Estimation of the Linear System via Optimal Transportation and Its Application for Missing Data Observations

Jiayi Kang[†], Xiaopei Jiao[†] and Stephen S.-T. Yau[*], *Life Fellow, IEEE*

***Abstract*—In this paper, an optimal transportation particle method has been proposed to deal with data fusion smoothing problem. The proposed method can handle prediction, filtering, and smoothing problems uniformly more robustly and stably than traditional algorithms. Our main idea is to approximate the trajectory in Wasserstein space which is the set of probability distributions equipped with the Wasserstein metric. Recent literature has demonstrated the successful application of optimal transportation for prediction and filtering problems. In our paper, we derive an optimal transportation particle for solving the smoothing problem utilizing Mayne - Fraser's formula [1], [2]. Detailed convergence results are presented, and the proposed algorithms are tested on missing observation processes, showcasing their ability to solve hybrid data fusion problems. This work introduces a new approach to particle methods that expands their possibilities in data fusion applications.**

***Index Terms*—Kalman Filtering, Estimation, Optimal transportation, Stability of linear systems.**

I. Introduction

Data fusion [3] combines information from different sensors to achieve a more accurate representation of a quantity of interest. It is widely used in integrated navigation systems for maneuvering targets, such as airplanes, ships, cars, and robots. The state estimation problem, which aims to determine the state of a target under some observations, is central to data fusion. Such state estimation can be considered as three different types [4], [5], which are prediction, filtering, and smoothing.

Kalman proposed the well-known Kalman filter in 1960 [6], which is an optimal linear estimator for filtering problems. Soon, Rauch-Tung-Striebel's (RTS) optimal smoothing algorithm was proposed for smoothing problems [7]. The Mayne-Fraser two-filter (MFTF) formula is a prominent example in systems and control, where forward and backward filters are merged into a single estimate for fixed-interval smoothing problems [8], [9]. In 1979 [10], the forward-backward duality of MFTF was proposed. Using this duality, the interpolation formula in [11] was designed for a single interval of loss for observation cases. A recent breakthrough in [12] is that the forward-backward duality holds for any intermittent observation structure. This can be achieved by using a cascade of continuous and discrete-time forward and backward Kalman filters, which depends on the assumed information pattern.

Since the 1990s, Monte Carlo methods have been applied to estimation problems, which lead to the development of particle prediction (sampling), particle filtering (PF), and particle smoothing (PS) [13], [14]. Compared to the Kalman filter (KF) and Rauch-Tung-Striebel (RTS), PF and PS are more flexible in practical applications. However, these particle methods suffer from issues like particle degeneracy, which means that only a few particles have large weights, in numerical implementations. Ensemble methods were proposed to overcome particle degeneracy by setting equally weighted particles. The feedback particle filter (FPF) [15], [16] is a type of particle filter that directly samples from posterior distributions of filtering problems using a controlled interacting particle system. The ensemble Kalman filter (EnKF) [17], [18], [19] and ensemble Kalman smoother (EnKS) [20], [21] are widely used in various applications such as weather prediction, earth physics, and industry. To elucidate their efficiency, it is essential to establish a framework for these ensemble algorithms. In the case of linear-Gaussian problems, the mean-field limit of EnKF and EnKS algorithms have been demonstrated in [19], [21], [22]. The mean-field limits provide an insight into the behavior of high-dimensional stochastic models by approximating the original models with a simpler model that can capture the main structure of the original high-dimensional model. Therefore, this paper will focus only on the estimation problems with linear-Gaussian settings. However, even in these settings, applying the EnKF and EnKS algorithms to very high-dimensional problems (such as 10^7) presents significant computational challenges.

A unified framework for the sampling process was proposed in [23], which views different sampling processes as various paths in Wasserstein space (WS). However, Taghvaei highlighted in [22] that infinitely many flows can correspond to the same paths in WS, leading to non-uniqueness issues in filtering cases. Taghvaei and Mehta addressed this for filtering problems by establishing an optimal transportation (OT) formulation for EnKF [24] [22]. This method was recently extended to linear systems with correlated noises in [25]. This work aims to generalize Taghvaei's framework to prediction and smoothing problems. Inspired by the McKean-Vlasov stochastic differential equation (SDE) which is a type

[†] Equal contribution; [*] Corresponding author.

This work was supported by the Tsinghua University Education Foundation Fund (042202008).

Jiayi Kang is with Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China. (E-mail: kangjiayi@bimsa.cn)

Xiaopei Jiao is with the Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China. (E-mail: jiaoxiaopei@bimsa.cn).

Stephen S.-T. Yau is with with the Department of Mathematical Sciences, Tsinghua University, and the Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China. (E-mail: yau@uic.edu).

of controlled SDE, our work proposes a controlled ODE called the tangent flow. The tangent flow is closely related to the gradient flow. The tangent and gradient flow share a gradient form; however, the gradient flow represents a probabilistic trajectory derived from optimizing a specific energy functional [26]. In contrast, the tangent flow can be constructed for more general partial differential equation (PDE) and stochastic partial differential equation (SPDE) systems. Unlike the McKean-Vlasov SDE, the tangent flow system does not require additional random terms, which can achieve better numerical accuracy than traditional stochastic methods. From a mathematical perspective, the filtration of the tangent flow is identical, meaning our tangent flow does not introduce additional uncertainty.

We make the following contributions in this paper:

- Based on optimal transportation, novel particle-based algorithms have been first proposed for linear systems including prediction, filter, and smoothing problems.
- By applying the MFTF formula, OTPS can be formulated as a bi-directional filter process that can deal with smoothing problems with missing observation.
- Rigorous convergence analysis of new algorithms has been proposed well.

Notations: Let $\|\cdot\|_2$ represent the Euclidean norm of the vectors on $\mathbb{R}^n$, and $\mathbb{S}_n^+$ represent the set of all $n \times n$ real positive-defined symmetric matrices. $\mathrm{Tr}(A)$ is the matrix-trace of A. $\lambda_{min}(A)$ and $\lambda_{max}(A)$ are the minimum and maximum eigenvalues of the matrix A, respectively. For a matrix $A \in \mathbb{S}^n$, $\|A\|_F := \mathrm{Tr}(AA^\top)$ denotes the Frobenius norm, $\|A\|_2 := \sqrt{\lambda_{max}(AA^\top)}$ is the spectral norm. For any $A, B \in \mathbb{S}^n$, we denote $A > B$ if $A - B$ is positive definite. A Gaussian probability distribution with mean μ and covariance P will be denoted as $\mathcal{N}(\mu, P)$. We denote $\circ$ as Stratonovich integral. We define two new operators, $\bar{\nabla}(*)$ and $\bar{\nabla}\cdot(*)$ as follows,

$$\bar{\nabla}(\varphi_0(t,x)) := \left(\nabla\varphi_0^1(t,x), \quad \cdots, \quad \nabla\varphi_0^m(t,x)\right),$$

where $\varphi_0(t,x) := \left(\varphi_0^1(t,x), \quad \cdots, \quad \varphi_0^m(t,x)\right)$ is a m dimensional row vector value functions and $\nabla\varphi_0^i(t,x)$ is the divergence of the $\varphi_0^i(t,x)$ with $1 \le i \le m$;

$$\bar{\nabla}\cdot(\mathcal{K}(t,x)) := \left(\nabla\cdot(\mathcal{K}^1(t,x)), \quad \cdots, \quad \nabla\cdot(\mathcal{K}^m(t,x))\right),$$

where $\mathcal{K}(t,x) := (\mathcal{K}^1(t,x), \cdots, \mathcal{K}^m(t,x))$ and $\mathcal{K}^i(t,x)$ with $1 \le i \le m$ are all n dimensional column vector fields.

Combining the two new operators, we can define $\bar{\Delta}(*)$ as

$$\bar{\Delta}\varphi_0(t,x) := \bar{\nabla}\cdot(\bar{\nabla}\varphi_0(t,x)) = \left(\Delta\varphi_0^1(t,x), \quad \cdots, \quad \Delta\varphi_0^m(t,x)\right).$$

II. Background and Preliminary

In this paper, we shall focus on the general linear dynamical system as follows,

$$\begin{aligned} dx_t &= A(t)x_t dt + B(t)dv_t, \\ dy_t &= H(t)x_t dt + D(t)dv_t, \end{aligned} \tag{1}$$

where x_t, y_t are n dimensional and m dimensional stochastic processes, respectively, v_t is $p-$dimensional standard Brownian motion and $A(t) \in \mathbb{R}^{n\times n}$, $B(t) \in \mathbb{R}^{n\times p}$, $H(t) \in \mathbb{R}^{m\times n}$, $D(t) \in \mathbb{R}^{m\times p}$ are all smooth and matrix-value functions of time t. And we denote the $\mathcal{Y}_t := \sigma(y_s, 0 \le s \le t)$ as the $\sigma-$algebra generated by $(y_s)_{0\le s\le t}$.

***Remark** 2.1:* (1) is with independent noise if $BD^\top = 0$. Otherwise, (1) is with correlated noise.

Next, we shall introduce the DMZ equation, a SPDE, for (1) in filtering theory.

$$d\sigma(t,x) = L_0\sigma(t,x) + \mathcal{H}(\sigma(t,x)) \circ dy_t \ \forall t \in [0,T], \tag{2}$$

where

$$\begin{aligned} L_0(\cdot) :=& \frac{1}{2}\sum_{i,j=1}^{n}\left([BB^\top]_{i,j}(t)\right)\frac{\partial^2}{\partial x_i \partial x_j}(\cdot) - (A(t)x)^\top\nabla\cdot(\cdot) \\ &- \frac{1}{2}(H(t)x)^\top(DD^\top)^{-1}(t)H(t)x\cdot(\cdot) - \mathrm{Tr}(A(t))(\cdot), \end{aligned}$$

and the $\mathcal{H}(\cdot) := x^\top H^\top(t) \times (\cdot) - BD^\top(t)\cdot \mathrm{div}(\cdot)^\top$.

A. Optimal transportation

Firstly, we introduce the basic concepts of the OT. Let α and β be two probability measures on measure spaces Ω_X and Ω_Y, respectively. $\mathcal{P}(\Omega)$ denotes the set of probability measures on Ω. Let $c : \Omega_X \times \Omega_Y \to \mathbb{R}^+$ be a cost function and $c(x,y)$ measures the cost of transporting one unit of mass from $x \in \Omega_X$ to $y \in \Omega_Y$. The transport map is defined below.

***Definition** 2.1:* Let $\alpha \in P(\Omega_X)$ and $\beta \in P(\Omega_Y)$. We say that T is a transport map from α to β if

$$\beta(B) = \alpha\left(\mathcal{T}^{-1}(B)\right) \qquad \text{for all } \beta\text{-measurable sets } B. \tag{3}$$

Equivalently we write $\beta = \mathcal{T}_\#\alpha$.
Monge's problem is formulated as follows:

***Theorem** 2.1 (Monge's optimal transportation problem [27]):* Given $\alpha \in \mathcal{P}(\Omega_X)$, $\beta \in \mathcal{P}(\Omega_Y)$, let

$$I[\mathcal{T}] = \int_{\Omega_X} c(x, \mathcal{T}(x))\mathrm{d}\alpha(x), \tag{4}$$

where $\mathcal{T} : \Omega_X \to \Omega_Y$ is a transport map from α to β, i.e. $\beta = \mathcal{T}_\#\alpha$. Then Monge's optimal transportation problem is to minimize the above integral among all transport maps from α to β.

If we further assume that the density functions of α and β are C^2 smooth, the optimal transportation map is the gradient form of some function Φ, i.e. $\nabla\Phi(x) = \mathcal{T}(x)$. The function Φ is determined by the following PDE which is the so-called Monge-Ampère equation [27],

$$\det\nabla^2\Phi(x) = \frac{\beta_0(x)}{\alpha_0(\nabla\Phi(x))}, \tag{5}$$

where $\nabla^2\Phi(x)$ is the Hessian matrix and α_0, β_0 are the density functions of α and β, respectively [28].

***Definition** 2.2:* Let $(\mathbb{R}^n, d)$ be a standard Eulidean space and denote $\mathcal{P}_2(\mathbb{R}^n)$ as the set of all probability measures μ on $\mathbb{R}^n$ satisfying $\int_{\mathbb{R}^n} d^2(x, x_0)d\mu(x) < \infty$ for some $x_0 \in \mathbb{R}$. The $p-$Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^n)$ is defined by

$$W_2(\mu,\nu) = \left(\inf_{\gamma\in\Gamma[\mu,\nu]}\int_{\mathbb{R}^n\times\mathbb{R}^n} d^2(x_1,x_2)d\gamma(x_1,x_2)\right)^{\frac{1}{2}}, \tag{6}$$

where $\Gamma[\mu,\nu]$ is the set of joint measures on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals μ and ν.

B. The forward Fokker-Planck equation

***Lemma** 2.1 (Fokker-Planck equation):* A general controlled SDE is formulated as follows

$$dx_t = \mathcal{U}(t, x_t)dt + \mathcal{K}(t, x_t) \circ dI_t, \tag{7}$$

where $x_t \in \mathbb{R}^n$, I_t is $m-$dimensional Brownian motion, the $\mathcal{U}(t, x)$: $\mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function, $\mathcal{K}(t, x) := (\mathcal{K}^1(t, x), \cdots, \mathcal{K}^m(t, x))$ and $\mathcal{K}^i(t, x)$ with $1 \leq i \leq m$ are all n dimensional column vector fields.

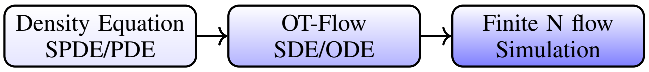
Then the forward Fokker-Planck density equation of (7) is determined by the following SPDE,

$$\begin{aligned} dp &= -\nabla \cdot (\mathcal{U}(t, x_t)p(t, x))dt \\ &- \bar{\nabla} \cdot (\mathcal{K}(t, x)p(t, x)) \circ dI_t. \end{aligned} \tag{8}$$

where $\bar{\nabla} \cdot (\mathcal{K}(t, x)p(t, x)) := (\nabla \cdot \left(\mathcal{K}^1(t, x_t)p(t, x)\right), \cdots, \nabla \cdot (\mathcal{K}^m(t, x)p(t, x)))$.

III. The tangent flow motivated by OT

The posterior of a continuous stochastic system, characterized by PDE or SPDE, traces a trajectory in WS once initialized [23]. The challenge lies in uniquely defining a flow via OT corresponding to this trajectory. In this section, we shall answer this question by proposing a new concept, which is called tangent flow. The graph of workflow is given as follows.



Here, the main steps of the proposed framework are as follows:

1) Determine the SPDE or PDE for the solution of the problem. For the prediction problem, the posterior distribution is governed by the Fokker-Planck equation, a PDE. The filtering problem is governed by the Kushner-Stratonovich equation, an SPDE. In the smoothing problem, the estimation can be viewed as a linear combination of two filters.
2) Construct the corresponding tangent flow for such PDE or SPDE. The detailed definitions are given in the following subsections. Theoretically, the tangent flow will not introduce any error because it admits the same density evolution dynamics.
3) Use finite particles to simulate the tangent flow we have, which is given in section IV.

A. Optimal transportation in PDE

In this subsection, we shall first derive the tangent flow for the PDE (9) based on the Monge-Ampère equation in OT.

$$\frac{\partial p}{\partial t} = \mathcal{D}(p), \quad t \in [0, S], \tag{9}$$

where $\mathcal{D}(\cdot)$ is some differential operator and $p(t, \cdot) \in \mathcal{P}_2(\mathbb{R}^d)$.

***Definition** 3.1:* The tangent flow of the PDE system (9) is defined as

$$dx_t = \nabla\varphi_1(t, x)dt, \tag{10}$$

where $x_0 \sim p(0, x)$. And $\varphi_1(t, x)$ is solution of the following PDE,

$$-p(t, x)\Delta\varphi_1(t, x) - \nabla(p(t, x)) \cdot \nabla\varphi_1(t, x) = \mathcal{D}(p(t, x)). \tag{11}$$

The tangent flow of the PDE system is motivated by many related works such as gradient flow [29], [30]. Here, we provide a new way to construct the tangent flow based on the Monge-Ampère equation and use the PDE expansion technique.

***Remark** 3.1:* In this remark, we will explain the well-defined property of the tangent flow of PDE. It is noticed that the posterior distribution of x_t governed by (10) satisfied the following Fokker-Planck equation,

$$\frac{\partial p(t, x)}{\partial t} = -\nabla \cdot (\nabla\varphi_1(t, x)p(t, x)). \tag{12}$$

And combining with the (11), we have

$$\begin{aligned} \frac{\partial p(t, x)}{\partial t} &= -\nabla \cdot (\nabla\varphi_1(t, x)p(t, x)) \\ &= -\Delta\varphi_1(t, x)p(t, x) - \nabla\varphi_1(t, x) \cdot \nabla p(t, x) \\ &= \mathcal{D}(p(t, x)). \end{aligned} \tag{13}$$

We have shown that the distribution of dynamical system (10) satisfies equation (9).

Next, we shall prove that the tangent flow is the unique flow in the sense of OT.

***Theorem** 3.1:* The tangent flow of the PDE system (9) is the deterministic and unique flow that corresponds to OT.

The proof is given in the appendix.

B. Optimal transportation in SPDE

In this subsection, we shall extend the tangent flow to the situation where density evolution satisfies a SPDE,

$$dp = \mathcal{D}(p)dt + \mathcal{H}(p) \circ dI_t \quad t \in [0, S]. \tag{14}$$

Here the $\mathcal{D}(\cdot)$ and $\mathcal{H}(\cdot)$ are some differential operators, and I_t is $m-$dimensional Gaussian process.

Similarly to the case of PDE, the tangent flow of the SPDE can be defined naturally as below.

***Definition** 3.2:* Let $p(t, x)$ be the solution of (14). Then, the tangent flow for general SPDE (14) is defined as

$$dx_t = \nabla\varphi_1(t, x_t)dt + \bar{\nabla}\varphi_0(t, x_t) \circ dI_t. \tag{15}$$

Here $\varphi_0 = (\varphi_0^1, \cdots, \varphi_0^m)$ is the vector value function and satisfies the following equations,

$$-p(t, x)\bar{\Delta}\varphi_0(t, x) - \nabla(p(t, x)) \cdot \bar{\nabla}\varphi_0(t, x) = \mathcal{H}(p), \tag{16}$$

where $\mathcal{H}$ is the m dimensional row vector operator. Then, φ_1 is the scale function and satisfies the following equation

$$-p(t, x)\Delta\varphi_1(t, x) - \nabla(p(t, x)) \cdot \nabla\varphi_1(t, x) = \mathcal{D}(p(t, x)). \tag{17}$$

***Remark** 3.2:* In this remark, we shall explain the above property of the tangent flow of SPDE is well-defined. It is noticed that the posterior distributions of x_t in (15) satisfy the forward Fokker-Planck equation in Lemma 2.1,

$$\begin{aligned} dp(t, x) &= -\nabla \cdot (\nabla\varphi_1(t, x)p(t, x))dt \\ &- \bar{\nabla} \cdot (\bar{\nabla}\varphi_0(t, x_t)p(t, x)) \circ dI_t. \end{aligned} \tag{18}$$

And combining with the (11), we have

$$\begin{aligned} dp(t,x) &= -\nabla \cdot (\nabla \varphi_1(t,x)p(t,x))dt \\ &\quad - \bar{\nabla} \cdot (\nabla \varphi_1(t,x)p(t,x)) \circ dI_t \\ &= (-\Delta \varphi_1(t,x)p(t,x) - \nabla \varphi_1(t,x) \cdot \nabla p(t,x))\, dt \\ &+ \big(-\bar{\Delta}\varphi_0(t,x)p(t,x) - \bar{\nabla}\varphi_0(t,x) \cdot \nabla p(t,x)\big) \circ dI_t \\ &= \mathcal{D}(p(t,x))dt + \mathcal{H}(p(t,x)) \circ dI_t. \end{aligned}$$

So, we show that the distributions of dynamical system (15) satisfy the (14).

Similarly, we can have the following theorem.

***Theorem** 3.2:* The tangent flow of the SPDE system (14) is the unique flow corresponding to OT.

The proof is given in the appendix.

***Remark** 3.3:* Here, in Theorem 3.2, we use an important technique to approximate the SPDE by a sequence of PDE. Such an idea is motivated by the recent works on solving filtering problems via OT [31]. And the mathematical foundations of this procedure are given as Wong-Zakai approximation [32], [33]. And we provide a modern version for the reference of this topic [34].

C. The geometry understanding of tangent flow

In this section, we aim to provide a geometric understanding of the proposed tangent flow. To do so, we first recall the tangent space of a probability distribution q in $\mathcal{P}_2(\mathbb{R}^n)$ [23], which can be defined as ([35], Theorem 13.8): $T_q\mathcal{P}_2(\mathbb{R}^n) = \overline{\{\nabla f | f \in C^\infty(\mathbb{R}^n)\}}^{L_q(\mathbb{R}^n)}$, where $L_q(\mathbb{R}^n)$ is an infinite-dimensional Hilbert space consisting of vector fields $V(x)$ that satisfy $\int_{\mathbb{R}^n} \|V(x)\|_2^2 dq < \infty$, and the overline indicates the closure of the set [23]. The tangent space $T_q\mathcal{P}_2(\mathbb{R}^n)$ inherits an inner product from $L_q(\mathbb{R}^n)$ given by: $\langle V_1(x), V_2(x)\rangle_{L_q(\mathbb{R}^n)} := \int_{\mathbb{R}^n} V_1(x) \cdot V_2(x) dq$, which defines the Riemannian structure on $\mathcal{P}_2(\mathbb{R}^n)$.

For any smooth curve $(p(t,x))_{t\geq 0}$ on $\mathcal{P}_2(\mathbb{R}^n)$, there exists a unique, almost everywhere time-dependent vector field $v(t,x)$ on $\mathbb{R}^n$ that satisfies $\frac{\partial}{\partial t}p(t,x) + \nabla \cdot (v(t,x)p(t,x)) = 0$ for all $t \in \mathbb{R}$, along with $v(t,\cdot) \in T_{p(t,\cdot)}\mathcal{P}_2(\mathbb{R}^n)$ ([26] , Theorem 8.3.1).

***Remark** 3.4:* Assuming that the smooth flow $(p(t,x))_{t\geq 0}$ is the solution of PDE (9), we can derive that the unique time-dependent vector field $v(t,x)$ satisfies the equation below:

$$\frac{\partial p}{\partial t} + \nabla \cdot (v(t,x)p(t,x)) = \mathcal{D}(p) + \nabla \cdot (v(t,x)p(t,x)) = 0, \tag{19}$$

where $v(t,x) \in T_{p(t,x)}\mathcal{P}_2(\mathbb{R}^n)$ and $v(t,x) = \nabla\varphi_1$. We can rewrite equation (19) as (11), giving us a better understanding of why we refer to it as the tangent flow. This concept can also be extended to SPDE cases, as presented in Def. 3.2.

IV. Linear Estimation via particle methods

In this section, we shall introduce several linear estimation frameworks via particle methods, which are OTPP, OTPF, and OTPS.

A. Prediction Via Optimal transportation

The prediction is to estimate the conditional density $p(t,x|\mathcal{Y}_s)$, where $s < t$. The particle method provides a natural solution which is simply push-forward the density according to the dynamical system, i.e.,

$$dx_t = A(t)x_t dt + B(t)dv_t, x_s \sim \mathcal{N}(\mu_s, P_s). \tag{20}$$

As for (20), the means and covariance have explicit solutions which are as follows.

***Lemma** 4.1 ([4]):* The mean and the covariance of (20) are determined by following ODEs,

$$\begin{aligned} \frac{d\mu(t)}{dt} &= A(t)\mu(t) \\ \frac{dP(t)}{dt} &= A(t)P(t) + P(t)A^\top(t) + BB^\top(t). \end{aligned} \tag{21}$$

Then, the posterior density functions of (20) is given as $p(t,x) = c\exp\left(-\frac{1}{2}(x-\mu(t))^\top P^{-1}(t)(x-\mu(t))\right)$, where c is a nomalization factor.

***Remark** 4.1:* There are several research directions related to the density evolution of single linear SDE (20):

- (Monte Carlo Sampling) The goal of sampling is to obtain particles from a density function. One important sampling method in Bayesian inference and a related field is dynamics-based Markov chain Monte Carlo (MCMC), which uses dynamics simulation for state transitions in a Markov chain [36], [37]. Notably, the evolution of (20) can be seen as a type of MCMC. From the computational side, stochastic MCMC will take a longer time to get convergence results [38]. Therefore, an essential question is to design an equivalent deterministic MCMC with the same probability density evolution [39], [23].
- (Stochastic Control with Linear Dynamic Priors) Stochastic control refers to determining the optimal control policy that appeared in the dynamical system to minimize the objective energy function. A typical example is the linear stochastic control shown as follows:

  $$dx_t = A^{(prior)}(t)x_t dt + u(t)dt + B(t)dv_t, \tag{22}$$

  with $A^{(prior)}(t) \in \mathbb{R}^{n\times n}$, $u(t) \in \mathbb{R}^n$, and $B(t) \in \mathbb{R}^{n\times p}$ continuously varying over time t. The optimal control of such minimum energy $\mathbb{E}\int_0^t \|u(t)\|^2 dt$ problem is in a linear feedback form, i.e. $u(t) := A^{(input)}(t)x_t$[40], [41], [42].

Next, we shall introduce the tangent flow of (22), which can be considered as a dynamical flow designed based on OT.

***Theorem** 4.1:* Consider the system (20). We can construct the associated tangent flow as the follow ODE,

$$\begin{aligned} dx_t^{(ot)} &= A(t)x_t^{(ot)}dt + \frac{1}{2}BB^\top(t)P^{-1}(t)(x_t^{(ot)} - \mu(t))dt \\ &+ \Omega(t)P^{-1}(t)(x_t^{(ot)} - \mu(t))dt \end{aligned} \tag{23}$$

where $\Omega(t)$ is uniquely obtained by the following matrix equation:

$$\begin{aligned} \Omega(t)P^{-1}(t) + P^{-1}(t)\Omega(t) &= A(t) - A^\top(t) \\ &+ BB^\top(t)P^{-1}(t) - P^{-1}(t)BB^\top(t), \end{aligned} \tag{24}$$

and $A(t) + \frac{1}{2}BB^\top(t)P^{-1}(t) + \Omega(t)P^{-1}(t)$ is a symmetric matrix.

The proof is given in the appendix.

***Remark** 4.2:* In (23), the stochastic term $B(t)dv_t$ is replaced by the deterministic term $\frac{1}{2}BB^\top(t)P^{-1}(t)(x_t^{(ot)} - \mu(t))dt$. This can be viewed as a deterministic MCMC, offering a better convergence rate than a stochastic one [36]. The ODE (23) includes an additional term with the skew-symmetric matrix $\Omega(t)$. It is a unique choice defined in (24) that acts as a correction term, making the dynamics symmetric and thus optimal in the sense of optimal transportation.

B. Filtering Via Optimal transportation

In this subsection, we shall consider the filtering problem defined in (1). It is well-known that the optimal estimate of the state in (1) is given by the Kalman-Bucy filter (KBF). Let $\mu_-(t) := E[x_t|\mathcal{Y}_t]$ and $P_-(t) := E[(x_t - \mu_-(t))(x_t - \mu_-(t))^\top|\mathcal{Y}_t]$. Then the evolution equations of the conditional expectation $\mu_-(t)$ and the conditional covariance $P_-(t)$ are given in the following lemma.

For simplicity, we define the Riccati operator of filtering (1) in the following definition.

***Definition** 4.1:* The Riccati operator $\mathrm{Ricc}(\cdot)$ of filtering (1) is defined as

$$\begin{aligned}\mathrm{Ricc}(\tilde{P}) &:= (A(t) - CH(t))\tilde{P} \\ &+ \tilde{P}(A(t) - CH(t))^\top + R(t) - \tilde{P}S(t)\tilde{P},\end{aligned} \tag{25}$$

where $R(t) := BB^\top(t) - BD^\top(t)(DD^\top(t))^{-1}DB^\top(t)$, $C(t) := BD(t)^\top(DD^\top(t))^{-1}$, and $S(t) := H^\top(t)(DD^\top(t))^{-1}H(t)$.

***Lemma** 4.2:* [6] The KBF of system (1) is as follows:

$$\begin{aligned}d\mu_-(t) =&A(t)\mu_-(t)dt \\ &+ (K(t) + C(t))(dy_t - H(t)\mu_-(t)dt),\end{aligned} \tag{26}$$

$$\frac{dP_-(t)}{dt} = \mathrm{Ricc}(P_-(t)) \tag{27}$$

where $K(t) := [P_-(t)H^\top(t)](DD^\top(t))^{-1}$ is called Kalman gain, and $\mathrm{Ricc}(\cdot)$ is the Riccati operator of filtering (1).

***Remark** 4.3:* In Lemma 4.2, the stochastic integration is Itô form. However, in this case, the Stratonovich integration is equivalent to the Itô integration, i.e., $(K(t) + C(t))dy_t = (K(t) + C(t)) \circ dy_t$

In the following, similar to prediction part, the tangent flow of KBF can be constructed.

***Theorem** 4.2:* For the system (1), the associated tangent flow is given as follows,

$$\begin{aligned}dx_t^{(ot)} =&A(t)x_t^{(ot)}dt + C(t)(dy_t - H(t)x_t^{(ot)}dt) \\ &+ \frac{1}{2}R(t)P_-^{-1}(t)(x_t^{(ot)} - \mu_-(t))dt \\ &+ K(t)(dy_t - \frac{H(t)x_t^{(ot)} + H(t)\mu_-(t)}{2}dt) \\ &+ \Omega(t)P_-^{-1}(t)(x_t^{(ot)} - \mu_-(t)))dt,\end{aligned} \tag{28}$$

where $K(t) = [P_-(t)H^\top(t)](DD^\top(t))^{-1}$ is called Kalman gain, $R(t) := BB^\top(t) - BD^\top(t)(DD^\top(t))^{-1}DB^\top(t)$, $C(t) = BD(t)^\top(DD^\top(t))^{-1}$, and $\Omega(t)$ is the solution of

$$\begin{aligned}\Omega(t)P_-(t)^{-1}+&P_-(t)^{-1}\Omega(t) = \\ &(A(t) - C(t)H(t))^\top - (A(t) - C(t)H(t)) \\ &+ \frac{1}{2}S(t)P_-^{-1}(t) - \frac{1}{2}P_-^{-1}(t)S(t) \\ &+ \frac{1}{2}(R(t)P_-(t)^{-1} - P_-(t)^{-1}R(t)).\end{aligned}$$

The proof is given in the appendix.

***Remark** 4.4:* Introduced in [24], the OT-modified linear FPF addresses the non-uniqueness problem through an error process and proposes a unique control law via an optimization time-stepping method. In this paper, the derivation is based on tangent flow which combines the Monge-Ampère equation and the PDE expansion technique. The two derivations are ultimately equivalent but different in tools. (28) firstly appeared in [25], where an alternative derivation method is given here. In the linear-Gaussian case, both EnKF and FPF can overcome the curse of dimensionality, whereas PF cannot [22]. Similar to the prediction process, the role of $\Omega(t)$ matrix is to keep symmetric property and optimality.

C. Smoothing Via Optimal transportation

This subsection will focus on the smoothing problem with (1). The smoothing problem is to estimate the $p(t, x|\mathcal{Y}_T)$ by utilizing all observation data. And for linear systems, there is a MFTF formula [10] designed for smoothing problems, which means the estimates generated by two different filters are merged into a combined and more reliable estimate in fixed time interval, i.e.,

Step 1: Do Filtering $p(t, x|\mathcal{Y}_t)$ from $t = 0$ to $t = T$

Step 2: Reverse Smoothing $p(t, x|\mathcal{Y}_T)$ from $t = T$ to $t = 0$

Here, we can see that Step 1 is the Kalman-Bucy filter. Step 2 is the backward Kalman-Bucy filter for system (1) which is defined in the following Lemma 4.3. Firstly, we recall observation history is defined as $\mathcal{Y}_{[0,t]} := \sigma(\{y_s|0 \leq s \leq t\})$ and $\mathcal{Y}_{(t,T]} := \sigma(\{y_s|t < s \leq T\})$.

***Assumption** 1:* The matrices $BB^\top(t)$ and $S(t)$ are in $\mathbb{S}_n^+$, $\mathbb{S}_m^+$, respectively and uniformly bounded, i.e.,

$$\begin{aligned}0 < \inf_{t\geq 0} BB^\top(t) \leq \sup_{t\geq 0} BB^\top(t) < \infty \\ 0 < \inf_{t\geq 0} S(t) \leq \sup_{t\geq 0} S(t) < \infty\end{aligned}$$

The direct result of Assumption 1 is that $P(t)$ defined in (21) is positive definite for any $t \geq 0$. Then, we shall define the dual system of (1).

***Definition** 4.2 (The time-reverse system [10]):* If the linear system x_t satisfies the Assumption 1, the dual system of (1) is defined as

$$\begin{aligned}d\bar{x}_t &= -A^\top(t)\bar{x}_t dt + \bar{B}(t)d\bar{v}_t, \quad \bar{x}_T = 0 \\ dy_t &= \bar{H}(t)\bar{x}_t dt + D(t)d\bar{v}_t\end{aligned} \tag{29}$$

where $\bar{B}(t) := P^{-1}(t)B(t)$, $\bar{H}(t) = H(t)P(t) + D(t)B(t)$, $d\bar{v}_t = dv_t - \bar{B}^\top(t)x_t dt$.

Similarly, with the Definition 4.1, we can define the Riccati operator for dual system (29).

***Definition** 4.3:* The Riccati operator $\mathrm{Ricc}(\cdot)$ of dual filtering system of (1) (defined in (29)) is defined as

$$\begin{aligned}\widetilde{\mathrm{Ricc}}(\tilde{P}) :=& (-A^\top(t) - \bar{C}\bar{H}(t))\tilde{P} \\ &+ \tilde{P}(-A^\top(t) - \bar{C}\bar{H}(t))^\top + \bar{R}(t) - \tilde{P}\bar{S}(t)\tilde{P},\end{aligned} \tag{30}$$

where $\bar{R}(t) := \bar{B}\bar{B}^\top(t) - \bar{B}D^\top(t)(DD^\top(t))^{-1}D\bar{B}^\top(t)$, $\bar{C}(t) = \bar{B}D(t)^\top(DD^\top(t))^{-1}$, and $\bar{S}(t) := \bar{H}^\top(t)(t)(DD^\top(t))^{-1}\bar{H}(t)$.

***Lemma** 4.3:* Consider the filtering problem (1), then the time-reverse continuous system is given in (29). Identically, a cascade of backward Kalman filters generates a process $\bar{\mu}(t)$ with covariance $\bar{P}(t)$ based on the backward stochastic realization (29) and the observation windows $[t,T]$,

$$\begin{aligned}d\bar{\mu}_+(t) =& - A^\top(t)\bar{\mu}_+(t)dt \\ &+ (\bar{K}(t) + \bar{C}(t))(dy_t - \bar{H}(t)\bar{\mu}_+(t)dt),\end{aligned} \tag{31}$$

$$\frac{d\bar{P}_+(t)}{dt} =\widetilde{\mathrm{Ricc}}(\bar{P}_+(t)) \tag{32}$$

where $\bar{K}(t) := [\bar{P}_+(t)\bar{H}^\top(t)](DD^\top(t))^{-1}$ is called Kalman gain, and the terminal condition are $\bar{\mu}_+(T) = 0$ and $\bar{P}_+(T) = P_-(T)$.

Similar to Theorem 4.2, we can construct a backward optimal transportation dynamical flow.

***Theorem** 4.3:* For the dual system (29), the associated tangent flow is given as follows

$$\begin{aligned}d\bar{x}_t^{(ot)} =& - A^\top(t)\bar{x}_t^{(ot)}dt + \bar{C}(t)(dy_t - \bar{H}(t)\bar{x}_t^{(ot)}dt) \\ &+ \frac{1}{2}\bar{R}(t)\bar{P}^{-1}(t)(\bar{x}_t^{(ot)} - \bar{\mu}(t))dt \\ &+ \bar{K}(t)(dy_t - \frac{\bar{H}(t)\bar{x}_t^{(ot)} + \bar{H}(t)\bar{\mu}(t)}{2}dt) \\ &+ \bar{\Omega}(t)\bar{P}^{-1}(t)(\bar{x}_t^{(ot)} - \bar{\mu}(t))dt,\end{aligned} \tag{33}$$

where $\bar{x}_T^{(ot)} = 0$, $\bar{K}(t) := [\bar{P}(t)\bar{H}^\top(t)](DD^\top(t))^{-1}$, $\bar{R}(t) := \bar{B}\bar{B}^\top(t) - \bar{B}D^\top(t)(DD^\top(t))^{-1}D\bar{B}^\top(t)$, $\bar{C}(t) = \bar{B}D(t)^\top(DD^\top(t))^{-1}$, and $\bar{\Omega}(t)$ is the solution of

$$\begin{aligned}\bar{\Omega}(t)P(t)^{-1}+&P(t)^{-1}\bar{\Omega}(t) = \\ &(-A^\top(t) - \bar{C}(t)\bar{H}(t))^\top - (-A^\top(t) - \bar{C}(t)\bar{H}(t)) \\ &+ \frac{1}{2}\bar{S}(t)\bar{P}^{-1}(t) - \frac{1}{2}\bar{P}^{-1}(t)\bar{S}(t) \\ &+ \frac{1}{2}\bar{R}(t)\bar{P}(t)^{-1} - \bar{P}(t)^{-1}\bar{R}(t)).\end{aligned} \tag{34}$$

Then, let $\hat{\mu}(t)$ be the smooth estimation $E[x_t|\mathcal{Y}_T]$ and $\hat{P}(t) := E[(x_t - \hat{\mu}(t))(x_t - \hat{\mu}(t))^\top|\mathcal{Y}_T]$ which will be calculated by the following result.

***Theorem** 4.4 ([12]):* Consider the stochastic system (1) with observation and its associated dual system (29). Then, for $t \in [0,T]$, the Mayne-Fraser smoothing formula is given as

$$\hat{\mu}(t) = \hat{P}(t)[\bar{P}_+^{-1}\bar{\mu}_+(t) + P_-^{-1}(t)\mu_-(t)] \tag{35}$$

where $\hat{P}$ satisfies the following equation

$$\begin{aligned}\hat{P}^{-1}(t) = \frac{1}{2}\Big(&\bar{P}_+^{-1}P_-^{-1}(t) + P_-^{-1}\bar{P}_+^{-1}(t) - \bar{P}_+^{-1}P^{-2}P_-^{-1}(t) \\ &+ P_-^{-1}P^{-2}\bar{P}_+^{-1}(t) + (P_-^{-1} + \bar{P}_+^{-1})P(t) \\ &+ P(t)(P_-^{-1} + \bar{P}_+^{-1}) - 2I\Big).\end{aligned} \tag{36}$$

***Remark** 4.5:* Theorem 4.4 holds for any intermittent observations structure. Once the observations are missed in some intervals, we can simplily assume the $K(t)$, $C(t)$, $\bar{K}(t)$, and $\bar{C}(t)$ are all zero in the Theorem 4.2 and 4.3. This fact was first pointed out by [12].

D. The algorithms and the Finite N formulation

In this subsection, we shall introduce a numerical implementation. We shall simulate N independent stochastic processes (particles) $\left\{x_t^{(ot,i)}, 1 \leq i \leq N\right\}$ according to (23). However, the $\mu(t)$ and $P(t)$ in (23) shall be approximated by the particles $\left\{x_t^{(ot,i)}, 1 \leq i \leq N\right\}$, which satisfy

$$\begin{aligned}\mu_*(t) &\approx \mu^{(N)}(t) = \frac{1}{N}\sum_{i=1}^N x_t^{(ot,i)}, \\ P_*(t) &\approx P^{(N)}(t) \\ &= \frac{1}{N-1}\sum_{i=1}^N (x_t^{(ot,i)} - \mu^{(N)}(t))^\top(x_t^{(ot,i)} - \mu^{(N)}(t)),\end{aligned} \tag{37}$$

1) Finite N formulation for OTPP: As for OTPP, we combine (37) and (23). Then, we shall have

$$\begin{aligned}dx_t^{(ot,i)} =& A(t)x_t^{(ot,i)}dt \\ &+ \frac{1}{2}BB^\top(t)(P^{(N)})^{-1}(t)(x_t^{(ot,i)} - \mu^{(N)}(t))dt \\ &+ \Omega^{(N)}(t)(P^{(N)})^{-1}(t)(x_t^{(ot,i)} - \mu^{(N)}(t))dt,\end{aligned} \tag{38}$$

where the $\Omega^{(N)}$ is the solution of

$$\begin{aligned}&\Omega^{(N)}(t)(P^{(N)})^{-1}(t) + (P^{(N)})^{-1}(t)\Omega^{(N)}(t) = A(t) - A^\top(t) \\ &+ BB^\top(t)(P^{(N)})^{-1}(t) - (P^{(N)})^{-1}(t)BB^\top(t)\end{aligned}$$

We call (38) as the finite N system for (23). And the OTPP algorithm is summarised in Alg. 1.

Algorithm 1 Finite N formulation for OTPP

```
Initialization
for $i := 1$ to $N$ do
    Sample $x_i^0$ from $p(0,x)$
end for
Assign value $t := 0$
Iteration [from $t$ to $t + \Delta t$]
Calculate $\mu^{(N)}(t)$ and $P^{(N)}(t)$ by using (37).
for $i := 1$ to $N$ do
    Update $x_{t+\Delta t}^{(ot,i)}$ by using forward Euler scheme of (38).
end for
Update $t := t + \Delta t$.
```

2) Finite N formulation for OTPF: We combine (37) and (28), so that we have,

$$\begin{aligned}dx_t^{(ot,i)} =&A(t)x_t^{(ot,i)}dt + C(t)(dy_t - H(t)x_t^{(ot,i)}dt) \\ &+ \frac{1}{2}R(t)(P_-^{(N)})^{-1}(t)(x_t^{(ot,i)} - \mu_-^{(N)}(t))dt \\ &+ K^{(N)}(t)(dy_t - \frac{H(t)x_t^{(ot,i)} + H(t)\mu_-^{(N)}(t)}{2}dt) \\ &+ \Omega^{(N)}(t)P^{-1}(t)(x_t^{(ot,i)} - \mu_-^{(N)}(t))dt,\end{aligned} \tag{39}$$

where $K^{(N)}(t) = [P_-^{(N)}(t)H^\top(t)](DD^\top(t))^{-1}$ (Kalman Gain), $R(t) := BB^\top(t) - BD^\top(t)(DD^\top(t))^{-1}DB^\top(t)$, $C(t) = BD(t)^\top(DD^\top(t))^{-1}$, and $\Omega^{(N)}(t)$ is the solution of

$$\begin{aligned}\Omega^{(N)}(t)P_-^{(N)}(t)^{-1} + P_-^{(N)}(t)^{-1}\Omega^{(N)}(t) = \\ (A(t) - C(t)H(t))^\top - (A(t) - C(t)H(t)) \\ + \frac{1}{2}S(t)(P_-^{(N)})^{-1}(t) - \frac{1}{2}(P_-^{(N)})^{-1}(t)S(t) \\ + \frac{1}{2}R(t)(P_-^{(N)})^{-1}(t) - (P_-^{(N)})^{-1}(t)R(t)).\end{aligned}$$

We call (39) as the finite N system for (28). And the OTPF algorithm is summarised in Alg. 2.

Algorithm 2 Finite N formulation for OTPF

```
Initialization
for $i := 1$ to $N$ do
    Sample $x_0^i$ from $p(0,x)$
end for
Assign value $t := 0$
Iteration [from $t$ to $t + \Delta t$]
Calculate $\mu^{(N)}(t)$ and $P^{(N)}(t)$ by using (37).
for $i := 1$ to $N$ do
    Update $x_{t+\Delta t}^{(ot,i)}$ by using forward Euler scheme of (39).
end for
Update $t := t + \Delta t$.
```

3) Finite N formulation for OTPS: Similar to the filtering problem, we combine (37) and (33), so that we have,

$$\begin{aligned}d\bar{x}_t^{(ot,i)} =& - A^\top(t)\bar{x}_t^{(ot,i)}dt \\ &+ \frac{1}{2}\bar{B}\bar{B}^\top(t)(\bar{P}^{(N)}(t))^{-1}(t)(\bar{x}_t^{(ot,i)} - \bar{\mu}_+^{(N)}(t))dt \\ &+ \Omega^{(N)}(t)(\bar{P}^{(N)}(t))^{-1}(t)(\bar{x}_t^{(ot,i)} - \bar{\mu}_+^{(N)}(t))dt \\ &+ \bar{K}^{(N)}(t)(dy_t - \frac{H(t)\bar{x}_t^{(ot)} + H(t)\bar{\mu}_+^{(N)}(t)}{2}dt) \\ &+ \Omega^{(N)}(t)P^{-1}(t)(\bar{x}_t - \bar{\mu}_+^{(N)}(t)))dt \qquad (40)\end{aligned}$$

where $\bar{K}^{(N)}(t) := [\bar{P}_+^{(N)}(t)\bar{H}^\top(t)](DD^\top(t))^{-1}$, $\bar{R}(t) := \bar{B}\bar{B}^\top(t) - \bar{B}D^\top(t)(DD^\top(t))^{-1}D\bar{B}^\top(t)$, $\bar{C}(t) = \bar{B}D(t)^\top(DD^\top(t))^{-1}$, and $\bar{\Omega}^N(t)$ is the solution of

$$\begin{aligned}\bar{\Omega}^N(t)P(t)^{-1} + P(t)^{-1}\bar{\Omega}^N(t) = \\ (-A^\top(t) - \bar{C}(t)\bar{H}(t))^\top - (-A^\top(t) - \bar{C}(t)\bar{H}(t)) \\ + \frac{1}{2}\bar{S}(t)(\bar{P}_+^N)^{-1}(t) - \frac{1}{2}(\bar{P}_+^N)^{-1}(t)S(t) \\ + \frac{1}{2}\bar{R}(t)(\bar{P}_+^N)^{-1}(t) - (\bar{P}_+^N)^{-1}(t)\bar{R}(t)).\end{aligned}$$

And the Algorithm is summarized in Alg. 3.

V. THE CONVERGENCE ANALYSES OF PROPOSED ALGORITHMS

In this section, we shall provide the convergence analysis of proposed algorithms, including OTPP, OTPF, and OTPS.

Algorithm 3 Finite N formulation for OTPS

```
Initialization
for $i := 1$ to $N$ do
    Sample $x_0^i$ from $p(0,x)$, and sample $\bar{x}_0^i$ from $\bar{p}(0,x)$.
end for
Assign value $t := 0$
Iteration [from $t$ to $t + \Delta t$]
Calculate $\mu^{(N)}(t)$ and $P^{(N)}(t)$ by using (37).
for $i := 1$ to $N$ do
    Update $x_{t+\Delta t}^{(ot,i)}$ by using forward Euler scheme of (39).
    Update $\bar{x}_{t+\Delta t}^{(ot,i)}$ by using forward Euler scheme of (40).
end for
Update $\hat{\mu}_{t+\Delta t}$ by using $\{x_{t+\Delta t}^{(ot,i)}, \bar{x}_{t+\Delta t}^{(ot,i)} | 1 \le i \le N\}$, and (35).
Update $t := t + \Delta t$.
```

A. Main assumptions and stability for KBF.

In this subsection, we shall introduce the main assumptions for the filtering problem. By considering the constant matrix M, we can define the transition matrix as $\mathcal{E}_{s,t} = e^{M(t-s)}$, and for general time-vary flow, we define in the following.

Definition 5.1 ([43]): The transition matrix associated with a smooth matrix value flow $M : u \to M(u) \in \mathbb{R}^{n \times n}$ with $u \in [0, \infty)$ is defined as the solution of the following matrix value differential equation,

$$\frac{\partial}{\partial t}\mathcal{E}_{s,t}(M) = M(t)\mathcal{E}_{s,t}(M),\ \frac{\partial}{\partial s}\mathcal{E}_{s,t}(M) = -\mathcal{E}_{s,t}(M)M(s),$$

for any $s \le t$, with $\mathcal{E}_{s,s} = I_n$, the identity matrix.

Then, we shall introduce the observability Gramian, $\mathcal{O}_{s,t}$, and controllability Gramian, $\mathcal{C}_{s,t}$ of system (1) which are defined by

$$\mathcal{O}_{s,t} := \int_s^t \mathcal{E}_{r,t}(CH - A)S(t)\mathcal{E}_{r,t}(CH - A)^\top dr$$

where $S(t)$ is defined in Def. 4.1 and

$$\mathcal{C}_{s,t} := \int_s^t \mathcal{E}_{r,t}(A - CH)BB^\top(r)\mathcal{E}_{r,t}(A - CH)^\top dr.$$

In control theory, the system (1) is observable/controllable if the observability/controllability Gramian is positive-defined. We recommend readers refer to [43] for details.

Assumption 2: For the observability and controllability Gramians, there exists parameters $v, \omega_\pm^o > 0, \omega_\pm^c > 0$ such that

$$\begin{aligned}\omega_-^c I_n \le \mathcal{C}_{t,t+v} \le \omega_+^c I_n \\ \omega_-^o I_n \le \mathcal{O}_{t,t+v} \le \omega_+^o I_n\end{aligned}$$

uniformly for all $t \ge 0$. Here, parameter v is called the interval of observability/controllability.

Assumption 3: The $A(t)$ in system (20) and the $A(t) - CH(t)$ in (4.2) are uniformly bounded, i.e., for some λ

$$\sup_{t \ge 0} \max\{|\lambda_{max}(A(t))|, |\lambda_{min}(A(t))|\} \le \lambda$$

$$\sup_{t \ge 0} \max\{|\lambda_{max}(A(t) - CH(t))|, |\lambda_{min}(A(t) - CH(t))|\} \le \lambda$$

The following Lemmas are important in the convergence of the OTPF and OTPS.

***Lemma** 5.1:* Consider the following Riccati system 4.1

$$\frac{d\tilde{P}(t)}{dt} = \operatorname{Ricc}(\tilde{P}(t)). \tag{41}$$

Let $\Phi_t(\cdot)$ be the solution operator of the Riccati system, i.e. $\Phi_t(\tilde{P}(0)) = \tilde{P}(t)$. For any $s \leq t$ and $Q \in \mathbb{S}_n^+,$ we can define a smooth flow set $u \to A(u) - CH(u) - \Phi_u(Q)S(u)$. By using the Def.5.1, the transition matrix associated with a smooth flow $A(u) - CH(u) - \Phi_u(Q)S(u)$ is given as $\mathcal{E}_{s,t}(A(t) - CH(t) - \Phi_t(Q)S(t))$. For simplicity, we define a new operator $E_{s,t}(Q)$, i. e.,

$$E_{s,t}(Q) := \mathcal{E}_{s,t}(A(t) - CH(t) - \Phi_t(Q)S(t)).$$

And the Assumption 2 is satisfied for some $v > 0$. Then

1) (Theorem 1.1 in [44]) For any $t \geq v$ and $\tilde{P}_1(0) \in \mathbb{S}_n^+$, there exist two matrices $\Lambda_{\min}, \Lambda_{\max} \in \mathbb{S}_n^+$ such that
$$\Lambda_{\min} \leq \Phi_t(\tilde{P}_1(0)) \leq \Lambda_{\max}. \tag{42}$$
2) (Theorem 1.2 in [44]) For any $\tilde{P}_1(0), \tilde{P}_1(0) \in \mathbb{S}_n^+$ and $t \geq 0$, there is
$$\begin{aligned} &\|\Phi_t(\tilde{P}_1(0)) - \Phi_t(\tilde{P}_2(0))\|_2 \\ &\qquad \leq \alpha_1 \exp(-\beta t)\|\tilde{P}_1(0) - \tilde{P}_1(0)\|_2, \end{aligned} \tag{43}$$
where the α_1 depends on $\tilde{P}_1(0), \tilde{P}_1(0)$ and
$$\begin{aligned} \beta := \frac{1}{2(\omega_+^0 + 1/\omega_-^c)} \Bigg[&\inf_{t \geq 0]} \lambda_{min}(S(t)) \\ &+ \frac{\inf_{t \geq 0} \lambda_{min}(BB^\top(t))}{(\omega_+^c + 1/\omega_-^o)^2} \Bigg]. \end{aligned}$$
3) (Theorem 4.2 in [44]) For any $t \geq s \geq v$ and $Q \in \mathbb{S}_n^+$, we have
$$\|E_{s,t}(Q)\|_2 \leq \alpha_2 \exp(-\beta(t-s)),$$
where $\alpha_2^2 := \frac{\omega_+^o(\mathcal{C}) + 1/\omega_-^c}{\omega_+^c(\mathcal{C}) + 1/\omega_-^o}$ and β is same as those in 2) above.

Here, we shall introduce a lemma on the general convergence of Monte-Carlo methods.

***Lemma** 5.2:* [45] Let the n-dimensional random vectors x_i, which are independently sampled from $\mathcal{N}(\mu, P)$ with $i = 1, \cdots, N$. Define

$$\begin{aligned} \mu^{(N)} &:= \frac{1}{N}\sum_{i=1}^{N} x_i \\ P^{(N)} &:= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu^{(N)})(x_i - \mu^{(N)})^\top. \end{aligned} \tag{44}$$

Then, we have

$$\mathbb{E}[\|\mu - \mu^{(N)}\|^2] \leq \frac{c(n)}{N}, \mathbb{E}[\|P - P^{(N)}\|^2] \leq \frac{c(n)}{N}, \tag{45}$$

where c_n is some constant depending on n.

B. The main convergence Theorem

In this subsection, the main convergence Theorems are given.

***Theorem** 5.1 (Convergence for OTPP):* Consider the dynamical system (20) with Gaussian initial density. Let $(\mu^{(N)}(t), P^{(N)}(t))$ be the empirical mean and covariance obtained from the OTPP system (38) while $(\mu(t), P(t))$ is mean and covariance obtained from (21). Under the Assumptions 1, 2, and 3, then, for any $t \geq 0$, there exist constants c_1, c_2 such that

$$\begin{aligned} \|\mu(t) - \mu^{(N)}(t)\|_2^2 &\leq \frac{c_1}{N}, \\ \|P(t) - P^{(N)}(t)\|_F^2 &\leq \frac{c_2}{N}. \end{aligned} \tag{46}$$

***Lemma** 5.3 ([25]):* The evolution of empirical mean and covariance $(\mu^{(N)}, P^{(N)})$ satisfy

$$\begin{aligned} d\mu^{(N)}(t) &= A(t)\mu^{(N)}(t)dt \\ &\quad + (K(t) + C(t))(dy_t - H(t)\mu^{(N)}(t)dt) \\ \frac{dP^{(N)}(t)}{dt} &= \operatorname{Ricc}(P^{(N)}(t)), \end{aligned} \tag{47}$$

where $K^{(N)}(t) := [P^{(N)}(t)H^\top(t)](DD^\top(t))^{-1}$ as Kalman gain, $R(t) := BB^\top(t) - BD^\top(t)(DD^\top(t))^{-1}DB^\top(t)$ and $C(t) = BD(t)^\top(DD^\top(t))^{-1}$.

***Theorem** 5.2 (Convergence for OTPF):* Consider the dynamical system (1) with initial Gaussian. Let $(\mu_-^{(N)}(t), P_-^{(N)}(t))$ be the empirical mean and covariance obtained from the OTPF system (39) while $(\mu_-(t), P_-(t))$ is the mean and covariance obtained from KBF. Under the Assumptions 1, 2, and 3, for any $t \geq 0$, there exist constants c_1, c_2 such that

$$\begin{aligned} \|\mu_-(t) - \mu_-^{(N)}(t)\|_2^2 &\leq c_1(1 + \frac{2n^2\alpha^2\lambda}{\beta})e^{-2\beta t}\frac{1}{N} \\ \|P_-(t) - P_-^{(N)}(t)\|_F^2 &\leq c_2(1 + \frac{2n^2\alpha^2\lambda}{\beta})e^{-2\beta t}\frac{1}{N}. \end{aligned} \tag{48}$$

To analyze the convergence OTPS, we shall combine the forward and backward filters. Since the constant of convergence analysis depends on the covariance of the system, we can simply assume that both the state system and dual system covariances I_n for any time t. This method is widely used in previous works, known as the normalized trick.

***Theorem** 5.3 (Convergence for OTPS):* Consider the dynamical system (1) with initial Gaussian. Let $(\mu_-^{(N)}(t), P_-^{(N)}(t))$ and $(\bar{\mu}_+^{(N)}(t), \bar{P}_+^{(N)}(t))$ the empirical means and covariances obtained from the OTPF system (39) and revise OTPF (40), respectively. Then, the empirical mean and covariance $(\hat{\mu}^{(N)}(t), \hat{P}^{(N)}(t))$ of the OTPS system for (1) can be computed using the MFTF formula (35) and (36) while $(\hat{\mu}(t), \hat{P}(t))$ is the mean and covariance obtained from RTS smoother. Under the Assumptions 1, 2, and 3, for any $t \geq 0$, there exist constants c_1, c_2 such that

$$\begin{aligned} \|\hat{\mu}(t) - \hat{\mu}^{(N)}(t)\|_2^2 &\leq c_\mu \frac{1}{N}, \\ \|\hat{P}(t) - \hat{P}^{(N)}(t)\|_F^2 &\leq c_{\hat{P}} \frac{1}{N}, . \end{aligned} \tag{49}$$

The proof is given in the appendix.

VI. Numerical Simulation

This section will present an extensive numerical study of the estimation problems via OT and compare the filter results with some traditional algorithms such as KF, PF, EnKF, RTS, PS, and EnKS. To compare the performance of different methods, we introduce the mean squared error (MSE) and the mean of the mean squared error (MMSE) based on 100 realizations, which are defined as follows:

$$\mathrm{MSE}(k\cdot\Delta t) := \frac{1}{100}\sum_{i=1}^{100}\|x_{k\cdot\Delta t}^{[i]} - \hat{x}_{k\cdot\Delta t}^{[i]}\|_2,$$

$$\mathrm{MMSE} := \frac{1}{100}\frac{1}{Sp}\sum_{k=1}^{Sp}\sum_{i=1}^{100}\|x_{k\cdot\Delta t}^{[i]} - \hat{x}_{k\cdot\Delta t}^{[i]}\|_2,$$

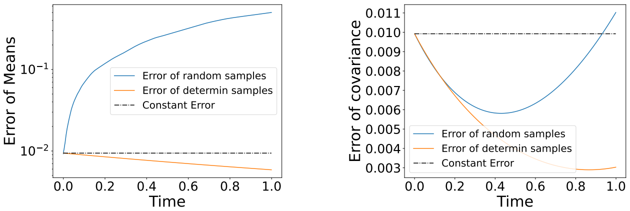
where $x_{k\cdot\Delta t}^{[i]}$ is the real state at discrete time instant $k\cdot\Delta t$ in $i-$th experiment, $\hat{x}_{k\cdot\Delta t}^{[i]}$ is the estimation of $x_{k\cdot\Delta t}^{[i]}$, and Sp is the total steps. The mean time (MT) is introduced for numerical complexity.

A. The prediction problem

In this subsection, we present a high-dimensional sampling problem. Here, we consider the high dimensional sampling problem which can be modeled as follows,

$$dx_t = Ax_tdt + dv_t,$$

where the $x_t \in \mathbb{R}^{100}$, v_t is 100-dimensional standard Brownian motion, and $A = (A_{i,j})_{i,j=1}^{100}$ are defined as $A_{i,j} = 0.1$ if $i-j=1$, $A_{i,j} = -0.5$ if $i-j=0$. The posterior distribution at $t=1$ is the target distribution.



(a) Error of the expectation

(b) Error of the covariance

Fig. 1. Sampling results. The error of the expectation is shown in (a) and the error of the covariance is shown in (b). Constant error is a horizontal reference line.

Based on Fig. 1, it can be found that deterministic sampling results in smaller errors. From the standpoint of expectation, the error of deterministic sampling decreases over time, while the error of random sampling increases. From the perspective of covariance, the error of deterministic sampling is also gradually decreasing, while the results of random sampling, although improving at first, ultimately become worse.

B. The filtering problem with missing observation

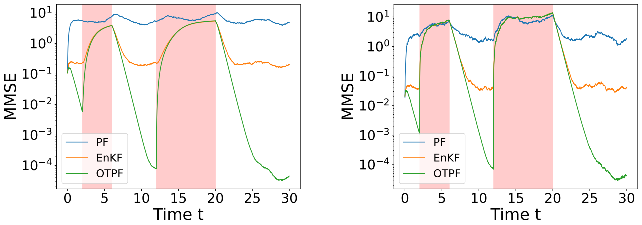
In this subsection, we present the following example for both filtering experiments and compare the proposed algorithm with EnKF, PF, and KF.

$$\begin{aligned} dx_t &= A_n(t)x_tdt + (0.4I_n, 1.6I_n)dv_t,\\ dy_t &= I_nx_tdt + (0_{n\times n}, I_n)dv_t, \end{aligned} \tag{50}$$

where $x_t, y_t \in \mathbb{R}^n$, v_t is the $2n$-dimensional standard Brownian motion and the $A(t) := (A_{i,j}(t))_{i,j=1}^n$ is define as $(A_n(t))_{i,j} = 0.1\cos(t)$ if $i-j=1$, $(A_n(t))_{i,j} = -0.5(1+0.1\cos(2t))$ if $i-j=0$, $(A_n(t))_{i,j} = 0.15$ if $i-j=-1$, and $(A_n(t))_{i,j} = 0$ for otherwise. Numerical simulation over time interval $[0,30]$ produces a time-function dy_t, which is sampled with integer multiples of $\Delta t = 0.01$ (units). The interval $[0,30]$ is partitioned into 5 pieces, i.e., $[0,30] = \bigcup_{i=1}^{5}[t_i, t_{i+1}]$, where $t_{i+1} - t_i = 2*i$. And when $t \in [t_i, t_{i+1}]$ with $i = 2, 4$, the sensors are blocked, so there is no observation data dy_t for this filter system.

TABLE I
Filter experiment results in 10-dimensional cases with different simulated particle number N

Method	KF(M)	PF	EnKF	OTPF
MMSE	3.3179	15.1744	5.2758	3.6475
MT	0.2932	0.9960	0.5144	0.5911
N	–	10	10	10
Method	KF(M)	PF	EnKF	OTPF
MMSE	3.3179	13.5943	4.0294	3.32957
MT	0.2932	1.3214	0.93029	1.0886
N	–	20	20	20
Method	KF(M)	PF	EnKF	OTPF
MMSE	3.3179	10.3675	3.6126	3.3265
MT	0.2932	2.2632	1.9247	2.096501
N	–	50	50	50
Method	KF(M)	PF	EnKF	OTPF
MMSE	3.3179	8.0009	3.49359	3.3209
MT	0.2932	4.0108	3.7084	3.7475
N	–	100	100	100
Method	KF(M)	PF	EnKF	OTPF
MMSE	3.3179	4.3683	3.3280	3.3187
MT	0.2932	16.7924	15.02346	16.3594
N	–	500	500	500



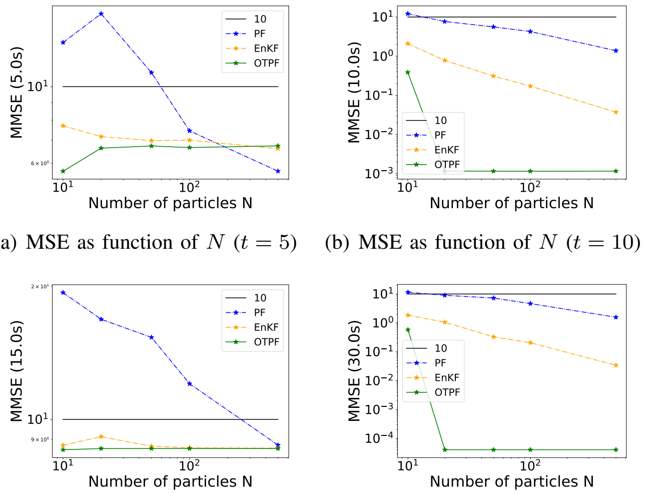
(a) 100 Particles MSE of time

(b) 500 Particles MSE of time

Fig. 2. The MSE of three different filter algorithms are shown. There is no observation in time periods marked in red. The dimension of the state and observation is $n = m = 10$.

The simulation results of filtering problems are discussed next:

1) **MMSE Comparison Between Optimal KF and Particle Algorithms with Missing Observations** The 10-dimensional numerical results are shown in Figure 2, with 100 particles in (a) and 500 particles in (b). OTPF provides a more accurate approximation than other algorithms, demonstrating the fastest convergence rate and best convergence order during observation intervals. Even with 100 particles, OTPF is highly efficient, whereas PF with 100 particles is nearly ineffective, and



(a) MSE as function of N ($t = 5$) (b) MSE as function of N ($t = 10$)

(c) MSE as function of N ($t = 15$) (d) MSE as function of N ($t = 30$)

Fig. 3. MSE at different times as the function of N. The dimension of the state and observation is $n = m = 10$.

EnKF with 500 particles still significantly lags behind OTPF with 100 particles.

2) **MSE vs. Particle numbers.** We conduct a simulation on the 10-dimensional system in (50) with varying particle numbers $N \in \{10, 20, 50, 100, 500\}$. We report the average results of 20 runs in Table I. The KF(M) is the Kalman filter with missing observations, where the Kalman gain is set to zero during the missing observation period. Our OTPF algorithm outperforms PF with 500 particles and is comparable to EnKF with 500 particles, even with only 20 particles. Additionally, as the number of particles increases, the computation time of OTPF slightly increases but remains much lower than PF while being slightly larger than EnKF.

3) **MSE vs. time** We conduct a simulation to study the performance of the EnKF and OTPF algorithms with different particle numbers N and at different time intervals t. Specifically, we plotted the MMSE as a function of N at times $t \in \{5, 10, 15, 30\}$s for a 10-dimensional system, as shown in Figure 3. In the absence of observation periods, both EnKF and OTPF perform similarly. However, during observation periods, OTPF has a faster convergence speed and achieves better results than those of EnKF and PF, especially with a small number of particles.

C. The smoothing problem

In this subsection, we shall consider the following example for smoothing experiments and compare the proposed algorithm with EnKS, PS, and RTS.

$$\begin{cases} dx_t = \tilde{A}_n(t)x_t dt + (I_n, 0_{n\times n})dv_t, \\ dy_t = x_t dt + (0_{n\times n}, I_n)dv_t, \end{cases} \tag{51}$$

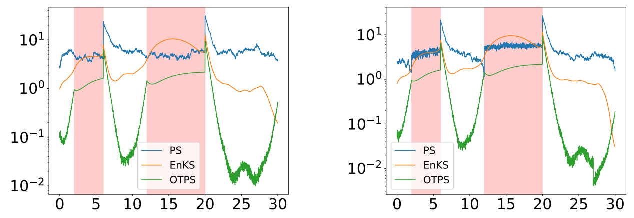
where $x_t, y_t \in \mathbb{R}^n$, $x_0 \sim \mathcal{N}(0, I_n)$, v_t is the $2n$-dimensional standard Brownian motion, and $\tilde{A}_n(t) := ((A_n)_{i,j}(t))_{i,j=1}^n$ is defined as $(\tilde{A}_n(t))_{i,j} = 0.1\cos(t)$ if $i - j = 1$, $(\tilde{A}_n(t))_{i,j} = -0.5$ if $i = j$ $(\tilde{A}_n(t))_{i,j} = -0.1\cos(t)$ if $i - j = -1$ and $(\tilde{A}_n(t))_{i,j}$ for otherwise. Then, the dual system of (51) is as follows,

$$\begin{cases} d\bar{x}_t = -A_n^\top(t)\bar{x}_t dt + \big(I_n, 0_{n\times n}\big)d\bar{v}_t, \\ d\bar{y}_t = I_n\bar{x}_t dt + \big(0_{n\times n}, I_n\big)d\bar{v}_t. \end{cases} \tag{52}$$

where $\bar{x}(T) = 0$.

TABLE II
SMOOTHING EXPERIMENT RESULTS IN 10-DIMENSIONAL CASES WITH DIFFERENT SIMULATED PARTICLE NUMBER N

Method	RTS(M)	PS	EnKS	OTPS
MMSE	1.5307	11.60939	8.2758	2.04826
MT	0.4091	4.09136	1.39132	1.92688
N	–	10	10	10
Method	RTS(M)	PS	EnKS	OTPS
MMSE	1.5307	10.61721	5.28977	1.54802
MT	0.4091	4.9669	1.96412	2.65917
N	–	20	20	20
Method	RTS(M)	PS	EnKS	OTPS
MMSE	1.5307	7.8642	4.245651	1.53312
MT	0.4091	7.850553	3.58582	4.8880045
N	–	50	50	50
Method	RTS(M)	PS	EnKS	OTPS
MMSE	1.5307	6.00583	3.78255	1.53228
MT	0.4091	12.7001	6.2978	8.5595
N	–	100	100	100
Method	RTS(M)	PS	EnKS	OTPS
MMSE	1.5307	3.12928	3.004476	1.53254
MT	0.4091	51.31694	27.7580	38.1201
N	–	500	500	500

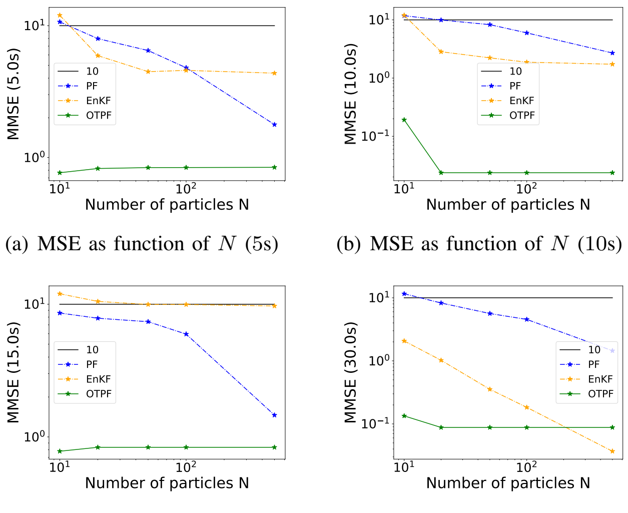


(a) 100 Particles MSE (b) 500 Particles MSE

Fig. 4. The MSE of three different smoother algorithms are shown. There is no observation in time periods marked in red. The dimension of the state and observation is $n = m = 10$.

The simulation results of smoothing problems are discussed next:

1) **MSE Comparison Between Optimal RTS and Particle Algorithms with Missing Observations** The 10-dimensional numerical results are shown in Figure 4, with 100 particles in (a) and 500 particles in (b). OTPS provides a more accurate approximation than other algorithms, showing the fastest convergence rate and best convergence order during observation intervals. Even with 100 particles, OTPS is highly efficient, whereas PS with 100 particles is nearly ineffective, and EnKS with 500 particles still significantly lags behind OTPS with 100 particles. The accuracy of OTPS is slightly less than OTPS, primarily because OTPS relies



(a) MSE as function of N (5s)

(b) MSE as function of N (10s)

(c) MSE as function of N (15s)

(d) MSE as function of N (30s)

Fig. 5. MSE at different times as the function of N. The dimension of the state and observation is $n = m = 10$.

on solving a covariance matrix equation reconstructed by particles, and the convergence of the covariance matrix is weaker than the order of convergence of expectations.

2) **MSE vs. Particle Numbers** In this simulation, we test the 10-dimensional case system of (50) with different particle numbers $N \in \{10, 20, 50, 100, 500\}$. The average results of 20 times are shown in Table I. The KS(M) is the KS with missing observation, where the Kalman gain is set to zero during the missing observation period. The OTPS with 20 particles performs better than PS with 500 particles and it almost performs the same as EnKS with 500 particles. The computation time of OTPS growth is much slower than PS and slightly larger than EnKS when the number of particles increases.

3) **MSE vs. time** We perform a simulation to compare the performance of EnKS and OTPS algorithms with varying particle numbers N and at different time intervals t in a 10-dimensional system. We plotted the MSE as a function of N at times $t \in \{5, 10, 15, 30\}$s in Figure 5. The EnKS had a larger error at low particle numbers in the absence of observations, while PS with resampling achieved results close to OTPS as the number of particles increased. OTPS's error did not change much with increasing particle numbers. However, during observation periods, OTPS had a faster convergence speed and achieved better results than those of EnKS and PS, particularly with a small number of particles.

VII. Conclusion

In this paper, we extend the OTPF to prediction and smoothing problems inspired by the work of Georgiou and Lindquist. Numerical experiments for prediction, filtering, and smoothing problems are conducted to demonstrate the effectiveness of our proposed approach. The proposed methods offer two advantages. 1) The geometric meaning of OT constructs a unique transport mapping that minimally alters the particles, resulting in a more robust and stable algorithm. 2) Additionally, minimal filtration eliminates the need for additional sampling, which improves the algorithm's accuracy and reduces reliance on extra noise.

Appendix
Proofs in the Section III.A

Proofs for Theorem 3.1 Considering the optimal transport problem between two distribution $p(t, x)$ and $p(t + \Delta t, x)$, we will get following Monge-Ampère equation by letting the $f(x) = p(t + \Delta t, x)$ and $g(x) = p(t, x)$:

$$\det D^2\Phi_t = \frac{p(t + \Delta t, x)}{p(t, \nabla\Phi_t(x))}. \tag{53}$$

The Monge-Ampère equation is a nonlinear PDE that is hard to solve, but luckily the $p(t, x)$ and $p(t+\Delta t, x)$ are close when Δt is small. So $D^2\Phi_t$ should be close to I_n, we can expand both sides of equation (53) in the asymptotic sense of $\Delta t \to 0$. Firstly, we can expand $p(t + \Delta t, x)$ according (9),

$$p(t + \Delta t, x) = p(t, x) + \mathcal{D}(p(t, x))\Delta t + O(\Delta t^2). \tag{54}$$

Divide $p(t, x)$ for the both sides, which yields,

$$\frac{p(t + \Delta t, x)}{p(t, x)} = 1 + \frac{\mathcal{D}(p(t, x))}{p(t, x)}\Delta t + O(\Delta t^2). \tag{55}$$

And we consider to expand the $\Phi_t(x)$. Since $D^2\Phi_t$ should be close to I_n, then $\Phi_t(x)$ should be close to $\frac{|x|^2}{2}$. So, we have

$$\Phi_t(x) = \frac{|x|^2}{2} + \varphi_1(t, x) \cdot \Delta t + O(\Delta t^2), \tag{56}$$

where φ_1 is an undetermined function.

So, submit (56) and (55) into the equation (53) and take $\Delta t \to 0$, we will get equation satisfied by φ_1

$$\Delta\varphi_1(t, x) + \nabla(\log p(t, x))\nabla\varphi_1(t, x) = -\frac{\mathcal{D}(p(t, x))}{p(t, x)}, \tag{57}$$

and

$$\nabla\Phi_t(x) = x + \varphi_1(t, x)\Delta t + O(\Delta t^2). \tag{58}$$

Thus, we get optimal transport $\nabla\Phi_t(x)$ in Monge's optimal transportation problem in Theorem 2.1. Then we can design probability flow $x_{t+\Delta t} = \nabla\Phi_t(x_t)$ and by using (58) which leads to

$$x_{t+\Delta t} = x_t + \nabla\varphi_1(t, x_t)\Delta t + O(\Delta t^2) \tag{59}$$

In the asymptotic sense for (59), we get **tangent flow:**

$$dx_t = \nabla\varphi_1(t, x)dt, \tag{60}$$

where $x_0 \sim p(0, x)$. □

Appendix
Proofs in the Section III.B

Proofs for Theorem 3.2 Similarly, we consider a discretization of (14). Notice that if we want to apply the Monge-Ampère equation, $p(t+\Delta t, x)$ and $p(t,x)$ should be determined explicitly, which means that the path of dI_t is given in the analysis. In the following, we take the realization of a stochastic process $I_t = I_t(\omega)$ by fixing ω. For a time sequence $\{0 = t_0 \leq t_1 \leq \cdots \leq t_{2^n} = S\}$ with $n \in \mathbb{Z}^+$ and $t_i = iS2^{-n}$, we approximate the $\frac{dI_t}{dt} \approx \dot{I_t}^{(n)} = 2^n(I_{t_{i+1}} - I_{t_i})$ for $t \in [t_i, t_{i+1})$. So, we have

$$\frac{\partial p^{(n)}}{\partial t} = \mathcal{D}(p^{(n)}) + \mathcal{H}(p^{(n)}) \cdot \dot{I_t}^{(n)}, \tag{61}$$

where the $\dot{I_t}^{(n)} = 2^n(I_{t_{i+1}} - I_{t_i})$ if $t \in [t_i, t_{i+1})$. As before we need to solve equation (53) and assume its solution has the following form.

$$\tilde{\Phi}_t(x) = \frac{|x|^2}{2} + \varphi_1(t,x)\cdot\Delta t + \varphi_0(t,x)\dot{I}_t^{(n)}\Delta t + O(\Delta t^2), \tag{62}$$

where φ_1, φ_0 are undetermined function. Similarly with with (10), after getting solution $\tilde{\Phi}_t$, by using Theorem 2.1, we can get tangent flow (61)

$$\frac{dx_t^{(n)}}{dt} = \nabla\varphi_1(t, x_t^{(n)}) + \nabla\varphi_0(t, x_t^{(n)})\dot{I_t}^{(n)}, \tag{63}$$

where the $\dot{I_t}^{(n)} = 2^n(I_{t_{i+1}} - I_{t_i})$ if $t \in [t_i, t_{i+1})$. Here, the $\nabla\varphi_0(t, x_t^{(n)})$ is a slightly abuse of symbols. In fact $\nabla\varphi_0(t,x) = \big(\nabla\varphi_0^1(t,x), \;\; \cdots, \;\; \nabla\varphi_0^m(t,x)\big)$. In the following, we substitute the specific form of $\tilde{\Phi}_t$ to (53) and expand the right-hand side of (53) as a series in terms of Δt.

$$\begin{aligned}\det D^2\Phi^{(n)}(t, x_t^{(n)}) =& 1 + \Delta\varphi_0(t, x_t^{(n)}) \cdot \Delta I_t \\ &+ \Delta\varphi_1(t, x_t^{(n)}) \cdot \Delta t + O(\Delta t^2)\end{aligned} \tag{64}$$

$$\begin{aligned}&\frac{p^{(n)}(t+\Delta t, x)}{p^{(n)}(t, x_{t+\Delta t}^{(n)})} \\ &= (1 - \frac{1}{p^{(n)}(t, x_t^{(n)})}\nabla p^{(n)}(t, x_t^{(n)})\frac{dx_t^{(n)}}{dt}\Delta t + O(\Delta t^2)) \\ &\cdot (1 + \frac{1}{p^{(n)}}\mathcal{D}(p^{(n)})\Delta t + \frac{1}{p^{(n)}}\mathcal{H}(p^{(n)}) \cdot \dot{I_t}^{(n)}\Delta t + O(\Delta t^2)) \\ &= 1 + \mathcal{D}(p^{(n)})\Delta t + \frac{1}{p^{(n)}}\mathcal{H}(p^{(n)}) \cdot \dot{I_t}^{(n)}\Delta t \\ &- \frac{1}{p^{(n)}(t, x_t^{(n)})}\nabla p^{(n)}(t, x_t^{(n)})\frac{dx_t^{(n)}}{dt}\Delta t + O(\Delta t^2))\end{aligned} \tag{65}$$

Then we take the limit $n \to \infty$ which put $\Delta t \to 0$, the ODE (63) will converge to SDE according to Wong-Zakai approximation [46],

$$dx_t = \nabla\varphi_1(t, x_t)dt + \nabla\varphi_0(t, x_t) \circ dI_t. \tag{66}$$

And the left-hand side of (53) will become

$$\det D^2\Phi_t(x_t) = 1 + \Delta\varphi_0(t, x_t) \circ dI_t + \Delta\varphi_1(t, x_t)dt, \tag{67}$$

the right-hand side of (53) will become

$$\begin{aligned}&\frac{p(t,x) + dp(t,x)}{p(t,x)} = 1 + \frac{1}{p(t, x_t)}[\mathcal{D}(p(t, x_t)dt + \mathcal{H}(p(t, x_t)) \circ dI_t \\ &- \nabla p(t, x_t)(\nabla\varphi_1(t, x_t)dt + \nabla\varphi_0(t, x_t) \circ dI_t)]\end{aligned} \tag{68}$$

Therefore, substituting (67) and (68) to (53), we will get a constraint equation satisfied by φ_1 and φ_0.

Appendix
Proofs in the Section IV

Proofs for Theorem 4.1

At first, we shall submit the posterior density function $p(t,x) = c\exp\big(-\frac{1}{2}(x-\mu(t))^\top P^{-1}(x-\mu(t))\big)$ into the PDE (11), which yields,

$$\Delta\varphi_1(t,x) - (x-\mu(t))^\top P^{-1}(t)\nabla\varphi_1(t,x) = -(\frac{\mathcal{D}(p)}{p}). \tag{69}$$

By using $\frac{\partial p}{\partial t} = \mathcal{D}(p)$, we can transform (69) as

$$\begin{aligned}\Delta\varphi_1(t,x) - (x-\mu(t))^\top P^{-1}(t)\nabla\varphi_1(t,x) &= -(\frac{1}{p}\frac{\partial p}{\partial t}) \\ &= \frac{1}{2}\frac{d}{dt}\big((x-\mu(t))^\top P^{-1}(t)(x-\mu(t))\big)\end{aligned} \tag{70}$$

Next, we shall calculate the right-hand side of (70), which yields,

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\big((x-\mu(t))^\top P^{-1}(x-\mu(t))\big) = \\ &-\frac{1}{2}\left[\frac{d\mu(t)}{dt}^\top P^{-1}(t)(x-\mu(t)) + (x-\mu(t))^\top P^{-1}(t)\frac{d\mu(t)}{dt}\right] \\ &+ \frac{1}{2}(x-\mu(t))^\top \frac{dP^{-1}(t)}{dt}(x-\mu(t)).\end{aligned} \tag{71}$$

Then, we can verify that there is a linear function $\nabla\varphi_1(t,x)$ which solves (70). We assume that $\nabla\varphi_1(t,x) = U(t)x + l(t)$, where $U(t)$ is assumed to be symmetric. Since the right-hand side of (70) is a quadratic function, we can solve for the PDE (69) by verifying the coefficients of the quadratic functions on both sides.

We start with the quadratic coefficients, and we can have,

$$\begin{aligned}&-\frac{1}{2}(U(t)P^{-1}(t) + P^{-1}(t)U(t)) = \frac{1}{2}\frac{dP^{-1}(t)}{dt} \\ &= \frac{1}{2}\big(-P^{-1}(t)A(t) - A^\top(t)P^{-1}(t) - P^{-1}(t)BB^\top P^{-1}(t)\big)\end{aligned} \tag{72}$$

Multiply both sides by $2P(t)(\cdot)P(t)$, we can get

$$U(t)P(t) + P(t)U(t) = A(t)P(t) + P(t)A^\top(t) + BB^\top(t). \tag{73}$$

Similarly, we can calculate the linear coefficients for both sides of (69), which yields,

$$\begin{aligned}&-P^{-1}(t)l(t) + U(t)P^{-1}(t)\mu(t) = -P^{-1}(t)\frac{d\mu(t)}{dt} \\ &\big(P^{-1}(t)A(t) + A^\top(t)P^{-1}(t) + P^{-1}(t)BB^\top P^{-1}(t)\big).\end{aligned} \tag{74}$$

By using (73), we can have $U(t)P^{-1}(t)\mu(t) - (P^{-1}(t)A(t) + A^\top(t)P^{-1}(t) + P^{-1}(t)BB^\top P^{-1}(t))\mu(t) = P^{-1}U(t)\mu(t)$. So, (74) can be simplified as

$$l(t) = -U(t)\mu(t) + \frac{d\mu(t)}{dt} = (A(t) - U(t))\mu(t). \tag{75}$$

Finally, we assume $U(t) := A(t) + \frac{1}{2}BB^\top(t)P^{-1}(t) + \Omega(t)P^{-1}(t)$, then we submit it into (73), which yields,

$$\begin{aligned}\Omega(t)P^{-1}(t) + P^{-1}(t)\Omega(t) &= A(t) - A^\top(t)\\ &+ BB^\top(t)P^{-1}(t) - P^{-1}(t)BB^\top(t).\end{aligned} \tag{76}$$

Here, we finish the proof. □

Proof for Theorem 4.2

The proof of Theorem 4.2 is similar to the proof of Theorem 4.1. At first, we shall differential the posterior density function $p(t,x) = c(t)\exp\left(-\frac{1}{2}(x-\mu(t))^\top P^{-1}(t)(x-\mu(t))\right)$, where the $\mu(t), P(t)$ is the conditional mean and conditional variance in KBF, which yields

$$\begin{aligned}dp(t,x) = &-\frac{1}{2}p(t,x)d\left((x-\mu(t))^\top P^{-1}(x-\mu(t))\right)\\ &+\frac{dc(t)}{dt}\exp\left(-\frac{1}{2}(x-\mu(t))^\top P^{-1}(t)(x-\mu(t))\right).\end{aligned} \tag{77}$$

Here, we consider the drift condition of the tangent flow of (77), and we have

$$\Delta\varphi_1(t,x) - (x-\mu(t))^\top P^{-1}(t)\nabla\varphi_1(t,x) = -(\frac{\mathcal{D}(p)}{p}). \tag{78}$$

By using (77) and the drift term of $d\mu(t)$, we can have

$$\begin{aligned}&-(\frac{\mathcal{D}(p)}{p}) =\\ &\frac{1}{2}\left((-A(t)\mu(t) + (K(t)+C(t))H(t)\mu(t))^\top P^{-1}(t)(x-\mu(t))\right)\\ &+\frac{1}{2}\left((x-\mu(t))^\top P^{-1}(t)(-(A(t)\mu(t) + K(t) + C(t)H(t)\mu(t)))\right)\\ &+\frac{1}{2}\left((x-\mu(t))^\top \frac{dP^{-1}(t)}{dt}(x-\mu(t)))\right) - \frac{dc(t)}{dt}\frac{1}{c(t)}\end{aligned} \tag{79}$$

Similarly, we can verify that there is a linear function $\nabla\varphi_1(t,x)$ which solves (78). We assume that $\nabla\varphi_1(t,x) = U(t)x + l(t)$. Since the right-hand side of (79) is a quadratic function, we can solve (78) by verifying the coefficients of the quadratic functions on both sides. We start with the quadratic coefficients, which yields,

$$\begin{aligned}-\frac{1}{2}(U(t)P^{-1}(t)+P^{-1}(t)U(t)) &= \frac{1}{2}\frac{dP^{-1}(t)}{dt}\\ &= -\frac{1}{2}P^{-1}(t)\frac{dP(t)}{dt}P^{-1}(t)\end{aligned} \tag{80}$$

Multiply both sides by $2P(t)(\cdot)P(t)$ and combining with KBF, we can derive equation satisfied by $U(t)$,

$$\begin{aligned}U(t)P(t) + P(t)U(t) =&(A(t) - CH(t))P(t)\\ &+ P(t)(A(t) - CH(t))^\top + R(t)\\ &- K(t)(DD^\top(t))K(t)^\top,\end{aligned}$$

where $K(t) := [P(t)H^\top(t)](DD^\top(t))^{-1}$ (Kalman Gain), $R(t) := BB^\top(t) - BD^\top(t)(DD^\top(t))^{-1}DB^\top(t)$ and $C(t) = BD(t)^\top(DD^\top(t))^{-1}$.

Similarly, we can calculate the linear coefficients for both sides of (78), which yields equation satisfied by $l(t)$

$$\begin{aligned}&-P^{-1}(t)l(t) + U(t)P^{-1}(t)\mu(t) =\\ &-P^{-1}(t)(A(t)\mu(t) - KH(t)\mu(t))\\ &-\frac{dP^{-1}(t)}{dt}\mu(t).\end{aligned} \tag{81}$$

By using (80), we can have $U(t)P^{-1}(t)\mu(t) + \frac{dP^{-1}(t)}{dt}\mu(t) = P^{-1}U(t)\mu(t)$. So, (81) can be simplified as

$$l(t) = -U(t)\mu(t) + (A(t) - CH(t) - KH(t))\mu(t). \tag{82}$$

Finally, shall assume specific solution form as $U(t) := A(t) - CH(t) + \frac{1}{2}RP^{-1}(t) - \frac{1}{2}KH(t) + \Omega(t)P^{-1}(t)$, then we submit it into (81), which yields,

$$\begin{aligned}\Omega(t)P(t)^{-1}+P(t)^{-1}\Omega(t) =&\\ &(A(t) - C(t)H(t))^\top - (A(t) - C(t)H(t))\\ &+\frac{1}{2}K(t)(DD^\top(t))K^\top(t)P^{-1}(t)\\ &-\frac{1}{2}P^{-1}(t)K(t)(DD^\top(t))K^\top(t)\\ &+\frac{1}{2}R(t)P(t)^{-1} - P(t)^{-1}R(t)).\end{aligned} \tag{83}$$

We can notice that only $d\mu(t)$ contains the stochastic term. So, by using (77), the stochastic term is $(x-\mu(t))P^{-1}(K(t) + C(t))\circ dy_t$. Then, according to $-\bar{\Delta}\varphi_0(t,x) + \nabla(p(t,x))\cdot\bar{\nabla}\varphi_0(t,x) = \frac{\mathcal{H}(p)}{p}$,, we can verify that the $\nabla\phi_0(x) := K(t) + C(t)$. Finally, we finish the proof. □

Appendix
Proofs in Section V

Proofs for Theorem 5.1 Let $e(t) = \mu^{(N)}(t) - \mu(t)$ and by using (38), we have

$$d\mu^{(N)}(t) = A(t)\mu_t^{(N)}dt$$

So, the evolution equation of $e(t)$ satisfies,

$$de(t) = A(t)e(t)dt$$

Here, we can have

$$\begin{aligned}d\|e(t)\|^2 &= d\langle e(t), e(t)\rangle\\ &= \langle (A(t) + A^\top(t))e(t), e(t)\rangle dt\\ &\le 2\lambda\|e(t)\|^2 dt.\end{aligned} \tag{84}$$

And by using Gronwall's inequality, we get,

$$\|e(t)\|^2 \le c_1 e^{\lambda t}\|e(0)\|^2. \tag{85}$$

Similarly, we can get

$$\frac{dP^{(N)}}{dt} = A(t)P^{(N)} + P^{(N)}A^\top(t) + BB^\top. \tag{86}$$

Let $\Theta(t) = P^{(N)}(t) - P(t)$, which yields,

$$\frac{d\Theta(t)}{dt} = A(t)\Theta(t) + \Theta(t)A^\top(t). \tag{87}$$

We further consider to take differential of $\|\Theta(t)\|_F^2$, which yields,

$$\begin{aligned}\frac{d\|\Theta(t)\|_F^2}{dt} \leq & 2\lambda_{max}(A(t)+A^\top(t))\|\Theta(t)\|_F^2 \\ & \leq 4\lambda\|\Theta(t)\|_F^2\end{aligned} \tag{88}$$

And by using Gronwall's inequality, we get,

$$\|\Theta(t)\|_F^2 \leq c_1 e^{\lambda t}\|\Theta(0)\|^2. \tag{89}$$

Proofs for Theorem 5.2

By using the Lemma 5.3 and 4.2, we can get the

$$\begin{aligned}& d\|P_-(t)-P^{(N)}(t)\|_F^2 = \\ & 2\mathrm{Tr}\Bigg\{\Bigg[(A(t)-C(t)H(t))+(A(t)-C(t)H(t))^\top \\ & -\frac{1}{2}(P_-(t)+P^{(N)}(t))S(t)-\frac{1}{2}S(t)(P_-(t)+P^{(N)}(t))\Bigg] \\ & \qquad \times (P_-(t)-P^{(N)}(t))^2\Bigg\}\end{aligned} \tag{90}$$

Since the $P(t), P^{(N)}, S(t)$ are all positive defined matrix and $S(t)$ is scaler matrix, which yields,

$$\begin{aligned}& d\|P_-(t)-P^{(N)}(t)\|_F^2 \leq \\ & 2\mathrm{Tr}\Bigg\{\left[(A(t)-C(t)H(t))+(A(t)-C(t)H(t))^\top\right] \\ & \qquad \times (P_-(t)-P^{(N)}(t))^2\Bigg\} \\ & \leq 4\lambda(\mathrm{Tr}(P_-(t)-P^{(N)}(t))^2 = 4\lambda\|P_-(t)-P^{(N)}(t)\|_F^2 dt.\end{aligned} \tag{91}$$

Then, according to the second term in Lemma 5.1, the $P_-(t)-P^{(N)}(t)$ can be rewritten as $\Phi_t(P_-(0))-\Phi_t(P^{(N)}(0))$ and we get

$$\begin{aligned}& d\|P_-(t)-P^{(N)}(t)\|_F^2 \leq 4\lambda\|P_-(t)-P^{(N)}(t)\|_F^2 dt \\ & \leq 4\lambda\alpha^2\exp(-2\beta t)n^2\|P_-(0)-P^{(N)}(0)\|_2^2 dt\end{aligned} \tag{92}$$

And by using Gronwall's inequality, we can have

$$\|P_-(t)-P^{(N)}(t)\|_F^2 \leq (1+\frac{4n^2\lambda\alpha}{\beta})e^{-2\beta t}\|P_-(0)-P^{(N)}(0)\|_F^2 \tag{93}$$

Next, we start to estimate the $\mu_-(t)-\mu^{(N)}(t)$. Similarly, by using the Lemma 5.3 and 4.2, we can have

$$\begin{aligned}d(\mu_-(t)-\mu^{(N)}(t)) = & (A(t)-C(t)H(t)-P^{(N)}S(t)) \\ & \cdot(\mu_-(t)-\mu^{(N)}(t))dt \\ & +(P^{(N)}-P_-(t))H^\top(t)(DD^\top(t))^{-1}dI_t,\end{aligned} \tag{94}$$

where $dI_t = dy_t - H(t)\mu_-(t)dt$ which is called the innovation process and it is a martingale with quadratic variation $d\langle I_t\rangle := DD^\top(t)$. The solution of (94) is given by

$$\begin{aligned}\mu_-(t)-\mu^{(N)}(t) = & E_t(P^{(N)(0)})(\mu_-(0)-\mu^{(N)}(0)) \\ & +\int_0^t \mathcal{E}_{s,t}(P^{(N)}(s)-P_-(s))H^\top(s)(DD^\top(s))^{-1}dI_s.\end{aligned}$$

The norm of the first term is bounded by:

$$\begin{aligned}& \mathbb{E}[\|E_t(P^{(N)}(0))(\mu_-(0)-\mu^{(N)}(0))\|_2^2] \\ & \qquad \leq \alpha_2^2 e^{-2\beta t}\mathbb{E}[\|\mu_-(0)-\mu^{(N)}(0)\|_2^2] \\ & \qquad \leq \alpha_2^2 e^{-2\beta t}\frac{Tr(P_0)}{N}.\end{aligned}$$

The norm of the second term is bounded by:

$$\begin{aligned}& \mathbb{E}\Big[\Big\|\int_0^t \mathcal{E}_{s,t}(P^{(N)}(s)-P_-(s))H^\top(t)(DD^\top(s))^{-1}dI_s\Big\|_2^2\Big] = \\ & \int_0^t \mathbb{E}\Big[\mathcal{E}_{s,t}(P^{(N)}(s)-P_-(s))S(t)\mathcal{E}_{s,t}(P^{(N)}(s)-P_-(s))^\top\Big]ds \\ & \leq \sup_{t\geq 0}\|S(t)\|_2\int_0^t \alpha_2^2 e^{-2\beta(t-s)}e^{-2\beta s)}c\mathbb{E}[\|P^{(N)}-P_-\|^2] \\ & \leq \tilde{c}\frac{1}{N}\end{aligned}$$

Adding the two bounds, the proof is finished. □

Proofs for Theorem 5.3 First, we can approximate the error for the inverse matrix.

$$\begin{aligned}& \|P_-^{(N)^{-1}}(t)-P_-^{-1}(t)\|_F \leq \frac{1}{\lambda_{\min}(P_-(t))}\|P_-(t)P_-^{(N)^{-1}}(t)-I_n\|_F \\ & \leq \frac{1}{\lambda_{\min}(P_-(t))}\|(P_-(t)-P_-^{(N)}(t))P_-^{(N)^{-1}}(t)\|_F \\ & \leq \frac{\|P_-^{(N)^{-1}}(t)\|}{\lambda_{\min}(P_-(t))}\|(P_-(t)-P_-^{(N)}(t))\|_F.\end{aligned} \tag{95}$$

We can conclude that

$$\|P_-^{(N)^{-1}}(t)-P_-^{-1}(t)\|^2 \leq c_1(P_-(t), P_-^{(N)}(t))\frac{1}{N}. \tag{96}$$

Similarly, we can get

$$\|P_+^{(N)^{-1}}(t)-P_+^{-1}(t)\|^2 \leq c_2(P_+(t), \bar{P}_+^{(N)}(t))\frac{1}{N}. \tag{97}$$

Once we assume that the $P(t) = I_n$, then the calculation of the (36) will be simplified as

$$\hat{P}^{-1}(t) = P_-^{-1}(t)+\bar{P}_+^{-1}(t)-I_n. \tag{98}$$

So, if the $P_-{}^{-1}(t)$, $\bar{P}_+^{-1}(t)$ are approximated by the finite-N formulation, solving (98), the $(\hat{P}^{(N)})^{-1}(t)$ can be obtained. Now, we shall estimate the $(\hat{P}^{(N)})^{-1}(t)-\hat{P}^{-1}(t)$, which yields

$$(\hat{P}^{(N)})^{-1}(t)-\hat{P}^{-1}(t) = (P_-^{(N)})^{-1}(t)-P_-^{-1}(t)+(\bar{P}_+^{(N)})^{-1}(t)-\bar{P}_+^{-1}(t). \tag{99}$$

So, we have

$$\|(\hat{P}^{(N)})^{-1}(t)-\hat{P}^{-1}(t)\|_F^2 \leq 2(c_1+c_2)\frac{1}{N}. \tag{100}$$

Then, according to the same method (95) at the beginning of the proof, there is

$$\|(\hat{P}^{(N)})(t)-\hat{P}(t)\|_F^2 \leq c\frac{1}{N}, \tag{101}$$

where c depends on the $\hat{P}(t), P_-(t), \bar{P}_+(t), P_-^{(N)}$,and $\bar{P}_+^{(N)}$.

Secondly, we shall estimate the $\hat{\mu}(t) - \hat{\mu}^{(N)}(t)$ and the $\hat{\mu}^{(N)}(t)$ is defined in

$$\hat{\mu}^{(N)}(t) := \hat{P}^{(N)}(t)[(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t) + (P_-^{(N)})^{-1}(t)\mu_-^{(N)}(t)].$$

Easily, we can divide the $\hat{P}(t)(\bar{P}_+)^{-1}\bar{\mu}_+(t) - \hat{P}^{(N)}(t)(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t)$ into following three terms

$$\begin{aligned}
&\hat{P}(t)(\bar{P}_+)^{-1}\bar{\mu}_+(t) - \hat{P}^{(N)}(t)(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t)\\
=&\hat{P}(t)(\bar{P}_+)^{-1}(\bar{\mu}_+(t) - \bar{\mu}_+^{(N)}(t))\\
&+ \hat{P}(t)((\bar{P}_+)^{-1} - (\bar{P}_+^{(N)})^{-1})\bar{\mu}_+^{(N)}(t)\\
&+ (\hat{P}(t) - \hat{P}^{(N)}(t))(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t).
\end{aligned}$$

Using the convergence results of $\bar{P}_+^{(N)}, \hat{P}^{(N)}$ and $\bar{\mu}_+^{(N)}(t)$, we can estimate the following upper bound

$$\begin{aligned}
&\|\hat{P}(t)(\bar{P}_+)^{-1}(\bar{\mu}_+(t) - \bar{\mu}_+^{(N)}(t))\|_2\\
&\quad\le \|\hat{P}(t)(\bar{P}_+)^{-1}\|_2\|\bar{\mu}_+(t) - \bar{\mu}_+^{(N)}(t))\|_2\\
&\quad\le c_{01}\frac{1}{N},
\end{aligned}\tag{102}$$

$$\begin{aligned}
&\|\hat{P}(t)((\bar{P}_+)^{-1} - (\bar{P}_+^{(N)})^{-1})\bar{\mu}_+^{(N)}(t)\|_2\\
&\qquad\le \|\hat{P}(t)(\bar{P}_+)^{-1}\|_2\|\bar{\mu}_+(t) - \bar{\mu}_+^{(N)}(t))\|_2\\
&\qquad\le c_{02}\frac{1}{N},
\end{aligned}\tag{103}$$

and

$$\begin{aligned}
&\|(\hat{P}(t) - \hat{P}^{(N)}(t))(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t)\|_2\\
&\le \|(\hat{P}(t) - \hat{P}^{(N)}(t))(\bar{P}_+^{(N)})^{-1}\|_2\|\bar{\mu}_+(t) - \bar{\mu}_+^{(N)}(t))\|_2\\
&\le c_{03}\frac{1}{N},
\end{aligned}\tag{104}$$

By summing up the (102),(103), and (104), we get

$$\begin{aligned}
&\Big\|\hat{P}(t)(\bar{P}_+)^{-1}\bar{\mu}_+(t)\\
&- \hat{P}^{(N)}(t)(\bar{P}_+^{(N)})^{-1}\bar{\mu}_+^{(N)}(t)\Big\|_2 \le c_{00}\frac{1}{N},
\end{aligned}\tag{105}$$

where $c_{00} := c_{01} + c_{02} + c_{03}$ depends on $\bar{P}_+^{(N)}, \hat{P}^{(N)}$ and $\bar{\mu}_+^{(N)}(t)$. Similarly, $\hat{P}(t)(P_-)^{-1}\bar{\mu}_+(t) - \hat{P}^{(N)}(t)(P_-^{(N)})^{-1}\bar{\mu}_-^{(N)}(t)$ can be bounded in the same way. Similarly, the convergence analysis of $\hat{P}(t)(P_-^{-1}(t)\mu_-^{(N)}(t) - \hat{P}^{(N)}(t)(P_-^{(N)})^{-1}(t)\mu_-^{(N)}(t)$ can be derivated as well. Finally, the convergence of $\hat{\mu}(t) - \hat{\mu}^{(N)}(t)$ can be given. □

REFERENCES

[1] D. Q. Mayne, “A solution of the smoothing problem for linear dynamic systems,” *Automatica*, vol. 4, no. 2, pp. 73–92, 1966.

[2] D. Fraser and J. Potter, “The optimum linear smoother as a combination of two optimum linear filters,” *IEEE Transactions on automatic control*, vol. 14, no. 4, pp. 387–390, 1969.

[3] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[4] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[5] A. Bain and D. Crisan, *Fundamentals of stochastic filtering*. Springer, 2009, vol. 3.

[6] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[7] H. E. Rauch, F. Tung, and C. T. Striebel, “Maximum likelihood estimates of linear dynamic systems,” *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[8] D. Q. Mayne, “A solution of the smoothing problem for linear dynamic systems,” *Automatica*, vol. 4, no. 2, pp. 73–92, 1966.

[9] D. Fraser and J. Potter, “The optimum linear smoother as a combination of two optimum linear filters,” *IEEE Transactions on automatic control*, vol. 14, no. 4, pp. 387–390, 1969.

[10] F. Badawi, A. Lindquist, and M. Pavon, “A stochastic realization approach to the smoothing problem,” *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 878–888, 1979.

[11] M. Pavon, “Optimal interpolation for linear stochastic systems,” *SIAM journal on Control and Optimization*, vol. 22, no. 4, pp. 618–629, 1984.

[12] T. Georgiou and A. Lindquist, “Optimal estimation with missing observations via balanced time-symmetric stochastic models,” *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5590–5603, 2017.

[13] N. J. Gordon, D. J. Salmond, and A. F. Smith, “Novel approach to nonlinear/non-gaussian bayesian state estimation,” in *IEE proceedings F (radar and signal processing)*, vol. 140, no. 2. IET, 1993, pp. 107–113.

[14] A. Doucet, A. M. Johansen, *et al.*, “A tutorial on particle filtering and smoothing: Fifteen years later,” *Handbook of nonlinear filtering*, vol. 12, no. 656-704, p. 3, 2009.

[15] T. Yang, P. G. Mehta, and S. P. Meyn, “Feedback particle filter,” *IEEE Transactions on Automatic control*, vol. 58, no. 10, pp. 2465–2480, 2013.

[16] T. Yang, R. S. Laugesen, P. G. Mehta, and S. P. Meyn, “Multivariable feedback particle filter,” *Automatica*, vol. 71, pp. 10–23, 2016.

[17] G. Evensen, “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics,” *Journal of Geophysical Research: Oceans*, vol. 99, no. C5, pp. 10 143–10 162, 1994.

[18] ——, “The ensemble Kalman filter: Theoretical formulation and practical implementation,” *Ocean dynamics*, vol. 53, pp. 343–367, 2003.

[19] ——, “Sampling strategies and square root analysis schemes for the EnKF,” *Ocean dynamics*, vol. 54, no. 6, pp. 539–560, 2004.

[20] G. Evensen and P. J. Van Leeuwen, “An ensemble Kalman smoother for nonlinear dynamics,” *Monthly Weather Review*, vol. 128, no. 6, pp. 1852–1867, 2000.

[21] P. N. Raanes, “On the ensemble Rauch-Tung-Striebel smoother and its equivalence to the ensemble kalman smoother,” *Quarterly Journal of the Royal Meteorological Society*, vol. 142, no. 696, pp. 1259–1264, 2016.

[22] A. Taghvaei and P. G. Mehta, “An optimal transport formulation of the ensemble kalman filter,” *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 3052–3067, 2020.

[23] C. Liu, J. Zhuo, and J. Zhu, “Understanding MCMC dynamics as flows on the Wasserstein space,” in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. Long Beach, California USA: PMLR, 09–15 Jun 2019, pp. 4093–4103.

[24] A. Taghvaei and P. G. Mehta, “An optimal transport formulation of the linear feedback particle filter,” in *American Control Conference*, 2016.

[25] J. Kang, X. Chen, Y. Tao, and S. S.-T. Yau, “Optimal transportation particle filter for linear filtering systems with correlated noises,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 6, pp. 5190–5203, 2022.

[26] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[27] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003, vol. 58.

[28] Y. Brenier, “Polar factorization and monotone rearrangement of vector-valued functions,” *Communications on pure and applied mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

[29] F. Santambrogio, “{Euclidean, metric, and Wasserstein} gradient flows: an overview,” *Bulletin of Mathematical Sciences*, vol. 7, pp. 87–154, 2017.

[30] W. Li and L. Ying, “Hessian transport gradient flows,” *Research in the Mathematical Sciences*, vol. 6, no. 4, p. 34, 2019.

[31] A. Taghvaei and P. G. Mehta, “Optimal transportation methods in nonlinear filtering,” *IEEE Control Systems Magazine*, vol. 41, no. 4, pp. 34–49, 2021.

[32] E. Wong and M. Zakai, “On the convergence of ordinary integrals to stochastic integrals,” *The Annals of Mathematical Statistics*, vol. 36, no. 5, pp. 1560–1564, 1965.

[33] W. Eugene and Z. Moshe, "On the relation between Ordinary and stochastic differential equations," *International Journal of Engineering Science*, vol. 3, no. 2, pp. 213–229, 1965.

[34] K. Twardowska, "Wong-Zakai approximations for stochastic differential equations," *Acta Applicandae Mathematica*, vol. 43, pp. 317–359, 1996.

[35] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[36] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[37] C. Liu, J. Zhu, and Y. Song, "Stochastic gradient geodesic mcmc methods," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/77f959f119f4fb2321e9ce801e2f5163-Paper.pdf

[38] A. E. Raftery and S. M. Lewis, "Implementing mcmc," *Markov chain Monte Carlo in practice*, pp. 115–130, 1996.

[39] Q. Liu, "Stein variational gradient descent as gradient flow," *Advances in neural information processing systems*, vol. 30, 2017.

[40] Y. Chen, T. T. Georgiou, and M. Pavon, "Optimal steering of a linear stochastic system to a final probability distribution, part i," *IEEE Transactions on Automatic Control*, vol. 61, no. 5, pp. 1158–1169, 2016.

[41] ——, "Optimal steering of a linear stochastic system to a final probability distribution, part ii," *IEEE Transactions on Automatic Control*, vol. 61, no. 5, pp. 1170–1180, 2016.

[42] ——, "Optimal steering of a linear stochastic system to a final probability distribution—part iii," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 3112–3118, 2018.

[43] C.-T. Chen, *Linear system theory and design*. Saunders college publishing, 1984.

[44] A. N. Bishop and P. Del Moral, "On the stability of Kalman–Bucy diffusion processes," *SIAM Journal on Control and Optimization*, vol. 55, no. 6, pp. 4015–4047, 2017.

[45] X. Chen and S. S.-T. Yau, "On the stability of linear feedback particle filter," *Asian Journal of Mathematics*, vol. (to appear), 2023.

[46] F. Konecny, "On wong-zakai approximation of stochastic differential equations," *Journal of multivariate analysis*, vol. 13, no. 4, pp. 605–611, 1983.

Jiayi Kang received the B.S. degree from the college of mathematics, Sichuan University, Chengdu, China, in 2019 and Ph.D. degree from Department of Mathematical Sciences at Tsinghua University, China in 2024. His research interests include machine learning, nonlinear filtering and bioinformatics.

Xiaopei Jiao received B.S. degree in applied physics and a dual B.S. degree in computer science from Shanghai Jiao Tong University, China in 2017 and a Ph.D. degree in applied mathematics from Tsinghua University, China in 2022. After his Ph.D., he joined as a postdoctoral researcher at the Beijing Institute of Mathematical Sciences and Applications (BIMSA) and subsequently at the University of Twente, Netherlands. Now, he is an assistant professor at BIMSA, China. His research interests include nonlinear filter theory and numerical application, Physics-Informed neural networks (PINNs), and numerical methods in partial differential equations.

Stephen S.-T. Yau (F' 03) received the Ph.D. degree in mathematics from the State University of New York at Stony Brook, NY, USA in 1976.

He was a Member of the Institute of Advanced Study at Princeton from 1976-1977 and 1981-1982, and a Benjamin Pierce Assistant Professor at Harvard University during 1977-1980. After that, he joined the Department of Mathematics, Statistics and Computer Science (MSCS), University of Illinois at Chicago (UIC), and served for over 30 years. During 2005-2011, he became a joint Professor with the Department of Electrical and Computer Engineering at the MSCS, UIC. After his retirement in 2012, he joined Tsinghua University, Beijing, China, where he is a full-time professor in the Department of Mathematical Sciences. His research interests include nonlinear filtering, bioinformatics, complex algebraic geometry, CR geometry, and singularities theory.

Dr. Yau 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. He was the General Chairman of the IEEE International Conference on Control and Information, which was held in the Chinese University of Hong Kong in 1995. He was awarded the Sloan Fellowship in 1980, the Guggenheim Fellowship in 2000, and the AMS Fellow Award in 2013. In 2005, he was entitled the UIC Distinguished Professor.