The Extended Kalman Filter as a Local Asymptotic Observer for Nonlinear Discrete-Time Systems[†]

Yongkyu Song
AERO and EECS Departments

J. W. Grizzle
EECS Department

University of Michigan
Ann Arbor, MI 48109–2122
USA

Abstract

The convergence aspects of the extended Kalman filter, when used as a deterministic observer for a nonlinear discrete-time system, are analyzed. To a certain extent, the results parallel those of [1] for continuous-time systems. However, in addition to the analysis done in [1], the case of systems with nonlinear output maps is treated and the conditions needed to ensure the uniform boundedness of certain Riccati equations are related to the observability properties of the underlying nonlinear system. Furthermore, we show the convergence of the filter without any *a priori* boundedness assumptions on the error covariances whenever the states stay within a convex compact domain.

[†] Work supported by the National Science Foundation under contract NSF ECS-88-96136 with matching funds provided by the FORD MO. CO.

September 13, 1991

1. Introduction

Designing an observer for a nonlinear system is quite a challenge. Thus, as a first step, it is interesting to see how classical linearization techniques work with nonlinear systems and what their limitations are. In [1], Baras et al. describe a method for constructing observers for dynamic systems as asymptotic limits of filters. They discuss the method as applied to the linear case, and a class of nonlinear systems with linear observations, in continuous-time domain. Essentially the extended Kalman filter(EKF) is used as their observer[1,6].

Motivated by their work, we analyze the convergence aspects of the EKF when it is used as a deterministic observer for a nonlinear discrete-time system. That is, we will consider the system:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k), \quad x_0 \text{ given}, \\ y_k &= h(x_k, u_k), \end{aligned} \tag{1.1}$$

and the EKF for the associated "noisy" system:

$$\begin{aligned} z_{k+1} &= f(z_k, u_k) + Nw_k, \\ \xi_k &= h(z_k, u_k) + Rv_k. \end{aligned} \tag{1.2}$$

Throughout the paper $x, w \in \boldsymbol{R}^n$ and $y, v \in \boldsymbol{R}^p$ and f, h are assumed to be at least twice differentiable. As usual, z_0, v_k, and w_k are assumed jointly Gaussian and mutually independent. Furthermore $z_0 \sim \mathcal{N}(\bar{x}_0, \bar{Q}_0), w_k \sim \mathcal{N}(0, I_n)$, and $v_k \sim \mathcal{N}(0, I_p)$. We also assume that N has rank n and R and $\bar{Q}_0$ are positive definite.

We denote by $|\cdot|$, the Euclidean norm of a vector, and by $\|\cdot\|$ and $|||\cdot|||$, the induced norms on matrices and tensors. The symbol ":=" means that the RHS is defined to be equal to the LHS; the reverse holds for " =: ".

Section 2 discusses the convergence of the Kalman filter for the class of linear systems; it is emphasized that, especially for time-invariant systems, it is not necessary to *assume* the uniform boundedness of the error covariances (cf. condition (28) in [1]) since it is *implied* by the usual detectability condition and the invertibility of the system matrix. In Section 3, we consider the case of nonlinear systems with nonlinear output maps. The conditions needed to

ensure the uniform boundedness of certain Riccati equations are related to the observability properties of the underlying nonlinear system in Section 4. In Section 5, convergence of the EKF without any boundedness assumption on the error covariances is proven whenever the states stay within a convex compact set, which is not necessarily small. These results show that the EKF is a quasi-local observer[14]. Conclusions are made in Section 6.

2. A global asymptotic observer for linear time-varying systems

In this Section we explicitly show that the Kalman filter for linear systems with artificial noises can be used as a global asymptotic observer for the underlying deterministic system. The results summarized here are essential for setting up the analysis on nonlinear systems, which is done in Section 3 through Section 5.

Consider the linear system:

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \quad x_0 \text{ given}, \\ y_k &= C_k x_k, \end{aligned} \tag{2.1}$$

where A_k is assumed invertible, and consider also the associated "*noisy*" system:

$$\begin{aligned} z_{k+1} &= A_k z_k + B_k u_k + N w_k, \\ \xi_k &= C_k z_k + R v_k, \end{aligned} \tag{2.2}$$

where the design parameters N and R are assumed positive definite. Then the Kalman filter equations for (2.2) are given as follows[3].

Measurement update:

$$\hat{x}_k = \hat{x}_k^- + K_k(\xi_k - C_k \hat{x}_k^-), \tag{2.3a}$$

$$Q_k^{-1} = (Q_k^-)^{-1} + C_k^T (RR^T)^{-1} C_k, \tag{2.3b}$$

Time update:

$$\hat{x}_{k+1}^- = A_k \hat{x}_k + B_k u_k, \tag{2.4a}$$

$$Q_{k+1}^- = A_k Q_k A_k^T + NN^T, \tag{2.4b}$$

where

$$K_k = Q_k C_k^T (RR^T)^{-1} = Q_k^- C_k^T (C_k Q_k^- C_k^T + RR^T)^{-1}$$

and Q_k^-, Q_k are the *a priori* and *a posteriori* error covariances, respectively. The filter is initiated by setting $\hat{x}_0^- = \bar{x}_0$ and $Q_0^- = \bar{Q}_0$; $\bar{Q}_0$ is used as a design parameter, assumed also positive definite.

To obtain an error dynamics, let's rewrite the Kalman filter in terms of the *a priori* variables. ¿From (2.3) and (2.4) we have, noting that we use y_k instead of ξ_k,

$$\hat{x}_{k+1}^- = A_k(I - K_k C_k)\hat{x}_k^- + B_k u_k + A_k K_k y_k, \tag{2.5}$$

$$Q_{k+1}^- = A_k(I - K_k C_k)Q_k^- A_k^T + NN^T. \tag{2.6}$$

If we define the error as $e_k = x_k - \hat{x}_k^-$, then the error dynamics is given as

$$e_{k+1} = A_k(I - K_k C_k)e_k. \tag{2.7}$$

The associated Riccati equations for the error covariances are

$$Q_{k+1}^- = A_k[(Q_k^-)^{-1} + C_k^T (RR^T)^{-1} C_k]^{-1} A_k^T + NN^T, \tag{2.8}$$

$$Q_{k+1}^{-1} = [A_k Q_k A_k^T + NN^T]^{-1} + C_k^T (RR^T)^{-1} C_k. \tag{2.9}$$

We note that taking $Q_0^- = \bar{Q}_0 > 0$ and rank $N = n$ implies $Q_k^- > 0$ and $Q_k > 0$ for all $k \geq 0$.

Since we are interested in the asymptotic behavior of the error, e_k, it is necessary to obtain bounds for $\|Q_k^-\|$ and $\|Q_k^{-1}\|$.

Deyst and Price [11] obtained a sufficient condition which gives lower and upper bounds of Q_k. Consider the following "noisy" system:

$$\begin{aligned} x_{k+1} &= A_k x_k + N w_k, \\ y_k &= C_k x_k + R v_k, \end{aligned} \tag{2.10}$$

Suppose that there are real numbers $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that the following conditions hold for all $k \geq M$ and for some finite $M \geq 0$:

$$\alpha_1 I \geq \sum_{i=k-M}^{k-1} \Phi(k, i+1) NN^T \Phi^T(k, i+1) \geq \alpha_2 I, \quad 0 < \alpha_1, \alpha_2 < \infty, \tag{2.11}$$

$$\beta_1 I \leq \sum_{k-M}^{k} \Phi^T(i,k) C_k^T (RR^T)^{-1} C_k \Phi(i,k) \leq \beta_2 I, \quad 0 < \beta_1, \beta_2 < \infty; \tag{2.12}$$

then

$$\frac{1}{\beta_2 + 1/\alpha_2} I \leq Q_k \leq (\alpha_1 + 1/\beta_1) I,$$

where

$$\Phi(k,i) = A_{k-1} A_{k-2} \cdots A_i.$$

Thus from (2.4b)

$$\|Q_k^-\| \leq (\alpha_1 + 1/\beta_1)\|A\|^2 + \|N\|^2.$$

The conditions (2.11) and (2.12) imply that the "noisy" system (2.10) is stochastically controllable and observable[12]. The condition (2.11) is immediately satisfied with nonsingular design parameter N. On the other hand, let's take $R = I$, R being a design parameter; then condition (2.12) is satisfied if the deterministic part of the system (2.10), i.e., the pair (A_k, C_k), is uniformly completely reconstructible[13].

Baras et al.[1] have also obtained bounds for the error covariances in continuous-time, using dual optimal control problems. Similar methods yield bounds for the error covariances in discrete-time. The bounds for the case of linear time invariant systems are explicitly shown in the Appendix, and follow from the detectability of the pair (A, C) and the invertibility of A. We will discuss how observability is related to the boundedness of the error covariances in the extended Kalman filter later in Section 4. For now, we make the following assumption, which, we note, is implied by the uniform observability of (A_k, C_k).

Assumption 2.1 *The error covariances of the Kalman filter (2.3) and (2.4) are uniformly bounded , i.e., there exists $q < \infty$ and $p < \infty$ such that $\|Q_k^-\| \leq q$ and $\|Q_k^{-1}\| \leq p$ for all $k \geq 0$.*

Now with these bounds we can show that the error converges to zero asymptotically. Before proceeding we need a lemma from Lyapunov stability theory [2, Theorem 4.8.3].

Definition 2.2 A function ϕ is said to be of class K if it is continuous in $[0, a)$, strictly increasing and $\phi(0) = 0$. Let $\boldsymbol{N}^+$ be the set of nonnegative integers, $\boldsymbol{R}^+$ the set of positive

reals, and B_a the open ball having center at 0 and radius a.

Lemma 2.3 Assume for some $a > 0$ that there exists a function V such that

(1) $V : \boldsymbol{N}^+ \times B_a \mapsto \boldsymbol{R}^+; V(k,0) = 0; V$ is positive definite and continuous with respect to the second argument;

(2) $\Delta V(k, e_k) = V(k+1, e_{k+1}) - V(k, e_k) \leq -\mu(|e_k|)$, where μ is of class K. Then the origin of (2.7) is asymptotically stable.

Theorem 2.4 *Consider the system (2.1) and the Kalman filter equations (2.3) and (2.4) for the associated system (2.2). Suppose that A_k is invertible for all k and that Assumption 2.1 holds. Suppose further that $\|A\| := sup\{\|A_k\| : k = 0, 1, \cdots\}$ and $\|C\| := sup\{\|C_k\| : k = 0, 1, \cdots\}$ are bounded. Then the Kalman filter for the noisy system (2.2) is a global asymptotic observer for the deterministic system (2.1), as long as N has rank n and R and $\bar{Q}_0$ are positive definite.*

Proof: Let $P_K^- = (Q_k^-)^{-1}$. From (2.4b)

$$A_k^{-1} Q_{k+1}^- A_k^{-T} = Q_k + A_k^{-1} N N^T A_k^{-T}.$$

Inverting the above equation

$$A_k^T P_{k+1}^- A_k = Q_k^{-1} - Q_k^{-1}(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} Q_k^{-1}.$$

If we note $Q_k = (I - K_k C_k) Q_k^-$ or $Q_k^{-1} = (Q_k^-)^{-1}(I - K_k C_k)^{-1}$ then

$$A_k^T P_{k+1}^- A_k = \{P_k^- - P_k^-(I - K_k C_k)^{-1}(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^-\}(I - K_k C_k)^{-1} \quad (2.13)$$

Thus, from (2.7) and (2.13)

$$\begin{aligned} e_{k+1}^T P_{k+1}^- e_{k+1} &= e_k^T (I - K_k C_k)^T A_k^T P_{k+1}^- A_k (I - K_k C_k) e_k \\ &= e_k^T (I - K_k C_k)^T \{P_k^- - P_k^-(I - K_k C_k)^{-1}(Q_k^{-1} \\ &\quad + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^-\} e_k. \end{aligned}$$

Since $Q_k = (I - K_k C_k)Q_k^-$ is symmetric, $(I - K_k C_k)^T = P_k^-(I - K_k C_k)Q_k^-$. Therefore,

$$\begin{aligned} e_{k+1}^T P_{k+1}^- e_{k+1} &= e_k^T \{P_k^-(I - K_k C_k) - P_k^-(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^-\} e_k \\ &= e_k^T P_k^- e_k - e_k^T \{P_k^- K_k C_k + P_k^-(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^-\} e_k. \end{aligned}$$

Now if we let $V(k, e_k) = e_k^T P_k^- e_k$ then V satisfies the conditions given in Lemma 2.3. Moreover, noting $P_k^- K_k C_k = C_k^T (C_k Q_k^- C_k^T + RR^T)^{-1} C_k$,

$$\begin{aligned} \Delta V(k, e_k) &= e_{k+1}^T P_{k+1}^- e_{k+1} - e_k^T P_k^- e_k \\ &= -e_k^T \{C_k^T (C_k Q_k^- C_k^T + RR^T)^{-1} C_k + P_k^-(Q_k^{-1} \\ &\quad + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^-\} e_k \\ &\leq -e_k^T P_k^-(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^- e_k. \end{aligned}$$

Since

$$\|Q_k^{-1} + A_k^T (NN^T)^{-1} A_k\| \leq \|Q_k^{-1}\| + \|N^{-1} A_k\|^2 \leq p + \|N^{-1}\|^2 \|A\|^2 =: r,$$

$$e_k^T P_k^-(Q_k^{-1} + A_k^T (NN^T)^{-1} A_k)^{-1} P_k^- e_k \geq \frac{1}{r} |P_k^- e_k|^2.$$

If we use $|P_k^- e_k| \geq \frac{1}{q} |e_k|$

$$\Delta V(k, e_k) \leq -\frac{1}{rq^2} |e_k|^2 \leq -\frac{1}{rq^2 p_1} V(k, e_k),$$

where we used the bounds given in Assumption 2.1 and $p_1 := p + \|R^{-1}\|^2 \|C\|^2$. Therefore by Lemma 2.3, e_k converges to zero asymptotically.

3. General Nonlinear Systems

In this Section the results for linear systems are extended to general nonlinear systems of the form:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k), \quad x_0 \text{ given}, \\ y_k &= h(x_k, u_k), \end{aligned}$$

where f and h are at least twice continuously differentiable. For simplicity of notation[1], we consider a system without controls:

$$\begin{aligned} x_{k+1} &= f(x_k), \quad x_0 \text{ given}, \\ y_k &= h(x_k), \end{aligned} \tag{3.1}$$

and its associated "*noisy*" system:

$$\begin{aligned} z_{k+1} &= f(z_k) + Nw_k, \\ \xi_k &= h(z_k) + Rv_k. \end{aligned} \tag{3.2}$$

The extended Kalman filter for the associated system is given as follows[3].

Measurement update:

$$\begin{aligned} \hat{x}_k &= \hat{x}_k^- + K_k(\xi_k - h(\hat{x}_k^-)), \\ Q_k^{-1} &= (Q_k^-)^{-1} + C_k^T(RR^T)^{-1}C_k, \end{aligned} \tag{3.3}$$

Time update:

$$\begin{aligned} \hat{x}_{k+1}^- &= f(\hat{x}_k), \\ Q_{k+1}^- &= A_k Q_k A_k^T + NN^T, \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} K_k &= Q_k^- C_k^T(C_k Q_k^- C_k^T + RR^T)^{-1}, \\ A_k &:= \frac{\partial f}{\partial x}(\hat{x}_k), \\ C_k &:= \frac{\partial h}{\partial x}(\hat{x}_k^-). \end{aligned}$$

The Riccati equations for the error covariances are given as follows

$$Q_{k+1}^- = A_k[(Q_k^-)^{-1} + H_k^T H_k]^{-1}A_k^T + NN^T, \tag{3.5}$$

$$Q_{k+1}^{-1} = [A_k Q_k A_k^T + NN^T]^{-1} + H_{k+1}^T H_{k+1}, \tag{3.6}$$

where $H_k = R^{-1}C_k$.

[1]The modifications necessasry to handle inputs are indicated at the end of the Section.

To begin with, we make the following assumptions for setting up the analysis; Section 4 addresses how Assumption 3.1.1 is implied by an observability property of (3.1).

Assumption 3.1

1. *The error covariances of the extended Kalman filter (3.3) and (3.4) are uniformly bounded, i.e., there exist* $q < \infty$ *and* $p_1 < \infty$ *such that, for all* $k \geq 0$, $\|Q_k^-\| \leq q$ *and* $\|Q_k^{-1}\| \leq p_1$.

2. $A(x) := \frac{\partial f}{\partial x}(x)$ *is invertible at each* $x \in \boldsymbol{R}^n$, *and* $\|A\| := \sup_{x \in R^n} \|A(x)\|$ *and* $\|A^{-1}\| := \sup_{x \in R^n} \|A^{-1}(x)\|$ *are bounded.*

3. $\|H\| := \sup_{x \in R^n} \|R^{-1} \dfrac{\partial h}{\partial x}(x)\|$ *is bounded.*

4. *Let* $g(x,y) := h(x) - h(y) - \frac{\partial h}{\partial x}(y)(x-y)$, *and suppose that there exists* $g < \infty$ *such that* $|g(x,y)| \leq g\||D^2 h\|| \, |x-y|^2$ *for all* $x, y \in \boldsymbol{R}^n$.

For later use we derive a few more bounds. From (3.3)

$$P_M^- = (Q_M^-)^{-1} = Q_M^{-1} - H_k^T H_k,$$

thus giving

$$\|P_M^-\| \leq \|Q_M^{-1}\| + \|H\|^2 \leq p_1 + \|H\|^2 := p.$$

Also from (3.3)

$$\|Q_M\| \leq \|Q_M^-\| \leq q.$$

Furthermore,

$$\|I - K_M C_M\| = \|Q_M (Q_M^-)^{-1}\| \leq pq$$

and

$$\|K_M\| = \|Q_M C_M^T (RR^T)^{-1}\| \leq q\|H\|\|R^{-1}\|^2 =: \delta.$$

Now to prove convergence, set

$$V(k, e_k) = e_k^T P_k^- e_k, \||D^2 f\|| = \sup_{x \in R^n} \||D^2 f(x)\||, \ \||D^2 h\|| = \sup_{x \in R^n} \||D^2 h(x)\||,$$

and

$$\begin{aligned}
\phi(|e_k|, \|\|D^2 f\||, \|\|D^2 h\||) &= \delta g \|\|D^2 h\|\| \, \|A\| + \frac{1}{2}\|\|D^2 f\|\|(pq + \delta g \|\|D^2 h\|\| \, |e_k|)^2, \\
\varphi(|e_k|, \|\|D^2 f\||, \|\|D^2 h\||) &= -\frac{1}{rq^2} + p|e_k|\phi(|e_k|, \|\|D^2 f\||, \|\|D^2 h\||)\{2pq\|A\| \\
&\quad + \phi(|e_k|, \|\|D^2 f\||, \|\|D^2 h\||) \, |e_k|\}.
\end{aligned}$$

Theorem 3.2 *Consider the system (3.1) and the extended Kalman filter equations (3.3) and (3.4) for the associated system (3.2). Suppose that Assumption 3.1 holds. Then, if* $|e_0|, \|\|D^2 f\||,$ *and*$\|\|D^2 h\||$ *are such that for some* $\gamma > 0$,

$$\varphi(q^{\frac{1}{2}} V^{\frac{1}{2}}(0, e_0), \|\|D^2 f\||, \|\|D^2 h\||) \leq -\gamma$$

the extended Kalman filter for the noisy system (3.2) is a local asymptotic observer for the deterministic system (3.1), as long as the design variables N, R *and* $\bar{Q}_0$ *have been chosen such that* N *has rank* n *and* R *and* $\bar{Q}_0$ *are positive definite.*

Proof: Let $e_k = x_k - \hat{x}_k^-$. Then

$$\begin{aligned}
e_{k+1} &= f(x_k) - f(\hat{x}_k) \\
&= \int_0^1 Df(\hat{x}_k + s\tilde{e}_k) ds \tilde{e}_k
\end{aligned}$$

where

$$\tilde{e}_k = x_k - \hat{x}_k = x_k - \hat{x}_k^- - K_k(h(x_k) - h(\hat{x}_k^-)).$$

Note also

$$\begin{aligned}
\tilde{e}_k &= e_k - K_k(C_k e_k + g_k) \\
&= (I - K_k C_k) e_k - K_k g_k.
\end{aligned}$$

Thus, using the above equation,

$$e_{k+1} = [A_k + \int_0^1 (Df(\hat{x}_k + s\tilde{e}_k) - Df(\hat{x}_k)) ds] \tilde{e}_k$$

$$
\begin{aligned}
&= [A_k + \int_0^1\int_0^1 D^2 f(\hat{x}_k + rs\tilde{e}_k)s\tilde{e}_k dr ds]\tilde{e}_k \\
&= [A_k + B_k]\tilde{e}_k \\
&= A_k[(I - K_k C_k)e_k - K_k g_k] + B_k\tilde{e}_k \\
&= A_k(I - K_k C_k)e_k + l_k,
\end{aligned}
$$

where

$$
\begin{aligned}
B_k &= \int_0^1\int_0^1 D^2 f(\hat{x}_k + rs\tilde{e}_k)s\tilde{e}_k dr ds \\
l_k &= -A_k K_k g_k + B_k\tilde{e}_k.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
e_{k+1}^T P_{k+1}^- e_{k+1} &= (e_k^T(I - K_k C_k)^T A_k^T + l_k^T)P_{k+1}^-(A_k(I - K_k C_k)e_k + l_k) \\
&= e_k^T(I - K_k C_k)^T A_k^T P_{k+1}^- A_k(I - K_k C_k)e_k + l_k^T P_{k+1}^- A_k \\
&\quad \times (I - K_k C_k)e_k + e_k^T(I - K_k C_k)^T A_k^T P_{k+1}^- l_k + l_k^T P_{k+1}^- l_k.
\end{aligned}
$$

Using the linear results,

$$
\begin{aligned}
\Delta V(k, e_k) &= e_{k+1}^T P_{k+1}^- e_{k+1} - e_k^T P_k^- e_k \\
&\leq -e_k^T P_k^-(Q_k^{-1} + A^T(NN^T)^{-1}A)^{-1}P_k^- e_k + l_k^T P_{k+1}^- A_k(I - K_k C_k)e_k \\
&\quad + e_k^T(I - K_k C_k)^T A_k^T P_{k+1}^- l_k + l_k^T P_{k+1}^- l_k.
\end{aligned}
$$

With the definition of $g_k = g(x_k, \hat{x}_k^-)$, since

$$
\begin{aligned}
|\tilde{e}_k| &= |(I - K_k C_k)e_k - K_k g_k| \\
&\leq \|I - K_k C_k\|\ |e_k| + \|K_k\|\ |g_k| \\
&\leq (pq + \delta g \||D^2 h|\|\ |e_k|)|e_k|,
\end{aligned}
$$

and

$$
\begin{aligned}
\|B_k\| &= \|\int_0^1\int_0^1 D^2 f(\hat{x}_k + rs\tilde{e}_k)s\tilde{e}_k dr ds\| \\
&\leq \int_0^1\int_0^1 \||D^2 f|\|s dr ds|\tilde{e}_k| = \frac{1}{2}\||D^2 f|\|\ |\tilde{e}_k|,
\end{aligned}
$$

it follows that

$$\begin{aligned} |l_k| &= |-A_k K_k g_k + B_k \tilde{e}_k| \\ &\leq \phi(|e_k|, \|\|D^2 f\|\|, \|\|D^2 h\|\|)|e_k|^2 \end{aligned}$$

and

$$\begin{aligned} & l_k^T P_{k+1}^- A_k (I - K_k C_k) e_k + e_k^T (I - K_k C_k)^T A_k^T P_{k+1}^- l_k + l_k^T P_{k+1}^- l_k \\ & \quad \leq \|P_{l+1}^-\| \, |l_k| (2\|A\| \, \|I - K_k C_k\| \, |e_k| + |l_k|) \\ & \quad \leq p|e_k|^3 \phi(|e_k|, \|\|D^2 f\|\|, \|\|D^2 h\|\|)\{2pq\|A\| + \phi(|e_k|, \|\|D^2 f\|\|, \|\|D^2 h\|\|)|e_k|\}. \end{aligned}$$

Therefore,

$$\Delta V(k, e_k) \leq \varphi(|e_k|, \|\|D^2 f\|\|, \|\|D^2 h\|\|)|e_k|^2. \tag{3.7}$$

A simple argument shows that if $\varphi(q^{\frac{1}{2}} V^{\frac{1}{2}}(0, e_0), \|\|D^2 f\|\|, \|\|D^2 h\|\|) \leq -\gamma$ then $\Delta V(k, e_k) \leq -\gamma |e_k|^2$ for all $k \geq 0$. Thus e_k converges to zero asymptotically by Lemma 2.3.

Remark 3.3

(a) If the observation map is linear, i.e., $h(x) = Cx$, then $D^2 h \equiv 0$. It follows that $\varphi(|e_k|, \|\|D^2 f\|\|, \|\|D^2 h\|\|) = -\frac{1}{rq^2} + \frac{p^4 q^3}{2}|e_k| \cdot \|\|D^2 f\|\|(2\|A\| + \frac{pq}{2}|e_k| \cdot \|\|D^2 f\|\|)$. Thus if we let ζ^+ be the real positive solution of the equation $-\frac{1}{rq^2} + \frac{p^4 q^3}{2}\zeta(2\|A\| + \frac{pq}{2}\zeta) = -\gamma, 0 < \gamma < \frac{1}{rq^2}$, then ζ^+ is a function of the design variables $N, R, \bar{Q}_0, \gamma$. Therefore, under Assumption 3.1, if

$$|e_0| \cdot \|\|D^2 f\|\| \leq \max_{N,R,\bar{Q}_0,\gamma} \frac{\zeta^+}{(pq)^{1/2}}, \tag{3.8}$$

the extended Kalman filter (3.3) and (3.4) with $\xi_k = y_k$ is a local asymptotic observer for the deterministic system (3.1) with linear observations. We note that the condition (3.8) can be satisfied if either $|e_0|$ or $\|\|D^2 f\|\|$ is small enough, in other words, if either the estimate of the initial state is close enough to the true value or f is only weakly nonlinear.

(b) If we know the controls we can construct in the same way a local asymptotic observer for systems with inputs:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k), \quad x_0 \text{ given}, \\ y_k &= h(x_k, u_k), \end{aligned} \tag{3.9}$$

using the extended Kalman filter for the associated "*noisy*" system:

$$\begin{aligned} z_{k+1} &= f(z_k, u_k) + Nw_k, \\ \xi_k &= h(z_k, u_k) + Rv_k. \end{aligned} \tag{3.10}$$

The extended Kalman filter equations and the Riccati equations for the covariances of the associated system (3.10) are the same as (3.3), (3.4),(3.5),(3.6) with $f(\hat{x}_k), h(\hat{x}_k^-)$ replaced by $f(\hat{x}_k, u_k), h(\hat{x}_k^-, u_k)$. For known u, let $f^u(x) := f(x, u)$ and $h^u(x) := h(x, u)$. Now suppose that Assumption 3.1 holds with the following bounds:

$$\|A\| := sup\{\|\frac{\partial f}{\partial x}(x, u)\| : x \in \boldsymbol{R}^n, u \in \boldsymbol{R}^m\},$$

$$\|A^{-1}\| := sup\{\|[\frac{\partial f}{\partial x}(x, u)]^{-1}\| : x \in \boldsymbol{R}^n, u \in \boldsymbol{R}^m\},$$

$$\|H\| := sup\{\|R^{-1}\frac{\partial h}{\partial x}(x, u)\| : x \in \boldsymbol{R}^n, u \in \boldsymbol{R}^m\}.$$

Then Theorem 3.2 holds with the appropriate replacements.

4. Observability conditions of a nonlinear system and its linearization

In this Section we discuss the observability condition in relation to the EKF. First, consider the system (2.10). If we use $R = I$, the observability condition (2.12) becomes

$$\beta_1 I \leq \sum_{k-M}^{k} \Phi^T(i, k) C_k^T C_k \Phi(i, k) \leq \beta_2 I, \quad 0 < \beta_1, \beta_2 < \infty. \tag{4.1}$$

If we assume further that $A_k^T A_k \geq \nu I > 0 \ \forall k$, then condition (4.1) is equivalent to the following condition, for some $\gamma_1, \gamma_2,\ 0 < \gamma_1 \leq \gamma_2 < \infty$,

$$\gamma_1 I \leq O^T(k - M, k) O(k - M, k) \leq \gamma_2 I, \tag{4.2}$$

where

$$O(k - M, k) := \begin{bmatrix} C_{k-M} \\ C_{k-M+1} A_{k-M} \\ \vdots \\ C_k A_{k-1} \cdots A_{k-M} \end{bmatrix}.$$

In order to apply this linear observability condition to the EKF (3.3) and (3.4) and, ultimately, to relate this to observability properties of the underlying nonlinear system, let's represent $O(k-M,k)$ in terms of the EKF variables, i.e.,

$$\begin{aligned} O_e(k-M,k) &:= \begin{bmatrix} C(\hat{x}_{k-M}^-) \\ C(\hat{x}_{k-M+1}^-)A(\hat{x}_{k-M}) \\ \vdots \\ C(\hat{x}_k^-)A(\hat{x}_{k-1})\cdots A(\hat{x}_{k-M}) \end{bmatrix} \\ &=: O_e(\hat{x}_{k-M}^-, \hat{x}_{k-M}, \cdots, \hat{x}_{k-1}, \hat{x}_k^-). \end{aligned} \tag{4.3}$$

Define the map $H : \boldsymbol{R}^n \to (\boldsymbol{R}^p)^n$ by

$$H(x) := (h(x), h(f(x)), \cdots, h(f^{n-1}(x))) \tag{4.4}$$

The system is said to satisfy the *observability rank condition* at x_0[15] if the rank[2] of the map H at x_0 equals n; The system satisfies the *observability rank condition on* $\mathcal{O}$ if this is true for every $x \in \mathcal{O}$; if $\mathcal{O} = \boldsymbol{R}^n$, then one says that the system satifies the obsevability rank condition. By the chain rule,

$$\begin{aligned} \frac{\partial H}{\partial x}(x_0) &= \begin{bmatrix} \frac{\partial h}{\partial x}(x_0) \\ \frac{\partial h}{\partial x}(x_1)\frac{\partial f}{\partial x}(x_0) \\ \vdots \\ \frac{\partial h}{\partial x}(x_{n-1})\frac{\partial f}{\partial x}(x_{n-2})\cdots\frac{\partial f}{\partial x}(x_0) \end{bmatrix} \\ &=: \frac{\partial H}{\partial x}(x_0, x_1, \cdots, x_{n-1}) \end{aligned} \tag{4.5}$$

where $x_{k+1} = f(x_k),\ k = 0, 1, \cdots, n-2$. It follows that rank $\mathcal{O}_e = rank\ \frac{\partial H}{\partial x}$ if $\hat{x}_k^-$ and $\hat{x}_k$ are equal to the true state x_k, for $k = 0, 1, \cdots, n-1$. By continuity we can argue that if the system (3.1) satisfies the observability rank condition, then its associated EKF satisfies the observability condition (4.2), for $M = n-1$, whenever the estimates $\hat{x}_k^-$ and $\hat{x}_k$ are "sufficiently" close to the true state x_k. The boundedness of the error covariances would

[2]Recall that the rank of H at x_0 equals the rank of $\frac{\partial H}{\partial x}(x)$ evaluated at x_0.

then follow from Deyst and Price, [11]. This line of reasoning is made precise in Proposition 4.1 below and in Section 5.

Proposition 4.1 *Suppose that the system (3.1) satisfies the observability rank condition on a compact subset $K \subset \mathbb{R}^n$. Then there exist $\gamma_1, \gamma_2, 0 < \gamma_1 \leq \gamma_2 < \infty$ and $\delta_1 > 0$ such that*

$$\gamma_1 I \leq \frac{\partial H}{\partial x}(\hat{x}_0, \cdots, \hat{x}_{n-1})^T \frac{\partial H}{\partial x}(\hat{x}_0, \cdots, \hat{x}_{n-1}) \leq \gamma_2 I \tag{4.6}$$

for all $\hat{x}_l$ such that $|\hat{x}_l - x_l| \leq \delta_1, l = 0, \cdots, n-1$, and for each $x_0 \in K$.

Proof: By the observability rank condition,

$$\frac{\partial H}{\partial x}(x_0, x_1, \cdots, x_{n-1})^T \frac{\partial H}{\partial x}(x_0, x_1, \cdots, x_{n-1}) > 0$$

for all $x_0 \in K$. Since $\frac{\partial H}{\partial x}$ is continuous and K is compact, there exist $\beta_1 > 0, \beta_2 > 0$ such that, for all $x_0 \in K$,

$$\beta_1 I \leq \frac{\partial H}{\partial x}(x_0, x_1, \cdots, x_{n-1})^T \frac{\partial H}{\partial x}(x_0, x_1, \cdots, x_{n-1}) \leq \beta_2 I.$$

Then, once again invoking continuity and compactness, there exist $\gamma_1, \gamma_2, 0 < \gamma_1 \leq \beta_1 \leq \beta_2 \leq \gamma_2 < \infty$ and $\delta_1 > 0$ such that (4.6) holds.

Remark 4.2 If one assumes that $[\frac{\partial f}{\partial x}(x_0)]^{-1}$ exists for each $x_0 \in K$, $K \subset \mathbb{R}^n$ compact, then, as long as $\frac{\partial f}{\partial x}$ is continuous, it follows that there exist $\nu_1, \nu_2, 0 < \nu_1 \leq \nu_2 < \infty$, such that

$$\nu_1 I \leq \frac{\partial f}{\partial x}(x_0)^T \frac{\partial f}{\partial x}(x_0) \leq \nu_2 I.$$

Recall that this is important for linking (4.1) and (4.2).

Remark 4.3 Suppose that the system (3.1) satisfies the observability rank condition and that the output y is scalar valued. Then $\bar{x} = H(x)$ is a local diffeomorphism about the origin. In the $\bar{x}$- coordinates, (3.1) is transformed into a local, observer canonical form:

$$\begin{aligned} \bar{x}_1(k+1) &= \bar{x}_2(k) \\ &\vdots \\ \bar{x}_{n-1}(k+1) &= \bar{x}_n(k) \\ \bar{x}_n(k+1) &= \phi(\bar{x}_1(k), \cdots, \bar{x}_n(k)) \\ y &= \bar{x}_1(k). \end{aligned} \tag{4.7}$$

A simple computation shows that the linearized observability condition (4.2) is always satisfied for a system in the form (4.7); indeed, $O(k-M,k) \equiv I_n$ for $M = n-1$. This is a marked contrast to the situation analyzed in Proposition 4.1, and underlines the coordinate dependence of the Kalman filter in general, and the linearized observability condition (4.2) in particular.

5. Applicability of EKF as an observer for nonlinear systems

In this Section we seek to remove the boundedness assumption on the error covariances that was used in Section 3. By applying the EKF on a convex compact subset of the state space, this can be done. Before we begin, a few notations are mentioned. Let $\mathcal{O}$ be a convex compact subset of $\boldsymbol{R}^n$, $\sim \mathcal{O}$ the complement of $\mathcal{O}$, and $\epsilon > 0$ be a positive constant. Define $d(x, \sim \mathcal{O}) = inf\{|x-y| : y \in \sim \mathcal{O}\}$, and $\mathcal{O}_\epsilon = \{x \in \mathcal{O} : d(x, \sim \mathcal{O}) \geq \epsilon\}$. Since $\mathcal{O}$ is compact, $\|A\| := \sup_{x \in \mathcal{O}} \|\frac{\partial f}{\partial x}(x)\|$ and $\|Dh\| := \sup_{x \in \mathcal{O}} \|\frac{\partial h}{\partial x}(x)\|$ are bounded. Let $a = max(1, \|A\|)$ and

$$\begin{aligned} b_k &= (1 + \|\bar{Q}_0\|\|Dh\|^2\|R^{-1}\|^2)a^k \prod_{l=1}^{k} \{1 + \|Dh\|^2\|R^{-1}\|^2 \\ &\quad \times [\|A\|^{2l}\|\bar{Q}_0\| + \|N\|^2(\|A\|^{2(l-1)} + \|A\|^{2(l-2)} + \cdots + 1)]\}. \end{aligned}$$

First we consider a sufficient condition for keeping the estimates $\hat{x}_k^-$ and $\hat{x}_k$ near the true state x_k.

Theorem 5.1 *Consider the system (3.1) and its associated EKF (3.3) and (3.4). Suppose that the following conditions hold.*

1. $x_k \in \mathcal{O}_\epsilon$, *for some* $\epsilon > 0, \quad 0 \leq k \leq M$.

2. $|e_0| = |\hat{x}_0^- - x_0| \leq \frac{\delta}{b_M}$ *for some* $0 < \delta \leq \epsilon/2$.

Then for $k = 0, 1, \cdots, M$,

$$|\hat{x}_k^- - x_k| \leq \delta \quad \text{and} \quad |\hat{x}_k - x_k| \leq \delta.$$

Proof: We show the closeness by induction. First, by assumption, $|\hat{x}_0^- - x_0| \leq \delta$, thus $\hat{x}_0^- \in \mathcal{O}_{\epsilon/2}$. Now

$$\begin{aligned} |\hat{x}_0 - x_0| &= |\hat{x}_0^- + K_0(h(x_0) - h(\hat{x}_0^-)) - x_0| \\ &\leq |e_0| + \|K_0\| \cdot |\int_0^1 Dh(\hat{x}_0^- + s(x_0 - \hat{x}_0^-))ds\ e_0| \\ &\leq (1 + \|K_0\|\ \|Dh\|)|e_0|, \end{aligned}$$

where we used the fact that, by convexity, $\hat{x}_0^- + s(x_0 - \hat{x}_0^-) \in \mathcal{O}$ for $0 \leq s \leq 1$. Since $C_0 = \frac{\partial h}{\partial x}(\hat{x}_0^-)$, $\|K_0\| \leq \|\bar{Q}_0\|\ \|Dh\|\ \|R^{-1}\|^2$. Thus

$$|\hat{x}_0 - x_0| \leq (1 + \|\bar{Q}_0\|\ \|Dh\|^2\|R^{-1}\|^2)|e_0| \leq \delta.$$

For $k = 1$,

$$\begin{aligned} |\hat{x}_1^- - x_1| &= |f(\hat{x}_0) - f(x_0)| \\ &= |\int_0^1 Df(x_0 + s(\hat{x}_0 - x_0))ds(\hat{x}_0 - x_0)| \\ &\leq \|A\|\ |\hat{x}_0 - x_0| \leq \|A\|(1 + \|\bar{Q}_0\|\ \|Dh\|^2\|R^{-1}\|^2)|e_0| \leq \delta. \end{aligned}$$

In the same way as for $k = 0$,

$$|\hat{x}_1 - x_1| \leq (1 + \|K_1\|\ \|Dh\|)|e_1|.$$

Using the fact that $\|K_1\| \leq (\|A\|^2\|\bar{Q}_0\| + \|N\|^2)\|Dh\|\ \|R^{-1}\|^2$,

$$|\hat{x}_1 - x_1| \leq b_1|e_0| \leq \delta.$$

Now suppose that $|\hat{x}_l^- - x_l| \leq \delta$, and $|\hat{x}_l - x_l| \leq \delta$ for $0 \leq l \leq k-1$. Then

$$\begin{aligned} |\hat{x}_k^- - x_k| &= |f(\hat{x}_{k-1}) - f(x_{k-1})| \\ &\leq \|A\|\ |\hat{x}_{k-1} - x_{k-1}| \\ &\leq \|A\|(1 + \|K_{k-1}\|\ \|Dh\|)|\hat{x}_{k-1}^- - x_{k-1}| \\ &\leq \|A\|(1 + \|K_{k-1}\|\ \|Dh\|)\|A\|(1 + \|K_{k-2}\|\ \|Dh\|) \\ &\quad \cdots \|A\|(1 + \|K_0\|\ \|Dh\|)|e_0|. \end{aligned}$$

Note also that for $1 \leq l \leq k-1$,

$$\begin{aligned}
\|Q_l\| &= \|[(Q_l^-)^{-1} + C_l^T (RR^T)^{-1} C_l]^{-1}\| \leq \|Q_l^-\|, \\
\|Q_l^-\| &= \|A_{l-1} Q_{l-1} A_{l-1}^T + NN^T\| \\
&\leq \|A\|^2 \|Q_{l-1}^-\| + \|N\|^2 \\
&\leq \|A\|^{2l} \|\bar{Q}_0\| + \|N\|^2 (\|A\|^{2(l-1)} + \|A\|^{2(l-2)} + \cdots + 1), \\
\|K_l\| &\leq \|Q_l^-\| \, \|Dh\| \, \|R^{-1}\|^2.
\end{aligned}$$

Therefore, for $2 \leq k \leq M$,

$$|\hat{x}_k^- - x_k| \leq b_{k-1} |e_0| \leq \delta.$$

Also,

$$\begin{aligned}
|\hat{x}_k - x_k| &\leq (1 + \|K_k\| \cdot \|Dh\|) |\hat{x}_k^- - x_k| \\
&\leq b_k |e_0| \leq \delta.
\end{aligned}$$

This completes the proof.

Since we have conditions which keep the EKF estimates close to the true state, we can now use the results of Theorem 3.2, Proposition 4.1, and Theorem 5.1 to show the convergence of the EKF on a convex compact set without Assumption 3.1.

Note that on a compact set $\mathcal{O} \subset \boldsymbol{R}^n$, $\||D^2 f\|| := \sup_{x \in \mathcal{O}} \|| \dfrac{\partial^2 f}{\partial x^2}(x) \||$ and $\||D^2 h\|| := \sup_{x \in \mathcal{O}} \|| \dfrac{\partial^2 h}{\partial x^2}(x) \||$ are bounded, and Assumption 3.1.4 holds for all $x, y \in \mathcal{O}$. Let $\alpha_1 = \|N\|^2 (1 + \|A\|^2 + \|A\|^4 + \cdots + \|A\|^{2(n-2)})$, $\alpha_2 =$ minimum eigenvalue of NN^T, $a = max(1, \|A\|)$, and

$$\begin{aligned}
\beta_k &= (1 + \|\bar{Q}_0\| \, \|Dh\|^2) a^k \prod_{l=1}^{k} \{1 + \|Dh\|^2 \\
&\quad \times [\|A\|^{2l} \|\bar{Q}_0\| + \|N\|^2 (\|A\|^{2(l-1)} + \|A\|^{2(l-2)} + \cdots + 1)]\}.
\end{aligned}$$

Theorem 5.2 *Suppose that the system (3.1) satisfies the observability rank condition on a convex compact set $\mathcal{O}$. Let $\delta_1 > 0$ be a constant which satisfies the inequality (4.6) for some*

$0 < \gamma_1 \leq \gamma_2$. *Let* $p = (\gamma_2 + 1/\alpha_2), \quad q = a^2(\alpha_1 + 1/\gamma_1) + \|N\|^2$. *Let* $\delta_2 > 0$ *be such that* $\varphi((pq)^{1/2}\delta_2, \||D^2 f\||, \||D^2 h\||) \leq -\gamma$ *for some* $\gamma > 0$, *where* φ *is defined in Section 3,* M *be the smallest integer which satisfies*

$$[1 + (q\|A\|^2 + \|N\|^2)\|Dh\|^2]\|A\|(1 + q\|Dh\|^2)(1 - \frac{\gamma}{p})^{M/2}(pq)^{1/2} < 1,$$

and $\delta = min(\epsilon/2, \delta_1, \delta_2)$ *for some* $\epsilon > 0$. *Suppose further that* $|e_0| \leq \frac{\delta}{\beta_{n+M-1}}$. *Then we have the following results:*

1. $|\hat{x}_k^- - x_k| \leq \delta$ *and* $|\hat{x}_k - x_k| \leq \delta \quad \forall k \geq 0$.

2. *The linearized system around* $\hat{x}_k^-$ *and* $\hat{x}_k$, *i.e.,* $z_{k+1} = \frac{\partial f}{\partial x}(\hat{x}_k)z_k, \quad y_k = \frac{\partial h}{\partial x}(\hat{x}_k^-)z_k$, *satisfies the observability condition (4.2) for* $k \geq n-1$. *Thus there exist* $q < \infty, p < \infty$ *such that* $\|Q_l\| \leq q$, *and* $\|Q_k^-\| \leq p \quad \forall k \geq n-1$.

3. *The error is bounded by* δ *and after time step* $n-1$, *converges to zero, i.e., for* $k \leq n-1$, $|e_k| \leq \delta$, *and for* $k > n-1$, $|e_k| \leq min(\delta, (1 - \frac{\gamma}{p})^{(k-n+1)/2}(pq)^{1/2}\delta)$.

Proof: Since the assumptions satisfy the sufficient condition which bounds $\hat{x}_k^-$ and $\hat{x}_k$ near x_k for $k = 0, \cdots, n+M-1$, it follows that for $k = 0, \cdots, n+M-1$,

$$|\hat{x}_k^- - x_k| \leq \delta \leq \epsilon/2 \quad \text{and} \quad |\hat{x}_k - x_k| \leq \delta \leq \epsilon/2.$$

Therefore, the EKF (3.3), (3.4) satisfies the observability condition (4.2) with $R = I$; i.e., for $n-1 \leq k \leq n+M-1$,

$$\gamma_1 I \leq \sum_{i=k-n+1}^{k-1} \Phi^T(i, k-n+1)C_i^T C_i \Phi(i, k-n+1) \leq \gamma_2 I. \tag{5.1}$$

Since N is nonsingular and $\mathcal{O}$ is compact, it follows clearly that for $n-1 \leq k \leq n+M-1$,

$$\alpha_1 I \geq \sum_{i=k-M}^{k-1} \Phi(k, i+1)NN^T\Phi^T(k, i+1) \geq \alpha_2 I. \tag{5.2}$$

Hence by Deyst and Price[11], for $n-1 \leq k \leq n+M-1$,

$$\|Q_k\| \leq \alpha_1 + 1/\gamma_1 \quad \text{and} \quad \|Q_k^{-1}\| \leq p,$$

thereby giving the following bounds for $n-1 \le k \le n+M-1$,

$$1/p \le \|Q_k\| \le \|Q_k^-\| \le q \quad \text{and} \quad 1/q \le \|(Q_k^-)^{-1}\| \le \|Q_k^{-1}\| \le p.$$

Using $|e_{n-1}| \le \delta$, we have

$$\varphi((pq)^{1/2}|e_{n-1}|, \||D^2 f\||, \||D^2 h\||) \le -\gamma.$$

Accordingly, though we apply EKF from $k=0$, we have the convergence results only after $k=n-1$, i.e.,

$$|e_l| \le (pq)^{1/2}(1-\frac{\gamma}{p})^{(l-n+1)/2}|e_{n-1}|, \quad l \ge n-1.$$

Now we show the remaining part by induction. That is,

$$\begin{aligned}
|\hat{x}_{n+M}^- - x_{n+M}| &\le \|A\|\,|\hat{x}_{n+M-1} - x_{n+M-1}| \\
&\le \|A\|(1+\|K_{n+M-1}\|\,\|Dh\|)|e_{n+M-1}| \\
&\le \|A\|(1+q\|Dh\|^2)(pq)^{1/2}(1-\frac{\gamma}{p})^{M/2}|e_{n-1}| \le \delta.
\end{aligned}$$

Note that $R=I$ is used as a design variable. Also,

$$\begin{aligned}
|\hat{x}_{n+M} - x_{n+M}| &\le (1+\|K_{n+M}\|\,\|Dh\|)|e_{n+M}| \\
&\le [1+(q\|A\|^2+\|N\|^2)\|Dh\|^2]\|A\|(1+q\|Dh\|^2) \\
&\quad \times(1-\frac{\gamma}{p})^{M/2}(pq)^{1/2}|e_{n-1}| \le \delta.
\end{aligned}$$

In addition, we have

$$\hat{x}_{n+M}^- \in \mathcal{O}_{\epsilon/2} \quad \text{and} \quad \hat{x}_{n+M} \in \mathcal{O}_{\epsilon/2}.$$

Thus the conditions (5.1) and (5.2) are also met for $k=n+M$. Hence $\|Q_{n+M}\| \le \alpha_1 + 1/\gamma_1$ and $\|Q_{n+M}^{-1}\| \le p$. Therefore by induction it can be shown that for $k \ge n+M$,

1. $|\hat{x}_k^- - x_k| \le \delta \le \epsilon/2, \quad |\hat{x}_k - x_k| \le \delta \le \epsilon/2.$
2. $\|Q_k^-\| \le q, \quad \|Q_k^{-1}\| \le p.$
3. $|e_k| \le \delta(pq)^{1/2}(1-\frac{\gamma}{p})^{(k-n+1)/2}.$

Remark 5.3

(a) In order to satisfy the observability condition, it was necessary to keep the estimates $\hat{x}_k^-$ and $\hat{x}_k$ near x_k for $0 \leq k \leq n-1$, thus requiring a very good initial guess.

(b) We also needed to have an initializing period ($n-1 \leq k \leq n+M-1$) for the EKF in order to build up the observability condition; after this, the recursions proceeded automatically.

6. Conclusion

We have analyzed in detail how the EKF works when it is applied to a deterministic nonlinear system for the purpose of observation. With *a priori* bounded error covariances, it can be shown that the EKF works as a quasi-local observer[14]. To obtain the convergence, it is generally necessary either to have a very good initial guess or to have a weak nonlinearity in the sense that $\|\|D^2 f\|\|$ and $\|\|D^2 h\|\|$ should be sufficiently small. This part of the analysis was rather standard and followed the work of [1]. In order to *estabilish* the boundedness of the error covariances in the EKF, an observability condition must be imposed on the linearization of the nonlinear system along the estimated trajectory. Conditions under which the observability of the underlying nonlinear system implied that of the linearized system were identified in Section 4. In Section 5, it was then shown how this could be used to prove the boundedness of the error covariances.

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, Oct. 1988, pp. 1147-1158.

[2] V. Lakshmikantham and D. Trigiante, *Theory of Difference Equations : Numerical Methods and Applications,* Boston:Academic, 1988.

[3] B. D. O. Anderson and J. B. Moore, *Optimal Filtering,* Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

[4] John M. Fitts, "On the global observability of nonlinear systems," Ph.D. Dissertation, Univ. of California, Los Angeles, 1970.

[5] Frank L. Lewis, *Optimal Estimation,* New York:John Wiley & Sons, 1986.

[6] J. S. Baras and P. S. Krishnaprasad, "Dynamic observers as asymptotic limits of recursive filters," in *Proc. 21st IEEE Conf. Decision Contr.,* Orlando, FL, Dec. 1982, pp. 1126-1127.

[7] E. A. Misawa and J. K. Hedrick, "Nonlinear observers - a state-of-the-art survey," *ASME Journal of dynamic systems, measurement, and control,* Vol. 111, Sept. 1989, pp. 344-352.

[8] A. E. Bryson, Jr. and Y. C. Ho, *Applied Optimal Control,* New York;Hemisphere, 1975.

[9] O. Hijab, "Asymptotic Baysean estimation of a first order equation with small diffusion," *Annals. Probab.,* 12(1984), pp. 890-902.

[10] Frank L. Lewis, *Optimal Control,* New York:John Wiley & Sons, 1986.

[11] J. J. Deyst, Jr. and C. F. Price," Conditions for asymptotic stability of the discrete minimum-variance linear estimator," *IEEE Trans. Automatic Control,* Vol.13, No.6, Dec. 1968, pp. 702-705.

[12] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems,* Wiley-Interscience, New York, 1972.

[13] M. Aoki, *Optimization of Stochastic Systems,* Academic Press, New York, 1967.

[14] J. W. Grizzle and P. E. Moraal, "Newton, Observers and Nonlinear Discre-time Control," in *Proc. 29th IEEE Conf. Decision Contr.,* Hawaii, Dec. 1990, pp. 760-767.

[15] H. Nijmeijer, " Observability of autonomous discrete time non-linear systems: a geometric approach," INT. J. CONTROL, 1982, VOL. 36, NO. 5, 867-874

Appendix: Error covariance bounds for linear time invariant systems

Motivated by [1], we interpret Q_k and Q_k^{-1} in terms of dual optimal control problems which give the same Riccati equations as (2.8)and (2.9). Let $H = R^{-1}C$ and let $M > 0$ be a fixed integer. Consider

$$\eta_k = A^T\eta_{k+1} + H^T v_{k+1}, \qquad k = 0, \cdots, M-1, \tag{A.1}$$

where η_M is given and v is the control. The cost to minimize is

$$J_1 = \frac{1}{2}\eta_0^T Q_0^- \eta_0 + \frac{1}{2}\sum_{k=0}^{M-1}(\eta_{k+1}^T NN^T\eta_{k+1} + v_{k+1}^T v_{k+1}). \tag{A.2}$$

Then the necessary conditions are given as follows in terms of a two-point boundary-value problem:

$$\begin{aligned} \eta_k &= A^T\eta_{k+1} + H^T v_{K+1}, \quad \eta_M \text{ given}, \\ \lambda_{k+1} &= A\lambda_k + NN^T\eta_{k+1}, \quad \lambda_0 = Q_0^-\eta_0, \end{aligned}$$

and the optimal control is given as

$$v_{k+1} = -H\lambda_k.$$

If we set $\lambda_k = Q_k^-\eta_k$ then by the "*sweep method*"[8] it can be shown that Q_k^- satisfies the Riccati equation (2.8). Moreover, the minimum cost is

$$J_1^* = \frac{1}{2}\eta_M^T Q_M^-\eta_M = \frac{1}{2}\eta_M^T\lambda_M$$

Similarly, set $P_k = Q_k^{-1}$ and consider

$$\lambda_{k+1} = A\lambda_k + Nv_k, \quad k = 0, \cdots, M-1, \tag{A.3}$$

where λ_M is given and v is the control. The cost to minimize is

$$J_2 = \frac{1}{2}\lambda_0^T(P_0 - H^TH)\lambda_0 + \frac{1}{2}\sum_{k=0}^{M-1}(\lambda_k^T H^T H\lambda_k + v_k^T v_k). \tag{A.4}$$

Then the necessary conditions are again given as follows in terms of a two-point boundary-value problem:

$$\begin{aligned} \lambda_{k+1} &= A\lambda_k + Nv_k, \quad \lambda_M \text{ given}, \\ \eta_k &= A^T\eta_{k+1} + H^T H\lambda_{k+1}, \quad \eta_0 = -A^{-T}P_0\lambda_0, \end{aligned}$$

where $A^{-T} = (A^{-1})^T$ and the optimal control is given as

$$v_k = -N^T\eta_k.$$

If we set $\eta_k = -A^{-T}P_k\lambda_k$, then it can also be shown that P_k satisfies the Riccati equaiton (2.9). Moreover, the minimum cost is

$$J_2^* = \frac{1}{2}\lambda_M^T(P_M - H^T H)\lambda_M.$$

Now we show that $\|Q_M^-\|$ and $\|P_M\|$ are bounded for all M. Since R is assumed positive definite and N has rank n ,the pair (H, A) is detectable and the pair (A, N) is controllable.

Theorem A.1 *Consider the system (2.1), the Kalman filter equations (2.3) and (2.4) for the associated system (2.2) and the above two optimal control problems. Suppose that N has rank n and $\bar{Q}_0, R$ are positive definite. Then for any $\Lambda, ,$ chosen such that all the eigenvalues of $(A + \Lambda H)$ and $(A + N,)^{-1}$ are inside the unit disk and nonzero, let*

$$\begin{aligned} \Phi^Q &= \sum_{k=0}^{\infty}(A + \Lambda H)^k\{(A + \Lambda H)^T\}^k, \\ \Phi^P &= \sum_{k=0}^{\infty}\{(A + N,)^{-T}\}^k(A + N,)^{-k}. \end{aligned}$$

Then Φ^Q and Φ^P are well-defined positive definite matrices. Moreover,

$$\|Q_M^-\| \leq \{\|\bar{Q}_0\| \cdot \frac{\lambda_{max}(\Phi^Q)}{\lambda_{min}(\Phi^Q)} + (\|N\|^2 + \|\Lambda\|^2) \cdot \lambda_{max}(\Phi^Q)\} =: q, \tag{A.5}$$

$$\begin{aligned} \|P_M\| \leq\ & \{\|H\|^2 + (\|\bar{Q}_0^{-1}\| + \|H\|^2 + \|, \|^2) \cdot \frac{\lambda_{max}(\Phi^P)}{\lambda_{min}(\Phi^P)} + \\ & (\|H\|^2 + \|, \|^2) \cdot \lambda_{max}(\Phi^P)\} =: p, \end{aligned} \tag{A.6}$$

where $\lambda_{min}(\cdot), \lambda_{max}(\cdot)$ denote the minimum and maximum eigenvalues, respectively.

Proof: First, consider in (A.1) a feedback control

$$v_k = \Lambda^T \eta_k.$$

Then from (A.2)

$$\begin{aligned} 2J_1^* = \eta_M^T Q_M^- \eta_M &\leq \eta_0^T Q_0^- \eta_0 + \sum_{k=0}^{M-1} \eta_{k+1}^T (NN^T + \Lambda\Lambda^T)\eta_{k+1} \\ &\leq \|Q_0^-\||\eta_0|^2 + (\|N\|^2 + \|\Lambda\|^2) \sum_{k=0}^{M-1} |\eta_{k+1}|^2. \end{aligned} \tag{A.7}$$

Now by $\eta_k = (A + \Lambda H)^T \eta_{k+1}$

$$\begin{aligned} \eta_M^T \Phi^Q \eta_M &= \sum_{k=0}^{M-1} (\eta_{k+1}^T \Phi^Q \eta_{k+1} - \eta_k^T \Phi^Q \eta_k) + \eta_0^T \Phi^Q \eta_0 \\ &= \sum_{k=0}^{M-1} \eta_{k+1}^T [\Phi^Q - (A + \Lambda H)\Phi^Q (A + \Lambda H)^T] \eta_{k+1} + \eta_0^T \Phi^Q \eta_0. \end{aligned}$$

Since the pair (H, A) is detectable, we can find Λ such that all the eigenvalues of $(A + \Lambda H)$ are inside the unit disk.Then there exists a unique positive definite matrix Φ^Q that satisfies the Liapunov equation

$$\Phi^Q - (A + \Lambda H)\Phi^Q (A + \Lambda H)^T = I.$$

Indeed,the solution is given as

$$\Phi^Q = \sum_{k=0}^{\infty} (A + \Lambda H)^k \{(A + \Lambda H)^T\}^k.$$

With this Φ^Q,

$$\eta_M^T \Phi^Q \eta_M = \sum_{k=0}^{M-1} |\eta_{k+1}|^2 + \eta_0^T \Phi^Q \eta_0.$$

Therefore,

$$\sum_{k=0}^{M-1} |\eta_{k+1}|^2 \leq \eta_M^T \Phi^Q \eta_M \leq \lambda_{max}(\Phi^Q)|\eta_M|^2 \tag{A.8}$$

and

$$\lambda_{min}(\Phi^Q)|\eta_0|^2 \leq \eta_0^T \Phi^Q \eta_0 \leq \lambda_{max}(\Phi^Q)|\eta_M|^2;$$

thus

$$|\eta_0|^2 \leq \frac{\lambda_{max}(\Phi^Q)}{\lambda_{min}(\Phi^Q)} \cdot |\eta_M|^2. \tag{A.9}$$

Substituting (A.8) and (A.9) into (A.7) gives (A.5).

Similarly, consider in (A.3) a feedback control

$$v_k = , \lambda_k .$$

Then from (A.4),

$$2J_2^* = \lambda_M^T (P_M - H^T H)\lambda_M \leq \lambda_0^T (P_0 - H^T H)\lambda_0 + \sum_{k=0}^{M-1} (\lambda_k^T H^T H \lambda_k + v_k^T v_k)$$

or from (2.3b) noting $P_0 - H^T H = \bar{Q}_0^{-1}$

$$\begin{aligned} \lambda_M^T P_M \lambda_M &\leq \lambda_M^T H^T H \lambda_M + \lambda_0^T (\bar{Q}_0^{-1} + H^T H + , ^T ,)\lambda_0 + \sum_{k=1}^{M-1} \lambda_k^T (H^T H + , ^T ,)\lambda_k \\ &\leq \|H\|^2 |\lambda_M|^2 + (\|\bar{Q}_0^{-1}\| + \|H\|^2 + \|, \|^2)|\lambda_0|^2 \\ &\quad + (\|H\|^2 + \|, \|^2) \sum_{k=1}^{M-1} |\lambda_k|^2. \end{aligned} \tag{A.10}$$

Now $\lambda_k = (A + N,)\lambda_{k-1}$, and thus $\lambda_{k-1} = (A + N,)^{-1}\lambda_k$; therefore

$$\begin{aligned} \lambda_M^T \Phi^P \lambda_M &= \sum_{k=1}^{M} (\lambda_k^T \Phi^P \lambda_k - \lambda_{k-1}^T \Phi^P \lambda_{k-1}) + \lambda_0^T \Phi^P \lambda_0 \\ &= \sum_{k=1}^{M} \lambda_k^T [\Phi^P - (A + N,)^{-T} \Phi^P (A + N,)^{-1}]\lambda_k + \lambda_0^T \Phi^P \lambda_0 \end{aligned}$$

Since all the eigenvalues of $(A + N,)^{-1}$ are inside the unit disk and nonzero, there exists a unique positive definite matrix Φ^P satisfying

$$\Phi^P - (A + N,)^{-T} \Phi^P (A + N,)^{-1} = I.$$

Moreover, the solution is given as

$$\Phi^P = \sum_{k=0}^{\infty} \{(A + N,)^{-T}\}^k (A + N,)^{-k}.$$

Since

$$\lambda_M^T \Phi^P \lambda_M = \sum_{k=1}^{M} |\lambda_k|^2 + \lambda_0^T \Phi^P \lambda_0 \geq \sum_{k=1}^{M-1} |\lambda_k|^2 + \lambda_0^T \Phi^P \lambda_0,$$

it follows that

$$\sum_{k=1}^{M-1} |\lambda_k|^2 \leq \lambda_{max}(\Phi^P)|\lambda_M|^2$$

and

$$|\lambda_0|^2 \leq \frac{\lambda_{max}(\Phi^P)}{\lambda_{min}(\Phi^P)} \cdot |\lambda_M|^2.$$

Finally, with (A.10) this gives (A.6).