the linear least squares estimate.

1 Introduction

The class of nonlinear autoregressive systems with exogenous inputs (NARX systems) (Sjöberg et al., 1995) is a straightforward generalization of linear ARX systems that has been used in many applications. For an NARX system, the optimal one step ahead predictor is a nonlinear function of a finite number of past output and input components. Using a version of the prediction-error method (Ljung, 1999), we will here simultaneously estimate both a nonparametric NARX model and a parametric ARX model such that their sum give an as good prediction of the output as possible. Related model structures have been used in semiparametric or partially linear models (see, for example, Heckman, 1988; Chen et al., 2001).

It is interesting to consider nonparametric methods for nonlinear system identification since the assumptions about the true system are usually weaker for such methods than for parametric methods. For a nonlinear system, it can be hard to tell in advance whether a specific assumption about, for example, the shape of the nonlinearities is reasonable or not. In this paper, the only assumption about the true NARX system is that its nonlinearities are Lipschitz continuous.

This assumption makes it possible to use an approach where the identification problem is formulated as a quadratic programming (QP) problem. By solving this problem, both the parameters of the linear ARX model and the nonparametric NARX model can be estimated at the same time. A version of this idea, without the linear, parametric part, has previously been used for

nonparametric regression and for maximum likelihood estimation of unknown parameters in probability density functions (Bertsimas et al., 1999). Other methods for nonparametric regression can be found in, for example, Fan and Gijbels (1996).

Lipschitz conditions are a common way to guarantee that a function, or some of its derivatives, will be smooth. For example, functions with a Lipschitz continuous gradient can be identified using local modeling such that the worst-case mean-square error is minimized (Roll et al., 2005).

A small nonlinear system component can have a large influence on an estimated linear approximation of the system if standard methods for linear identification are used (Mäkilä, 2005; Enqvist and Ljung, 2004). In some cases, this behavior can be understood if the nonlinear contribution to the system output is viewed as a nonlinear disturbance (Pintelon et al., 2001; Schoukens et al., 1998). The method presented in this paper will make the estimate of the linear model more robust against nonlinearities in the system since the nonparametric NARX model can compensate for some of the nonlinear effects. A related concept is the notion of unknown but bounded noise and set membership identification (Garulli et al., 1999), since a bounded nonlinearity might affect the system output in a similar way as such a noise term.

2 NARX Identification

Consider an NARX system with input $u(t)$ and output $y(t)$ that can be written

$$y(t) = \theta_0^T \varphi(t) + r_0(\varphi(t)) + e(t), \tag{1}$$

where

$$\varphi(t) = \begin{pmatrix} -y(t-1) \\ \vdots \\ -y(t-n_a) \\ u(t-n_k) \\ \vdots \\ u(t-n_k-n_b) \end{pmatrix} \tag{2}$$

is a regression vector and where $e(t)$ is white noise. The constant vector θ_0 defines a linear ARX part of the system while the function r_0 can be nonlinear. Assume that $e(t)$ and $\varphi(t)$ are independent for all t and that r_0 is a Lipschitz continuous function with Lipschitz constant L_0, i.e., that

$$|r_0(\varphi_1) - r_0(\varphi_2)| \leq L_0 \|\varphi_1 - \varphi_2\|_2, \ \forall \varphi_1, \varphi_2 \in \mathbb{R}^n, \tag{3}$$

where $n = n_a + n_b + 1$. Furthermore, assume that a dataset $(\varphi(t), y(t))_{t=1}^N$ consisting of N measurements of the regression vector and the system output is available.

Using this dataset, estimates $\hat{\theta}_N$ and $\hat{r}_N$ of θ_0 and r_0, respectively, can be obtained by solving the QP problem

$$\begin{array}{ll} \underset{\theta_N, \rho_N}{\text{minimize}} & \frac{1}{N} \sum_{t=1}^N (y(t) - \theta_N^T \varphi(t) - \rho_N(t))^2 \\ \text{subject to} & \rho_N(t) - \rho_N(s) \leq L \|\varphi(t) - \varphi(s)\|_2 \\ & \forall s, t \in \{1, 2, \ldots, N\}. \end{array} \tag{4}$$

In this problem, ρ_N is a vector with N elements $\rho_N(t)$ which can be viewed as estimates of $r_0(\varphi(t))$. The constraints on the variables $\rho_N(t)$ imply that these variables will satisfy

$$|\rho_N(t) - \rho_N(s)| \leq L\|\varphi(t) - \varphi(s)\|_2$$

for all $s, t \in \{1, 2, \ldots, N\}$. If the variables $\rho_N(t)$ are viewed as samples from some function, this implies that a Lipschitz condition holds for the sample points $(\varphi(t))_{t=1}^N$. Note that N of the constraints in (4) are trivial ($0 \leq 0$) and present in (4) only for notational convenience. These constraints can be removed without changing the solution of the problem.

An optimal solution $(\hat{\theta}_N, \hat{\rho}_N)$ to the problem (4) can be used to construct one step ahead predictions

$$\hat{y}_N(\varphi(t)) = \hat{\theta}_N^T \varphi(t) + \hat{\rho}_N(t) \tag{5}$$

of the system output for the observed regression vectors $(\varphi(t))_{t=1}^N$. In order to obtain a predictor which can be used for an arbitrary regression vector, the nonparametric function estimate $\hat{\rho}_N$ has to be interpolated.

When $\varphi(t)$ is a scalar, linear interpolation is probably the most natural type of interpolation. However, for $\varphi(t) \in \mathbb{R}^n$ with $n > 1$, linear interpolation of the variables $\hat{\rho}_N(t)$ will in general not result in a function that satisfies the Lipschitz condition for the choice of L used in (4). Instead, for $n > 1$, an estimate $\hat{r}_N$ of r_0 can be defined as

$$\hat{r}_N(\varphi) = \frac{1}{2} \max_{1 \leq t \leq N} (\hat{\rho}_N(t) - L\|\varphi - \varphi(t)\|_2) + \frac{1}{2} \min_{1 \leq t \leq N} (\hat{\rho}_N(t) + L\|\varphi - \varphi(t)\|_2) \tag{6}$$

using a similar construction as in Bertsimas et al. (1999). The function $\hat{r}_N$ is Lipschitz continuous since it is the mean of two Lipschitz continuous functions. Using $\hat{\theta}_N$ and $\hat{r}_N$, a general one step ahead predictor

$$\hat{y}_N(\varphi) = \hat{\theta}_N^T \varphi + \hat{r}_N(\varphi) \tag{7}$$

can be constructed. At first sight, it might seem that the $N + n$ variables used in the problem (4) and for the construction of the model (7) are too many since there are only N measurements. However, thanks to the randomness of the disturbance $e(t)$ in (1), the constraints in (4) will impose an averaging effect on the nonparametric function estimate.

Without these constraints, one optimal solution to (4) is $\theta_N = 0$, $\rho_N(t) = y(t)$ for $t = 1, 2, \ldots, N$. Of course, since the measurements of the output are noisy, such a solution does not give a good model of the true system. By adding constraints like in (4), two variables $\rho_N(t)$ and $\rho_N(s)$ are allowed to differ only marginally from each other if the distance $\|\varphi(t) - \varphi(s)\|_2$ between the corresponding regression vectors is small. In this way, the ρ variables are imposed to have similar properties as samples from the true Lipschitz continuous function r_0. If the set of regression vectors gets more dense when N increases, $\hat{\theta}_N^T \varphi(t) + \hat{\rho}_N(t)$ will approach $\theta_0^T \varphi(t) + r_0(\varphi(t))$. For an intuitive understanding of this convergence, consider a small region in $\mathbb{R}^n$ which contains many regression vectors. The corresponding ρ variables will with a high probability be close to the mean of $y(t) - \hat{\theta}_N^T \varphi(t)$ since the constraints in (4) implies that the ρ variables

should have values close to each other. The consistency of the predictor function estimator (7) will be discussed in Section 3.

Several types of extensions can be made to the identification method presented here. For example, if any prior knowledge about the true system can be written as linear constraints on θ_N and ρ_N, this knowledge can easily be incorporated in the QP problem (4). Examples of such prior knowledge are:

- Bounds on the function r_0 are known in a subset of its domain.
- Different Lipschitz constants can be used in different parts of the domain of r_0.
- The function r_0 is known to be odd or even.
- An expression for the function r_0 is known in a subset of its domain.

Sometimes it could also be interesting to consider the case when only a Lipschitz continuous function should be estimated (setting $\theta_N = \theta_0 = 0$ in (1) and (4)). Analogously to (4), we can handle this case by solving a QP

$$\begin{array}{ll} \underset{\rho_N}{\text{minimize}} & \frac{1}{N}\sum_{t=1}^{N}(y(t)-\rho_N(t))^2 \\ \text{subject to} & \rho_N(t)-\rho_N(s) \leq L\|\varphi(t)-\varphi(s)\|_2 \\ & \forall s,t \in \{1,2,\ldots,N\}. \end{array} \tag{8}$$

The construction of $\hat{r}_N$ using the interpolation method (6) can be used also in this case. In the next section, the consistency of both presented nonparametric identification methods will be shown.

3 Consistency

Before we consider the consistency of the approaches, let us study the behavior of the mean of the predicted outputs at $(\varphi(t))_{t=1}^N$. As the following lemmas show, it is quite simple to show consistency for these.

Lemma 3.1
The optimum of (4) *satisfies*

$$\frac{1}{N}\sum_{t=1}^{N}\left(\hat{\theta}_N^T\varphi(t)+\hat{\rho}_N(t)\right) = \frac{1}{N}\sum_{t=1}^{N} y(t), \tag{9}$$

and, for NFIR systems,

$$\mathrm{E}\left[\frac{1}{N}\sum_{t=1}^{N}\left(\hat{\theta}_N^T\varphi(t)+\hat{\rho}_N(t)\right)\middle|(\varphi(t))_{t=1}^N\right] = \frac{1}{N}\sum_{t=1}^{N}\theta_0^T\varphi(t)+r_0(\varphi(t)). \tag{10}$$

Proof: The Lagrangian of (4) (see Boyd and Vandenberghe (2004)) can be written

$$\begin{aligned}\mathcal{L}(\theta_N,\rho_N;\lambda) = &\frac{1}{N}\sum_{t=1}^{N}(y(t)-\theta_N^T\varphi(t)-\rho_N(t))^2 \\ &-\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij}\left(L\|\varphi(i)-\varphi(j)\|_2-\rho_N(i)+\rho_N(j)\right).\end{aligned} \tag{11}$$

The optimum should satisfy $\frac{\partial \mathcal{L}}{\partial \rho_N(k)} = 0$ for $k = 1, \ldots, N$:

$$-\frac{2}{N}(y(k) - \hat{\theta}_N^T \varphi(k) - \hat{\rho}_N(k)) + \sum_{i=1}^{N}(\hat{\lambda}_{ki} - \hat{\lambda}_{ik}) = 0. \tag{12}$$

Summing (12) over k gives (9). Taking expectations over both sides of (9) then gives (10). □

Lemma 3.2
The optimum of (8) *satisfies*

$$\frac{1}{N}\sum_{t=1}^{N}\hat{\rho}_N(t) = \frac{1}{N}\sum_{t=1}^{N} y(t), \tag{13}$$

and, for NFIR systems,

$$\mathrm{E}\left[\frac{1}{N}\sum_{t=1}^{N}\hat{\rho}_N(t)\middle|(\varphi(t))_{t=1}^{N}\right] = \frac{1}{N}\sum_{t=1}^{N} r_0(\varphi(t)). \tag{14}$$

Proof: As for Lemma 3.1. □

As it now turns out, the identification methods given by (4) and (8), respectively, have fairly attractive properties. We will now show consistency of the estimator given by the QP problem

$$\begin{array}{ll} \underset{\theta_N, \rho_N}{\text{minimize}} & \frac{1}{N}\sum_{t=1}^{N}(y(t) - \theta_N^T\varphi(t) - \rho_N(t))^2 \\ \text{subject to} & \rho_N(t) - \rho_N(s) \le L\|\varphi(t) - \varphi(s)\|_2, \\ & \qquad \forall s, t \in \{1, 2, \ldots, N\} \\ & \pm\theta_N \preccurlyeq m_\theta, \end{array} \tag{15}$$

where $\preccurlyeq$ denotes component-wise inequality. This optimization problem is a version of (4) where bounds on θ_N have been added. It should be pointed out that these bounds can be chosen arbitrarily large. Hence, the restriction compared to (4) is not very severe in practice.

Since (8) is a special case of (15), the consistency of the nonparametric function estimator without a linear part follows from the more general case. A related consistency result for the estimator defined by (8) is shown in Bertsimas et al. (1999) using results from Vapnik (1998). However, the result in the following theorem is based on different assumptions and is shown using an alternative proof.

Theorem 3.1
Consider data sets $(\varphi(t), y(t))_{t=1}^{N}$ *generated from the nonlinear system*

$$y(t) = \theta_0^T\varphi(t) + r_0(\varphi(t)) + e(t) \triangleq f_0(\varphi(t)) + e(t), \tag{16}$$

where $e(t)$ *is a white stationary stochastic process with zero mean and bounded variance* σ^2 *and where* $\varphi(t)$ *is a stationary stochastic process. For each data set, let* $\hat{\theta}_N$ *and* $\hat{\rho}_N(t)$ *be the optimal solution to* (15). *Furthermore, let* $\hat{f}_N$ *be the predictor function given by this solution, i.e.,*

$$\hat{f}_N(\varphi) = \hat{\theta}_N^T\varphi(t) + \hat{r}_N(\varphi),$$

where $\hat{r}_N$ *is defined in* (6). *Suppose that*

1. *$\varphi(t) \in \Phi$, where Φ is a compact set such that the probability density function $p(\varphi)$ for $\varphi(t)$ is positive for all $\varphi \in \Phi$ and that for any $\varepsilon > 0$, Φ can be partitioned*

$$\Phi = \bigcup_{i=1}^{d} \Phi_i, \tag{17}$$

where $\varphi_1, \varphi_2 \in \Phi_i \Rightarrow |\varphi_1 - \varphi_2| \leq \varepsilon$ and $p_i = P(\varphi(t) \in \Phi_i) > 0$ for all $i = 1, 2, \ldots, d$,

2. *the stochastic process $\varphi(t)$ is such that $N_i/N \to p_i$ when $N \to \infty$ w.p.1 for all i in any ε-partitioning* (17) *where*

$$N_i = \text{card}(T_i) \text{ and } T_i = \{t \mid \varphi(t) \in \Phi_i,\ t \leq N\}, \tag{18}$$

3. *$e(t)$ and $\varphi(t)$ are independent, but $\varphi(t)$ may depend on past $e(s)$,*

4. *$|r_0(\varphi_1) - r_0(\varphi_2)| \leq L\|\varphi_1 - \varphi_2\|_2$ for all $\varphi_1, \varphi_2 \in \Phi$,*

5. *$|f_0(\varphi_1) - f_0(\varphi_2)| \leq \tilde{L}\|\varphi_1 - \varphi_2\|_2$ for all $\varphi_1, \varphi_2 \in \Phi$, where $\tilde{L} = L + M_\theta$ and*

$$M_\theta^2 = \sum_{i=1}^{n_a+n_b+1} m_{\theta,i}^2.$$

Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} (\hat{f}_N(\varphi(t)) - f_0(\varphi(t)))^2 = 0 \quad \textit{w.p.1} \tag{19}$$

and

$$\hat{f}_N(\varphi) \to f_0(\varphi) \textit{ uniformly on } \Phi \textit{ as } N \to \infty \quad \textit{w.p.1.} \tag{20}$$

Proof: Take an arbitrary $\varepsilon > 0$ and consider an ε-partitioning such that the first assumption is satisfied. Let $I_{\Phi_i}(\varphi)$ be the indicator function for the set Φ_i, i.e.,

$$I_{\Phi_i}(\varphi) = \begin{cases} 1, & \varphi \in \Phi_i, \\ 0, & \text{otherwise.} \end{cases}$$

Consider arbitrary realizations of the processes $\varphi(t)$ and $e(t)$. With probability one, these realizations are such that $N_i/N \to p_i$ as $N \to \infty$ and that

$$\lim_{N\to\infty} \frac{1}{p_i N} \sum_{t\in T_i} e(t) = \lim_{N\to\infty} \frac{1}{p_i N} \sum_{t=1}^{N} I_{\Phi_i}(\varphi(t))e(t) = 0, \tag{21a}$$

$$\lim_{N\to\infty} \frac{1}{p_i N} \sum_{t\in T_i} |e(t)| = \lim_{N\to\infty} \frac{1}{p_i N} \sum_{t=1}^{N} I_{\Phi_i}(\varphi(t))|e(t)| \leq C, \tag{21b}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} e(t)^2 = \sigma^2 \tag{21c}$$

for some constant C and for all $i = 1, 2, \ldots, d$. The limits (21) follow from the law of large numbers. Let

$$\tilde{e}(t) = I_{\Phi_i}(\varphi(t))e(t),$$

and consider two arbitrary different time instants s and t. Without loss of generality, we can assume that $s > t$. Since $\varphi(t)$, $\varphi(s)$ and $e(t)$ are all independent of $e(s)$, this implies that

$$\mathrm{E}(\tilde{e}(t)\tilde{e}(s)) = \mathrm{E}\big(I_{\Phi_i}(\varphi(t))e(t)I_{\Phi_i}(\varphi(s))\big)\mathrm{E}\big(e(s)\big) = 0,$$

i.e., that $\tilde{e}(t)$ and $\tilde{e}(s)$ are uncorrelated. Similarly, with

$$e^*(t) = I_{\Phi_i}(\varphi(t))(|e(t)| - \mathrm{E}(|e(t)|)),$$

$e^*(t)$ and $e^*(s)$ can be shown to be uncorrelated. Furthermore, the variances of $\tilde{e}(t)$ and $e^*(t)$ are finite. Hence, the version of the strong law of large numbers in Theorem 5.1.2 in Chung (1974) imply that (21a) and (21b) hold. The convergence of the series in (21c) follows from the strong law of large numbers for independent variables (Chung, 1974, Theorem 5.4.2).

For two fixed realizations of $\varphi(t)$ and $e(t)$ for which (21) holds, we can thus find an $N'(\varepsilon)$ such that

$$\left|\frac{1}{p_i N}\sum_{t\in T_i} e(t)\right| \le \varepsilon, \quad \forall i \in \{1,2,\dots,d\}, \tag{22a}$$

$$\frac{1}{p_i N}\sum_{t\in T_i} |e(t)| \le 2C, \quad \forall i \in \{1,2,\dots,d\}, \tag{22b}$$

$$\frac{1}{N}\sum_{t=1}^{N} e(t)^2 \le 2\sigma^2 \tag{22c}$$

for all $N > N'(\varepsilon)$. This result follows since the partitioning is finite for any ε.

The fourth and fifth assumption in the theorem imply that

$$f_N(\varphi(t)) \triangleq \theta_N^T\varphi(t) + \rho_N(t) = f_0(\varphi(t))$$

is a feasible choice of function in the optimization problem (15), either with $\rho_N(t) = r_0(\varphi(t))$ and $\theta_N = \theta_0$ or sometimes with some smaller θ_N and larger $\rho_N(t)$:s. Hence, we have

$$\frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{f}_N(\varphi(t)))^2 \le \frac{1}{N}\sum_{t=1}^{N}(y(t) - f_0(\varphi(t))^2 = \frac{1}{N}\sum_{t=1}^{N} e(t)^2,$$

which means that

$$\frac{1}{N}\sum_{t=1}^{N}(f_0(\varphi(t)) - \hat{f}_N(\varphi(t)))^2 \le \left|\frac{2}{N}\sum_{t=1}^{N} e(t)(f_0(\varphi(t)) - \hat{f}_N(\varphi(t)))\right|. \tag{23}$$

Note first that, by applying Cauchy-Schwarz inequality to the right hand side, we obtain that

$$\frac{1}{N}\sum_{t=1}^{N}(f_0(\varphi(t)) - \hat{f}_N(\varphi(t)))^2 \le \frac{4}{N}\sum_{t=1}^{N} e(t)^2.$$

Since f_0 is bounded by $C_{f_0} = \sup_{\varphi\in\Phi}|f_0(\varphi)|$ and (22c) holds for all $N > N'(\varepsilon)$, $\hat{f}_N(\varphi)$ as defined by (6) and (7) must be bounded too. Hence, we can choose a constant $C_{\hat{f}}$, such that for $N > N'(\varepsilon)$ we have $C_{\hat{f}} > \sup_{\varphi\in\Phi}|\hat{f}_N(\varphi)|$.

Since $N_i \to \infty$ for all i, there will be many $\varphi(t)$ in every set Φ_i in the ε-partitioning. Choose $t_i^* \in T_i$ and let

$$f_i = f_0(\varphi(t_i^*)),$$
$$\hat{f}_{N,i} = \hat{f}_N(\varphi(t_i^*)).$$

This means that for $t \in T_i$, it holds that

$$|f_0(\varphi(t)) - f_i| \leq \tilde{L}\varepsilon$$

and

$$|\hat{f}_N(\varphi(t)) - \hat{f}_{N,i}| \leq \tilde{L}\varepsilon.$$

Inserting this into the expression (23) gives

$$\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N}(f_0(\varphi(t)) - \hat{f}_N(\varphi(t)))^2 \\
&\leq \left|\frac{2}{N}\sum_{t=1}^{N} e(t)(f_0(\varphi(t)) - \hat{f}_N(\varphi(t)))\right| \\
&= \left|\frac{2}{N}\sum_{i=1}^{d}\sum_{t\in T_i} e(t)(f_0(\varphi(t)) - f_i + f_i - \hat{f}_N(\varphi(t)) + \hat{f}_{N,i} - \hat{f}_{N,i})\right| \\
&= \left|2\sum_{i=1}^{d} p_i\left(\left(\frac{1}{p_i N}\sum_{t\in T_i} e(t)\right)\left(f_i - \hat{f}_{N,i}\right)\right.\right. \\
&\quad \left.\left. + \left(\frac{1}{p_i N}\sum_{t\in T_i} e(t)(f_0(\varphi(t)) - f_i - \hat{f}_N(\varphi(t)) + \hat{f}_{N,i})\right)\right)\right| \\
&\leq 2\sum_{i=1}^{d} p_i\left(\varepsilon \max_i |f_i - \hat{f}_{N,i}| + \frac{1}{p_i N}\sum_{t\in T_i} |e(t)|2\tilde{L}\varepsilon\right) \\
&\leq C'\varepsilon \qquad \text{for } N > N'(\varepsilon),
\end{aligned}$$

where $C' = 2C_{f_0} + 2C_{\hat{f}} + 8\tilde{L}C$. Since ε and the realizations were arbitrary, (19) has been proven.

We will now prove that the result (19) implies the uniform convergence in (20). First, we will consider pointwise convergence. Consider arbitrary realizations of $\varphi(t)$ and $e(t)$. With probability one these realizations are such that the second assumption in the theorem is satisfied and that the convergence in (19) holds. Consider two arbitrary realizations where these limits hold and assume that $\hat{f}_N$ does not converge pointwise to $f_0(\varphi)$ on Φ, i.e., that there exists a φ_0 in Φ, a $\delta > 0$ and an infinite strictly increasing sequence of integers K_j, $j \in \mathbb{Z}_+$, such that

$$|\hat{f}_{K_j}(\varphi_0) - f_0(\varphi_0)| > \delta$$

for all j. Consider a $\delta/4\tilde{L}$-partitioning of Φ such that the two first assumptions are satisfied, i.e., a partitioning where

$$\varphi_1,\ \varphi_2 \in \Phi_i \Rightarrow |\varphi_1 - \varphi_2| < \frac{\delta}{4\tilde{L}}$$

and where $N_i/N \to p_i$ for all $i = 1, 2, \ldots, d$. Consider the set Φ_l that contains φ_0. For every φ in Φ_l, it holds that

$$\begin{aligned}|\hat{f}_{K_j}(\varphi) - f_0(\varphi)| &\geq |\hat{f}_{K_j}(\varphi_0) - f_0(\varphi_0)| - |\hat{f}_{K_j}(\varphi) - \hat{f}_{K_j}(\varphi_0)| - |f_0(\varphi_0) - f_0(\varphi)| \\ &\geq \delta - \tilde{L}\frac{\delta}{4\tilde{L}} - \tilde{L}\frac{\delta}{4\tilde{L}} = \frac{\delta}{2}. \end{aligned} \tag{24}$$

Moreover, from the second assumption in the theorem, it holds that $K_{j,l}/K_j \to p_l$ when $j \to \infty$, where $K_{j,l}$ is the number of $\varphi(t)$ in Φ_l when the total number of measurements is K_j. The convergence

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} (\hat{f}_N(\varphi(t)) - f_0(\varphi(t)))^2 = 0$$

implies that

$$\lim_{j\to\infty} \frac{1}{K_j} \sum_{t=1}^{K_j} (\hat{f}_{K_j}(\varphi(t)) - f_0(\varphi(t)))^2 = 0.$$

However, using (24), it follows that

$$\begin{aligned}\frac{1}{K_j} \sum_{t=1}^{K_j} (\hat{f}_{K_j}(\varphi(t)) - f_0(\varphi(t)))^2 &\geq \frac{1}{K_j} \sum_{t\in T_l} (\hat{f}_{K_j}(\varphi(t)) - f_0(\varphi(t)))^2 \\ &\geq \frac{1}{K_j} \sum_{t\in T_l} \frac{\delta^2}{4} = \frac{\delta^2 K_{j,l}}{4K_j}. \end{aligned}$$

Since

$$\frac{\delta^2 K_{j,l}}{4K_j} \to \frac{\delta^2 p_l}{4} > 0, \quad j \to \infty,$$

we have a contradiction. Thus, $\hat{f}$ must converge pointwise to f_0 on Φ.

It turns out that pointwise convergence gives uniform convergence in this case. Take an arbitrary $\tilde{\varepsilon} > 0$ and assume that $\hat{f}_N(\varphi)$ converges pointwise to $f_0(\varphi)$ on Φ. Select a finite number of points $\tilde{\varphi}_k$, $k = 1, 2, \ldots, d_{\tilde{\varepsilon}}$ in Φ such that for every point φ in Φ,

$$|\varphi - \tilde{\varphi}_k| < \frac{\tilde{\varepsilon}}{3\tilde{L}}$$

for some k. Choose an $N_{\tilde{\varepsilon}}$ such that for all k it holds that

$$|\hat{f}_N(\tilde{\varphi}_k) - f_0(\tilde{\varphi}_k))| < \frac{\tilde{\varepsilon}}{3}, \quad \forall N > N_{\tilde{\varepsilon}}.$$

Hence, for an arbitrary point φ in Φ, there is a k such that

$$\begin{aligned}|\hat{f}_N(\varphi) - f_0(\varphi)| &\leq |\hat{f}_N(\varphi) - \hat{f}_N(\tilde{\varphi}_k)| + |f_0(\tilde{\varphi}_k) - f_0(\varphi)| + |\hat{f}_N(\tilde{\varphi}_k) - f_0(\tilde{\varphi}_k)| \\ &< \tilde{L}\frac{\tilde{\varepsilon}}{3\tilde{L}} + \tilde{L}\frac{\tilde{\varepsilon}}{3\tilde{L}} + \frac{\tilde{\varepsilon}}{3} = \tilde{\varepsilon}, \quad \forall N > N_{\tilde{\varepsilon}}, \end{aligned}$$

where we have used that both f_0 and $\hat{f}_N$ satisfy a Lipschitz condition with Lipschitz constant $\tilde{L}$. Since $\tilde{\varepsilon}$ and φ were arbitrary it follows that $\hat{f}_N$ converges uniformly to f_0. □

Remark 3.1. The theorem is still true if $\{e(t)\}$ is a mixing process, independent of the process $\{\varphi(t)\}$, since the only thing that matters is that (21) holds.

From the fourth and fifth assumption in Theorem 3.1, it can be seen that the constant L in the QP problem must be an upper bound on the true Lipschitz constant for the nonlinear function r_0 and that $\tilde{L} = L + M_\theta$ must be an upper bound on the Lipschitz constant of the true predictor function f_0 for the consistency result to hold. These conditions are quite intuitive since the estimated function must be allowed to vary at least as much as the true function. However, it is interesting to see that the choice $\theta_N = \theta_0$ does not have to be a feasible point to (15) since the linear part of the system (1) can be modeled by the nonparametric function $\hat{r}_N$, provided L is large enough. However, although a too hard bound on θ_N might not ruin the consistency, an unnecessarily large L will give a less smooth function estimate with a finite number of measurements. One of the main benefits of including also a linear parametric term in the model structure is that the smoothness of the estimate of the nonlinear part in that case will not depend on how large the linear part of the system is.

The fact that the linear part of the nonlinear system can be described by the nonparametric nonlinear part of the model explains why Theorem 3.1 does not discuss consistency for the individual linear and nonlinear estimators. In general, these estimators are not consistent since the separation of the system into a linear and a nonlinear part is not unique if a too large Lipschitz constant L is used. However, when $m_{\theta,i} = 0$, it follows that $\hat{r}_N$ converges uniformly to r_0.

Some further properties of the proposed mixed parametric and nonparametric method will be discussed in the examples in the next section.

4 Examples

The previously presented method for combined parametric and nonparametric estimation of NARX systems has been used in a couple of numerical examples. The first example concerns identification of a static nonlinearity.

Example 4.1

Consider the system

$$y(t) = 0.4u(t) + r_0(u(t)) + e(t), \tag{25}$$

where both $u(t)$ and $e(t)$ are white noise processes and independent of each other. The input $u(t)$ has uniform distribution on the interval $[-10, 10]$ while the noise $e(t)$ is normally distributed such that its mean is zero and its variance is 25. The nonlinearity in this system is

$$r_0(u(t)) = \frac{40}{5 + |u(t)|}\left(\frac{u(t)}{1 + |u(t)|} - \frac{u(t) - 3}{1 + |u(t) - 3|} - \frac{u(t) + 6}{1 + |u(t) + 6|} + \frac{3}{28}\right). \tag{26}$$

This function is Lipschitz continuous with $L_0 = 7.4$ and bounded since $|r_0'(x)| < 7.4$ and $|r_0(x)| < 3.1$, for all $x \in \mathbb{R}$.

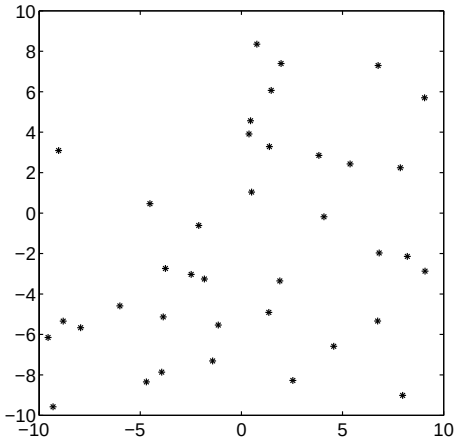


Figure 1: *The values of $y(t)$ plotted against $u(t)$ for the dataset with 40 measurements used in Example 4.1.*

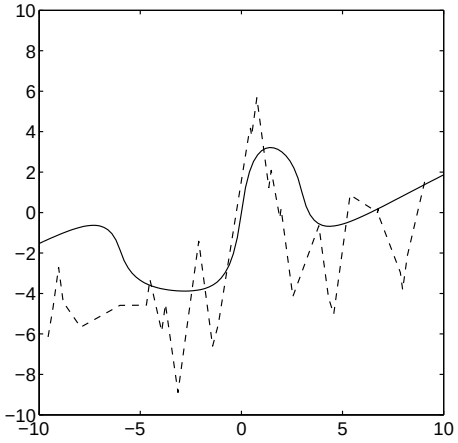
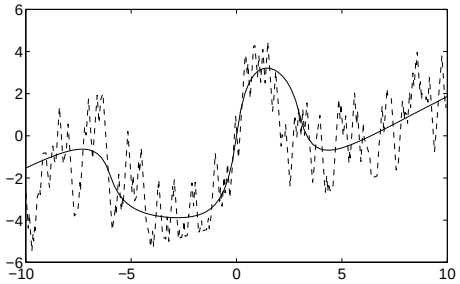


Figure 2: *The predictor function estimated from 40 measurements (dashed) and the true predictor function (solid) from Example 4.1.*

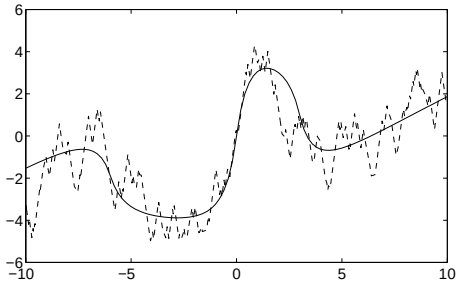
A small dataset consisting of 40 realizations of the input and output in (25) has been generated and is shown in Figure 1. Note that the shape of the nonlinear function is not obvious in this figure. The method (4) with $L = 7.4$ has been used with this dataset and linear interpolation has been used to construct $\hat{r}_N$. The resulting predictor function $\hat{y}_N(\varphi)$ is shown in Figure 2. From this figure, it seems that the function estimate has managed to pick up some key features of the true function, despite the small number of measurements.

In this case, the L value used in the method is equal to L_0. In a more realistic example, the true Lipschitz constant would typically be unknown. An alternative would then be to divide the dataset into estimation data and validation data and try different values of L. By evaluating the predictor (7) on the validation data for different choices of Lipschitz constant, it would be possible to find a good choice of L.

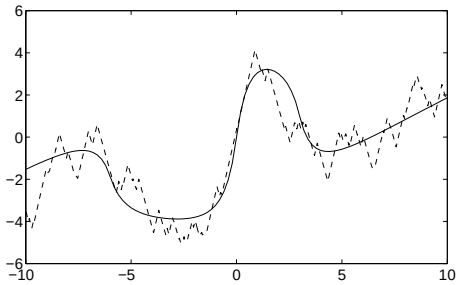
A larger dataset consisting of 500 realizations of the input and output in (25)



(a) $L = 15$



(b) $L = 7.4$



(c) $L = 4$

Figure 3: *The predictor function estimated from 500 measurements for three choices of L (dashed) and the true predictor function (solid) from Example 4.1.*

has also been generated and a couple of models have been estimated using an extended version of (4) where bounds $\pm\rho_N(t) \leq 4$ have been added. One model was estimated using $L = 15$ and the resulting predictor function is shown in Figure 3a. The choices $L = 7.4$ and $L = 4$ gave the results shown in Figure 3b and 3c, respectively. From these figures, it seems that the function estimates contain no significant systematic errors and that a larger value of L gives more variations. Note that for $L = 4$, the true function r_0 is not a feasible solution to the identification problem. However, the obtained function estimate gives a rather good approximation of r_0 anyway.

In the case with $L = 15$, the obtained estimate of the linear regression parameter $\theta_0 = 0.4$ was $\hat{\theta}_N = 0.31$ while $L = 7.4$ gave $\hat{\theta}_N = 0.33$ and $L = 4$ gave $\hat{\theta}_N = 0.39$. Using the same dataset but with a completely linear model, the least-

squares method gave an estimate $\hat{\theta}_{\mathrm{LS}} = 0.23$. Hence, it seems that including a bounded nonlinear Lipschitz continuous term in the model sometimes can improve the estimate of the linear part.

The method (4) combined with the interpolation (6) has also been used on a NARX system where the regression vector consists of two past output components and one input component. The results of this numerical experiment are described in the following example.

Example 4.2

Consider the following NARX system:

$$\begin{aligned} y(t) = &-y(t-1) - 0.2y(t-2) + u(t-1) \\ &+ \arctan(u(t-1) + y(t-1)) + \sin(y(t-2)) + e(t). \end{aligned} \tag{27}$$

This system can be viewed as being composed by a linear part $\theta_0^T \varphi(t)$ (with $\theta_0 = (1\ 0.2\ 1)^T$) and a nonlinear part $r_0(t)$ with Lipschitz constant $L_0 = \sqrt{3}$. Furthermore, $|r_0(t)| \leq \pi/2 + 1$. The noise terms are independent, normally distributed variables with unit variance.

The system has been estimated using an estimation dataset of 500 samples generated from $u(t) \in N(0, 4)$. Three Lipschitz constants have been tried: $L = 4$, $L = \sqrt{3}$ and $L = 1.4$, together with the upper bound on $r_0(t)$. The obtained models have been evaluated on a validation dataset of 500 samples generated under the same conditions as the estimation data. As quality measure, the fit has been calculated according to

$$\left(1 - \sqrt{\frac{\sum_t (y(t) - \hat{y}_N(\varphi(t)))^2}{\sum_t (y(t) - \bar{y})^2}}\right) \cdot 100\%, \tag{28}$$

where $\hat{y}_N(\varphi(t))$ is the output value predicted by the model and $\bar{y}$ is the arithmetic mean of $(y(t))_{t=1}^N$.

The results are given in Table 1. As comparison, a linear ARX model has also been identified. Furthermore, the fit has been calculated for a one step ahead predictor using the true parameter values and nonlinearities. Clearly, the NARX models outperform the linear ARX model. They also get rather close in performance to the true model. Note that the NARX model with a "too small" Lipschitz constant performs best. The reason for this is that in the region where data is available, we can decrease the Lipschitz constant of the nonlinear part by "tilting it" and properly adjust the linear part of the model.

5 Discussion

As mentioned previously, it is easy to incorporate various kinds of prior knowledge into the identification problem. In fact, we can regard the presented ap-

Table 1: *Fits for the estimated models in Example 4.2.*

Model	Fit (validation data)
NARX, $L = 1.4$	69.625
NARX, $L = \sqrt{3}$	69.050
NARX, $L = 4$	66.231
ARX	63.743
True model	72.039

proach as a special instance of the more general identification problem

$$\begin{array}{ll} \underset{\theta_N,\rho_N}{\text{minimize}} & \frac{1}{N}\sum_{t=1}^{N}(y(t)-\theta_N^T\varphi(t)-\rho_N(t))^2 \\ \text{subject to} & A\begin{pmatrix}\rho_N\\ \theta\end{pmatrix} \preccurlyeq b. \end{array} \tag{29}$$

This is still a convex QP problem. An interesting special case of (29) is

$$\begin{array}{ll} \underset{\theta_N,\rho_N}{\text{minimize}} & \frac{1}{N}\sum_{t=1}^{N}(y(t)-\theta_N^T\varphi(t)-\rho_N(t))^2 \\ \text{subject to} & |\rho_N(t)| \le M \\ & \forall t \in \{1,2,\ldots,N\}. \end{array} \tag{30}$$

It turns out that minimizing (30) gives exactly the same linear part as using an *ε-insensitive norm* for identification of ARX models, i.e.,

$$\underset{\theta_N}{\text{minimize}} \ \frac{1}{N}\sum_{t=1}^{N}\left|y(t)-\theta_N^T\varphi(t)\right|_\varepsilon^k \tag{31}$$

with

$$|x|_\varepsilon = \begin{cases} 0 & |x| \le \varepsilon \\ |x|-\varepsilon & |x| > \varepsilon \end{cases}$$

and with $k = 2$ and $\varepsilon = M$. This norm (or the corresponding norm with $k = 1$) is often used in support vector machines (Vapnik, 1998), and similar approaches are also used in robust adaptive control (Peterson and Narendra, 1982). To see the equivalence between (30) and (31), define

$$\bar{r}(t,\theta) = \begin{cases} M & y(t)-\theta^T\varphi(t) > M, \\ y(t)-\theta^T\varphi(t) & -M \le y(t)-\theta^T\varphi(t) \le M, \\ -M & y(t)-\theta^T\varphi(t) < -M. \end{cases}$$

Then we can write (31) as

$$\underset{\theta_N}{\text{minimize}} \ \frac{1}{N}\sum_{t=1}^{N}\left|y(t)-(\theta_N^T\varphi(t)+\bar{r}(t,\theta_N))\right|^k.$$

On the other hand it is easy to see that, for a given θ, the minimum of (30) is obtained precisely when $\rho_N(t) = \bar{r}(t,\theta)$. Since $\bar{r}(t,\theta)$ automatically has a magnitude not greater than M, the desired equivalences follow.

The advantage with using the formulation (30) instead of (31) is that the explicit representation of ρ_N again makes it possible to combine different types of requirements on the nonlinearity, just as was done in Example 4.2.

Instead of assuming a nonlinearity in the system, we can also interpret the terms $\hat{\rho}_N$ as estimates of deterministic noise terms ρ_0. These could for instance be bounded (unknown but bounded noise) like in (30) or satisfy a Lipschitz condition as in (4). Another option would be that their variation over time could be bounded, i.e.,

$$|\rho_N(t+1) - \rho_N(t)| \leq L_t.$$

6 Conclusions

In this paper, NARX systems that can be written as the sum of a linear ARX part and a nonlinear, Lipschitz continuous, NARX part have been studied. It has been shown that a model with a linear, parametric ARX part and a nonparametric NARX part of such an NARX system can be estimated by solving a quadratic programming problem. A novel proof of the consistency of this method has been presented. It should be noted that the consistency does not rely on knowledge of the true Lipschitz constant L_0. In fact, the only knowledge necessary is an upper bound of L_0. The tighter the upper bound, however, the faster convergence to the true function we can expect. The examples indicate that the method is fairly robust to incorrect values of L.

The examples also show that the introduction of a nonlinear term in the model sometimes can improve the estimate of the linear ARX term. Furthermore, the described method can produce NARX models that can predict the output in a validation dataset much better than an ARX model. The method can be useful also when the dataset is relatively small.

References

D. Bertsimas, D. Gamarnik, and J.N. Tsitsiklis. Estimation of time-varying parameters in statistical models: An optimization approach. *Machine Learning*, 35(3):225–245, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Xiaohong Chen, Jeffrey Racine, and Norman R. Swanson. Semiparametric ARX neural-network models with an application to forecasting inflation. *IEEE Transactions on Neural Networks*, 12(4):674–683, 2001.

K. L. Chung. *A Course in Probability Theory*. Academic Press, New York, second edition, 1974.

Martin Enqvist and Lennart Ljung. LTI approximations of slightly nonlinear systems: Some intriguing examples. In *Proc. NOLCOS 2004 - IFAC Symposium on Nonlinear Control Systems*, Stuttgart, Germany, 2004.

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

A. Garulli, A. Tesi, and A. Vicino, editors. *Robustness in Identification and Control*. Lecture Notes in Control and Information Sciences. Springer-Verlag, 1999.

Nancy E. Heckman. Minimax estimates in a semiparametric model. *Journal of the American Statistical Association*, 83(404):1090–1096, 1988.

L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, New Jersey, second edition, 1999.

P. M. Mäkilä. LTI modelling of NFIR systems: near-linearity and control, LS estimation and linearization. *Automatica*, 41(1):29–41, 2005.

B. B. Peterson and K. S. Narendra. Bounded error adaptive control. *IEEE Transactions on Automatic Control*, 27(6):1161–1168, 1982.

R. Pintelon, J. Schoukens, W. Van Moer, and Y. Rolain. Identification of linear systems in the presence of nonlinear distortions. *IEEE Transactions on Instrumentation and Measurement*, 50(4):855–863, 2001.

Jacob Roll, Alexander Nazin, and Lennart Ljung. Nonlinear system identification via direct weight optimization. *Automatica*, 41(3):475–490, 2005.

J. Schoukens, T. Dobrowiecki, and R. Pintelon. Parametric and nonparametric identification of linear systems in the presence of nonlinear distortions-a frequency domain approach. *IEEE Transactions on Automatic Control*, 43(2): 176 – 190, 1998.

J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2005-12-22

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English

☐ _____

Rapporttyp
Report category

☐ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☒ Övrig rapport
☐ _____

ISBN
—

ISRN
—

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

URL för elektronisk version

http://www.control.isy.liu.se

LiTH-ISY-R-2721

Titel
Title

Consistent Nonparametric Estimation of NARX Systems Using Convex Optimization

Författare
Author

Jacob Roll, Martin Enqvist, Lennart Ljung

Sammanfattning
Abstract

In this paper, a nonparametric method based on quadratic programming (QP) for identification of nonlinear autoregressive systems with exogenous inputs (NARX systems) is presented. We consider a mixed parametric/nonparametric model structure. The output is assumed to be the sum of a parametric linear part and a nonparametric Lipschitz continuous part. The consistency of the estimator is shown assuming only that an upper bound on the true Lipschitz constant is given. In addition, different types of prior knowledge about the system can easily be incorporated. Examples show that the method can give accurate estimates also for small data sets and that the estimate of the linear part sometimes can be improved compared to the linear least squares estimate.

Nyckelord
Keywords

System identification, NARX systems, nonparametric models