[3] J. B. Lasserre, "A trace inequality for the matrix product," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 1500–1501, 1995.

[4] M. Mrabti and M. Benseddik, "Bounds for the eigenvalues of the solution of the unified algebraic Riccati matrix equation," *Syst. Contr. Lett.*, vol. 24, pp. 345–349, 1995.

[5] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. Orlando, FL: Academic, 1979.

[6] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

Convergence Analysis of the Extended Kalman Filter Used as an Observer for Nonlinear Deterministic Discrete-Time Systems

M. Boutayeb, H. Rafaralahy, and M. Darouach

***Abstract*—In this paper, convergence analysis of the extended Kalman filter (EKF), when used as an observer for nonlinear deterministic discrete-time systems, is presented. Based on a new formulation of the first-order linearization technique, sufficient conditions to ensure local asymptotic convergence are established. Furthermore, it is shown that the design of the arbitrary matrix, namely R_k in the paper, plays an important role in enlarging the domain of attraction and then improving the convergence of the modified EKF significantly. The efficiency of this approach, compared to the classical version of the EKF, is shown through a nonlinear identification problem as well as a state and parameter estimation of nonlinear discrete-time systems.**

***Index Terms*—Convergence analysis, deterministic nonlinear discrete-time systems, extended Kalman filter, Lyapunov approach.**

I. Introduction

Since the 1960's, many research activities have been developed to deal with the problem of state estimation for nonlinear dynamical systems. The main motivation for doing so is that most physical processes are described by nonlinear mathematical models. Thus, several nonlinear state estimation methods have been performed to increase the accuracy and performances of the control system design.

Generally, we distinguish two approaches for nonlinear observers' design. The first one is based on some nonlinear state transformation using the Lie algebra. This approach brings the original system into a canonical form from which the design of state observers is performed using linear techniques in the new coordinates. Necessary and sufficient conditions for a standard nonlinear system to be state equivalent to the nonlinear canonical form, for both continuous and discrete-time cases, have been established in [2], [8], [9], and [14]. We notice that only a few classes of forced nonlinear systems are considered by this nonlinear transformation.

The second approach is without transformation and is based on the linearized model. In spite of the local convergence of this method, it is widely used in practice and generally gives good results under less restrictive conditions than the first approach [6], [7], [12], [13], [16]. One of the most popular estimation techniques largely investigated for state estimation of nonlinear systems is the extended Kalman filter (EKF).

The EKF consists of using the classical Kalman filter equations to the first-order approximation of the nonlinear model about the last estimate. The literature is vast about this subject, and we refer the reader to [7], [11], [15], [16], and [19] and the references therein. However, few works have been performed to analyze the stability and convergence of the filter. The main difficulty arises from the fact that the EKF equations are only approximate ones. So, the corresponding propagation equations are available only if the estimate belongs to a neighborhood of the actual state.

An interesting study of asymptotic behavior of the EKF, used for the joint parameter and state estimation of linear time-invariant systems, was developed by Ljung in [11]. For this particular case, some modifications, such as coupling between parameters vector and the Kalman gain, were introduced to enhance convergence of the EKF. More recently, in a synthesis work on nonlinear discrete-time systems, sufficient conditions of the EKF for noisy systems used as a local asymptotic observer for the deterministic case have been established by Song *et al.* [15]. They have also shown that conditions needed to ensure the uniform boundedness of certain Riccati equations are related to the observability properties of the considered nonlinear system.

Recently, Cicarrella *et al.* [18] have developed a robust observer, which uses n output values and the inverse of the observability matrix at each step, and the local convergence analysis for a multi-input/single-output (MISO) nonlinear discrete-time system was studied.

Motivated by the identification problem of time-invariant nonlinear systems [3], [4], we address here the problem of stability and the convergence of the EKF when used as a deterministic observer for multi-input/multi-output nonlinear discrete-time systems written in their general form. Based on a new formulation for the exact linearization technique, we introduce instrumental time-varying matrices for the stability and convergence analysis. It is pointed out that, under mild conditions, asymptotic behavior of the EKF may be improved significantly even for bad first-order approximation. To show accuracy and performances of this technique, we use the modified EKF at first as a parameter estimator for the Hammerstein model and secondly as a simultaneous state and parameter estimator of a nonlinear model.

II. Problem Formulation

The nonlinear systems considered here are of the form

$$x_{k+1} = f(x_k, u_k) \tag{1a}$$

$$y_k = h(x_k, u_k) \tag{1b}$$

where $u_k \in R^r$ and $y_k \in R^p$ are the input and output vectors at time instant k. We assume that $f(x_k, u_k)$ and $h(x_k, u_k)$ are differentiable on R^n. The EKF for the associated noisy system that we use here as an observer of (1a), (1b) is:

1) measurement update

$$\hat{x}_{k+1} = \hat{x}_{k+1/k} + K_{k+1}e_{k+1} \tag{2}$$

$$K_{k+1} = P_{k+1/k}H_{k+1}^T(H_{k+1}P_{k+1/k}H_{k+1}^T + R_{k+1})^{-1} \tag{3}$$

$$P_{k+1} = (I - K_{k+1}H_{k+1})P_{k+1/k} \tag{4}$$

2) time update

$$\hat{x}_{k+1/k} = f(\hat{x}_k, u_k) \tag{5}$$

$$P_{k+1/k} = F_kP_kF_k^T + Q_k \tag{6}$$

Manuscript received June 5, 1995; revised October 17, 1995.

The authors are with the CRAN CNRS URA, Cosnes et Romain 54400, France.

Publisher Item Identifier S 0018-9286(97)02050-3.

where

$$e_{k+1} = y_{k+1} - h(\hat{x}_{k+1/k}, u_{k+1}) \tag{7}$$

$$F_k = F(\hat{x}_k, u_k) = \left.\frac{\partial f(x_k, u_k)}{\partial x_k}\right|_{x_k=\hat{x}_k} \tag{8}$$

$$H_{k+1} = H(\hat{x}_{k+1/k}, u_{k+1}) = \left.\frac{\partial h(x_{k+1}, u_{k+1})}{\partial x_{k+1}}\right|_{x_{k+1}=\hat{x}_{k+1/k}}. \tag{9}$$

When used as an observer for linear deterministic systems, Q_k and R_k are arbitrarily chosen, for example, as 0_n and I_p, respectively. In the case of linear stochastic systems, optimal filtering in the maximum likelihood sense is obtained when Q_k and R_k are the covariance matrices of the system and measurement noises, respectively. However, for nonlinear noisy systems, optimality has not been proved, but in general we continue to consider Q_k and R_k as covariance matrices.

Hereafter, we will show that the design of R_k plays an important role in improving the convergence of the EKF when used as an observer of (1a), (1b). However, since (1a), (1b) is a deterministic system, we set $Q_k = 0$.

III. Convergence Analysis

In this section we give a simple approach for setting up the convergence analysis of the considered nonlinear systems. It is emphasized that under the local reconstructibility condition [19], the asymptotic convergence of the EKF is ensured when the arbitrary matrix R_k is adequately chosen.

Let us note by $\tilde{x}_{k+1}$ and $\tilde{x}_{k+1/k}$ the state estimation and state prediction error vectors respectively defined by

$$\tilde{x}_{k+1} = x_{k+1} - \hat{x}_{k+1} \tag{10}$$

$$\tilde{x}_{k+1/k} = x_{k+1} - \hat{x}_{k+1/k} \tag{11}$$

and the candidate Lyapunov function V_{k+1} as

$$V_{k+1} = \tilde{x}^T_{k+1} P^{-1}_{k+1} \tilde{x}_{k+1}. \tag{12}$$

Our aim here is to determine conditions for which $\{V_k\}_{k=1\cdots}$ is a decreasing sequence and to show the EKF limitations when the first-order approximation is used. One classical approach consists of using the convergence analysis performed in the linear case when e_{k+1} and $\tilde{x}_{k+1/k}$ are approximated as

$$e_{k+1} \approx H_{k+1}\tilde{x}_{k+1/k} \tag{13}$$

and

$$\tilde{x}_{k+1/k} \approx F_k \tilde{x}_k. \tag{14}$$

However, this approximation is available only if $\hat{x}_{k+1/k}$ and $\hat{x}_k$ belong to a neighborhood of x_{k+1} and x_k, respectively, otherwise the EKF diverges.

For a rigorous convergence study we have to show that $\{V_k\}_{k=1\cdots}$ decreases without any approximation. We notice that there always exist residues, due to the first-order linearization technique, of each output error prediction component e_{ik+1} $(i = 1, \cdots, p)$ of e_{k+1} and each state error prediction component $\tilde{x}_{jk+1/k}$ $(j = 1, \cdots, n)$ of $\tilde{x}_{k+1/k}$, for all k. To take these residues into account, in order to obtain an exact equality, we introduce here unknown diagonal matrices α_{k+1} and β_k so that

$$H_{ik+1}\tilde{x}_{k+1/k} = \alpha_{ik+1} \cdot e_{ik+1} \tag{15}$$

$$\tilde{x}_{jk+1/k} = \beta_{jk} F_{jk} \tilde{x}_k. \tag{16}$$

H_{ik+1} and F_{jk} are the ith and jth rows of H_{k+1} and F_k, respectively.

The exact equality represented here by (15) and (16) is the key point of our approach since the convergence study will be performed without any approximation while e_{ik+1} and $\tilde{x}_{jk+1/k}$ are written in a first-order representation. Now if we introduce the signal vector, we obtain

$$\alpha_{k+1}e_{k+1} = H_{k+1}\tilde{x}_{k+1/k} \tag{17}$$

$$\tilde{x}_{k+1/k} = \beta_k F_k \tilde{x}_k \tag{18}$$

where $\alpha_{k+1} \in R^{p \cdot p}$ and $\beta_k \in R^{n \cdot n}$ are unknown time-varying diagonal matrices

$$\alpha_{k+1} = \mathrm{diag}\{\alpha_{1k+1}, \cdots, \alpha_{pk+1}\} \tag{19}$$

and

$$\beta_k = \mathrm{diag}\{\beta_1, \cdots, \beta_{nk}\}. \tag{20}$$

By subtracting both sides of (2) from x_{k+1}, we obtain

$$\tilde{x}_{k+1} = \tilde{x}_{k+1/k} - P_{k+1/k}H_{k+1}(H_{k+1}P_{k+1/k}H^T_{k+1} + R_{k+1})^{-1}e_{k+1}. \tag{21}$$

On the other hand, from (3) and (4), we have

$$P_{k+1}H^T_{k+1}R^{-1}_{k+1} = P_{k+1/k}H^T_{k+1}(H_{k+1}P_{k+1/k}H^T_{k+1} + R_{k+1})^{-1} \tag{22}$$

and

$$P^{-1}_{k+1} = P^{-1}_{k+1/k} + H^T_{k+1}R^{-1}_{k+1}H_{k+1}. \tag{23}$$

Substituting (22) into (21) and (21) into (12), the quadratic function becomes

$$V_{k+1} = (\tilde{x}_{k+1/k} - P_{k+1}H^T_{k+1}R^{-1}_{k+1}e_{k+1})^T \cdot P^{-1}_{k+1}(\tilde{x}_{k+1/k} - P_{k+1}H^T_{k+1}R^{-1}_{k+1}e_{k+1}) \tag{24}$$

or

$$\begin{aligned} V_{k+1} = {} & \tilde{x}_{k+1/k}P^{-1}_{k+1}\tilde{x}_{k+1/k} - \tilde{x}^T_{k+1/k}H^T_{k+1}R^{-1}_{k+1}e_{k+1} \\ & - e^T_{k+1}R^{-1}_{k+1}H_{k+1}\tilde{x}_{k+1/k} \\ & + e^T_{k+1}R^{-1}_{k+1}H_{k+1}P_{k+1}H^T_{k+1}R^{-1}_{k+1}e_{k+1} \end{aligned} \tag{25}$$

and substituting (23) into (25)

$$\begin{aligned} V_{k+1} = {} & V_{k+1/k} + \tilde{x}^T_{k+1/k}H^T_{k+1}R^{-1}_{k+1}H_{k+1}\tilde{x}_{k+1/k} \\ & - \tilde{x}^T_{k+1/k}H^T_{k+1}R^{-1}_{k+1}e_{k+1} + e^T_{k+1}R^{-1}_{k+1}H_{k+1}\tilde{x}_{k+1/k} \\ & + e^T_{k+1}R^{-1}_{k+1}H_{k+1}P_{k+1}H^T_{k+1}R^{-1}_{k+1}e_{k+1} \end{aligned} \tag{26}$$

with

$$V_{k+1/k} = \tilde{x}^T_{k+1/k}P^{-1}_{k+1/k}\tilde{x}_{k+1/k}. \tag{27}$$

From (17) and (18), (26) becomes

$$\begin{aligned} V_{k+1} = {} & V_{k+1/k} + e^T_{k+1}(\alpha_{k+1}R^{-1}_{k+1}\alpha_{k+1} \\ & - \alpha_{k+1}R^{-1}_{k+1} - R^{-1}_{k+1}\alpha_{k+1} \\ & + R^{-1}_{k+1}H_{k+1}P_{k+1}H^T_{k+1}R^{-1}_{k+1})e_{k+1}. \end{aligned} \tag{28}$$

On the other hand, $V_{k+1/k}$ may be written as

$$V_{k+1/k} = \tilde{x}^T_k F^T_k \beta_k (F_k P_k F^T_k)^{-1} \beta_k F_k \tilde{x}_k. \tag{29}$$

A decreasing sequence $\{V_k\}_{k=1\cdots}$ means that

$$V_{k+1} - V_k = V_{k+1} - V_{k+1/k} + V_{k+1/k} - V_k \leq 0 \tag{30}$$

or equivalently

$$\begin{aligned} V_{k+1} - V_k = {}& e_{k+1}^T (\alpha_{k+1} R_{k+1}^{-1} \alpha_{k+1} \\ & - \alpha_{k+1} R_{k+1}^{-1} - R_{k+1}^{-1} \alpha_{k+1} \\ & + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) e_{k+1} \\ & + \tilde{x}_k^T (F_k^T \beta_k (F_k P_k F_k^T)^{-1} \beta_k F_k \\ & - P_k^{-1}) \tilde{x}_k \leq 0. \end{aligned}$$

A sufficient condition to ensure that is

$$\begin{aligned} & \alpha_{k+1} R_{k+1}^{-1} \alpha_{k+1} - \alpha_{k+1} R_{k+1}^{-1} - R_{k+1}^{-1} \alpha_{k+1} \\ & \quad + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1} \leq 0 \end{aligned} \tag{31}$$

and

$$F_k^T \beta_k (F_k P_k F_k^T)^{-1} \beta_k F_k - P_k^{-1} \leq 0. \tag{32}$$

Before we give sufficient conditions to ensure convergence of the EKF, two lemmas are established for intermediate results.

Lemma 1: If we assume that each α_{ik+1} satisfies the following condition:

$$\begin{aligned} & 1 - \sqrt{1 - \Delta_{k+1}} < \alpha_{ik+1} < 1 + \sqrt{1 - \Delta_{k+1}} \\ & \quad \text{for} \quad i = 1, \cdots, p \end{aligned} \tag{33}$$

with

$$\begin{aligned} \delta_{k+1} = {}& \lambda_{\max}(R_{k+1}) \lambda_{\max}(R_{k+1}^{-1} H_{k+1} P_{k+1/k} H_{k+1}^T \\ & \cdot (H_{k+1} P_{k+1/k} H_{k+1}^T + R_{k+1})^{-1}) \end{aligned} \tag{34}$$

where $\lambda_{\max}(\cdot)$ represents the maximum eigenvalue and R_{k+1} is chosen such that $\Delta_{k+1} \leq 1$, then (31) is verified.

Proof: As α_{k+1} is a diagonal matrix, its eigenvalues are given by α_{ik+1} and verify

$$\alpha_{k+1} s_i = \alpha_{ik+1} s_i \tag{35}$$

and

$$s_i^T \alpha_{k+1} = \alpha_{ik+1} s_i^T \quad \text{for} \quad i = 1, \cdots, p \tag{36}$$

where s_i is the associated eigenvector.

Pre- and post-multiplying the left side of (31) by s_i^T and s_i, respectively

$$\begin{aligned} & s_i^T (\alpha_{k+1} R_{k+1}^{-1} \alpha_{k+1} - \alpha_{k+1} R_{k+1}^{-1} - R_{k+1}^{-1} \alpha_{k+1} \\ & \quad + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) s_i \leq 0 \end{aligned} \tag{37}$$

and using relations (35) and (36) in (37) yields to

$$\begin{aligned} & s_i^T (\alpha_{ik+1}^2 R_{k+1}^{-1} - 2\alpha_{ik+1} R_{k+1}^{-1} \\ & \quad + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) s_i \leq 0 \end{aligned} \tag{38}$$

or

$$\begin{aligned} & \alpha_{ik+1}^2 R_{k+1}^{-1} - 2\alpha_{ik+1} R_{k+1}^{-1} \\ & \quad + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1} \leq 0. \end{aligned} \tag{39}$$

By using the measure of matrix properties [17], corresponding to norm 2, the left side of (39) can be bounded as

$$\begin{aligned} & \mu((\alpha_{ik+1}^2 - 2\alpha_{ik+1}) R_{k+1}^{-1} + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) \\ & \quad \leq \mu((\alpha_{ik+1}^2 - 2\alpha_{ik+1}) R_{k+1}^{-1}) \\ & \qquad + \mu(R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) \end{aligned} \tag{40}$$

where $\mu(\cdot)$ is the measure of matrix defined by

$$\mu(A) = \lambda_{\max}\left(\frac{A + A^T}{2}\right).$$

As R_{k+1}^{-1} and $R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}$ are symmetric matrices and as the interval $]1 - \sqrt{1 - \Delta_{k+1}}, 1 + \sqrt{1 - \Delta_{k+1}}[\subset]0, 2[$, this implies that $\alpha_{ik+1}^2 - 2\alpha_{ik+1} < 0$, and (40) yields to

$$\begin{aligned} & \lambda_{\max}((\alpha_{ik+1}^2 - 2\alpha_{ik+1}) R_{k+1}^{-1} + R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) \\ & \quad \leq -(\alpha_{ik+1}^2 - 2\alpha_{ik+1}) \lambda_{\max}(R_{k+1}^{-1}) \\ & \qquad + \lambda_{\max}(R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1}) \end{aligned} \tag{41}$$

with

$$\begin{aligned} & R_{k+1}^{-1} H_{k+1} P_{k+1} H_{k+1}^T R_{k+1}^{-1} \\ & \quad = R_{k+1}^{-1} H_{k+1} P_{k+1/k} H_{k+1}^T (H_{k+1} P_{k+1/k} H_{k+1}^T R_{k+1})^{-1}. \end{aligned} \tag{42}$$

Therefore, under (33) and by the use of (41) and (42), it follows that (31) is satisfied.

Lemma 2: If we assume that

$$F_k \text{ is a bounded nonsingular matrix} \tag{43}$$

and each β_{jk} satisfies the following condition:

$$-1 \leq \beta_{jk} \leq 1 \quad \text{for} \quad j = 1, \cdots, n \tag{44}$$

then (32) is verified.

Proof: As β_k is a diagonal matrix, its eigenvalues are given by β_{jk} and verify

$$\beta_k m_j = \beta_{jk} m_j \tag{45}$$

and

$$m_j^T \beta_k = \beta_{jk} m_j^T \quad \text{for} \quad j = 1, \cdots, n \tag{46}$$

where m_j is the associated eigenvector.

Under assumption (43), (32) is equivalent to

$$\beta_k F_k^{-T} P_k^{-1} F_k^{-1} \beta_k - F_k^{-T} P_k^{-1} F_k^{-1} \leq 0. \tag{47}$$

Pre- and post-multiplying the left side of (47) by m_j^T and m_j, respectively

$$m_j^T (\beta_k F_k^{-T} P_k^{-1} F_k^{-1} \beta_k - F_k^{-T} P_k^{-1} F_k^{-1}) m_j \leq 0. \tag{48}$$

Using (45) and (46) in (48) yields

$$\beta_{jk}^2 F_k^{-T} P_k^{-1} F_k^{-1} - F_k^{-T} P_k^{-1} F_k^{-1} \leq 0 \tag{49}$$

thus under assumption (44), it is easy to show that (32) is also verified.

Unfortunately, the convergence of the EKF is not ensured even if $\{V_k\}_{k=1\cdots}$ is a decreasing sequence. So, some additional conditions in relation to the reconstructibility are needed. For nonlinear systems, the local reconstructibility may be defined in the same way as in the linear case [19]. Let us recall this property, which will be partly investigated in our developments. Under (43), if there exist positive real numbers η_1 and η_2 so that for all $k \geq M$ and for some finite $M \geq 0$ we have

$$\eta_1 I_n \leq O_e^T(k-M,k) \Re(k-M,k) O_e(k-M,k) \leq \eta_2 I_n \tag{50}$$

with

$$O_e(k-M,k) = \begin{bmatrix} H_{k-M} F_{k-M}^{-1} F_{k-M+1}^{-1} \cdots F_{k-1}^{-1} \\ H_{k-M+1} F_{k-M+1} \cdots F_{k-1} \\ \cdot \\ \cdot \\ H_k \end{bmatrix} \tag{51}$$

$$\Re(k-M,k) = \mathrm{Diag}(R_{k-M}^{-1}, \cdots, R_k^{-1}) \tag{52}$$

where F_k and H_k are defined by (8) and (9), respectively, then the system (and its associated EKF for $\hat{x}_{i/i-1}$ and $\hat{x}_i$ sufficiently close to the true state x_i) is said to be reconstructible. Thus, we obtain the following lemma.

Lemma 3: If we assume that (1a), (1b) is reconstructible (50), then we have

$$\lim_{k\to\infty} \lambda_{\min}(P_k^{-1}) = \infty \tag{53}$$

and

$$\text{Sup} \lim_{k\to\infty} \frac{\lambda_{\max}(P_k^{-1})}{\lambda_{\min}(P_k^{-1})} < \infty. \tag{54}$$

Proof: From (23) and under (43) it is easy to show, by induction, that

$$P_{k+1}^{-1} = O_e^T(1, k+1)\Re(1, k+1)O_e(1, k+1) + \Psi(0, k) \tag{55}$$

with

$$\Psi(0, k) = F_k^{-T} F_{k-1}^{-T} \cdots F_0^{-T} P_0^{-1} F_0^{-1} \cdots F_{k-1}^{-1} F_k^{-1}. \tag{56}$$

$O_e(1, k+1)$ and $\Re(1, k+1)$ are defined in (51) and (52), respectively.

On a horizon time kM, if the reconstructibility Gramian $O_e^T(1, kM)\Re(1, kM)O_e(1, kM)$ is decomposed into k block matrices of horizon M, we obtain

$$\begin{aligned} P_{kM}^{-1} = \sum_{i=1}^{k} [O_e^T((i-1)M+1, iM)\Re((i-1)M \\ + 1, iM)O_e((i-1)M+1, iM)] + \psi(0, kM-1). \end{aligned} \tag{57}$$

Since $P_{kM}^{-1} - \Psi(0, kM-1)$ is the sum of k reconstructibility Gramians, then under (50) we obtain the following inequalities:

$$o < k\eta_1 \le \lambda(P_{kM}^{-1} - \Psi(0, kM-1)) \le k\eta_2. \tag{58}$$

Thus, as $\lambda(P_{kM}^{-1} - \Psi(0, kM-1))$ is bounded by $k\eta_1$ and $k\eta_2$, it is easy to show that (53) and (54) follow from (58), where λ represents the eigenvalue symbol of $P_{kM}^{-1} - \Psi(0, kM-1)$.

Now, we propose a simple method to prove local convergence of the EKF applied to deterministic nonlinear systems.

Theorem: Suppose that (33), (44), and (50) hold, then the EKF (2)–(6), when used as an observer for the nonlinear discrete-time system (1a)–(1b), ensures that

$$\lim_{k\to\infty} (x_k - \hat{x}_k) = 0.$$

Proof: Under (33) and (44), it has been shown (according to Lemmas 1 and 2) that $\{V_k\}_{k=1\cdots}$ is a decreasing sequence which converges to a positive scalar V, i.e.,

$$\lim_{k\to\infty} V_k = V.$$

On the other hand, we have

$$\frac{V_k}{\text{tr}(P_k^{-1})} \ge \frac{\lambda_{\min}[P_k^{-1}]\tilde{x}_k^T \tilde{x}_k}{n\lambda_{\max}[P_k^{-1}]} \ge 0. \tag{59}$$

According to (53) one obtains

$$\lim_{k\to\infty} \text{tr}(P_k^{-1}) = \infty \tag{60}$$

then

$$\lim_{k\to\infty} \frac{\lambda_{\min}[P_k^{-1}]\tilde{x}_k^T \tilde{x}_k}{n\lambda_{\max}[P_k^{-1}]} = 0 \tag{61}$$

thus (54) yields to

$$\lim_{k\to\infty} \tilde{x}_k = 0. \tag{62}$$

Some Concluding Remarks

- α_{ik} and β_{jk} are unknown factors introduced to evaluate the linearity of the model and, consequently, to control the gain matrix K_{k+1} by an adequate choice of R_k in order to ensure stability of the algorithm, particularly when $e_{k+1} \ne H_{k+1/k}\tilde{x}_{k+1/k}$ and $\tilde{x}_{k+1/k} \ne F_k\tilde{x}_k$. Indeed, the sufficient conditions (33) and (44) mean that $\{V_k\}$ is a decreasing sequence for all approximations in the form

  $$\alpha_{ik+1}e_{ik+1} = H_{ik+1}\tilde{x}_{k+1/k} \quad \text{and} \quad \tilde{x}_{jk+1/k} = \beta_{jk}F_{jk}\tilde{x}_k$$

  with

  $$\alpha_{ik+1} \in]1 - \sqrt{1-\Delta_{k+1}}, 1 + \sqrt{1-\Delta_{k+1}}[$$
  $$\text{and} \qquad \beta_{jk} \in [-1, +1].$$

  As α_{ik+1} and β_{jk} are unknown factors, R_{k+1} have to be chosen so that the interval $]1 - \sqrt{1-\Delta_{k+1}}, 1 + \sqrt{1-\Delta_{k+1}}[$ is large enough to fulfil (33). An *a priori* choice of R_{k+1} is to set it much higher than $H_{k+1}P_{k+1/k}H_{k+1}^T$. Notice that within the maximum limits we have

  $$]1 - \sqrt{1-\Delta_{k+1}}, 1 + \sqrt{1-\Delta_{k+1}}[\rightarrow]0, 2[.$$

  However, setting high values of R_{k+1} leads to a very slow convergence rate (the gain matrix K_{k+1} goes to zero). A tradeoff between stability and rate of convergence of the proposed algorithm leads us to set (in our numerical simulations)

  $$R_{k+1} = \mu H_{k+1}P_{k+1/k}H_{k+1}^T + \zeta I_p \text{ where } \mu > 0 \quad \text{and}$$
  $$\zeta > 0 \text{ are fixed by the user.}$$

- For MISO nonlinear systems, R_k is a positive scalar, and then (33) becomes

  $$\Delta_{k+1} = H_{k+1}P_{k+1/k}H_{k+1}^T(H_{k+1}P_{k+1/k}H_{k+1}^T + R_{k+1})^{-1}$$

  with $1 - \Delta_{k+1} > 0$ for all k.

- For linear time-varying systems we have $\alpha_k = 1$ and $\beta_k = I_n$. Then (31) and (32) become

  $$-1 + H_{k+1}P_{k+1/k}H_{k+1}^T(H_{k+1}P_{k+1/k}H_{k+1}^T + R_{k+1})^{-1} \le 0$$

  and

  $$F_k^T(F_kP_kF_k^T)^{-1}F_k - P_k^{-1} \le 0.$$

  For $R_k > 0, \{V_k\}$ is a decreasing sequence.

IV. Numerical Examples

In order to show the efficiency of the proposed approach, we consider two numerical examples.

Example 1: The first example concerns the identification problem of nonlinear time-invariant systems described by the MISO Hammerstein model. This kind of model is composed by a static nonlinearity followed by a linear dynamic system. Here we consider two interconnected subsystems (high order and high values of g_{ik}) with the following pulse transfer functions:

$$\begin{aligned} V_{1k} &= g_{11}u_{1k} + g_{12}u_{1k}^2 + g_{13}u_{1k}^3 + g_{14}u_{1k}^4 + g_{15}u_{1k}^5 \\ V_{2k} &= g_{21}u_{2k} + g_{22}u_{2k}^2 + g_{23}u_{2k}^3 \\ &\quad + g_{24}u_{2k}^4 + g_{25}u_{2k}^5 + g_{26}u_{2k}^6 \\ \frac{B_1}{A_1} &= \frac{q^{-1} + b_{11}q^{-2}}{1 + a_{11}q^{-1} + a_{12}q^{-2}} \\ \frac{B_2}{A_2} &= \frac{q^{-1} + b_{21}q^{-2} + b_{22}q^{-3}}{1 + a_{21}q^{-1} + a_{22}q^{-2}} \end{aligned}$$

where q^{-1} is the delay operator and u_{ik} is the ith input of the system at time instant $k, i = 1, 2,$ with

$$\begin{aligned} a_{11} &= 0.4; \quad a_{12} = .65; \quad a_{21} = 0.75; \quad a_{22} = 0.9 \\ b_{11} &= 0.5; \quad b_{21} = -0.6; \quad b_{22} = 0.7 \\ g_{11} &= 5.2; \quad g_{12} = -2.0; \quad g_{13} = 5.2; g_{14} = -3.5 \\ g_{12} &= 6.5; \quad g_{21} = 6.3; \quad g_{22} = 2.8; \quad g_{23} = -0.02 \\ g_{24} &= 3.1; \quad g_{25} = -2.3; \quad g_{26} = 5.6. \end{aligned}$$

Parameters of the static nonlinearity V_1 and V_2 are independent of those of $A_1, A_2, B_1,$ and B_2.

The invariant parameter vector x (i.e., $x_{k+1} = x_k$) to be estimated from the input output data $(y_k, u_{1k}, u_{2k})_{k=1,\cdots}$, is defined as

$$\begin{aligned} x = (&a_{11} \quad a_{12} \quad a_{21} \quad a_{22} \quad b_{11} \quad b_{21} \quad b_{22} \quad g_{11} \\ &\cdots \quad g_{15} \quad \cdots \quad g_{21} \quad \cdots \quad g_{26})^T \in R^{18}. \end{aligned}$$

A nonlinear state-space representation of the input–output Hammerstein model may be written as

$$\begin{aligned} x_{k+1} &= f(x_k, u_{1k}, u_{2k}) = x_k \\ y_k &= h(x_k, u_{1k}, u_{2k}) \\ &= -(a_{11} + a_{21})y_{k-1} - (a_{11}a_{21} + a_{22} + a_{12})y_{k-2} \\ &\quad - (a_{11}a_{22} + a_{21}a_{12})y_{k-3} - a_{12}a_{22}y_{k-4} \\ &\quad + V_{1k-1} + (b_{11} + a_{21})V_{1k-2} \\ &\quad + (b_{11}a_{21} + a_{22})V_{1k-3} \\ &\quad + b_{11}a_{22}V_{1k-4} + V_{2k-1} + (b_{21} + a_{11})V_{2k-2} \\ &\quad + (b_{21}a_{11} + a_{12} + b_{22})V_{2k-3} \\ &\quad + (a_{11}b_{22} + b_{21}a_{12})V_{2k-4} + a_{12}b_{22}V_{2k-5}. \end{aligned}$$

In order to fulfil (33) and (44) with a good convergence rate, we take

$$R_{k+1} = 2H_{k+1}P_{k+1/k}H_{k+1}^T + 1.$$

Owing to a lack of space here, we consider a worst case of very bad initialization. $\hat{x}_0 = 100$, i.e., all parameters are initialized at 100 with $P_0 = 10^7\boldsymbol{I}_{18}$. The input signals (u_{1k}) and (u_{2k}) are zero mean white noise sequences with standard deviations $\sigma_1 = 0.3$ and $\sigma_2 = 0.4$, respectively.

Example 2: The second numerical example, which was worked in [18], consists of a combined parameter and state estimation of the following discrete-time system:

$$\begin{bmatrix} x_{1k+1} \\ x_{2k+1} \\ x_{3k+1} \\ x_{4k+1} \\ x_{5k+1} \end{bmatrix} = \begin{bmatrix} x_{2k} \\ -a_0x_{1k} - a_1x_{2k} + bu_k \\ x_{3k} \\ x_{4k} \\ x_{5k} \end{bmatrix}$$

$$\begin{aligned} y_k &= x_{1k}x_{2k} \quad \text{with} \\ a_0 &= 0.3 + 0.1 \ \sin(x_{3k}) \\ a_1 &= 1.1 + 0.1 \ \sin(x_4k) \\ b &= 2.4 + 0.1 \ \sin(x_{5k}) \end{aligned}$$

and

$$u_k = 5 + 2 \ \sin(0.8k) + 2 \ \sin(1.8k).$$

The initial conditions are

$$\begin{aligned} x_{10} &= 4, \quad x_{20} = 5, \quad x_{30} = 0, \quad x_{40} = 0, \quad x_{50} = 0 \\ \hat{x}_{10} &= 20, \quad \hat{x}_{20} = 20, \quad \hat{x}_{30} = 1, \quad \hat{x}_{40} = 1, \quad \hat{x}_{50} = 1 \end{aligned}$$

with

$$R_{k+1} = 3H_{k+1}P_{k+1/k}H_{k+1}^T + 1 \quad \text{and} \quad P_0 = 10^{20}\boldsymbol{I}_5$$

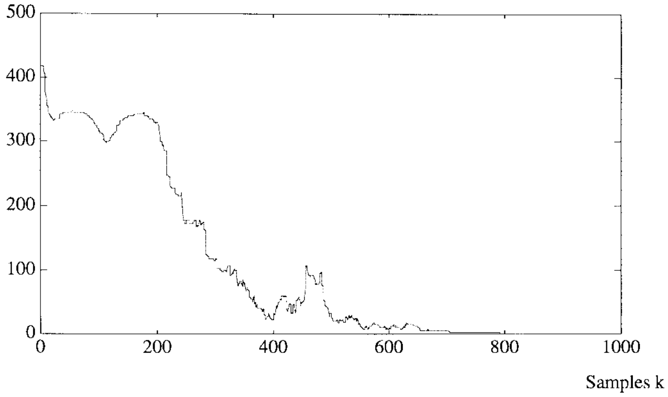

Fig. 1. Rate of convergence of $\|\hat{x}_k - x_k\|$—Example 1.

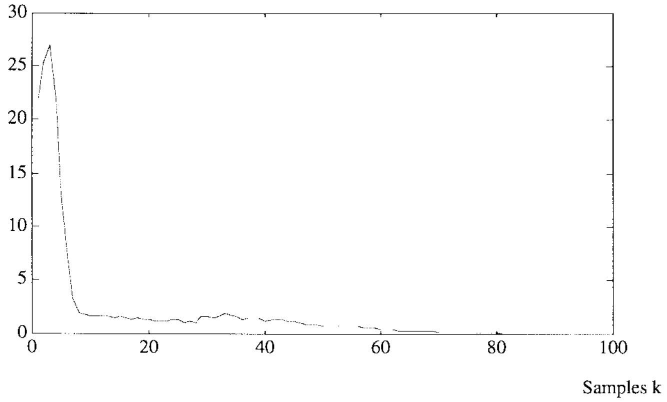

Fig. 2. Rate of convergence of $\|\hat{x}_k - x_k\|$—Example 2.

Figs. 1 and 2 show the rate of convergence of the state error norm $\|\hat{x}_k - x_k\|$ with respect to samples k when R_k is adequately chosen for both examples, while the classical EKF diverges for $R_k = 1$. We notice that in spite of bad initialization and with a large scale system (example 1), we obtain excellent results with our approach: the actual values of parameters are reached approximately at 800 samples for the first example and at 80 samples for the second one.

As we expect, the reconstructibility condition for nonlinear systems depends closely on the input signals especially for identification of nonlinear systems. We notice that in the case of parameter estimation, reconstructibility of the system is equivalent to the persistently exciting condition since $x_{i+1} = x_i$ (for time-invariant systems where x_i represents the parameters vector to be estimated), and therefore we have $F_i = F_i^{-1} = I, I$ is the identity matrix, for all i, and

$$\begin{aligned} O_e(k-M, k) &= \begin{bmatrix} H_{k-M}F_{k-M}^{-1}F_{k-M+1}^{-1} \cdots F_{k-1}^{-1} \\ H_{k-M+1}F_{k-M+1} \cdots F_{k-1}^{-1} \\ \cdot \\ \cdot \\ H_k \end{bmatrix} \\ &= \begin{bmatrix} H_{k-M} \\ H_{k-M+1} \\ \cdot \\ \cdot \\ H_k \end{bmatrix}. \end{aligned}$$

V. Conclusion

In this paper, convergence analysis of the EKF used as an observer for nonlinear deterministic discrete-time systems was considered. It is shown, from the theoretical point of view, that under mild conditions and with an appropriate choice of the arbitrary matrix R_k, convergence of the EKF may be improved significantly in the sense that the domain of attraction is enlarged. One of the main results in this paper consists of introducing instrumental matrices α_k and β_k to evaluate the linearity of the model in order to control both stability and convergence of the EKF.

Acknowledgment

The authors wish to thank the anonymous referees for their insightful comments.

References

[1] J. S. Baras, A. Bensoussan, and M. R. James, "Dynamic observers as asymptotic limits of recursive filters: Special cases," *Siam J. Appl. Math.*, vol. 48, no. 5, pp. 1147–1158, 1988.

[2] D. Bestle and M. Zeitz, "Canonical form observer design for non linear time-variable systems," *Int. J. Contr.*, vol. 38, pp. 419–431, 1983.

[3] M. Boutayeb and M. Darouach, "Recursive identification method for Hammerstein model—Extension to the nonlinear MISO case," *Contr. Theory Advanced Technol.*, vol. 10, no. 1, pp. 57–72, 1994.

[4] ——, "Recursive identification method for MISO Wiener-Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. 40, no. 2, pp. 287–291, 1995.

[5] J. J. Deyst and C. F. Price, "Conditions for asymptotic stability of the discrete minimum-variance linear estimator," *IEEE Trans. Automat. Contr.*, pp. 702–705, 1968.

[6] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1974.

[7] A. H. Jaswinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[8] A. J. Krener and A. Isidori, "Linearization by output injection and nonlinear observers," *Syst. Contr. Lett.*, vol. 3, pp. 47–52, 1983.

[9] A. J. Krener and W. Respondek, "Nonlinear observers with linearizable error dynamics," *Siam J. Contr. Optim.*, vol. 23, pp. 197–216, 1985.

[10] H. J. Kushner, "Approximations to optimal nonlinear filters," *IEEE Trans. Automat. Contr.*, vol. AC-12, no. 5, pp. 546–556, 1967.

[11] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *IEEE Trans. Automat. Contr.*, AC-24, pp. 36–50, 1979.

[12] T. P. McGarty, *Stochastic Systems and State Estimation.* New York: Wiley, 1973.

[13] R. K. Mehra, "A comparison of several nonlinear filters for reentry vehicle tracking," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 307–319, 1971.

[14] H. Nijmeijer, "Observability of autonomous discrete-time nonlinear systems: A geometric approach," *Int. J. Contr.*, vol. 36, pp. 867–874, 1982.

[15] Y. Song and J. W. Grizzle, "The extended Kalman filter as a local asymptotic observer for nonlinear discrete-time systems," in *Proc. Amer. Contr. Conf.*, 1992, pp. 3365–3369.

[16] Special issue on applications of Kalman filtering, *IEEE Trans. Automat. Contr.*, vol. AC-28, no. 3, 1983.

[17] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] G. Ciccarella, M. Dalla Mora, and A. Germani, "A robust observer for discrete time nonlinear systems," *Syst. Contr. Lett.*, vol. 24, pp. 291–300, 1995.

[19] P. E. Moraal and J. W. Grizzle, "Observer design for nonlinear systems with discrete-time measurements," *IEEE Trans. Automat. Contr.*, vol. 40, no. 3, 1995.

Dynamics and Convergence Rate of Ordinal Comparison of Stochastic Discrete-Event Systems

Xiaolan Xie

***Abstract*—This paper addresses ordinal comparison in the simulation of discrete-event systems. It examines dynamic behaviors of ordinal comparison in a fairly general framework. It proves that for regenerative systems, the probability of obtaining a desired solution using ordinal comparison approaches converges at exponential rate, while the variances of the performance measures converge at best at rate $O(1/t^2)$, where t is the simulation time. Heuristic arguments are provided to explain that exponential convergence holds for general systems.**

I. Introduction

Optimization in discrete solution space becomes more and more important for discrete-event dynamic systems. The only general tool for evaluating such systems is the simulation. A straightforward and widely used approach consists of simulating all candidate designs to obtain accurate estimation of the performance measures and selecting the best design. Its main drawback is the requirement of a long simulation run to obtain accurate performance estimators. Variances of performance measures converge typically at rate $O(1/t)$ in time t. and these rates are usually unsatisfactory when the number of candidate designs is important.

Ordinal optimization approaches first proposed in [10] (see also [11] for an overview) reduce the computation burden by "combining some mind-set changes" concerning the problem of optimization of discrete-event systems. The primary concern of the ordinal optimization is the rankings of candidate solutions instead of their criterion values. Simulations conducted by various authors for a wide range of problems have shown that the rankings stabilize before the convergence of performance estimates. To achieve further reduction of simulation time, ordinal optimization approaches typically relax the goal of optimization to the isolation of a set of good candidate solutions. Experiments indicate that it is possible to determine whether a solution is good or bad very early in the simulation with high probability. Recent research [2], [3], [12] has demonstrated that impressive improvement in computation efficiency can be achieved using ordinal optimization approaches.

The purpose of this paper is to provide theoretical evidence of the efficiency of ordinal optimization approaches. It is an extension of a recent work [4] which considers the convergence of the probability that the best observed design is indeed a "good" design. We consider the following fundamental indicator: the probability that at least k of the observed top-s designs are the actual top-g designs (i.e., satisfactory designs). It is called alignment probability and is meaningful in the perspective of simulation budget allocation. Our contributions include:

1) monotonicity properties of the alignment probability with respect to s, g, and k are established. They show that goal relaxation improves the computation efficiency;
2) the association, a kind of positive correlation, of simulated systems improves the convergence rate of the probability that the observed best design is the actual best one;
3) informal arguments of the exponential convergence rate of the alignment probability in the general case;

Manuscript received July 14, 1995; revised June 12, 1996.

The author is with INRIA, Technopole Metz 2000, 57070 Metz, France.

Publisher Item Identifier S 0018-9286(97)02044-8.