Automatica 94 (2018) 9–17

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

The suboptimal method via probabilists' Hermite polynomials to solve nonlinear filtering problems☆

Xue Luo [a], Stephen S.-T. Yau [b,*]

[a] School of Mathematics and Systems Science, Beihang University, Beijing, 100191, PR China
[b] Department of mathematical sciences, Tsinghua University, Beijing, 100084, PR China

ARTICLE INFO

Article history:
Received 30 May 2016
Received in revised form 26 November 2017
Accepted 16 March 2018

Keywords:
Estimation theory
Filtering
Suboptimal method
Carleman approach

ABSTRACT

In this paper, we shall investigate a novel suboptimal nonlinear filtering with augmented states via probabilists' Hermite polynomials (HP). The estimation of the original state can be extracted from the augmented one. Our method is motivated by the so-called Carleman approach (Germani et al., 2007). The novelty of our paper is to augment the original state with its probabilists' HPs, instead of its powers as in Carleman approach. Then we form a bilinear system of the first ν generalized Hermite polynomials (gHP) to yield the degree-ν approximation. We demonstrate that the neglect of the probabilists' gHPs with high degree is more reasonable by showing that the expectation of the HPs with degree n tends to zero, as n goes to infinity, if the density function belongs to certain function class. Moreover, we discuss the choice of the scaling and translating factors to yield better resolution. The benchmark example, 1d cubic sensor problem with zero initial condition, has been numerically solved by various methods, including the most widely used extended Kalman filter and particle filter. Our method with adaptive scaling factor outperforms the other methods in accuracy.

1. Introduction

The nonlinear filtering (NLF) problem has a long history back to 1960s, immediately after the discovery of the optimal estimation of linear filtering problem, the so-called Kalman filter (Kalman, 1960) and Kalman–Bucy filter (Kalman & Bucy, 1961). Many physical and engineering problems are naturally modeled by nonlinear stochastic dynamic systems. The general NLF problem can be expressed by the Itô stochastic differential equation (SDE):

$$\begin{cases} dx_t = f(x_t, t)dt + g(x_t, t)dw_t \\ dy_t = h(x_t, t)dt + dv_t, \end{cases} \tag{1.1}$$

where x_t is the state vector in $\mathbb{R}^n$, y_t is the observation in $\mathbb{R}^m$. $w_t \in \mathbb{R}^p$ and $v_t \in \mathbb{R}^m$ are independent standard Brownian motions with respect to an increasing family of σ-algebra, i.e. $\{\mathcal{F}_t, t \geq 0\}$. We further assume that $f : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^{n\times p}$ and $h : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^m$ are smooth nonlinear maps.

It is well known that in general the optimal estimation of NLF in the sense of minimal mean square error (MSE) cannot be obtained by finite many statistical quantities. The typical example is the cubic sensor problem (Hazewinkel, Marcus, & Sussmann, 1983), on which we shall investigate as a benchmark example in this paper for numerical comparison purpose. The finite-dimensional NLF problems have been studied in a series of papers by the second author in 1990s, see Wong and Yau (1999) and the references therein. Since 1966, Duncan (1967), Mortensen (1966) and Zakai (1969) have independently derived the SDE of the unnormalized conditional density function, which is so-called DMZ or Zakai's equation in the literature nowadays. To overcome the heavy computation, one has to use certain suboptimal implementable filtering algorithm (Ito, 1996). Some further improvements in this direction include the splitting-up method (Gyongy & Krylov, 2003), the S^3 algorithm (Lototsky, Mikulevicius, & Rozovskii, 1997), Yau–Yau's on- and off-line algorithm (Luo & Yau, 2013a, b; Yau & Yau, 2008), etc. Besides the DMZ equation, the most popular method is the particle filter (PF) (Arulampalam, Markell, Gordon, & Clapp, 2002), which is originated from Monte Carlo simulations. In the literature, one refers those methods which directly approximate the posterior conditional density function of the states as the global

☆ X. Luo acknowledges the support of National Natural Science Foundation of China (NSFC, grant no. 11501023). Both X. Luo and S. S.-T. Yau would like to thank the financial support from NSFC (grant no. 11471184). This work is also supported by S. S.-T. Yaus start-up fund from Tsinghua University. S. S.-T. Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done. The material in this paper was partially presented at the 8th International Congress on Industrial and Applied Mathematics (ICIAM2015), August 10–14, 2015, Beijing, P. R. China. This paper was recommended for publication in revised form by Associate Editor Martin Enqvist under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: xluo@buaa.edu.cn (X. Luo), yau@uic.edu (S.S.-T. Yau).

https://doi.org/10.1016/j.automatica.2018.04.004
0005-1098/

approaches, see exhaustive discussion in the survey paper of the first author (Luo, 2014).

Compared with the global approaches, the more handy and efficient ones are the so-called local approaches, such as the extended Kalman filter (EKF) (Gelb, 1984), Gaussian sum filter (Ito & Xiong, 2000), unscented Kalman filter (Julier & Uhlmann, 2004), ensemble Kalman filter (Evensen, 2003), etc. Recently, Germani, Manes, and Palumbo (2007) used the Carleman approximation to form a bilinear system (linear drift and multiplicative noise) of the appropriate augmented states. The suboptimal estimation of the bilinear SDE has been developed in Carravetta, Germani, and Shuakayev (2000). The suboptimal linear estimate for continuous–discrete bilinear system has been studied in Chen, Luo, and Yau (2017). To form a closed system, one has to ignore all the higher moments. The moments of degree greater than $\nu+1$ are ignored to yield the νth approximation. Intuitively, the larger ν one chooses, the more moments one keeps, and the more accurate estimation should be obtained. However, it may be inappropriate to do so for most NLF problems. Actually, as early as in 1967, it is Kushner (1967) who considered the moment sequences (see further discussions in Akhiezer & Krein, 1962), which states that even for the one-dimensional random variable, the moment sequence has to satisfy the following inequalities:

$$m_2 > 0,\ m_4 > m_2^2,\ m_6 > m_4^2/m_2, \ldots,$$

where m_s, $s = 1, 2, \ldots$, represent the s-moment of the random variable. In particular, in the case that the standard Gaussian distribution is obeyed, all the higher moments of this random variable can be computed explicitly:

$$m_s = \begin{cases} 0, & \text{all odd } s \geq 1 \\ (s-1)!!\, m_2^s, & \text{all even } s \geq 2, \end{cases} \tag{1.2}$$

where $(s-1)!! = 1 \cdot 3 \cdot 5 \cdots (s-1)$. It is easy to see that no matter how small m_2 is, the even moments grow without bound as $s \to \infty$. Therefore, it is inappropriate to let all the higher moments to be zero. In Luo, Jiao, Chiou, and Yau (2015, 2016), the authors considered to augment the original states with their central moments instead.

In this paper, we shall focus on the system of one dimensional state, i.e. $x \in \mathbb{R}$, and improve the Carleman approach by augmenting the original states via its gHPs $\left\{He_j^{\alpha,\beta}(x)\right\}_{j=0}^{\infty}$ (defined in (2.3)). The new state is defined as

$$He_{1:\nu}^{\alpha,\beta}(x_t) = \left[He_1^{\alpha,\beta}(x_t)\, He_2^{\alpha,\beta}(x_t)\ \cdots\ He_\nu^{\alpha,\beta}(x_t)\right]^T.$$

We derive the evolution system of $He_{1:\nu}^{\alpha,\beta}(x_t)$ and obtain the estimation of $He_{1:\nu}^{\alpha,\beta}(x_t)$, instead of the original state HPs x_t. It can be shown that $\mathcal{E}\left[He_j^{\alpha,\beta}(\xi)\right]$ tends to zero, as $j \to \infty$, if α, β are chosen properly, when the density function of the random variable ξ belongs to certain class of functions. Thus, compared with the Carleman approach, we believe that it is more appropriate by letting all $He_j^{\alpha,\beta}(\xi) \equiv 0, j > \nu$ to be zero in our degree-ν approximation, and our method can yield more accurate estimations. Furthermore, we have two more parameters, the scaling factor α and the translating factor β, to be tuned to yield even better results.

Theoretically, this algorithm can also be applied to the high-dimensional state, by augmenting the original state $x_t \in \mathbb{R}^d$ with its Kronecker product of HPs:

$$\mathbf{He}_{1:\nu}^{\alpha,\beta} = \otimes_{i=1}^{d} He_{1:\nu,i}^{\alpha_i,\beta_i},$$

where $He_{1:\nu,i}^{\alpha_i,\beta_i} = \left[He_1^{\alpha_i,\beta_i}(x_i)\, He_2^{\alpha_i,\beta_i}(x_i)\ \cdots\ He_{\nu_i}^{\alpha_i,\beta_i}(x_i)\right]^T$, $i = 1, 2, \ldots, d$. However, due to the fact that the product of two HPs is no longer a HP, see (2.7), the matrices (3.27)–(3.30) may not be evaluated efficiently, compared to the Carleman approach, where the Kronecker algebra serves as a powerful tool.

Our paper is organized as follows. In Section 2, we state some facts of the gHPs. In Section 3, we formulate our method of degree-ν in detail, and we show the fact that the expectation of the HPs in random variable ξ tends to zero, as the degree approaches infinity, when the density function of ξ belongs to the exponential decay class. Section 4 is devoted to the numerical experiments. The cubic sensor problem has been solved numerically as a benchmark example by different methods. The conclusions are drawn in the last section.

2. Preliminaries

In this section, we shall give some basic and useful facts of the generalized probabilists' Hermite polynomials (gHP). We call

$$He_n^{\alpha,\beta}(x) = He_n\left(\alpha(x-\beta)\right) \tag{2.3}$$

the gHP, where $\{He_n(x)\}_{n=0}^{\infty}$ are the Hermite polynomials (HP), $\alpha > 0$ is the scaling factor and $\beta \in \mathbb{R}$ is the translating factor. The following properties hold:

$$\left(He_n^{\alpha,\beta}(x)\right)' = \alpha n\, He_{n-1}^{\alpha,\beta}(x), \tag{2.4}$$

$$He_{n+1}^{\alpha,\beta}(x) = \alpha(x-\beta)He_n^{\alpha,\beta}(x) - n\, He_{n-1}^{\alpha,\beta}(x), \tag{2.5}$$

and

$$\int_{\mathbb{R}} He_n^{\alpha,\beta}(x)He_m^{\alpha,\beta}(x)w_{\alpha,\beta}(x)dx = \sqrt{2\pi}\, n!\delta_{nm}, \tag{2.6}$$

where $w_{\alpha,\beta}(x) = e^{-\frac{\alpha^2(x-\beta)^2}{2}}$. For any nonnegative integers n and m, we have

$$He_n^{\alpha,\beta}(x)He_m^{\alpha,\beta}(x) = \sum_{p \leq n\wedge m} A_{n,m,p} He_{n+m-2p}^{\alpha,\beta}(x), \tag{2.7}$$

where

$$A_{n,m,p} = \frac{n!m!}{p!(n-p)!(m-p)!}. \tag{2.8}$$

The next lemma is the key observation, which explains why we choose to augment the original state by its gHPs. Essentially, it says that $\mathcal{E}\left(He_j^{\alpha,\beta}(\xi)\right) \to 0$, as $j \to \infty$, with properly chosen α and β, if the density function of the random variable ξ obeys Gaussian distribution. Similar statement and proof can be found in Lemma 2.3, Luo (2006).

Lemma 1. *Suppose that the random variable $\xi \sim \mathcal{N}(a, b^2)$. Then for any $\mu \in \mathbb{R}$, we have*

$$\mathcal{E}\left[He_n^{1/b,a}(\xi+\mu)\right] = \left(\frac{\mu}{b}\right)^n.$$

In particular, for any $|\mu| < b$,

$$\lim_{n\to\infty} \mathcal{E}\left[He_n^{1/b,a}(\xi+\mu)\right] = 0. \tag{2.9}$$

The above lemma suggests that if the appropriate scaling factor $\frac{1}{b}$ is chosen according to μ, then the expectation of HPs can be arbitrarily small, provided the order of the polynomial is high enough. That is to say, the higher order terms are negligible in some sense.

3. Degree-ν approximation via gHPs

In this section, we shall derive the degree-ν suboptimal method via gHPs. We begin with showing that the expectation of the gHP tends to zero as the degree goes to infinity, if the density function of the state belongs to the exponential decay class (see Definition 1).

3.1. Functions of exponential decay class

Let us define the so-called functions of exponential decay and show that if the density function of the random variable ξ is in this class, the similar result as (2.9) also holds.

For any function $f(x) \in L^2(\mathbb{R})$, it can be represented as

$$f(x) = \sum_{i=0}^{\infty} \hat{f}_i H_i^{\alpha,\beta}(x), \tag{3.10}$$

where $H_i^{\alpha,\beta}(x)$ are the generalized Hermite function

$$H_i^{\alpha,\beta}(x) := \frac{1}{\sqrt{2\pi}\, i!}\, He_i^{\alpha,\beta}(x) e^{-\frac{\alpha^2(x-\beta)^2}{2}},$$

and the so-called Fourier–Hermite coefficients are

$$\hat{f}_i \overset{(2.6)}{:=} \int_{\mathbb{R}} f(x)\, He_i^{\alpha,\beta}(x) dx. \tag{3.11}$$

Definition 1. We say the density function $p(x) \in L^2(\mathbb{R})$ belongs to the **exponential decay class** with respect to (α, β), if for any $|\mu| < \frac{1}{\alpha}$, there exists some constant $C > 0$ and $\eta \in (0, 1 - \alpha|\mu|)$, such that

$$|\hat{p}_i| \le C\eta^i, \tag{3.12}$$

where $\hat{p}_i$, $i = 0, 1, 2, \ldots$, are Fourier–Hermite coefficients of $p(x)$.

Remark 1.

(1) It is not hard to see that the exponential decay class is non-empty, since the standard Gaussian belongs to this class with respect to $(1, 0)$, due to the fact that $\hat{p}_0 = 1$ and $\hat{p}_i = 0$, for all $i \ge 1$. Actually, there are infinitely many functions in this class. For instance, all $H_i^{\alpha,\beta}(x)$, $i = 1, 2, \ldots$, belong to the exponential decay class with respect to (α, β), since

$$\widehat{(H_i^{\alpha,\beta})}_j = \begin{cases} 1, & j = i \\ 0, & \text{otherwise.} \end{cases}$$

(2) Condition (3.12) is not so restrictive as it seems to be. Similar as Riemann–Lebesgue lemma claims, the coefficients approach to zero rapidly, provided the function is smooth enough. In Boyd (1984), Boyd claimed that if $f(z)$ is an entire function in complex plane and decays as super-Gaussian at $\pm\infty$ on the real-axis, i.e. $f(x) \sim \mathcal{O}(e^{-C|x|^k})$, $x \in \mathbb{R}$, for some $k > 2$, then the convergence rate of the Nth Hermite coefficient (3.11) is faster than the order $\mathcal{O}(e^{-CN^r})$, for some constant $C > 0$ and $r \le \frac{k}{2(k-1)}$.

Theorem 1. *Given $\alpha > 0$, $\beta \in \mathbb{R}$. If random variable ξ has density function $p(x) \in L^2(\mathbb{R})$ and belongs to the exponential decay class with respect to (α, β), then for any $|\mu| < \frac{1}{\alpha}$, we have*

$$\lim_{n\to\infty} \left|\mathcal{E}\left[He_n^{\alpha,\beta}(\xi + \mu)\right]\right| = 0. \tag{3.13}$$

Proof. With similar computation in Lemma 2.3, Luo (2006), we have

$$\begin{aligned} &\mathcal{E}\left[He_n^{\alpha,\beta}(\xi + \mu)\right] \\ &= \sum_{i=0}^{n} \frac{(\alpha\mu)^i}{i!} \frac{n!}{(n-i)!} \mathcal{E}\left[He_{n-i}^{1,0}(\alpha(\xi - \beta))\right] \\ &= \sum_{i=0}^{n} \frac{(\alpha\mu)^i}{i!} \frac{n!}{(n-i)!} \hat{p}_{n-i}, \end{aligned} \tag{3.14}$$

where the last equality follows from the fact that

$$\mathcal{E}\left[He_k^{\alpha,\beta}(\xi)\right] \overset{(3.10)}{=} \int_{\mathbb{R}} He_k^{\alpha,\beta}(x) \sum_{i=0}^{\infty} \hat{p}_i H_i^{\alpha,\beta}(x) dx = \hat{p}_k.$$

Since the density function $p(x)$ belongs to the exponential decay class with respect to (α, β), we have for any $|\mu| < \frac{1}{\alpha}$, there exists some constant $C > 0$ and $\eta \in (0, 1 - \alpha|\mu|)$ such that (3.12) holds. Hence, we have

$$\begin{aligned} \left|\mathcal{E}\left[He_n^{\alpha,\beta}(\xi + \mu)\right]\right| &\overset{(3.14)}{\le} C \sum_{i=0}^{n} \frac{(\alpha|\mu|)^i}{i!} \frac{n!}{(n-i)!} \eta^{n-i} \\ &= C(\alpha|\mu| + \eta)^n. \end{aligned} \tag{3.15}$$

Eq. (3.13) follows immediately by taking limit on both sides of (3.15) and the fact that $\eta \in (0, 1 - \alpha|\mu|)$. □

3.2. Derivation of degree-ν approximation for 1-d NLF problems

The basic idea of our method is to augment the original state via its gHPs. The augmented states satisfy an infinite-dimensional bilinear system. To form a closed system, we ignore all the HPs of degree greater than ν.

It is well known that $\{He_k(x)\}_{k=0}^{\infty}$ form an orthogonal basis of $L^2(\mathbb{R})$, so do $\{He_k^{\alpha,\beta}(x)\}_{k=0}^{\infty}$, for given $\alpha > 0$ and $\beta \in \mathbb{R}$. Assume that f, g and $h \in C^{\infty}\left([0, T]; L^2(\mathbb{R})\right)$. One can always expand these functions with respect to the gHPs:

$$\circ(x, t) = \sum_{k=0}^{\infty} \circ_k^{\alpha,\beta}(t) He_k^{\alpha,\beta}(x), \tag{3.16}$$

where $\circ_k^{\alpha,\beta}$ are smooth functions of t, which can be computed by

$$\circ_k^{\alpha,\beta}(t) = \frac{1}{\sqrt{2\pi}\, n!} \int_{\mathbb{R}} \circ(x, t)\, He_k{}^{\alpha,\beta}(x) e^{-\frac{\alpha^2(x-\beta)^2}{2}} dx, \tag{3.17}$$

where $\circ$ represents f, g and h. For conciseness of notation, we will drop the superscript α, β in $f_k^{\alpha,\beta}, g_k^{\alpha,\beta}$ and $h_k^{\alpha,\beta}$ in the sequel, if there is no confusion.

According to Itô lemma (Jazwinski, 1970), we have

$$\begin{aligned} &d\, He_j^{\alpha,\beta}(x_t) \\ &\overset{(1.1),(2.4)}{=} \left[\alpha j\, He_{j-1}^{\alpha,\beta}(x_t) f + \frac{1}{2}\alpha^2 j(j-1)\, He_{j-2}^{\alpha,\beta}(x_t)(gQg)\right] dt \\ &\quad + \alpha j\, He_{j-1}^{\alpha,\beta}(x_t) g dw_t \\ &\overset{(3.16)}{=} \alpha j \sum_{k=0}^{\infty} f_k\, He_{j-1}^{\alpha,\beta}(x_t) He_k^{\alpha,\beta}(x_t) dt \\ &\quad + \frac{1}{2}\alpha^2 j(j-1) Q \\ &\qquad \cdot \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} g_i g_k\, He_{j-2}^{\alpha,\beta}(x_t) He_i^{\alpha,\beta}(x_t) He_k^{\alpha,\beta}(x_t) dt \\ &\quad + \alpha j \sum_{k=0}^{\infty} g_k\, He_{j-1}^{\alpha,\beta}(x_t) He_k^{\alpha,\beta}(x_t) dw_t \\ &=: \mathrm{I} + \mathrm{II} + \mathrm{III}, \end{aligned} \tag{3.18}$$

for $j = 0, 1, 2, \ldots$, with the convention that $He_j^{\alpha,\beta}(x_t) \equiv 0$, when $j < 0$.

In the proposition below, we shall merge the double summation into one by reordering.

Proposition 1. *For any integer $j \geq 0$, (3.18) can be rewritten as*

$$\begin{aligned} & d\,He_j^{\alpha,\beta}(x_t) \\ &= \alpha j \sum_{l=-j+1}^{\infty} \mathcal{B}_{j-1,l}(f)\, He_{j-1+l}^{\alpha,\beta}(x_t)dt \\ &\quad + \frac{1}{2}\alpha^2 j(j-1)Q \sum_{r=-j+2}^{\infty} \mathcal{C}_{j,r}(g)\, He_{j-2+r}^{\alpha,\beta}(x_t)dt \\ &\quad + \alpha j \sum_{l=-j+1}^{\infty} \mathcal{B}_{j-1,l}(g)\, He_{j-1+l}^{\alpha,\beta}(x_t)dw_t, \end{aligned} \tag{3.19}$$

where

$$\mathcal{C}_{j,r}(g) = \sum_{l=-(j-2)}^{\infty} \mathcal{B}_{j-2,l}(g)\mathcal{B}_{j-2+l,r-l}(g), \tag{3.20}$$

and

$$\mathcal{B}_{j,l}(\circ) = \begin{cases} \sum_{p=-l}^{j} (\circ)_{l+2p} A_{j,l+2p,p}, & \text{if } -j \leq l \leq 0 \\ \sum_{p=0}^{j} (\circ)_{l+2p} A_{j,l+2p,p}, & \text{if } 1 \leq l, \end{cases} \tag{3.21}$$

where $(\circ)_k$ and $A_{n,m,p}$ are defined in (3.17) and (2.8), respectively. With convention, $\mathcal{B}_{j,l}(\circ) = 0$, if $j < 0$.

Proof. Term I and term III on the right-hand side of (3.18) can be written as

$$\begin{aligned} \mathrm{I} &\overset{(2.7)}{=} \alpha j \sum_{k=0}^{\infty} f_k \sum_{p \leq k \wedge (j-1)} A_{j-1,k,p}\, He_{k+j-1-2p}^{\alpha,\beta}(x_t)dt \\ &= \alpha j \sum_{l=-(j-1)}^{\infty} \mathcal{B}_{j-1,l}(f) He_{j-1+l}^{\alpha,\beta}(x_t)dt, \end{aligned} \tag{3.22}$$

where $\mathcal{B}_{j,l}(f)$ is in (3.21). Similarly, we have

$$\mathrm{III} = \alpha j \sum_{l=-(j-1)}^{\infty} \mathcal{B}_{j-1,l}(g)\, He_{j-1+l}^{\alpha,\beta}(x_t)dw_t. \tag{3.23}$$

By re-ordering the summations twice, term II becomes

$$\begin{aligned} \mathrm{II} &\overset{(2.7)}{=} \frac{1}{2}\alpha^2 j(j-1)Q \\ &\quad \cdot \sum_{k=0}^{\infty} g_k \sum_{l=-(j-2)}^{\infty} \mathcal{B}_{j-2,l}(g)\, He_{j-2+l}^{\alpha,\beta}(x_t) He_k^{\alpha,\beta}(x_t)dt \\ &= \frac{1}{2}\alpha^2 j(j-1)Q \sum_{l=-(j-2)}^{\infty} \mathcal{B}_{j-2,l}(g) \\ &\qquad \sum_{s=-(j-2+l)}^{\infty} \mathcal{B}_{j-2+l,s}(g)\, He_{j-2+l+s}^{\alpha,\beta}(x_t)dt \\ &= \frac{1}{2}\alpha^2 j(j-1)Q \sum_{l=-(j-2)}^{\infty} \sum_{r=-(j-2)}^{\infty} \\ &\qquad \mathcal{B}_{j-2,l}(g)\mathcal{B}_{j-2+l,r-l}(g)\, He_{j-2+r}^{\alpha,\beta}(x_t)dt \\ &= \frac{1}{2}\alpha^2 j(j-1)Q \sum_{r=-(j-2)}^{\infty} \mathcal{C}_{j,r}(g)\, He_{j-2+r}^{\alpha,\beta}(x_t)dt, \end{aligned} \tag{3.24}$$

by letting $r = l + s$, where $\mathcal{C}_{j,r}(g)$ is in (3.20). Combining (3.22)–(3.24), (3.19) follows directly. □

It is clear to see from (3.19) that $d\,He_j^{\alpha,\beta}(x_t)$, $j \geq 1$ forms an infinite-dimensional system, which cannot be solved, unless certain approximation is used. Let us introduce the degree-ν approximation (for any $\nu \geq 2$) by keeping only the first ν equations in (3.19).

Let us denote

$$He_{1:\nu}^{\alpha,\beta}(x_t) = \left[He_1^{\alpha,\beta}(x_t)\, He_2^{\alpha,\beta}(x_t)\, \cdots\, He_\nu^{\alpha,\beta}(x_t)\right]^T. \tag{3.25}$$

From (3.19) and the degree-ν approximation, $He_{1:\nu}^{\alpha,\beta}(x_t)$ satisfies the following bilinear system:

$$\begin{cases} d\,He_{1:\nu}^{\alpha,\beta}(x_t) = \left(\mathbf{F}_\nu He_{1:\nu}^{\alpha,\beta}(x_t) + \mathbf{F}_{0,\nu}\right)dt \\ \qquad\qquad + \left(\mathbf{G}_\nu He_{1:\nu}^{\alpha,\beta}(x_t) + \mathbf{G}_{0,\nu}\right)dw_t \quad , \\ dy_t = \left(H_\nu He_{1:\nu}^{\alpha,\beta}(x_t) + H_0\right)dt + dv_t \end{cases} \tag{3.26}$$

where

$$\begin{aligned} \mathbf{F}_\nu(j,:) &= \alpha j \mathcal{B}_{j-1,-(j-1)+1:\nu-(j-1)}(f) \\ &\quad + \frac{1}{2}\alpha^2 j(j-1)Q\mathcal{C}_{j,-(j-2)+1:\nu-(j-2)}(g), \end{aligned} \tag{3.27}$$

$$\begin{aligned} \mathbf{F}_{0,\nu}(j) &= \alpha j \mathcal{B}_{j-1,-j+1}(f) \\ &\quad + \frac{1}{2}\alpha^2 j(j-1)Q\mathcal{C}_{j,-j+2}(g), \end{aligned} \tag{3.28}$$

$$\mathbf{G}_\nu(j,:) = \alpha j \mathcal{B}_{j-1,-(j-1)+1:\nu-(j-1)}(g), \tag{3.29}$$

$$\mathbf{G}_{0,\nu}(j) = \alpha j \mathcal{B}_{j-1,-(j-1)}(g), \tag{3.30}$$

for $1 \leq j \leq \nu$, and

$$H_\nu = [h_1\, h_2\, \cdots\, h_\nu], \quad H_0 = h_0, \tag{3.31}$$

with h_k in (3.16). The notations $*(j,:)$ and $(*)_{j,a:b}$ represent the jth row of the matrix $*$ and the a to b's column in jth row of matrix $*$, respectively.

It is clear to see that (3.26) is a bilinear system of $(\nu+1)$ equations. The suboptimal estimation for bilinear system has been investigated in Carravetta et al. (2000). For the sake of clarity, we state the theorem below and its proof is given in Theorem 4.4, Carravetta et al. (2000).

Theorem 2. *For any given $\alpha > 0$ and $\beta \in \mathbb{R}$. Let us denote the optimal estimate conditioned on the observation history of $He_{1:\nu}^{\alpha,\beta}(x_t)$ as $\widehat{He_{1:\nu}^{\alpha,\beta}}(x_t) := \mathcal{E}\left(He_{1:\nu}^{\alpha,\beta}(x_t)\middle|\mathcal{F}_t\right)$, satisfies the equation*

$$\begin{aligned} & d\widehat{He_{1:\nu}^{\alpha,\beta}}(x_t) \\ &= \left(\mathbf{F}_\nu \widehat{He_{1:\nu}^{\alpha,\beta}}(x_t) + \mathbf{F}_{0,\nu}\right)dt \\ &\quad + \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu} + \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T\right) \\ &\quad \cdot R^{-1}\left[dy - \left(H_\nu \widehat{He_{1:\nu}^{\alpha,\beta}}(x_t) + H_0\right)dt\right], \end{aligned} \tag{3.32}$$

where $m_\nu^{\alpha,\beta} := \mathcal{E}\left(He_{1:\nu}^{\alpha,\beta}(x_t)\right)$ satisfying the following equations

$$\dot{m}_\nu^{\alpha,\beta}(t) = \mathbf{F}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{F}_{0,\nu} \tag{3.33}$$

with the initial values $m_\nu^{\alpha,\beta}(0) = \mathcal{E}\left(He_{1:\nu}^{\alpha,\beta}(x_0)\right)$, and $\mathbf{P}_\nu^{\alpha,\beta}(t)$ is the conditional error covariance matrix

$$\begin{aligned} \mathbf{P}_\nu^{\alpha,\beta}(t) = \mathcal{E}\Bigg[& \left(He_{1:\nu}^{\alpha,\beta}(x_t) - \widehat{He_{1:\nu}^{\alpha,\beta}}(x_t)\right) \\ & \cdot \left(He_{1:\nu}^{\alpha,\beta}(x_t) - \widehat{He_{1:\nu}^{\alpha,\beta}}(x_t)\right)^T \Bigg| \mathcal{F}_t \Bigg] \end{aligned}$$

evolving according to the equation

$$\begin{aligned} \dot{\mathbf{P}}_\nu^{\alpha,\beta}(t) &= \mathbf{F}_\nu \mathbf{P}_\nu^{\alpha,\beta}(t) + \mathbf{P}_\nu^{\alpha,\beta}(t)\mathbf{F}_\nu^T + Q(t) \\ &\quad - \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu} + \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T\right) \\ &\quad \cdot R^{-1}\left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu} + \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T\right)^T, \end{aligned} \tag{3.34}$$

with $\mathbf{P}_\nu^{\alpha,\beta}(0) = \Psi_\nu^{\alpha,\beta}(0)$.

Remark 2. For $\nu = 1$, (3.32)–(3.34) coincide with the extended Kalman–Bucy filter (EKBF). However, the degree-2 approximation in (3.32)–(3.34) is different from the so-called second order EKBF. In the second order EKBF, a second order Taylor approximation is used in the state process (Gelb, 1984), while both in our method and the degree-2 Carleman approximation (Germani, Manes, & Palumbo, 2005; Germani et al., 2007), the second order state increments are substituted with the components of the error covariance matrix provided by the Riccati equation (3.34).

The estimation of the augmented states can be used to construct that of the original state, due to the fact that

$$\alpha(x_t - \beta) = [1\ 0_{\nu-1}] \cdot He^{\alpha,\beta}_{1:\nu}(x_t), \tag{3.35}$$

where $[1\ 0_{\nu-1}]$ is a row vector of size $1 \times \nu$ with all its elements 0, except the first one being 1. That is, if the estimation $\widehat{He^{\alpha,\beta}_{1:\nu}(x_t)}$ of the extended state $He^{\alpha,\beta}_{1:\nu}(x_t)$ is obtained, then $\hat{x}_t = \beta + \widehat{He^{\alpha,\beta}_{1:\nu}}(1)/\alpha$ is the estimation of the original state, where $\widehat{He^{\alpha,\beta}_{1:\nu}}(1)$ represents the first component of the vector $\widehat{He^{\alpha,\beta}_{1:\nu}}$.

3.3. Discussion on the choice of scaling and translating factor

We shall solve (3.33), (3.32) and (3.34) numerically by Euler forward scheme in the total experimental time $[0, T]$. The estimations are obtained at equi-distant discrete time t_k, $k = 0, 1, \ldots, K$, where $t_0 = 0$, $t_K = T$. Let us denote the new state $X_k := He^{\alpha_k,\beta_k}_{1:\nu}(x_t)$ in each time interval $[t_k, t_{k+1}]$, $k = 0, 1, \ldots, K$.

As we mentioned in (3.35), $\hat{x}_{t_k}$ can be obtained by $\hat{X}_{k-1}\big|_{t_k}$. That is,

$$\hat{x}_{t_k} = \frac{\hat{X}_{k-1}\big|_{t_k}(1)}{\alpha_{k-1}} + \beta_{k-1}, \tag{3.36}$$

where $\hat{X}_{k-1}\big|_{t_k}(1)$ represents the first component of $\hat{X}_{k-1}$ at time t_k. At the beginning of the time interval $[t_k, t_{k+1}]$, we have

$$\hat{X}_k\big|_{t_k}(1) = \frac{\alpha_k}{\alpha_{k-1}} \hat{X}_{k-1}\big|_{t_k}(1) + \alpha_k(\beta_{k-1} - \beta_k), \tag{3.37}$$

which follows from (3.36) and $\hat{x}_{t_k} = \frac{\hat{X}_k|_{t_k}(1)}{\alpha_k} + \beta_k$ also holds. We suggest to let $\beta_k = \hat{x}_{t_k}$. The reason is from the following: if $|\hat{x}_{t_k}| > 1$ and $x_{t_k} \sim \mathcal{N}(\hat{x}_{t_k}, 1)$, then

$$\begin{aligned}\mathcal{E}\left[He^{1,0}_n(x_{t_k})\right] &= \mathcal{E}\left[He^{1,0}_n\left((x_{t_k} - \hat{x}_{t_k}) + \hat{x}_{t_k}\right)\right] \\ &= \hat{x}^n_{t_k} \nrightarrow 0,\end{aligned}$$

as $n \to \infty$, where the last equality follows by Lemma 1. However, if we let $\beta_k = \hat{x}_{t_k}$ at t_k, then

$$\mathcal{E}\left[He^{1,\beta_k}_n(x_{t_k})\right] = \mathcal{E}\left[He^{1,0}_n(x_{t_k} - \hat{x}_{t_k})\right] = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Intuitively, even if $x_{t_k} \nsim \mathcal{N}(\hat{x}_{t_k}, 1)$, as long as its density is not far from $\mathcal{N}(\hat{x}_{t_k}, 1)$, it may still hold that $\mathcal{E}\left[He^{1,\beta_k}_n(x_{t_k})\right] \to 0$, as $n \to \infty$.

Next, we discuss the choice of the scaling factor. Although it seems a good idea to change the scaling factor at every time step, it needs to update the matrices $\mathbf{F}_\nu$, $\mathbf{F}_{0,\nu}$, $\mathbf{G}_\nu$ and $\mathbf{G}_{0,\nu}$ in (3.33), (3.32) and (3.34), and H, H_0 in (3.31).

Let us compare the computational complexity with that of Carleman approximation. Suppose we have obtained the bilinear system in Carleman approach:

$$dX_{\nu,ex} = A_\nu(\alpha, \beta)X_{\nu,ex}dt + B_\nu(\alpha, \beta)X_{\nu,ex}dw_t,$$

and the one-to-one transformation between gHPs and the powers

$$He^{\alpha,\beta}_{0:\nu} = P_\nu X_{\nu,ex},$$

where $X_{\nu,ex} = \begin{bmatrix} 1 \\ \alpha(x-\beta) \\ \vdots \\ \alpha^\nu(x-\beta)^\nu \end{bmatrix}$ and $He^{\alpha,\beta}_{0:\nu} = \begin{bmatrix} 1 \\ He^{\alpha,\beta}_1 \\ \vdots \\ He^{\alpha,\beta}_\nu \end{bmatrix}$ as defined in (3.25), then it can be easily derived that

$$\begin{aligned}dHe^{\alpha,\beta}_{0:\nu} &= P_\nu A_\nu(\alpha, \beta)P_\nu^{-1}He^{\alpha,\beta}_{0:\nu}dt \\ &\quad + P_\nu B_\nu(\alpha, \beta)P_\nu^{-1}He^{\alpha,\beta}_{0:\nu}dw_t.\end{aligned} \tag{3.38}$$

The only extra computation in our algorithm is the multiplication of the transformation matrix P_ν and its inverse in the evolution system (3.38).

We propose to set a threshold value so that the scaling factor is only altered if necessary. It is called the adaptive scaling factor technique in the sequel. Notice that the choice of the scaling factor α_k in X_k is closely related to the covariance, so does $X_{k-1}|_{t_k}(2) = \alpha^2_{k-1}(x_{t_k} - \beta_{k-1})^2 - 1$, where $\beta_{k-1} = \hat{x}_{t_{k-1}}$. Therefore, we suggest that if

$$\left|X_{k-1}|_{t_k}(2)\right| > C_{threshold} \times \alpha^2_{k-1}, \tag{3.39}$$

then let $\alpha_k = \sqrt{1 + 0.5 \times \left|X_{k-1}|_{t_k}(2)\right|}$, where adding 1 in the previous expression is to avoid singularity of the covariance.[1] Numerical investigation on the choice of $C_{threshold}$ is in Section 4.1. We suggest to set $C_{threshold} \in (0, 1.2]$. In a sum, the initial data at the time interval $[t_k, t_{k+1}]$ for (3.33), (3.34) and (3.32) are

$$\begin{aligned} m_{\nu,k}\big|_{t_k} &= \frac{\alpha_k}{\alpha_{k-1}} m_{\nu,k-1}\big|_{t_k}, \\ \mathbf{P}^{\alpha_k,\beta_k}_{\nu,k} &= X_k|_{t_k} \cdot X_k|^T_{t_k}, \end{aligned}$$

and

$$\begin{aligned} X_k|_{t_k} &= \frac{\alpha_k}{\alpha_{k-1}} X_{k-1}|_{t_k} \\ &\quad + \alpha_{k-1}\left[He^{1,0}_{1:\nu}(\hat{x}_{k-1}) - He^{1,0}_{1:\nu}(\hat{x}_k)\right], \end{aligned}$$

respectively.

The algorithm with adaptive scaling factor has been summarized in Table 1.

4. Numerical simulations

In this section, we shall compare our algorithms with existing NLF methods, such as the most widely used EKF, PF and the Carleman approximation (Germani et al., 2005, 2007). To obtain better resolutions of the Carleman approach, we set $\bar{x} = \hat{x}_{t_k}$ in Lemma 2, Germani et al. (2007), in each time interval $[t_k, t_{k+1}]$, similar as the translating factor β in our algorithm.

In Hazewinkel et al. (1983), the cubic sensor problem has been shown to be essentially infinite-dimensional. That is, it cannot be solved exactly by any finite-dimensional statistical quantities. This problem is in the form

$$\begin{cases} dx_t = dw_t \\ dy_t = x_t^3 dt + dv_t, \end{cases} \tag{4.40}$$

where $\mathcal{E}(dw_t dw_t^T) = \mathcal{E}(dv_t dv_t^T) = 1$. The initial state x_0 has been chosen to be 0 and the initial covariance is 0.1. The true state is generated by the Euler–Maruyama method (Higham, 2001) in the time interval $[0, 5]$ with time discretization $\triangle t = 5 \times 10^{-4}$.

[1] The scalar 1 can be replaced by other positive numbers. Here, we choose 1, due to the fact that if $\left|X_{k-1}|_{t_k}(2)\right|$ is close to zero, then α_k is reset to some value close to 1, the same value as α_0.

Table 1
Algorithm with adaptive scaling factor.

Algorithm with adaptive scaling factor
Initiate $\alpha_0 = 1$ and $\beta_0 = x_0$.
Start with $m_\nu = X_0 = He^{\alpha,\beta}_{1;\nu}(x_0)$ and $\mathbf{P}_{\nu,0} = 0.1 \times I_\nu$, where I_ν is the identity matrix with size $\nu \times \nu$.
For $k = 0 : K, K =$ the total time steps in T.
− Solve (3.33), (3.32) and (3.34) using Euler forward method.
Denote the solution as X_k.
− The estimation of $\hat{x}_{k+1} = X_k\|_{t_{k+1}}(1)/\alpha_k + \beta_k$ is obtained, where $X_k\|_{t_{k+1}}(1)$ is the first component of $X_k\|_{t_{k+1}}$.
− If $\left\|X_k\|_{t_{k+1}}(2)\right\| > C_{threshold} \times \alpha_k^2$ $\alpha_{k+1} = \sqrt{1 + 0.5 \times \left\|X_k\|_{t_{k+1}}(2)\right\|}$ (△) else $\alpha_{k+1} = \alpha_k$ End
− Let $\beta_{k+1} = \hat{x}_{k+1}$.
− Update $m_{\nu,k+1} = \alpha_{k+1}/\alpha_k m_{\nu,k}$, $X_{k+1}\|_{t_{k+1}} = \alpha_{k+1}/\alpha_k \cdot X_k\|_{t_{k+1}} + \alpha_k \left[He^{1,0}_{1;\nu}(\hat{x}_{k-1}) - He^{1,0}_{1;\nu}(\hat{x}_k)\right]$, and $\mathbf{P}_{\nu,k+1} = X_k\|_{t_{k+1}} \cdot X_k\|^T_{t_{k+1}}$.
− Update the coefficients in $\mathbf{F}_\nu$, $\mathbf{F}_{0,\nu}$, $\mathbf{G}_\nu$ and $\mathbf{G}_{0,\nu}$ in (3.33), (3.32) and (3.34), and H, H_0 in (3.32), if necessary.
End

The simplest numerical scheme, Euler forward method, is used to solve ODEs in EKF, Carleman approach and our algorithms. The performance of each realization is evaluated by the MSE:

$$E_{MSE} := \frac{1}{K+1}\sum_{k=0}^{K}(\hat{x}_{t_k} - x_{t_k})^2, \tag{4.41}$$

where $K = T/\triangle t$ and $\hat{x}_{t_k}$ is the estimation obtained by various methods at t_k. The mean and standard deviation of MSE over N_{MC} Monte Carlo runs are given by

$$Mean_{MSE} := \frac{1}{N_{MC}}\sum_{i=1}^{N_{MC}} E_{MSE,i},$$
$$Std_{MSE} := \sqrt{\frac{1}{N_{MC}}\sum_{i=1}^{N_{MC}}(E_{MSE,i} - Mean_{MSE})^2}, \tag{4.42}$$

where $E_{MSE,i}$ is the MSE of the ith realization in the Monte Carlo runs.

4.1. The choice of $C_{threshold}$ in (3.39)

In this subsection, we take cubic sensor problem (4.40) as an example to test the choice of the threshold $C_{threshold}$ in (3.39). In Table 2, we list several important indicators, such as $Mean_{MSE}$, Std_{MSE} etc., versus different $C_{threshold}$. Sometimes the NaN in Matlab is obtained, in which case, we say this method fails or diverges in this realization and we record the number of failures of different methods out of 500 runs. In Section 4.3, we shall discuss the divergence phenomena in detail. When $C_{threshold} = 0.1, 0.5, 0.8$ and 1.2, the number of failures are either 0 or 1, and the mean and standard deviation of MSE over successful Monte Carlo runs are nearly the same. The smaller $C_{threshold}$ is, the more frequent the scaling factor changes, and the heavier computation it would be, if the truncation ν is large. In Table 2, with $C_{threshold} = 0.1$, the averaged changing times of α can be as many as 9300 out of $N = 10\,000$. So one expect as large $C_{threshold}$ as possible if the number of failures and the MSE keep small. It seems that our algorithm is robust at least within certain range of threshold, say $C_{threshold} \in (0, 1.2]$.

4.2. Comparison with different methods

We compared our algorithms with and without adaptive scaling factor with EKF, PF with 50 particles and Carleman approximation. The PF used in our experiment is the SIR algorithm, see Algorithm 4, Arulampalam et al. (2002). It is worth noting that there have been much progresses after (Arulampalam et al., 2002). The SIR algorithm is used only for comparison purpose. In both Carleman approach and our method, the cubic sensor problem reformulated as (3.26) with $\nu = 3$. All the methods have been run 500 times, where the realizations of the true states are generated by *randn('state',s)* with $s = 1$ to 500 in MatLab. The $Mean_{MSE}$, Std_{MSE}, CPU times and the number of failures within 500 runs are listed in Table 3. The averaged CPU times, $Mean_{MSE}$ and Std_{MSE} are computed within the successful runs. Our algorithm with adaptive scaling factor has the least failures. The EKF fails completely, since no matter what true states are, it always yields 0 as its estimation all the time, as long as it starts with 0. Recall the mechanics of EKF, the linearization at the estimation 0 gives $F = H = 0$, if the estimation is zero at the initial time. In this case, the Kalman gain is zero. Consequently, the optimal estimation is obtained without any innovation and keeps zero if $F = 0$. The more general situation is that the state estimation keeps constant (not necessarily to be zero), if the critical point of the nonlinear drift function f is the same as that of the observation function h, and the initial state estimation is exactly at the critical point. In other words, the state without drifting stayed at the critical point is not observable. PF gives as bad estimation as EKF, since their $Mean_{MSE}$ and Std_{MSE} are comparably large.

In Fig. 1, the absolute values of errors E_{AE} averaged over all successful Monte Carlo runs from different methods versus time are displayed in various colors.

$$E_{AE}(t_k) := \frac{1}{N_{MC}}\sum_{i=1}^{N_{MC}}\left|\hat{x}_{t_k} - x_{t_k}\right|, \tag{4.43}$$

$k = 1, \ldots, K$. It indicates that our algorithm with/without adaptive scaling factor yields the best averaged estimation. In Fig. 2, we display the E_{AE} versus time obtained by Carleman approach, our algorithm with fixed scaling factor $\alpha = 1$ and adaptive scaling factor with $C_{threshold} = 0.5$, when the true state is generated by *randn('state',130)*. It is clear to see that our algorithm with adaptive scaling factor is more stable with $E_{MSE} \approx 0.4620$, though the scaling factor has been changed 6768 times, while both the Carleman approach and our algorithm with fixed scaling factor fail completely around $t = 1.5$. Both Carleman approach and our algorithm with fixed scaling factor yield NaN in MatLab simultaneously for $\hat{x}_{t_k}$ at $k = 2924$.

Table 2
Different thresholds affect the performances of our degree-3 algorithm with adaptive scaling factor in the cubic sensor problem (4.40).

Threshold $C_{threshold}$ in (3.39)	♯ of failures in 500 runs	$Mean_{MSE}$	Std_{MSE}	Averaged changing times of α
0.1	1	0.4602	0.2744	9300
0.5	1	0.4503	0.2267	4867
0.8	1	0.4498	0.2690	435
1.2	0	0.4834	0.3669	265
2	45	0.6608	0.6802	159

Table 3
The number of failures, averaged CPU times, $Mean_{MSE}$ and Std_{MSE} from different methods within 500 Monte Carlo runs are listed.

	PF with 50 particles	EKF	Carleman	Our algorithm	
				With fixed $\alpha = 1$	With adaptive scaling factor
♯ of failures in 500 runs	0	500[a]	15	6	1
Averaged CPU times	6.0623	0.0042	1.6895	2.7397	2.7959
$Mean_{MSE}$	2.7942	2.4935	0.8892	0.4782	0.4503
Std_{MSE}	2.9497	2.7867	0.7942	0.2487	0.2267

[a] The estimation given by EKF is 0 for all realization of true states, provided the initial state is 0.

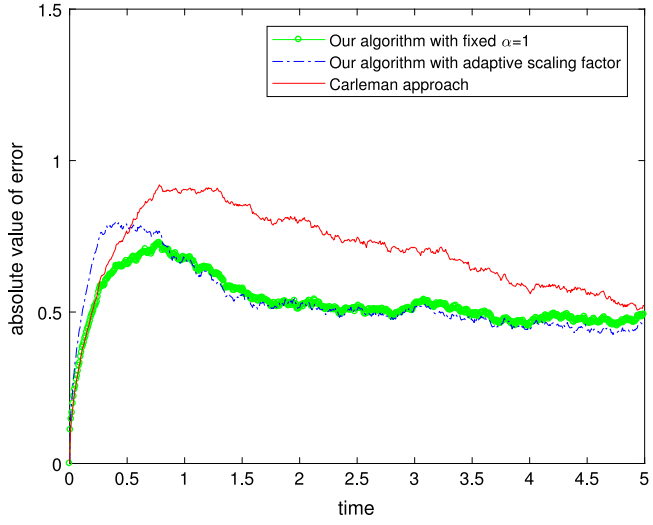

Fig. 1. The y-axis is the absolute value of error averaged over all successful Monte Carlo runs (within 500) from Carleman approach and our algorithm with and without adaptive scaling factor have been displayed, while the x-axis is the time.

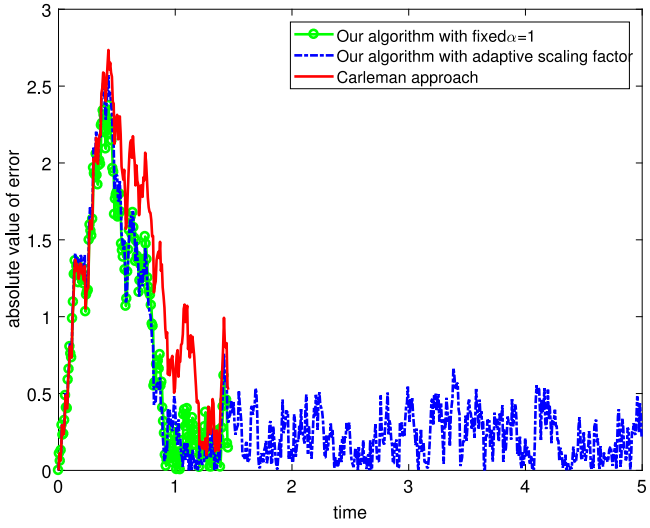

Fig. 2. The y-axis is the absolute value of error obtained by Carleman approach and our algorithm with fixed scaling factor $\alpha = 1$ are compared with our algorithm with adaptive scaling factor, when the realization is generated by *randn('state',130)*, while the x-axis is the time.

4.3. Discussions on divergence

In this subsection, we shall investigate the possible reason of divergence in the numerical experiment for cubic sensor problem. Motivated by Fitzgerald (1971), we examine the evolution equation of $\mathbf{P}_\nu^{\alpha,\beta}$ (3.34) and rewrite it in the following form:

$$\begin{aligned}\dot{\mathbf{P}}_\nu^{\alpha,\beta}(t) = &\left[\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu\right]\mathbf{P}_\nu^{\alpha,\beta}(t)\\ &+ \mathbf{P}_\nu^{\alpha,\beta}(t)\left[\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu\right]^T\\ &- \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T R^{-1}H_\nu\mathbf{P}_\nu^{\alpha,\beta}(t)\\ &+ Q(t) - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}(\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)^T. \qquad (4.44)\end{aligned}$$

As mentioned in Fitzgerald (1971), the Potter's "regularity" conditions

(1) No eigenvector of $\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu$ whose eigenvalue has a non-negative real part is a null vector of $R^{-1}H_\nu$.
(2) No eigenvector of $\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu$ whose eigenvalue has a non-negative real part is a null vector of $Q(t) - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}(\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)^T$.

guarantee the unique positive semidefinite critical point, i.e. $\dot{\mathbf{P}}_{\nu,s}^{\alpha,\beta} = \mathbf{0}$. However, in the cubic sensor problem it is easy to see that rank $\left(\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu\right) \leq 2$. According to MatLab, Potter's "regularity" condition (1) is violated, i.e. the eigenvector corresponding to the zero eigenvalue is a null vector of $R^{-1}H_\nu$. But fortunately, rank $\left(\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu\right) = 2$ and the other two eigenvalues are negative all the time until the divergence. The possible consequence of the violation is that the critical point $\mathbf{P}_{\nu,s}^{\alpha,\beta}(t)$ is not unique, and heavily depends on the initial condition $\mathbf{P}_\nu^{\alpha,\beta}(0)$. In fact, for arbitrary $r \in \mathbb{R}$, $\mathbf{P}_{\nu,s}^{\alpha,\beta} + r\mathbf{e}\mathbf{e}^T$ is always a steady state of (4.44), provided that $\left[\mathbf{F}_\nu - (\mathbf{G}_\nu m_\nu^{\alpha,\beta} + \mathbf{G}_0)R^{-1}H_\nu\right]\mathbf{e} = R^{-1}H_\nu\mathbf{e} = 0$. Thus, $\|\mathbf{P}_{\nu,s}^{\alpha,\beta} + r\mathbf{e}\mathbf{e}^T\|_\infty \to \infty$, as $r \to \infty$, where $\|\cdot\|_\infty$ denotes the L^∞ norm of the matrix. This undesirable fact may cause the divergence of the estimation. Let us rewrite (3.32) in the following form:

$$\begin{aligned}&\widehat{dHe_{1:\nu}^{\alpha,\beta}}(x_t)\\ &= \left[\mathbf{F}_\nu - \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu}\right)R^{-1}H_\nu\right]\widehat{He_{1:\nu}^{\alpha,\beta}}(x_t)dt\\ &\quad - \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T R^{-1}H_\nu\widehat{He_{1:\nu}^{\alpha,\beta}}(x_t)dt\\ &\quad + \left[\mathbf{F}_{\mathbf{0},\nu} - \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu} + \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T\right)R^{-1}H_0\right]dt\\ &\quad + \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu} + \mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T\right)R^{-1}dy. \qquad (4.45)\end{aligned}$$

The divergence may be caused by the last two terms on the right-hand side of (4.45), instead of the first two. Let us analyze term by term. The first term is under control since in the numerical experiment $[\mathbf{F}_\nu - \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu}\right)R^{-1}H_\nu]$ has one zero and two negative eigenvalues, and the unstable eigenvector is a null vector of $R^{-1}H_\nu$. The coefficient in front of $\widehat{He_{1:\nu}^{\alpha,\beta}}(x_t)$ in the second term is negative semi-definite all the time. Taking a close look at the terms $\mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T R^{-1}H_0$ and $\mathbf{P}_\nu^{\alpha,\beta}(t)H_\nu^T R^{-1}$ in the third and fourth term on the right-hand side of (4.45), let us replace $\mathbf{P}_\nu^{\alpha,\beta}$ with $\mathbf{P}_{\nu,s}^{\alpha,\beta} + r\mathbf{e}\mathbf{e}^T$. If without any round-off error in $\mathbf{e}$ and H_ν, we have $\mathbf{e}^T H_\nu^T R^{-1} = 0$. Unfortunately, with a little perturbation in $\mathbf{e}^T$ or H_ν, this term can go to infinity, as $r \to \infty$. This situation may happen highly dependent on the initial condition of $P_\nu^{\alpha,\beta}$.

Let us take a closer look at the possible cause of the divergence in the experiment in Fig. 2. To examine the explosion of various

Table 4
We list three important quantities in the analysis of divergence of various methods: the eigenvector $\mathbf{e}$ of the unstable mode, the examination of the null vector of $R^{-1}H_\nu$ and the covariance $\mathbf{P}_\nu^{\alpha,\beta}$.

Time		1.4575			1.4595		
Carleman approach	$\mathbf{e}$	1.3322e−17	1.2000e−01	−9.9277e−01	−8.9026e−17	5.3458e−01	−8.4511e−01
	$\mathbf{e}R^{-1}H_\nu$	2.2204e−16			−2.2204e−16		
	$\mathbf{P}_\nu^{\alpha,\beta}$	−4.8850e−18	−1.9296e−02	−1.4492e−01	0	−1.1490e+05	−8.7693e+05
		−1.9296e−02	1.7696e+00	1.2715e+01	−1.1490e+05	−2.6404e+13	−2.0152e+14
		−1.4492e−01	1.2715e+01	9.1358e+01	−8.7693e+05	−2.0152e+14	−1.5380e+15
Our algorithm with fixed $\alpha = 1$	$\mathbf{e}^T$	−1.1595e−16	1.3925e−01	−9.9026e−01	3.0285e−17	2.7278e−01	−9.6208e−01
	$\mathbf{e}R^{-1}H_\nu$	−3.3307e−16			2.2204e−16		
	$\mathbf{P}_\nu^{\alpha,\beta}$	1.9970e−04	−2.6916e−02	−1.6297e−01	−1.6544e+07	−6.1527e+14	−3.8488e+15
		−2.6916e−02	1.7025e+00	1.0042e+01	−6.1527e+14	−2.2882e+22	−1.4313e+23
		−1.6297e−01	1.0042e+01	5.9254e+01	−3.8488e+15	−1.4313e+23	−8.9537e+23
Our algorithm with adaptive scaling factor	α	5.1343			5.4750		
	$\mathbf{e}^T$	1.3732e−17	2.6039e−02	−9.9966e−01	1.7026e−18	2.4537e−02	−9.9970e−01
	$\mathbf{e}R^{-1}H_\nu$	2.6021e−18			8.6736e−19		
	$\mathbf{P}_\nu^{\alpha,\beta}$	1.1124e−30	3.0154e−14	1.3230e−13	5.0093e−31	−1.9718e−14	−8.2313e−14
		3.0154e−14	8.1737e+02	3.5862e+03	−1.9718e−14	7.7611e+02	3.2399e+03
		1.3230e−13	3.5862e+03	1.5734e+04	−8.2313e−14	3.2399e+03	1.3525e+04

methods in Fig. 2, we record three important quantities in Table 4 and have the following observations:

(1) The eigenvector $\mathbf{e}$ corresponding to the zero eigenvalue of $\mathbf{F}_\nu - \left(\mathbf{G}_\nu m_\nu^{\alpha,\beta}(t) + \mathbf{G}_{0,\nu}\right) R^{-1}H_\nu$ keeps almost the same until the explosion.
(2) The eigenvector $\mathbf{e}$ is approximately the null vector of $R^{-1}H_\nu$, yet there is always a round-off error. By tuning the scaling factor in our algorithm, $\mathbf{e}$ is more likely to reside in the null space of $R^{-1}H_\nu$.
(3) The covariance becomes larger and larger before explosion.

In other word, the choice of the initial value in (4.44) is crucial. The discussion on the choice of scaling and translating factors in Section 3.3 provides an effective way to avoid the explosion to some degree.

4.4. The performance with respect to the truncation ν

Intuitively, the larger truncation mode ν is, the more accurate the approximation $\hat{x}_t$ is. Instead of the cubic sensor problem (4.40), we numerically solve a slightly different one (4.46), since the cubic sensor problem gives exactly the same $\mathbf{F}_\nu$, $\mathbf{F}_{0,\nu}$, $\mathbf{G}_\nu$, $\mathbf{G}_{0,\nu}$, H_ν and H_0 in (4.40) for all $\nu \geq 3$.

Let us consider

$$\begin{cases} dx_t = x_t^2 dt + dw_t \\ dy_t = x_t^3 dt + dv_t, \end{cases} \tag{4.46}$$

where $\mathcal{E}\left(dw_t dw_t^T\right) = \mathcal{E}\left(dv_t dv_t^T\right) = 1$. The initial state x_0 has been chosen to be 0 and the covariance is 0.1.

The stochastic realization of system (4.46) is generated by *randn('state',1)*, according to the Euler–Maruyama method (Higham, 2001). The total experiment time is $T = 2$ with time discretization $\triangle t = 2 \times 10^{-4}$. The simplest numerical method, the Euler forward method, is used to solve (3.33), (3.32) and (3.34). In Fig. 3, the absolute value of errors E_{AE} of our method with adaptive scaling factor to different truncation modes $\nu = 3, 4$ and 5 are plotted with respect to time. The MSE for $\nu = 3, 4$ and 5 are 0.3310, 0.2182, and 0.1413, respectively. It verifies our intuition that the larger ν should give more accurate estimation.

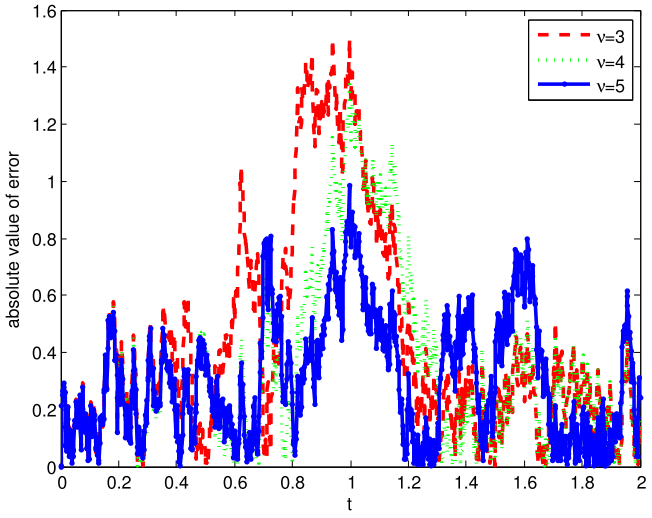

Fig. 3. The absolute value of errors from different truncation $\nu = 3, 4$ and 5 versus time are displayed, when the realization is generated by *randn('state',1)*.

5. Conclusions

In this paper we investigated a novel suboptimal method for the NLF problem by augmenting the original state via its gHPs. The augmented state after truncation satisfies a bilinear system, whose suboptimal filtering has been developed in Carravetta et al. (2000). Our paper is motivated by the key observation that in the Carleman approach (Germani et al., 2007) it is in general inappropriate to neglect all the higher moments. And we show that the more proper way to augment the state is via its gHPs, due to the fact that the expectation of these polynomials tends to zero as the degree goes to infinity, if the density function of the original states is in the exponential decay class. This makes the neglect of the gHPs with high enough degree more reasonable. The numerical simulation of the 1 d cubic sensor problem with zero/nearly zero initial condition is presented to show the superiority of our algorithms to the most widely used methods, including EKF, PF and Carleman approximation. Our algorithm with adaptive scaling factor may be more adequate for the NLF problems, whose linearized counterpart has unstable and unobservable state.

References

Akhiezer, N., & Krein, M. (1962). In W. Fleming, & D. Prill (Eds.), *Some problems in the theory of moments*. Providence, R. I.: American Mathematics Society, Tranl.

Arulampalam, M. S., Markell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.

Boyd, J. (1984). Asymptotic coefficients of Hermite function series. *Journal of Computational Physics*, *54*(3), 382–410.
Carravetta, F., Germani, A., & Shuakayev, M. (2000). A new suboptimal approach to the filtering problem for bilinear stochastic differential systems. *SIAM Journal on Control and Optimization*, *38*(4), 1171–1203.
Chen, X., Luo, X., & Yau, S. S.-T. (2017). Suboptimal linear estimation for continuous-discrete bilinear systems. submitted for publication.
Duncan, T. (1967). *Probability densities for diffusion processes with applications to nonlinear filtering theory* (Ph. D. thesis), United States: Stanford University.
Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, *54*, 343–367.
Fitzgerald, R. (1971). Divergence of the Kalman filter. *IEEE Transactions on Automatic Control*, *16*(6), 736–747.
Gelb, A. (1984). *Applied optimal estimation*. Cambridge: MIT Press.
Germani, A., Manes, C., & Palumbo, P. (2005). Filtering of differential nonlinear systems via a carleman approximation approach. In *Proceedings of the 44th IEEE conference on decision and control, and the european control conference*, Seville, Spain (pp. 5917–5922).
Germani, A., Manes, C., & Palumbo, P. (2007). Filtering of stochastic nonlinear differential systems via a carleman approximation approach. *IEEE Transactions on Automatic Control*, *52*(11), 2166–2172.
Gyongy, I., & Krylov, N. (2003). On the splitting-up method and stochastic partial differential equation. *The Annals of Probability*, *31*, 564–591.
Hazewinkel, M., Marcus, S., & Sussmann, H. (1983). Nonexistence of finite-dimensional filters for conditional statistics of the cubic sensor problem. *Systems & Control Letters*, *3*, 331–340.
Higham, D. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, *43*(3), 525–546.
Ito, K. (1996). Approximation of the Zakai equation for nonlinear filtering. *SIAM Journal on Control and Optimization*, *34*, 620–634.
Ito, K., & Xiong, K. (2000). Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, *45*(5), 910–927.
Jazwinski, A. (1970). *Stochastic processes and filtering theory*. New York: Academic Press.
Julier, S., & Uhlmann, J. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, *92*(3), 401–422.
Kalman, R. (1960). A new approach to linear filtering and prediction problem. *Transactions of the ASME. Series D, Journal of Basic Engineering*, *82*, 34–45.
Kalman, R., & Bucy, R. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME. Series D, Journal of Basic Engineering*, *83*, 95–107.
Kushner, H. (1967). Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, *12*(5), 546–556.
Lototsky, S., Mikulevicius, R., & Rozovskii, B. (1997). Nonlinear filtering revisited: a spectral approach. *SIAM Journal on Control and Optimization*, *35*, 435–461.
Luo, W. (2006). *Wiener chaos expansion and numerical solutions of stochastic partial differential equations* (Ph. D. thesis), United States: California Institute of Technology.
Luo, X. (2014). On recent advance of nonlinear filtering theory: emphases on global approaches. *Pure and Applied Mathematics Quarterly*, *10*(4), 685–721.
Luo, X., Jiao, Y., Chiou, W.-L., & Yau, S. S.-T. (2015). A novel suboptimal method for solving polynomial filtering problems. *Automatica. A Journal of IFAC*, *62*, 26–31.
Luo, X., Jiao, Y., Chiou, W.-L., & Yau, S. S.-T. (2016). Novel suboptimal filter via higher order central moments. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(4), 2030–2038.
Luo, X., & Yau, S. S.-T. (2013a). Complete real time solution of the general nonlinear filtering problem without memory. *IEEE Transactions on Automatic Control*, *58*(10), 2563–2578.
Luo, X., & Yau, S. S.-T. (2013b). Hermite spectral method to 1-D forward Kolmogorov equation and its application to nonlinear filtering problems. *IEEE Transactions on Automatic Control*, *58*(10), 2495–2507.
Mortensen, R. (1966). *Optimal control of continuous time stochastic systems* (Ph. D. thesis), United States: University of California at Berkeley.
Wong, W.-S., & Yau, S. S.-T. (1999). The estimation algebra of nonlinear filtering systems. In J. Baillieul, & J. C. Willems (Eds.), *Mathematical control theory* (pp. 33–65). Springer.
Yau, S.-T., & Yau, S. S.-T. (2008). Real time solution of the nonlinear filtering problem without memory. II. *SIAM Journal on Control and Optimization*, *47*(1), 163–195.
Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *11*(3), 230–243.

Xue Luo received her first Ph.D. degree in mathematics from East China Normal University (ECNU), Shanghai, P.R. China in 2010 and her second Ph.D. degree in applied mathematics from University of Illinois at Chicago (UIC) in 2013. During her study as a Ph.D. candidate in ECNU, she visited the department of Mathematics, University of Connecticut in 2008–2009 and the department of mathematics, statistics and computer science, UIC in 2009–2010, as a visiting scholar respectively. After her graduation from UIC, she joined in Beihang University (BUAA), Beijing, P.R. China. She is currently an associated professor in School of Mathematics and System Sciences, BUAA. She was elevated as IEEE senior member in 2015.

Dr. Luo's research interests include nonlinear filtering theory, numerical analysis of spectral methods, analysis of partial differential equations, sparse grid algorithm and fluid mechanics.

Stephen S.-T. Yau received the Ph.D. degree in mathematics from the State University of New York at Stony Brook, NY, US, in 1976. He was a member of Institute of Advanced Study at Princeton 1976–1977 and 1981–1982, and a Benjamin Pierce Assistant Professor at Harvard University during 1977–1980. After that, he joined the department of mathematics, statistics and computer science (MSCS), University of Illinois at Chicago (UIC), and served for over 30 years. He was awarded Sloan Fellowship in 1980, Guggenheim Fellowship in 2000, IEEE Fellow Award in 2003 and AMS Fellow Award in 2013. In 2005, he was entitled the UIC distinguished professor. During 2005–2011, he became a joint-professor of department of electrical and computer engineering and MSCS, UIC. After his retirement in 2012, he joined Tsinghua University, Beijing, P.R. China, where he is a full-time professor in department of mathematical science.

Dr. Yau's research interests include nonlinear filtering, bioinformatics, complex algebraic geometry, CR geometry and singularities theory.

Dr. Yau is the Managing Editor and founder of Journal of Algebraic Geometry from 1991, and the Editors-in-Chief and founder of Communications in Information and Systems from 2000 till now. He was the General Chairman of IEEE International Conference on Control and Information, which was held in the Chinese University of Hong Kong in 1995.