Kernel-based local order estimation of nonlinear nonparametric systems


Published in:
Automatica

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2015 Elsevier Ltd. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/ which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Kernel-based local order estimation of nonlinear nonparametric systems

Wenxiao Zhao\textsuperscript{a}, Han-Fu Chen\textsuperscript{a}, Er-wei Bai\textsuperscript{b}, Kang Li\textsuperscript{c}

\textsuperscript{a} Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
\textsuperscript{b} Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA
\textsuperscript{c} School of Electronics, Electrical Engineering and Computer Science, Queen’s University, Belfast, UK

Abstract

We consider the local order estimation of nonlinear autoregressive systems with exogenous inputs (NARX), which may have different local dimensions at different points. By minimizing the kernel-based local information criterion introduced in this paper, the strongly consistent estimates for the local orders of the NARX system at points of interest are obtained. The modification of the criterion and a simple procedure of searching the minimum of the criterion, are also discussed. The theoretical results derived here are tested by simulation examples.

1. Introduction

Consider a single-input–single-output (SISO) nonlinear autoregressive system with exogenous input (NARX),

\[ y_{k+1} = f(y_k, \ldots, y_{k+1-M}, u_k, \ldots, u_{k+1-M}) + \epsilon_{k+1}, \]

where \( u_k \) and \( y_k \) are the system input and output, respectively, \( \epsilon_k \) is the driven noise, \( M \) is the known upper bound of the true system order and \( f(\cdot) \) is the unknown function representing the system dynamics.

In recent years identification of system (1) has been an active research topic, estimating not only the nonlinear function \( f(\cdot) \) itself (Bai, 2010; Bai, Tempo, & Liu, 2007; Ninness & Henriksen, 2010; Roll, Nazin, & Ljung, 2005; Schon, Wills, & Ninness, 2011; Zhao, Chen, & Zheng, 2010; Zhao, Zheng, & Bai, 2013) but also the system orders (Autinl, Biey, & Haslerl, 1992; Bai, Li, Zhao, & Xu, 2014; He & Asada, 1993; Hong, Mitchell, & Chen, 2008; Rhodes & Morari, 1998; Roll, Lind, & Ljung, 2006). As far as the estimation of the nonlinear function \( f(\cdot) \) is of concern, the approaches can roughly be divided into two categories, the parametric approach (Ninness & Henriksen, 2010; Schon et al., 2011; You, accepted for publication; You, Xie, & Song, 2013; Zhou, Duan, & Lin, 2011; Zhou, Zheng, & Duan, 2011) and the nonparametric approach (Bai, 2010; Bai et al., 2007; Roll et al., 2005; Zhao et al., 2010, 2013), according to the description of \( f(\cdot) \). In the former, it is usually assumed that \( f(\cdot) = f(\cdot, \theta) \) with a known structure of \( f(\cdot) \) and an unknown parameter \( \theta \), and consequently identification of \( f(\cdot) \) is transformed into a parametric optimization problem for \( \theta \) while in the latter approach, it is often to estimate the values of \( f(\cdot) \) at the points of interest referred to as Model on Demand in the literature (see, e.g., Bai, 2010; Bai et al., 2007, Fan & Gijbels, 1996, Ninness & Henriksen, 2010, Roll et al., 2005; Zhao et al., 2010). The direct weight optimization (Roll et al., 2005), the local linear estimator (Bai, 2010) and its recursive version (Zhao et al., 2013), the stochastic approximation algorithm (Zhao et al., 2010) all belong to this class. Notice that most of the nonparametric identification algorithms are the weighted local average algorithms in a certain sense, and in order to derive the reliable estimates it requires to obtain the adequate measurements around the given points. In some applications, a global description of an unknown nonlinear system is too complicated both in structure and in dimension. This makes identification unreliable and the obtained model practically useless. Typical examples can be easily found in the fields of biology, atmospheres, geophysics, economy, engineering, communication, etc. An efficient and practical way is to split the task into a number of manageable pieces either in structure/dimension or in both. This is
the idea of local modeling including local polynomial modeling, a hot topic in statistics. This paper studies the problem of the order of the local modeling.

Over the last few decades considerable progress has been made on the order estimation as well as variable selection of linear stochastic systems. For example, the Akaike's information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) and their generalizations (Chen & Guo, 1991), the recursive algorithms (Chen & Zhao, 2010), the so-called LASSO (Zou, 2006), are a few among many others. But these approaches are not applicable to system (1) due to its nonparametric and nonlinear description. The order estimation for nonlinear systems has also been studied in recent years, e.g., Aulin et al. (1992), Bomberger and Seborg (1998), He and Asada (1993), Kennel, Brown, and Abarbanel (1992), Mao and Billings (2006), Peduzzi (1980), Rhodes and Morari (1998) and Roll et al. (2006). In Aulin et al. (1992) an approach to estimate the orders of the linearized nonlinear system is introduced. The so-called Lipschitz number approach and false nearest neighbors approach are proposed in He and Asada (1993) and Kennel et al. (1992), respectively, and successive research appeared in Bomberger and Seborg (1998), Ramdania, Castiesa, and Boucharab et al. (1992), and Rhodes and Morari (1998), etc. These two approaches do not identify the nonlinearity \( f(\cdot) \) itself, while estimating the orders. The methods in Aulin et al. (1992), He and Asada (1993), and Kennel et al. (1992) are however sensitive to the system noises, and, to the authors' knowledge, their convergence and consistency are unclear. The stepwise approach and the analysis of variance (ANOVA) approach are suggested in Peduzzi (1980) and Roll et al. (2006) based on hypothesis tests for the parameterized nonlinear systems. For these approaches a review is given in Hong et al. (2008). Note that the order estimation in the above papers is in a global sense, i.e., the true order is unique over the whole domain. In contrast to this, sometimes the true orders of a nonlinear system are not unique and may vary from point to point. To this end, let us consider examples given below.

Example (1): A piecewise linear system is defined by

\[
y_{k+1} = f_{i}(y_{k}, \ldots, y_{k+1-M}, u_{k}, \ldots, u_{k+1-M}) + \epsilon_{k+1},
\]

with

\[
f_{i}(y_{k}, \ldots, y_{k+1-M}, u_{k}, \ldots, u_{k+1-M}) = \begin{cases} a_{1}(y_{1} + \cdots + a_{p}(y_{k+1-p} + b_{1}(u_{k} + \cdots + b_{q}(u_{k+1-q}), \quad \text{if } [y_{k}, \ldots, y_{k+1-M}, u_{k}, \ldots, u_{k+1-M}] \in \mathcal{A}_{1}, \\ \vdots, \\ a_{s}(y_{1} + \cdots + a_{p}(y_{k+1-p} + b_{1}(u_{k} + \cdots + b_{q}(u_{k+1-q}) \in \mathcal{A}_{s}, \\ [y_{k}, \ldots, y_{k+1-M}, u_{k}, \ldots, u_{k+1-M}] \in \mathcal{A}_{s}, \end{cases}
\]

where \( \mathcal{A}_{i}, i = 1, \ldots, s \) is a partition of \( \mathbb{R}^{2M} \).

Example (2): The finite impulse response system is given by

\[
y_{k+1} = f_{i}(u_{k}, u_{k-1}, u_{k-2}) + \epsilon_{k+1},
\]

where

\[
f_{i}(u_{k}, u_{k-1}, u_{k-2}) = u_{k}u_{k-1}u_{k-2}, \text{ if } u_{k} > 1; u_{k}u_{k-1}, \text{ if } -1 \leq u_{k} \leq 1; \text{ and } u_{k}, \text{ if } u_{k} < -1.
\]

These two examples demonstrate a need for the local order estimation at points of interest. To the authors' knowledge, there has not much been done on this topic, though in Bai et al. (2014) a forward/backward approach was proposed. The numerical simulations seem to suggest that the forward/backward approach works well in terms of variable selection, but determination of the system order and its theoretical study remain open.

The contribution of the paper is as follows. First, a kernel-based local information criterion, for simplicity of reference, named as the local information criterion (LIC), is proposed for the local order estimation of system (1). Under moderate conditions, the estimates generated from LIC converge almost surely to the true local orders of system (1) at the points of interest. Second, a modification of LIC and a simple procedure of searching the minimum of LIC are suggested, and the strong consistency of the estimates is established as well.

The rest of the paper is arranged as follows. The LIC and the strong consistency of the estimates are given in Section 2. A modification of LIC is discussed in Section 3. Two simulation examples are given in Section 4 and some concluding remarks are addressed in Section 5. Some technical proofs are placed in the Appendix.

2. Local order estimation

2.1. Local information criterion for order estimation

We further introduce the following notations. Notice that the nonlinear function \( f(\cdot) \) in (1) is defined on \( \mathbb{R}^{2M} \). The regressor and the point of interest in \( \mathbb{R}^{2M} \) are denoted by \( \phi_{k}(M, M) \) and \( x^{*}(2M) \), respectively.

\[
\phi_{k}(M, M) = [y_{k} \cdots y_{k+1-M} u_{k} \cdots u_{k+1-M}]^{T},
\]

\[
x^{*}(2M) = [x_{1}^{*}, \ldots, x_{M}^{*}]^{T}.
\]

Similar to (4), for any fixed \( 1 \leq p \leq M \) and \( 1 \leq q \leq M \) let us define

\[
\phi_{k}(p, q) = [y_{k} \cdots y_{k+1-p} u_{k} \cdots u_{k+1-q}]^{T},
\]

\[
x^{*}(p, q) = [x_{1}^{*}, \ldots, x_{p}^{*}, x_{M+1}^{*}, \ldots, x_{M+q}^{*}]^{T}.
\]

From the examples given in the introduction, it is seen that the orders of nonlinear systems may be varying from point to point. This is a different picture from linear systems. The question is how to define and estimate the local order of \( f(\cdot) \) at the given \( x^{*}(2M) \) based on the observations \( [y_{k}, u_{k}]_{k\geq 1} \). A direct approach is to define the local order of \( f(\cdot) \) at \( x^{*}(2M) \) as the number of variables that contribute to the function value \( f(x^{*}(2M)) \). However, if the system order is defined in such a manner, it is difficult to choose the quantitative information based on which the algorithms estimating the local order can be designed, since \( f(\cdot) \) is nonlinear and nonparametric. On the other hand, it is clear that the function \( f(x) \) can be well approximated by a local linear model if \( x \) is close to \( x^{*}(2M) \), i.e.,

\[
f(x^{*}(2M)) = f(x^{*}(2M)) + \nabla f(x^{*}(2M))^{T}
\]

\[
\cdot [x(2M) - x^{*}(2M)] + O\left(\|x(2M) - x^{*}(2M)\|^{2}\right),
\]

\[
\forall \left\|x(2M) - x^{*}(2M)\right\| \leq \varepsilon \text{ for small enough } \varepsilon > 0. \text{ Denote the gradient of } f(\cdot) \text{ at } x^{*}(2M) \text{ by }
\]

\[
\nabla f(x^{*}(2M)) \triangleq \begin{bmatrix}
\frac{df}{dx_{1}^{*}} & \cdots & \frac{df}{dx_{M}^{*}} & \cdots & \frac{df}{dx_{M+q}^{*}}
\end{bmatrix}^{T} \in \mathbb{R}^{2M}\text{ if it exists. It is clear that if } f(x^{*}(2M)) \text{ depends only on } (p_{0} + q_{0}) \text{ variables, i.e.,}
\]

\[
x^{*}(2M) = x_{1}^{*}, \ldots, x_{p}^{*}, x_{M+1}^{*}, \ldots, x_{M+q}^{*}
\]

\[
\begin{align*}
&= f(x_{1}^{*}, \ldots, x_{p}^{*}) x^{T}(M - p_{0}), \\
x_{M+1}^{*}, \ldots, x_{M+q}^{*}, x^{T}(M - q_{0})
\end{align*}
\]
∀ x(M − p0) ∈ R^{M−p0} and ∀ x(M − q0) ∈ R^{M−q0}, then \frac{df}{dx^p} = 0 for i = p0 + 1, \ldots, M and M + q0 + 1, \ldots, 2M, i.e.,

\[ f(x^*(2M)) = \left[ \begin{array}{c} \frac{df}{dx_1} \cdots \frac{df}{dx_{p0}} 0 \cdots 0 \\ \frac{df}{dx_{p0+1}} \cdots \frac{df}{dx_{2M-q0}} 0 \cdots 0 \end{array} \right] \cdot \left( \begin{array}{c} M-p0 \\ M-q0 \end{array} \right).

From (8) and (10) it is seen that if we can find a local linear model of \( f(\cdot) \) at \( x^*(2M) \), then we can estimate the local order by determining the biggest p and q such that \( \frac{df}{dx_p} \neq 0 \), \( 1 \leq p \leq M \) and \( \frac{df}{dx_{M+q}} \neq 0 \), \( 1 \leq q \leq M \).

To this end, we further impose the following assumptions.

(A1) The finite upper bound M for orders (p, q) is known;

(A2) \( |x(x)| \leq c_1\|x\|^2 + c_2 \), \( x \in R^{2M} \) for some positive constants \( c_1 \), \( c_2 \) and p and f(\cdot) is twice differentiable at \( x^*(2M) \). Further, \( \frac{df}{dx_p} \neq 0 \) and \( \frac{df}{dx_{M+q}} \neq 0 \) for some \( p = 1, \ldots, M \) and \( q = 1, \ldots, M \).

**Definition 1.** The local order of \( f(\cdot) \) at \( x^*(2M) \) is defined as \((s_0,t_0)\), where

\[ s_0 = \max \left\{ p = 1, \ldots, M \mid \frac{df}{dx_p} \neq 0 \right\}, \quad t_0 = \max \left\{ q = 1, \ldots, M \mid \frac{df}{dx_{M+q}} \neq 0 \right\}.

It is natural to ask why \((s_0,t_0)\) rather than \((p_0,q_0)\) given in (9) is defined as the local order of \( f(\cdot) \) at \( x^*(2M) \)? Do we need to take the second order derivatives into consideration? By the Taylor expansion, we know that a local linear estimator approximates \( f(\cdot) \) at \( x^*(2M) \) well if \( x(2M) \in R^{2M} \) is close to \( x^*(2M) \) and the second order terms can be neglected. In this regard, it is reasonable to find the local order of \( f(\cdot) \) at \( x^*(2M) \) from its local linear approximations. On the other hand, it is clear that if \( \frac{df}{dx_p} \neq 0 \) and \( \frac{df}{dx_{M+q}} \neq 0 \) for all \( s_0 \leq p, q \leq t_0 \), then \((s_0,t_0)\) is \((p_0,q_0)\). But sometimes, the local order given by **Definition 1** is smaller than \((p_0,q_0)\). We next provide two examples to illustrate **Definition 1**.

**Example (iii):** For the linear system \( y_{k+1} = a_1 y_{k+1} + \cdots + a_{p0} y_{k+1-p0} + b_0 u_k + \cdots + b_{q0} u_{k+q0} + \varepsilon_{k+1} \) with \( a_0 \neq 0, b_0 \neq 0 \) we have \( f(x(2M)) = a_1 x_1 + \cdots + a_{p0} x_{p0} + b_0 x_{M+1} + \cdots + b_{q0} x_{M+q0} \). It is clear that \( \frac{df}{dx_{p0}} \neq 0, \frac{df}{dx_{M+q0}} \neq 0 \), and \( \frac{df}{dx_i} = 0, i = p_0 + 1, \ldots, M, q_0 + 1, \ldots, 2M \). Thus for this example the system order \((s_0,t_0)\) derived by **Definition 1** equals \((p_0,q_0)\), which is consistent with the linear system theory.

**Example (iv):** For the nonlinear system \( y_{k+1} = a y_k + b u_k + u_{k-1} + \varepsilon_{k+1} \) with \( a \neq 0, b \neq 0 \), we have \( f(x(4)) = f(x_1, x_2, x_3, x_4) = a x_1 + b x_2 + b x_3 + b x_4 \). At the fixed point \( x(4) = [0 1 0 1]^T \in R^4 \), it is clear that \( \frac{df}{dx(4)} = [0 0 0 0]^T \), and by the Taylor expansion \( f(x(4)) = f(x(4)) + a \frac{df}{dx_1} (x_1 - x_1) + b \frac{df}{dx_2} (x_2 - x_1) \) for all \( x(4) \) close to \( x(4) \). This implies that the local order at the given point should be \((s_0,t_0)\) = \((1,1)\).

Based on the above discussion the key step of our approach to estimate the local order is to find the local linear model of \( f(\cdot) \) at \( x^*(2M) \). In [Bai (2010)] and [Zhao et al. (2013)], the kernel function-based local linear estimator (KLE) and its recursive version (RLLE) are considered, which estimate the values of the nonlinear function at fixed points together with their gradients. Let us first reformulate the RLLE introduced in [Zhao et al. (2013)], on which the order estimation algorithm is essentially based. Notice that the RLLE in [Zhao et al. (2013)] is with known system orders, but here the orders (p, q) in the algorithm may vary in the set \( \{(p, q) : 1 \leq p \leq M, 1 \leq q \leq M\} \).

With the given order (p, q) and measurements \( \{u_k, y_k\}_{k=1}^N \), the RLLE estimate of \( f(\cdot) \) at time \( N + 1 \) is given by

\[
\begin{align*}
\theta_{N+1}(p, q) & = \left[ \theta_{N+1}(p, q) \quad \theta_{N+1}(q, p) \right]^T \\
& \leq \arg\min_{\theta_N(p, q)} \sum_{k=1}^N w_k(x^*(2M))(y_{k+1} - \theta_0(p, q)) \\
& - \theta_1(p, q)^T (\phi_k(p, q) - x^*(p, q))^2,
\end{align*}
\]

where the kernel function \( w_k(x^*(2M)) \) is given by

\[ w_k(x^*(2M)) = \frac{1}{b_k^2} \exp \left\{ - \frac{1}{2b_k^2} \right\} \]

Notice that \( \theta_{N+1}(p, q) = \left[ \theta_{N+1}(p, q) \quad \theta_{N+1}(q, p) \right]^T \) with the given order \( (p, q) \), \( \theta_{N+1}(p, q) \) serves as the estimate for \( f(x^*(M, M)) \) while \( \theta_{N+1}(q, p) \) for \( \phi(x^*(M, M)) \).

Set

\[ X_k(p, q) = \begin{bmatrix} 1 \end{bmatrix} \left[ \phi_k(p, q) - x^*(p, q) \right]. \]

By some simple manipulations, RLLE in (11) can be expressed by

\[ \theta_{N+1}(p, q) = \left( \sum_{k=1}^N w_k(x^*(2M))X_k(p, q)X_k(p, q)^T \right)^{-1} \times \left( \sum_{k=1}^N w_k(x^*(2M))X_k(p, q) \right)^T. \]

if the matrices \( \sum_{k=1}^N w_k(x^*(2M))X_k(p, q)X_k(p, q)^T \), \( N \geq 1 \) are nonsingular. Notice that by the matrix inverse lemma, \( \theta_{N+1}(p, q) \) given by (14) can be computed in a recursive way.

**Remark 1.** A widely used kernel is the Gaussian pdf, and in this case we have

\[ w_k(x^*(2M)) = \frac{1}{(2\pi)^{M/2}} \frac{1}{b_k^M} \exp \left\{ - \frac{1}{2} \sum_{j=1}^M \frac{(y_{k+1-j} - x^*_j)^2}{b_k^2} \right\}. \]

Other important kernels include the rectangle kernel, triangle kernel, Epanechnikov kernel, etc.

**Remark 2.** From the above example we see that the kernel function plays the role like a weight: The regressors \( \phi_k(M, M) \) close to \( x^*(M, M) \) are taken into considerably higher account in comparison with those far away from \( x^*(M, M) \), because the kernel \( w_k(x^*(2M)) \) rapidly vanishes as the regressors deviate from \( x^*(M, M) \). As for the sequence \( \{b_k\}_{k=1}^N \), it is usually required that \( b_k \rightarrow 0 \) but \( b_k^M N \rightarrow \infty \) as \( N \rightarrow \infty \). Thus, the number of data around \( x^*(2M) \) is increasing, and the estimates \( \theta_N(p, q) \) and \( \theta_{1,N}(p, q) \) generated by (11) are approaching to \( f(x^*(M, M)) \) and \( \phi(x^*(M, M)) \), respectively, as \( N \rightarrow \infty \), provided the orders (p, q) match the true system orders well.

We introduce the following assumption which is adopted in [Zhao et al. (2013)] for the convergence analysis of RLLE.
(A3) Select $b_k = 1/k^d$ for some $\delta \in (0, 1/(2(2M + 1)))$; $w(\cdot)$ is chosen as a symmetric probability density function (pdf) with $w(x) = 0$ for $\|x\| > 1$, and $\int_{\mathbb{R}^2} w(x) dx > 0$.

For estimating the local order $(s_0, t_0)$, we introduce the following local information criterion (LIC) $L_{N+1}(p , q)$:

$$L_{N+1}(p , q) = \sum_{k=1}^{N} u_k(x^k(2M)) \times \left( y_{k+1} - \theta_0, N, N+1(p,q) - \theta_1, N+1(p,q) \right)^T \left( \phi(p,q) - x^k(p,q) \right)^2,$$

where $\sigma_{N+1}(p,q) = \sum_{k=1}^{N} u_k(x^k(2M))$.

Remark 3. Notice that $\theta_0(p,q) = \{\theta_0, N, N+1(p,q) : N \geq 0\}$ decreases as $p$ and $q$ increase but the performance may not change much for $p \geq s_0$ and $q \geq t_0$. On the other hand, $(p + q)$ increases as $p$ and $q$ increases. This indicates that (17) with appropriately chosen $(s_0, t_0)$ defines a reasonable estimate for $(s_0, t_0)$.

We list some further conditions used for convergence analysis of the order estimates. Note that (1) is an infinite impulse response nonlinear system, and the second order statistics may not contain adequate information for its identification. So ergodicity and mixing properties are often required, see, e.g., Fan and Gijbels (1996) in statistics literature.

(A4) $\{\xi_k\}_{k \geq 0}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $E[\xi_k] = 0, 0 < E[\xi_k^2] < \infty$ for some $\eta \in (0, 2); \{\phi_k(M, M)\}$ and $\delta_k$ are mutually independent for each $k \geq 1$.

(A5) The sequence $\{\phi_k(M, M)\}_{k \geq 1}$ is geometrically ergodic, i.e., there exists an invariant probability measure $P_N(\cdot)$ on $(\mathbb{R}^M, \mathbb{B}(\mathbb{R}^M))$ and some constants $c_1 > 0$ and $0 < \rho_1 < 1$ such that $\|P_N(\cdot) - P_N(\cdot)\|_{\text{var}} \leq c_1 \rho_1^k$, where $P_N(\cdot)$ is the marginal distribution of $\phi_k(M, M)$. $P_N(\cdot)$ is with a bounded pdf, denoted by $f_{\delta_k}(\cdot)$, which is with a continuous second order derivative at $x^k(2M)$.

(A6) $\{\phi_k(M, M)\}_{k \geq 1}$ on $x^k(2M)$ is an $\alpha$-mixing with mixing coefficients $|\alpha(k)| \leq c_2k^{1-\delta}$ for some $c_2 > 0$ and $0 < \rho_2 < 1$ and $E[\phi_k(M, M)]^{1/2} < \infty$ for $k \geq 1$, where the constant $\delta$ is specified in assumption (A2).

(A7) The sequence $\{a_N\}_{N \geq 1}$ satisfies

$$N^{1-4\delta}/a_N \rightarrow 0, \quad a_N/N^{1-2\delta} \rightarrow 0,$$

where $\delta > 0$ is given in (A3).

The conditions (A5) and (A6), in fact, are on the asymptotical independency and stationarity of the sequence $\{\phi_k(M, M)\}_{k \geq 1}$, and they can be guaranteed by assuming stability of the system with input excited in a certain sense as shown in Zhao et al. (2010, 2013). The conditions given in Zhao et al. (2010, 2013) cover a large class of systems, including the ARX system, the Hammerstein systems, and the Wiener system, etc. So for ease of presentation, in this paper we assume that $\{\phi_k(M, M)\}_{k \geq 1}$ is a mixing process with an asymptotically stationary distribution.

The convergence of (17) is considered in the next section.

2.2. Strong consistency of estimates

For any fixed $1 \leq p \leq M$ and $1 \leq q \leq M$, define

$$\forall f(x^p, q) \triangleq \left[ \begin{array}{c} \frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_p} \frac{\partial f}{\partial x_{M+1}} \cdots \frac{\partial f}{\partial x_{M+q}} \end{array} \right]^T, (19)$$

and

$$\bar{\theta}_{1, N+1}(p,q) \triangleq \left[ \theta_{1, N+1}(p,q)(1:p)^T 0 \cdots 0 \right]_{M-p},$$

$$\tilde{\theta}_{1, N+1}(p,q) \triangleq \left[ f(x^k(2M)) - \theta_{0, N+1}(p,q) \right] \left[ \phi(p,q) - x^k(p,q) \right] \in \mathbb{R}^{2M}. (20)$$

Denote the maximal and minimal eigenvalues of $\sum_{k=1}^{N} w_i(x^k(2M)) X_i(p,q)X_i(p,q)^T$ by $\lambda_{\text{max}}^p(N)$ and $\lambda_{\text{min}}^p(N)$, respectively.

Theorem 1. Under conditions (A1)–(A7), the order estimate $(p_N, q_N)$ given by (17) is strongly consistent,

$$(p_N, q_N) \rightarrow (s_0, t_0) \quad \text{a.s.} (22)$$

Proof. See Appendix. □

2.3. A simple procedure for searching the minimum of LIC

To obtain estimates defined by (17) it is required to calculate $M^2$ function values of $L_{N+1}(p,q)$ and then to find the minimum among them. In this section we introduce a simple procedure for searching the minimum of (17) for which the computational complexity is $O(M)$.

Define

$$\hat{p}_N \triangleq \arg\min L_{N+1}(p, M),$$

$$\hat{q}_N \triangleq \arg\min L_{N+1}(p, M). (23)$$

where $L_{N+1}(p,q)$ is defined by (15).

Theorem 2. Assume (A1)–(A7) hold. Then

$$\hat{p}_N \rightarrow s_0 \quad \text{a.s.} (25)$$

$$\hat{q}_N \rightarrow t_0 \quad \text{a.s.} (26)$$

Proof. Here we just sketch the proof. The proof is divided into two steps. First, the strong consistency of $\hat{p}_N$ is proved. This can be done by carrying out almost the same discussion as that given in Theorem 1. Second, based on that $\hat{p}_N = s_0$ and hence $L_{N+1}(\hat{p}_N, 1) = L_{N+1}(s_0, q)$ for all $N$ large enough, the convergence of $\hat{q}_N$ is established via a similar derivation as that for (25). □
Remark 4. The order estimates can also be defined by
\[
\hat{q}_{N+1} = \min_{1 \leq q \leq M} L_{N+1}(M, q), \tag{27}
\]
\[
\hat{p}_{N+1} = \min_{1 \leq p \leq M} L_{N+1}(p, \hat{q}_{N+1}), \tag{28}
\]
which are strongly consistent under (A1)-(A7).

3. Modified LIC

In the last section, based on LIC the strongly consistent estimate for the system order at a fixed point is obtained. We now introduce a modified LIC as follows:
\[
\tilde{L}_{N+1}(p, q) \overset{d}{=} N \log \sigma_{N+1}(p, q) + a_N \cdot (p + q), \tag{29}
\]
where \( \sigma_{N+1}(p, q) \) is given by (16).

The estimate \((\hat{p}_{N+1}, \hat{q}_{N+1})\) for \((s_0, t_0)\) is given by minimizing \(\tilde{L}_{N+1}(p, q)\), i.e.,
\[
(\hat{p}_{N+1}, \hat{q}_{N+1}) \overset{d}{=} \min_{1 \leq p \leq M, 1 \leq q \leq M} \tilde{L}_{N+1}(p, q). \tag{30}
\]

**Theorem 3.** Under conditions (A1)-(A7), the order estimate \((\hat{p}_{N}, \hat{q}_{N})\) given by (30) is strongly consistent,
\[
(\hat{p}_{N}, \hat{q}_{N}) \overset{d}{\rightarrow} (s_0, t_0) \quad \text{a.s.} \tag{31}
\]
**Proof.** See Appendix. \(\square\)

Define
\[
\tilde{p}_{N+1} = \min_{1 \leq p \leq M} \tilde{L}_{N+1}(p, M), \tag{32}
\]
\[
\tilde{q}_{N+1} = \min_{1 \leq q \leq M} \tilde{L}_{N+1}(\tilde{p}_{N+1}, q), \tag{33}
\]
where \(\tilde{L}_{N+1}(p, q)\) is defined by (29).

Similar to Theorem 2, the following result holds.

**Theorem 4.** Assume (A1)-(A7) hold. Then
\[
\tilde{p}_{N} \overset{d}{\rightarrow} s_0 \quad \text{a.s.} \tag{34}
\]
\[
\tilde{q}_{N} \overset{d}{\rightarrow} t_0 \quad \text{a.s.} \tag{35}
\]

**Remark 5.** The order estimates can also be defined by
\[
\tilde{q}_{N+1} = \min_{1 \leq q \leq M} \tilde{L}_{N+1}(M, q), \tag{36}
\]
\[
\tilde{p}_{N+1} = \min_{1 \leq p \leq M} \tilde{L}_{N+1}(p, \tilde{q}_{N+1}), \tag{37}
\]
which are strongly consistent under (A1)-(A7).

**Remark 6.** LIC and its modification considered in this paper look similar to the well known AIC, BIC, and their generalizations. However, AIC, BIC, and others are in a global sense and thus they are inapplicable to the local order estimation. While for LIC the kernel function \(w_k(x^*(2M))\) plays a bandwidth like role to stress those measurements which are close to the given point and to take their average. The sequence \(\{a_N\}\) in AIC, BIC, and their generalizations can be chosen as \(N^\alpha\) for any \(0 < \alpha < 1\), or \(\log^{1+\beta} N\) for some \(\beta \geq 0\), or even a constant (Chen & Guo, 1987, 1991), but here in LIC the choice of \(\{a_N\}\) is more delicate.

4. Discussions and simulations

In the above sections, we have introduced two criteria, i.e., \(L_N(p, q)\) defined by (15) and \(\hat{L}_N(p, q)\) defined by (29), respectively. Theoretically, any \(a_N\) that meets the requirement in assumption (A7), for example, \(a_N = cN^{1-\delta}\) for any constant \(c > 0\), guarantees the a.s. convergence of the estimates generated by (15) and (29). However, from the numerical calculation point of view, there exists some difference between \(L_N(p, q)\) and \(\hat{L}_N(p, q)\).

(i) Let us take \(a_N = cN^{1-\delta}\) for some constant \(c > 0\) as an example. As required in assumption (A3), the parameter \(s\) usually is small and thus even for the integer \(N > 0\) large enough it still holds that \(N^{1-\delta} \approx N\). On the other hand, since the kernel functions \(w_k(x^*(2M))\) is involved in the residual term, i.e.,
\[
\sigma_{N+1}(p, q) \overset{d}{=} \sum_{k=1}^{N} w_k(x^*(2M)) \left(y_{k+1} - \theta_{0, N+1}(p, q)\right)^2 + \theta_{1, N+1}(p, q)^T (y_{k+1} - x^*(p, q)), \tag{38}
\]

it often holds that \(\sigma_{N+1}(p, q) = o(N)\) and thus \(a_N(p + q)\) is the dominated term in \(L_N(p, q)\), i.e.,
\[
L_N(p, q) = |\sigma_{N+1}(p, q) + a_N \cdot (p + q) \approx \mathcal{O}(N \cdot (p + q)). \tag{39}
\]

This indicates that for convergence of the estimates generated from the criterion \(L_N(p, q)\), in order to balance the penalty term \(a_N(p + q)\) it usually requires the number of data be large enough, and thus the convergence rate is slow. To speed up the convergence rate, one may choose, for example, \(a_N = cN^{1-\delta}\) for some \(c > 0\) small enough to reduce the effect of the penalty term \(a_N \cdot (p + q)\) in \(L_N(p, q)\).

(ii) By noticing the first term in \(\tilde{L}_N(p, q)\) defined by (29), it can be found that \(N = o(N \log \sigma_{N+1}(p, q))\) and thus \(a_N \cdot (p + q)\) with \(a_N\) satisfying (A7) is a moderate penalty term in \(\tilde{L}_N(p, q)\). So the convergence rate of estimates generated from \(\tilde{L}_N(p, q)\) should be faster than that of (17).

In the following we present the numerical simulations to verify the theoretical analysis.

**Example 1.** Consider an FIR system
\[
y_{k+1} = f(u_k, u_{k-1}, u_{k-2}) = f_{k+1}, \tag{38}
\]
\[
f(u_k, u_{k-1}, u_{k-2}) = \begin{cases} u_k + u_{k-1} + u_{k-2}, & \text{if } u_k > 1, \\ u_k + u_{k-1}, & \text{if } -1 < u_k < 0, \\ u_k, & \text{if } u_k < -1. \end{cases} \tag{39}
\]

where the inputs \(\{u_k\}_{k=1}^{N}\) and the noises \(\{\varepsilon_k\}_{k=3}^{N}\) are mutually independent i.i.d. Gaussian random variables with distributions \(\mathcal{N}(0, \sigma^2)\) and \(\mathcal{N}(0, 0.1^2)\), respectively. It is noticed that the right-hand side of (38) is free of the system output, so \(x^*(2M)\) defined by (5) changes to \(x^*(M) \overset{d}{=} \left[u_{1}, \ldots, u_M\right]^T\), and \(L_{N+1}(p, q)\) and \(\sigma_{N+1}(p, q)\) defined by (15) and (16) correspondingly change to functions \(L_{N+1}(q)\) and \(\sigma_{N+1}(q)\), respectively. Assume the upper bound \(M\) of system orders is 4. Thus, \(x^*(M) = \left[u_{1}, \ldots, u_{4}\right]^T\). We choose two points for test, \(x^*_M(M) = [2 \ 1 \ 1 \ 0]^T\) and \(x^*_M(M) = [0 \ 0 \ 0 \ 0]^T\). Note that the true local orders are different at the two points.

More than 30 simulations have been performed. Here we only present one of them since the performance of others is almost the same. Tables 1 and 2 show the performance of the proposed estimator with the data set \(\{u_k, y_{k+1}\}_{k=1}^{N} = 1000, 2000, 3000, 4000\), respectively, where
\[
L_N(q) = \sigma_{N+1}(q) + 0.005N^{1-\delta}, \tag{39}
\]
\[
\tilde{L}_N(q) = N \log \sigma_{N+1}(q) + 0.5N^{1-\delta} \cdot q, \tag{40}
\]
with \(\delta = 0.05\).
It can be found that the criterion $\tilde{L}_{N+1}(q)$ always gives the correct order estimates 3 and 2 for the local orders of the system at $x_1^q(M)$ and $x_2^q(M)$, respectively. It can also be found that the convergence rate of estimates generated from $\tilde{L}_{N+1}(q)$ is faster than that generated from $L_{N+1}(q)$.

**Example 2.** Consider a benchmark problem for nonlinear system identification (Bai et al., 2007; Zhao et al., 2013):

$$x_1(k + 1) = \frac{x_1(k)}{1 + x_1^2(k)} + 2 \sin x_2(k),$$

$$x_2(k + 1) = x_2(k) \cos x_1(k) + x_1(k) \exp \left( \frac{-x_1^2(k) + x_2^2(k)}{8} \right) + u_k^2 + 0.5 \cos(x_1(k) + x_2(k)),$$

$$y_k = \frac{x_1(k)}{1 + 0.5 \sin x_2(k)} + \frac{x_2(k)}{1 + 0.5 \sin x_1(k)} + \epsilon_k,$$

where $u_k$ and $y_k$ are the system input and output, respectively, $\epsilon_k$ is the system noise with the Gaussian distribution $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.1$, and $x_1(k)$ and $x_2(k)$ are the unmeasured system states.

The NARX system

$$y_{k+1} = f(y_k, \ldots, y_{k-M}, u_k, \ldots, u_{k-M}) + \epsilon_{k+1}$$

is used to approximate the unknown system. Notice that in existing literature (Bai et al., 2007; Zhao et al., 2013), a common choice for the order $M$ is $M = 3$. Here we adopt $M = 3$ as the upper bound for the system order.

First, $N(=1000)$ samples $(u_k, y_k)_{k=1}^{1000}$ are generated by i.i.d. $u_k$ with the Gaussian distribution $\mathcal{N}(0, 1)$. The local orders as well as the values of the function $f(\cdot)$ and its gradients $\nabla f(\cdot)$ are estimated based on $(u_k, y_k)_{k=1}^{1000}$. Then the input signals $u_k = \sin \frac{\pi k}{N} + \sin \frac{\pi k}{2}, k = N + 1, \ldots, N + 100$ are fed into the estimated model to calculate the one-step predicted output. Specifically, the intervals $[-3, 3]$ and $[-2, 2]$ are equally divided into 5 and 4 subintervals, respectively, and the domain of interest $S = \{(y_1, y_2, y_1, u_3, u_2, u_1) \in \mathbb{R}^6 \mid y_3 \in [-3, 3], y_2 \in [-3, 3], y_1 \in [-3, 3], u_3 \in [-2, 2], u_2 \in [-2, 2], u_1 \in [-2, 2]\}$ is uniformly divided into 8000 disjoint small cubes $S = \bigcup_{i=1}^{8000} \bar{S}_i$ and from each $\bar{S}_i$, a point $\psi^*_i$ is randomly chosen, $i = 1, \ldots, 8000$. Then with $\delta = 0.04$ and $\alpha_q = N^{1-3\delta}$, the algorithms (14) and (29) are applied to estimate the local orders denoted by $(p_{N,i}, q_{N,i})$, and parameters denoted by $f_N(\psi^*_i(p_{N,i}, q_{N,i}))$ and $f_N(\psi^*_i(p_{N,i}, q_{N,i}))$ at each $\psi^*_i$, $i = 1, \ldots, 8000$, where $\psi^*_i(p_{N,i}, q_{N,i})$ is a $(p_{N,i} + q_{N,i})$-vector defined by (7). Then the one-step predictions are given as follows,

$$\hat{y}_{k+1} = \hat{f}_N(\psi^*_i(p_{N,i}, q_{N,i})) + \nabla f_N(\psi^*_i(p_{N,i}, q_{N,i})) \left( \hat{\psi}_N(p_{N,i}, q_{N,i}) - \psi^*_i(p_{N,i}, q_{N,i}) \right),$$

with regressor

$$\hat{\psi}_N(p_{N,i}, q_{N,i}) = \left[ \hat{y}_k, \ldots, \hat{y}_{k-p_{N,i}}, u_k, \ldots, u_{k-q_{N,i}} \right]^T,$$

if $\psi^*_i(3, 3) \in \bar{S}_i$ for some $i = 1, \ldots, 8000$ where $k = N + 1, \ldots, N + 100$.

Ten simulations are performed. Figs. 1 and 2 show one of the simulations. In Fig. 1 the solid lines are the actual output $y_k$, $k = N + 1, \ldots, N + 100$, the dotted line is the predicted output generated by (41) and the dashed line is the predicted output generated by (42) without order estimation, i.e.,

$$\hat{y}_{k+1} = \hat{f}_N(\psi^*_i(3, 3)) + \nabla f_N(\psi^*_i(1, 3)) \left( \hat{\psi}_N(3, 3) - \psi^*_i(3, 3) \right).$$

Fig. 2 shows the estimated orders at the 8000 given points. Notice that the blocks at the bottom of the figure represent the 1st to the 100th points while those at the top of the figure represent the 7901st to the 8000th points. The estimated orders are indicated with different depths of color.

To test the performance of algorithm, the following quality of fit (QOF) is calculated

$$1 - \frac{\sum_{k=N+1}^{N+100} (y_k - \hat{y}_k)^2}{\sum_{k=N+1}^{N+100} \left( y_k - \frac{1}{N+100} \sum_{t=N+1}^{N+100} y_t \right)^2} \times 100\%,$$

where $\hat{y}_k$ is the predicted output.

Table 3 shows the average of QOF of the ten simulations and the standard deviation. From Figs. 1 and 2 and Table 3 we find that

<table>
<thead>
<tr>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1^q(M)$</td>
<td>$x_2^q(M)$</td>
<td>$x_1^q(M)$</td>
<td>$x_2^q(M)$</td>
</tr>
<tr>
<td>$q$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$L_0(q)$, N = 1000</td>
<td>36.8984</td>
<td>29.9966</td>
<td>26.5952</td>
</tr>
<tr>
<td>$L_0(q)$, N = 2000</td>
<td>86.5703</td>
<td>69.4839</td>
<td>60.8297</td>
</tr>
<tr>
<td>$L_0(q)$, N = 3000</td>
<td>105.2002</td>
<td>132.5021</td>
<td>112.9479</td>
</tr>
<tr>
<td>$L_0(q)$, N = 4000</td>
<td>203.3250</td>
<td>183.6043</td>
<td>155.0399</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
<th>$L_0(q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1^q(M)$</td>
<td>$x_2^q(M)$</td>
<td>$x_1^q(M)$</td>
<td>$x_2^q(M)$</td>
</tr>
<tr>
<td>$q$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$L_0(q)$, N = 1000</td>
<td>3.5178 × 10^1</td>
<td>3.4255 × 10^1</td>
<td>3.3769 × 10^1</td>
</tr>
<tr>
<td>$L_0(q)$, N = 2000</td>
<td>9.1149 × 10^1</td>
<td>8.8993 × 10^1</td>
<td>8.9046 × 10^1</td>
</tr>
<tr>
<td>$L_0(q)$, N = 3000</td>
<td>1.5305 × 10^2</td>
<td>1.4963 × 10^2</td>
<td>1.4753 × 10^2</td>
</tr>
<tr>
<td>$L_0(q)$, N = 4000</td>
<td>2.2067 × 10^3</td>
<td>2.1511 × 10^4</td>
<td>2.1141 × 10^4</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Zhao, W., et al., Kernel-based local order estimation of nonlinear nonparametric systems. Automatica (2014), http://dx.doi.org/10.1016/j.automatica.2014.10.069
Table 3

<table>
<thead>
<tr>
<th></th>
<th>QOF with order estimation</th>
<th>QOF without order estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>70.36%</td>
<td>71.55%</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.0516</td>
<td>0.0423</td>
</tr>
</tbody>
</table>

The performance of algorithm (41) is similar to that of algorithm (42). However, from Fig. 2 we find that the estimated local orders are reduced at many of the 8000 given points. Thus the benefit of algorithm (41) is that to apply the order estimation technique the complexity of the identified model for the benchmark problem is reduced and therefore a more precise system model is obtained.

5. Concluding remarks

In the paper LIC is suggested for the local order estimation of NARX systems and the consistency of the estimates is established. Some important issues connected with LIC are summarized as follows.

1. Theoretically, the order estimation algorithm requires to compute the local order at each point of interest. For some special systems, for example, the piecewise-defined systems, the number of data points needed can be significantly reduced. In this case, by implementing the proposed algorithms, fewer local orders have to be estimated and better models of the system can be obtained.

2. LIC is based on the recursive locally linear estimator introduced in Zhao et al. (2013). We can also use its nonrecursive version investigated in Bai (2010) to construct LIC and to carry out corresponding convergence analysis.

3. The results in the paper can easily be extended to the case $1 \leq s_0 \leq M_1$ and $1 \leq t_0 \leq M_2$ for some known but different $M_1$ and $M_2$. For future research, it is of interest to remove the upper bound assumption on the true system orders.

4. The order estimation algorithms in the paper are nonrecursive, i.e., for each $N \geq 1$ we need to calculate the function $L_{N+1}(p, q)$, $1 \leq p \leq M$, $1 \leq q \leq M$ and then to find the minimum to serve as the estimate. It is interesting to consider the recursive way to obtain the order estimates.

5. The closed-loop order estimation of NARX systems also deserves further research.

Acknowledgments

Wenxiao Zhao would like to thank Miss Jinlong Lei for her help in numerical simulation. The research of Wenxiao Zhao was supported by the National Key Basic Research Program of China (973 program) under Grant No. 2014CB845301 and the National Natural Science Foundation (NSF) of China under Grants No. 61104052, 61273193, 61272902, and 61134013. The research of Han-Fu Chen was supported by the 973 program of China under Grant No. 2014CB845301 and the NSF of China under Grants No. 61273193, 61120106011, and 61134013. The research of Er-Wei Bai was supported in part by NSF CNS-1329057 and DOE DEFG52-09NA29364. The research of Kang Li was supported by EPSRC project under EP/L001063/1.

Appendix

Lemma 1. Assume that (A1)–(A7) hold. Then the following estimates take place:

\[
\frac{1}{N} \sum_{k=1}^{N} w_k(x^*(2M)) \left( f(\psi_k(M, M)) - f(x^*(2M)) \right) - \nabla f(x^*(2M))^T (\psi_k(M, M) - x^*(2M)) \\
= \frac{1}{2(1 - 2\delta)^2} b_2^2 \int_{x \geq m} w(x) x \nabla^2 f(x^*(2M)) x \, dx \\
\cdot \cdot v_0(x^*(2M)) + o(b_2^2) + o \left( 1 / \left( N^{1 - \epsilon} b_2^M \right) \right) \quad \text{a.s.} \tag{44}
\]

Appendix

\[
\frac{1}{N} \sum_{k=1}^{N} w_k(x^*(2M)) (\psi_k(M, M) - x^*(2M)) \\
\cdot \left( f(\psi_k(M, M)) - f(x^*(2M)) \right)
\]
where $I \in \mathbb{R}^{2M \times 2M}$ and $A_N = \begin{bmatrix} A_N(1, 1) & A_N(1, 2) \\ A_N(2, 1) & A_N(2, 2) \end{bmatrix}$ with elements $A_N(1, 1), A_N(1, 2), A_N(2, 1),$ and $A_N(2, 2)$ defined as follows:

\[
A_N(1, 1) = \frac{1}{N} \sum_{k=1}^{N} w_k(x^*(2M)), \quad A_N(1, 2) = A_N(1, 2)^T
\]

\[
A_N(2, 2) = \frac{1}{N^{1-\alpha}} \sum_{k=1}^{N} w_k(x^*(2M))(\psi_k(M, M) - x^*(2M))^T.
\]

At given $x^*(2M)$, define

\[
\xi_{k+1} \triangleq f(\psi_k(M, M)) - f(x^*(2M)) - \text{sgn}(x^*(2M))\).
\]

Lemma 3. Assume (A1) holds. Then the function $\sigma_{N+1}(p, q)$ defined by (16) with any $1 \leq p \leq M$ and $1 \leq q \leq M$ takes the following expression:

\[
\sigma_{N+1}(p, q) = \frac{N}{N+1} \sum_{i=1}^{N} w_i(x^*(2M))X_i(M, M)
\]

\[
\times X_i(M, M) - x^*(2M)) + 2\delta_{N+1}(p, q)^T \sum_{i=1}^{N} w_i(x^*(2M))X_i(M, M)\xi_{i+1}
\]

\[
+ \sum_{i=1}^{N} w_i(x^*(2M))\xi_{i+1}^{2}.
\]

Lemma 2 (Zhao et al., 2013). Assume that (A1)–(A7) hold. Then

\[
\sum_{k=1}^{N} w_k(x^*(2M))X_k(M, M)X_k(M, M)^T
\]

\[
\overset{N}{\longrightarrow} \sigma_{N+1}(p, q)^T \left( \sum_{i=1}^{N} w_i(x^*(2M))X_i(p, q)X_i(p, q)^T \right)^{-1}
\]

\[
+ \sum_{i=1}^{N} w_i(x^*(2M))\xi_{i+1}^{2}.
\]

Proof. By the definition of $\sigma_{N+1}(p, q)$ we have that

\[
\sigma_{N+1}(p, q) = \sum_{i=1}^{N} w_i(x^*(2M)) \left( y_{i+1} - \theta_{i, N+1}(p, q) \right)
\]

\[
- \theta_{i, N+1}(p, q)^T (\psi_i(p, q) - x^*(p, q))
\]

\[
= \sum_{i=1}^{N} w_i(x^*(2M)) \left( f(\psi_i(M, M)) + \xi_{i+1} - \theta_{i, N+1}(p, q) \right)
\]

\[
- \theta_{i, N+1}(p, q)^T (\psi_i(p, q) - x^*(p, q))
\]

\[
= \sum_{i=1}^{N} w_i(x^*(2M)) \left( f(x^*(2M))
\]

\[
\overset{N}{\longrightarrow} \sigma_{N+1}(p, q)^T \left( \sum_{i=1}^{N} w_i(x^*(2M))X_i(p, q)X_i(p, q)^T \right)^{-1}
\]

\[
+ \sum_{i=1}^{N} w_i(x^*(2M))\xi_{i+1}^{2}.
\]
\[ + \nabla f(x^*(2M))^T (\varphi_i(M, M) - x^*(2M)) \]
\[ - \theta_{0, N+1}(p, q) - \tilde{\theta}_{1, N+1}(p, q)^T (\varphi_i(M, M) - x^*(2M)) \]
\[ + f(\varphi_i(M, M)) - f(x^*(2M)) \]
\[ - \nabla f(x^*(2M))^T (\varphi_i(M, M) - x^*(2M)) + \varepsilon_{k+1} \]^2
\[ = \sum_{i=1}^N w_i(x^*(2M))\left( \tilde{\theta}_{N+1}(p, q)x_i(M, M) + \xi_{k+1} \right)^2. \tag{58} \]

where \( \xi_{k+1} \) is defined by (55) and \( \tilde{\theta}_{N+1}(p, q) \) is given by (21). From (58) we then obtain (56).

We now consider the case \( p \geq s_0 \) and \( q \geq t_0 \). By Lemma 2, the matrices \( \sum_{k=1}^N u_k(x^*(2M))X_k(M, M)x_k(M, M)^T \) are nonsingular for all \( N \) large enough. This ensures \( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)x_k(p, q)^T > 0 \) for \( 1 \leq p \leq M, 1 \leq q \leq M \) and all \( N \) large enough by noticing the definition of \( X_k(p, q) \). Without losing generality, we assume that these matrices are positive definite for \( N \geq 1 \). Then the formula (14) takes place. From (14) and by noticing \( p \geq s_0 \) and \( q \geq t_0 \) we have
\[ \theta_{N+1}(p, q) = \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)x_k(p, q)^T \right)^{-1} \]
\[ \times \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)y_{k+1} \right) \]
\[ = \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)x_k(p, q)^T \right)^{-1} \]
\[ \cdot \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)\left( f(x^*(2M)) \right) \right. \]
\[ + \nabla f(x^*(p, q))^T (\varphi_i(p, q) - x^*(p, q)) \]
\[ + f(\varphi_i(M, M)) - f(x^*(2M)) \]
\[ \left. - \nabla f(x^*(p, q))^T (\varphi_i(p, q) - x^*(p, q)) + \varepsilon_{k+1} \right) \tag{59} \]

which, by noticing the definition of \( \xi_{k+1} \) given by (55), implies
\[ \left( f(x^*(2M)) \right. \nabla f(x^*(p, q))^T \left. \right) - \theta_{N+1}(p, q) \]
\[ = - \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)x_k(p, q)^T \right)^{-1} \]
\[ \cdot \left( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)y_{k+1} \right). \tag{60} \]

Noticing (21) we find that
\[ \left( f(x^*(2M)) \right. \nabla f(x^*(p, q))^T \left. \right) - \theta_{N+1}(p, q) \]
\[ = \tilde{\theta}_{N+1}(p, q)x_i(M, M), \tag{61} \]

which combining with (56) and (60) yields (57). \( \square \)

**Proof of Theorem 1.** The proof is motivated by Chen and Guo (1987) for the order estimation of linear systems. Here we present it in detail in a nonlinear and nonparametric description. By Lemma 2, we may assume \( \sum_{k=1}^N u_k(x^*(2M))X_k(p, q)x_k(p, q)^T > 0 \), for \( 1 \leq p \leq M, 1 \leq q \leq M \) and all \( N \geq 1 \).

Because all \( p, q, s_0 \), and \( t_0 \) are positive integers between 1 and \( M \), for (22) it suffices to show that any limit point of \((p_N, q_N)\) \( \geq 1 \) coincides with \( (s_0, t_0) \). Assume that \((p', q')\) is a limit point of \((p_N, q_N)\) \( \geq 1 \), i.e., there exists a subsequence of \((p_N, q_N)\) \( \geq 1 \) denoted by \((p_{N_k}, q_{N_k})\) \( k \geq 1 \), such that \((p_{N_k}, q_{N_k}) \rightarrow (p', q')\) as \( k \rightarrow \infty \). Since \((p_{N_k}, q_{N_k})\) \( \geq 1 \) and \((p', q')\) are nonnegative integers, there exists \( K > 0 \) such that
\[ (p_{N_k}, q_{N_k}) = (p', q'), \quad \forall k \geq K. \tag{62} \]

For (22) we need to prove the impossibility of the following cases: (i) \( p' < s_0 \); (ii) \( q' < t_0 \); (iii) \( p' + q' > s_0 + t_0 \).

We first consider case (i). By Lemmas 1 and 2, it follows that
\[ \lambda_{\max}^M(N) \sim N, \quad \lambda_{\min}^M(N) \sim N^{1-2\delta}. \tag{63} \]

Define
\[ M_{N+1} = \frac{\tilde{\theta}_{N+1}(p', q')^T}{\theta_{N+1}(p', q')} \]
\[ \cdot \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)x_i(M, M)^T \]
\[ + 2\tilde{\theta}_{N+1}(p', q')\sum_{i=1}^N w_i(x^*(2M))X_i(M, M)\xi_{k+1}. \tag{64} \]

and
\[ \alpha_{N+1} = \left( \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)x_i(M, M)^T \right)^{-1} \]
\[ \cdot \tilde{\theta}_{N+1}(p', q'). \tag{65} \]

From Lemma 3, for all \( k \geq K \) it follows that
\[ \alpha_{N+1}(p', q') = M_{N+1} + \sum_{i=1}^N w_i(x^*(2M))\xi_{k+1}. \tag{66} \]

By the definition of \( \tilde{\theta}_{N+1}(p', q') \) and noticing \( p' < s_0 \), we know that
\[ \|\tilde{\theta}_{N+1}(p', q')\|^2 \geq \left( \frac{\partial f}{\partial x_{t_0}} \right)^2 \tag{67} \]
and the following equality takes place:
\[ M_{N+1} = \alpha_{N+1}^T \left( \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)x_i(M, M)^T \right)^{-1} \]
\[ + 2\sum_{i=1}^N w_i(x^*(2M))X_i(M, M)\xi_{k+1} \]
\[ \cdot \tilde{\theta}_{N+1}(p', q') \|\tilde{\theta}_{N+1}(p', q')\|^2 \cdot \tilde{\theta}_{N+1}(p', q')^T \]
\[ \cdot \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)x_i(M, M)^T \right)^{-1} \alpha_{N+1}. \tag{68} \]

For RLE, we have
\[ \left( \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)x_i(M, M)^T \right)^{-1} \]
\[ \cdot \sum_{i=1}^N w_i(x^*(2M))X_i(M, M)\xi_{k+1} \]
\[ = \left[ \begin{array}{c} N^\delta \\ 0 \end{array} \right] A_{N+1}(1, 1) A_{N+1}(1, 2)^{-1} \left[ \begin{array}{c} B_{N+1}(1) \\ B_{N+1}(2) \end{array} \right], \tag{69} \]
where \( A_{ij}, i, j = 1, 2 \) are defined in Lemma 2 and

\[
B_N(1) = \frac{1}{N^{1-\alpha}} \sum_{k=1}^{N} w_k(x^*(2M))
\]

\[
\cdot \left( f(\phi_k(M), M) - f(x^*(2M))
  - \nabla f(x^*(2M))^T (\phi_k(M, M) - x^*(2M)) + \varepsilon_{k+1} \right)
\]

\[
B_N(2) = \frac{1}{N^{1-\delta}} \sum_{k=1}^{N} w_k(x^*(2M))
\]

\[
\cdot \left( \phi_k(M, M) - x^*(2M) \right) \cdot \left( f(\phi_k(M, M), M) - f(x^*(2M))
  - \nabla f(x^*(2M))^T (\phi_k(M, M) - x^*(2M)) + \varepsilon_{k+1} \right).
\]

By Lemmas 1 and 2, it follows that

\[
\begin{bmatrix}
A_N(1, 1) & A_N(1, 2) \\
A_N(2, 1) & A_N(2, 2)
\end{bmatrix}
\rightarrow_{N \to \infty} f_W(x^*(2M))
\]

\[
\cdot \begin{bmatrix}
1 \\
0
\end{bmatrix}
\rightarrow_{N \to \infty} 0 \quad \text{a.s.}
\]

and

\[
\begin{bmatrix}
B_N(1) \\
B_N(2)
\end{bmatrix}
\rightarrow_{N \to \infty} 0 \quad \text{a.s.}
\]

Then by (69), (70), and (71), we have

\[
\left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(M, M)X_i(M, M)^T \right)^{-1}
\]

\[
\cdot \left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(M, M)\varepsilon_{i+1} \right) = o(1),
\]

and by noticing (67),

\[
M_{N_k+1} = \alpha_{N_k+1} \left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(M, M)X_i(M, M)^T \right)^{-1}
\]

\[
\cdot \alpha_{N_k+1} \cdot (1 + o(1))
\]

\[
\geq \frac{1}{2} \min_{(M, M)} (N_k) \left\| \nabla_{\varepsilon_{N_k+1}}(p', q') \right\|^2
\]

\[
\geq \frac{1}{2} \min_{(M, M)} (N_k) \left( \frac{\partial f}{\partial x_{N_0}} \right)^2,
\]

from which and (66) we have

\[
\sigma_{N_k+1}(p', q') \geq \frac{1}{2} \min_{(M, M)} (N_k) \left( \frac{\partial f}{\partial x_{N_0}} \right)^2
\]

\[
+ \sum_{i=1}^{N_k} w_i(x^*(2M))\varepsilon_{i+1}^2.
\]

Now we consider \( \sigma_{N_k+1}(s_0, t_0) \). By Lemma 3, it holds that

\[
\sigma_{N_k+1}(s_0, t_0) = - \left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(s_0, t_0)\varepsilon_{i+1} \right)^T
\]

\[
\cdot \left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(s_0, t_0)X_i(s_0, t_0)^T \right)^{-1}
\]

\[
= O \left( \sum_{i=1}^{N_k} w_i(x^*(2M))X_i(s_0, t_0)\varepsilon_{i+1} \right)^T
\]

\[
= O \left( \sum_{i=1}^{N_k} w_i(x^*(2M))\varepsilon_{i+1}^2 \right) = O(1).
\]

since \( \delta \in \left( 0, \frac{1}{2(M+1)} \right) \) and hence \( 0 < M \delta < 1/2 \). The estimate (81) implies (80) and hence (79).

Combining (75), (77), (78), and (79), we have

\[
\sigma_{N_k+1}(s_0, t_0) \leq \sum_{i=1}^{N_k} w_i(x^*(2M))\varepsilon_{i+1}^2 = O(N_k).
\]
By (63), (74), and (75) and paying attention to (A7) we have the following

\[ 0 \geq L_{N_{k}+1}(p', q') - L_{N_{k}+1}(s_0, t_0) \]

\[ = \sigma_{N_{k}+1}(p', q') - \sigma_{N_{k}+1}(s_0, t_0) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ \geq c_2^{(M, M)}(N_{k}) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ = \lambda_{\min}^{(M, M)}(N_{k}) (c + a_{N_{k}}(p' + q' - s_0 - t_0)) \rightarrow \infty, \quad \text{(83)} \]

where \( c > 0 \) may depend on sample paths. The contradiction ensures that \( p' \geq s_0 \). Similarly, we can prove that \( q' \geq t_0 \).

Finally, we consider the case (iii): \( p' + q' > s_0 + t_0 \). Since we have established \( p' \geq s_0 \) and \( q' \geq t_0 \), by Lemma 3, it follows that

\[ \sigma_{N_{k}+1}(p', q') = - \left( \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')\xi_{i+1} \right)^{T} \]

\[ \cdot \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')X_{i}(p', q')^{T} + \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')^{2} \xi_{i+1}. \quad \text{(84)} \]

By Lemmas 1 and 2, we have

\[ \left( \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')\xi_{i+1} \right)^{T} \]

\[ \cdot \left( \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')X_{i}(p', q')^{T} \right)^{-1} \]

\[ \cdot \left( \sum_{i=1}^{N_{k}} u_{i}(x' \cdot 2M)X_{i}(p', q')\xi_{i+1} \right) \]

\[ = O\left( \left[ N_{k}^{1-2\delta} \right. \left. N_{k}^{-\delta -1} \right] \right) \rightarrow O(N_{k}^{1-4\delta}), \quad \text{(85)} \]

where \( I \in \mathbb{R}^{p' \times q' + 1} \) and \( I \in \mathbb{R}^{(p' + q' + 1) \times (p' + q')} \).

From (75), (84), and (85) we have

\[ 0 \geq L_{N_{k}+1}(p', q') - L_{N_{k}+1}(s_0, t_0) \]

\[ = \sigma_{N_{k}+1}(p', q') - \sigma_{N_{k}+1}(s_0, t_0) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ \geq -cN_{k}^{1-4\delta} + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ = a_{N_{k}}(p' + q' - s_0 - t_0 - cN_{k}^{1-4\delta} a_{N_{k}}) \rightarrow \infty, \quad \text{(86)} \]

where the limit takes place by noticing (A7) and \( p' + q' > s_0 + t_0 \). The obtained contradiction indicates that \( p' = s_0 \) and \( q' = t_0 \). This completes the proof. □

**Proof of Theorem 3.** We only sketch the proof.

It suffices to show that any limit point of \( \{(\overline{p}_{N_{k}}, \overline{q}_{N_{k}})\}_{N_{k} \geq 1} \) coincides with \( (s_0, t_0) \). Assume that \( (p', q') \) is a limit point of \( \{(\overline{p}_{N_{k}}, \overline{q}_{N_{k}})\}_{N_{k} \geq 1} \), i.e., there exists a subsequence \( \{(\overline{p}_{N_{k}}, \overline{q}_{N_{k}})\}_{N_{k} \geq 1} \) denoted by \( \{(\overline{p}_{N_{k}}, \overline{q}_{N_{k}})\}_{N_{k} \geq 1} \), and \( K > 0 \) such that

\[ (\overline{p}_{N_{k}}, \overline{q}_{N_{k}}) = (p', q'), \quad \forall K \geq K. \quad \text{(87)} \]

For (31) we need to prove the impossibility of the following cases:

(i) \( p' < s_0 \); (ii) \( q' < t_0 \); (iii) \( p' + q' > s_0 + t_0 \).

We first consider the case (i). By (63), (74), and (82) we have

\[ 0 \geq L_{N_{k}+1}(p', q') - L_{N_{k}+1}(s_0, t_0) \]

\[ = N_{k} \log \left( 1 + \frac{\sigma_{N_{k}+1}(p', q') - \sigma_{N_{k}+1}(s_0, t_0)}{\sigma_{N_{k}+1}(s_0, t_0)} \right) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ \geq N_{k} \log \left( 1 + \frac{c^{(M, M)}(N_{k})}{N_{k}} \right) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ = \frac{c^{(M, M)}(N_{k})}{N_{k}} + N_{k} \cdot o \left( \frac{c^{(M, M)}(N_{k})}{N_{k}} \right) \]

\[ = \lambda_{\min}^{(M, M)}(N_{k}) \left( c + o(1) \right) \]

\[ \rightarrow \infty, \quad k \rightarrow \infty, \quad \text{(88)} \]

where \( c > 0 \) may depend on sample paths. The obtained contradiction ensures that \( p' \geq s_0 \). Similarly, we can prove that \( q' \geq t_0 \).

Finally, we consider the case (iii). From (82), (84), and (85) we have

\[ 0 \geq L_{N_{k}+1}(p', q') - L_{N_{k}+1}(s_0, t_0) \]

\[ = N_{k} \log \left( 1 + \frac{\sigma_{N_{k}+1}(p', q') - \sigma_{N_{k}+1}(s_0, t_0)}{\sigma_{N_{k}+1}(s_0, t_0)} \right) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ \geq N_{k} \log \left( 1 - cN_{k}^{1-4\delta} \right) + a_{N_{k}}(p' + q' - s_0 - t_0) \]

\[ = a_{N_{k}}(p' + q' - s_0 - t_0 - cN_{k}^{1-4\delta} a_{N_{k}}) \rightarrow \infty, \quad \text{(89)} \]

where \( c > 0 \) may depend on sample paths. Thus the case \( p' + q' > s_0 + t_0 \) is impossible, which in turn guarantees that (31) takes place. □

**References**


