On stability of a class of Kalman–Bucy filters for systems with non-linear dynamics

Toni Karvonen

Department of Electrical Engineering and Automation
Aalto University, Espoo, Finland

SFB 1294 Seminar, University of Potsdam
March 9, 2018

Aalto University
School of Electrical Engineering

Collaborators

Silvère Bonnabel
Mines ParisTech

Eric Moulines
École Polytechnique

Simo Särkkä
Aalto University

Inspiration

This work is inspired by a recent series of articles by P. Del Moral, A. N. Bishop, A. Kurtzmann, and J. Tugaut:

- DM, K, T (2016). **On the stability and the exponential concentration of extended Kalman-Bucy filters**. arXiv:1606.08251.
- DM, K, T (2017). On the stability and the uniform propagation of chaos of a class of extended ensemble Kalman–Bucy filters. *SIAM J. Control Optim.* 55(1):119–155.
- B, DM (2017). On the stability of Kalman–Bucy diffusion processes. *SIAM J. Control Optim.* 55(6):4015–4047.
- DM, T (2016). On the stability and the uniform propagation of chaos properties of ensemble Kalman-Bucy filters. arXiv:1605.09329.

Table of contents

Model: the linear case

Consider a linear SDE system with the **state** $X_t \in \mathbb{R}^{d_X}$ and **measurements** $Y_t \in \mathbb{R}^{d_Y}$:

$$\begin{aligned} \mathrm{d}X_t &= A_t X_t \,\mathrm{d}t + Q^{1/2}\,\mathrm{d}W_t, \\ \mathrm{d}Y_t &= H X_t \,\mathrm{d}t + R^{1/2}\,\mathrm{d}V_t, \end{aligned}$$

with W_t and V_t independent standard Brownian motions.

In filtering, we want to infer the state based on the potentially partial and noisy measurements. The aim is to compute the

$$\textbf{filtering distributions} \quad X_t \mid \mathcal{F}_t, \quad \mathcal{F}_t = \sigma(Y_s, 0 \le s \le t),$$

for each $t \geq 0$.

The Kalman–Bucy filter

If the initial state X_0 is Gaussian, the filtering distributions $X_t \mid \mathcal{F}_t$ are Gaussian and solved by the classical **Kalman–Bucy filter**.

The mean $\widehat{X}_t = \mathbb{E}[X_t \mid \mathcal{F}_t]$ of the filtering distribution evolves according to the SDE

$$\mathrm{d}\widehat{X}_t = A_t \widehat{X}_t \,\mathrm{d}t + P_t H^\mathsf{T} R^{-1} \big[\,\mathrm{d}Y_t - H \widehat{X}_t \,\mathrm{d}t\big],$$

where P_t is the **error covariance**

$$P_t = \mathbb{E}\big[\big(X_t - \widehat{X}_t\big)\big(X_t - \widehat{X}_t\big)^\mathsf{T} \mid \mathcal{F}_t\big].$$

It is solved from the **Riccati equation**

$$\partial_t P_t = A_t P_t + P_t A_t^\mathsf{T} + Q - P_t S P_t \quad (S = H^\mathsf{T} R^{-1} H).$$

Non-linear dynamics

We consider a (partially) non-linear extension

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + Q^{1/2}\,\mathrm{d}W_t,$$
$$\mathrm{d}Y_t = HX_t\,\mathrm{d}t + R^{1/2}\,\mathrm{d}V_t,$$

where the *drift* $f : \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ is now a non-linear function with a bounded Jacobian.

The filtering distributions are not going to remain Gaussian in this case. Some kind of approximation must be used.

Kalman–Bucy filters for non-linear systems are based on *pretending* that the filtering distributions are Gaussian using some sort of linearisation.

The extended Kalman–Bucy filter

The **extended Kalman–Bucy filter** (EKF) employs simple first-order linearisations to approximate the filtering mean and covariance:

$$\begin{aligned} \mathrm{d}\widehat{X}_t &= f(\widehat{X}_t)\,\mathrm{d}t + P_t H^\mathsf{T} R^{-1}\big[\,\mathrm{d}Y_t - H\widehat{X}_t\,\mathrm{d}t\big], \\ \partial_t P_t &= J_f(\widehat{X}_t)P_t + P_t J_f(\widehat{X}_t)^\mathsf{T} + Q - P_t S P_t, \end{aligned}$$

where $J_f(x) \in \mathbb{R}^{d_X \times d_X}$ is the Jacobian of f at point $x \in \mathbb{R}^{d_X}$.

In contrast to the linear case, the Riccati equation now depends on the measurements Y_t through the Jacobian $J_f(\widehat{X}_t)$.

Gaussian assumed density filters

In the **Gaussian assumed density filter**, the point evaluation of the EKF are replaced with Gaussian expectations:

$$
\begin{aligned}
\mathrm{d}\widehat{X}_t &= \mathbb{E}_{\mathcal{N}(\widehat{X}_t, P_t)}(f)\,\mathrm{d}t + P_t H^{\mathsf{T}} R^{-1}\big[\,\mathrm{d}Y_t - H\widehat{X}_t\,\mathrm{d}t\big], \\
\partial_t P_t &= \mathbb{E}_{\mathcal{N}(\widehat{X}_t, P_t)}(J_f) P_t + P_t \mathbb{E}_{\mathcal{N}(\widehat{X}_t, P_t)}(J_f)^{\mathsf{T}} + Q - P_t S P_t,
\end{aligned}
$$

where

$$
\mathbb{E}_{\mathcal{N}(x,P)}(g) \coloneqq \int_{\mathbb{R}^{d_X}} g(z)\,\mathcal{N}(z \mid x, P)\,\mathrm{d}z
$$

is element-wise expectation.

This filter can be implemented in practice only rarely because the integrals typically lack analytical solutions. *Use numerical integration*.

Gaussian numerical integration filters

A (Gaussian) **numerical integration filter** uses numerical integration to approximate the expectations:

$$\sum_{i=1}^{N} w_i g(x) \approx \mathbb{E}_{\mathcal{N}(0,I)}(g) \quad \text{and} \quad \sum_{i=1}^{N} w_i g\big(x + \sqrt{P}\xi_i\big) \approx \mathbb{E}_{\mathcal{N}(x,P)}(g)$$

for some *weights* $w_i \in \mathbb{R}$ and *unit sigma-points* $\xi_i \in \mathbb{R}^{d_X}$. The filter is

$$\begin{aligned} \mathrm{d}\widehat{X}_t &= \sum_{i=1}^{n} w_i f\big(\widehat{X}_t + \sqrt{P_t}\xi_i\big)\,\mathrm{d}t + P_t H^\mathsf{T} R^{-1}\big[\,\mathrm{d}Y_t - H\widehat{X}_t\,\mathrm{d}t\big], \\ \partial_t P_t &= \sum_{i=1}^{n} w_i \Big[f\big(\widehat{X}_t + \sqrt{P_t}\xi_i\big)\xi_i^\mathsf{T}\sqrt{P_t} + \sqrt{P_t}\xi_i f\big(\widehat{X}_t + \sqrt{P_t}\xi_i\big)^\mathsf{T}\Big] \\ &\quad + Q - P_t S P_t. \end{aligned}$$

Numerical integration filters: examples

- *Unscented Kalman–Bucy filter* (UKF): $2d_X + 1$ sigma-points

$$\xi_{2d_X+1} = 0, \quad \xi_i = e_i, \quad \xi_{i+d_X} = -e_i \text{ for } i = 1, \ldots, d_X.$$

- *Cubature Kalman–Bucy filter* (CKF): UKF but without the central point.
- *Gauss–Hermite Kalman–Bucy filter* (GHKF): tensor product grids based on the univariate Gauss–Hermite quadrature rule.
- And a myriad of others: stochastic integration, Monte Carlo, sparse grids, Bayesian cubature ... (all these are not necessarily covered by our stability analysis)

General class of Kalman–Bucy filters

Let $L_{x,P}$ be a linear functional that is applied element-wise to vector-valued functions. We consider a class of *generic* filters computing the conditional mean approximation as

$$\mathrm{d}\widehat{X}_t = L_{\widehat{X}_t,P_t}(f)\,\mathrm{d}t + P_t H^\mathsf{T} R^{-1}\big(\mathrm{d}Y_t - H\widehat{X}_t\,\mathrm{d}t\big).$$

We *do not* require that P_t is a solution to a Riccati-type equation. However, a time-uniform upper bound on $\mathrm{tr}(P_t)$ will be assumed later.

Examples:

- EKF: $L_{x,P}(g) = g(x)$.
- Gaussian assumed density filter: $L_{x,P}(g) = \mathbb{E}_{\mathcal{N}(x,P)}(g)$.
- Numerical integration filters: $L_{x,P}(g) = \sum_{i=1}^{N} w_i g\big(x + \sqrt{P}\xi_i\big)$.

Assumption on $L_{x,P}$

Define the **logarithmic norms**

$$\nu(A) = \frac{1}{2}\lambda_{\min}(A + A^{\mathsf{T}}) \quad \text{and} \quad \mu(A) = \frac{1}{2}\lambda_{\max}(A + A^{\mathsf{T}})$$

of a square matrix A and the "Lipschitz constants"

$$N(g) = \inf_z \nu[J_g(z)] \quad \text{and} \quad M(g) = \sup_z \mu[J_g(z)]$$

that satisfy

$$N(g)\,\|x - y\|^2 \leq \langle x - y, g(x) - g(y) \rangle \leq M(g)\,\|x - y\|^2 .$$

Assumption

For every differentiable $g\colon \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ there is a constant $C_g \geq 0$, varying continuously with $M(g)$ and $N(g)$, such that

$$\langle x - y, g(x) - L_{y,P}(g) \rangle \leq M(g)\,\|x - y\|^2 + C_g\,\mathrm{tr}(P)$$

for any points $x, y \in \mathbb{R}^{d_X}$ and any matrix $P \in \mathbb{R}^{d_X \times d_X}$.

Table of contents

Filtering error I

Consider the EKF and recall that

$$\begin{aligned}
\mathrm{d}X_t &= f(X_t)\,\mathrm{d}t + Q^{1/2}\,\mathrm{d}W_t, \\
\mathrm{d}Y_t &= HX_t\,\mathrm{d}t + R^{1/2}\,\mathrm{d}V_t, \\
\mathrm{d}\widehat{X}_t &= f(\widehat{X}_t)\,\mathrm{d}t + P_t H^\mathsf{T} R^{-1}\big[\,\mathrm{d}Y_t - H\widehat{X}_t\,\mathrm{d}t\big].
\end{aligned}$$

The filtering error $E_t \coloneqq X_t - \widehat{X}_t$ is

$$\begin{aligned}
\mathrm{d}E_t = \big[f(X_t) - f(\widehat{X}_t) - P_t H^\mathsf{T} R^{-1} H(X_t - \widehat{X}_t)\big]\,\mathrm{d}t \\
+ Q^{1/2}\,\mathrm{d}W_t - P_t H^\mathsf{T} R^{-1/2}\,\mathrm{d}V_t.
\end{aligned}$$

Filtering error II

Itô's lemma yields

$$\begin{aligned}\mathrm{d}\|E_t\|^2 &= 2\big\langle f(X_t) - f(\widehat{X}_t) - P_t S(X_t - \widehat{X}), X_t - \widehat{X}_t \big\rangle \, \mathrm{d}t \\ &\quad + 2\big[\operatorname{tr}(Q) + \operatorname{tr}(SP_t^2)\big] \, \mathrm{d}t + \mathrm{d}M_t \\ &\leq 2M\big(f - P_t S\big) \, \|E_t\|^2 + 2\big[\operatorname{tr}(Q) + \operatorname{tr}(SP_t^2)\big] \, \mathrm{d}t + \mathrm{d}M_t\end{aligned}$$

for a zero-mean martingale M_t. For expectation,

$$\partial_t \mathbb{E}(\|E_t\|^2) \leq 2M(f - P_t S)\mathbb{E}(\|E_t\|^2) + 2\big[\operatorname{tr}(Q) + \operatorname{tr}(SP_t^2)\big].$$

Grönwall's inequality produces

$$\mathbb{E}(\|E_t\|^2) \leq \mathbb{E}(\|E_0\|^2) \, \mathrm{e}^{2\alpha_t t} + \frac{\mathrm{e}^{2\alpha_t t} - 1}{\alpha_t} \Big[\operatorname{tr}(Q) + \lambda_{\max}(S) \sup_{\tau \leq t} \operatorname{tr}(P_\tau^2) \Big]$$

where

$$\alpha_t = \sup_{\tau \leq t} M\big(f - P_t S\big).$$

Main assumptions

We impose strong time-uniform boundedness and contractivity conditions.

Assumption I

There exists $\lambda_P \geq 0$ such that $\sup_{t \geq 0} \operatorname{tr}(P_t) \leq \lambda_P$.

Assumption II

There exists a time $T \geq 0$ and $\lambda > 0$ such that

$$M(f - P_t S) = \sup_{x \in \mathbb{R}^{d_X}} \mu\big[J_f(x) - P_t S\big] \leq -\lambda < 0$$

holds for every $t \geq T$.

On the assumptions

These assumptions are very restrictive:

- For linear Kalman–Bucy filters, it can be shown (under appropriate conditions) that $A_t - P_t S$ defines an exponentially stable system. However, it does not necessarily follow that $\mu(A_t - P_t S) < 0$.
- For time-invariant linear systems ($A_t = A$), the limit $\lim_{t\to\infty} P_t = P$ satisfies the time-invariant stability condition

$$\alpha(A - PS) \coloneqq \max_{i=1,\dots,d_X} \mathrm{Re}\big[\lambda_i(A - PS)\big] < 0.$$

  We can only say that $\alpha(A - PS) \leq \mu(A - PS)$.
- In practice, P_t is the solution to a non-linear Riccati-type equation. Such equations are difficult to control for all but fully observed models.

Some implications

Recall that (still for the EKF)

$$\partial_t \mathbb{E}(\|E_t\|^2) \leq 2M(f - P_t S)\mathbb{E}(\|E_t\|^2) + 2\big[\operatorname{tr}(Q) + \operatorname{tr}(SP_t^2)\big].$$

Under our assumptions,

$$\partial_t \mathbb{E}(\|E_t\|^2) \leq 2\big[-\lambda \mathbb{E}(\|E_t\|^2) + u\big]$$

with $u = \operatorname{tr}(Q) + \lambda_{\max}(S)\lambda_P^2$ for $t \geq T$. Thus, by Grönwall,

$$\mathbb{E}(\|E_t\|^2) \leq \mathbb{E}(\|E_T\|^2)\,\mathrm{e}^{-2\lambda(t-T)} + u/\lambda \leq \mathbb{E}(\|E_T\|^2) + u/\lambda.$$

More can be done with the use of Bernstein's inequality.

Bernstein's inequality

Let X be a non-negative random variable. Suppose that there exists $\alpha > 0$ such that

$$\mathbb{E}(X^n) \leq n^n \alpha^n$$

for every integer $n \geq 2$. Then

$$\mathbb{P}\big[X \geq \alpha\, \mathrm{e} \left(\sqrt{2\delta} + \delta\right)\big] \leq \mathrm{e}^{-\delta}$$

for any $\delta > 0$.

By applying Ito's lemma to $\|E_t\|^{2n}$ and deriving an affine differential inequality for $\mathbb{E}(\|E_t\|^{2n})^{1/n}$, it is not too difficult to show that

$$\mathbb{E}(\|E_t\|^{2n}) \leq \left(C_T\, \mathrm{e}^{-2\lambda(t-T)} + u/\lambda\right)^n n^n$$

for every $n \geq 1$ and for certain $C_T \geq 0$.

An exponential concentration inequality

Theorem

Consider any generic Kalman–Bucy filter and suppose that Assumptions I and II hold. Denote

$$\beta(\delta) = \mathrm{e}\,(\sqrt{2\delta} + \delta).$$

Then there are non-negative constants C_λ *and* C_T *such that, for any* $t \geq T$ *and* $\delta > 0$*, the probability of*

$$\left\|X_t - \widehat{X}_t\right\|^2 \geq \left(C_T\,\mathrm{e}^{-2\lambda(t-T)} + \frac{\mathrm{tr}(Q) + C_\lambda \lambda_P + \mathrm{tr}(S)\lambda_P^2}{\lambda} \right) \beta(\delta)$$

is smaller than $\mathrm{e}^{-\delta}$*.*

- C_λ depends on the filter. E.g., $C_\lambda = 0$ for the EKF.
- C_T is an upper bound on $\mathbb{E}(\|E_T\|^2)$ under the assumptions $M(f) < \infty$ and $\sup_{t\geq 0} \mathrm{tr}(P_t) \leq \lambda_P$.

Table of contents

Fully observed models

Definition

A model is **fully observed** if $S = H^\mathsf{T} R^{-1} H = sI$ for some $s > 0$.

For fully observed models,

$$M(f - P_t S) \leq M(f) + s\mu(-P_t) = M(f) - s\lambda_{\min}(P_t).$$

This is negative if

- The model is contractive: $M(f) < 0$. This implies that $x_t \to 0$ exponentially fast for the homogeneous system $\partial_t x_t = f(x_t)$.
- The error covariance is large enough:

$$\inf_{t \geq T} \lambda_{\min}(P_t) > M(f)/s.$$

Contractive models (Del Moral, Kurtzmann, Tugaut)

This is the case considered by Del Moral, Kurtzmann, and Tugaut[1].

Assume full observability and $M(f) \leq -\lambda < 0$. Then

$$\begin{aligned}
\mathrm{d}\|E_t\|^2 &= 2\Big[\big\langle f(X_t) - f(\widehat{X}_t), X_t - \widehat{X}_t\big\rangle - s\langle P_t(X_t - \widehat{X}_t), X_t - \widehat{X}_t\rangle\Big]\,\mathrm{d}t \\
&\quad + 2\big[\operatorname{tr}(Q) + s\operatorname{tr}(P_t^2)\big]\,\mathrm{d}t + \mathrm{d}M_t \\
&\leq -2\lambda\,\|E_t\|^2 + 2\big[\operatorname{tr}(Q) + s\operatorname{tr}(P_t^2)\big]\,\mathrm{d}t + \mathrm{d}M_t
\end{aligned}$$

and

$$\begin{aligned}
\partial_t \operatorname{tr}(P_t) &= \operatorname{tr}\left[J_f(\widehat{X}_t)P_t + P_t J_f(\widehat{X}_t)\right] + \operatorname{tr}(Q) - s\operatorname{tr}(P_t^2) \\
&\leq -2\lambda\operatorname{tr}(P_t) + \operatorname{tr}(Q).
\end{aligned}$$

Grönwall: $\operatorname{tr}(P_t)$ and $\mathbb{E}(\|E_t\|^2)$ remain uniformly bounded.

[1] On the stability and the exponential concentration of extended Kalman-Bucy filters. arXiv:1606.08251, 2016.

Bounds on the Riccati equation

For simplicity, consider again the EKF. The (tuned) Riccati equation for P_t is

$$\partial_t P_t = J_f(\widehat{X}_t)P_t + P_t J_f(\widehat{X}_t)^{\mathsf{T}} + Q_{\mathrm{tu}} - sP_t^2,$$

where Q_{tu} is positive-definite matrix that does not have to equal Q.

Proposition

At the limit $t \to \infty$,

$$\mathrm{tr}(P_t) \geq \frac{\lambda_{\min}(Q_{\mathrm{tu}})/d_X}{N(f) + \sqrt{s\lambda_{\min}(Q_{\mathrm{tu}})/d_X + N(f)^2}}$$

and

$$\mathrm{tr}(P_t) \leq \frac{M(f) + \sqrt{s\,\mathrm{tr}(Q_{\mathrm{tu}})/d_X + M(f)^2}}{s/d_X}.$$

Fully observable models with covariance inflation

Since

$$\mathrm{tr}(P_t) \geq \frac{\lambda_{\min}(Q_{\mathrm{tu}})/d_X}{N(f) + \sqrt{s\lambda_{\min}(Q_{\mathrm{tu}})/d_X + N(f)^2}},$$

$\inf_{t \geq T} \lambda_{\min}(P_t) > M(f)/s$ for large enough tuned model noise covariance Q_{tu}.

For the EKF, the probability of

$$\left\|X_t - \widehat{X}_t\right\|^2 \geq \left(C_T\, \mathrm{e}^{-2\lambda(t-T)} + \frac{\mathrm{tr}(Q) + \mathrm{tr}(S)\lambda_P^2}{\lambda} \right) \mathrm{e}\left(\sqrt{2\delta} + \delta\right)$$

is smaller than $\mathrm{e}^{-\delta}$.

Using inflated Q_{tu} guarantees stability but degrades accuracy as λ_P^2 becomes larger.

Integrated velocity models I

Consider the **integrated velocity model**

$$\mathrm{d}\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} A_1 X_{t,1} + A_2 X_{t,2} \\ -g(X_{t,2}) \end{pmatrix} \mathrm{d}t + Q^{1/2}\,\mathrm{d}W_t,$$
$$\mathrm{d}Y_t = H_1 X_{t,1}\,\mathrm{d}t + R^{1/2}\,\mathrm{d}V_t,$$

where the non-linear function $M(g) < 0$.

The state component $X_{t,2}$ has an equilibrium at the origin. *The model is essentially a linear one*.

Under "full detectability" (i.e., all state components that are not exponentially stable are fully observed), stability should be attainable.

Integrated velocity models II

If $S_1 = H_1^\mathsf{T} R^{-1} H_1 = sI$,

$$J_f(\widehat{X}_t) - P_t S = \begin{pmatrix} A_1 - sP_{t,11} & A_2 \\ -sP_{t,12}^\mathsf{T} & -J_g(\widehat{X}_{t,2}) \end{pmatrix}.$$

With large enough $P_{t,11}$, this matrix should be of negative logarithmic norm. *However, we have not been able to prove this*.

For the case $X_{t,1}, X_{t,2}, Y_t \in \mathbb{R}$, analysis is simpler and it can be shown that sufficient covariance inflation guarantees stability.

Conclusions

- Assumptions are very stringent and it is not easy to apply them to systems other than fully observed.
- Weakness primarily stems from uniform nature of the analysis. That is, all the stochastic terms difficult to analyse are "bounded away" from the filtering error SDE.
- Similar analysis can be carried out for discrete-time systems where $M(f - P_t S) < 0$ is replaced with

$$\sup_x \|J_f(x)\| \, \|I - K_k H\| < 1.$$

- Covering fully detectable systems is a desirable generalisation.

Thank you for your attention!