The Extended Direct Method for Generalized Time-varying Yau Filtering Systems

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

[a] *Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, 101408, P. R. China*

[b] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, P. R. China*

[c] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, P. R. China*

[d] *School of Mathematics, Renmin University of China, Beijing 100872, P. R. China*

[e] *Beijing Key Laboratory of Topological Statistics and Applications for Complex Systems, Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, 101408, P. R. China*

[f] *Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, P. R. China*

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.

Key words: Nonlinear filtering; Duncan-Mortensen-Zakai equation; Direct method; Time-varying Yau filtering system; Convergence analysis.

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-

⋆ Corresponding author. Tel: +86-10-62787874.

Email addresses: fengminli@bimsa.cn (Minli Feng), cxq0828@ruc.edu.cn (Xiuqiong Chen), yau@uic.edu (Stephen S.-T. Yau).

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) [16], which was published in 1960, and its continuous counterpart Kalman–Bucy filter [17]. 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) [13], and particle filter (PF) [10] [14]. In fact, EKF performs poorly when the 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 [11] [22] [40]. 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 [2] [3] [21]. 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 [4] [7] [36] [37] [38], and have been devoted to the study of non-maximal rank case [29] [27] [25] [15] [39].

The second is Yau-Yau algorithm proposed by Yau and Yau [32] [35]. 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 [20] [9], proper orthogonal decomposition [28], tensor training [18], etc.

The third is the direct method proposed by Yau and Yau [30] [31]. In [33], 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 [5] [26] [6]. In [6], 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 Guassian approximation proposed in [26], 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 [26], "the direct method is much stable and has theoretic convergence proof." The convergence of those using linear approximation is based on strict assumptions [23], and Example 3 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-poseness 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. Notations used throughout this paper are listed in Appendix A.

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 [11], 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 $U := \mathbb{R}^n \times [0,T]$:

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

where $G(t) := g(t)Q(t)g(t)^{\mathrm{T}}$. For each arrived observation, making an invertible exponential transformation [24]

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

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

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

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

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

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

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

and

$$K_k(x,t) := h(x,t)^{\mathrm{T}} S(t)^{-1} Y_{\tau_{k-1}}.$$

By [34], 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 equation (3).

In [19], 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 *[19, 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)

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

In [5] and [6], 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 $\tilde{\phi} : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$ 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 *[5, 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 $\tilde{v}_k$:

$$\begin{cases} \dfrac{\partial\tilde{v}_k}{\partial t} = \left(\dfrac{1}{2}\nabla^{\mathrm{T}}G\nabla - (L_1x + L_0)^{\mathrm{T}}\nabla \right. & \\ \qquad \left. +\tilde{p}\right)\tilde{v}_k & \text{in } U^{(k)}, \\ \tilde{v}_1(x,0) = \mathrm{e}^{-\phi(x,0)}\sigma_0(x) & \text{in } \mathbb{R}^n, \\ \tilde{v}_k\left(x,\tau_{k-1}\right) = \exp\left(h\left(x,\tau_{k-1}\right)^{\mathrm{T}} S\left(\tau_{k-1}\right)^{-1}\right. & \\ \qquad \left.\left(Y_{\tau_{k-1}} - Y_{\tau_{k-2}}\right)\right)\tilde{v}_{k-1}\left(x,\tau_{k-1}\right) & \\ \qquad \text{for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{6}$$

where

$$\begin{aligned}\tilde{p} :=& \frac{1}{2}\nabla^{\mathrm{T}}G\nabla\phi - \frac{1}{2}\nabla\phi^{\mathrm{T}}G\nabla\phi - (L_1x + L_0)^{\mathrm{T}}\nabla\phi \\ & - \Delta\tilde{\phi} - \frac{\partial\phi}{\partial t} - \frac{1}{2}h^{\mathrm{T}}S^{-1}h - \operatorname{tr}L_1.\end{aligned}$$

Proposition 3 *[6, 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:

$$\begin{cases} \dfrac{\partial v_k}{\partial t} = \left(\dfrac{1}{2}\Delta - (F_1x + F_0)^{\mathrm{T}}\nabla + p\right)v_k & \text{in } U^{(k)}, \\ v_1(z,0) = \mathrm{e}^{-\phi(B(0)z,0)}\sigma_0\left(B(0)z\right) & \text{in } \mathbb{R}^n, \\ v_k\left(z,\tau_{k-1}\right) = \exp\left(h\left(B\left(\tau_{k-1}\right)z,\tau_{k-1}\right)^{\mathrm{T}}\right. & \\ \qquad \left. S\left(\tau_{k-1}\right)^{-1}\left(Y_{\tau_{k-1}} - Y_{\tau_{k-2}}\right)\right) & \\ \qquad v_{k-1}\left(z,\tau_{k-1}\right) \text{ for } k \geq 2 & \text{in } \mathbb{R}^n, \end{cases} \tag{7}$$

where

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

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 [1] filtering systems as its special cases.

Proposition 4 *[6, 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

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

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 $A_2\left(\tau_{k-1}\right) = A_{2,\tau_{k-1}}$, $A_1\left(\tau_{k-1}\right) = A_{1,\tau_{k-1}}$ and $a_0\left(\tau_{k-1}\right) = a_{0,\tau_{k-1}}$, which satisfy the following system of

nonlinear ODEs, and A_2 is a symmetric-matrix-valued function:

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

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 [26], 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.

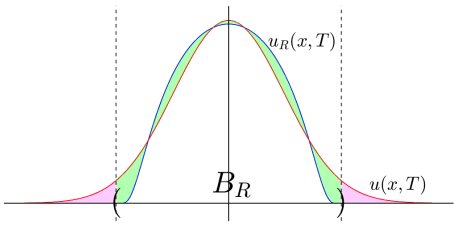

Fig. 1. Proof strategy diagram

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

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

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 does 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 [35], 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 Figure 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.

Algorithm 1 Extended Direct method

for $k \in \{1, \ldots, N\}$ **do**
  if $k = 1$ **then**
    Calculate $v_1(z, 0) = \mathrm{e}^{-\phi(B(0)z,0)}\sigma_0\left(B(0)z\right)$.
  else
    Calculate

$$\begin{aligned} v_k\left(z, \tau_{k-1}\right) = & \exp\Big(h\left(B\left(\tau_{k-1}\right)z, \tau_{k-1}\right)^{\mathrm{T}} \\ & S\left(\tau_{k-1}\right)^{-1}\left(Y_{\tau_{k-1}} - Y_{\tau_{k-2}}\right)\Big) \\ & v_{k-1}\left(z, \tau_{k-1}\right). \end{aligned}$$

  end if
  Get the Gaussian approximation of $v_k\left(z, \tau_{k-1}\right)$ via method proposed in [26]

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

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

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

    Solve (8) with the initial condition

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

    by Proposition 4.
  end for
  Calculate the approximate solution v_k to (7) $v_k \approx \sum_i^{N_k} \alpha_{k,i} v_{k,i}$.
  Calculate $\tilde{v}_k\left(x, \tau_k\right) = v_k\left(B\left(\tau_k\right)^{-1} x, \tau_k\right)$.
  Calculate $\tilde{u}_k\left(x, \tau_k\right) = \mathrm{e}^{\phi(x,\tau_k)} \tilde{v}_k\left(x, \tau_k\right)$.
  Calculate

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

  Calculate

$$\sigma\left(x, \tau_k\right) = \mathrm{e}^{h(x,\tau_k)^{\mathrm{T}} S(\tau_k)^{-1} Y_{\tau_k}} u_k\left(x, \tau_k\right).$$

  Calculate the conditional expectation of the state X_{τ_k}

$$E\left[X_{\tau_k} \mid \mathcal{Y}_{\tau_k}\right] \approx \frac{\displaystyle\int_{\mathbb{R}^n} x\sigma\left(x, \tau_k\right) \mathrm{d}x}{\displaystyle\int_{\mathbb{R}^n} \sigma\left(x, \tau_k\right) \mathrm{d}x}.$$

end for

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*

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

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

$$\int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}} u(x,T)\,\mathrm{d}x \leq \mathrm{e}^{C_1 T} \int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}} \varphi(x)\,\mathrm{d}x.$$

In particular, we have

$$\int_{\mathbb{R}^n \setminus B_R} u(x,T)\,\mathrm{d}x \leq \mathrm{e}^{-\sqrt{1+R^2}} \mathrm{e}^{C_1 T} \int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}} \varphi(x)\,\mathrm{d}x.$$

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

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

$$\begin{aligned} & \int_{B_R} \left(\mathrm{e}^{|x|^4/R^3 - 2|x|^2/R} - \mathrm{e}^{-R}\right) v(T)\,\mathrm{d}x \\ \leq & \frac{\mathrm{e}^{C_2 T} - 1}{C_2} C_3 \mathrm{e}^{-R} \mathrm{e}^{C_1 T} \int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}} \varphi(x)\,\mathrm{d}x. \end{aligned}$$

In particular, we have

$$\begin{aligned} & \int_{B_{R/2}} v(T)\,\mathrm{d}x \\ \leq & \frac{2\left(\mathrm{e}^{C_2 T} - 1\right)}{C_2} C_3 \mathrm{e}^{-9R/16} \mathrm{e}^{C_1 T} \int_{\mathbb{R}^n} \mathrm{e}^{\sqrt{1+|x|^2}} \varphi(x)\,\mathrm{d}x. \end{aligned}$$

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

Theorem 7 *Assume that*

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

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

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

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:

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

where $c \geq 0$ 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) $p(x,t) \leq 0, \forall (x,t) \in U$;
(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, so the seemingly lengthy assumptions, which are just a compilation of assumptions from all the preceding theorems, essentially only requires that the leading coefficient of p is sufficiently small. 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 $|D(x)| \leq d_0 + d_1|x|$, and $|\operatorname{div} D(x)| \leq e_0 + e_1|x| + e_2|x|^2$;
(4) w is a positive C^2 function with $\nabla w(x) = w(x)D(x)$, and $w\varphi, w\hat{\varphi}, x_i w\varphi, x_i w\hat{\varphi} \in \mathcal{I}_0$, where x_i is the i-th component of x.

Further assume that:

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

Write

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

Then for a given φ and any $\varepsilon > 0$, there exist positive numbers R and δ, such that for any $\hat{\varphi}$, we have $|x_{\mathrm{est}} - \hat{x}_{\mathrm{est}}| < \varepsilon$ 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 are 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) $|p(x,t) - \hat{p}(x,t)| \le \lambda_2, \forall (x,t) \in U_R$*,*

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*

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

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:

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

where:

- c is a constant;
- $T > 0$ is a fixed termination, with the sampling interval Δt for observations and its corresponding time sequence $\tau_k := k\Delta t, k = 0, 1, \ldots, N$;
- x_t, y_t^1, and y_t^2 are scalar stochastic process;
- 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 function is $\sigma_0 := \dfrac{\tilde{\sigma}_0}{\displaystyle\int_{\mathbb{R}} \tilde{\sigma}_0(x) \, \mathrm{d}x}$;
- $\tilde{\phi} : \mathbb{R} \times [0,T] \to \mathbb{R}$ is a C^∞ function.

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 estimation 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. The EKF was numerically implemented by the Euler's method. Besides, we approximate the dynamic system (13) by the Euler's method for the PF with different particles.

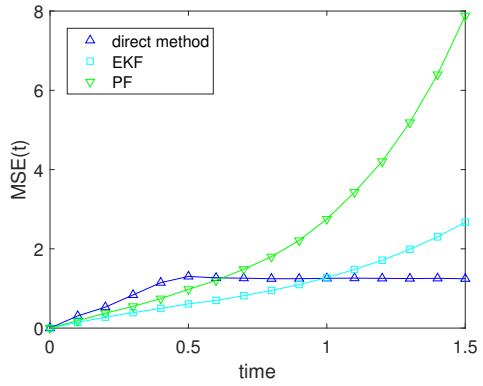

Fig. 2. MSE(t) based on 50 simulations of Example 1

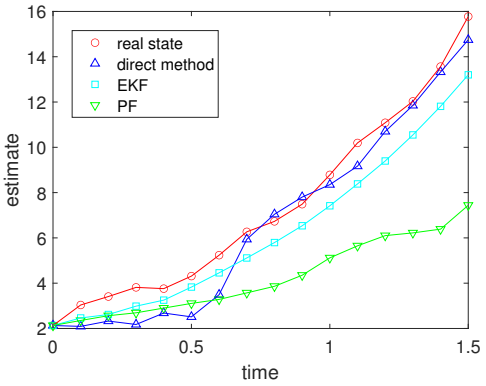

Fig. 3. A typical simulation of Example 1

Method	Time/s	MSE(T)
direct method	0.2229	1.2453
EKF	0.0002	2.6713
PF with 4600 particles	0.2210	7.8799

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

Example 1: We set $c = 1$, $T = 1.5$, $\Delta t = 0.1$ and $\tilde{\phi}(x,t) := 10^{-3}x^3t$. It can be computed that

$$\begin{aligned} p(z,t) = &-\left(4.5\times10^{-6}t^2\right)z^4 \\ &-\left(3\times10^{-3}t+10^{-3}\right)z^3 \\ &-\left(3\times10^{-3}t+\frac{1}{2\left(1+\sin(0.2t)\right)^2}\right)z^2 \\ &-\left(3\times10^{-3}t\right)z-1, \end{aligned}$$

which is approximated by

$$\begin{aligned} \hat{p}(z,t) = &-\left(3\times10^{-3}t+\frac{1}{2\left(1+\sin(0.2t)\right)^2}\right)z^2 \\ &-\left(3\times10^{-3}t\right)z-1 \end{aligned}$$

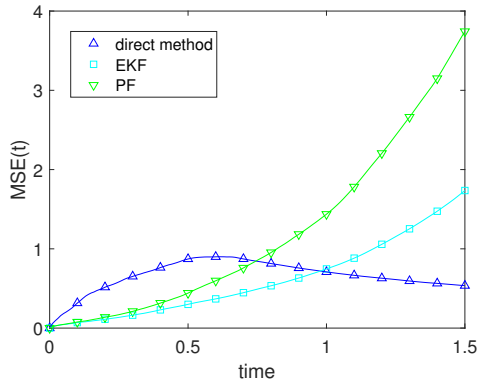

Fig. 4. MSE(t) based on 50 simulations of Example 2

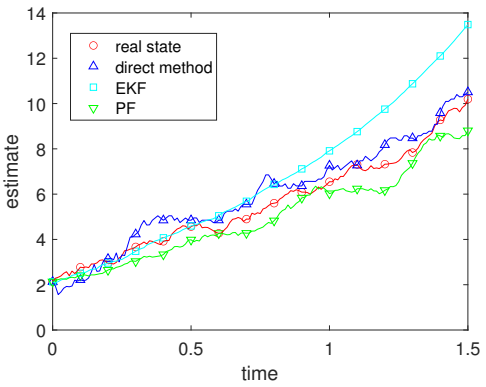

Fig. 5. A typical simulation of Example 2

Method	Time/s	MSE(T)
direct method	2.5498	0.5353
EKF	0.0015	1.7357
PF with 3800 particles	2.5401	3.7427

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

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 Figure 2, Figure 3 and Table 1. It can be seen that the direct method demonstrates a lower MSE(T) in comparison to the EKF, despite requiring a longer running time. When compared with the PF with 4600 particles, the direct method has a similar time consumption while exhibiting superior performance, which can be clearly seen in both MSE and trajectory in a typical simulation.

Example 2: We set $c = 1$, $T = 1.5$, $\Delta t = 0.01$ and

$\tilde{\phi}(x,t) := \mathrm{e}^{0.1x}$. It can be computed that

$$\begin{aligned} p(z,t) = &-0.005\mathrm{e}^{0.2z} - (0.1z + 0.105)\mathrm{e}^{0.1z} \\ &- \frac{1}{2\left(1+\sin(0.2t)\right)^2} z^2 - 1, \end{aligned}$$

which is approximated by

$$\begin{aligned} \hat{p}(z,t) = &-\left(0.010625 + \frac{1}{2\left(1+\sin(0.2t)\right)^2}\right) z^2 \\ &- 0.1115z - 1.11 \end{aligned}$$

when we apply Algorithm 1. Here, p no longer has a polynomial structure, but is again approximated well by $\hat{p}$.

The result is demonstrated in Figure 4, Figure 5 and Table 2. Similarly to the previous example, the direct method still performs best among these three methods in the example, where the trajectory of real state oscillates and is well tracked by the trajectory of estimation by the direct method.

Example 3: We set $c = -1$, $T = 15$, $\Delta t = 0.01$ and $\tilde{\phi}(x,t) := 0.01\sin\left(x^4 t\right)$. It can be computed that

$$\begin{aligned} p(z,t) = &0.08z^6t^2\sin\left(z^4t\right) - 0.0008z^6t^2\cos^2\left(z^3t\right) \\ &+ 0.04z^4t\cos\left(z^4t\right) - 0.01z^4\cos\left(z^4t\right) \\ &- 0.04z^3t\cos\left(z^4t\right) - 0.06z^2t\cos\left(z^4t\right) \\ &- \frac{1}{2\left(1+\sin(0.2t)\right)^2} z^2 + 1, \end{aligned}$$

which is approximated by

$$\hat{p}(z,t) = -\left(0.06t + \frac{1}{2\left(1+\sin(0.2t)\right)^2}\right) z^2 + 1$$

when we apply Algorithm 1. This time, the approximation for p results in fairly large deviation, on account of high nonlinearity of the system.

The result is demonstrated in Figure 6, Figure 7 and Table 3. In this example, result of the EKF does not shown due to its instability, which causes its trajectory to exceed the vertical axis scale and makes the rest of these figures difficult to read. Despite this, the direct method performs better than the PF in the long run. More specifically, the direct method has an average time cost near to that of the PF with 3200 particles, but it achieves the lowest MSE(t) when $t > 8$.

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 require-

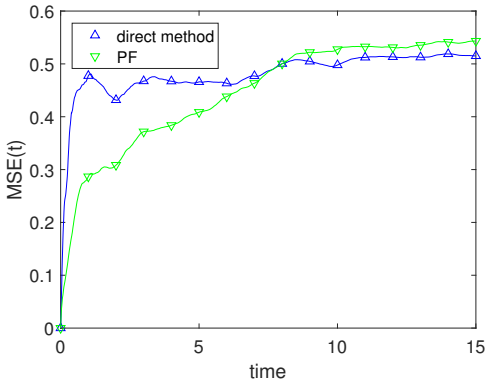

Fig. 6. MSE(t) based on 50 simulations of Example 3

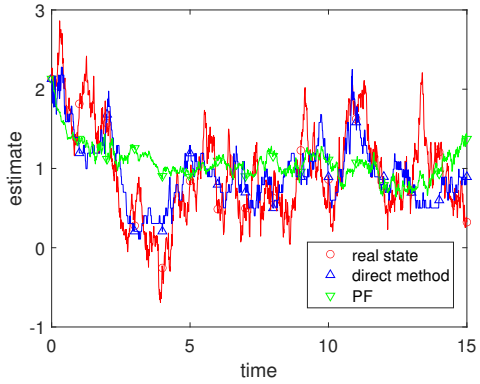

Fig. 7. A typical simulation of Example 3

Method	Time/s	MSE(T)
direct method	16.9354	0.5150
PF with 3200 particles	17.1910	0.5433

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

ments, 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.

A Notations

- **Operators:**
  $*^{\mathrm{T}}$: transposition of a matrix;
  $\nabla_x := \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)^{\mathrm{T}}$: gradient of a multivariate scalar function w.r.t. x, where the subscript is usually omitted if it causes no ambiguity (the same below);
  $\mathrm{div}_x := \nabla_x^{\mathrm{T}}$: divergence of a vector field w.r.t. x;
  $\Delta_x := \mathrm{div}_x \circ \nabla_x$: Laplace operator w.r.t. x;
  $\nabla_x^2 := \left(\frac{\partial^2}{\partial x_i \partial x_j}\right)_{i,j}$: Hessian matrix w.r.t. x;
  tr: trace of a square matrix;
  $|\cdot|$: Euclidean norm of a vector;
  $\|\cdot\|$: norm of a function, depending on its subscript.
- **Relations:**
  $\rightharpoonup$: the left side converges to the right side in weak topology;
  $\lesssim$: the left side is no more than the right side times a universal constant.
- **Subsets of Euclidean space:**
  $M_{n\times n}(\mathbb{R})$: spaces of all square matrices of order n with real entries;
  $U := \mathbb{R}^n \times [0,T]$;
  $U^{(k)} := \mathbb{R}^n \times [\tau_{k-1}, \tau_k]$;
  $B_R := \{x \in \mathbb{R}^n : |x| < R\}$;
  $\Gamma_R := \partial B_R \times [0,T]$.
- **Function Spaces:**
  $C^p(\Omega)$: space of all scalar functions on Ω whose derivatives of order no more than p all exist and are continuous, where Ω is an open connected subset of an Euclidean space (the same below);
  $C^\infty(\Omega)$: space of all scalar functions on Ω whose derivatives of arbitrary order all exist and are continuous;
  $C_c^\infty(\Omega)$: space of all C^∞ scalar functions on Ω whose support is compact in Ω;
  $L^p(\Omega)$: space of all scalar functions on Ω whose p-th power is absolutely integrable;
  $H^1(\Omega)$: space of all square-integrable scalar functions on Ω whose weak derivatives of order 1 are all square-integrable;
  $H_0^1(\Omega)$: completion of $C_c^\infty(\Omega)$ as a subspace of $H^1(\Omega)$;
  $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$: space of all functionals from $[0,T]$ to $H_0^1\left(\mathbb{R}^n\right)$ whose norm belongs to $L^p\left((0,T)\right)$ as a scalar function on $(0,T)$, where we use bold and light letters to represent its element and corresponding function from $[0,T]\times\mathbb{R}^n$ to $\mathbb{R}$, respectively;

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

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

- **Notations about probability theory:**
  $\mathcal{Y}_t := \sigma\{Y_s : 0\le s\le t\}$: sigma algebra generated by a family of random variables $\{Y_s : 0\le s\le t\}$;
  $\mathbb{E}\left[X_t \mid \mathcal{Y}_t\right]$: conditional expectation of a random variable X_t given a sigma algebra $\mathcal{Y}_t$.

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

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

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

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

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

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

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

Let $\boldsymbol{u}_R \in L^2\left(0,T;H_0^1\left(B_R\right)\right)$. Then its extension by zero

$$\left(\tilde{\boldsymbol{u}}_R(t)\right)(x) := \begin{cases} \left(\boldsymbol{u}_R(t)\right)(x), & \text{in } B_R, \\ 0, & \text{in } \mathbb{R}^n \backslash B_R \end{cases}$$

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

Lemma 13 *Assume that:*

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

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 $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$.*

PROOF. Fix $R > 0$. Write

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

Taking the derivative of E_0, we have

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

which implies

$$E_0(t) \le \mathrm{e}^{C_1 t} E_0(0) \le \mathrm{e}^{C_1 T} E_0(0). \tag{B.1}$$

Taking the derivative of E_1, we have

$$\begin{aligned} E_1'(t) =& \int_{B_R} 2\left(\nabla u_R\right)^{\mathrm{T}} \nabla \left(\frac{\partial u_R}{\partial t}\right) \mathrm{d}x \\ =& -\int_{B_R} 2\left(\Delta u_R\right) \frac{\partial u_R}{\partial t} \,\mathrm{d}x \\ =& -\int_{B_R} \left(\Delta u_R\right)^2 \,\mathrm{d}x + \int_{B_R} 2\Delta u_R F^{\mathrm{T}} \nabla u_R \,\mathrm{d}x \\ &- \int_{B_R} 2pu_R \Delta u_R \,\mathrm{d}x \\ \le& \int_{B_R} \left|F^{\mathrm{T}} \nabla u_R\right|^2 \,\mathrm{d}x - \int_{B_R} 2u_R (\nabla p)^{\mathrm{T}} \nabla u_R \,\mathrm{d}x \\ &+ \int_{B_R} 2p |\nabla u_R|^2 \,\mathrm{d}x \\ \le& \int_{B_R} \left(|F|^2 + 2p\right) |\nabla u_R|^2 \,\mathrm{d}x + \int_{B_R} (\Delta p) u_R^2 \,\mathrm{d}x \\ \le& C_2 E_1(t) + C_3 E_0(t) \\ \le& C_2 E_1(t) + C_3 \mathrm{e}^{C_1 T} E_0(0), \end{aligned}$$

which implies

$$\begin{aligned} E_1(t) &\le \mathrm{e}^{C_2 t} E_1(0) + \frac{\mathrm{e}^{C_2 t} - 1}{C_2} C_3 \mathrm{e}^{C_1 T} E_0(0) \\ &\le \mathrm{e}^{C_2 T} E_1(0) + \frac{\mathrm{e}^{C_2 T} - 1}{C_2} C_3 \mathrm{e}^{C_1 T} E_0(0). \end{aligned} \tag{B.2}$$

By (B.1) and (B.2), we have

$$\begin{aligned} &\|\boldsymbol{u}_R\|^2_{L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)} \\ =& \int_0^T \left(E_0(t) + E_1(t)\right) \mathrm{d}t \\ \le& T\left(\mathrm{e}^{C_1 T} E_0(0) + \mathrm{e}^{C_2 T} E_1(0) + \frac{\mathrm{e}^{C_2 T} - 1}{C_2} C_3 \mathrm{e}^{C_1 T} E_0(0)\right) \\ <& +\infty. \end{aligned}$$

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

PROOF. Thanks to [8, Chapter VII, Section 6], $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$ is a reflexive space, since $H_0^1\left(\mathbb{R}^n\right)$ has the Radon-Nikodym property. So any bounded subset of $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$ 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 $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$. 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 $L^2\left(0,T;H_0^1\left(\mathbb{R}^n\right)\right)$ 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

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

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*

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

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 $\tilde{U} := \mathbb{R}^n \times [0,T_0]$. The definition of a weak solution implies that for any $\boldsymbol{\psi} \in L^2(0,T_0; C_c^\infty(\mathbb{R}^n))$,

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

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

$$\begin{aligned}
&\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|} \boldsymbol{u}(T_0) \boldsymbol{\psi}(T_0) \, \mathrm{d}x \\
=& -\frac{1}{2} \int_{\tilde{U}} \mathrm{e}^{4c|x|} (\nabla \boldsymbol{u})^{\mathrm{T}} \nabla \boldsymbol{\psi} \, \mathrm{d}x \, \mathrm{d}t - 2c \int_{\tilde{U}} \mathrm{e}^{4c|x|} \boldsymbol{\psi} \frac{x^{\mathrm{T}}}{|x|} \nabla \boldsymbol{u} \, \mathrm{d}x \, \mathrm{d}t \\
&- \int_{\tilde{U}} \mathrm{e}^{4c|x|} \left(F^{\mathrm{T}} \nabla \boldsymbol{u} \right) \boldsymbol{\psi} \, \mathrm{d}x \, \mathrm{d}t + \int_{\tilde{U}} \mathrm{e}^{4c|x|} p\boldsymbol{u}\boldsymbol{\psi} \, \mathrm{d}x \, \mathrm{d}t \\
&+ \int_{\tilde{U}} \mathrm{e}^{4c|x|} \boldsymbol{u}\boldsymbol{\psi}' \, \mathrm{d}x \, \mathrm{d}t.
\end{aligned}$$

Approximating $\boldsymbol{u}$ by $\boldsymbol{\psi}$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^n} \mathrm{e}^{4c|x|} \boldsymbol{u}(T_0)^2 \, \mathrm{d}x \\
=& \int_{\tilde{U}} \mathrm{e}^{4c|x|} \left(-|\nabla \boldsymbol{u}|^2 - 2\boldsymbol{u} \left(F + c\frac{x}{|x|} \right)^{\mathrm{T}} \nabla \boldsymbol{u} \right. \\
&\left. + 2p\boldsymbol{u}^2 \right) \mathrm{d}x \, \mathrm{d}t \\
\leq& \int_{\tilde{U}} \mathrm{e}^{4c|x|} \left(\left| F + c\frac{x}{|x|} \right|^2 + 2p \right) \boldsymbol{u}^2 \, \mathrm{d}x \, \mathrm{d}t, \\
\leq& C \int_{\tilde{U}} \mathrm{e}^{4c|x|} \boldsymbol{u}^2 \, \mathrm{d}x \, \mathrm{d}t.
\end{aligned}$$

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 $CT_0 < CT < 1$. 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 of the three lemma above.

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 Lemma 13 and Lemma 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

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

which implies

$$\int_{B_R}\mathrm{e}^{\gamma}u_R\,\mathrm{d}x \le \mathrm{e}^{C_1 t}\int_{B_R}\mathrm{e}^{\gamma}\varphi\,\mathrm{d}x. \tag{B.3}$$

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

$$\begin{aligned}
\int_{\mathbb{R}^n}\eta_r u\,\mathrm{d}x &= \lim_{k\to+\infty}\int_{\mathbb{R}^n}\eta_r u_{R_k}\,\mathrm{d}x \le \overline{\lim_{k\to+\infty}}\int_{\mathbb{R}^n}\mathrm{e}^{\gamma}u_{R_k}\,\mathrm{d}x \\
&= \overline{\lim_{k\to+\infty}}\int_{B_{R_k}}\mathrm{e}^{\gamma}u_{R_k}\,\mathrm{d}x \\
&\le \overline{\lim_{k\to+\infty}}\,\mathrm{e}^{C_1 t}\int_{B_{R_k}}\mathrm{e}^{\gamma}\varphi\,\mathrm{d}x \\
&= \mathrm{e}^{C_1 t}\int_{\mathbb{R}^n}\mathrm{e}^{\gamma}\varphi\,\mathrm{d}x.
\end{aligned}$$

By monotone convergence theorem, we obtain

$$\int_{\mathbb{R}^n}\mathrm{e}^{\gamma}u\,\mathrm{d}x = \lim_{r\to+\infty}\int_{\mathbb{R}^n}\eta_r u\,\mathrm{d}x \le \mathrm{e}^{C_1 t}\int_{\mathbb{R}^n}\mathrm{e}^{\gamma}\varphi\,\mathrm{d}x,$$

which was to be shown.

PROOF. [of Theorem 6] Thanks to the maximum principle of solutions to second-order parabolic equations [12, 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

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{B_R}\eta v\,\mathrm{d}x \\
=&\int_{B_R}\eta\left(\frac{1}{2}\Delta v - F^{\mathrm{T}}\nabla v + pv\right)\,\mathrm{d}x \\
=&\frac{1}{2}\int_{B_R} v\Delta\eta\,\mathrm{d}x - \frac{1}{2}\int_{\partial B_R} v\frac{\partial\eta}{\partial\nu}\,\mathrm{d}S + \frac{1}{2}\int_{\partial B_R}\eta\frac{\partial v}{\partial\nu}\,\mathrm{d}S \\
&+\int_{B_R}\mathrm{div}\,(\eta F)\,v\,\mathrm{d}x - \int_{\partial B_R}\eta v F^{\mathrm{T}}\nu\,\mathrm{d}S(x) + \int_{B_R}\eta p v\,\mathrm{d}x \\
=&\frac{1}{2}\int_{B_R}\mathrm{e}^{\gamma}v\left(\Delta\gamma + |\nabla\gamma|^2\right)\,\mathrm{d}x \\
&+\int_{B_R}\mathrm{e}^{\gamma}v\left((\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F\right)\,\mathrm{d}x \\
&+\int_{B_R}\eta p v\,\mathrm{d}x \\
=&\int_{B_R}\eta v\left(\frac{1}{2}\Delta\gamma + \frac{1}{2}|\nabla\gamma|^2 + (\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F + p\right)\,\mathrm{d}x \\
&+\mathrm{e}^{-R}\int_{B_R} v\left(\frac{1}{2}\Delta\gamma + \frac{1}{2}|\nabla\gamma|^2 + (\nabla\gamma)^{\mathrm{T}}F + \mathrm{div}\,F\right)\,\mathrm{d}x \\
\le&C_2\int_{B_R}\eta v\,\mathrm{d}x + C_3\mathrm{e}^{-R}\int_{B_R}\mathrm{e}^{\sqrt{1+|x|^2}}v\,\mathrm{d}x \\
\le&C_2\int_{B_R}\eta v\,\mathrm{d}x + C_3\mathrm{e}^{-R}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}u\,\mathrm{d}x \\
\le&C_2\int_{B_R}\eta v\,\mathrm{d}x + C_3\mathrm{e}^{-R}\mathrm{e}^{C_1 T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi\,\mathrm{d}x,
\end{aligned}$$

which implies

$$\begin{aligned}
\int_{B_R}\eta v(x,T)\,\mathrm{d}x \le& \mathrm{e}^{C_2 T}\int_{B_R}\eta v(x,0)\,\mathrm{d}x \\
&+\frac{\mathrm{e}^{C_2 T}-1}{C_2}C_3\mathrm{e}^{-R}\mathrm{e}^{C_1 T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi\,\mathrm{d}x \\
=&\frac{\mathrm{e}^{C_2 T}-1}{C_2}C_3\mathrm{e}^{-R}\mathrm{e}^{C_1 T}\int_{\mathbb{R}^n}\mathrm{e}^{\sqrt{1+|x|^2}}\varphi\,\mathrm{d}x.
\end{aligned}$$

B.3 Estimation by the initial value

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

Lemma 17 *[35, 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

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

which implies

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

Similarly,

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

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

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

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

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

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

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

Thus we have

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

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 $\|u(x,T) - \hat{u}(x,T)\|_{L^1(\mathbb{R}^n)} < \varepsilon$.

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

$$\begin{aligned}&\tilde{F}(x,t) := F(x,t) + D(x), \\ &\tilde{p}(x,t) := p(x,t) + \frac{1}{2}|D(x)|^2 - \frac{1}{2}\operatorname{div} D(x) - F(x,t)^{\mathrm{T}}D(x).\end{aligned}$$

PROOF. The condition $\nabla w(x) = w(x)D(x)$ yields

$$\begin{aligned}\Delta w(x) &= \operatorname{div}(w(x)D(x)) \\ &= (\nabla w(x))^{\mathrm{T}} D(x) + w(x)\operatorname{div} D(x) \\ &= \left(|D(x)|^2 + \operatorname{div} D(x)\right) w(x), \\ w(x)\nabla u(x,t) &= \nabla v(x,t) - u(x,t)\nabla w(x) \\ &= \nabla v(x,t) - v(x,t)D(x).\end{aligned}$$

We compute

$$\begin{aligned}
&w(x)\Delta u(x,t)\\
=&\Delta v(x,t)-2\left(\nabla w(x)\right)^{\mathrm{T}}\nabla u(x,t)-\left(\Delta w(x)\right)u(x,t)\\
=&\Delta v(x,t)-2D(x)^{\mathrm{T}}\left(w(x)\nabla u(x,t)\right)\\
&-\left(|D(x)|^2+\operatorname{div}D(x)\right)v(x,t)\\
=&\Delta v(x,t)-2D(x)^{\mathrm{T}}\left(\nabla v(x,t)-v(x,t)D(x)\right)\\
&-\left(|D(x)|^2+\operatorname{div}D(x)\right)v(x,t)\\
=&\Delta v(x,t)-2D(x)^{\mathrm{T}}\nabla v(x,t)\\
&+\left(|D(x)|^2-\operatorname{div}D(x)\right)v(x,t).
\end{aligned}$$

Thus we have

$$\begin{aligned}
\frac{\partial v}{\partial t}(x,t)=&\frac{1}{2}w(x)\Delta u(x,t)-F(x,t)^{\mathrm{T}}\left(w(x)\nabla u(x,t)\right)\\
&+p(x,t)v(x,t)\\
=&\frac{1}{2}\Delta v(x,t)-\tilde{F}(x,t)^{\mathrm{T}}\nabla v(x,t)+\tilde{p}(x,t)v(x,t).
\end{aligned}$$

PROOF. [of Theorem 9] It is straightforward to compute that

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

Thus it is sufficient to show that

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

are sufficiently small, since $\int_{\mathbb{R}^n}wu(x,T)\,\mathrm{d}x$ and x_{est} are fixed when φ is given. Take $\left\|w\left(u(x,T)-\hat{u}(x,T)\right)\right\|_{L^1(\mathbb{R}^n)}$ as an example: By Theorem 8 and Lemma 18, $\left\|w\left(u(x,T)-\hat{u}(x,T)\right)\right\|_{L^1(\mathbb{R}^n)}$ can be made arbitrarily small if $\left\|w\left(\varphi-\hat{\varphi}\right)\right\|_{L^1(B_R)}$ 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

$$\left\|w\left(u(x,T)-\hat{u}(x,T)\right)\right\|_{L^1(\mathbb{R}^n)}\leq\sup_{x\in\overline{B}_R}|w(x)|\left\|\varphi-\hat{\varphi}\right\|_{L^1(B_R)}$$

can be made arbitrarily small if $\left\|\varphi-\hat{\varphi}\right\|_{L^1(B_R)}$ is sufficiently small, since w is fixed and so is $\sup_{x\in\overline{B}_R}|w(x)|$. All that remains is a direct calculation for translating Assumptions 1)-9) in Theorem 8 to Assumptions 5)-8) in this theorem:

(1) It reduces to 5) and 6).
(2)
$$\begin{aligned}
\left|\tilde{F}\right|^2+2\tilde{p}=&|F|^2+2p+2|D|^2+\operatorname{div}D\\
\lesssim&x^{\mathrm{T}}\left(|F_1|^2+2p_2+2d_1^2+e_2\right)Ix,
\end{aligned}$$
which coincides with Assumption 5) in this theorem.
(3) It coincides with Assumption 6) in this theorem.
(4) It reduces to 2), since $cx/|x|$ is bounded. It is why the constant c in Theorem 8 can be set to 0 in this theorem.
(5) It reduces to 7).
(6) It coincides with Assumption 7) in this theorem.
(7)
$$\begin{aligned}
&2\left(4|x|\left|\tilde{F}\right|+\operatorname{div}\tilde{F}+\tilde{p}\right)\\
=&8|x||F+D|+2\operatorname{div}F+2p+|D|^2+\operatorname{div}D\\
&-2F^{\mathrm{T}}D\\
\lesssim&x^{\mathrm{T}}\left(8\left(|F_1|+d_1\right)+2p_2+d_1^2+e_2+2\left|F_1\right|d_1\right)Ix,
\end{aligned}$$
which coincides with Assumption 8) in this theorem.
(8) It is trivial, since $12+2n+4|x||F|+\operatorname{div}F$ can be estimated by a quadratic polynomial.
(9) It reduces to 5).

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*

$$\begin{aligned}
&\int_{B_R}|\nabla u_R|^2\,\mathrm{d}x\\
\leq&\mathrm{e}^{C_2T}\int_{B_R}\varphi^2\,\mathrm{d}x+\frac{\mathrm{e}^{C_2T}-1}{C_2}C_3\mathrm{e}^{C_1T}\int_{B_R}|\nabla\varphi|^2\,\mathrm{d}x.
\end{aligned}$$

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 Lemma 17 and Lemma 19, we have

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

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

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

Similarly,

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

Therefore, we have

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

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

Acknowledgements

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

References

[1] V. E. Beněs. Exact finite dimensional filters for certain diffusions with nonlinear drift stochastics. *Stochastics An International Journal of Probability and Stochastic Processes*, 5:65–92, 1981.

[2] R. W. Brockett. *Nonlinear Systems and Nonlinear Estimation Theory*, pages 441–477. Springer, Dordrecht, 1981.

[3] R. W. Brockett and J. M. C. Clark. *The Geometry of the Conditional Density Functions*, pages 299–309. Academic Press, New York, 1980.

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

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

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

[7] Wen-Lin Chiou and Stephen Shing-Toung Yau. 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, 1994.

[8] J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, Rhode Island, 1977.

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

[10] A. Doucet, N. de Freitas, and N. Gordon. *An Introduction to Sequential Monte Carlo Methods*, pages 3–14. Springer, New York, 2001.

[11] T. E. Duncan. Probability densities for diffusion processes with applications to nonlinear filtering theory and detection theory, 1967.

[12] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, 2nd edition, 2010.

[13] A. Gelb. *Applied Optimal Estimation*. The M. I. T. Press, Cambridge, 1974.

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

[15] Xiaopei Jiao and Stephen Shing-Toung Yau. 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, 2020.

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

[17] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83:95–108, 1961.

[18] Shijing Li, Zhongjian Wang, Zhiwen Zhang, and Stephen Shing-Toung Yau. Solving nonlinear filtering problems using

a tensor train decomposition method. *IEEE Transactions on Automatic Control*, 68(7):4405–4412, 2013.

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

[20] Xue Luo and Stephen Shing-Toung Yau. 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, 2013.

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

[22] N. E. Mortensen. Optimal control of continuous-time stochastic systems, 1966.

[23] J. Picard. Efficiency of the extended kalman filter for nonlinear systems with small noise. *SIAM J. Appl. Math.*, 51(3):843–885, 1991.

[24] B. L. Rozovsky. Stochastic partial differential equations arising in nonlinear filtering problems. *Uspekhi Matematicheskikh Nauk*, 27:213–214, 1972.

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

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

[27] Ji Shi and Stephen Shing-Toung Yau. 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, 2017.

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

[29] Xi Wu and Stephen Shing-Toung Yau. Classification of estimation algebras with state dimension 2. *SIAM J. Control and Optimization*, 45(3):1039–1073, 2006.

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

[31] Shing-Tung Yau and Stephen Shing-Toung Yau. 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, 1997.

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

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

[34] Shing-Tung Yau and Stephen Shing-Toung Yau. Solution of filtering problem with nonlinear observations. *SIAM J. Control and Optimization*, 44(3):1019–1039, 2005.

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

[36] Stephen Shing-Toung Yau. 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, 1994.

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

[38] Stephen Shing-Toung Yau and Guo-Qing Hu. 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, 2005.

[39] Hongyu Yu, Xiaopei Jiao, and Stephen Shing-Toung Yau. 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, 2024.

[40] M. Zakai. On the optimal filtering of diffusion processes. *Z. Wahrsch. Verw. Gebiete*, 11:230–243, 1969.