Identification of Wiener Systems with Binary-Valued Output Observations

Yanlong Zhao, [a] Le Yi Wang, [b] G. George Yin, [c] Ji-Feng Zhang [a]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China.*
Email: ylzhao@amss.ac.cn. jif@iss.ac.cn.

[b] *Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan 48202, U.S.A.*
Email: lywang@ece.eng.wayne.edu.

[c] *Department of Mathematics, Wayne State University, Detroit, Michigan 48202, U.S.A. Email: gyin@math.wayne.edu.*

Abstract

This work is concerned with identification of Wiener systems whose outputs are measured by binary-valued sensors. The system consists of a linear FIR (Finite Impulse Response) subsystem of known order, followed by a nonlinear function with a known parametrization structure. The parameters of both linear and nonlinear parts are unknown. Input design, identification algorithms, and their essential properties are presented under the assumptions that the distribution function of the noise is known and the nonlinearity is continuous and invertible. It is shown that under scaled periodic inputs, identification of Wiener systems can be decomposed into a finite number of core identification problems. The concept of joint identifiability of the core problem is introduced to capture the essential conditions under which the Wiener system can be identified with binary-valued observations. Under scaled full-rank conditions and joint identifiability, a strongly convergent algorithm is constructed. The algorithm is shown to be asymptotically efficient for the core identification problem, hence achieving asymptotic optimality in its convergence rate. For computational simplicity, recursive algorithms are also developed.

Key words: Identification, binary-valued observations, Wiener systems, parameter estimation, sensor thresholds, periodic inputs, joint identifiability.

1 Introduction

Binary-valued sensors are commonly employed in practical systems since they are more cost effective than regular sensors. In some applications, they are the only ones available during real-time operations (Wang *et al.*, 2002a). More importantly, binary-valued observations are the fundamental building blocks for quantized observations that are an integrated part of communication channels.

Wiener systems are typical nonlinear systems, which represent a nonlinear dynamic system with a dynamic linear part, followed by a memoryless nonlinear function, shown schematically in Fig. 1. Wiener systems have been successfully used to represent systems in many practical applications such as biological systems (Hunter & Korenberg, 1986; Wang *et al.*, 2002b, 2004), signal processing, communications and control (Norquay *et al.*, 1999), etc.

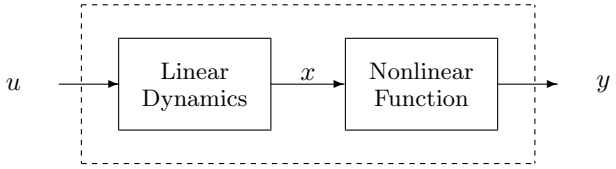

Fig. 1. Wiener systems

This paper focuses on identification of Wiener systems with binary-valued output observations. Since finite quantization may be regarded as a finite cascading of binary sensors, binary-valued observations are building blocks for quantization. When the output of a Wiener system must be measured by a binary-valued sensor or sent through a communication channel, it can be represented as a Wiener system with binary-valued or quantized output observations. Consequently, understanding identification of Wiener systems with binary-valued observations will be essential for studying both identification of nonlinear systems and impact of communication channels on system models. Binary-valued

observations supply very limited information on the system outputs, and hence introduce difficulties in system modeling, identification, and control. In Wang *et al.* (2003), we investigated the identification errors, time complexity, input design, and impact of disturbances and unmodeled dynamics on identification accuracy and complexity for linear systems that are modeled by impulse responses with binary-valued observations. The work was extended to rational models and unknown noise distributions in Wang *et al.* (2006a). Recently, the methodologies have been extended to system identification with quantized observations in Wang *et al.* (2006b, 2006c). Most significantly, the optimality of the identification algorithms has been established by showing the Cramér-Rao lower bound is asymptotically achieved Wang *et al.* (2006b).

Related works on identification with quantized output measurements can be found in Krishnamarty (1995) and Wigren (1995, 1998). Wigren (1995) considered the identification problem of linear IIR (Finite Impulse Response) systems with quantized output measurements and obtained local convergent parameter estimates by using a recursive search algorithm with approximate gradients. Based on this algorithm, Wigren (1998) studied linear FIR systems and gave global convergent identification results. Krishnamarty (1995) dealt with ARMA model with binary-valued observations, where the density of the noise is assumed to be symmetric about zero. However, the aforementioned research and our early investigations were limited to linear systems.

There have been substantial efforts in nonlinear system identification. The reader is referred to Sjoberg *et al.* (1995), Bai (2003), and Roll *et al.* (2005) for extensive exposition of the existing literature. Within nonlinear system identification, Wiener/Hammerstein structures have drawn much attention due to their structural simplicity and connections to linear systems (Hunter & Korenberg, 1986; Schoukens, 2003; Verhaegen & Westwick, 1996). Identification methodologies used for Wiener structures may be loosely classified by iterative algorithms (Hunter & Korenberg, 1986; Korenberg & Hunter, 1998), correlation techniques (Billings, 1980), least-squares estimation and singular value decomposition methods (Bai, 1998; Lacy & Bernstein, 2002), stochastic recursive algorithms (Chen, 2006; Hu & Chen, 2005), etc. Several identification algorithms were analyzed in Wigren (1994) for their convergence and error bounds. Frequency-domain identification methods for Wiener/Hammerstein structures were explored in Bai (1998) and Ninness and Gibson (2002). All these approaches require output measurements by regular sensors.

Our work in this paper has essential differences with previous research, mainly due to introduction of binary measurements into the system configuration. Interaction between the nonlinear sensor and the nonlinear subsystem imposes a challenge in relating empirical measures to system parameters and ensuring identifiability, which can be readily established for linear systems (Wang *et al.*, 2003). Different from the traditional gradient methods, we use the methods of empirical measures, periodic input design, and recursive algorithms to develop strongly convergent algorithms for Wiener system identification with binary-valued observations. One of the advantages of this approach is: We are able to establish the optimality of the algorithms by using the Fisher information. Assuming that the noise distribution function is known, it is shown that under scaled periodic inputs, identification of Wiener systems can be decomposed into a finite number of core identification problems. The concept of joint identifiability is introduced to capture the essential conditions, under which the Wiener system can be identified with binary-valued observations. Under joint identifiability and input full-rank conditions, a global convergent algorithm is constructed. The algorithm is shown to be asymptotically efficient for the core identification problem, hence achieving asymptotically optimal convergence rate. For computational simplicity, simplified recursive algorithms are also discussed.

The rest of the paper is organized as follows. The structure of Wiener systems using binary-valued observations is formulated in Section 2. It is shown in Section 3 that under scaled periodic inputs, identification of Wiener systems can be decomposed into a number of core identification problems. Basic properties of periodic signals and the concepts of joint identifiability are introduced in Section 4. Based on the algorithms for the core problems, Section 5 presents the main identification algorithms for Wiener systems. Under scaled full-rank inputs and joint identifiability, the identification algorithms for Wiener systems are shown to be strongly convergent (in the sense of convergence with probability one). Identification algorithms for the core problems are constructed in Section 6. By comparing the estimation errors with the Cramér-Rao lower bound, the algorithms are shown to be asymptotically efficient, hence achieving asymptotically optimal convergence speed. For simplicity, recursive algorithms are discussed in Section 7 that can be used to find system parameters under certain stability conditions. Illustrative examples are presented in Section 8 to demonstrate input design, identification algorithms, and convergence results of the methodologies discussed in this paper. Section 9 provides a brief summary of the findings of this paper.

2 Problem formulation

Consider the system in Fig. 2, in which

$$\begin{cases} x(k) = \sum_{i=0}^{n-1} a_i u(k-i), \\ y(k) = H(x(k), \eta) + d(k), \end{cases} \tag{1}$$

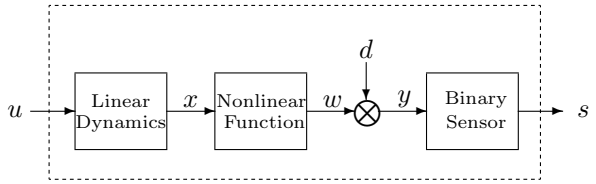

Fig. 2. Wiener systems with binary-valued observations

where $u(k)$ is the input, $x(k)$ the intermediate variable, and $d(k)$ the measurement noise. $H(\cdot,\eta)$: $\mathcal{D}_H \subseteq \mathbb{R} \to \mathbb{R}$, is a parameterized static nonlinear function with domain $\mathcal{D}_H$ and vector-valued parameter $\eta \in \Omega_\eta \subseteq \mathbb{R}^m$. Both n and m are known. By defining $\phi(k) = [u(k), \ldots, u(k-n+1)]^T$ and $\theta = [a_0, \ldots, a_{n-1}]^T$, the linear dynamics can be expressed compactly as $x(k) = \phi(k)^T\theta$.

Assumption A1. The noise $\{d(k)\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with finite variance. The distribution function $F(\cdot)$ of $d(1)$ is known, which is continuously differentiable together with a continuously differentiable inverse $F^{-1}(\cdot)$ and a bounded density $f(\cdot)$ with $f(x) \neq 0$ for $x \neq 0$.

Assumption A2. For any given $\eta \in \Omega_\eta$, $H(x,\eta)$ is bounded for any finite x, continuous and invertible in x.

The output $y(k)$ is measured by a binary sensor with threshold C. That is, the sensor output $s(k) = \mathcal{S}(y(k))$ is a function of $y(k)$ indicating only whether $y(k) \leq C$ or $y(k) > C$, where C is known. We use the indicator function

$$s(k) = \mathcal{S}(y(k)) = I_{\{y(k)\leq C\}} = \begin{cases} 1, & \text{if } y(k) \leq C, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

to represent the sensor.

Remark 1. The threshold C has significant impact on identification accuracy. For the binary series estimation algorithm employed in Wigren (1995), it was shown that for improving identification accuracy, it is necessary to have $C \neq 0$ and is desirable that signal energy is centered around C. In Wang *et al* (2003), $C \neq 0$ is also shown to be required in the worst-case identification framework. However, this constraint is no longer relevant under the stochastic framework introduced in Wang *et al.* (2003).

This paper employs the approach of empirical measures. While this approach requires the knowledge of noise distribution functions (Wang *et al.*, 2003), or its estimation (Wang *et al.*, 2006a), the essential requirement is that the input and C are selected such that $C - y(k)$ lies within the support of the noise density $f(\cdot)$ (otherwise, $s_k \equiv 0$ or $s_k \equiv 1$, w.p.1., and the sensor does not provide useful information on y_k). If $f(\cdot)$ has infinite support, such as the normal distribution, theoretically any C is valid. However, the Cramér-Rao lower bound, which will be derived subsequently, will characterize precisely the impact of C selection on identification accuracy in this approach. In other words, it is desirable to design C and inputs to minimize the Cramér-Rao lower bound. Threshold selection is studied in depth in (Wang *et al.*, 2006c) and will not be explored further in this paper.

Parametrization of the static nonlinear function $H(\cdot,\eta)$ depends on specific applications. Often, the structures of actual systems can provide guidance in selecting function forms whose parameters carry physical meanings (Wang *et al.*, 2002b, 2004). On the other hand, when a black-box approach is employed, namely, the models represent input/output relationships based on data only, one may choose some generic structures for theoretical and algorithm development. For instance, a common structure is $H(x,\eta) = \sum_{i=0}^{m-1} b_i h_i(x)$, where $h_i(x), i = 0, \ldots, m-1$, are base functions and $\eta = [b_0, b_1, \ldots, b_{m-1}]^T \in \mathbb{R}^m$ is a vector of m unknown parameters. For example, the typical polynomial structure is

$$H(x,\eta) = \sum_{i=0}^{m-1} b_i x^i. \tag{3}$$

In this paper, we will discuss input design, derive joint estimators of θ and η, and establish their identifiability, convergence, convergence rates, and efficiency (optimality in convergence rate).

3 Basic input design and core identification problems

We first outline the main ideas of using $2n(m+1)$-scaled periodic inputs and empirical measures to identify Wiener systems under binary-valued observations. It will be shown that this approach leads to a core identification problem, for which identification algorithms and their key properties will be established.

Unlike adaptive controls, the purpose of this paper is only on parameter estimation. And the key rule for selecting inputs is to provide sufficient information to support identification accuracy. The persistent and decaying excitation conditions described in, for instance, Chen and Guo (1991), are typical conditions for traditional identification algorithms to ensure convergence and consistency of the parameter estimators. It is well established that there are many classes of persistent excitation signals. The periodic inputs are particularly useful in supporting the identification method of this paper, although theoretically many other signals can potentially be used also. Not only they provide sufficient information, but they lead to much simplified identification algorithms and well established convergence properties.

The input signal, that will be used to identify the system, is a $2n(m+1)$-periodic signal u whose one-period values are $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_m v, \rho_m v)$, where $v = (v_1, \ldots, v_n)$ is to be specified. [1] The scaling factors $\{\rho_0, \rho_1, \ldots, \rho_m\}$ are assumed to be nonzero and distinct.

If under the $2n$ input values $u = (v, v)$, the linear subsystem has the following n consecutive output values at $n, \ldots, 2n-1$

$$\delta_i = a_0 u(n+i) + \cdots + a_{n-1} u(1+i), \quad i = 0, \ldots, n-1,$$

then the output under the scaled input (qv, qv) is

$$x(n+i) = q\delta_i, \quad i = 0, \ldots, n-1.$$

Without loss of generality, assume $\delta_0 \neq 0$. [2] The output of the linear subsystem contains the following $(m+1)$-periodic subsequence with its single period values $\{\rho_0\delta_0, \rho_1\delta_0, \ldots, \rho_m\delta_0\}$:

$$x(n) = \rho_0\delta_0, \quad x(3n) = \rho_1\delta_0, \quad \ldots$$
$$x((2m+1)n) = \rho_m\delta_0, \ \ldots$$

By concentrating on this subsequence of $x(k)$, under a new index l with $l = 1, 2, \ldots$, the corresponding output of the nonlinear part may be rewritten as

$$\begin{aligned} \widetilde{y}(l(m+1)) &= H(\rho_0\delta_0, \eta) + \widetilde{d}(l(m+1)), \\ \widetilde{y}(l(m+1)+1) &= H(\rho_1\delta_0, \eta) + \widetilde{d}(l(m+1)+1), \\ &\vdots \\ \widetilde{y}(l(m+1)+m) &= H(\rho_m\delta_0, \eta) + \widetilde{d}(l(m+1)+m). \end{aligned} \tag{4}$$

The equations in (4) form the basic observation relationship for identifying η and δ_0.

For $\rho = [\rho_0, \ldots, \rho_m]^T$ and a scalar δ, we denote

$$\mathbf{H}(\rho\delta, \eta) = [H(\rho_0\delta, \eta), \ldots, H(\rho_m\delta, \eta)]^T. \tag{5}$$

Then, (4) can be expressed as

$$\widetilde{Y}(l) = \mathbf{H}(\rho\delta, \eta) + \widetilde{D}(l), \quad l = 0, 1, \ldots, \tag{6}$$

where $\delta \neq 0$, $\widetilde{Y}(l) = [\widetilde{y}(l(m+1)), \ldots, \widetilde{y}(l(m+1)+m)]^T$ and $\widetilde{D}(l) = [\widetilde{d}(l(m+1)), \ldots, \widetilde{d}(l(m+1)+m)]^T$. Correspondingly, the outputs of the binary-valued sensor on $\widetilde{Y}(l)$ are $\widetilde{S}(l) = \mathcal{S}(\widetilde{Y}(l))$, $l = 0, 1, \ldots$.

[1] The reason for repeating each scaled vector, such as $\rho_0 v, \rho_0 v$, etc., is to simplify algorithm development and convergence analysis, not a fundamental requirement.

[2] It will become clear that when v is full rank, to be discussed later, there exists at least one i such that $\delta_i \neq 0$.

Let $\tau = [\tau_0, \ldots, \tau_m]^T \triangleq [\delta, \eta^T]^T$. We introduce the following identification problem.

Core Identification Problem: Estimate the parameter τ from observations on $\widetilde{S}(l)$.

Denote $\zeta_i = H(\rho_i\delta, \eta), i = 0, 1, \ldots, m$. Then $\zeta = [\zeta_0, \ldots, \zeta_m]^T = \mathbf{H}(\rho\delta, \eta)$ and (6) can be rewritten as

$$\widetilde{Y}(l) = \zeta + \widetilde{D}(l), \quad l = 0, 1, \ldots \tag{7}$$

The main idea of solving the core identification problem is first to estimate ζ, and then to solve the interpolation equations

$$\zeta_i = H(\rho_i\delta, \eta), \quad i = 0, 1, \ldots, m \tag{8}$$

for δ and η. The basic properties on signals and systems that ensure solvability of the core identification problem will be discussed next.

4 Properties of inputs and systems

We first establish some essential properties of periodic signals and present the idea of joint identifiability, which will play an important role in the subsequent development. Some related ideas can be found in Horn and Johnson (1985), Lancaster and Tismenetsky (1985), Wang *et al.* (2006a).

4.1 Generalized circulant matrices and periodic inputs

An $n \times n$ generalized circulant matrix (Lancaster & Tismenetsky, 1985)

$$T = \begin{bmatrix} v_n & v_{n-1} & v_{n-2} & \cdots & v_1 \\ qv_1 & v_n & v_{n-1} & \ddots & v_2 \\ qv_2 & qv_1 & v_n & \ddots & v_3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ qv_{n-1} & qv_{n-2} & qv_{n-3} & \cdots & v_n \end{bmatrix} \tag{9}$$

is completely determined by its first row $[v_n, \ldots, v_1]$ and q, which will be denoted by $T(q, [v_n, \ldots, v_1])$. In the special case of $q = 1$, the matrix in (9) is called circulant matrix and will be denoted by $T([v_n, \ldots, v_1])$.

Definition 1. An n-periodic signal generated from its single-period values $(v_1, \ldots, v_n)$ is said to be *full rank* if the circulant matrix $T([v_n, \ldots, v_1])$ is full rank.

An important property of circulant matrices is the following frequency-domain criterion.

Lemma 1. (Horn & Johnson, 1985) *If* $T = T(q, [v_n, \ldots,$

$v_1]$) *is a generalized circulant matrix, then the eigenvalues of* T *are* $\{q\gamma_k, k = 1, \ldots, n\}$ *and the determinant of* T *is*

$$\det(T) = \prod_{k=1}^{n} q\gamma_k, \tag{10}$$

where γ_k *is the discrete Fourier transform (DFT) of* $v_j q^{-\frac{j}{n}}, j = 1, ..., n$.

$$\gamma_k = \sum_{j=1}^{n} v_j q^{\frac{-j}{n}} e^{-i\omega_k j}, \quad \omega_k = \frac{2\pi k}{n}, \quad k = 1, \ldots, n.$$

Hence, T *is full rank if and only if* $\gamma_k \neq 0$, $k = 1, \ldots, n$.

Proof. Let $P = \begin{bmatrix} 0 & I_{n-1} \\ q & 0 \end{bmatrix}$, whose characteristic polynomial is $\det(\lambda I_n - P) = \lambda^n - q$ and eigenvalues are $q^{\frac{1}{n}} e^{i\omega_k}, k = 1, \ldots, n$. Then T can be represented by $T = \sum_{j=1}^{n} v_j P^{n-j}$. For P and $k = 1, \ldots, n$, if λ_k is the corresponding eigenvector of $q^{\frac{1}{n}} e^{i\omega_k}$, then

$$T\lambda_k = \sum_{j=1}^{n} v_j P^{n-j} \lambda_k = \sum_{j=1}^{n} v_j (q^{\frac{1}{n}} e^{i\omega_k})^{n-j} \lambda_k = q\gamma_k \lambda_k.$$

Therefore, $q\gamma_k$ is an eigenvalue of T and (10) is confirmed. By hypothesis, $q \neq 0$. Hence T is full rank if and only if $\gamma_k \neq 0$, $k = 1, \ldots, n$. □

For the special case of $q = 1$, we have the following property.

Corollary 1. *An* n*-periodic signal generated from* $v = (v_1, \ldots, v_n)$ *is full rank if and only if its discrete Fourier transform* $\gamma_k = V(\omega_k) = \sum_{j=1}^{n} v_j e^{-i\omega_k j}$ *is nonzero at* $\omega_k = \frac{2\pi k}{n}$, $k = 1, \ldots, n$.

Recall that $\mathcal{F}[v] = \{\gamma_1, \ldots, \gamma_n\}$ is the frequency samples of the n-periodic signal u, where $\mathcal{F}[\cdot]$ is the discrete Fourier transform. Hence, Definition 1 may be equivalently stated as "an n-periodic signal v is said to be full rank if its frequency samples do not contain 0." In other words, the signal contains n nonzero frequency components.

Definition 2. A $2n(m+1)$-periodic signal u is called a scaled full rank signal if its single-period values are $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_m v, \rho_m v)$, where $v = (v_1, \ldots, v_n)$ is full rank, i.e., $0 \notin \mathcal{F}[v]$; $\rho_j \neq 0$, $j = 1, \ldots, m$, and $\rho_i \neq \rho_j$, $i \neq j$. We use $\mathcal{U}$ to denote the class of such signals.

Definition 3. An $n(m+1)$-periodic signal u is called an exponentially scaled full rank signal if its single-period values are $(v, qv, \ldots, q^m v)$, where $q \neq 0$ and $q \neq 1$, and $v = (v_1, \ldots, v_n)$ is full rank. We use $\mathcal{U}_e$ to denote this class of input signals.

4.2 Joint identifiability

Joint identifiability conditions mandate that the unknown parameters δ and η can be uniquely and jointly determined by the interpolation conditions (8).

Prior information. The prior information on the unknown parameters $\tau = [\delta, \eta^T]^T$ for the core identification problem is $\tau \in \Omega \subseteq \mathbb{R}^{m+1}$. Denote $\mathbb{R}_d^m = \{\rho = [\rho_1, \ldots, \rho_m]^T \in \mathbb{R}^m : \rho_j \neq 0, \forall j; \rho_i \neq \rho_j, i \neq j\}$, namely the set of all vectors in $\mathbb{R}^m$ that contain non-zero and distinct elements.

Definition 4. Suppose that $\Upsilon \subseteq \mathbb{R}_d^{m+1}$. $H(x;\eta)$ is said to be *jointly identifiable* in Ω with respect to Υ, if for any $\rho = [\rho_0, \ldots, \rho_m]^T \in \Upsilon$, $\mathbf{H}(\rho\delta;\eta)$ is invertible in Ω, namely $\zeta = \mathbf{H}(\rho\delta;\eta)$ has a unique solution $\tau \in \Omega$. In this case, elements in Υ are called *sufficiently rich scaling factors.*

Depending on the parametric function forms $H(\cdot,\eta)$ and the domain $\mathcal{D}_H$, the set of sufficiently rich scaling factors can vary significantly. For example, the polynomial class of functions of a fixed order has a large set Υ. The polynomial class has been used extensively as the nonlinear part of Wiener systems and their approximations in Celka *et al.* (2001), Norquay *et al.* (1999), Wigren (1994).

When the base functions are polynomials of order m, $H(x,\eta)$ can be expressed as

$$H(x,\eta) = \sum_{j=0}^{m} b_j x^j, \text{ with } b_m \neq 0. \tag{11}$$

Then $H(\rho_i\delta,\eta) = \sum_{j=0}^{m} b_j \delta^j \rho_i^j$, $i = 0, 1, \ldots, m$. Apparently, one cannot uniquely determine $m+2$ parameters $\delta, b_0, \ldots, b_m$ from $m+1$ coefficients of the polynomial. A typical remedy to this well-known fact is normalization of the parameter set by assuming one parameter, say, $b_l = 1$ for some l. In this case, the coefficient equations become $b_j\delta^j = c_j, j \neq l$ and $\delta^l = c_l$. For given c_j, to ensure uniqueness of solutions $b_j, j \neq l$ and δ to the equations, l must be an odd number.

We now show that $H(x,\eta)$ that satisfies Assumption A2 contains at least one non-zero odd-order term. Indeed, if $H(x,\eta)$ contains only even-order terms, it must be an even function. It follows that $H(x,\eta) = H(-x,\eta)$, namely it is not an invertible function. This contradicts Assumption 2.

Since $H(x,\eta)$ contains at least one non-zero odd-order term $b_l x^l$ for some odd integer l, without loss of gener-

ality we assume $b_l = 1$. The reduced parameter vector is $\eta_0 = [b_0, \ldots, b_{l-1}, b_{l+1}, \ldots, b_m]^T$, which contains only m unknowns. Such polynomials will be called "normalized polynomial functions of order m."

Proposition 1. *Under Assumption A2, all normalized polynomial functions of order m are jointly identifiable with respect to $\mathbb{R}_d^m$.*

Proof. For any given $\rho = [\rho_0, \ldots, \rho_m] \in \mathbb{R}_d^{m+1}$, the interpolation equations

$$\sum_{j=0}^{m} c_j \rho_i^j = \zeta_i, \text{ for } i = 0, \ldots, m \tag{12}$$

can be rewritten as $\Re c = \zeta$, where ζ is defined in (7) and

$$\Re = \begin{pmatrix} 1 & \rho_0 & \cdots & \rho_0^m \\ 1 & \rho_1 & \ddots & \rho_1^m \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \rho_m & \cdots & \rho_m^m \end{pmatrix}, \; c = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_m \end{pmatrix}.$$

Since the determinant of the Vandermonde matrix

$$\det \Re = \prod_{0 \le i < j \le m-1} (\rho_j - \rho_i) \neq 0$$

for distinct $\rho_i, i = 0, \ldots, m-1$, we have $c = \Re^{-1}\zeta$. Furthermore, the equation $\delta^l = c_l$ yields the unique solution $\delta = (c_l)^{1/l} \neq 0$ by hypothesis. Then, $b_j = c_j/\delta^j$, $j \neq l$, solve uniquely for the remaining parameters. Consequently, $\mathbf{H}(\rho\delta; \eta_0)$ is invertible as a joint function of δ and η_0. This implies that $H(x, \eta_0)$ is jointly identifiable with respect to any vector in $\mathbb{R}_d^m$. □

Other basis can also be used. For instance, $H(x, \eta) = \eta + e^x$, where $\eta \neq 0$. Under the prior information $\Omega = \{[\delta, \eta^T]^T : \delta > 0, \eta \neq 0\}$, consider $\Upsilon = \{(\rho_0, \rho_1) : \rho_0 > 0, \rho_1 < 0\}$. The interpolation equations are

$$\begin{cases} \eta + e^{\rho_0 \delta} = \zeta_0 \\ \eta + e^{\rho_1 \delta} = \zeta_1. \end{cases} \tag{13}$$

These imply

$$e^{\rho_0 \delta} - e^{\rho_1 \delta} = \zeta_0 - \zeta_1. \tag{14}$$

It is easily seen that for $\rho_0 > 0$ and $\rho_1 < 0$, the derivative of (14) is $\frac{\mathrm{d}(e^{\rho_0\delta} - e^{\rho_1\delta})}{\mathrm{d}\delta} = \rho_0 e^{\rho_0\delta} - \rho_1 e^{\rho_1\delta} > 0$. Hence, (14) has a unique solution, which indicates that $H(x, \eta)$ is jointly identifiable with respect to Υ.

Joint identifiability is certainly not a trivial condition. For the above function form $H(x, \eta) = \eta + e^x$, Υ cannot be expanded to $\mathbb{R}_d^2$. Indeed, if one selects $\rho_0 = -2$, $\rho_1 = -1$, $\zeta_0 = 1.075$, $\zeta_1 = 1.2$, then (13) becomes $\begin{cases} \eta + e^{-2\delta} = 1.075 \\ \eta + e^{-\delta} = 1.2, \end{cases}$ Both $\delta = 1.921$, $\eta = 1.054$ and $\delta = 0.158$, $\eta = 0.346$ solve the equations. By definition, $H(x, \eta) = \eta + e^x$ is not jointly identifiable with respect to $\mathbb{R}_d^2$.

5 Identification algorithms

Based on periodic inputs and joint identifiability, we now derive algorithms for parameter estimates and prove their convergence.

Assumption A3.

i) The prior information on θ and η is that $\theta \neq 0$, $\eta \neq 0$, $\theta \in \Omega_\theta$ and $\eta \in \Omega_\eta$ such that under Ω_θ and Ω_η, the set Υ of sufficiently rich scaling factors is non-empty. $C - y(k)$ lies within the support of the noise density $f(\cdot)$ for $k = 1, 2, \ldots$

ii) $H(x, \eta)$ is jointly identifiable with respect to Υ and continuously differentiable with respect to both x and η.

By using the vector notation, for $j = 1, 2, \ldots$,

$$\begin{aligned} X(j) &= [\, x(2(j-1)(m+1)n + n), \ldots, \\ &\qquad x(2j(m+1)n + n - 1)]^T, \\ Y(j) &= [\, y(2(j-1)(m+1)n + n), \ldots, \\ &\qquad y(2j(m+1)n + n - 1)]^T, \\ \widetilde{\Phi}(j) &= [\, \phi(2(j-1)(m+1)n + n), \ldots, \\ &\qquad \phi(2j(m+1)n + n - 1)]^T, \\ D(j) &= [\, d(2(j-1)(m+1)n + n), \ldots, \\ &\qquad d(2j(m+1)n + n - 1)]^T, \\ S(j) &= [\, s(2(j-1)(m+1)n + n), \ldots, \\ &\qquad s(2j(m+1)n + n - 1)]^T, \end{aligned} \tag{15}$$

the observations can be rewritten in block form as

$$\begin{cases} Y(j) = \mathbf{H}(X(j), \eta) + D(j), \\ X(j) = \widetilde{\Phi}(j)\theta. \end{cases}$$

The input is a scaled $2n(m+1)-$periodic signal with single period values

$$(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_m v, \rho_m v),$$

where $v = (v_1, \ldots, v_n)$ is full rank.

By periodicity, $\widetilde{\Phi}(j) = \widetilde{\Phi}$, for all j and $\widetilde{\Phi}$ can be decomposed into $2(m+1)$ submatrices Φ_i, $i = 1, \ldots, 2(m+1)$, of dimension $n \times n$: $\widetilde{\Phi} = [\Phi_1^T, \Phi_2^T, \ldots, \Phi_{2(m+1)}^T]^T$. Denote the $n \times n$ circulant matrix $\Phi = T([v_n, \ldots, v_1])$. Then the odd-indexed block matrices [3] satisfy the simple scaling relationship

$$\Phi_1 = \rho_0 \Phi, \quad \Phi_3 = \rho_1 \Phi, \quad \ldots, \quad \Phi_{2m+1} = \rho_m \Phi. \tag{16}$$

Remark 2. In $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_m v, \rho_m v)$, there are always two identical subsequences $\rho_i v, i = 0, \ldots, m$ appearing consecutively. The main reason for this input structure is to generate block matrices that satisfy the above scaling relationship (16).

Remark 3. We use the following notation for element-wise vector functions. For a scalar function $g(\cdot)$ and a vector $x = [x_1, \ldots, x_l]^T \in \mathbb{R}^l$, the boldface symbol $\mathbf{g}(x)$ represents

$$\mathbf{g}(x) = [g(x_1), \ldots, g(x_l)]^T \in \mathbb{R}^l. \tag{17}$$

In addition, if $g(x)$ is invertible, $\mathbf{g}^{-1}(x)$ represents the component-wise inverse

$$\mathbf{g}^{-1}(x) = [g^{-1}(x_1), \ldots, g^{-1}(x_l)]^T \in \mathbb{R}^l. \tag{18}$$

Similarly, for $\alpha = [\alpha_1, \ldots, \alpha_l]^T \in \mathbb{R}^l$ and $c = [c_1, \ldots, c_l]^T \in \mathbb{R}^l$, we use the vector notation $\mathbf{I}_{\{\alpha \le c\}} = [I_{\{\alpha_1 \le c_1\}}, \ldots, I_{\{\alpha_l \le c_l\}}]^T$. $\mathbb{1}_\ell$ and $\mathbf{0}_\ell \in \mathbb{R}^\ell$ will denote column vectors with all components being 1 and 0, respectively. For a given threshold C, $\mathbf{C}_l = C\mathbb{1}_l \in \mathbb{R}^l$.

5.1 Identification algorithms for the core problem

For the core problem (7), let

$$\begin{aligned}\widetilde{z}(N) &= \frac{1}{N}\sum_{l=0}^{N-1} \widetilde{S}(l) \\ &= \frac{1}{N}\sum_{l=0}^{N-1} \mathbf{I}\{\widetilde{D}(l) \le \mathbf{C}_{m+1} - \mathbf{H}(\rho\delta, \eta)\},\end{aligned}$$

which is the empirical distribution of $\widetilde{D}(k)$ at $\mathbf{C}_{m+1} - \mathbf{H}(\rho\delta, \eta)$. Define [4]

$$\widetilde{\xi}(N) = \begin{cases} \widetilde{z}(N), & \text{if} \quad 0 < \widetilde{z}(N) < 1; \\ \frac{1}{N}, & \text{if} \quad \widetilde{z}(N) = 0; \\ \frac{N-1}{N}, & \text{if} \quad \widetilde{z}(N) = 1. \end{cases} \tag{19}$$

Then, by the strong law of large numbers,

$$\widetilde{\xi}(N) \to p = \mathbf{F}(\mathbf{C}_{m+1} - \mathbf{H}(\rho\delta, \eta)), \quad \text{w.p.1.} \tag{20}$$

By Assumption A1, F has a continuous inverse. Hence,

$$\begin{aligned}&\zeta(N) = \mathbf{C}_{m+1} - \mathbf{F}^{-1}(\widetilde{\xi}(N)) \\ &\quad \to \zeta = \mathbf{C}_{m+1} - \mathbf{F}^{-1}(p) = \mathbf{H}(\rho\delta, \eta) \ \text{ w.p.1.}\end{aligned}$$

By Assumption A3, $\mathbf{H}$ is invertible as a function of $\tau = [\delta, \eta^T]^T$. As a result, $\tau(N) = \mathbf{H}^{-1}(\zeta(N)) \to \tau$ w.p.1. In summary, we have the following theorem.

Theorem 1. *Under Assumptions A1-A3, let* $\tau(N) = \mathbf{H}^{-1}(\zeta(N)) = \mathbf{H}^{-1}(\mathbf{C}_{m+1} - \mathbf{F}^{-1}(\widetilde{\xi}(N)))$. *Then*

$$\tau(N) \to \tau \ \ \textit{w.p.1 as} \ \ N \to \infty. \tag{21}$$

Proof. Under Assumptions A1 and A2, $\mathbf{H}^{-1}$ and $\mathbf{F}^{-1}$ are continuous. By the above analysis, we have

$$\begin{aligned}&\tau(N) = \mathbf{H}^{-1}(\mathbf{C}_{m+1} - \mathbf{F}^{-1}(\widetilde{\xi}(N))) \\ &\to \mathbf{H}^{-1}(\mathbf{C}_{m+1} - \mathbf{F}^{-1}(p)) = \mathbf{H}^{-1}(\zeta) = \tau \ \text{ w.p.1.} \ \square\end{aligned}$$

5.2 Parameter estimates of the original problem

Parameter estimates are generated as follows. Define $z(N) = \frac{1}{N}\sum_{l=0}^{N-1} S(l)$ and

$$\xi(N) = \begin{cases} z(N), & \text{if} \quad 0 < z(N) < 1; \\ \frac{1}{N}, & \text{if} \quad z(N) = 0; \\ \frac{N-1}{N}, & \text{if} \quad z(N) = 1. \end{cases} \tag{22}$$

Then, the strong law of large numbers yields that

$$\xi(N) \to \xi = \mathbf{F}(\mathbf{C}_{2(m+1)n} - \mathbf{H}(\widetilde{\Phi}\theta, \eta)) \ \text{ w.p.1.} \tag{23}$$

[3] The even-indexed block matrices are not be used in the proof.

[4] This modification is to avoid the points $\widetilde{z}(N) = 0$ or $\widetilde{z}(N) = 1$ since the distribution function $F(\cdot)$ is not invertible at these points. Since the probability of of these points are asymptotically zero as $N \to \infty$, the consequent analysis and conclusions will not be affected by this modification. As a result, this modification will not be explicitly stated in the subsequent proofs and development.

Equations in (23) for system (1) contain the following equations by extracting the odd-indexed blocks

$$\mathbf{H}(\rho_j \Phi\theta; \eta) = \mathbf{C}_n - \mathbf{F}^{-1}(\xi^{2j}),\ j = 0, \ldots, m.$$

We now show that this subset of equations are sufficient to determine θ and η uniquely.

Theorem 2. *Suppose $u \in \mathcal{U}$. Under Assumptions A1-A3,*

$$\xi = \mathbf{F}(\mathbf{C}_{2n(m+1)} - \mathbf{H}(\widetilde{\Phi}\theta, \eta)) \tag{24}$$

has a unique solution (θ^*, η^*).

Proof. Consider the first block $\Phi_1\theta$ of $\widetilde{\Phi}\theta$. Since v is full rank, Φ_1 is a full rank matrix. It follows that for any nonzero θ, $\Phi_1\theta \neq \mathbf{0}_n$. Without loss of generality, suppose that the i^*th element δ of $\Phi_1\theta$ is nonzero. By construction of $\widetilde{\Phi}$, we can extract the following m nonzero elements from $\widetilde{\Phi}\theta$: the $(2nl+i^*)$th element, $l = 0, \ldots, m$, is $\rho_l\delta$. Extracting these rows from the equation $\mathbf{H}(\widetilde{\Phi}\theta, \eta) = \mathbf{C}_{2n(m+1)} - \mathbf{F}^{-1}(\xi)$ leads to a core problem

$$\mathbf{H}(\rho\delta, \eta) = \mathbf{C}_n - \mathbf{F}^{-1}(\widetilde{\xi}), \tag{25}$$

where $\rho = [\rho_0, \rho_1, \ldots, \rho_m]^T$. Since $\delta \neq 0$ and ρ has distinct elements, $\rho\delta$ has distinct elements. By hypothesis, $H(x; \eta)$ is jointly identifiable. It follows that (25) has a unique solution (δ^*, η^*).

From the derived η^*, we denote the first n equations of $\mathbf{H}(\widetilde{\Phi}\theta, \eta) = \mathbf{C}_{2n(m+1)} - \mathbf{F}^{-1}(\xi)$ by

$$\mathbf{H}(\Phi\theta, \eta^*) = \mathbf{C}_n - \mathbf{F}^{-1}(\xi^1). \tag{26}$$

By Assumption A2, $\mathbf{H}^{-1}(x; \eta^*)$ exists (as a function of x). Since v is full rank, $\Phi = T([v_n, \ldots, v_1])$ is invertible. As a result, $\theta^* = \Phi^{-1}\mathbf{H}^{-1}(\mathbf{C}_n - \mathbf{F}^{-1}(\xi^1), \eta^*)$ is the unique solution to (26). This completes the proof. □

A particular choice of the scaling factors ρ_j is $\rho_j = q^j$, $j = 0, 1, \ldots, m$ for some $q \neq 0$ and $q \neq 1$. In this case, the period of input u can be shortened to $n(m+1)$ under a slightly different condition.

Let $\xi(N)$ be defined as in (22), with dimension changed from $2n(m+1)$ to $n(m+1)$. By the strong law of large numbers, as $N \to \infty$,

$$\xi(N) \to \xi = \mathbf{F}(\mathbf{C}_{n(m+1)} - \mathbf{H}(\Phi'\theta, \eta)) \ \text{ w.p.1} \tag{27}$$

for some $(n(m+1)) \times n$ matrix Φ'. Partition Φ' into $(m+1)$ submatrices Φ'_i, $i = 1, \ldots, m+1$, of dimension $n \times n$:

$$\Phi' = [(\Phi'_1)^T, (\Phi'_2)^T, \ldots, (\Phi'_{m+1})^T]^T. \tag{28}$$

If $u \in \mathcal{U}_e$, then it can be directly verified that $\Phi'_{l+1} = q^l\Phi' = q^l T(q, [v_n, \ldots, v_1])$, $l = 0, 1, \ldots, m$. We have the following result, whose proof is similar to that of Theorem 2 and hence is omitted.

Theorem 3. *Suppose $u \in \mathcal{U}_e$. Under Assumptions A1-A3,*

$$\xi = \mathbf{F}(\mathbf{C}_{n(m+1)} - \mathbf{H}(\Phi'\theta, \eta)) \tag{29}$$

has a unique solution (θ^*, η^*).

5.3 Identification algorithms and convergence of estimates

The $\xi(N) = [\xi_0(N), \ldots, \xi_{2n(m+1)-1}(N)]^T$ in (22) has $2n(m+1)$ components for a scaled full rank signal $u \in \mathcal{U}$. But there are only $n+m$ unknown parameters. Consider $\Phi\theta = [\delta_0, \ldots, \delta_{n-1}]^T$. We separate the components to n groups, for $i = 0, \ldots, n-1$, $\varepsilon^i(N) = [\xi_i(N), \xi_{i+2n}(N), \ldots, \xi_{i+2nm}(N)]^T$. Let $\delta_i(N)$ and $\eta_i(N)$ satisfy

$$\begin{aligned} \varepsilon^i(N) &= [\varepsilon^i_0(N), \ldots, \varepsilon^i_m(N)]^T \\ &= \mathbf{F}(\mathbf{C}_{m+1} - \mathbf{H}(\rho\delta_i(N), \eta_i(N))). \end{aligned} \tag{30}$$

Then, by (23) we have

$$\varepsilon^i(N) \to \varepsilon_i = \mathbf{F}(\mathbf{C}_{m+1} - \mathbf{H}(\delta_i\rho, \eta)). \tag{31}$$

If $\delta_i \neq 0$, (31) becomes a core identification problem. Furthermore, since $\theta \neq \mathbf{0_n}$ and Φ is full rank, there exists i^* such that $\delta_{i^*} \neq 0$. The identification algorithms include the following steps:

1: Calculate $i^* = \text{argmax}_i|\delta_i|$ to choose nonzero δ_{i^*}. If there exists $j \neq k$ such that $\varepsilon^i_j(N) = \varepsilon^i_k(N)$, then let $\delta_i(N) = 0$ and $\eta_i(N) = \mathbf{0_m}$. Otherwise, $\delta_i(N)$ and $\eta_i(N)$ are solved from (30). Let $i^*(N) = \text{argmax}_i|\delta_i(N)|$, where "argmax" means the argument of the maximum.
2: Estimate η from core identification problem. $\eta(N) = \eta_{i^*}(N)$.
3: Estimate θ. $\theta(N) = \Phi^{-1}\mathbf{H}^{-1}(\mathbf{C}_n - \mathbf{F}^{-1}(\xi^*(N)), \eta(N))$, where $\xi^*(N) = [\xi_0(N), \xi_1(N), \ldots, \xi_{n-1}(N)]^T$.

Theorem 4. *Suppose $u \in \mathcal{U}$. Under Assumptions A1, A2, and A3,*

$$\theta(N) \to \theta,\ \eta(N) \to \eta \ \ \textit{w.p.1 as}\ \ N \to \infty. \tag{32}$$

Proof. By Assumption A2, $\delta_i(N)$ and $\eta_i(N)$ can be solved from step 1. By core identification problems, if $\delta_i \neq 0$, $\delta_i(N) \to \delta_i$ w.p.1 as $N \to \infty$. Hence,

$$i^*(N) = \text{argmax}_i|\delta_i(N)| \to i^* = \text{argmax}_i|\delta_i|, \quad \text{w.p.1.}$$

Since there exists $\delta_i \neq 0$, we have $\delta_{i^*} \neq 0$. By (21), we have $\delta(N) \to \delta_{i^*}$, $\eta(N) \to \eta$, as w.p.1 as $N \to \infty$. For $\xi^*(N) = [\xi_0(N), \xi_1(N), \ldots, \xi_{n-1}(N)]^T$, $\xi^*(N) \to \xi^* = \mathbf{F}(\mathbf{C}_n - \mathbf{H}(\Phi\theta, \eta))$ w.p.1, so as $N \to \infty$,

$$\theta(N) = \Phi^{-1}\mathbf{H}^{-1}(\mathbf{C}_n - \mathbf{F}^{-1}(\xi^*(N)), \eta(N))$$
$$\to \Phi^{-1}\mathbf{H}^{-1}(\mathbf{C}_n - \mathbf{F}^{-1}(\xi^*), \eta) = \theta, \quad \text{w.p.1.}$$

□

Similarly, for an exponentially scaled full rank signal $u \in \mathcal{U}_e$, the identification algorithms can be constructed and its convergence can be derived similarly.

6 Asymptotic efficiency of the core identification algorithms

The identification of the core problem contains the main idea of the algorithms constructed in Section 5. In this section, the efficiency of the core identification algorithms will be established by comparing the error variance with the Cramér-Rao lower bound.

6.1 Asymptotic analysis of empirical measures

Suppose that $F_N(x)$ is the N-sample empirical distribution of the noise d at $x \in \mathbb{R}$. Let $\nu_N(x) = \sqrt{N}(F_N(x) - F(x))$.

Lemma 2. *Under Assumption A1, the following assertions hold.*

a) For any compact subset $S \subset \mathbb{R}$, $\sup_{x\in S} |F_N(x) - F(x)| \to 0$ w.p.1 as $N \to \infty$.
b) $\nu_N(\cdot)$ converges weakly to $\nu(\cdot)$, a stretched Brownian bridge process such that the covariance of $\nu(\cdot)$ is given by $E\nu(x)\nu(y) = \min\{F(x), F(y)\} - F(x)F(y), \forall x, y \in \mathbb{R}$.

Remark 4. In the above, Assertion a) is the well-known Glivenko-Cantelli Theorem (p. 103, Billings, 1968), whereas b) is a rate of convergence result on the sampling distribution. Lemma 2 b) indicates that $\nu_N(\cdot)$ converges to $\nu(\cdot)$. By virtue of the Skorohod representation (p. 230, Kushner & Yin, 2003, with a slight abuse of notation), we may assume that $\nu_N(\cdot) \to \nu(\cdot)$ w.p.1 and the convergence takes place uniformly on any compact set.

From (19), the i-th component $\widetilde{\xi}_i(N)$ of $\widetilde{\xi}(N)$ is the N-sample empirical distribution of $\widetilde{d}(k)$ at $C - H(\rho_i\delta, \eta)$, denote $\mu_i(N) = \sqrt{N}(\widetilde{\xi}_i(N) - p_i)$. Since $\widetilde{d}(i), i = 1, 2, \ldots$ are i.i.d., for $0 \leq i \leq m$,

$$P\{\widetilde{s}(k(m+1)+i) = 1\} = P\{\widetilde{s}(i) = 1\} = p_i,$$
$$P\{\widetilde{s}(k(m+1)+i) = 0\} = P\{\widetilde{s}(i) = 0\} = 1 - p_i.$$

Hence, the expectation $E\widetilde{s}(i) = p_i$, $E(\widetilde{s}(i) - p_i)^2 = p_i(1 - p_i)$, and for $0 \leq i < j \leq m$, $E(\widetilde{s}(i) - p_i)(\widetilde{s}(j) - p_j) = 0$.

Since $\widetilde{d}(i), i = 1, 2, \ldots$ are i.i.d, for $i \neq j$, $\mu_i(N)$ and $\mu_j(N)$ are independent, hence $E\mu_i(N)\mu_j(N) = 0$. Also, $E(\mu_i(N))^2 = NE(\widetilde{\xi}_i(N) - p_i)^2 = E(\widetilde{s}(i) - p_i)^2 = p_i(1 - p_i)$. Let $\mu(N) = [\mu_1(N), \ldots, \mu_{m+1}(N)]^T$. Then, the above expressions imply that

$$\begin{aligned} E\mu(N)\mu(N)^T &\to V \text{ as } N \to \infty \\ &= \text{diag}(p_0(1-p_0), \ldots, p_m(1-p_m)). \end{aligned} \tag{33}$$

In view of Lemma 2,

$$\mu(N) \sim N(0, V) \text{ as } N \to \infty. \tag{34}$$

That is, $\mu(N)$ converges in distribution to a normal random vector with mean 0 and covariance V.

6.2 Asymptotic analysis of identification errors

The following analysis of identification errors is generic, and hence is described without reference to specific algorithms. For simplicity, for $x \in \mathbb{R}$, denote $B(x) = C - F^{-1}(x)$. Then, by (20) we have

$$p = [p_0, \ldots, p_m]^T = \mathbf{F}(\mathbf{C}_{m+1} - \zeta) = \mathbf{B}^{-1}(\zeta), \tag{35}$$

where ζ is denoted as $\zeta = [\zeta_0, \ldots, \zeta_m]^T$. Let $\mathbf{g}(\zeta) = [g_0(\zeta), \ldots, g_m(\zeta)]^T = \mathbf{H}^{-1}(\zeta)$. Then, $\zeta(N)$, $\tau(N)$ in Theorem 1 and $\tau = [\tau_0, \ldots, \tau_m]^T$ can be written as

$$\zeta(N) = \mathbf{B}(\widetilde{\xi}(N)), \ \tau(N) = \mathbf{g}(\zeta(N)), \ \tau = \mathbf{g}(\mathbf{B}(p)). \tag{36}$$

The estimation error for τ is $e(N) = [e_0(N), \ldots, e_m(N)]^T = \tau(N) - \tau$.

For $\tau = \mathbf{g}(\zeta)$, the Jacobian matrix is

$$J(\mathbf{g}(\zeta)) = \frac{\partial \mathbf{g}(\zeta)}{\partial \zeta} = \begin{bmatrix} \frac{\partial g_0(\zeta)}{\partial \zeta_0} & \cdots & \frac{\partial g_0(\zeta)}{\partial \zeta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\zeta)}{\partial \zeta_0} & \cdots & \frac{\partial g_m(\zeta)}{\partial \zeta_m} \end{bmatrix},$$

and for $\zeta = \mathbf{H}(\tau)$,

$$J(\mathbf{H}(\tau)) = \frac{\partial \mathbf{H}(\tau)}{\partial \tau} = \begin{bmatrix} \frac{\partial h_0(\tau)}{\partial \tau_0} & \cdots & \frac{\partial h_0(\tau)}{\partial \tau_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m(\tau)}{\partial \tau_0} & \cdots & \frac{\partial h_m(\tau)}{\partial \tau_m} \end{bmatrix}.$$

Since $\zeta = \mathbf{H}(\tau)$, we have

$$J(\mathbf{g}(\zeta))J(\mathbf{H}(\tau)) = \frac{\partial \mathbf{g}(\zeta)}{\partial \zeta}\frac{\partial \mathbf{H}(\tau)}{\partial \tau} = \frac{\partial \mathbf{g}(\mathbf{H}(\tau))}{\partial \tau} = \frac{\partial \tau}{\partial \tau} = I_{m+1}.$$

As a result, $J(\mathbf{g}(\zeta)) = J(\mathbf{H}(\tau))^{-1}$. From (35), we have $\zeta_i = B(p_i)$, $i = 0, 1, \ldots, m$. It follows that the Jacobian matrix for $\zeta = \mathbf{B}(p)$ is

$$J(\mathbf{B}(p)) = \text{diag}(\frac{\partial B(p_0)}{\partial p_0}, \ldots, \frac{\partial B(p_m)}{\partial p_m}),$$

and for $p = B^{-1}(\zeta)$,

$$J(\mathbf{B}^{-1}(p)) = \text{diag}(\frac{\partial B^{-1}(\zeta_0)}{\partial \zeta_0}, \ldots, \frac{\partial B^{-1}(\zeta_m)}{\partial \zeta_m}).$$

Theorem 5. [5] *Under Assumptions A1, A2, and A3, $N\sigma^2(e(N)) = NEe(N)e(N)^T \to \Lambda$, as $N \to \infty$, where $\Lambda = WVW^T$ with $W = J(\mathbf{g}(\zeta))J(\mathbf{B}(p))$ and V being given by (33).*

Proof. See Appendix A.

6.3 Cramér-Rao lower bound and asymptotic efficiency

Consider N blocks of $m+1$ observations for the core identification problem. We first derive the Cramér-Rao lower bound based on these $N(m+1)$ observation data. The Cramér-Rao lower bound is denoted as $\sigma^2_{CR}(N)$. To proceed, we first derive a lemma and then Theorem 6 follows.

Lemma 3. *The Cramér-Rao lower bound for estimating the parameter τ, based on observations of $\{\widetilde{S}(k)\}$, is $\sigma^2_{CR}(N) = \Lambda/N$.*

Proof. See Appendix B.

Theorem 6. *Under Assumptions A1, A2, and A3, $N[\sigma^2(e(N)) - \sigma^2_{CR}(N)] \to 0$ as $N \to \infty$.*

Proof. This follows directly from Theorem 5 and Lemma 3. □

[5] The convergence in Theorem 5 is valid for disturbances whose probability density functions are in an exponential class: For some $\alpha > 0$ and $\beta > 0$, $f(x) \geq \beta e^{-\alpha x^2}$. This implies that $f(x)$ does not go to zero faster than the exponential function of x^2 as $x \to \infty$. Since all commonly encountered density functions are in this class, for clarity and simplicity of presentation, we will not state this condition explicitly.

7 Recursive algorithms and convergence

This section develops a recursive algorithm for estimating (θ^*, η^*). The essence is to treat the parameters (θ, η) jointly. Define $\Theta = [\theta^T, \eta^T]^T \in \mathbb{R}^{(n+m)\times 1}$. For an $(n+m) \times 2n(m+1)$ matrix M, and for each $\widetilde{\xi}$, define

$$G(\Theta, \widetilde{\xi}) = M[\widetilde{\xi} - \mathbf{F}(C_{2n(m+1)} - \mathbf{H}(\widetilde{\Phi}\theta, \eta)]. \tag{37}$$

It is easily seen that the purpose of the matrix M is to make the function under consideration "compatible" with the dimension of the vector Θ. We use the following recursive algorithm for parameter estimation

$$\begin{aligned} \xi(k+1) &= \xi(k) - \frac{1}{k+1}\xi(k) + \frac{1}{k+1}S(k+1), \\ \Theta(k+1) &= \Theta(k) + \beta_k G(\Theta(k), \xi(k)), \quad k = 0, 1, \ldots, \end{aligned} \tag{38}$$

where $S(k+1)$ is defined in (15). In the above algorithm, β_k is a sequence of step sizes satisfying $\beta_k \geq 0$, $\sum_{k=1}^{\infty} \beta_k = \infty$, $\beta_k \to 0$, and

$$\frac{\beta_k - \beta_{k+1}}{\beta_k} = O(\beta_k) \text{ as } k \to \infty. \tag{39}$$

Take for instance, $\beta_k = 1/k^\alpha$ with $0 < \alpha \leq 1$. Then, the condition (39) is satisfied. Commonly used step sizes include $\beta_k = O(1/k^\alpha)$ with $(1/2) < \alpha \leq 1$.

Associated with (38), consider an ordinary differential equation (ODE)

$$\dot{\Theta} = \overline{G}(\Theta), \tag{40}$$

where $\overline{G}(\Theta) = M(\xi - \mathbf{F}(C_{2(m+1)n} - \mathbf{H}(\widetilde{\Phi}\theta, \eta))$. Θ^* is the unique stationary point of (40). To proceed, we assume the following assumption holds.

Assumption A4. The ODE (40) has a unique solution for each initial condition; $\Theta^* = (\theta^*, \eta^*)$ is an asymptotically stable point of (40); $\mathbf{H}(\cdot)$ is continuous in its arguments together with its inverse.

Remark 5. A sufficient condition to ensure the asymptotic stability of (40) can be obtained by linearizing $M[\xi - \mathbf{F}(\mathbf{C}_{2n(m+1)} - \mathbf{H}(\widetilde{\Phi}\theta, \eta))]$ about its stationary point Θ^*. Under this linearization, if the Jacobian matrix $-M(\partial \mathbf{F}(\mathbf{C}_{2n(m+1)} - \mathbf{H}(\widetilde{\Phi}\theta^*, \eta^*))/\partial\Theta)$ is a stable matrix (that is, all of its eigenvalues are on the left-hand side of the complex plane), the required asymptotic stability follows.

Theorem 7. *Under Assumptions A1–A4, $\xi(k) \to \xi$ and $\Theta(k) \to \Theta^*$ w.p.1 as $k \to \infty$.*

Proof. Note that we have already proved that $\xi(k) \to \xi$ w.p.1. Thus, to obtain the desired result, we need only to establish the convergence of $\{\Theta(k)\}$. To this end, we use the ODE methods to complete the proof.

We will use the basic convergence theorem (Theorem 6.1.1, p. 166 in Kushner & Yin, 2003). Thus, all needed is to verify the conditions in the aforementioned theorem hold. Note that we do not have a projection now, but in our recursion $\mathbf{F}$ is used and is uniformly bounded. In view of Assumptions A1–A4, as explained in (Section 6.2, p. 170 of Kushner & Yin, 2003), to verify the conditions in the theorem, we need only show that a "rate of change" condition (see p. 137 in Kushner & Yin, 2003, for a definition) is satisfied. Thus, the remaining proof is to verify this condition.

Define $t_0 = 0$, $t_k = \sum_{i=0}^{k-1} \beta_i$, and let $m(t)$ be the unique value k such that $t_k \le t < t_{k+1}$ when $t \ge 0$, and set $m(t) = 0$ when $t < 0$. Define the piecewise constant interpolation as $\Theta^0(t) = \Theta(k)$ for $t_k \le t < t_{k+1}$, and define the shifted sequence by $\Theta^k(t) = \Theta^0(t + t_k)$, $t \in (-\infty, \infty)$. Using the ODE methods, we can show the sequence of functions $\Theta^k(\cdot)$ converges to the solution of desired limit ODE. For $m = 1, 2, \ldots$, and a fixed Θ, denote

$$\Xi(m) = \sum_{i=0}^{m-1} [G(\Theta, \xi(i)) - \overline{G}(\Theta)],$$

and $\Xi_0 = 0$. In view of (37), $G(\cdot, \cdot)$ is a continuous function in both variables Θ and $\widetilde{\xi}$.

We note that by a partial summation, for any $m, j \ge 0$,

$$\sum_{i=j}^{m} \beta_i [G(\Theta, \xi(i)) - \overline{G}(\Theta)] = \beta_m \Xi(m+1) - \beta_m \Xi(j) + \sum_{i=j}^{m-1} [\Xi(i+1) - \Xi(j)](\beta_i - \beta_{i+1}).$$

Taking $m = m(t) - 1$ and $j = 0$, and recalling $\Xi_0 = 0$, we obtain

$$\sum_{i=0}^{m(t)-1} \beta_i [G(\Theta, \xi(i)) - \overline{G}(\Theta)] = \beta_{m(t)} \Xi(m(t)) + \sum_{i=0}^{m(t)-2} \Xi(i+1) \frac{\beta_i - \beta_{i+1}}{\beta_i} \beta_i.$$

It is readily seen that as $k \to \infty$, $\beta_k \Xi(k) \to 0$ w.p.1. Thus, the asymptotic rate of change of $\sum_{i=0}^{m(t)-1} \beta_i [G(\Theta, \xi_i) - \overline{G}(\Theta)]$ is zero w.p.1. Then by virtue of Theorem 6.1.1 in Kushner and Yin (2003), the limit ODE is precisely (40). The asymptotic stability of the ODE then leads to the desired result. □

Remark 6. Note that in (38), we could include additional random noises (representing the measurement noise and other external noise). The treatment remains essentially the same. We choose the current setup for notational simplicity.

8 Illustrative examples

In this section, we illustrate convergence of estimates from the algorithms developed in this paper. The noise is gaussian distributed zero mean and known variance, although the algorithms are valid for all distribution functions that satisfy Assumption A1. The identification algorithm of Section 5 is shown in Example 1, and the asymptotic efficiency is also illustrated for the core identification problem. Example 2 illustrates the recursive algorithm. The estimates of parameters are shown to be convergent in both cases.

Example 1. Consider

$$\begin{cases} y(k) = H(x(k), \eta) + d(k) = b_0 + e^{x(k)} + d(k), \\ x(k) = a_1 u(k-1) + a_2 u(k-2), \end{cases} \tag{41}$$

where the noise $\{d(k)\}$ is a sequence of i.i.d. normal random variables with $Ed_1 = 0$, $\sigma_d^2 = 1$. For normal distribution, the support is $(-\infty, \infty)$. The output is measured by a binary-valued sensor with threshold $C = 3$. The linear subsystem has order $n = 2$. The nonlinear function is parameterized as $b_0 + e^x$. The prior information on b_0, and a_i, $i = 1, 2$ is that $b_0, a_i \in [0.5, 5]$. Suppose the true values of unknown parameters are $\theta = [a_1, a_2] = [0.7, 0.63]$ and $\eta = b_0 = 1.1$.

For $n = 2$ and $m = 1$, the input should be $2n(m+1) = 8-$periodic with single period $u = [\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v]$. By Section 4.2, $H(x, \eta)$ is jointly identifiable with respect to $\Upsilon = \{(\rho_0, \rho_1) : \rho_0 > 0, \rho_1 < 0\}$. Let $v = [1, 1.2]$, $\rho_0 = 1$ and $\rho_1 = -1$. Define the block variables $X(j), Y(j), \widetilde{\Phi}(j), D(j)$ and $S(j)$, in the case of an 8-periodic input, $\widetilde{\Phi}(j) = \widetilde{\Phi} = [\Phi_1^T, \ldots, \Phi_4^T]^T$, where $\Phi_1 = \rho_0 \Phi = \Phi = \begin{bmatrix} v_2 & v_1 \\ v_1 & v_2 \end{bmatrix}$ and $\Phi_3 = \rho_1 \Phi$. Using (22), we can construct the algorithms in Section 5.3.

The estimates of θ and η are shown in Fig. 3, where the errors are measured by the Euclidean norm. The algorithms are simulated for five times. It is shown that both parameter estimates of the linear and nonlinear subsystems converge to their true values. In this simulation $\eta(N)$ demonstrates a higher convergence speed than $\theta(N)$. A possible explanation is that $\eta(N)$ is updated

first, and then used to obtain $\theta(N)$. As a result, convergence on $\theta(N)$ can occur only after the error $\eta(N) - \eta$ is reduced.

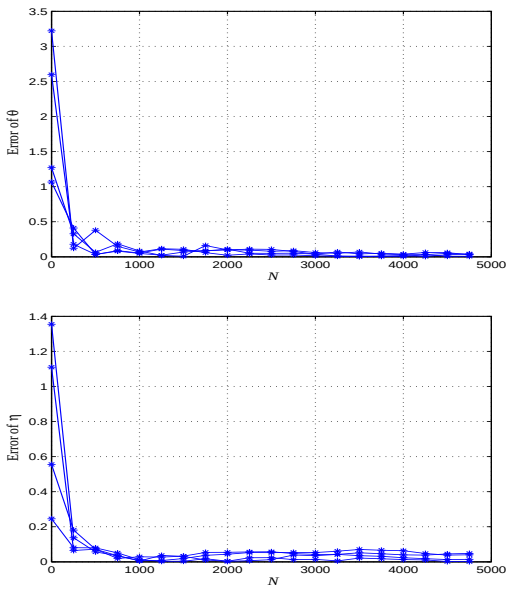

Fig. 3. Joint identification errors of θ and η

To understand reliability of the estimation schemes, the estimation algorithms are performed 500 times of total data length 2000 each. Estimation errors for each run are recorded at $N = 500$, $N = 1000$, and $N = 2000$. The error distributions are calculated by histograms in Figure 4, which illustrate improved estimation accuracy with respect to data length N and are consistent with the theoretical analysis.

Consider the core identification problem of (41)

$$\widetilde{Y}(l) = H(\rho\delta, \eta) + \widetilde{D}(l) = b_0 \mathbb{1}_2 + e^{\rho\delta} + \widetilde{D}(l),$$

where $\delta = a_0 v_2 + a_1 v_1 \neq 0$ and $\rho = [\rho_0, \rho_1]^T$. The convergence of $N[\sigma^2(e(N)) - \sigma^2_{CR}(N)]$ in Theorem 6 is shown in Fig. 5, where the error is measured by the Frobenius norm.

Example 2. We use the same system and inputs as in Example 1. The recursive algorithms in Section 7 are now used.

Let $\rho_1 = 0.5$ and $\Theta = [\theta^T, \eta^T]^T$. For system (41), the ODE (40) becomes

$$\dot{\Theta} = M[\xi - \mathbf{F}(\mathbf{C}_8 - b\mathbb{1}_8 - \exp(\widetilde{\Phi}\theta))].$$

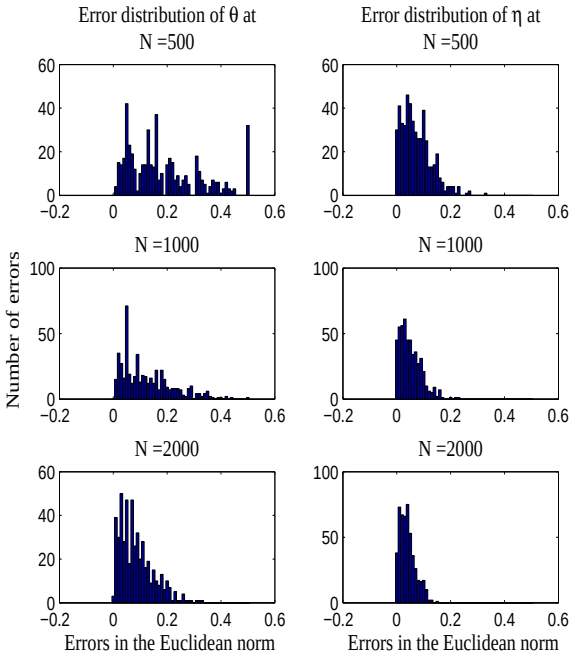

Fig. 4. Estimation error distributions

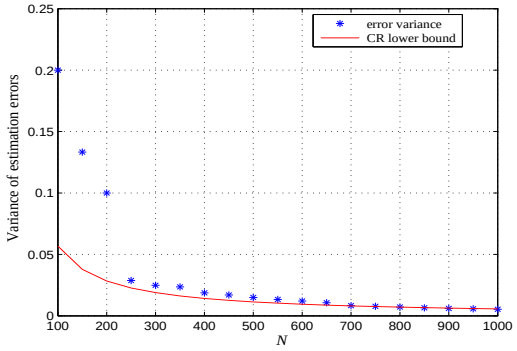

Fig. 5. Asymptotic efficiency

Choose $\beta_k = \frac{1}{k}$ and

$$M = -\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Then the Jacobian matrix can be calculated to be

$$\begin{aligned} J(\Theta) &= -M[\partial \mathbf{F}(\mathbf{C}_8 - b\mathbb{1}_8 - \exp(\widetilde{\Phi}\theta))/\partial\Theta] \\ &= \begin{bmatrix} -0.660 & -0.247 & -0.429 \\ -0.242 & -0.645 & -0.434 \\ -0.210 & -0.079 & -0.397 \end{bmatrix}. \end{aligned}$$

The eigenvalues of $J(\Theta)$ are $[-1.08, -0.402, -0.220]$,

which are all less than 0. As a result, the Jacobin matrix $J(\eta)$ is stable.

Let $\Theta(k) = [\theta(k)^T, \eta(k)^T]^T$ be the estimates of $\Theta = [\theta^T, \eta^T]^T$. Then the recursive algorithms can be constructed as follows: First, set $\beta_k = 1/k$, $\Theta(1) = [1.5, 1.5, 1.5]^T$, and $\xi_1 = \mathbf{0}_8$. The estimates are then updated according to (38). Convergence of Θ is shown in Fig. 6, where the errors are measured by the Euclidean norm and the algorithms are simulated for five times.

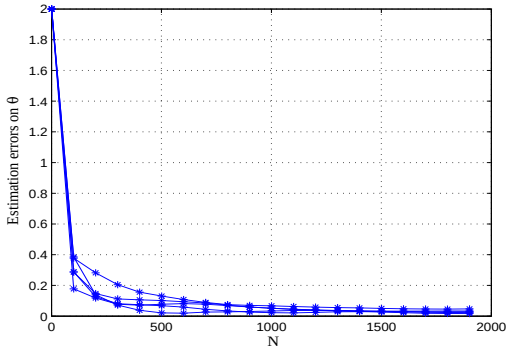

Fig. 6. Estimation errors of Θ using recursive algorithms

Remark 7. It is easy to see that when $\theta(N) \to \theta$ and $\eta(N) \to \eta$, the prediction of the output $H(\Phi\theta, \eta)$ converges to the true output $H(\Phi\theta, \eta)$. This implies that one can use the parameter estimation errors as a good indicator for output prediction errors. For this reason, the output prediction errors are not plotted here.

9 Concluding remarks

In this paper, identification of Wiener systems with binary-valued output observations is studied. Unlike traditional approximate gradient methods or covariance analysis, we employ the methods of empirical measures. Under assumptions of known disturbance distribution function, invertible nonlinearity and joint identifiability, identification algorithms, convergence properties, and identification efficiency are derived.

We have assumed that the structure and order of the linear dynamics and nonlinear function are known. The issues of unmodelled dynamics (for the linear subsystem when the system order is higher than the model order) and model mismatch (for the nonlinear part when the nonlinear function does not belong to the model class) are not included in this paper, mainly due to page limitations. Irreducible identification errors due to unmodelled dynamics were characterized in Wang, et al (2003). The impact of model mismatch on identification errors were presented in Yin, et al (2006).

There are many potential extensions of the results in this paper. For example, when the sensor threshold value and/or the noise distribution function are unknown, combined identification of systems, distribution functions and sensor thresholds is of practical importance. Some related results can be found in Wang, et al (2006a). For other typical nonlinear structures, such as Hammerstein systems and kernel systems, similar identification problems can be pursued.

A Appendices

Appendix A: Proof of Theorem 5. Consider

$$e_i(N) = \tau_i(N) - \tau_i = g_i(\zeta(N)) - g_i(\zeta), \quad i = 0, \ldots, m,$$

where $\zeta(N) = [\zeta_0(N), \ldots, \zeta_m(N)]^T$, $\tau(N) = [\tau_0(N), \ldots, \tau_m(N)]^T$, τ and ζ are given by (36) and (7), respectively. Denote

$$\Omega(N) = [\min\{\zeta_0(N), \zeta_0\}, \max\{\zeta_0(N), \zeta_0\}] \times \cdots \times [\min\{\zeta_m(N), \zeta_m\}, \max\{\zeta_m(N), \zeta_m\}]$$

as the Cartesian product (p. 3, Royden, 1988) of sets $[\min\{\zeta_i(N), \zeta_i\}, \max\{\zeta_i(N), \zeta_i\}]$, for $i = 0, \ldots, m$.

For $j = 0, \ldots, m-1$, denote

$$\widetilde{\zeta}_j(N) = [\zeta_0, \ldots, \zeta_j, \zeta_{j+1}(N), \ldots, \zeta_m(N)]^T,$$

$\widetilde{\zeta}_{-1}(N) = [\zeta_0(N), \ldots, \zeta_m(N)]^T$ and $\widetilde{\zeta}_m(N) = \zeta$. Then

$$\begin{aligned} e_i(N) &= g_i(\zeta(N)) - g_i(\zeta) \\ &= \sum_{j=-1}^{m-1} [g_i(\widetilde{\zeta}_j(N)) - g_i(\widetilde{\zeta}_{j+1}(N))]. \end{aligned}$$

Since $H(\cdot)$ is continuous, by the well-known mean value theorem, there exists $\lambda_{ij}(N) \in \Omega(N)$ for $j = 0, \ldots, m$ such that

$$g_i(\widetilde{\zeta}_j(N)) - g_i(\widetilde{\zeta}_{j+1}(N)) = \frac{\partial g_i(\lambda_{ij}(N))}{\partial \zeta_j}(\zeta_j(N) - \zeta_j),$$

which implies

$$\begin{aligned} e_i(N) &= \sum_{j=0}^{m} \frac{\partial g_i(\lambda_{ij}(N))}{\partial \zeta_j}(\zeta_j(N) - \zeta_j) \\ &= [\frac{\partial g_i(\lambda_{i0}(N))}{\partial \zeta_0}, \ldots, \frac{\partial g_i(\lambda_{im}(N))}{\partial \zeta_m}](\zeta(N) - \zeta). \end{aligned}$$

Thus

$$e(N) = L(N)(\zeta(N) - \zeta), \tag{A.1}$$

where

$$L(N) = \begin{bmatrix} \frac{\partial g_0(\lambda_{00}(N))}{\partial \zeta_0} & \cdots & \frac{\partial g_0(\lambda_{0m}(N))}{\partial \zeta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\lambda_{m0}(N))}{\partial \zeta_0} & \cdots & \frac{\partial g_m(\lambda_{mm}(N))}{\partial \zeta_m} \end{bmatrix}.$$

Since $\zeta_i(N) = B(\widetilde{\xi}_i(N))$, $i = 0, 1, \ldots, m$, by the mean value theorem, there exists $\kappa_i(N)$ on the line segment $\widetilde{\xi}_i(N)$ and p_i such that

$$\zeta(N) - \zeta = \mathrm{diag}(\frac{\partial B(\kappa_0(N))}{\partial p_0}, \cdots, \frac{\partial B(\kappa_m(N))}{\partial p_m}) \times(\widetilde{\xi}(N) - p). \quad \text{(A.2)}$$

Moreover, as $N \to \infty$, w.p.1,

$$L(N)\mathrm{diag}(\frac{\partial B(\kappa_0(N))}{\partial p_0}, \cdots, \frac{\partial B(\kappa_m(N))}{\partial p_m}) \to W. \quad \text{(A.3)}$$

Using (A.1), and by virtue of (34), (A.2), and (A.3), as $N \to \infty$, $NEe(N)e(N)^T \to WVW^T = \Lambda$. □

Appendix B: Proof of Lemma 3. Let $x(k)$ take values in {0,1}. The likelihood function, which is the joint distribution of $\widetilde{s}(1), \ldots, \widetilde{s}(N(m+1))$, depending on $\tau = [\tau_0, \ldots, \tau_m]^T = [\delta, \eta^T]^T$, is given by

$$\begin{aligned} & l(N) \\ &= P\{\widetilde{s}(1) = x(1), \ldots, \widetilde{s}(N(m+1)) = x(N(m+1)); \tau\} \\ &= \prod_{k=0}^{m} P\{\widetilde{s}(kN+1) = x(kN+1), \\ &\qquad \ldots, \widetilde{s}(kN+m+1) = x((k+1)N); \tau\}. \end{aligned}$$

Replace $x(k)$'s by their corresponding random elements $\widetilde{s}(k)$'s, and denote the resulting quantity by l in short. Then, we have

$$\begin{aligned} & \log l(N) \\ &= \log\Big[\prod_{k=0}^{m} p_k(\tau)^{N\widetilde{\xi}_k(N)}(1 - p_k(\tau))^{N(1-\widetilde{\xi}_k(N))}\Big] \\ &= N\sum_{k=0}^{m}[\widetilde{\xi}_k(N)\log p_k(\tau) + (1 - \widetilde{\xi}_k(N))\log(1 - p_k(\tau))], \end{aligned}$$

$$\frac{\partial \log l(N)}{\partial \tau_i} = N\sum_{k=0}^{m}(\frac{\widetilde{\xi}_k(N)}{p_k} - \frac{1 - \widetilde{\xi}_k(N)}{1 - p_k})\frac{\partial p_k}{\partial \zeta_k}\frac{\partial \zeta_k}{\partial \tau_i},$$

$$\frac{\partial \log l(N)}{\partial \tau} = [\frac{\partial \log l(N)}{\partial \tau_0}, \ldots, \frac{\partial \log l(N)}{\partial \tau_m}]^T.$$

Furthermore, for $i, j = 0, \ldots, m$,

$$\begin{aligned} \frac{\partial^2 \log l(N)}{\partial \tau_i \partial \tau_j} &= N\sum_{k=0}^{m}[(-\frac{\widetilde{\xi}_k(N)}{p_k^2} - \frac{1 - \widetilde{\xi}_k(N)}{(1 - p_k)^2})\frac{\partial p_k}{\partial \tau_i}\frac{\partial p_k}{\partial \tau_j} \\ &\quad +(\frac{\widetilde{\xi}_k(N)}{p_k} - \frac{1 - \widetilde{\xi}_k(N)}{1 - p_k})\frac{\partial^2 p_k}{\partial \tau_i \partial \tau_j}]. \end{aligned}$$

As a result,

$$\begin{aligned} & E\frac{\partial^2 \log l(N)}{\partial \tau_i \partial \tau_j} \\ &= NE\sum_{k=0}^{m}[(-\frac{\widetilde{\xi}_k(N)}{p_k^2} - \frac{1 - \widetilde{\xi}_k(N)}{(1 - p_k)^2})\frac{\partial p_k}{\partial \tau_i}\frac{\partial p_k}{\partial \tau_j} \\ &\quad +(\frac{\widetilde{\xi}_k(N)}{p_k} - \frac{1 - \widetilde{\xi}_k(N)}{1 - p_k})\frac{\partial^2 p_k}{\partial \tau_i \partial \tau_j}] \\ &= -N\sum_{k=0}^{m}\frac{1}{p_k(1 - p_k)}\frac{\partial p_k}{\partial \tau_i}\frac{\partial p_k}{\partial \tau_j} \\ &= -N\sum_{k=0}^{m}\frac{1}{p_k(1 - p_k)}(\frac{\partial p_k}{\partial \zeta_k})^2\frac{\partial \zeta_k}{\partial \tau_i}\frac{\partial \zeta_k}{\partial \tau_j}, \end{aligned}$$

and

$$E\frac{\partial^2 \log l(N)}{\partial \tau \partial \tau} = -NW^{-1}V^{-1}(W^T)^{-1}.$$

The Cramér-Rao lower bound is then given by

$$\sigma_{CR}^2(N) = -(E\frac{\partial^2 \log l(N)}{\partial \tau \partial \tau})^{-1} = \frac{WVW^T}{N} = \frac{\Lambda}{N}.$$

□

Acknowledgements

The research of Yanlong Zhao and Ji-Feng Zhang was supported by the National Natural Science Foundation of China under grants 60221301, 60674038. The research of Le Yi Wang was supported in part by the National Science Foundation under ECS-0329597 and DMS-0624849, in part by the Michigan Economic Development Council, and in part by Wayne State University Research Enhancement Program. The research of George Yin was supported in part by the National Science Foundation under DMS-0603287 and DMS-0624849, and in part by Wayne State University Research Enhancement Program.

References

Bai, E. W. (1998). An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear system, *Automatica*, 34, 333-338.

Bai, E. W. (2003). Frequency domain identification of Wiener models, *Automatica*, 39, 1521-1530.

Billings, S. (1968). *Convergence of probability measures*, J. Wiley, New York.

Billings, S. (1980). Identification of nonlinear systems-A survey, *Proc. of IEE*, Part D, 127, 272-285.

Celka, P., Bershad, N. J., & Vesin, J. M. (2001). Stochastic gradient identification of polynomial Wiener systems: Analysis and application, *IEEE Trans. Signal Process.*, 49, 301-313.

Chen, H. F. (2006). Recursive identification for Wiener model with discontinuous piece-wise linear function, *IEEE Trans. Autom. Control*, 51, 390-400.

Chen, H. F., & Guo, L. (1991). *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston.

Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*, Cambridge University Press, Cambridge.

Hu, X. L., & Chen, H. F. (2005). Strong consistence of recursive identification for Wiener systems, *Automatica*, 41, 1095-1916.

Hunter, I. W., & Korenberg, M. J. (1986). The identification of nonlinear biological systems: Wiener and Hammerstein cascade models, *Bio Cybernetics*, 55, 135-144.

Feller, W. (1968 & 1971). *An introduction to probability theory and its applications*, vol. I, 3rd ed. Wiley, New York; & vol. II, 2nd ed. Wiley, New York.

Korenberg, M. J., & Hunter, I. W. (1998). Two methods for identifying Wiener cascades having noninvertible static nonlinearities, *Annals of Biomedical Engineering*, 27, 793-804.

Krishnamarty, V. (1995). Estimation of quantized linear errors-in-variables models, *Automatica*, 31, 1459-1464.

Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications*, 2nd Ed., Springer-Verlag, New York.

Lacy, S. L., & Bernstein, D. S. (2002). Identification of FIR Wiener systems with unknown, noninvertible polynomial nonlinearities, *Proc. Amer. Contr. Conf.*, 893-899.

Lancaster, P., & Tismenetsky, M. (1985). *The theory of matrices*, 2nd Ed., Academic Press.

Ninness, B., & Gibson, S., (2002). Quantifying the accuracy of Hammerstein model estimation, *Automatica*, 38, 2037-2051.

Norquay, S. J., Palazoglu, A., & Romagnoli, J. A. (1999). Application of Wiener model predictive control (WMPC) to pH neutralization experiment, *IEEE Trans. Control Sys. Technology*, 7, 437-445.

Roll, J., Nazin, A. & Ljung L. (2005). Nonlinear system identification via direct weight optimization, *Automatica*, 41, 475-490.

Royden, H. L., (1988). *Real Analysis*, 3rd Ed. New York: Macmillan.

Schoukens, J., Nemeth, J. G., Crama, P., Rolain, Y., & Pintelon, R. (2003) Fast approximate identification of nonlinear systems, *Automatica*, 39, 1267-1274.

Sjoberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., Hjalmarsson H., & Juditskys, A. (1995). Nonlinear black-box modeling in system identification: a unified overview, *Automatica*, 31, 1691-1724.

Verhaegen, M., & Westwick, D. (1996). Identifying MIMO Wiener systems using subspace model identification methods, *Signal Processing*, 52, 235-258.

Wigren, T. (1994). Convergence analysis of recursive identification algorithms based on the nonlinear Wiener model, *IEEE Trans. Automatic Control*, 39, 2191-2206.

Wigren, T. (1995). Approximate gradients, convergence and positive realness in recursive identification of a class of nonlinear systems, *Int. J. Adaptive Contr. Signal Processing*, 9, 325-354.

Wigren, T. (1998). Adaptive filtering using quantized output measurements, *IEEE Trans. Signal Processing*, 46, 3423-3426.

Wang, L. Y., Kim, Y., & Sun, J (2002a). "Prediction of oxygen storage capacity and stored NOx using HEGO sensor model for improved LNT control strategies", *2002 ASME International Mechanical Engineering Congress and Exposition*, New Orleans, Nov. 17-22.

Wang, L. Y., & Wang, H. (2002b). Control-oriented modeling of BIS-based patient response to anesthesia infusion, *2002 Internat. Conf. Math. Eng. Techniques in Medicine and Bio. Sci.*, Las Vegas, June 24-27.

Wang, L. Y., Yin, G., & Wang, H. (2004). Identification of Wiener models with anesthesia applications, *Int. J. Pure & Appl. Sci*, 1, 35-61.

Wang, L. Y., Yin, G., & Zhang, J. F. (2006a). Joint identification of plant rational models and noise distribution functions using binary-valued observations, *Automatica*, 42, 535-547.

Wang, L. Y., Yin, G. (2006b). Asymptotically efficient parameter estimation using quantized output observations, to appear in *Automatica*.

Wang, L. Y., Yin, G., & Zhang, J. F. (2006c). Space and time complexities, sensor threshold selection, and input design of quantized identification, submitted.

Wang, L. Y., Zhang, J. F., & Yin, G. (2003). System identification using binary sensors, *IEEE Trans. Automat. Control*, 48, 1892-1907.

G. Yin, S. Kan, and L.Y. Wang. (2006) Identification error bounds and asymptotic distributions for systems with structural uncertainties, *Journal of Systems Science and Complexity*, Vol. 19, 22-35.