Automatica 187 (2026) 112879

Contents lists available at ScienceDirect

Automatica

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

The extended direct method for generalized time-varying Yau filtering systems☆

Minli Feng [a,b,c], Xiuqiong Chen [d], Stephen S.-T. Yau [e,*]

[a] Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, 101408, PR China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, PR China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, PR China
[d] School of Mathematics, Renmin University of China, Beijing, 100872, PR China
[e] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, PR China

ARTICLE INFO

Article history:
Received 24 November 2024
Received in revised form 23 December 2025
Accepted 7 January 2026

Keywords:
Nonlinear filtering
Duncan–Mortensen–Zakai equation
Direct method
Time-varying Yau filtering system
Convergence analysis

ABSTRACT

The goal of nonlinear filtering is to determine the conditional mean of the state given the observation history, and one way is to solve the Duncan–Mortensen–Zakai equation in real time and in a memoryless manner. One of our approaches is the direct method which works exceptionally well for time-varying Yau filtering system under the assumption that a certain function p is quadratic. In this paper, we eliminate this assumption, thereby extending the direct method to generalized time-varying Yau filtering systems. Furthermore, we provide a theoretical proof that, under very mild conditions, the error of its estimation result in the original framework is derived from the Gaussian approximation for a non-Gaussian initial distribution, and can be made arbitrary small if the error between this distribution and its Gaussian approximation is sufficiently small in $L^1(B_R)$ sense for a sufficiently large ball B_R, which facilitates numerical computation. Additionally, the extended direct method can still behave well provided that p can be approximated properly by its Taylor polynomial of degree 2. We also present three numerical experiments demonstrating the superior efficiency of the extended direct method compared to the extended Kalman filter and the particle filter.

© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Estimating the state of a stochastic dynamical system from noisy observations is of central importance in engineering. Filtering serves as a powerful tool for estimating unobservable stochastic processes that occur across various applied fields. The continuous time-varying filtering problem we address can be described as follows:

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\mathrm{d}V_t, & X_0 = \xi, \\ \mathrm{d}Y_t = h(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t, & Y_0 = 0, \end{cases} \quad t \in [0, T], \tag{1}$$

where:

- $T > 0$ is a fixed termination;
- $X := \{X_t : 0 \le t \le T\}$ is the $\mathbb{R}^n$-valued state process we would like to track;
- $Y := \{Y_t : 0 \le t \le T\}$ is the $\mathbb{R}^m$-valued noisy observation to the state process X;
- $V := \{V_t : 0 \le t \le T\}$ and $W := \{W_t : 0 \le t \le T\}$ are mutually independent Brownian motion processes, with $E\left[\mathrm{d}V_t \mathrm{d}V_t^{\mathrm{T}}\right] = Q(t)\mathrm{d}t$ and $E\left[\mathrm{d}W_t \mathrm{d}W_t^{\mathrm{T}}\right] = S(t)\mathrm{d}t$, respectively, where Q and S are C^∞ positive-definite-matrix-valued functions;
- ξ is a random variable with probability density function σ_0, which is independent of V and W;
- f and h are C^∞ functions, possibly nonlinear, and g is a C^∞ matrix-valued function.

Interest in filtering problem can be dated back almost two centuries to the work of Gauss and later, the names of Wiener and Kalman are associated with advances in filtering theory. The most influential work in filtering theory are the classical Kalman filter (KF) (Kalman, 1960), which was published in 1960, and its continuous counterpart Kalman–Bucy filter (Kalman & Bucy, 1961). Since most systems considered in real applications are nonlinear, there have been a lot of work which extend the filtering results to the nonlinear filtering (NLF) problems, such as the extended KF (EKF) (Gelb, 1974), and particle filter (PF) (Doucet et al., 2001; Gordon et al., 1993). In fact, EKF performs poorly when the

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Yanlong Zhao under the direction of Editor Alessandro Chiuso.

* Corresponding author.
E-mail addresses: fengminli@bimsa.cn (M. Feng), cxq0828@ruc.edu.cn (X. Chen), yau@uic.edu (S.S.-T. Yau).

https://doi.org/10.1016/j.automatica.2026.112879
0005-1098/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

dynamic system is significantly nonlinear and is very sensitive to initial value due to its reliance on linear approximation.

Since our focus is on the conditional mean, which is the minimum variance estimate, an alternative approach to NLF problem is to derive the conditional probability of the state X_t given the observation history $\mathcal{Y}_t := \sigma(\{Y_s : 0 \le s \le t\})$. It is known that the unnormalized conditioned probability density function of the state satisfies the Duncan–Mortensen–Zakai (DMZ) equation (Duncan, 1967; Mortensen, 1966; Zakai, 1969). However, we usually cannot get the explicit solution of the DMZ equation in most situations, and listed below are some methods to solve it numerically.

The first is the estimation algebra method proposed by Brockett, Clark and Mitter in the 1970s (Brockett, 1981; Brockett & Clark, 1980; Mitter, 1979). Once estimation algebra of system is a finite-dimensional Lie algebra, Wei–Norman approach will reduce the solution of DMZ equation to a Kolmogorov equation, a system of ordinary differential equations (ODEs) and several first-order linear partial differential equations (PDEs), and thus DMZ equation can be solved completely. Through persistent efforts, Yau and his collaborators have completely classified all finite dimensional estimation algebras of maximal rank (Chen & Yau, 1996; Chiou & Yau, 1994; Yau, 1994, 2003; Yau & Hu, 2005), and have been devoted to the study of non-maximal rank case (Ji Shi et al., 2017; Jiao & Yau, 2020; Shi & Yau, 2017; Wu & Yau, 2006; Yu et al., 2024).

The second is Yau-Yau algorithm proposed by Yau and Yau (2000, 2008). It separates the filtering process to on-line and off-line parts, and thus we can compute the solution to the DMZ equation numerically in a memoryless and real-time manner: it uses each new observation to update a distribution without referring back to any earlier observations, and it makes the decisions of the state on the spot while the observation data keep coming in. The off-line procedure is numerically solving a Kolmogorov-type PDE, and various kinds of methods have been proposed, such as spectral methods (Dong et al., 2021; Luo & Yau, 2013b), proper orthogonal decomposition (Wang et al., 2020), tensor training (Li et al., 2013), etc.

The third is the direct method proposed by Yau and Yau (1996, 1997). In Yau and Yau (2004), they proceed and restrict the system to finite-dimensional case, named Yau filtering system, with arbitrary initial condition, and obtain the fundamental solution of Schrödinger equation by solving ODEs. This technique is extended to time-varying cases in Chen et al. (2019, 2017), Ji Shi and Yau (2018). In Chen et al. (2019), there are essentially only two assumptions: f in (1) is of the form (5), and p in (8) is quadratic. Under these assumptions, the error of its estimation result is totally derived from the Gaussian approximation proposed in Ji Shi and Yau (2018), where we approximate the non-Gaussian initial distribution by the sum of several Gaussian distributions.

The purpose of this paper is to extend the direct method by eliminating the assumption that p is a quadratic polynomial, thus generalizing the method to a broader class of time-varying Yau filtering systems. Unlike previous work, we provide a rigorous convergence analysis for both the original and the extended direct method, showing that under very mild conditions, the estimation error of the original method can be made arbitrarily small when the error between the true distribution and its Gaussian approximation is sufficiently small in the $L^1(B_R)$ sense for a sufficiently large ball B_R. Additionally, the extended direct method can still behave well provided that p can be approximated properly by its Taylor polynomial of degree 2.

As stated in the Introduction of Ji Shi and Yau (2018), "the direct method is much stable and has theoretic convergence proof". The convergence of those using linear approximation is based on strict assumptions (Picard, 1991), and examples in Section 4 demonstrates situations where EKF fails due to high nonlinearity. In contrast, the proposed method, like the original direct method, shows superior performance both theoretically and numerically.

The key contribution of this paper is to transform the error analysis problem of the direct method into a well-posedness problem in the context of parameter perturbations in a class of PDEs. This allows us to apply classical tools from PDE theory to solve the problem more effectively. Our extension goes beyond simply replacing p with a Taylor expansion; it focuses on how perturbations in the solution to a second-order parabolic PDE can be controlled by perturbations in the coefficients. This advancement makes the method applicable to a broader range of time-varying Yau filtering systems. Moreover, it ensures the robustness of the method in practical numerical applications, even when the approximations are not perfect.

Furthermore, we provide three numerical experiments that demonstrate the superior efficiency of the extended direct method compared to existing methods such as the EKF and PF. These experiments further highlight the practical advantages of the proposed method in terms of computational efficiency and accuracy in nonlinear filtering tasks.

This paper is organized as follows. Section 2 provides the preliminaries and summarizes the main procedure of the direct method. Our main results are stated in 3, while detailed proofs of convergence analysis are postponed to Appendix B. We will demonstrate numerical simulation results in Section 4 and draw our conclusion in Section 5.

2. Preliminaries

In this section, we elaborate on the development of the direct method: starting from the DMZ equation, through several transformations, we convert the solution to the DMZ equation into that to a second-order parabolic PDE, and point out that when the coefficients of the equation satisfy certain conditions, its solution under Gaussian initial value will also always take the form of a Gaussian distribution. At this point, the problem is reduced to solving a system of ODEs. The propositions involved have been rigorously derived in their respective original literature, and thus will not be repeated here.

In Duncan (1967), we know that the unnormalized density function $\sigma(x,t)$ of X_t conditioned on the observation history $\mathcal{Y}_t$ satisfied the DMZ equation in $U := \mathbb{R}^n \times [0,T]$:

$$\begin{cases} \mathrm{d}\sigma = \left(\frac{1}{2}\nabla^{\mathrm{T}} G \nabla - f^{\mathrm{T}}\nabla - \mathrm{div} f\right)\sigma \mathrm{d}t \\ \qquad\quad + h^{\mathrm{T}} S^{-1}\sigma \mathrm{d}Y_t & \text{in } U, \\ \sigma(x,0) = \sigma_0(x) & \text{in } \mathbb{R}^n, \end{cases}$$

where $G(t) := g(t)Q(t)g(t)^{\mathrm{T}}$. For each arrived observation, making an invertible exponential transformation (Rozovsky, 1972)

$$u(x,t) = \mathrm{e}^{-h(x,t)^{\mathrm{T}} S(t)^{-1} Y_t}\sigma(x,t),$$

the DMZ equation is transformed into a deterministic PDE with stochastic coefficients:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \mathcal{L}u & \text{in } U, \\ u(x,0) = \sigma_0(x) & \text{in } \mathbb{R}^n, \end{cases} \tag{2}$$

where

$$\begin{aligned} \mathcal{L} := & \frac{1}{2}\nabla^{\mathrm{T}} G \nabla - (G\nabla K - f)^{\mathrm{T}}\nabla + \Bigg(-\frac{\partial}{\partial t}\left(h^{\mathrm{T}} S^{-1}\right)^{\mathrm{T}} Y_t \\ & + \frac{1}{2}\nabla^{\mathrm{T}} G \nabla K + \frac{1}{2}(\nabla K)^{\mathrm{T}} G \nabla K - f^{\mathrm{T}}\nabla K - \mathrm{div} f \\ & - \frac{1}{2} h^{\mathrm{T}} S^{-1} h\Bigg), \end{aligned}$$

and

$$K(x,t) := h(x,t)^{\mathrm{T}} S(t)^{-1} Y_t.$$

However, the exact solution to (2), generally speaking, does not have a closed form.

Let us denote the observation time sequence as

$$\mathcal{P} := \{0 = \tau_0 < \tau_1 < \cdots < \tau_N = T\}.$$

Let u_k be a solution to (2) in $U^{(k)} := \mathbb{R}^n \times [\tau_{k-1}, \tau_k]$ with $Y_t = Y_{\tau_{k-1}}$ on the time interval $[\tau_{k-1}, \tau_k]$, $k = 1, \ldots, N$:

$$\begin{cases} \dfrac{\partial u_k}{\partial t} = \mathcal{L}_k u_k & \text{in } U^{(k)}, \\ u_1(x,0) = \sigma_0(x) & \text{in } \mathbb{R}^n, \\ u_k(x, \tau_{k-1}) = u_{k-1}(x, \tau_{k-1}) \text{ for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{3}$$

where

$$\begin{aligned} \mathcal{L}_k :=& \frac{1}{2}\nabla^{\mathrm{T}} G \nabla - (G\nabla K_k - f)^{\mathrm{T}} \nabla \\ &+ \Big(-\frac{\partial}{\partial t}\left(h^{\mathrm{T}} S^{-1}\right)^{\mathrm{T}} Y_{\tau_{k-1}} + \frac{1}{2}\nabla^{\mathrm{T}} G \nabla K_k \\ &+ \frac{1}{2}(\nabla K_k)^{\mathrm{T}} G \nabla K_k - f^{\mathrm{T}} \nabla K_k - \operatorname{div} f \\ &- \frac{1}{2} h^{\mathrm{T}} S^{-1} h \Big), \end{aligned}$$

and $K_k(x,t) := h(x,t)^{\mathrm{T}} S(t)^{-1} Y_{\tau_{k-1}}$. By Yau and Yau (2005), we know that in both pointwise sense and L^2 sense,

$$u(x,\tau) = \lim_{\sup_{1\leq k\leq N}(\tau_k - \tau_{k-1}) \to 0} u_k(x,\tau), \forall \tau \in [\tau_{k-1}, \tau_k].$$

Therefore, u_k is a good approximation of u in the interval $[\tau_{k-1}, \tau_k]$. We only need to seek a solution to DMZ Eq. (3).

In Luo and Yau (2013a), an on- and off-line algorithm is proposed. The key observation is that the heavy computation of solving PDE can be moved to off-line by the following proposition.

Proposition 1 (*Luo & Yau, 2013a, Proposition 2.1*). *For each $t \in [\tau_{k-1}, \tau_k]$, $k = 1, \ldots, N$, $u_k(x,t)$ satisfies (3) if and only if*

$$\tilde{u}_k(x,t) = \mathrm{e}^{h(x,t)^{\mathrm{T}} S(t)^{-1} Y_{\tau_{k-1}}} u_k(x,t)$$

satisfied the Kolmogorov forward equation (KFE)

$$\begin{cases} \dfrac{\partial \tilde{u}_k}{\partial t} = \Big(\dfrac{1}{2}\nabla^{\mathrm{T}} G \nabla - f^{\mathrm{T}} \nabla - (\operatorname{div} f \\ \qquad + \dfrac{1}{2} h^{\mathrm{T}} S^{-1} h \Big) \Big) \tilde{u}_k & \text{in } U^{(k)}, \\ \tilde{u}_1(x,0) = \sigma_0(x) & \text{in } \mathbb{R}^n, \\ \tilde{u}_k(x, \tau_{k-1}) = \tilde{u}_{k-1}(x, \tau_{k-1}) \text{ for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{4}$$

In Chen et al. (2017) and Chen et al. (2019), the results for time-invariant Yau filtering systems are extended to the more general time-varying Yau filtering systems:

$$f(x,t) = L_1(t)x + L_0(t) + \nabla\tilde{\phi}(x,t), \tag{5}$$

where $L_1 : [0,T] \to M_{n\times n}(\mathbb{R})$, $L_0 : [0,T] \to \mathbb{R}^n$ and $\tilde{\phi} : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ are C^∞ functions. When $n = 1$, this assumption is trivial, since any continuous function from $\mathbb{R}$ to $\mathbb{R}$ has a primitive function. As for the general case $n \geq 2$, it is a generalization of the linear case, so the direct method can behave better than those using linear approximation.

Proposition 2 (*Chen et al., 2017, Proposition 2*). *Assume that G is a positive-definite-matrix-valued function. Suppose $\tilde{u}_k$ is a solution to (4), and f is of the form (5). Let*

$$\tilde{u}_k(x,t) = \mathrm{e}^{\phi(x,t)} \tilde{v}_k(x,t),$$

where ϕ satisfies $\nabla\phi(x,t) = G^{-1}(t)\nabla_x\tilde{\phi}(x,t)$, then we have the following equation for $\tilde{v}_k$:

$$\begin{cases} \dfrac{\partial \tilde{v}_k}{\partial t} = \Big(\dfrac{1}{2}\nabla^{\mathrm{T}} G \nabla - (L_1 x + L_0)^{\mathrm{T}} \nabla \\ \qquad + \tilde{p} \Big) \tilde{v}_k & \text{in } U^{(k)}, \\ \tilde{v}_1(x,0) = \mathrm{e}^{-\phi(x,0)} \sigma_0(x) & \text{in } \mathbb{R}^n, \\ \tilde{v}_k(x, \tau_{k-1}) = \exp\Big(h(x, \tau_{k-1})^{\mathrm{T}} S(\tau_{k-1})^{-1} \\ \qquad \left(Y_{\tau_{k-1}} - Y_{\tau_{k-2}}\right)\Big) \tilde{v}_{k-1}(x, \tau_{k-1}) \\ \qquad \text{for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{6}$$

where

$$\begin{aligned} \tilde{p} :=& \frac{1}{2}\nabla^{\mathrm{T}} G \nabla\phi - \frac{1}{2}\nabla\phi^{\mathrm{T}} G \nabla\phi - (L_1 x + L_0)^{\mathrm{T}} \nabla\phi \\ &- \Delta\tilde{\phi} - \frac{\partial\phi}{\partial t} - \frac{1}{2} h^{\mathrm{T}} S^{-1} h - \operatorname{tr} L_1. \end{aligned}$$

Proposition 3 (*Chen et al., 2019, Theorem 1*). *Assume that G is a positive-definite-matrix-valued function. Suppose $\tilde{v}_k$ is a solution to (6), and let*

$$\tilde{v}_k(x,t) = v_k(z,t),$$

where $z = B(t)^{-1}x$, and B is a positive-definite-matrix-valued function such that $G(t) = B(t)B(t)^{\mathrm{T}}$. Then v_k is a solution to the following equation:

$$\begin{cases} \dfrac{\partial v_k}{\partial t} = \Big(\dfrac{1}{2}\Delta - (F_1 z + F_0)^{\mathrm{T}} \nabla. \\ \qquad + p \Big) v_k & \text{in } U^{(k)}, \\ v_1(z,0) = \mathrm{e}^{-\phi(B(0)z,0)} \sigma_0(B(0)z) & \text{in } \mathbb{R}^n, \\ v_k(z, \tau_{k-1}) = \exp\Big(h(B(\tau_{k-1})z, \tau_{k-1})^{\mathrm{T}} \\ \qquad S(\tau_{k-1})^{-1} \left(Y_{\tau_{k-1}} - Y_{\tau_{k-2}}\right)\Big) \\ \qquad v_{k-1}(z, \tau_{k-1}) \text{ for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{7}$$

where

$$\begin{aligned} F_1(t) &:= \left(B(t)^{-1}\right)' B(t) + B(t)^{-1} L_1(t) B(t), \\ F_0(t) &:= B(t)^{-1} L_0(t), \end{aligned}$$

and $p(z,t) := \tilde{p}(B(t)z, t)$.

Therefore, what we shall pay much attention to is the following second-order parabolic equation

$$\begin{aligned} \frac{\partial v_k}{\partial t}(z,t) =& \frac{1}{2}\Delta v_k(z,t) - (F_1(t)z + F_0(t))^{\mathrm{T}} \nabla v_k(z,t) \\ &+ p(z,t) v_k(z,t) \qquad \text{in } U^{(k)}. \end{aligned} \tag{8}$$

In the original direct method, we need to assume that p in (8) is quadratic with respect to (w.r.t.) z. Though the assumption about p seems restrictive, it includes Kalman-Bucy and Beněs (Beněs, 1981) filtering systems as its special cases.

Proposition 4 (*Chen et al., 2019, Theorem 2*). *Assume that p in (8) is of the form:*

$$p(z,t) = z^{\mathrm{T}} P_2(t) z + P_1(t) z + p_0(t),$$

where $P_2 : [\tau_{k-1}, \tau_k] \to M_{n\times n}(\mathbb{R})$, $P_1 : [\tau_{k-1}, \tau_k] \to \mathbb{R}^n$ and $p_0 : [\tau_{k-1}, \tau_k] \to \mathbb{R}$ are C^∞ functions, and P_2 is a symmetric-matrix-valued function. Then with Gaussian initial condition

$$v_k(z, \tau_{k-1}) = \exp\left(z^{\mathrm{T}} A_{2,\tau_{k-1}} z + A_{1,\tau_{k-1}}^{\mathrm{T}} z + a_{0,\tau_{k-1}} \right),$$

where $A_{2,\tau_{k-1}} \in M_{n\times n}(\mathbb{R})$ is a symmetric matrix, $A_{1,\tau_{k-1}} \in \mathbb{R}^n$, and $a_{0,\tau_{k-1}} \in \mathbb{R}$, a solution to (8) is of the following form

$$v_k(z,t) = \exp\left(z^{\mathrm{T}} A_2(t) z + A_1(t)^{\mathrm{T}} z + a_0(t) \right),$$

where $A_2 : [\tau_{k-1}, \tau_k] \to M_{n\times n}(\mathbb{R})$, $A_1 : [\tau_{k-1}, \tau_k] \to \mathbb{R}^n$ and $a_0 : [\tau_{k-1}, \tau_k] \to \mathbb{R}$ are C^∞ functions with $A_2(\tau_{k-1}) = A_{2,\tau_{k-1}}$, $A_1(\tau_{k-1}) = A_{1,\tau_{k-1}}$ and $a_0(\tau_{k-1}) = a_{0,\tau_{k-1}}$, which satisfy the following system of nonlinear ODEs, and A_2 is a symmetric-matrix-valued function:

$$\begin{cases} A_2' = 2A_2^2 - 2A_2F_1 + P_2, \\ \left(A_1^{\mathrm{T}}\right)' = 2A_1^{\mathrm{T}}A_2 - A_1^{\mathrm{T}}F_1 - 2F_0^{\mathrm{T}}A_2 + P_1^{\mathrm{T}}, \\ a_0' = \operatorname{tr} A_2 + \dfrac{1}{2}A_1^{\mathrm{T}}A_1 - F_0^{\mathrm{T}}A_1 + P_0. \end{cases}$$

However, the initial value $v_k(x, \tau_{k-1})$ in (7) in every step usually cannot be Gaussian. Therefore we need to derive its Gaussian approximation. It is well known that any non-Gaussian density function can be well approximated by finite linear combination of Gaussian distributions and the most widely used technique is expectation–maximization algorithm. However, in Ji Shi and Yau (2018), a new and original way to do Gaussian approximation is proposed, which is very effective as verified by the numerical experiments.

3. Main results

3.1. Algorithm

In the previous section, we have covered all the preliminaries of the original direct method for time-varying Yau filtering systems with arbitrary initial distributions. For generalized time-varying Yau filtering systems, i.e. p in (8) is not required to be quadratic, a natural idea is to approximate p by its Taylor polynomial of degree 2.

An outline of extended direct method is given in Algorithm 1. Clearly, the error of the output

$$\frac{\displaystyle\int_{\mathbb{R}^n} x w_k(x) v_k\left(B(\tau_k)^{-1}x, \tau_k\right) \mathrm{d}x}{\displaystyle\int_{\mathbb{R}^n} w_k(x) v_k\left(B(\tau_k)^{-1}x, \tau_k\right) \mathrm{d}x}$$

is entirely due to the Gaussian approximation of $v_k(z, \tau_{k-1})$ and the polynomial approximation of p, where

$$w_k(x) := \mathrm{e}^{\phi(x,\tau_k) + h(x,\tau_k)^{\mathrm{T}} S(\tau_k)^{-1}\left(Y_{\tau_k} - Y_{\tau_{k-1}}\right)}. \tag{9}$$

More precisely, assume v_k and $\hat{v}_k$ are solutions to (8) with the initial conditions $v_k(\cdot, \tau_{k-1}) = \varphi$ and $\hat{v}_k(\cdot, \tau_{k-1}) = \hat{\varphi}$, and coefficients p and $\hat{p}$, respectively. We expect that

$$\left| \frac{\displaystyle\int_{\mathbb{R}^n} x w_k(x) v_k(z, \tau_k)\, \mathrm{d}x}{\displaystyle\int_{\mathbb{R}^n} w_k(x) v_k(z, \tau_k)\, \mathrm{d}x} - \frac{\displaystyle\int_{\mathbb{R}^n} x w_k(x) \hat{v}_k(z, \tau_k)\, \mathrm{d}x}{\displaystyle\int_{\mathbb{R}^n} w_k(x) \hat{v}_k(z, \tau_k)\, \mathrm{d}x} \right|$$

can be made arbitrary small if the error between φ and $\hat{\varphi}$ is sufficiently small in some sense, and so is the error between p and $\hat{p}$, where w_k is independent of the initial value and the coefficient, and $z := B(\tau_k)^{-1}x$. Without loss of generality, we may assume $B(\tau_k)$ is an identity matrix in the discussion that follows.

3.2. Convergence analysis

The proofs of all results in this section can be found in Appendix B.

Consider the following second-order parabolic equation

$$\begin{aligned} \frac{\partial u}{\partial t}(x,t) = &\frac{1}{2}\Delta u(x,t) - F(x,t)^{\mathrm{T}}\nabla u(x,t) \\ &+ p(x,t)u(x,t) \qquad \text{in } U \end{aligned} \tag{10}$$

Algorithm 1 Extended Direct method

for $k \in \{1, \dots, N\}$ **do**
  Calculate $v_k(z, \tau_{k-1})$ according to (7).
  Get the Gaussian approximation of $v_k(z, \tau_{k-1})$ via method proposed in Ji Shi and Yau (2018)

$$v_k(z, \tau_{k-1}) \approx \sum_i^{N_k} \alpha_{k,i} \mathrm{e}^{z^{\mathrm{T}}A_{2,k,i}z + A_{1,k,i}^{\mathrm{T}}z + a_{0,k,i}}.$$

  for $i \in \{1, \dots, N_k\}$ **do**
    Get the polynomial approximation of p in (8)

$$p(z,t) \approx \frac{1}{2}z^{\mathrm{T}}\nabla^2 p(0,t)z + (\nabla p(0,t))^{\mathrm{T}}z + p(0,t).$$

    Solve (8) with the initial condition

$$v_{k,i}(z, \tau_{k-1}) = \mathrm{e}^{z^{\mathrm{T}}A_{2,k,i}z + A_{1,k,i}^{\mathrm{T}}z + a_{0,k,i}}$$

    by Proposition 4.
  end for
  Assemble $v_k \approx \sum_i^{N_k} \alpha_{k,i} v_{k,i}$.
  Recover $\sigma(x, \tau_k)$ by reversing the transformations in Section 2.
  Calculate the conditional expectation of the state X_{τ_k}.
end for

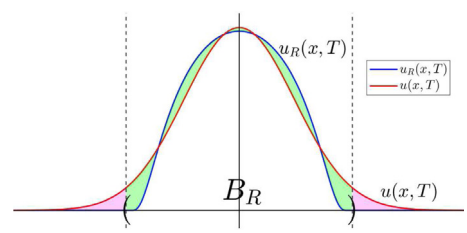

Fig. 1. Proof strategy diagram.

with the initial condition $u(\cdot, 0) = \varphi$, where F and p are sufficiently smooth functions, and $\varphi \in H_0^1(\mathbb{R}^n)$. We relate it to the following initial boundary value problem in $U_R := B_R \times [0, T]$

$$\begin{cases} \dfrac{\partial u_R}{\partial t} = \dfrac{1}{2}\Delta u_R - F^{\mathrm{T}}\nabla u_R + p u_R & \text{in } U_R, \\ u_R(x,0) = \varphi(x) & \text{in } B_R, \\ u_R(x,t) = 0 & \text{on } \Gamma_R, \end{cases} \tag{11}$$

where $R > 0$. Note that φ may not belong to $H_0^1(B_R)$, but we may as well assume it holds by the use of mollification.

In numerical computations, we inevitably need to restrict the domain from the entire space to a bounded domain B_R. To this end, following the idea proposed in Yau and Yau (2008), we want the solution u_R to (11), which is defined on B_R, to approximate the solution u to (10) with the initial condition $u(\cdot, 0) = \varphi$, which is defined on the entire space, provided R is sufficiently large. To be concrete, we hope that u_R contributes negligibly outside B_R and that it closely matches u inside B_R.

The above strategy can be illustrated in Fig. 1. Theorems 5 and 6 will prove that the areas of the green and pink shadows can be arbitrarily small, respectively.

At first, we show that the contribution of u outside B_R becomes negligible as R increases.

Theorem 5. *Let u be a non-negative solution to (10) with the initial condition $u(\cdot, 0) = \varphi$. Under the assumptions of Corollary 16 (see*

Appendix A), as well as

$$\frac{n+1}{2}+|F(x,t)|+\operatorname{div}F(x,t)+p(x,t)\le C_1,\ \forall(x,t)\in U,$$

where $C_1\ge 0$ is a constant, we have

$$\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}u(x,T)\mathrm{d}x\le \mathrm{e}^{C_1T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x.$$

In particular, we have

$$\int_{\mathbb{R}^n\setminus B_R}u(x,T)\mathrm{d}x\le \mathrm{e}^{C_1T-\sqrt{1+R^2}}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x.$$

Next, we show u and u_R differ by a small amount in the sense of integral over a sufficiently large closed ball.

Theorem 6. *Let u and u_R be non-negative solutions to (10) with the initial condition $u(\cdot,0)=\varphi$ and (11), respectively. Under the assumptions and notations of Theorem 5, as well as:*

(1) $p(x,t)\le 0,\ \forall(x,t)\in U$;
(2) $12+2n+4|x|\,|F(x,t)|+\operatorname{div}F(x,t)+p(x,t)\le C_2,\ \forall(x,t)\in U$;
(3) $\mathrm{e}^{-\sqrt{1-|x|^2}}\left(12+2n+4|x|\,|F(x,t)|+\operatorname{div}F(x,t)\right)\le C_3$, $\forall(x,t)\in U$,

where C_2 and C_3 are non-negative constants. Then $v:=u-u_R$ is non-negative, and we have

$$\int_{B_R}\left(\mathrm{e}^{|x|^4/R^3-2|x|^2/R}-\mathrm{e}^{-R}\right)v(T)\mathrm{d}x$$
$$\le\frac{\mathrm{e}^{C_2T}-1}{C_2}C_3\mathrm{e}^{-R}\mathrm{e}^{C_1T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x.$$

In particular, we have

$$\int_{B_{R/2}}v(T)\mathrm{d}x$$
$$\le\frac{2\left(\mathrm{e}^{C_2T}-1\right)}{C_2}C_3\mathrm{e}^{-9R/16}\mathrm{e}^{C_1T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x.$$

Then we show that the solution to (11) can be linearly controlled by its initial value in $L^1(B_R)$ sense.

Theorem 7. *Assume that*

$$\operatorname{div}F(x,t)+p(x,t)\le C,\ \forall(x,t)\in U_R,$$

where $C\ge 0$ is a constant. Let u_R be a solution to (11). Then we have

$$\int_{B_R}|u_R(x,T)|\,\mathrm{d}x\le \mathrm{e}^{CT}\int_{B_R}|\varphi(x)|\,\mathrm{d}x.$$

To ensure that the right-hand side of the inequalities in the above theorems is finite, we restrict the initial value function to the following subset:

$$\mathcal{I}_c:=\left\{\varphi\in H_0^1\left(\mathbb{R}^n\right)\cap C^2\left(\mathbb{R}^n\right):\int_{\mathbb{R}^n}\mathrm{e}^{c|x|}\varphi(x)^2\mathrm{d}x\right.$$
$$\left.<+\infty,\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x<+\infty,\varphi\ge 0\right\},$$

where $c\ge 0$ is a constant. This restriction essentially requires that the initial value function decays at a certain rate as it moves away from the origin. Most functions in practical applications, such as the Gaussian probability density function, satisfy this requirement, so it will not offset the generalization of the proposed method.

The next theorem is the central result about the original direct method, essentially aligning with our expectations from the previous subsection.

Theorem 8. *Let u and $\hat{u}$ be non-negative solutions to (10) with initial conditions $u(\cdot,0)=\varphi$ and $\hat{u}(\cdot,0)=\hat{\varphi}$, respectively, with $\varphi,\hat{\varphi}\in\mathcal{I}_{4c}$. Assume that:*

(1) $\operatorname{div}F(x,t)+2p(x,t)$ is upper bounded in U;
(2) $|F(x,t)|^2+2p(x,t)$ is upper bounded in U;
(3) $\Delta p(x,t)$ is upper bounded in U;
(4) $|F(x,t)+cx/|x||^2+2p(x,t)$ has a upper bound in U less than $1/T$;
(5) $|F(x,t)|+\operatorname{div}F(x,t)+p(x,t)$ is upper bounded in U;
(6) $p(x,t)\le 0,\ \forall(x,t)\in U$;
(7) $4|x|\,|F(x,t)|+\operatorname{div}F(x,t)+p(x,t)$ is upper bounded in U;
(8) $\mathrm{e}^{-\sqrt{1+|x|^2}}\left(12+2n+4|x|\,|F(x,t)|+\operatorname{div}F(x,t)\right)$ is upper bounded in U.

Then for any $\varepsilon>0$, there exist positive numbers R and δ, such that $\|u(\cdot,T)-\hat{u}(\cdot,T)\|_{L^1(\mathbb{R}^n)}<\varepsilon$ provided $\|\varphi-\hat{\varphi}\|_{L^1(B_R)}<\delta$.

Note that F is linear and p is quadratic within the framework of the direct method. Under these assumptions, most expressions simplify to quadratic polynomials. Therefore, as long as the leading coefficient of p is sufficiently small, the expressions will have an upper bound. For simplicity, we rewrite the previous theorem in the following form.

Theorem 9. *Let u and $\hat{u}$ be non-negative classical solutions to (10) with initial conditions $u(\cdot,0)=\varphi$ and $\hat{u}(\cdot,0)=\hat{\varphi}$, respectively. Assume that:*

(1) F is a $C^{1,0}$ function of the form $F(x,t)=F_0(t)+F_1(t)x$;
(2) p is a $C^{2,0}$ function of the form $p(x,t)=p_0(t)+p_1(t)^{\mathrm{T}}x+x^{\mathrm{T}}p_2(t)x$;
(3) D is a C^1 function with $|D(x)|\le d_0+d_1|x|$, and $|\operatorname{div}D(x)|\le e_0+e_1|x|+e_2|x|^2$;
(4) w is a positive C^2 function with $\nabla w(x)=w(x)D(x)$, and $w\varphi,w\hat{\varphi},x_iw\varphi,x_iw\hat{\varphi}\in\mathcal{I}_0$, where x_i is the ith component of x.

Further assume that:

(5) $|F_1(t)|^2+2p_2(t)+2d_1^2+e_2\le 0,\ \forall t\in[0,T]$;
(6) $\Delta\left(|D|^2-\operatorname{div}D-2F^{\mathrm{T}}D\right)$ is upper bounded in U;
(7) $2p(x,t)+|D(x)|^2-\operatorname{div}D(x)-F(x,t)^{\mathrm{T}}D(x)\le 0,\ \forall(x,t)\in U$;
(8) $(2d_1+8)\,|F_1(t)|+2p_2(t)+d_1^2+8d_1+e_2\le 0,\ \forall t\in[0,T]$.

Write

$$x_{\mathrm{est}}:=\int_{\mathbb{R}^n}xw(x)u(x,T)\mathrm{d}x\Big/\int_{\mathbb{R}^n}w(x)u(x,T)\mathrm{d}x,$$
$$\hat{x}_{\mathrm{est}}:=\int_{\mathbb{R}^n}xw(x)\hat{u}(x,T)\mathrm{d}x\Big/\int_{\mathbb{R}^n}w(x)\hat{u}(x,T)\mathrm{d}x.$$

Then for a given φ and any $\varepsilon>0$, there exist positive numbers R and δ, such that for any $\hat{\varphi}$, we have $\left|x_{\mathrm{est}}-\hat{x}_{\mathrm{est}}\right|<\varepsilon$ provided $\|\varphi-\hat{\varphi}\|_{L^1(B_R)}<\delta$.

Remark 10. Assumptions (1) and (2) are requirements of the direct method. Assumptions (3), (4), and (6) are also natural, since they just require that each w_k in (9) does not vary too drastically, and each initial valued in (8) to be similar to rapidly decreasing function. Although Assumptions (5), (7), and (8) may appear technical, in essence they only require $p_2(t)$ to be small enough, which is typically achievable in practice. Therefore, we think these assumptions are very mild.

Now we consider the extended case. Using the same idea, we just need to pay attention to the perturbation of solution to equation of the form (10) when its coefficients vary. It is obvious that such variation can introduce errors into the solution, but we

will show that this error is typically small if the perturbations themselves are small. As shown in Theorem 8, the perturbation of the solution can be controlled as long as the perturbation in the initial value is sufficiently small. As a corollary of Theorem 8, Theorem 9 demonstrates that the error of the estimates given by the direct method remains small, provided that the Gaussian approximation is sufficiently accurate. Parallelly, the following theorem will deal with the perturbation of p.

We might as well proceed by proving a more general result. The key point is to estimate the $L^1(B_R)$ norm of the difference between (11) and the following initial boundary value problem:

$$\begin{cases} \dfrac{\partial \hat{u}_R}{\partial t} = \dfrac{1}{2}\Delta \hat{u}_R - \hat{F}^{\mathrm{T}}\nabla \hat{u}_R + \hat{p}\hat{u}_R & \text{in } U_R, \\ \hat{u}_R(x,0) = \varphi(x) & \text{in } B_R, \\ \hat{u}_R(x,t) = 0 & \text{on } \Gamma_R. \end{cases} \tag{12}$$

Theorem 11. *Assume that:*

(1) $\operatorname{div} F(x,t) + p(x,t) \le C, \forall (x,t) \in U_R$;
(2) $\|F(\cdot,t) - \hat{F}(\cdot,t)\|_{L^2(B_R)} \le \lambda_1, \forall t \in [0,T]$;
(3) $\left|p(x,t) - \hat{p}(x,t)\right| \le \lambda_2, \forall (x,t) \in U_R$,

where C, λ_1, and λ_2 are non-negative constants. Let u_R and $\hat{u}_R$ be non-negative solutions to (11) and (12), respectively. Then we have

$$\int_{B_R} \left|u_R(x,T) - \hat{u}_R(x,T)\right| \mathrm{d}x \le K(\lambda_1, \lambda_2),$$

where $K : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function with $K(\lambda_1,\lambda_2) \to 0$ as either $\lambda_1 \to 0$ or $\lambda_2 \to 0$.

In particular, the extended direct method can behave well if $\lambda_1 = 0$ and λ_2 are small enough when R is as large as we need, or equivalently, the Taylor polynomial of degree 2 of p in (8) is a sufficiently good approximation to itself.

Remark 12. Furthermore, we have proved that the assumption that f in (1) is of the form (5) can be relaxed provided that f is close enough to the gradient of some function. In fact, through a variable substitution, we can show that the equation for $\tilde{u}_k$ in (4) can be transformed into the form of (10), and the perturbation of F and p in (10) is minimal when f in (4) undergoes only slight variations. By use of the conclusion above, the extended direct method remains effective in this case.

4. Simulation

Now we use three numerical examples to verify the efficiency of Algorithm 1, and the filtering system here is as follows:

$$\begin{cases} \mathrm{d}x_t = \left(Lx_t + l + \dfrac{\partial \tilde{\phi}}{\partial x}(x_t,t)\right)\mathrm{d}t + \mathrm{d}v_t, & x_0 = \xi, \\ \mathrm{d}y_t^1 = h_1(x_t)\,\mathrm{d}t + \mathrm{d}w_t^1, & y_0^1 = 0, \\ \mathrm{d}y_t^2 = h_2(x_t)\,\mathrm{d}t + \mathrm{d}w_t^2, & y_0^2 = 0, \\ & t \in [0,T], \end{cases} \tag{13}$$

where:

- L and l are constants;
- $T = 10$ is a fixed termination, with the sampling interval $\Delta t = 0.01$ for observations and its corresponding time sequence $\tau_k := k\Delta t, k = 0, 1, \ldots, 1000$;
- x_t, y_t^1, and y_t^2 are scalar stochastic processes;
- v_t, w_t^1, and w_t^2 are scalar independent Brownian motions, with variances 1, $(1+\sin(0.2t))^2$, and $(1+\sin(0.2t))^2$, respectively;
- ξ is a scalar random variable with unnormalized probability density function
$$\tilde{\sigma}_0(x) := \exp\left(-x\sin x - 0.5x\cos x - x^2 + 3x + 2\right),$$
i.e., its probability density is $\sigma_0 := \dfrac{\tilde{\sigma}_0}{\int_{\mathbb{R}} \tilde{\sigma}_0(x)\mathrm{d}x}$;
- $\tilde{\phi}, h_1, h_2$ are C^∞ functions.

To compare the average performance of different methods, we introduce the MSE. The MSE for m repeated realizations at instant τ_k is defined as follows:

$$\mathrm{MSE}(\tau_k) := \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=0}^{k}\left(x_{\tau_j}^i - \hat{x}_{\tau_j}^i\right)^2,$$

where $x_{\tau_j}^i$ is the real state at instant τ_j in the ith realization and $\hat{x}_{\tau_j}^i$ is the estimate of $x_{\tau_j}^i$ by different filtering methods.

Recall that we shall approximate p in (8) by its Taylor polynomial of degree 2:

$$\hat{p}(z,t) := \frac{1}{2}z^{\mathrm{T}}\left(\nabla^2 p(0,t)\right)z + \left(\nabla p(0,t)\right)^{\mathrm{T}} z + p(0,t).$$

Though in theory it may cause a fairly significant error when the difference between p and $\hat{p}$ is too large, the direct method still behaves well compared to other widely used methods in real applications considering the trade-off between the MSE and running time.

In the following three examples, the real dynamic system (13) is approximated by Euler's method with time step, and all methods are numerically implemented by the Euler's method as well. For Unscented KF (UKF), we used the parameters $\alpha = 1, \beta = \kappa = 2$. For Iterated EKF (IEKF), we terminated the iteration when the update magnitude was less than 10^{-5} or the number of iterations reached 10. For PF, we conducted experiments with 2000 and 10 000 particles, respectively.

Example 1. We set $L = l = 0.1$, $\tilde{\phi}(x,t) := 10^{-4}x^3 t$, $h_1(x) = x\sin x$ and $h_2(x) = x\cos x$. It can be computed that

$$\begin{aligned} p(z,t) = &-\left(4.5\times 10^{-8}t^2\right)z^4 \\ &-\left(3\times 10^{-4}t + 10^{-4}\right)z^3 \\ &-\left(3\times 10^{-4}t + \frac{1}{2\left(1+\sin(0.2t)\right)^2}\right)z^2 \\ &-\left(3\times 10^{-4}t\right)z - 1, \end{aligned}$$

which is approximated by

$$\begin{aligned} \hat{p}(z,t) = &-\left(3\times 10^{-4}t + \frac{1}{2\left(1+\sin(0.2t)\right)^2}\right)z^2 \\ &-\left(3\times 10^{-4}t\right)z - 1 \end{aligned}$$

when we apply Algorithm 1. Here, p is a polynomial of degree 4 with relatively small coefficients of high-order term, which causes $\hat{p}$ to be a good approximation to p.

The result is demonstrated in Figs. 2, 3 and Table 1. It can be seen that the proposed method achieves the lowest MSE, provides ideal tracking of the state trajectory, and meets the real-time requirements. In comparison, the methods based on the KF, although having minimal time cost, shows instability in some simulations for this nonlinear model. On the other hand, the PF exhibits a large MSE with a small number of particles, and fails to meet real-time requirements even with a slight increase in particle count. Therefore, the proposed method demonstrates superior performance.

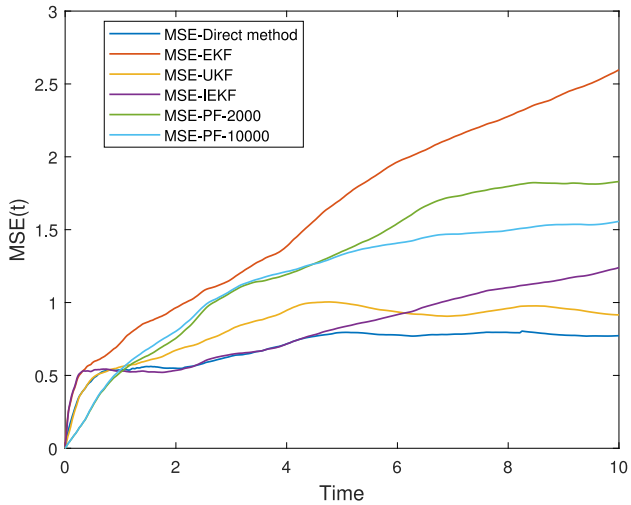

Fig. 2. MSE(t) based on 50 simulations of Example 1.

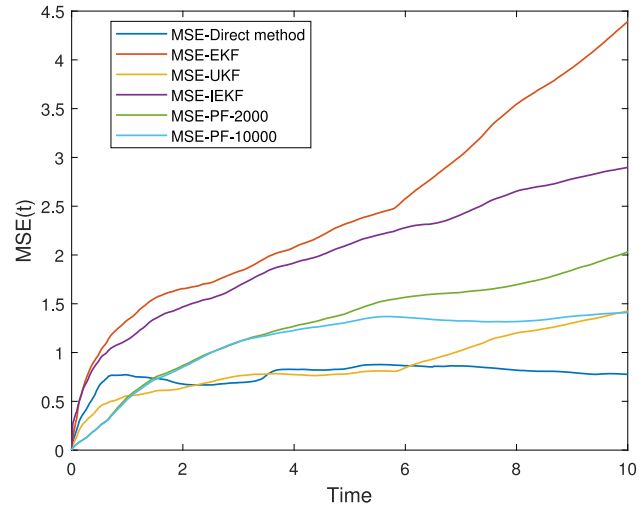

Fig. 4. MSE(t) based on 50 simulations of Example 2.

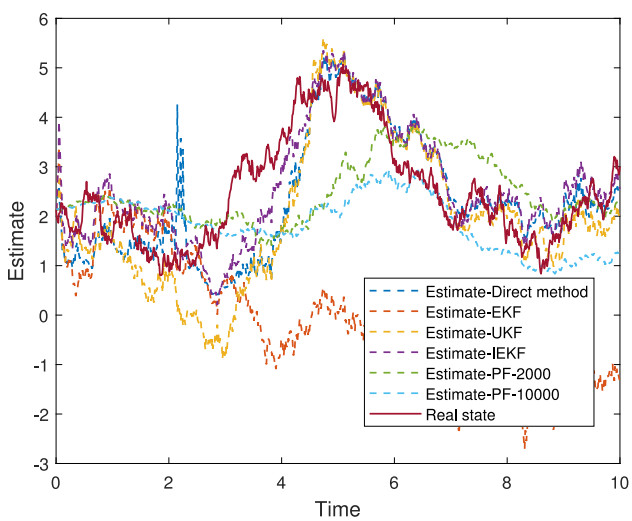

Fig. 3. A typical simulation of Example 1.

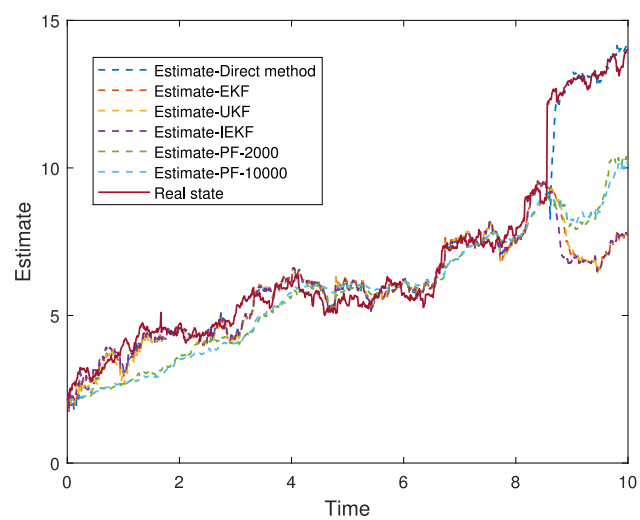

Fig. 5. A typical simulation of Example 2.

Table 1
Average time cost and MSE(T) of Example 1.

Method	Time/s	MSE(T)
Direct method	1.8685	0.7725
EKF	0.0058	2.5975
UKF	0.0115	0.9152
IEKF	0.0174	1.2401
PF with 2000 particles	3.1213	1.8302
PF with 10 000 particles	17.2285	1.5574

Table 2
Average time cost and MSE(T) of Example 2.

Method	Time/s	MSE(T)
Direct method	1.8640	0.8380
EKF	0.0053	4.3933
UKF	0.0112	1.4742
IEKF	0.0159	2.8973
PF with 2000 particles	9.7452	2.1108
PF with 10 000 particles	54.0549	1.4190

Example 2. In order to check the robustness of each algorithm to noise, we replace the noise terms $\mathrm{d}v_t, \mathrm{d}w_t^1, \mathrm{d}w_t^2$ from the previous example with a t-distribution with parameter $\nu = 3$, multiplying by a factor $\sqrt{(\nu-2)/\nu}$ to keep their variance unchanged.

The result is demonstrated in Figs. 4, 5 and Table 2. Both the proposed method and the KF-based methods are built upon Gaussian noise, so their performance deteriorates in the presence of non-Gaussian noise. However, it is evident that the proposed method exhibits better robustness and outperforms other methods, including the PF.

Example 3. Let us consider a real-world example: A planet moves in an elliptical orbit around a star under the influence of gravitational attraction, with a semi-major axis length of a and an eccentricity e. The parametric equation of the orbit is given by $y(x) = \left(a\cos x, a\sqrt{1-e^2}\sin x\right)$. According to Kepler's second law, the planet's velocity is given by

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{C}{1-e\cos x},$$

where C is a constant. Moreover,

$$\frac{\mathrm{d}s}{\mathrm{d}x} = \left|y'(x)\right| = a\sqrt{1-e^2\cos^2 x},$$

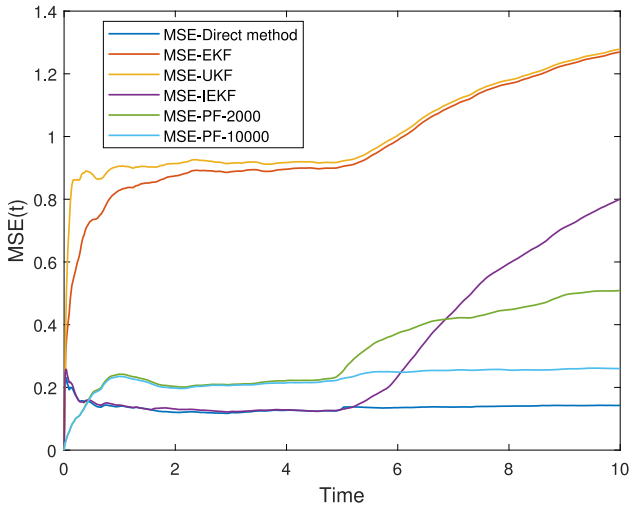

Fig. 6. MSE(t) based on 50 simulations of Example 3.

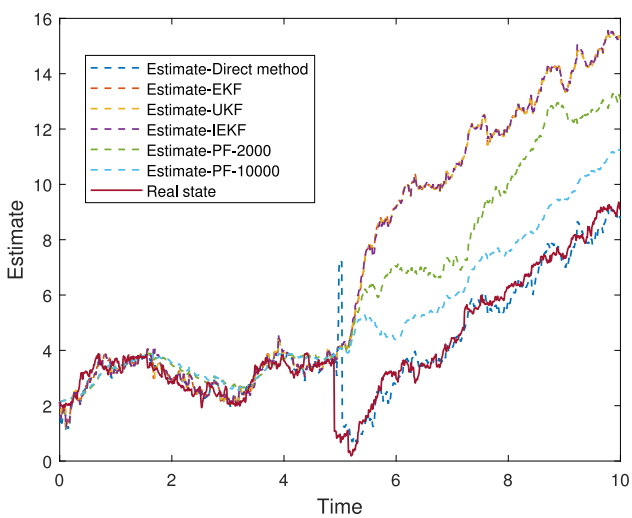

Fig. 7. A typical simulation of Example 3.

Table 3
Average time cost and MSE(T) of Example 2.

Method	Time/s	MSE(T)
Direct method	0.5822	0.1426
EKF	0.0054	1.2699
UKF	0.0107	1.2791
IEKF	0.0152	0.8006
PF with 2000 particles	8.1409	0.5088
PF with 10 000 particles	49.7394	0.2601

which implies that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{C}{a(1-e\cos x)\sqrt{1-e^2\cos^2 x}} := f(x).$$

Therefore, the filtering system can be considered under the following parameters in Eq. (13): $L = l = 0$, $\tilde{\phi}(x,t) = \int_{x_0}^{x} f(x)\mathrm{d}x$, $h_1(x) = -a\sin x$ and $h_2(x) = a\sqrt{1-e^2}\sin x$, where we choose $C = a = 10$ and $e = 0.1$.

The result is demonstrated in Figs. 6, 7 and Table 3. In most simulations, the proposed method performs just as well as the KF-based methods, while the PF, constrained by real-time requirements, performs slightly worse. In a few simulations, at certain moments, the noise is particularly large, causing interference to all methods. However, the proposed method demonstrates significant robustness among these methods.

The comparison of these methods is presented in Table 4.

5. Conclusion

In this paper, we extend the direct method for time-varying Yau filtering systems by removing the assumption that p in (8) is quadratic, and provide a convergence analysis to show that under very mild assumptions, Algorithm 1 can give accurate estimates with arbitrary precision for the conditional mean of state process given the observations, provided the Gaussian approximation

and the polynomial approximation are accurate enough. Specifically, we demonstrate that the key to the extension is not just the approximation of p but rather the ability to control perturbations in the solution through perturbations in p. This extension allows us to generalize the direct method to a much broader class of systems, significantly improving its applicability.

Furthermore, we present three numerical experiments demonstrating its efficiency compared to the EKF and PF. EKF has the advantage of low computational overhead and performs well in almost linear cases, but it becomes unstable in cases of high nonlinearity. PF can handle nonlinear situations, but under real-time requirements, its accuracy is not as high as that of the proposed method. Therefore, the capability of the extended direct method to solve very general nonlinear filtering problems is theoretically and numerically verified.

Although the proposed method demonstrates superiority in the one-dimensional case, its performance in high-dimensional scenarios is limited by the time complexity associated with Gaussian approximations. To enhance the efficiency and scalability of our method in high-dimensional settings, one promising direction for future research is the development of more efficient Gaussian approximation techniques.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 42450242, No. 12201631), and Tsinghua University Education Foundation.

Appendix A. Notations

$U := \mathbb{R}^n \times [0,T]$, $U^{(k)} := \mathbb{R}^n \times [\tau_{k-1}, \tau_k]$.
$B_R := \{x \in \mathbb{R}^n : |x| < R\}$, $\Gamma_R := \partial B_R \times [0,T]$.

$$\mathcal{S}_c := \Big\{\boldsymbol{u} \in L^2\left(0,T; H_0^1\left(\mathbb{R}^n\right)\right) : \sup_{0\le t\le T}\int_{\mathbb{R}^n} \mathrm{e}^{c|x|}u(x,t)^2\mathrm{d}x < +\infty\Big\}.$$

$$\mathcal{I}_c := \Big\{\varphi \in H_0^1\left(\mathbb{R}^n\right)\cap C^2\left(\mathbb{R}^n\right) : \varphi \ge 0, \int_{\mathbb{R}^n} \mathrm{e}^{c|x|}\varphi(x)^2\mathrm{d}x < +\infty, \int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}}\varphi(x)\mathrm{d}x < +\infty\Big\}.$$

Appendix B. Detailed proofs of theorems in Section 3

B.1. Weak convergence

In the first place, we present some foundational concepts and facts that will be necessary for the understanding of subsequent results and proofs.

For $p \geq 1$ and a Banach space $(X, \|\cdot\|_X)$, write

$$L^p(0,T;X) := \left\{ \boldsymbol{u} : [0,T] \to X : \|\boldsymbol{u}\|_{L^p(0,T;X)} := \|\|\boldsymbol{u}(\cdot)\|_X\|_{L^p((0,T))} < +\infty \right\},$$

which naturally becomes a normed vector space. For $\boldsymbol{u} \in L^1(0,T;X)$, say $\boldsymbol{v} \in L^1(0,T;X)$ is the *weak derivative* of $\boldsymbol{u}$, written $\boldsymbol{u}' = \boldsymbol{v}$, if

$$\int_0^T \boldsymbol{u}(t)\psi'(t)\mathrm{d}t = -\int_0^T \boldsymbol{v}(t)\psi(t)\mathrm{d}t, \forall \psi \in C_c^\infty([0,T]).$$

Say $\boldsymbol{u}_R \in L^2\left(0,T;H_0^1(B_R)\right)$, or its corresponding function $u_R(x,t) := (\boldsymbol{u}_R(t))(x)$, is a *weak solution* of (11), if $\boldsymbol{u}_R(0) = \varphi$, and $\boldsymbol{u}_R' \in L^2\left(0,T;H_0^1(B_R)\right)$ satisfies

$$\int_{B_R} \boldsymbol{u}_R'\psi\mathrm{d}x = \int_{B_R} \left(-\frac{1}{2}(\nabla \boldsymbol{u}_R)^{\mathrm{T}} \nabla\psi + \left(-F^{\mathrm{T}}\nabla \boldsymbol{u}_R + p\boldsymbol{u}_R\right)\psi \right)\mathrm{d}x, \text{ a.e. } t \in [0,T], \quad \forall \psi \in H_0^1(B_R).$$

A weak solution to (10) (with a certain initial condition) can be defined similarly. Clearly, a classical solution is a weak solution.

Let $\boldsymbol{u}_R \in L^2\left(0,T;H_0^1(B_R)\right)$. Then its extension by zero belongs to $L^2\left(0,T;H^1(\mathbb{R}^n)\right) = L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$. In this sense, the weak solutions to (10) and (11) lie in the same space $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$, which should cause no confusion.

Lemma 13. *Assume that:*

(1) $\operatorname{div} F(x,t) + 2p(x,t) \leq C_1, \forall (x,t) \in U$;
(2) $|F(x,t)|^2 + 2p(x,t) \leq C_2, \forall (x,t) \in U$;
(3) $\Delta p(x,t) \leq C_3, \forall (x,t) \in U$,

where C_1, C_2 and C_3 are non-negative constants. Let $\boldsymbol{u}_R$ be a weak solution to (11). Then $\{\boldsymbol{u}_R : R > 0\}$ is a bounded subset of $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$.

Proof. Fix $R > 0$. Write

$$E_0(t) := \int_{B_R} u_R^2(x,t)\mathrm{d}x, \quad E_1(t) := \int_{B_R} |\nabla u_R(x,t)|^2\,\mathrm{d}x.$$

Taking the derivative of E_0, we have

$$\begin{aligned} E_0'(t) &= \int_{B_R} \left(u_R\Delta u_R - 2u_R F^{\mathrm{T}} + 2pu_R^2\right)\mathrm{d}x \\ &= -\int_{B_R} |\nabla u_R|^2\,\mathrm{d}x + \int_{B_R} (\operatorname{div} F + 2p)u_R^2\mathrm{d}x \\ &\leq C_1 E_0(t), \end{aligned}$$

which implies

$$E_0(t) \leq \mathrm{e}^{C_1 T} E_0(0) < +\infty. \tag{B.1}$$

Taking the derivative of E_1, we have

$$\begin{aligned} E_1'(t) &= -\int_{B_R} (\Delta u_R)^2\,\mathrm{d}x + \int_{B_R} 2\Delta u_R F^{\mathrm{T}}\nabla u_R\mathrm{d}x \\ &\quad - \int_{B_R} 2pu_R\Delta u_R\mathrm{d}x \\ &\leq \int_{B_R} \left|F^{\mathrm{T}}\nabla u_R\right|^2\mathrm{d}x - \int_{B_R} 2u_R(\nabla p)^{\mathrm{T}}\nabla u_R\mathrm{d}x \\ &\quad + \int_{B_R} 2p\,|\nabla u_R|^2\,\mathrm{d}x \\ &\leq \int_{B_R} \left(|F|^2 + 2p\right)|\nabla u_R|^2\,\mathrm{d}x + \int_{B_R} (\Delta p)u_R^2\mathrm{d}x \\ &\leq C_2 E_1(t) + C_3 E_0(t), \end{aligned}$$

which implies

$$E_1(t) \leq \mathrm{e}^{C_2 T} E_1(0) + \frac{\mathrm{e}^{C_2 T} - 1}{C_2} C_3 \mathrm{e}^{C_1 T} E_0(0) < +\infty. \tag{B.2}$$

By (B.1) and (B.2), we have $\|\boldsymbol{u}_R\|^2_{L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)} < +\infty$.

Lemma 14. *Under the assumptions and notations of Lemma 13, there exists an increasing sequence of positive numbers, written $\{R_k\}$, such that $\lim_{k\to+\infty} R_k = +\infty$, and $\left\{\boldsymbol{u}_{R_k}\right\}$ weakly converges in $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$ to a weak solution $\boldsymbol{u}$ of (10) with the initial condition $\boldsymbol{u}(0) = \varphi$.*

Proof. Thanks to Diestel and Uhl (1977, Chapter VII, Section 6), $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$ is a reflexive space, since $H_0^1(\mathbb{R}^n)$ has the Radon–Nikodym property. So any bounded subset of $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$ has a weakly convergent sequence. In particular, by Lemma 13, $\{\boldsymbol{u}_R\}$ has a weakly convergent sequence, written $\boldsymbol{u}_{R_k} \rightharpoonup \boldsymbol{u}$ in $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$. Let us verify that $\boldsymbol{u}$ is indeed a weak solution to 13 with the initial condition $\boldsymbol{u}(0) = \varphi$: Clearly, $\boldsymbol{u}_{R_k} \rightharpoonup \boldsymbol{u}$ in $L^2\left(0,T;H_0^1(\mathbb{R}^n)\right)$ implies that for any $t \in [0,T]$, $\boldsymbol{u}_{R_k}(t) \rightharpoonup \boldsymbol{u}(t)$, $\nabla\boldsymbol{u}_{R_k}(t) \rightharpoonup \nabla\boldsymbol{u}(t)$ and $\boldsymbol{u}_{R_k}'(t) \rightharpoonup \boldsymbol{u}'(t)$ in $H_0^1(\mathbb{R}^n)$, so we have

$$\begin{aligned} &\int_{\mathbb{R}^n} \left(-\frac{1}{2}(\nabla\boldsymbol{u})^{\mathrm{T}}\nabla\psi + \left(-F^{\mathrm{T}}\nabla\boldsymbol{u} + p\boldsymbol{u}\right)\psi \right)\mathrm{d}x \\ &= \lim_{k\to+\infty} \int_{\mathbb{R}^n} \left(-\frac{1}{2}\left(\nabla\boldsymbol{u}_{R_k}\right)^{\mathrm{T}}\nabla\psi \right. \\ &\quad \left. + \left(-F^{\mathrm{T}}\nabla\boldsymbol{u}_{R_k} + p\boldsymbol{u}_{R_k}\right)\psi \right)\mathrm{d}x \\ &= \lim_{k\to+\infty} \int_{\mathbb{R}^n} \boldsymbol{u}_{R_k}'\psi\,\mathrm{d}x \\ &= \int_{\mathbb{R}^n} \boldsymbol{u}'\psi\,\mathrm{d}x, \text{ a.e. } t \in [0,T], \forall \psi \in H_0^1(\mathbb{R}^n), \end{aligned}$$

and $\boldsymbol{u}(0) = \lim_{k\to+\infty} \boldsymbol{u}_{R_k}(0) = \varphi$, since weak convergence implies pointwise convergence.

For technical reasons, we restrict the weak solution to $\mathcal{S}_c$ defined in Appendix A, where $c \geq 0$ is a constant.

Lemma 15. *Let $c \geq 0$ be a constant. Assume that*

$$\left|F(x,t) + c\frac{x}{|x|}\right|^2 + 2p(x,t) \leq C < \frac{1}{T}, \forall (x,t) \in U,$$

Table 4
Comparison of the existing methods.

Method	Direct method	KF-based methods	PF
Performance	Best accuracy	Optimal for linear and Gaussian systems	High accuracy if particles is sufficient
Computational efficiency	Moderate	High	Low
Applicability	Broader than linear	Linear or almost linear	General

where $C \geq 0$ is a constant. Then there is a unique weak solution to (10) in $\mathcal{S}_{4c}$.

Proof. It is sufficient to show $u = 0$ if $\varphi = 0$. Let $T_0 \in (0, T]$ be arbitrary, and denote $\tilde{U} := \mathbb{R}^n \times [0, T_0]$. For any $\boldsymbol{\psi} \in L^2\left(0, T_0; C_c^\infty(\mathbb{R}^n)\right)$,

$$\int_{\tilde{U}} \boldsymbol{u}'\boldsymbol{\psi}\mathrm{d}x\mathrm{d}t = \int_{\tilde{U}} \left(-\frac{1}{2}(\nabla \boldsymbol{u})^{\mathrm{T}} \nabla \boldsymbol{\psi} - \left(F^{\mathrm{T}}\nabla \boldsymbol{u}\right)\boldsymbol{\psi} + p\boldsymbol{u}\boldsymbol{\psi}\right)\mathrm{d}x\mathrm{d}t$$

Replacing $\boldsymbol{\psi}$ by $\mathrm{e}^{4c|x|}\boldsymbol{\psi}$ and then approximating $\boldsymbol{u}$ by $\boldsymbol{\psi}$, we have

$$\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_0)^2\,\mathrm{d}x \leq C\int_{\tilde{U}} \mathrm{e}^{4c|x|}\boldsymbol{u}^2\mathrm{d}x\mathrm{d}t.$$

By the mean value theorem, there exists $T_1 \in (0, T_0)$ such that

$$\int_{\tilde{U}} \mathrm{e}^{4c|x|}\boldsymbol{u}^2\mathrm{d}x\mathrm{d}t = T_0\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_1)^2\,\mathrm{d}x.$$

So

$$\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_0)^2\,\mathrm{d}x \leq CT_0\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_1)^2\,\mathrm{d}x.$$

Repeating the above process, for any positive integer k, there exists $T_k \in (0, T_0)$ such that

$$\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_0)^2\,\mathrm{d}x \leq (CT_0)^k\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|}\boldsymbol{u}(T_k)^2\,\mathrm{d}x,$$

where $CT_0 < 1$. Let $k \to +\infty$, and we obtain $\boldsymbol{u}(T_0) = 0$ since $\boldsymbol{u} \in \mathcal{S}_{4c}$. Thus $\boldsymbol{u} = 0$ since T_0 is arbitrary.

The following result is an immediate consequence.

Corollary 16. *Let u and u_R be classical solutions to (10) with the initial condition $u(\cdot, 0) = \varphi$ and (11), respectively. Assume that $u \in \mathcal{S}_{4c}$, where $c \geq 0$ is a constant. Then under the assumptions of Lemmas 13 and 15, there exists an increasing sequence of positive numbers, written $\{R_k\}$, such that $\lim_{k\to+\infty} R_k = +\infty$, and $u_{R_k} \rightharpoonup u$ in $L^2\left(0, T; H_0^1(\mathbb{R}^n)\right)$.*

B.2. Approximation inside and outside a disc

Proof of Theorem 5. Let $\boldsymbol{u}_R$ be a weak solution to (11). Writing $\gamma(x) := \sqrt{1+|x|^2}$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{B_R} \mathrm{e}^{\gamma}u_R\mathrm{d}x \\
=&\frac{1}{2}\int_{B_R} \mathrm{div}\left(\mathrm{e}^{\gamma}\nabla\gamma\right)u_R\mathrm{d}x + \frac{1}{2}\int_{\partial B_R} \mathrm{e}^{\gamma}\frac{\partial u_R}{\partial \nu}\mathrm{d}S \\
&+ \int_{B_R} \left(\mathrm{e}^{\gamma}u_R(\nabla\gamma)^{\mathrm{T}}F + \mathrm{e}^{\gamma}(\mathrm{div}\,F)u_R + \mathrm{e}^{\gamma}pu_R\right)\mathrm{d}x \\
\leq& \int_{B_R} \mathrm{e}^{\gamma}u_R\left(\frac{\Delta\gamma + |\nabla\gamma|^2}{2} - (\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F + p\right)\mathrm{d}x \\
\leq& C_1\int_{B_R} \mathrm{e}^{\gamma}u_R\mathrm{d}x,
\end{aligned}$$

which implies

$$\int_{B_R} \mathrm{e}^{\gamma}u_R\mathrm{d}x \leq \mathrm{e}^{C_1t}\int_{B_R} \mathrm{e}^{\gamma}\varphi\mathrm{d}x. \tag{B.3}$$

Fix $r > 0$. Consider the truncated function $\eta_r := \mathrm{e}^{\gamma}\mathbf{1}_{B_r}$, and suppose $\eta_r \in H_0^1(\mathbb{R}^n)$ by improving its regularity via a convolution with a mollifier. Then we have

$$\begin{aligned}
\int_{\mathbb{R}^n} \eta_r u\mathrm{d}x &= \lim_{k\to+\infty}\int_{\mathbb{R}^n} \eta_r u_{R_k}\mathrm{d}x \\
&\leq \overline{\lim_{k\to+\infty}}\int_{B_{R_k}} \mathrm{e}^{\gamma}u_{R_k}\mathrm{d}x \\
&\leq \overline{\lim_{k\to+\infty}}\,\mathrm{e}^{C_1t}\int_{B_{R_k}} \mathrm{e}^{\gamma}\varphi\mathrm{d}x \\
&= \mathrm{e}^{C_1t}\int_{\mathbb{R}^n} \mathrm{e}^{\gamma}\varphi\mathrm{d}x.
\end{aligned}$$

By monotone convergence theorem, we obtain

$$\int_{\mathbb{R}^n} \mathrm{e}^{\gamma}u\mathrm{d}x = \lim_{r\to+\infty}\int_{\mathbb{R}^n} \eta_r u\mathrm{d}x \leq \mathrm{e}^{C_1t}\int_{\mathbb{R}^n} \mathrm{e}^{\gamma}\varphi\mathrm{d}x.$$

Proof of Theorem 6. Thanks to the maximum principle of solutions to second-order parabolic equations (Evans, 2010, Chapter 7.1, Theorem 12), v is non-negative in U by Assumption (1), since $v = u$ is non-negative in $U\backslash U_R$, which in particular holds on Γ_R, and $v(\cdot, 0) = 0$. Writing $\gamma(x) := |x|^4/R^3 - 2|x|^2/R$, $\eta := \mathrm{e}^{\gamma} - \mathrm{e}^{R}$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{B_R} \eta v\mathrm{d}x \\
=&\frac{1}{2}\int_{B_R} v\Delta\eta\mathrm{d}x - \frac{1}{2}\int_{\partial B_R} v\frac{\partial\eta}{\partial\nu}\mathrm{d}S + \frac{1}{2}\int_{\partial B_R} \eta\frac{\partial v}{\partial\nu}\mathrm{d}S \\
&+ \int_{B_R} \left(\mathrm{div}\left(\eta F\right) + \eta p\right)v\mathrm{d}x - \int_{\partial B_R} \eta v F^{\mathrm{T}}\nu\mathrm{d}S(x) \\
=&\frac{1}{2}\int_{B_R} \mathrm{e}^{\gamma}v\left(\Delta\gamma + |\nabla\gamma|^2\right)\mathrm{d}x \\
&+ \int_{B_R} \mathrm{e}^{\gamma}v\left((\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F\right)\mathrm{d}x + \int_{B_R} \eta pv\mathrm{d}x \\
=&\int_{B_R} \eta v\left(\frac{\Delta\gamma + |\nabla\gamma|^2}{2} + (\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F + p\right)\mathrm{d}x \\
&+ \mathrm{e}^{-R}\int_{B_R} v\left(\frac{\Delta\gamma + |\nabla\gamma|^2}{2} + (\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F\right)\mathrm{d}x \\
\leq& C_2\int_{B_R} \eta v\mathrm{d}x + C_3\mathrm{e}^{-R}\int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}}u\mathrm{d}x,
\end{aligned}$$

which implies

$$\begin{aligned}
&\int_{B_R} \eta v(x, T)\mathrm{d}x \\
\leq& \mathrm{e}^{C_2T}\int_{B_R} \eta v(x, 0)\mathrm{d}x \\
&+ \frac{\mathrm{e}^{C_2T} - 1}{C_2}C_3\mathrm{e}^{-R}\mathrm{e}^{C_1T}\int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}}\varphi\mathrm{d}x \\
=&\frac{\mathrm{e}^{C_2T} - 1}{C_2}C_3\mathrm{e}^{-R}\mathrm{e}^{C_1T}\int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}}\varphi\mathrm{d}x.
\end{aligned}$$

B.3. Estimation by the initial value

To show Theorem 7, we first introduce a technical lemma without proof.

Lemma 17 (Yau and Yau 2008, Lemma 4.1)**.** *Let Ω be a bounded domain in $\mathbb{R}^n$ and let $v : \overline{\Omega} \times [0, T] \to \mathbb{R}$ be a C^1 function. Assume that $v(x, t) = 0$ for $(x, t) \in \partial\Omega \times [0, T]$. Let $\Omega_t^+ := \{x \in \Omega : v(x, t) \geq 0\}$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t^+} v(x, t)\mathrm{d}x = \int_{\Omega_t^+} \frac{\partial v}{\partial t}(x, t)\mathrm{d}x, \text{ a.e. } t \in [0, T].$$

Proof of Theorem 7. Write

$$\Omega_t^{\pm} := \{x \in B_R : \pm u_R(x,t) \geq 0\}.$$

Using Lemma 17, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t^+} u_R \mathrm{d}x \\
=&\int_{\Omega_t^+}\left(\frac{1}{2}\Delta u_R - F^{\mathrm{T}}\nabla u_R + p u_R\right)\mathrm{d}x \\
=&\frac{1}{2}\int_{\partial\Omega_t^+}\frac{\partial u_R}{\partial \nu}\mathrm{d}x + \int_{\Omega_t^+}(\operatorname{div} F + p)u_R \mathrm{d}x \\
\leq& C\int_{\Omega_t^+} u_R \mathrm{d}x, \text{ a.e. } t \in [0,T],
\end{aligned}$$

which implies

$$\int_{\Omega_t^+} u_R \mathrm{d}x \leq \mathrm{e}^{Ct}\int_{\Omega_0^+}\varphi \mathrm{d}x, \text{ a.e. } t \in [0,T].$$

The result regarding Ω_t^- can be similarly derived. Therefore, we have

$$\int_{B_R}|u_R|\,\mathrm{d}x \leq \mathrm{e}^{Ct}\int_{B_R}|\varphi|\mathrm{d}x, \text{ a.e. } t \in [0,T],$$

both sides of which are continuous w.r.t. t. Thus the conclusion is improved to hold in pointwise sense.

B.4. Error estimation for the original direct method

Proof of Theorem 8. Note that $\varphi, \hat{\varphi} \in \mathcal{I}_{4c}$ implies $\boldsymbol{u}, \hat{\boldsymbol{u}} \in \mathcal{S}_{4c}$. By Theorem 5, Assumptions (1)-(5) imply that there exists a decreasing function $C_1 : \mathbb{R}_+ \to \mathbb{R}_+$ with $C_1(+\infty) = 0$, such that

$$\int_{|x|\geq R} u(x,T)\mathrm{d}x, \int_{|x|\geq R}\hat{u}(x,T)\mathrm{d}x \leq C_1(R), \forall R > 0.$$

By Theorem 6, Assumptions (6)-(8) imply that there exists a decreasing function $C_2 : \mathbb{R}_+ \to \mathbb{R}_+$ with $C_2(+\infty) = 0$, such that

$$\int_{B_{R/2}}(u(x,T) - u_R(x,T))\,\mathrm{d}x,$$
$$\int_{B_{R/2}}\left(\hat{u}(x,T) - \hat{u}_R(x,T)\right)\mathrm{d}x \leq C_2(R), \forall R > 0.$$

By Theorem 7, there exists a function $C_3 : \mathbb{R}_+ \to \mathbb{R}_+$, such that

$$\int_{B_R}\left|u_R(x,T) - \hat{u}_R(x,T)\right|\mathrm{d}x \leq C_3(R)\int_{B_R}\left|\varphi - \hat{\varphi}\right|\mathrm{d}x.$$

Thus we have

$$\begin{aligned}
&\int_{\mathbb{R}^n}\left|u(x,T) - \hat{u}(x,T)\right|\mathrm{d}x \\
\leq& 2C_1\left(\frac{R}{2}\right) + 2C_2(R) + C_3(R)\delta.
\end{aligned}$$

For sufficiently large R and sufficiently small δ, we have $2C_1(R/2) + 2C_2(R) < \varepsilon/2$ and $C_3(R)\delta < \varepsilon/2$, which implies $\|u(x,T) - \hat{u}(x,T)\|_{L^1(\mathbb{R}^n)} < \varepsilon$.

A straightforward calculation yields the following result.

Lemma 18. *Let u be a solution to (10). Then for any C^2 function $w : \mathbb{R}^n \to \mathbb{R}$ with $\nabla w(x) = w(x)D(x)$, $v(x,t) := w(x)u(x,t)$ is a solution to the equation*

$$\frac{\partial v}{\partial t}(x,t) = \frac{1}{2}\Delta v(x,t) - \tilde{F}(x,t)^{\mathrm{T}}\nabla v(x,t) + \tilde{p}(x,t)v(x,t),$$

where $\tilde{F}(x,t) := F(x,t) + D(x)$, and

$$\tilde{p}(x,t) := p(x,t) + \frac{|D(x)|^2 - \operatorname{div} D(x)}{2} - F(x,t)^{\mathrm{T}}D(x).$$

Proof of Theorem 9. It is straightforward to compute that

$$\begin{aligned}
x_{\text{est}} - \hat{x}_{\text{est}} =& \left(1 - \frac{\int_{\mathbb{R}^n} w\left(u(x,T) - \hat{u}(x,T)\right)\mathrm{d}x}{\int_{\mathbb{R}^n} w u(x,T)\mathrm{d}x}\right)^{-1} \\
&\left(\frac{\int_{\mathbb{R}^n} x w\left(u(x,T) - \hat{u}(x,T)\right)\mathrm{d}x}{\int_{\mathbb{R}^n} w u(x,T)\mathrm{d}x}\right. \\
&\left. - \frac{\int_{\mathbb{R}^n} w\left(u(x,T) - \hat{u}(x,T)\right)\mathrm{d}x}{\int_{\mathbb{R}^n} w u(x,T)\mathrm{d}x} x_{\text{est}}\right).
\end{aligned}$$

Thus it is sufficient to show that

$$\|w\left(u(x,T) - \hat{u}(x,T)\right)\|_{L^1(\mathbb{R}^n)}$$

and

$$\|x_i w\left(u(x,T) - \hat{u}(x,T)\right)\|_{L^1(\mathbb{R}^n)}$$

are sufficiently small, since $\int_{\mathbb{R}^n} w u(x,T)\mathrm{d}x$ and x_{est} are fixed if φ is given. Take $\|w\left(u(x,T) - \hat{u}(x,T)\right)\|_{L^1(\mathbb{R}^n)}$ as an example: By Theorem 8 and Lemma 18, $\|w\left(u(x,T) - \hat{u}(x,T)\right)\|_{L^1(\mathbb{R}^n)}$ can be made arbitrarily small if $\|w\left(\varphi - \hat{\varphi}\right)\|_{L^1(B_R)}$ is sufficiently small, and $\tilde{F}$ and $\tilde{p}$ in Lemma 18, as substitutes of F and p, respectively, meet the assumptions listed in Theorem 8. Note that $\|w\left(u(x,T) - \hat{u}(x,T)\right)\|_{L^1(\mathbb{R}^n)}$ can be made arbitrarily small if $\|\varphi - \hat{\varphi}\|_{L^1(B_R)}$ is sufficiently small, since w is fixed and so is $\sup_{x\in\bar{B}_R}|w(x)|$. All that remains is a direct calculation for translating Assumptions (1)-(9) in Theorem 8 to Assumptions (5)-(8) in this theorem.

B.5. Error estimation for the extended direct method

Lemma 19. *Assume that:*

(1) $\operatorname{div} F(x,t) + 2p(x,t) \leq C_1, \forall (x,t) \in U_R$;
(2) $|F(x,t)|^2 + 2p(x,t) \leq C_2, \forall (x,t) \in U_R$;
(3) $\Delta p(x,t) \leq C_3, \forall (x,t) \in U_R$,

where C_1, C_2 and C_3 are non-negative constants. Let u_R be a non-negative solution to (10). Then we have

$$\begin{aligned}
&\int_{B_R}|\nabla u_R|^2\,\mathrm{d}x \\
\leq& \mathrm{e}^{C_2 T}\int_{B_R}\varphi^2\mathrm{d}x + \frac{\mathrm{e}^{C_2 T} - 1}{C_2}C_3\mathrm{e}^{C_1 T}\int_{B_R}|\nabla\varphi|^2\mathrm{d}x.
\end{aligned}$$

Proof. Similar to the proof of Lemma 13.

Proof of Theorem 11. Write $v := u_R - \hat{u}_R$ and

$$\Omega_t^{\pm} := \{x \in B_R : \pm v(x,t) \geq 0\}.$$

Using Lemmas 17 and 19, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t^+} v\mathrm{d}x \\
=&\frac{1}{2}\int_{\partial\Omega_t^+}\frac{\partial v}{\partial \nu}\mathrm{d}x+\int_{\Omega_t^+}(\operatorname{div}F+p)v\mathrm{d}x \\
&-\int_{\Omega_t^+}\left(F-\hat{F}\right)^{\mathrm{T}}\nabla\hat{u}_R\mathrm{d}x+\int_{\Omega_t^+}\left(p-\hat{p}\right)\hat{u}_R\mathrm{d}x \\
\leq& C\int_{\Omega_t^+} v\mathrm{d}x+\lambda_1\left(\int_{B_R}\left|\nabla\hat{u}_R\right|^2\mathrm{d}x\right)^{\frac{1}{2}}+\lambda_2\int_{B_R}\hat{u}_R\mathrm{d}x \\
\leq& C\int_{\Omega_t^+} v\mathrm{d}x+K_0\left(\lambda_1,\lambda_2\right),\ \text{a.e. } t\in[0,T],
\end{aligned}$$

where $K_0:\mathbb{R}_+\times\mathbb{R}_+\to\mathbb{R}_+$ is an increasing function, which is independent of t (and, in fact, depends on $\|\varphi\|_{H_0^1(\mathbb{R}^n)}$, $\|\varphi\|_{L^1(\mathbb{R}^n)}$, and a series of constants appeared in those lemmas and theorems we mentioned), with $K_0(\lambda_1,\lambda_2)\to 0$ as either $\lambda_1\to 0$ or $\lambda_2\to 0$. It implies that

$$\int_{\Omega_t^+} v\mathrm{d}x\leq\frac{\mathrm{e}^{Ct}-1}{C}K_0\left(\lambda_1,\lambda_2\right),\ \text{a.e. } t\in[0,T].$$

The result regarding Ω_t^- can be similarly derived. Therefore, we have

$$\int_{B_R}|v|\mathrm{d}x\leq\frac{\mathrm{e}^{Ct}-1}{C}K_0\left(\lambda_1,\lambda_2\right),\ \text{a.e. } t\in[0,T],$$

both sides of which are continuous w.r.t. t. Thus the conclusion is improved to hold in pointwise sense.

References

Beněs, V. E. (1981). Exact finite dimensional filters for certain diffusions with nonlinear drift stochastics. *Stochastics an International Journal of Probability and Stochastic Processes*, *5*, 65–92.

Brockett, R. W. (1981). *Nonlinear systems and nonlinear estimation theory* (pp. 441–477). Dordrecht: Springer.

Brockett, R. W., & Clark, J. M. C. (1980). *The geometry of the conditional density functions* (pp. 299–309). New York: Academic Press.

Chen, Xiuqiong, Ji Shi, & Yau, Stephen Shing-Toung (2019). Real-time solution of time-varying yau filtering problems via direct method and gaussian approximation. *IEEE Transactions on Automatic Control*, *64*(4), 1648–1654.

Chen, Xiuqiong, Luo, Xue, & Yau, Stephen Shing-Toung (2017). Direct method for time-varying nonlinear filtering problems. *IEEE Transactions on Aerospace and Electronic Systems*, *53*(2), 630–639.

Chen, Jie, & Yau, Stephen Shing-Toung (1996). Finite-dimensional filters with nonlinear drift. vi: Linear structure of ω. *Mathematics of Control, Signals, and Systems*, *9*(4), 370–385.

Chiou, Wen-Lin, & Yau, Stephen Shing-Toung (1994). Finite-dimensional filters with nonlinear drift. ii: Brockett's problem on classification of finite-dimensional estimation algebras. *SIAM Journal on Control and Optimization*, *32*(1), 297–310.

Diestel, J., & Uhl, J. J. (1977). *Vector measures*. Providence, Rhode Island: American Mathematical Society.

Dong, Wenhui, Luo, Xue, & Yau, Stephen Shing-Toung (2021). Solving nonlinear filtering problems in real-time by legendre galerkin spectral method. *IEEE Transactions on Automatic Control*, *66*(4), 1559–1572.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *An introduction to sequential Monte Carlo methods* (pp. 3–14). New York: Springer.

Duncan, T. E. (1967). Probability Densities for Diffusion Processes with Applications to Nonlinear Filtering Theory and Detection Theory. (Ph.D. thesis), Stanford, CA, USA: Stanford University.

Evans, Lawrence C. (2010). *Partial differential equations* (2nd ed.). Providence, Rhode Island: American Mathematical Society.

Gelb, A. (1974). *Applied optimal estimation*. Cambridge: The M.I.T. Press.

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar Signal Processing*, *140*(2), 107–113.

Ji Shi, Xiuqiong Chen, Dong, Wenhui, & Yau, Stephen Shing-Toung (2017). New classes of finite dimensional filters with non-maximal rank. *IEEE Control Systems Letters*, *1*(2), 233–237.

Ji Shi, Zhiyu Yang, & Yau, Stephen Shing-Toung (2018). Direct method for yau filtering system with nonlinear observations. *International Journal of Control*, *91*(3), 678–687.

Jiao, Xiaopei, & Yau, Stephen Shing-Toung (2020). New classes of finite dimensional filters with nonmaximal rank estimation algebra on state dimension n and linear rank $n-2$. *Yau, Stephen Shing-Toung*, *58*(6), 3413–3427.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, *83*, 95–108.

Li, Shijing, Wang, Zhongjian, Zhang, Zhiwen, & Yau, Stephen Shing-Toung (2013). Solving nonlinear filtering problems using a tensor train decomposition method. *IEEE Transactions on Automatic Control*, *68*(7), 4405–4412.

Luo, Xue, & Yau, Stephen Shing-Toung (2013a). Complete real time solution to the general nonlinear filtering problem without memory. *IEEE Transactions on Automatic Control*, *58*(10), 2563–2578.

Luo, Xue, & Yau, Stephen Shing-Toung (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.

Mitter, S. K. (1979). On the analogy between mathematical problems of nonlinear filtering and quantum physics. *Ricerche Automat.*, *10*(2), 163–216.

Mortensen, N. E. (1966). *optimal control of continuous-time stochastic systems* (Ph.D. thesis), Berkeley, CA, USA: University of California.

Picard, J. (1991). Efficiency of the extended kalman filter for nonlinear systems with small noise. *SIAM Journal on Applied Mathematics*, *51*(3), 843–885.

Rozovsky, B. L. (1972). Stochastic partial differential equations arising in nonlinear filtering problems. *Uspekhi Matematicheskikh Nauk*, *27*, 213–214.

Shi, Ji, & Yau, Stephen Shing-Toung (2017). Finite dimensional estimation algebras with state dimension 3 and rank 2. i: Linear structure of wong matrix. *SIAM Journal on Control and Optimization*, *55*(6), 4227–4246.

Wang, Zhongjian, Luo, Xue, & Yau, Stephen Shing-Toung (2020). Proper orthogonal decomposition method to nonlinear filtering problems in medium-high dimension. *IEEE Transactions on Automatic Control*, *65*(4), 1613–1624.

Wu, Xi, & Yau, Stephen Shing-Toung (2006). Classification of estimation algebras with state dimension 2. *SIAM Journal on Control and Optimization*, *45*(3), 1039–1073.

Yau, Stephen Shing-Toung (1994). Finite-dimensional filters with nonlinear drift. i: A class of filters including both kalman–bucy and beněs filters. *Journal of Mathematical Systems, Estimation, and Control*, *4*(2), 181–203.

Yau, Stephen Shing-Toung (2003). Complete classification of finite-dimensional estimation algebras of maximal rank. *International Journal of Control*, *76*(7), 657–677.

Yau, Stephen Shing-Toung, & Hu, Guo-Qing (2005). Classification of finite-dimensional estimation algebras of maximal rank with arbitrary state-space dimension and mitter conjecture. *International Journal of Control*, *78*(10), 689–705.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (1996). Explicit solution of a kolmogorov equation. *Applied Mathematics and Optimization*, *34*(3), 231–266.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (1997). Finite-dimensional filters with nonlinear drift iii: Duncan-mortensen-zakai equation with arbitrary initial condition for the linear filtering system and the beněs filtering system. *IEEE Transactions on Aerospace and Electronic Systems*, *33*(4), 1295–1308.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (2000). Real time solution to the nonlinear filtering problem without memory i. *Mathematical Research Letters*, *7*(6), 671–693.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (2004). Nonlinear filtering and time varying schrödinger equation. *IEEE Transactions on Aerospace and Electronic Systems*, *40*(1), 284–292.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (2005). Solution of filtering problem with nonlinear observations. *SIAM Journal on Control and Optimization*, *44*(3), 1019–1039.

Yau, Shing-Tung, & Yau, Stephen Shing-Toung (2008). Real time solution to the nonlinear filtering problem without memory ii. *SIAM Journal on Control and Optimization*, *47*(1), 163–195.

Yu, Hongyu, Jiao, Xiaopei, & Yau, Stephen Shing-Toung (2024). Complete classification of finite dimensional estimation algebras with state dimension n, linear rank $n-1$ and constant wong matrix. *IEEE Transactions on Automatic Control*, *69*(1), 295–302.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *11*, 230–243.

Minli Feng received the B.S. degree from the School of Mathematical Sciences, Peking University, Beijing, China, in 2017. He is currently pursuing the Ph.D. degree in applied mathematics with the Academy of Mathematics and Systems Science of the Chinese Academy of Sciences, Beijing, China. His research interests include nonlinear filtering.

Xiuqiong Chen received the B.S. degree from the School of Mathematical Sciences, Beihang University, Beijing, China, in 2014, and the Ph.D. degree in applied mathematics from the Department of Mathematical Sciences, Tsinghua University, Beijing, China, in 2019. From 2019 to 2021, she was a Postdoctoral Scholar at the Yau Mathematical Sciences Center, Tsinghua University. She joined Renmin University of China in 2021, where she is currently an Assistant Professor with the School of Mathematics. Her research interests include nonlinear filtering and deep learning.

Stephen S.-T. Yau (Life Fellow, IEEE) received the Ph.D. degree in mathematics from the State University of New York at Stony Brook, Stony Brook, NY, USA, in 1976.

He was a Member of the Institute of Advanced Study at Princeton, Princeton, NJ, USA, from 1976 to 1977 and from 1981 to 1982, and a Benjamin Pierce Assistant Professor with Harvard University, Cambridge, MA, USA, from 1977 to 1980. After that, he joined the Department of Mathematics, Statistics and Computer Science (MSCS), University of Illinois at Chicago (UIC), Chicago, IL, USA, and served for more than 30 years. During 2002–2011, he became a Joint Professor with the Department of Electrical and Computer Engineering, MSCS, UIC. After his retirement in 2012, he joined Tsinghua University, Beijing, China, where he is a Full-Time Professor with the Department of Mathematical Sciences. His research interests include nonlinear filtering, bioinformatics, complex algebraic geometry, Cauchy–Riemann (CR) geometry, and singularities theory.

Dr. Yau was a recipient of the Sloan Fellowship in 1980, the Guggenheim Fellowship in 2000, the AMS Fellow Award in 2013, and the AAIA Fellow in 2023. He was the General Chairperson of the IEEE International Conference on Control and Information, which was held in The Chinese University of Hong Kong, Hong Kong, in 1995. He is the Managing Editor and founder of the Journal of Algebraic Geometry since 1991, and the Editor-in-Chief and founder of Communications in Information and Systems from 2000 to the present. In 2005, he was entitled the UIC Distinguished Professor.