The Local Convergence of the Extended Kalman Filter

Arthur J. Krener [1]

July 10, 2004

[1]Research supported in part by NSF DMS-0204390 and AFOSR F49620-01-1-0202.

Chapter 1

The Convergence of the Extended Kalman Filter

Abstract

We demonstrate that the extended Kalman filter converges locally for a broad class of nonlinear systems. If the initial estimation error of the filter is not too large then the error goes to zero exponentially as time goes to infinity. To demonstrate this, we require that the system be C^2 and uniformly observable with bounded second partial derivatives.

1.1 Introduction

The extended Kalman filter is a widely used method for estimating the state $x(t) \in \mathbb{R}^n$ of a partially observed nonlinear dynamical system,

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= h(x,u) \\ x(0) &= x^0 \end{aligned} \tag{1.1.1}$$

from the past controls and observations $u(s) \in U \subset \mathbb{R}^m,\ y(s) \in \mathbb{R}^p, 0 \le s \le t$ and some information about the initial condition x^0. The functions f, h are known and assumed to be C^2.

An extended Kalman filter is derived by replacing (1.1.1) by its linear approximation around the trajectory $\hat{x}(t)$ and adding standard white Gaussian driving noise $w(t) \in \mathbb{R}^l$ and independent, standard white Gaussian observation noise $v(t) \in I\!R^l$,

$$\begin{aligned} \dot{z} &= f(\hat{x}(t), u(t)) + A(t)z + Gw \\ y &= h(\hat{x}(t), u(t)) + C(t)z + v \\ z(0) &= z^0 \end{aligned} \tag{1.1.2}$$

where G is a $n \times l$ matrix chosen by the designer,

$$A(t) = \frac{\partial f}{\partial x}(\hat{x}(t), u(t)), \qquad C(t) = \frac{\partial h}{\partial x}(\hat{x}(t), u(t)), \tag{1.1.3}$$

and z^0 is a Gaussian random vector independent of the noises with mean $\hat{x}^0$ and variance P^0 that are chosen by the designer.

The Kalman filter for (1.1.2) is

$$\begin{aligned} \dot{\hat{z}}(t) &= f(\hat{x}(t), u(t)) + A(t)\hat{z}(t) + P(t)C'\left(y(t) - h(\hat{x}(t), u(t)) - C(t)\hat{z}(t)\right) \\ \dot{P}(t) &= A(t)P(t) + P(t)A'(t) + \Gamma - P(t)C'(t)C(t)P(t) \\ \hat{z}(0) &= \hat{x}^0 \\ P(0) &= P^0 \end{aligned} \tag{1.1.4}$$

where $\Gamma = GG'$.

The extended Kalman filter for (1.1.1) is given by

$$\begin{aligned} \dot{\hat{x}}(t) &= f(\hat{x}(t), u(t)) + P(t)C'(t)\left(y(t) - h(\hat{x}(t), u(t))\right) \\ \dot{P}(t) &= A(t)P(t) + P(t)A'(t) + \Gamma - P(t)C'(t)C(t)P(t) \\ \hat{x}(0) &= \hat{x}^0 \\ P(0) &= P^0. \end{aligned} \tag{1.1.5}$$

Actually there are many extended Kalman filters for (1.1.1), depending on the choice of the design parameters $G,\ \hat{x}^0,\ P^0$. We could also broaden the class of extended Kalman filters for (1.1.1) by allowing $G = G(\hat{x}(t))$ and putting a similar coefficient in front of the observation noise in (1.1.2). We chose not to do so to simplify the discussion. For similar reasons we omit the discussion of time varying systems. We expect that our main theorem can be generalized to cover such systems. For more on the derivation of the extended Kalman filter, see Gelb [3].

Baras, Bensoussan and James [1] have shown that under suitable conditions, the extended Kalman filter converges locally, i.e., if the initial error $\tilde{x}(0) = x(0) - \hat{x}(0)$ is sufficiently small then $\tilde{x}(t) = x(t) - \hat{x}(t) \to 0$ as $t \to \infty$. Unfortunately their conditions are difficult to verify and may not be satisfied even by an observable linear system. Krener and Duarte have given a simple example where any extended Kalman filter fails to converge. More on these points later.

By modifying the techniques of [1] and incorporating techniques of the high gain observer of Gauthier, Hammouri and Othman [2] we shall show that under verifiable conditions that the extended Kalman filter converges locally. To state the main result we need a definition.

Definition 1.1.1 *[2] The system*

$$\begin{aligned} \dot{\xi} &= f(\xi, u) \\ y &= h(\xi, u) \end{aligned} \tag{1.1.6}$$

is uniformly observable for any input if there exist coordinates

$$\{x_{ij} : i = 1, \ldots, p,\ j = 1, \ldots, l_i\}$$

where $1 \leq l_1 \leq \ldots \leq l_p$ *and* $\sum l_i = n$ *such that in these coordinates the system takes the form*

$$\begin{aligned} y_i &= x_{i1} + h_i(u) \\ \dot{x}_{i1} &= x_{i2} + f_{i1}(\underline{x}_1, u) \\ &\vdots \\ \dot{x}_{ij} &= x_{ij+1} + f_{ij}(\underline{x}_j, u) \\ &\vdots \\ \dot{x}_{il_i-1} &= x_{il_i} + f_{il_i-1}(\underline{x}_{l_i-1}, u) \\ \dot{x}_{il_i} &= f_{il_i}(\underline{x}_{l_i}, u) \end{aligned} \tag{1.1.7}$$

for $i = 1, \ldots, p$ *where* $\underline{x}_j$ *is defined by*

$$\underline{x}_j = (x_{11}, \ldots, x_{1,j\wedge l_1}, x_{21}, \ldots, x_{pj}). \tag{1.1.8}$$

Notice that in $\underline{x}_j$ *the indices range over* $i = 1, \ldots, p;\ k = 1, \ldots, \min\{j, l_i\}$ *and the coordinates are ordered so that second index moves faster than the first.*

We also require that each f_{ij} *be Lipschitz continuous, there exists an* L *such that for all* $x, \xi \in I\!R^n, u \in U$,

$$|f_i(\underline{x}_j, u) - f_i(\underline{\xi}_j, u)| \leq L|\underline{x}_j - \underline{\xi}_j|. \tag{1.1.9}$$

The symbol $|\cdot|$ *denotes the Euclidean norm.*

Let

$$\bar{A}_i = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}^{l_i \times l_i} \qquad \bar{A} = \begin{bmatrix} \bar{A}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{A}_p \end{bmatrix}^{n \times n}$$

$$\bar{C}_i = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix}^{1\times l_i} \qquad \bar{C} = \begin{bmatrix} \bar{C}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{C}_p \end{bmatrix}^{p\times n}$$

$$\bar{f}_i(x,u) = \begin{bmatrix} f_{i1}(\underline{x}_1, u) \\ \vdots \\ f_{il_i}(\underline{x}_{l_i}, u) \end{bmatrix}^{l_i\times 1} \qquad \bar{f}(x,u) = \begin{bmatrix} \bar{f}_1(x,u) \\ \vdots \\ \bar{f}_p(x,u) \end{bmatrix}^{n\times 1}$$

$$\bar{h}(u) = \begin{bmatrix} h_1(u) \\ \vdots \\ h_p(u) \end{bmatrix}^{p\times 1}$$

then (1.1.7) becomes

$$\begin{aligned} \dot{x} &= \bar{A}x + \bar{f}(x,u) \\ y &= \bar{C}x + \bar{h}(u) \end{aligned} \tag{1.1.10}$$

A system such as (1.1.7) or, equivalently (1.1.10), is said to be in observable form [4].

We shall also require that the second derivative of $\bar{f}$ is bounded, i.e., for any $x, \xi \in I\!R^n, u \in U$,

$$\left| \frac{\partial^2 \bar{f}}{\partial x_i \partial x_j}(x,u)\xi_i\xi_j \right| \leq L|\xi|^2. \tag{1.1.11}$$

On the left we employ the convention of summing on repeated indices.

Theorem 1.1.1 (Main Theorem) *Suppose*

- *the system (1.1.1) is uniformly observable for any input and so without loss of generality we can assume that is in the form (1.1.10) and satisfies the Lipschitz conditions (1.1.9),*
- *the second derivative of $\bar{f}$ is bounded (1.1.11),*
- *$x(t), y(t)$ are any state and output trajectories generated by (1.1.10),*
- *G has been chosen to be invertible,*
- *$\hat{x}(t)$ and $P(t)$ are a solution of the extended Kalman filter (1.1.5) where $P(0)$ is positive definite and $\tilde{x}(0) = x(0) - \hat{x}(0)$ is sufficiently small,*

Then $|x(t) - \hat{x}(t)| \to 0$ exponentially as $t \to \infty$.

1.2 Proof of the Main Theorem

We extend the method of proof of [1]. Since the system is in observable form

$$\begin{aligned} A(t) &= \bar{A} + \tilde{A}(t) \\ C(t) &= \bar{C} \end{aligned} \tag{1.2.1}$$

where

$$\tilde{A}(t) = \frac{\partial \bar{f}}{\partial x}(\hat{x}(t), u(t)),$$

and

$$\frac{\partial \bar{f}_{ir}}{\partial x_{jk}}(\hat{x}(t)) = 0$$

if $k > r$.

First we show that there exists $m_1 > 0$ such that for all $t \geq 0$

$$P(t) \leq m_1 I^{n\times n}. \tag{1.2.2}$$

Consider the optimal control problem of minimizing

$$\xi'(0)P(0)\xi(0) + \int_0^t \xi'(s)\Gamma\xi(s) + \mu'(s)\mu(s)\ ds$$

subject to

$$\begin{aligned} \dot{\xi}(s) &= -A'(s)\xi(s) - \bar{C}'\mu(s) \\ \xi(t) &= \zeta. \end{aligned}$$

It is well-known that the optimal cost is

$$\zeta' P(t)\zeta$$

where $P(t)$ is the solution of (1.1.5).

Following [2] for $\theta > 0$ we define $S(\theta)$ as the solution of

$$\bar{A}'S(\theta) + S(\theta)\bar{A} - \bar{C}'\bar{C} = -\theta S(\theta). \tag{1.2.3}$$

It is not hard to see that $S(\theta)$ is positive definite for $\theta > 0$ as it satisfies the Lyapunov equation

$$\left(-\bar{A} - \frac{\theta}{2}I\right)' S(\theta) + S(\theta)\left(-\bar{A} - \frac{\theta}{2}I\right) = -\bar{C}'\bar{C}$$

where $\bar{C}, \left(-\bar{A} - \frac{\theta}{2} I\right)$ is an observable pair and $\left(-\bar{A} - \frac{\theta}{2} I\right)$ has all eigenvalues equal to $-\frac{\theta}{2}$. It follows from (1.2.3) that

$$S_{ij,\rho\sigma}(\theta) = \frac{S_{ij,\rho\sigma}(1)}{\theta^{j+\sigma-1}} = \frac{(-1)^{j+\sigma}}{\theta^{j+\sigma-1}} \begin{pmatrix} j+\sigma-2 \\ j-1 \end{pmatrix}.$$

Let $T(\theta) = S^{-1}(\theta) > 0$ then

$$T_{ij,\rho\sigma}(\theta) = \theta^{j+\sigma-1} T_{ij,\rho\sigma}(1)$$

and $T(\theta)$ satisfies the Riccati equation

$$-\bar{A}T(\theta) - T(\theta)\bar{A}' + T(\theta)\bar{C}'\bar{C}T(\theta) = \theta T(\theta).$$

We apply the suboptimal control $\mu = -\bar{C}T(\theta)\xi$ to the above optimal control problem and conclude that

$$\zeta' P(t)\zeta \leq \xi'(0)P(0)\xi(0) + \int_0^t \xi'(s)\left(\Gamma + T(\theta)\bar{C}'\bar{C}T(\theta)\right)\xi(s)\, ds \tag{1.2.4}$$

where

$$\begin{aligned} \dot{\xi}(s) &= \left(-A'(s) + \bar{C}'\bar{C}T(\theta)\right)\xi(s) \\ \xi(t) &= \zeta. \end{aligned}$$

Now

$$\begin{aligned} \frac{d}{ds}\xi'(s)T(\theta)\xi(s) &= \xi'(s)\left(\theta T(\theta) + T(\theta)\bar{C}'\bar{C}T(\theta)\right)\xi(s) \\ &\quad -\xi'(s)\left(\tilde{A}(s)T(\theta) + T(\theta)\tilde{A}'(s)\right)\xi(s). \end{aligned}$$

Because of the Lipschitz condition (1.1.9) we conclude that

$$|A(s)| \leq L$$

and

$$|\tilde{A}(s)| \leq L + |\bar{A}|.$$

From the form of $\tilde{A}(s)$ and $T(\theta)$ we conclude that

$$\left(\tilde{A}(s)T(\theta)\right)_{ij,\rho\sigma} = O(\theta)^{j+\sigma-1}$$

while on the other hand

$$\theta T_{ij,\rho\sigma}(\theta) = \theta^{j+\sigma} T_{ij,\rho\sigma}(1).$$

Hence we conclude that for any $\alpha > 0$ there exists θ sufficiently large so that

$$\theta T(\theta) + T(\theta)\bar{C}'\bar{C}T(\theta) - \tilde{A}(s)T(\theta) - T(\theta)\tilde{A}'(s) \geq \alpha I^{n\times n}.$$

Therefore for $0 \leq s \leq t$

$$\xi'(s)T(\theta)\xi(s) \quad \leq \quad e^{\alpha(s-t)}\zeta'\zeta$$

Now there exists $m_2(\theta) > 0$ such that

$$\xi'(s)\xi(s) \leq m_2(\theta)\xi'(s)T(\theta)\xi(s)$$

so we conclude that

$$\xi'(s)\xi(s) \leq m_2(\theta)e^{\alpha(s-t)}\zeta'\zeta.$$

There exist constants $m_3 > 0, m_4(\theta) > 0$ such that

$$\begin{aligned} P(0) &\leq m_3 I^{n\times n} \\ \Gamma + T(\theta)\bar{C}'\bar{C}T(\theta) &\leq m_4(\theta)I^{n\times n} \end{aligned}$$

From (1.2.4) we obtain the desired conclusion,

$$\begin{aligned} \zeta'P(t)\zeta &\leq m_3 e^{-\alpha t}\zeta'\zeta + m_4(\theta)\int_0^t e^{\alpha(s-t)}\zeta'\zeta\ ds \\ \zeta'P(t)\zeta &\leq m_3\zeta'\zeta + m_4(\theta)\int_{-\infty}^t e^{\alpha(s-t)}\zeta'\zeta\ ds \\ \zeta'P(t)\zeta &\leq \frac{m_3 + m_4(\theta)}{\alpha}\zeta'\zeta \end{aligned}$$

Define

$$Q(t) = P^{-1}(t)$$

then Q satisfies

$$\begin{aligned} \dot{Q}(t) &= -A'(t)Q(t) - Q(t)A(t) - Q(t)\Gamma Q(t) + \bar{C}'\bar{C} \\ Q(0) &= P^{-1}(0) > 0 \end{aligned} \qquad (1.2.5)$$

Next we show that there exists $m_5 > 0$ such that for all $t \geq 0$

$$Q(t) \quad \leq \quad m_5 I^{n\times n}.$$

This will imply that

$$P(t) \quad \geq \quad \frac{1}{m_5}I^{n\times n}. \qquad (1.2.6)$$

Consider the optimal control problem of minimizing

$$\xi'(0)Q(0)\xi(0) + \int_0^t \xi'(s)\bar{C}'\bar{C}\xi(s) + \mu'(s)\mu(s)\ ds$$

subject to

$$\begin{aligned}\dot{\xi}(s) &= A(s)\xi(s) + G\mu(s)\\ \xi(t) &= \zeta.\end{aligned}$$

It is well-known that the optimal cost is

$$\zeta' Q(t)\zeta$$

where $Q(t)$ is the solution of (1.1.5).

We use the suboptimal control

$$\mu(s) = G^{-1}(\alpha I^{n\times n} - A(s))\xi(s)$$

so that the closed loop dynamics is

$$\begin{aligned}\dot{\xi}(s) &= \alpha\xi(s)\\ \xi(t) &= \zeta\\ \xi(s) &= e^{\alpha(s-t)}\zeta.\end{aligned}$$

From this we obtain the desired bound

$$\begin{aligned}\zeta'Q(t)\zeta &\leq \xi'(0)Q(0)\xi(0)\\ &\quad + \int_0^t \xi'(s)\left(\bar{C}'\bar{C} + (\alpha I^{n\times n} - A'(s))\Gamma^{-1}(\alpha I^{n\times n} - A(s))\right)\xi(s)\ ds\\ \zeta'Q(t)\zeta &\leq e^{-2\alpha t}\zeta'Q(0)\zeta + \int_0^t e^{2\alpha(s-t)}\zeta'\left(\bar{C}'\bar{C} + (\alpha+L)^2\Gamma^{-1}\right)\zeta\ ds\\ \zeta'Q(t)\zeta &\leq \left(m_6 + \frac{m_7}{2\alpha}\right)\zeta'\zeta\end{aligned}$$

where

$$\begin{aligned}Q(0) &\leq m_6 I^{n\times n}\\ \bar{C}'\bar{C} + (\alpha+L)^2\Gamma^{-1} &\leq m_6 I^{n\times n}.\end{aligned}$$

Now let $x(t), u(t), y(t)$ be a trajectory of the system (1.1.10) starting at x^0. Let $\hat{x}(t)$ be the trajectory of the extended Kalman filter (1.1.5) starting at $\hat{x}^0$ and $\tilde{x}(t) = x(t) - \hat{x}(t)$, $\tilde{x}^0 = x^0 - \hat{x}^0$. Then

$$\begin{aligned}\frac{d}{dt}\tilde{x}'(t)Q(t)\tilde{x}(t) &= -\tilde{x}'(t)\left(\bar{C}'\bar{C} + Q(t)\Gamma Q(t)\right)\tilde{x}(t)\\ &\quad +2\tilde{x}'(t)Q(t)\left(\bar{f}(x(t),u(t)) - \bar{f}(\hat{x}(t),u(t)) - \tilde{A}(t)\tilde{x}(t)\right).\end{aligned}$$

Now following [1]

$$\begin{aligned}\bar{f}(x(t),u(t)) - \bar{f}(\hat{x}(t),u(t)) - \tilde{A}(t)\tilde{x}(t) &= \int_0^1\int_0^1 r\frac{\partial^2\bar{f}}{\partial x_i\partial x_j}(\hat{x}(t)+rs\tilde{x}(t),u(t))\tilde{x}_i(t)\tilde{x}_j(t)\;ds\;dr \\ &\leq L|\tilde{x}(t)|^2.\end{aligned}$$

Since G is invertible there exists $m_7 > 0$ such that

$$\Gamma \;\geq\; m_7 I^{n\times n}$$

and so

$$\begin{aligned}\frac{d}{dt}\tilde{x}'(t)Q(t)\tilde{x}(t) &\leq -\frac{m_7}{m_1^2}|\tilde{x}(t)|^2 + m_5 L|\tilde{x}(t)|^3 \\ &\leq -\frac{m_7}{m_1^2 m_5}\tilde{x}'(t)Q(t)\tilde{x}(t) + m_5 L(m_1\tilde{x}'(t)Q(t)\tilde{x}(t))^{\frac{3}{2}}\end{aligned}$$

If

$$(\tilde{x}'(t)Q(t)\tilde{x}(t))^{\frac{1}{2}} < \frac{m_7}{2m_1^{\frac{7}{2}}m_5^2 L}$$

then

$$\frac{d}{dt}\tilde{x}'(t)Q(t)\tilde{x}(t) \;\leq\; -\frac{m_7}{2m_1^2 m_5}\tilde{x}'(t)Q(t)\tilde{x}(t)$$

so $\tilde{x}'(t)Q(t)\tilde{x}(t) \to 0$ exponentially as $t \to \infty$. Therefore if

$$(\tilde{x}'(t)Q(t)\tilde{x}(0))^{\frac{1}{2}} < \frac{m_7}{2m_1^{\frac{7}{2}}m_5^2 L}$$

the extended Kalman filter converges.

1.3 Conclusions

The above result does not follow from that of Baras, Bensoussan and James [1]. To show local convergence of the extended Kalman filter they required "uniform detectability". They define this as follows. The system

$$\begin{aligned}\dot{x} &= f(x) \\ y &= Cx\end{aligned} \tag{1.3.1}$$

is uniformly detectable if there exists a bounded Borel matrix-valued function $\Lambda(x)$ and a constant $\alpha > 0$ such that for all $x, \xi \in I\!R^n$

$$\xi'\left(\frac{\partial f}{\partial x}(x) + \Lambda(x)C\right)\xi \;\leq\; -\alpha|\xi|^2.$$

This is a fairly restrictive condition as not all observable linear systems are uniformly detectable. Consider

$$\begin{aligned}\dot{x} &= \begin{bmatrix} 0 & 1 \\ a_1 & a_2 \end{bmatrix} x \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x.\end{aligned}$$

Suppose

$$\Lambda(x) = \begin{bmatrix} \lambda_1(x) \\ \lambda_2(x) \end{bmatrix}$$

then

$$\frac{\partial f}{\partial x}(x) + \Lambda(x)C = \begin{bmatrix} \lambda_1(x) & 1 \\ \lambda_2(x) + a_1 & a_2 \end{bmatrix}.$$

If $a_2 > 0$ and $\xi' = \begin{bmatrix} 0 & 1 \end{bmatrix}$ then

$$\xi' \left(\frac{\partial f}{\partial x}(x) + \Lambda(x)C \right) \xi = a_2 > 0$$

so the system is not uniformly detectable. This system does satisfies the conditions of Theorem 1.1.1 so an extended Kalman filter would converge locally. Since the system is linear, an extended Kalman filter is also a Kalman filter that converges globally

An example [5] of a highly nonlinear problem where an EKF may fail to converge is

$$\begin{aligned}\dot{x} &= f(x) = x(1 - x^2) \\ y &= h(x) = x^2 - x/2\end{aligned} \tag{1.3.2}$$

where $x, y \in I\!R$. The system is observable as $y, \dot{y}, \ddot{y}$ separate points but it is not uniformly observable. The dynamics has two stable equilibria at $x = \pm 1$ and an unstable equilibrium at $x = 0$. Under certain initial conditions, the extended Kalman filter fails to converge. Suppose the $x^0 = 1$ so $x(t) = 1$ and $y(t) = 1/2$ for all $t \geq 0$. But $h(-1/2) = 1/2$ so if $\hat{x}^0 \leq -1/2$ the extended Kalman filter will not converge. To see this notice that when $\hat{x}(t) = -1/2$, the term $y(t) - h(\hat{x}(t)) = 0$ so $\dot{\hat{x}} = f(\hat{x}(t)) = f(-1/2) = -3/8$. Therefore $\hat{x}(t) \leq -1/2$ for all $t \geq 0$.

1.4 Dedication

This paper is dedicated to my esteemed colleague and good friend, Professor Anders Lindquist on the occasion of his 60^{th} birthday.

Bibliography

[1] J. S. Baras, A. Bensoussan and M. R. James, *Dynamic observers as asymptotic limits of recursive filters: special cases,* SIAM J. on Applied Mathematics, 48 (1988), pp. 1147-1158.

[2] J. P. Gauthier, H. Hammouri and S. Othman, *A simple observer for nonlinear systems with applications to bioreactors,* IEEE Trans. on Automatic Control, 37 (1992), pp. 875-880.

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

[4] A. J. Krener, *Normal forms for linear and nonlinear systems,* Contemporary Mathematics, 68, *Differential Geometry, the Interface between Pure and Applied Mathematics*, American Mathematical Society, Providence, RI, (1987), pp. 157-189.

[5] A. J. Krener and A. Duarte, *A hybrid computational approach to nonlinear estimation,* in Proc. of 35th Conference on Decision and Control, Kobe, Japan, (1996) pp. 1815-1819.