NARX Models: Optimal Parametric Approximation of Nonparametric Estimators

Giancarlo Ferrari-Trecate†, Giuseppe De Nicolao‡
† Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Via Ferrata 1, I-27100 Pavia
Tel. +39-0382-505484, Fax +39-0382-505373. E-mail: {denicola,ferrari}@conpro.unipv.it

Abstract
Bayesian regression, a nonparametric identification technique with several appealing features, can be applied to the identification of NARX (nonlinear ARX) models. However, its computational complexity scales as $O(N^3)$ where $N$ is the data set size. In order to reduce complexity, the challenge is to obtain fixed-order parametric models capable of approximating accurately the nonparametric Bayes estimate avoiding its explicit computation. In this work we derive, optimal finite-dimensional approximations of complexity $O(N^2)$ focusing on their use in the parametric identification of NARX models.

Keywords: NARX models, nonparametric identification, parametric identification, Bayesian estimation, neural networks, Gaussian processes.

1 Introduction
In recent years a large stream of research focused on the black-box identification of nonlinear systems via NARX (Nonlinear Auto Regressive eXogenous) models of the type

$$y_i = f(y_{i-1}, y_{i-2}, \ldots, y_{i-n_y}, u_{i-1}, \ldots, u_{i-n_u}) + \epsilon_i, \quad (1)$$

where $i = 1, 2, \ldots, N$, $u_i$ and $y_i$ denote the scalar input and output, respectively; the integers $n_y, n_u$ are the maximum lags of past outputs and inputs entering the model, and the additive measurement errors $\epsilon_i$ are uncorrelated zero-mean Gaussian random variables with $\text{Var}[\epsilon_i] = \sigma^2$.

The success of NARX models is due both to their capability of capturing nonlinear dynamics and the availability of identification algorithms with a reasonable computational cost [2], [4]. Concerning the generality of such models, in [14] it is shown, under mild assumptions, that any finite-dimensional nonlinear system admits an input-output NARX representation, at least locally.

Letting,

$$x_i = [y_{i-1}, y_{i-2}, \ldots, y_{i-n_y}, u_{i-1}, \ldots, u_{i-n_u}]', \quad (2)$$

we have that

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, N \quad (3)$$

and the NARX identification problem amounts to reconstructing the nonlinear map $f(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$, $n = n_y + n_u$, from the noisy training set $D_N = \{(x_i, y_i), i = 1, 2, \ldots, N\}$. This is a regression problem that can be solved via parametric or nonparametric techniques.

In parametric regression, least squares minimization is used to perform identification within a given class of models, e.g. radial basis function [4], multilayer perceptrons (MLP), [14, 20], and polynomial models [2]. In the real world, however, it is not realistic to assume that the "true model" belongs to the given finite dimensional class. Therefore, to ensure a good approximation, the model class should be as wide as possible, but this raises the issue of overparametrization. A possible answer is to carefully design the model structure avoiding superfluous parameters, e.g. by resorting to pruning methods (for MLP models) or subset selection methods such as stepwise regression (for linear-in-parameter models) [4], although all these approaches are more or less heuristic.

The main alternative to parametric regression is offered by nonparametric methods. Among them, Bayesian regression describes the unknown map as a multidimensional stochastic process, whose statistics summarize the prior information that is available about the map. Recently, several contributions focused on the use of Gaussian processes since their flexibility allows one to describe a very wide class of maps, while remaining simple enough to be analytically tractable [24, 17]. The model is estimated by computing the posterior expectation given the training set, which in the Gaussian case amounts to solving and algebraic linear system (see Section 2). Bayesian regression is a nonparametric method because the number of model parameters coincides with the number of the data. Moreover, overfitting is avoided by the use of prior information.

Beside their sound statistical grounding, Bayesian regression admits an interesting interpretation in terms of Neural Networks. In fact, the Bayes estimate is a regularization network [12], a kind of RBF-like net-
work stemming from the solution of a Tychonov regularization problem. A remarkable property of regularization networks is that, under mild assumptions, they are consistent estimators [21, 5, 6] in infinite-dimensional Hilbert spaces. This means that they are much more flexible than parametric methods since, as the number of the data goes to infinity, they can asymptotically reconstruct every possible map that is not too irregular.

The main drawback of the Bayes estimate is its computational complexity amounting to \( O(N^3) \) operations, where \( N \) is the size of the dataset. This makes the procedure unsuitable for regression problems with large datasets. The aim of this paper is to derive optimal \( m \)-dimensional approximations of the Bayes estimate with \( m \ll N \). In Section 3, after recalling the main existing results, we show that the optimal approximation of the Bayes estimate within quite general \( m \)-dimensional classes can be computed in \( O(n^2) \) operations. Here "optimal" refers to the fact that such approximation minimizes the expected prediction error on a test set within the chosen finite-dimensional class. Formulas for the computation of the generalization error of the approximators are also provided. The use of such estimators for the identification of NARX models is discussed in Section 4.

2 Bayesian identification

In this Section, we consider the generic regression problem

\[
y_i = f(x_i) + \varepsilon_i,
\]

where \( f(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R} \). In Bayesian regression a statistical distribution (the so-called prior) for the unknown function \( f(\cdot) \) must be assigned. In particular, we model \( f(\cdot) \) as a zero-mean \( n \)-dimensional Gaussian process independent of the errors \( \varepsilon_i \) and with autocovariance function \( \mathbb{E}[f(\xi_1)f(\xi_2)] = C(\xi_1, \xi_2) \).

By properly choosing the autocovariance function, it is possible to encode information about the expected smoothness or the stationarity/nonstationarity of \( f \) [9, 17, 8]. For what concerns the training set, we assume that the \( x \)-points are sampled according to a probability density \( p(x) \). In any case, if the sampling schedule is deterministic it is possible to replace \( p(x) \) with the so-called fraction of time distribution of the \( x_k \) sequence [11, 6] without compromising the results of the paper. The Bayes estimate \( \hat{f}(x) \) is the posterior expectation given the training set \( D_N \), which admits the closed-form expression [23]

\[
\hat{f}(t) = \mathbb{E}[f|D_n](t) = \sum_{k=1}^{N} \theta_k C(x, x_k)
\]

where \( \theta_k \) is the Kronecker symbol and \( \bar{y} = (y_1 \ldots y_N)' \). Apparently, the Bayes estimate is a linear combination of the \( N \) regressors \( C(x, x_k) \). Since the number of regressors is not fixed a priori, but scales with the size of the data set, \( \hat{f} \) is a nonparametric estimate of \( f \). Note that \( \hat{f} \) does not depend on the density \( p(x) \).

The estimate \( \hat{f}(x) \) admits also an interesting interpretations in terms of neural networks. If \( C(x, x_k) \) is radially symmetric (i.e. \( C(\xi_1, \xi_2) = C(||\xi_1 - \xi_2||) \)), the Bayes estimate is just a Radial Basis Function neural network whose weights \( \theta \) can be computed as \( \theta = H^{-1}\bar{y} \). In the general case the Bayes estimate is a Regularization Network (RN) [12], a class of neural networks arising from Tychonov regularization theory. More specifically, RNs are obtained from the solution of the regularized problem

\[
f_{\gamma,N}(x) = \arg\min_{f \in \mathcal{Q}_C} \sum_{k=1}^{N} (y_k - f(x_k))^2 + \gamma\|f\|_{\mathcal{Q}_C}^2,
\]

where \( \|f\|_{\mathcal{Q}_C}^2 \) is the norm of the suitable Hilbert Space \( \mathcal{Q}_C \) defined by the autocovariance function \( C(\xi_1, \xi_2) \) (see [1, 22] for a thorough characterization of such spaces), and \( \gamma > 0 \) is the so-called regularization parameter.

Usually, the term \( \|f\|_{\mathcal{Q}_C}^2 \) can be viewed as a Sobolev norm that penalizes the energy of \( f \) and some of its derivatives [9]. The parameter \( \gamma \) is a tuning parameter that controls the balance between data-fit and smoothness of the solution. For instance, if \( \gamma \) is too small the estimate fits the training set very well at the cost of poor generalization performances. Conversely too large a \( \gamma \) limits the flexibility of the network which will yield an excessively smooth estimate that does not explain the data. The tuning of \( \gamma \) can be done in an automatic way by resorting to well-known algorithms such as Maximum Likelihood [15, 7] and Generalized Cross Validation [22]. The Bayes estimate and RNs are equivalent in the sense that \( \hat{f}(t) = f_{\gamma,N}(t) \) if \( \gamma = \sigma^2 \) [22].

Hereafter we are interested in approximating \( \hat{f} \) within the \( m \) dimensional space

\[
\mathcal{H} = \left\{ q(x) : q(x) = \sum_{i=1}^{m} \alpha_i \psi_i(x), \alpha_i \in \mathbb{R} \right\}
\]

where the \( \psi_i(x) \) are arbitrary functions. In order to accomplish this goal, we introduce two optimality criteria.
Definition 1 Let $s(x)$ be a predictor that uses only information from $D_n$ and gather the training $x$-data in the set $\bar{x} = \{x_1, \ldots, x_N\}$. The $\bar{x}$-error and the generalization error of $s(x)$ are respectively defined as

$$E_s(N, \bar{x}) = \mathbb{E}_{x, x'} \left[ (y^* - s(x^*))^2 \right]$$

$$E^g_s(N) = \mathbb{E}_\bar{x} \left[ E_s(N, \bar{x}) \right],$$

where $y^* = f(x^*) + e^*$, $x^*$ is an independent test point sampled with probability $p(x)$ and $e^*$ is the corresponding measurement error. An estimator $s^*(x)$ belonging to a class $\mathcal{H}$ is said "$\bar{x}$-optimal" or simply "optimal" if, respectively, $E_{s^*}(N, \bar{x}) < E_s(N, \bar{x})$ or $E^g_{s^*}(N) < E^g_s(N)$, for all $s(x) \in \mathcal{H}$ and all data sets $\bar{x}$.

Note that $\bar{x}$-optimality means optimality for any possible set $\bar{x}$ of data points. Obviously, if $s^*(x)$ is $\bar{x}$-optimal it is also optimal. The generalization error is a widely used measure of the mean performance of an estimator. It is basically derived through a two-step averaging: first with respect to a new test point $(x^*, y^*)$ thus evaluating the expected predictive performance, and then with respect to the training set. It follows that $E^g_s(N)$ depends only on the data set size $N$ but not on the particular test point or training set used.

We conclude the section by introducing the Mercer-Hilbert decomposition of the autocovariance function $C(x, s)$. Let $L^2_\mathbb{P}$ denote the space of square integrable functions with respect to the measure $d\mathbb{P}(x)$ induced by the probability density $p(x)$. Since each autocovariance function is symmetric, and positive-semidefinite (i.e. $\sum_{i=1}^n \sum_{j=1}^m a_ia_jC(x_i, x_j) \geq 0$, $\forall a_i \in \mathbb{R}$, $\forall x_i \in \mathbb{R}^n$, $\forall n \in \mathbb{N}_0$), if $C(\cdot, \cdot)$ is continuous and $C(s, x) \in L^2_\mathbb{P}$, then, by the theorems of Mercer, Hilbert, and Schmidt [19], the positive operator $\mathcal{C}$, defined by

$$\left[ \mathcal{C}f \right](x) = \int_{\mathbb{R}^2n} C(x, s)f(s)d\mathbb{P}(s), \quad f \in L^2_\mathbb{P}$$

has an $L^2_\mathbb{P}$-complete orthonormal system of (countable) eigenfunctions $\{\phi_i\}_{i=1}^{+\infty}$ and corresponding eigenvalues $\{\lambda_i\}_{i=1}^{+\infty}$, with $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ and $\sum_{i=1}^{+\infty} \lambda_i < +\infty$. Moreover $C(s, x)$ admits the uniformly convergent expansion

$$C(x, s) = \sum_{i=1}^{+\infty} \lambda_i \phi_i(x)\phi_i(s). \quad (6)$$

3 Finite Dimensional Approximations

The first result concerning the approximation of $f$ in a generic space $\mathcal{H}$ was due to Zhu and co-workers [25]. They showed that the optimal approximation can be computed by projecting the Bayes estimate on $\mathcal{H}$ according the $L^2_\mathbb{P}$ norm. This result, although remarkable from the theoretical point of view, suffers from two main drawbacks: first, the Bayes estimate must be obtained in order to compute the approximation; second, the projection of the Bayes estimate involves the evaluation of $n$-dimensional integrals. Therefore, the projection method tends to be computationally demanding if $n > 1$.

The use of finite-dimensional models leads naturally to the search for optimal $m$-dimensional approximating space. In [26] it was shown that, in the asymptotic regime (i.e. when the number of the data goes to infinity), one should use $\mathcal{H} = \text{Span} \{\phi_i(x), i = 1, \ldots, m\}$ where $\phi_i(x)$ are the first $m$ eigenfunctions of the covariance function $C(\xi_1, \xi_2)$. Moreover, the closed formula for the optimal approximator (called Projected Bayes Regression - PBR) was derived, showing that it can be computed directly from the dataset in $O(N)$ operations. However, PBR is only asymptotically optimal. In [10] an approximator was proposed that has the same structure of PBR but is optimal even for finite datasets. Moreover the complexity of this approximator scales with $O(N^2)$. A common feature of the last two approximations, is that one must use the first $m$ eigenfunctions of the covariance function. When they are not known (as often happens since they depend on the density $p(x)$) one must find the solutions $\phi_i$ of the functional equations $C(\cdot, \cdot)\phi_i = \lambda_i \phi_i$. Again, if $n > 1$ this task may be computationally prohibitive.

The aim of this paper is to derive optimal approximators by using a generic approximating basis $\psi_1(x), \ldots, \psi_m(x)$ instead of the one made by the covariance eigenfunctions. Without loss of generality, we assume that the $\psi_i(\cdot)$ functions are orthonormal with respect to the measure $d\mathbb{P}(x)$ induced by the probability density $p(x)$. Practically, one may choose the first $m$ functions of an $L^2$ orthonormal system and then divide them by $\sqrt{p(x)}$ (if $p(x)$ does not vanish a.e.). When it is not known in advance, the density $p(x)$ can be estimated as discussed in Section 4.

Before stating the main result, we recall a preliminary Lemma.

Lemma 1 ([18]) Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times r}$, $A > 0$. Then it holds that

$$\inf_{Z \in \mathbb{R}^{n \times n}} \text{Tr} \left[ (ZA'Z' - ZB - B'Z') \right] = \text{Tr} \left[ -B' A^{-1} B \right]$$

and the infimum is achieved for the matrix $Z^* = B' A^{-1}$.

We look for the optimal approximators in the classes $\mathcal{H}_1$.
defined as follows
\[ \mathcal{H}_1 \triangleq \left\{ g(x) = k'(x)L\bar{y}|L \in \mathbb{R}^{m \times n} \right\} \]
\[ \mathcal{H}_2 \triangleq \left\{ h(x) = k'(x)F\Psi \bar{y}|F \in \mathbb{R}^{m \times m} \right\} \]
\[ k(x) = \begin{bmatrix} \psi_1(x) & \cdots & \psi_m(x) \end{bmatrix}' \]
\[ \Psi = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_m(x_1) \\ \vdots & \ddots & \vdots \\ \psi_1(x_n) & \cdots & \psi_m(x_n) \end{bmatrix} \]

The main reason to work with the classes \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) is that their parametrized structure will allow us to derive the closed-form expression of the optimal approximators. However they do not coincide with the linear space \( \mathcal{H} \) since the following inclusions hold true:
\[ \mathcal{H} \supsetnot \mathcal{H}_1 \supsetnot \mathcal{H}_2. \]

**Theorem 1** The approximators \( g^o(x) \in \mathcal{H}_1 \) characterized by \( L = L^o = T'H^{-1} \) and \( h^o(x) \in \mathcal{H}_2 \) characterized by \( F = F^o = T'\Psi (\Psi' \Psi)^{-1} \), where
\[ T = \begin{bmatrix} T_1 & \cdots & T_m \end{bmatrix}, \]
\[ r(x^*) = \begin{bmatrix} r_1(x^*) & \cdots & r_N(x^*) \end{bmatrix}' \]
are \( \bar{v} \)-optimal \( \forall N \geq m \). Moreover,
\[ E_{g^o}(N, \bar{x}) = \sum_{i=1}^{+\infty} \lambda_i + \sigma^2 - \text{Tr} \left[ T'H^{-1}T \right], \]
\[ E_{h^o}(N, \bar{x}) = \sum_{i=1}^{+\infty} \lambda_i + \sigma^2 - \text{Tr} \left[ T'\Psi (\Psi' \Psi)^{-1} \Psi' \Psi \right]. \]

**Proof:** We start considering the approximator \( g^o(x) \). In view of the \( \bar{v} \)-error definition, it holds that
\[ E_{g^o, \bar{v}} \left[ (y^* - k'(x^*)L\bar{y})^2 \right] = \sigma^2 + C(x^*, x^*) + k'(x^*)LHL'k(x^*) - 2k'(x^*)Lr(x^*) \]
\[ = \sigma^2 + C(x^*, x^*) + \text{Tr} \left[ LHL'k'(x^*)k'(x^*) - 2Lr(x^*)k'(x^*) \right]. \]

To compute the \( \bar{v} \)-error, we compute the mean of (8) with respect to \( x^* \). Note that \( E_{x^*} \left[ k'(x^*)k'(x^*) \right] = I_m \) and, from (6), it holds that
\[ \int_{\mathbb{R}^d} C(s, s)dP(s) = \sum_{i=1}^{+\infty} \lambda_i. \]

Then, from (8), we obtain
\[ E_{g^o}(N, \bar{x}) = \sum_{i=1}^{+\infty} \lambda_i + \sigma^2 + s(L) \]
\[ s(L) = \text{Tr} \left[ LHL - 2LT \right], \]
and the next step is to minimize \( s(L) \) with respect to the matrix \( L \). By applying Lemma 1 with \( B = T, A = H > 0, Z = L \), one obtains
\[ \arg \min_L s(L) = L^o = T'H^{-1}, \]
\[ \min_L s(L) = -\text{Tr} \left[ T'H^{-1}T \right] \]
that is the desired result.

Coming to the approximation \( h^o(x) \), we apply Lemma 1 with \( L = F\Psi \) and then perform minimization of \( s(F\Psi) \), w.r.t. the matrix \( F \). This can be done as before noting that \( \Psi' \Psi > 0 \) if \( N \geq m \).

The estimators \( g^o(x) \) and \( h^o(x) \) can be written as
\[ g^o(x) = \sum_{i=1}^{m} \psi_i(x)\theta_i^g, \]
\[ h^o(x) = \sum_{i=1}^{m} \psi_i(x)\theta_i^h \]
where
\[ \left[ \theta_1^g \cdots \theta_m^g \right]' = T'\bar{y}, \quad H\bar{y} = \bar{y} \]
\[ \left[ \theta_1^h \cdots \theta_m^h \right]' = T'\Psi \bar{y}, \quad \Psi' \Psi \bar{y} = \Psi \bar{y} \]
From (11) it is apparent that the computational complexity of \( g^o(x) \) is \( O(N^3) \), that is the cost for obtaining \( \bar{y} \). Therefore, the computation of \( g^o(x) \) is not cheaper than the derivation of \( \bar{f}(x) \). On the other hand, the complexity of \( h^o(x) \) is \( O(N^2m^3) \), where \( m \) does not scale with \( N \) and usually \( N \ll m \) (otherwise one would prefer to compute directly the Bayes estimate). In fact, the calculation of \( \bar{y} \) takes \( O(m^3) \) operations and the term \( O(N^2) \) arises from the multiplications in \( \Psi' \Psi \).

4 A new identification procedure for NARX models

Consider the NARX model (1). Hereafter we focus on the approximator \( h^o(x) \) derived in Section 3 in order to propose a new parametric identification algorithm for NARX systems. The procedure is summarized in the following steps (recall that \( x_k \) is defined as in (2)):

1. Choose a covariance prior \( C(\xi_1, \xi_2) \);
2. Choose an \( L^2 \) orthonormal set of functions \( \left\{ \hat{\psi_i}(x) \right\}_{i=1}^{m} \);
3. Estimate \( \bar{p}(x) \) (an approximation of \( p(x) \)) from the collected datapoints \( x_k, k = 1, 2, \ldots, N \);
4. Compute
\[ \psi_i(x) = \frac{\hat{\psi_i}(x)}{\sqrt{\bar{p}(x)}}, \quad i = 1, 2, \ldots, m; \]
5. Compute the matrix

\[ \hat{T} = \frac{1}{N} \sum_{i=1}^{N} r(x_k) k'(x_k), \]

where

\[ r(x) = [C(x, x_1) \cdots C(x, x_N)]', \]

\[ k(x) = [\psi_1(x) \cdots \psi_m(x)]'. \]

6. Calculate the approximator \( \hat{h}^o(x) \) in the class

\[ \mathcal{H}_2 = \{ \hat{h}(x) = k'(x) F \Psi \hat{g} | F \in \mathbb{R}^{m \times m} \}, \]

through the formula

\[ \hat{h}^o(x) = k'(x) \hat{T}' \Psi (\Psi' H \Psi)^{-1} \Psi' \hat{g}. \]

where the matrix \( \Psi \) is defined by formula (7).

Steps 2 and 4 constitute an affordable way to obtain a set of orthonormal functions \( \{ \psi_i(x) \}_{i=1}^m \). In step 4, an approximation \( \hat{T} \) of the matrix \( T \) is computed. In step 6, the formula for \( \hat{h}^o(x) \) (see Theorem 1) is used replacing the \( T \) matrix and the \( p(x) \) density with their approximations \( \hat{T}, \hat{p}(x) \).

Obviously, the accuracy of this algorithm improves as the number \( N \) of the collected datapoints grows. Note that, it is sensible to assume that \( N \) is 'large'; otherwise, one would resort to the exact computation of the Bayes estimate.

The estimated density \( \hat{p}(x) \) (step 3) can be obtained in various ways (see [3] for a good review of the most common algorithms for density estimation). We suggest using Gaussian mixture models, where \( \hat{p}(x) \) is modeled as

\[ \hat{p}(x) = \sum_{j=1}^{M} \frac{\alpha_j}{(2\pi \sigma_j^2)^{d/2}} \exp \left( -\frac{||x - \mu_j||^2}{2\sigma_j^2} \right). \] (13)

and \( M \) is the number of Gaussian functions entering the mixture. All the parameters \( \alpha_j, \sigma_j, \mu_j \), can be tuned using fast algorithms like Expectation-Maximization or stochastic estimation [3]. Note that, in view of (13), it is apparent that \( \hat{p}(x) \) does not vanish a.e. as is required in Step 3 of our algorithm. Moreover, the cost of evaluating \( \hat{p}(\cdot) \) at a point \( x \) is independent of the size of the training set since \( M \) does not scale with \( N \). Then, it follows that the computational complexity of steps 4 and 5 is \( O(N^2) \) and this allows us to conclude that the complexity of the overall algorithm scales as \( O(N^2) \).

5 Concluding remarks

In standard statistically-based identification theory, consistency results and error bounds are derived under the assumption that there exists a "true model" belonging to a finite-dimensional class of models within which identification is performed. Since this assumption is not realistic, in recent years several researches focused on the problem of finding the best approximation to the true model within a finite-dimensional class that does not includes the true model itself. A solution is possible only exploiting a priori information that can be of statistical [13], or deterministic type [16]. When the problem is approached according to a Bayesian paradigm, the optimal approximation is obtained by projecting the nonparametric Bayes estimate onto the finite-dimensional space of parametric models [25]. However, this would require the explicit calculation of the Bayes estimate whose complexity scales with the cube of the number \( N \) of data. In this paper, it has been shown that for certain class of parametric models the optimal approximator (the one minimizing the the expected prediction error on a test set) can be computed directly in \( O(N^2) \) operations. The resulting estimation scheme can be applied to various identification problems, and in particular we have considered the identification of NARX models where a nontrivial multidimensional map has to be estimated.

Acknowledgments

This paper has been partially supported by MURST Project "Identification and Control of Industrial Systems", and ICE-CNR.

References

[7] G. De Nicolao, G. Sparacino, and C. Cobelli. Non parametric input estimation in physiological sys-


