2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Optimal Transport Particle Filters

Mohammad Al-Jarrah*, Bamdad Hosseini†, Amirhossein Taghvaei*

***Abstract*— This paper is concerned with the theoretical and computational development of a new class of nonlinear filtering algorithms called optimal transport particle filters (OTPF). The algorithm is based on a recently introduced variational formulation of the Bayes' rule, which aims to find the Brenier optimal transport map between the prior and the posterior distributions as the solution to a stochastic optimization problem. On the theoretical side, the existing methods for the error analysis of particle filters and stability results for optimal transport map estimation are combined to obtain uniform error bounds for the filter's performance in terms of the optimization gap in solving the variational problem. The error analysis reveals a bias-variance trade-off that can ultimately be used to understand if/when the curse of dimensionality can be avoided in these filters. On the computational side, the proposed algorithm is evaluated on a nonlinear filtering example in comparison with the ensemble Kalman filter (EnKF) and the sequential importance resampling (SIR) particle filter.**

I. Introduction

Optimal transportation (OT) theory has gained significant interest recently because it provides natural geometrical and mathematical tools for analysis and manipulation of probability distributions [1], [51], [36]. In particular, two geometric notions are of key importance: (i) a metric to measure the similarity/discrepancy between probability distributions, i.e., the Wasserstein metric, and (ii) a map to transport one distribution to the other, i.e., the OT or Monge map. In contrast to their information-theoretic counterparts (such as the Kullback-Liebler (KL) divergence), these OT metrics respect the geometry of the data and are often more robust against perturbations and errors. Due to these unique features, OT metrics and maps have been successfully employed in a variety of applications, including generative modeling and sampling [4], [47], domain adaptation [11], and image processing [24], [15], [31], [21], [41].

Given the significance of OT theory, there has been a growing line of research to apply OT tools for nonlinear filtering and Bayesian inference [17], [9], [35], [46]. The central idea here is to view the filtering/Bayesian update step as the problem of transporting the prior distribution (of the current state) to the posterior distribution (of the future state). This perspective has led to the development of new algorithms for Bayesian inference, namely learning triangular transport maps with polynomial, radial basis, and neural net parameterizations to sample from posteriors [17], [28], [25], ensemble transform particle filters [34], and OT interpretations of the feedback particle filter algorithm [52], [44], [45], [46].

This paper builds on the authors' recent work [43] where an OT-based variational formulation of the Bayes' law was introduced to learn the OT map from the prior to the posterior distribution for any value of the observation signal. In this formulation, the conditional distribution $P_{X|Y}$, of a hidden random variable X given the observation Y, is identified as $P_{X|Y}(\cdot|y) = \nabla_x \bar{f}(\cdot, y)\# P_X(\cdot)$, $\forall y$, i.e., the push-forward of the prior distribution P_X with respect to a map of the form $\nabla \bar{f}$ where $\bar{f}$ is a real-valued function that solves the optimization problem:

$$\bar{f} = \operatorname*{arg\,min}_{f \in \mathrm{CVX}_x} \mathbb{E}[f(\bar{X}, Y) + f^*(X, Y)]. \tag{1}$$

Here $\bar{X}$ is an independent copy of X and CVX_x denotes the set of functions $f(x; y)$ that are convex with respect to the x argument for any fixed y, and f^* is the convex conjugate of f with respect to the x argument.

The above variational formulation enjoys three key features that distinguish it from prior works: (i) It is simulation-based in the sense that it is possible to approximate the objective function in terms of samples from the joint distribution P_{XY} and does not require an explicit formula for the likelihood; (ii) The variational formulation enables new approximation methods for computing the posterior distribution by choosing different subsets/parameterizations of the set CVX_x; (iii) The problem (1) is stochastic and can be solved efficiently using recent machine learning techniques, for example, f can be parameterized as a deep neural network and trained using stochastic gradient descent. Problem (1) can be obtained as the dual form of a block-triangular Monge problem between the independent coupling $P_X \otimes P_Y$ and the joint distribution P_{XY}. Similar variational formulations arise in block-triangular transport of distributions in the context of conditional generative models; for example [40], [25], [32], [38], [39].

The objective of the current paper is to use the formulation (1) to develop a new nonlinear filtering algorithm, called *optimal transport particle filter* (OTPF), and provide preliminary theoretical analysis and numerical validation of the algorithm. The proposed algorithm can be viewed as a nonlinear and non-Gaussian generalization of the ensemble Kalman filter algorithm (EnKF) [18], [7] and the discrete-time counterpart of the feedback particle filter (FPF) algorithm [53], [52] (OTPF solves the gain function approximation and the numerical time discretization problems in the FPF altogether by solving the proposed variational

Mohammad Al-Jarrah and Amirhossein Taghvaei are supported by the National Science Foundation (NSF) award EPCN-2318977. Bamdad Hosseini is supported by the NSF award DMS-2208535

*Department of Aeronautics & Astronautics, University of Washington, Seattle; mohd9485@uw.edu, amirtag@uw.edu.

†Department of Applied Mathematics, University of Washington, Seattle bamdadh@uw.edu.

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383337

problem (1)).

The theoretical analysis of the paper is concerned with the error analysis of the proposed algorithm. In particular, we study how errors solving problem (1) at each time step affect the overall performance of the filtering algorithm. To do so, we adapt the existing methods for error analysis of particle filters (PF) to obtain a uniform bound on the filtering error in terms of the approximation error of the OT map [14], [49], [8], [13]. These results are based on a strong notion of uniform geometric filter stability [6], which is common in the analysis of PF. Next, we combine this with stability results for the estimation of OT maps [22] which relates the approximation error of the map to the optimization gap of (1) (see Lemma 2) . The error analysis is carried out for the mean-field limit of the algorithm and a variant of the particle system that involves an additional resampling step which makes the particles independent of each other and significantly simplifies the analysis.

The numerical experiments qualitatively and quantitatively evaluate the performance of the OTPF in comparison with the EnKF algorithm [18], [7] and the sequential importance resampling (SIR) PF [16]. In particular, we consider a linear stable dynamical system with three different observation functions: linear, quadratic, and cubic. The numerical results illustrate the versatile nature of the OTPF compared to the other two methods.

The rest of the paper is organized as follows: Section II reviews the filtering problem and equations, and introduces the notion of filter stability; Section III outlines the OTPF algorithm in detail; Section IV presents the error analysis; and Section V contains the numerical experiments.

II. Problem Formulation

A. Filtering problem

Consider a discrete-time stochastic dynamic system given by the update equations

$$X_t \sim a(\cdot|X_{t-1}), \quad X_0 \sim \pi_0 \tag{2a}$$
$$Y_t \sim h(\cdot|X_t) \tag{2b}$$

for $t = 1, 2, \dots$ where $X_t \in \mathbb{R}^n$ is the state of the system, $Y_t \in \mathbb{R}^m$ is the observation, π_0 is the probability distribution for the initial state X_0, $a(x'|x)$ is the probability kernel for the transition from the state x to the state x', and $h(y|x)$ is the likelihood distribution of an observation y given a state x. We assume that the update equation (2a) is realized with the stochastic map

$$X_t = \bar{a}(X_{t-1}, V_t) \tag{2c}$$

where $\{V_t\}_{t=1}^{\infty}$ is an i.i.d sequence and $\bar{a}(x, v)$ is Lipschitz in x for all v. Throughout the paper, we assume that all probability measures admit a density and use the same notation to refer to the distribution or the corresponding measure. If needed, the two notions will be distinguished depending on the context.

The filtering problem is to infer the conditional distribution of the state X_t given the history of the observations $\{Y_1, \dots, Y_t\}$, that is, the distribution

$$\pi_t := \mathbb{P}(X_t \in \cdot | Y_1, \dots, Y_t), \quad \text{for} \quad t = 1, 2, \dots,$$

often referred to as the *posterior* distribution.

B. Recursive update for the filter

The posterior distribution π_t admits a recursive update equation that is essential for the design of filtering algorithms. To present this recursive update, let us introduce the following operators:

$$\text{(propagation)} \quad \pi \mapsto \mathcal{A}\pi := \int_{\mathbb{R}^n} a(\cdot|x)\pi(x)dx \tag{3a}$$
$$\text{(conditioning)} \quad \pi \mapsto \mathcal{B}_y\pi := \frac{h(y|\cdot)\pi(\cdot)}{\int_{\mathbb{R}^n} h(y|x)\pi(x)dx} \tag{3b}$$

The first operator represents the update for the distribution of the state according to the dynamic model (2a). The second operator represents Bayes' rule that carries out the conditioning according to the observation model (2b). In terms of these two operators, the update law for the posterior is given by (e.g. see [8]):

$$\pi_t = \mathcal{T}_t\pi_{t-1} = \mathcal{B}_{Y_t}\mathcal{A}\pi_{t-1}. \tag{3c}$$

where we introduced $\mathcal{T}_t := \mathcal{B}_{Y_t}\mathcal{A}$. With slight abuse of notation, we further define the transition operator as

$$\mathcal{T}_{t,s} := \mathcal{T}_t \circ \cdots \circ \mathcal{T}_{s+1}, \quad \forall \quad t > s \geq 0.$$

We then have $\pi_t = \mathcal{T}_{t,s}\pi_s$ for all $t > s \geq 0$. Note that the transition operator $\mathcal{T}_{t,s}$ is stochastic in nature as it depends on the realization of the observation signal $\{Y_{s+1}, \dots, Y_t\}$. We suppress this dependence to simplify the presentation.

C. Filter stability

We use the following metric on (possibly random) probability measures μ, ν:

$$d(\mu, \nu) := \sup_{g \in \mathcal{G}} \sqrt{\mathbb{E}\left|\int g d\mu - \int g d\nu\right|^2} \tag{4}$$

where the expectation is over the possible randomness of the probability measures μ and ν, and $\mathcal{G} := \{g : \mathbb{R}^n \to \mathbb{R}; |g(x)| \leq 1, |g(x) - g(x')| \leq \|x - x'\|, \quad \forall x, x'\}$ is the space of functions that are uniformly bounded by one and uniformly Lipschitz with a constant smaller than one (this metric is also known as the dual bounded-Lipschitz distance). We use this metric to introduce a notion of uniform geometrical stability for the filter.

Definition 1 (Uniformly geometrically stable filter): The filter update (3) is uniformly geometrically stable if $\exists \lambda \in (0, 1)$ and positive constant $C > 0$ such that for all μ, ν and $t > s \geq 0$ it holds that

$$d(\mathcal{T}_{t,s}\mu, \mathcal{T}_{t,s}\nu) \leq C(1 - \lambda)^{t-s} d(\mu, \nu). \tag{5}$$

Remark 1: The uniform geometric stability property (5) is also used in the error analysis of PFs in [14], [13]. It can be verified if the dynamic transition kernel satisfies a

minorization condition, i.e., there exists a probability measure ρ and a constant $\epsilon > 0$ such that $a(x|x') \geq \epsilon\rho(x)$. The minorization is a mixing condition that ensures geometric ergodicity of the Markov process X_t [29]. We acknowledge that this condition is strong and can be verified for a restricted class of systems, e.g., X_t should belong to a compact set. A complete characterization of systems with uniform geometric stable filters is an open and challenging problem in the field. More insight is available for the weaker notion of asymptotic stability of the filter, i.e., $\lim_{t\to\infty} d(\mathcal{T}_{t,s}\mu, \mathcal{T}_{t,s}\nu) = 0$, which holds when the system is "detectable" in a sense that is suitable for nonlinear stochastic dynamical systems [50], [10], [48], [23]. This characterization of systems with asymptotic filter stability is in agreement with the existing results for the stability of the Kalman filter, which holds when the linear system is detectable in the classical sense [30]. A complete survey of existing filter stability results can be found in [12].

The following Lemma is useful for our error analysis.

Lemma 1: Let π be a (random) distribution and T and S two (random) measurable maps. Then,

$$d(T\#\pi; S\#\pi) \leq \mathbb{E}\left[\|T - S\|^2_{L^2(\pi)}\right]^{\frac{1}{2}},$$

where the expectation is over the possible randomness of the distribution π as well as the maps T, S.

Proof: The proof follows from a straightforward argument using the definition of the metric (4) and the Lipschitz property of the test functions g. ∎

III. Optimal Transport Particle Filters

The construction of OTPFs relies on the variational formulation (1). Consider the objective function

$$J(f, \pi) := \mathbb{E}[f(\bar{X}, Y) + f^\star(X, Y)], \tag{6}$$

where $X \sim \pi$, $Y \sim h(\cdot|X)$, and $\bar{X} \sim \pi$ is an independent copy of X, along with the optimization problem

$$\inf_{f \in \mathrm{CVX}_x} J(f, \pi). \tag{7}$$

It is shown in [43, Prop. 1] that the solution to this problem provides an OT characterization of the Bayes operator (3b). The result is reproduced here for completeness.

Proposition 1: Assume π admits a density with respect to the Lebesgue measure. Then, the objective function (6) has a unique (up to a constant shift) minimizer $\bar{f} \in \mathrm{CVX}_x$ and

$$\nabla_x \bar{f}(\cdot, y)\#\pi = \mathcal{B}_y\pi, \quad \text{for a.e. } y. \tag{8}$$

A. The exact mean-field process

We use the OT characterization of the conditional distribution to construct a (exact) mean-field process $\bar{X}_t$ whose distribution $\bar{\pi}_t$ is exactly equal to the posterior distribution π_t. Consider a process $\bar{X}_t$ with distribution $\bar{\pi}_t$ defined as

$$\begin{aligned} \bar{X}_t &= \nabla_x \bar{f}_t(\bar{a}(\bar{X}_{t-1}, \bar{V}_t), Y_t), \quad \bar{X}_0 \sim \bar{\pi}_0 \\ \bar{f}_t &= \operatorname*{arg\,min}_{f \in \mathrm{CVX}_x} J(f, \mathcal{A}\bar{\pi}_{t-1}), \end{aligned} \tag{9a}$$

where $\bar{V}_t$ is an independent copy of V_t in the dynamic model (2c). It is then straightforward to verify that

$$\bar{\pi}_t = \nabla_x \bar{f}_t(\cdot, Y_t)\#\mathcal{A}\bar{\pi}_{t-1} = \mathcal{B}_y\mathcal{A}\bar{\pi}_{t-1}, \tag{9b}$$

where the second identity is a consequence of Proposition 1. It then follows that whenever $\bar{\pi}_0 = \pi_0$ then $\bar{\pi}_t = \pi_t$. As such, the mean-field process $\bar{X}_t$ is called exact. The OTPF is obtained by approximating the exact mean-field process $\bar{X}_t$ in two steps, as described next.

B. The approximate mean-field process

The first approximation step consists of restricting the feasible set of the optimization problem (7) to a parameterized class of convex functions $\mathcal{F} \subset \mathrm{CVX}_x$. The resulting approximated distribution is denoted by $\pi^{\mathcal{F}}_t$ which follows the update rule:

$$\begin{aligned} \pi^{\mathcal{F}}_t &= \nabla_x f^{\mathcal{F}}_t(\cdot, Y_t)\#\mathcal{A}\pi^{\mathcal{F}}_{t-1}, \quad \pi^{\mathcal{F}}_0 = \pi_0 \\ f^{\mathcal{F}}_t &= \operatorname*{arg\,min}_{f \in \mathcal{F}} J(f, \mathcal{A}\pi^{\mathcal{F}}_{t-1}) \end{aligned} \tag{10a}$$

This update defines the approximate mean-field process

$$X^{\mathcal{F}}_t = \nabla_x f^{\mathcal{F}}_t(\bar{a}(X^{\mathcal{F}}_{t-1}, Y_t), \bar{V}_t), \quad X^{\mathcal{F}}_0 \sim \bar{\pi}_0. \tag{10b}$$

The approximation error between $\pi^{\mathcal{F}}_t$ and $\bar{\pi}_t$, due to the parameterization of the function f_t is studied in section IV-A.

C. The finite particle system

The second approximation step is to replace the mean-field process with an empirical distribution of a collection of particles $\{X^1_t, \ldots, X^N_t\}$, i.e., $\pi^{\mathcal{F}}_t \approx \frac{1}{N}\sum_{i=1}^N \delta_{X^i_t}$. The finite-$N$ discretization can be achieved through two different approaches leading to two different systems of particles. The first system (the particle system with resampling) is more amenable to error analysis, while the second system (the interacting particle system) is more practical.

(C.I) the particle system with resampling: Define the sampling operator

$$\pi \mapsto \mathcal{S}^N\pi := \frac{1}{N}\sum_{i=1}^N \delta_{X^i} \quad X^i \overset{\text{i.i.d.}}{\sim} \pi, \tag{11}$$

and approximate the mean-field distribution $\pi^{\mathcal{F}}_t$ by introducing the sampling operator $\mathcal{S}^N$ within the update equations:

$$\begin{aligned} \tilde{\pi}^{(\mathcal{F},N)}_t &= \nabla_x \tilde{f}^{(\mathcal{F},N)}_t(\cdot, Y_t)\#\mathcal{S}^N\mathcal{A}\tilde{\pi}^{(\mathcal{F},N)}_{t-1}, \quad \pi^{(\mathcal{F},N)}_0 = \pi_0 \\ \tilde{f}^{(\mathcal{F},N)}_t &= \operatorname*{arg\,min}_{f \in \mathcal{F}} J(f, \mathcal{S}^N\mathcal{A}\tilde{\pi}^{(\mathcal{F},N)}_{t-1}), \end{aligned} \tag{12a}$$

The presence of the sampling operator ensures that the distribution $\tilde{\pi}^{(\mathcal{F},N)}_t$ is an empirical distribution formed by a collection of particles, i.e. $\tilde{\pi}^{(\mathcal{F},N)}_t = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}^i_t}$. Equation (12a) further identifies an update law for the particles:

$$\tilde{X}^i_t = \nabla_x \tilde{f}^{(\mathcal{F},N)}_t(\bar{a}(\tilde{X}^{\sigma_i}_{t-1}, V^i_t), Y_t) \tag{12b}$$

where $\sigma_i \sim \mathrm{Unif}\{1, 2, \ldots, N\}$ and $\{V^i_t\}_{i=1}^N$ are independent copies of V_t. The sampling process is similar to the resampling stage in PFs, with the difference being that the

weights are uniform in this case. The resampling step makes the particles independent of each other, which significantly simplifies the error analysis, as seen in Section IV-B.

(C.II) the interacting particle system: The second approach to constructing the finite-N particle system is to discretize the update equation (10b) for the mean-field process $X_t^{\mathcal{F}}$ according to

$$\begin{aligned} X_t^i &= \nabla_x f_t^{(\mathcal{F},N)}(\bar{a}(X_t^i, V_t^i), Y_t) \\ f_t^{(\mathcal{F},N)} &= \arg\min_{f\in\mathcal{F}} J(f, \frac{1}{N}\sum_{i=1}^N \delta_{\bar{a}(X_t^i,V_t^i)}). \end{aligned} \tag{13}$$

The empirical distribution $\pi_t^{(\mathcal{F},N)} := \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$ does not follow an update-law similar to the update law for $\bar{\pi}_t^{(\mathcal{F},N)}$ due to the nature of the operator $\mathcal{A}$, which smooths out empirical distributions. Instead, the update for the interacting particle system can be expressed as

$$\pi_t^{(\mathcal{F},N)} = \nabla f^{(\mathcal{F},N)} \# \mathcal{A}^N \pi_{t-1}^{(\mathcal{F},N)}$$

where $\mathcal{A}^N$ is a stochastic operator that takes any empirical distribution $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and outputs $\frac{1}{N}\sum_{i=1}^N \delta_{a(x^i,V^i)}$. Moreover, in contrast to the previous construction (the particle system with resampling), the particles are now correlated, which makes the error analysis challenging (this is often studied under the propagation of chaos analysis [42]). We leave the error analysis of the interacting particle system as the subject of future work. However,we empirically validate the performance of this approximation in Section V.

IV. Error Analysis

The objective of this section is to study the approximation error of the OTPFs introduced above. We begin with the analysis for the approximate mean-field process before turning our attention to the particle system with resampling.

A. The mean-field analysis

The distance between the exact mean-field distribution $\bar{\pi}_t$ and the approximate distribution $\pi_t^{\mathcal{F}}$ is characterized by the following proposition.

Proposition 2: Consider $\bar{\pi}_t$ and $\pi_t^{\mathcal{F}}$ as in (9)-(10), respectively. Assume

1) The exact filter is stable according to Definition 1.
2) There exists $\epsilon_{\mathcal{F}} > 0$ such that

$$\inf_{f\in\mathcal{F}} J(f, \mathcal{A}\pi_t^{\mathcal{F}}) - \inf_{f\in \mathrm{CVX}_x} J(f, \mathcal{A}\pi_t^{\mathcal{F}}) \le \epsilon_{\mathcal{F}}, \quad \forall t. \tag{14}$$

3) For all y and t the function $f_t^{\mathcal{F}}(\cdot, y)$ is convex and $\nabla_x f_t^{\mathcal{F}}(\cdot, y)$ is β-Lipschitz.

Then, it holds that

$$d(\pi_t^{\mathcal{F}}, \pi_t) \le \frac{C\sqrt{2\beta\epsilon_{\mathcal{F}}}}{\lambda}, \quad \forall t, \tag{15}$$

with all constants independent of time.

Remark 2: The first assumption in the proposition is used to ensure the error produced at each step of the algorithm does not grow with time. The second assumption is related to the representation power of the function class $\mathcal{F}$ relative to the class of probability distributions introduced by the algorithm $\mathcal{A}\pi_t^{\mathcal{F}}$. For example, this error is zero when $\mathcal{F}$ is a class of convex and quadratic functions, and the filtering problem is based on a linear Gaussian dynamic and observation model. In this case, probability distributions $\pi_t^{\mathcal{F}}$ are Gaussian with the corresponding quadratic optimal function $\bar{f}_t$. In general, it is expected that the error is small when the distributions are inherently simple, e.g. when the problem exhibits low-dimensional structures or regularities. The analysis of these errors is the subject of representation theory [3], [37]. The last assumption is related to the regularity of the distributions $\mathcal{A}\pi_t^{\mathcal{F}}$ and the resulting posterior distributions and can be enforced by an appropriate choice of the class $\mathcal{F}$.

The following Lemma is useful for the proof of the Proposition 2.

Lemma 2: Consider the optimization problem (7) with the objective function (6). Assume π admits density. Let $\bar{f}$ be the optimal function and f be an arbitrary convex and β-smooth function. Then,

$$J(\pi, f) - J(\pi, \bar{f}) \ge \frac{1}{2\beta}\mathbb{E}[\|\nabla f(\bar{X}, Y) - \nabla \bar{f}(\bar{X}, Y)\|^2].$$

Proof: The proof is an extension of the result [22, Prop. 10] and omitted on the account of space. ∎

Proof: [Proof of Proposition 2] To simplify the presentation, we introduce the operator $\pi \mapsto \mathcal{T}_t^{\mathcal{F}}\pi := \nabla f_t^{\mathcal{F}}(\cdot, Y_t)\#\mathcal{A}\pi$ for all t, to denote the update law for the approximate mean-field distribution in (10a). The first step in the proof is to use the triangle inequality and the filter stability to bound the error between π_t and $\pi_t^{\mathcal{F}}$ as follows:

$$\begin{aligned} d(\pi_t, \pi_t^{\mathcal{F}}) &\le \sum_{k=1}^t d(\mathcal{T}_{t,k-1}\pi_{k-1}^{\mathcal{F}}, \mathcal{T}_{t,k}\pi_k^{\mathcal{F}}) \\ &\le \sum_{k=1}^t d(\mathcal{T}_{t,k}\mathcal{T}_k\pi_{k-1}^{\mathcal{F}}, \mathcal{T}_{t,k}\mathcal{T}_k^{\mathcal{F}}\pi_{k-1}^{\mathcal{F}}) \\ &\le \sum_{k=1}^t C(1-\lambda)^{t-k} d(\mathcal{T}_k\pi_{k-1}^{\mathcal{F}}, \mathcal{T}_k^{\mathcal{F}}\pi_{k-1}^{\mathcal{F}}) \\ &\le \frac{C}{\lambda}\max_{k\in\{1,2,\dots,t\}}\{d(\mathcal{T}_k\pi_{k-1}^{\mathcal{F}}, \mathcal{T}_k^{\mathcal{F}}\pi_{k-1}^{\mathcal{F}})\}. \end{aligned}$$

Next, we use Lemma 1 to bound the distance

$$\begin{aligned} &d(\mathcal{T}_k\pi_{k-1}^{\mathcal{F}}, \mathcal{T}_k^{\mathcal{F}}\pi_{k-1}^{\mathcal{F}}) \\ &\quad = d(\nabla \bar{f}_k(\cdot, Y_k)\#\mathcal{A}\pi_{k-1}^{\mathcal{F}}, \nabla f_k^{\mathcal{F}}(\cdot, Y_t)\#\mathcal{A}\pi_{k-1}^{\mathcal{F}}) \\ &\quad \le \mathbb{E}\left[\|\nabla \bar{f}_k(\cdot, Y_k) - \nabla f_k^{\mathcal{F}}(\cdot, Y_k)\|^2_{L^2(\mathcal{A}\pi_{k-1}^{\mathcal{F}})}\right]^{\frac{1}{2}} \end{aligned}$$

for all $k \ge 0$. Finally, we use the second and third assumptions in the proposition to obtain a uniform bound for the error between $\nabla \bar{f}_k$ and $\nabla f_k^{\mathcal{F}}$ using Lemma 2

$$\begin{aligned} &\mathbb{E}\left[\|\nabla \bar{f}_k(\cdot, Y_k) - \nabla f_k^{\mathcal{F}}(\cdot, Y_k)\|^2_{L^2(\mathcal{A}\pi_{k-1}^{\mathcal{F}})}\right] \\ &\quad \le 2\beta(J(f_k^{\mathcal{F}}, \mathcal{A}\pi_{k-1}^{\mathcal{F}}) - J(\bar{f}_k, \mathcal{A}\pi_{k-1}^{\mathcal{F}})) \\ &\quad \le 2\beta\epsilon_{\mathcal{F}} \end{aligned}$$

concluding the final bound (15). ∎

B. The particle-system-with-resampling analysis

Next, we analyze the error between the particle system (12) and the exact mean-field process (9). The process is similar to the mean-field analysis presented in the previous section, with an additional error due to the sampling operator and the empirical approximations.

Proposition 3: Consider the exact mean-field distribution $\bar{\pi}_t$ and the particle distribution $\tilde{\pi}_t^{(\mathcal{F},N)}$ defined in (9) and (12), respectively. Assume

1) The exact filter is stable according to Definition 1.
2) There exists a constant $\epsilon_{\mathcal{F},N} > 0$ such that for all t and N:
$$\inf_{f\in\mathcal{F}} J(f, \mathcal{S}^N\mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)}) - \inf_{f\in \mathrm{CVX}_x} J(f, \mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)}) \le \epsilon_{\mathcal{F},N}$$
3) For all y, t, and N, the function $f_t^{(\mathcal{F},N)}(\cdot, y)$ is convex and $\nabla_x f_t^{(\mathcal{F},N)}(\cdot, y)$ is β-Lipschitz.

Then, it holds that

$$d(\tilde{\pi}_t^{(\mathcal{F},N)}, \pi_t) \le \frac{C}{\lambda}\left(\sqrt{2\beta\epsilon_{\mathcal{F},N}} + \frac{1}{\sqrt{N}}\right), \quad \forall t, \tag{16}$$

where all constants are time-independent.

Proof: The proof is similar to that of Proposition 2. Define the operator $\tilde{\mathcal{T}}_t^{(\mathcal{F},N)} : \pi \mapsto \nabla \tilde{f}_t^{(\mathcal{F},N)}(\cdot, Y_t)\#\mathcal{S}^N\mathcal{A}\pi$. Then, the triangle inequality and filter stability imply

$$d(\pi_t, \pi_t^{(\mathcal{F},N)}) \le \frac{C}{\lambda} \max_{k\in\{1,2,\dots,t\}} \{d(\mathcal{T}_k\pi_{k-1}^{(\mathcal{F},N)}, \tilde{\mathcal{T}}_k^{(\mathcal{F},N)}\pi_{k-1}^{(\mathcal{F},N)})\}.$$

Applying the triangle inequality again, we can write

$$\begin{aligned}
&d(\mathcal{T}_k\pi_{k-1}^{(\mathcal{F},N)}, \tilde{\mathcal{T}}_k^{(\mathcal{F},N)}\pi_{k-1}^{(\mathcal{F},N)}) \\
&= d(\nabla \bar{f}_k(\cdot, Y_t)\#\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}, \nabla f_k^{(\mathcal{F},N)}\#\mathcal{S}^N\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}) \\
&\le d(\nabla \bar{f}_k(\cdot, Y_t)\#\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}, \nabla f_k^{(\mathcal{F},N)}\#\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}) \\
&+ d(\nabla f_k^{(\mathcal{F},N)}\#\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}, \nabla f_k^{(\mathcal{F},N)}\#\mathcal{S}^N\mathcal{A}\pi_{k-1}^{(\mathcal{F},N)}).
\end{aligned}$$

By application of Lemma 1 and Lemma 2, the first term is upper-bounded by the square-root of

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla \bar{f}_k(\cdot, Y_k) - \nabla f_k^{(\mathcal{F},N)}(\cdot, Y_k)\|^2_{L^2(\mathcal{A}\tilde{\pi}_{k-1}^{(\mathcal{F},N)})}\right] \\
&\le 2\beta(J(f_k^{(\mathcal{F},N)}, \mathcal{A}\tilde{\pi}_{k-1}^{(\mathcal{F},N)}) - J(\bar{f}_k, \mathcal{A}\tilde{\pi}_{k-1}^{(\mathcal{F},N)})) \\
&\le 2\beta\epsilon_{\mathcal{F},N}
\end{aligned}$$

where we used the second and the third assumptions. This gives the first term on the right-hand side of (16). The second term is due to the sampling error and upper-bounded by $\frac{1}{\sqrt{N}}$ since the test functions g in the definition of the metric d are uniformly bounded by one (e.g. see [33, Lemma 2.17]). Adding the two errors concludes the final bound. ∎

Remark 3: The assumptions of this proposition are similar to the assumptions in Proposition 2 with a slight difference in the second assumption. The bound in the second assumption can be decomposed into two terms:

$$\begin{aligned}
&\inf_{f\in\mathcal{F}} J(f, \mathcal{S}^N\mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)}) - \inf_{f\in\mathcal{F}} J(f, \mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)}) \\
&+ \inf_{f\in\mathcal{F}} J(f, \mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)}) - \inf_{f\in \mathrm{CVX}_x} J(f, \mathcal{A}\tilde{\pi}_t^{(\mathcal{F},N)})
\end{aligned}$$

The second term is similar to the one used in Proposition 2 and related to the representation power of $\mathcal{F}$. The first term corresponds to the statistical generalization errors due to approximating distributions with empirical samples and the subject of statistical generalization theory [37], [5], [26], [54]. The error is expected to scale according to $O(\frac{C_{\mathcal{F}}}{\sqrt{N}})$ where the constant $C_{\mathcal{F}}$ is a proxy for the complexity of the class of functions $\mathcal{F}$, and independent of the dimension d. The first term can also be interpreted as the variance, while the second term is the bias. Then our error analysis is a manifestation of the bias-variance trade-off dependent on the complexity of the function class $\mathcal{F}$. Similar bias-variance trade-offs also appear in the analysis of local PFs in [33].

V. Numerical Experiments

We use a numerical example to illustrate the proposed OTPF in comparison with two other filters: the Ensemble Kalman Filter (EnKF) [18], and the sequential importance resampling (SIR) PF [16].

For the OTPF, we solve a min-max formulation of the variational problem (7), as described in [43, Sec. III-B] and originally proposed in [27] for estimating OT maps. The min-max formulation involves optimization over an additional convex function ψ which is used to represent the convex conjugate f^* as follows:

$$\mathbb{E}[f^*(X,Y)] = \max_{\psi\in\mathrm{CVX}_x} \mathbb{E}[X^\top\nabla_x\psi(X,Y) - f(\nabla_x\psi(X,Y), Y)]$$

However, in our numerical experiments, we observed that relaxing the constraint and optimizing over a map $T(x;y)$ instead of $\nabla_x\psi(x;y)$ produces better numerical results due to the additional freedom in the parameterization. Therefore, we use the formulation

$$\mathbb{E}[f^*(X;Y)] = \max_T \mathbb{E}[X^\top T(X;Y) - f(T(X;Y);Y)]$$

Note that this does not change the optimization problem because the optimal T is of gradient form and equal to $\nabla_x f^*$. The final objective function takes the form

$$\begin{aligned}
\min_{f\in\mathrm{ICNN}} \max_{T\in\mathrm{ResNet}} \{&\mathbb{E}_{P_{XY}}[f(X,Y)] \\
&+ \mathbb{E}_{P_X\otimes P_Y}[X^T T(X,Y) - f(T(X,Y),Y)]\}
\end{aligned} \tag{17}$$

Remark 4: Note that we changed the role of source $P_X \otimes P_Y$ and the target P_{XY}, compared to the original formulation (1), so that T represents the transport map from the prior to the posterior, instead of $\nabla_x f$. This formulation leads to a more convenient parameterization of the map.

Similar relaxations to the above have also been found to be beneficial for computing Wasserstein barycenters [19] and Wasserstein gradient flows [20]. Here ICNN denotes the set of partially input convex neural networks [2].

To illustrate the performance of the filters, consider the following dynamics and observation model:

$$X_t = (1-\alpha)X_{t-1} + 2\sigma V_t, \quad X_0 \sim \mathcal{N}(0, I_n) \tag{18a}$$
$$Y_t = h(X_t) + \sigma W_t \tag{18b}$$

for $t = 1, 2, 3, \ldots$, where $X_t, Y_t \in \mathbb{R}^n$, $\{V_t\}_{t=1}^{\infty}$ and $\{W_t\}_{t=1}^{\infty}$ are i.i.d sequences of n-dimensional standard Gaussian random variables, $\alpha = 0.1$ and $\sigma = \sqrt{0.1}$. We use three observation functions:

$$h(x) = x, \quad h(x) = x \odot x, \quad h(x) = x \odot x \odot x$$

where $\odot$ denotes the element-wise (i.e., Hadamard) product when x is a vector.

In order to solve (17), we parameterize ICNN as

$$f(x; y) = \sum_{k=1}^{K} W_k (x^\top W_k^x + y^\top W_k^y + b_k)_+^2$$

where $W_k \geq 0$, $W_k^x, W_k^y \in \mathbb{R}^n$, $b_k \in \mathbb{R}$ for $k = 1, \ldots, K$, and $K = 32$ is the size of the network. The map T is modeled with a standard residual network with two blocks of size 32 and a ReLU-activation function.

We used the ADAM optimizer to solve the min-max problem with learning rate 10^{-2}, inner-loop iteration 10, and the total number of iterations 1024, which is divided by 2 after each time step (of the filtering problem) until it reaches 64. Each iteration involves a random selection of a batch of samples of size 32 from the total of $N = 1000$ particles $\{(X_t^1, Y_t^1), \ldots, (X_t^N, Y_t^N)\}$. Observation samples Y_t^i are produced using the observation model: $Y_t^i \sim h(\cdot|X_t^i)$. Samples from the independent coupling $P_X \otimes P_Y$ are generated by random shuffling. The number of particles N is the same for all algorithms. The details of the numerical code is available online[1].

The numerical results are presented in Figure 1 for a two-dimensional problem $n = 2$, while the figure only shows the first component (We choose $n = 2$ because the SIR and OT approach did not differ significantly when $n = 1$, while the difference became apparent with $n = 2$.) The figure shows the trajectory of the particles along with the trajectory of the hidden state. The first experiment, depicted in panel (a), illustrates the performance of a linear observation function. As expected, all three algorithms behave similarly as all of them are able to capture the exact solution, which is Gaussian in this case, and obtained using linear maps. The second experiment, depicted in panel (b), involves the quadratic observation function $h(x) = x \odot x$. This is an interesting case since the problem is not observable, and we expect to see a (symmetric) bimodal distribution. It is observed that EnKF fails to represent the bimodal distribution while both OT and SIR capture the two modes, although, SIR exhibits mode collapse in the time range of $t \in [2, 3.5]$. Finally, both SIR and OT perform better than EnKF for the cubic observation function $h(x) = x \odot x \odot x$, depicted in panel (c), as expected due to the strong nonlinearity in the observation model.

We also quantify the performance of all algorithms in these three experiments by computing the mean-squared-error (MSE) in estimating a function ϕ of the state:

$$\text{MSE}_t(\phi) = \mathbb{E} \| \frac{1}{N} \sum_{i=1}^{N} \phi(X_t^i) - \phi(X_t) \|^2. \qquad (19)$$

[1] https://github.com/Mohd9485/OT-EnKF-SIR

We use an empirical average over 100 independent simulations to approximate the expectation. The results are depicted in Figure 2. For the linear and cubic observation models, we used $\phi(x) = x$. For the quadratic case, we used $\phi(x) = \max(0, x)$ (comparing the estimated and true means is not a good criterion for the quadratic case because the distribution is symmetric with zero mean).

In Figure 2-(a), it is observed that both OT and SIR filters yield results that are close to the EnKF, which is asymptotically exact for the linear case. However, in Figure 2-(b), the OT method outperforms both EnKF and SIR for the quadratic case, while the difference between OT and SIR is not significant for the cubic case, depicted in Figure 2-(c). The performance of the OT filter is expected to improve with further fine-tuning, increasing the iteration number of training, and the number of parameters in the neural net, at the cost of higher computational effort. An appropriate analysis of the efficiency of the OT method, how it scales to high-dimensional problems, and its application to more realistic data, is the subject of future work.

VI. Discussion

In this paper, we presented the OTPF algorithm and provided preliminary theoretical error analysis and numerical results that demonstrated the competitive performance of our method in the presence of nonlinear observations and non-Gaussian states. We introduced several directions of future research: the verification of the geometric stability for dynamical systems e.g. of the form (18); error analysis of the optimization gap in solving the variational problem, both in terms of representation and generalization as discussed in Remark 3; error analysis of the interacting particle system without resampling; and extensive numerical experiments and comparison in truly high-dimensional settings.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[2] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[3] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

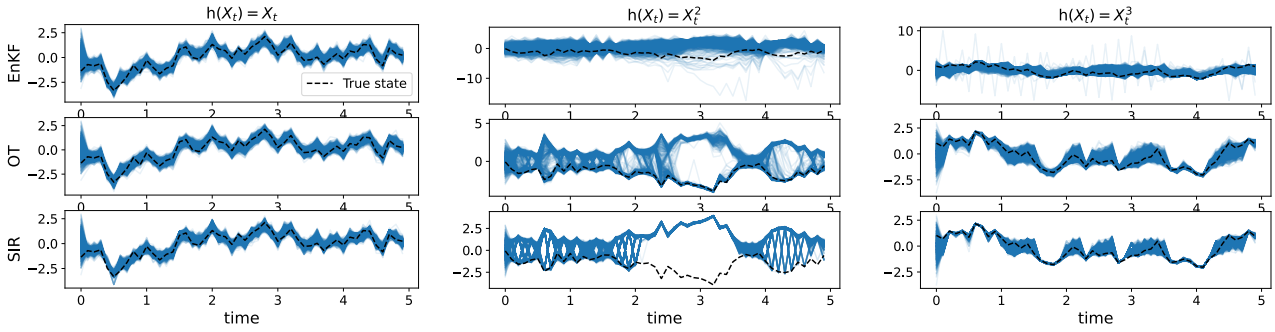
[5] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232. PMLR, 2017.

[6] Rami Atar and Ofer Zeitouni. Exponential stability for nonlinear filtering. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 33, pages 697–725. Elsevier, 1997.

[7] Edoardo Calvello, Sebastian Reich, and Andrew M Stuart. Ensemble Kalman methods: a mean field perspective. *arXiv preprint arXiv:2209.11371*, 2022.

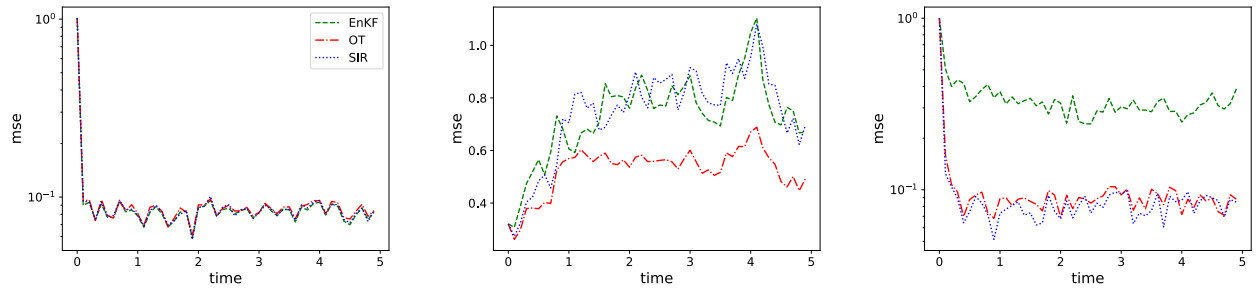
[8] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

[9] Yuan Cheng and Sebastian Reich. A McKean optimal transportation perspective on Feynman-Kac formulae with application to data assimilation. *arXiv preprint arXiv:1311.6300*, 2013.



(a) (b) (c)

Fig. 1. Numerical results for the application of the ensemble Kalman filter, optimal transport particle filter, and sequential importance resampling particle filter, denoted by EnKF, OT, and SIR in the figure, respectively, on the numerical example (18). The figure shows the trajectory of the particles $\{X_t^1, \ldots, X_t^N\}$ along with the trajectory of the true state X_t. The results include three observation functions: (a) $h(x) = x$, (b) $h(x) = x \odot x$, and (c) $h(x) = x \odot x \odot x$.



(a) (b) (c)

Fig. 2. Numerical results for the application of filters to the example (18) in a similar setting as Figure 1. The figure shows the MSE (19) in estimating a function of the state. In panels (a) and (c), the function $\phi(x) = x$ is used, while panel (b) is for $\phi(x) = \max(0, x)$. The MSE is evaluated by taking the empirical average over 100 independent simulations.

[10] Pavel Chigansky, Robert Liptser, and Ramon Van Handel. Intrinsic methods in filter stability. *Handbook of Nonlinear Filtering*, 2009.

[11] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *arXiv preprint arXiv:1507.00504*, 2015.

[12] Dan Crisan and Boris Rozovskii. *The Oxford handbook of nonlinear filtering*. Oxford University Press, 2011.

[13] Pierre Del Moral and Pierre Del Moral. *Feynman-Kac formulae*. Springer, 2004.

[14] Pierre Del Moral and Alice Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. In *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, volume 37, pages 155–194. Elsevier, 2001.

[15] Ayelet Dominitz and Allen Tannenbaum. Texture mapping via optimal mass transport. *IEEE transactions on visualization and computer graphics*, 16(3):419–433, 2010.

[16] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[17] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[18] Geir Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer Science & Business Media, 2006.

[19] Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1571–1581. PMLR, 18–24 Jul 2021.

[20] Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6185–6215. PMLR, 17–23 Jul 2022.

[21] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[22] Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

[23] Jin W Kim and Prashant G Mehta. Duality for nonlinear filtering i: Observability. *IEEE Transactions on Automatic Control*, 2023.

[24] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

[25] Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. Conditional sampling with monotone GANs. *arXiv preprint arXiv:2006.06755*, 2020.

[26] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[28] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

[29] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[30] Daniel Ocone and Etienne Pardoux. Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM Journal on Control and Optimization*, 34(1):226–243, 1996.

[31] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[32] Deep Ray, Harisankar Ramaswamy, Dhruv V Patel, and Assad A Oberai. The efficacy and generalizability of conditional GANs for posterior inference in physics-based inverse problems. *arXiv preprint arXiv:2202.07773*, 2022.

[33] Patrick Rebeschini and Ramon Van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

[34] Sebastian Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[35] Sebastian Reich. Data assimilation: The Schrödinger perspective. *Acta Numerica*, 28:635–711, 2019.

[36] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[37] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[38] Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pages 1792–1802. PMLR, 2022.

[39] Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp A Witte, and Felix J Herrmann. Preconditioned training of normalizing flows for variational inference in inverse problems. *arXiv preprint arXiv:2101.03709*, 2021.

[40] Alessio Spantini, Ricardo Baptista, and Youssef Marzouk. Coupling techniques for nonlinear ensemble filtering. *SIAM Review*, 64(4):921–953, 2022.

[41] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015.

[42] Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, 1464:165–251, 1991.

[43] Amirhossein Taghvaei and Bamdad Hosseini. An optimal transport formulation of Bayes' law for nonlinear filtering algorithms. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6608–6613. IEEE, 2022.

[44] Amirhossein Taghvaei and Prashant G Mehta. An optimal transport formulation of the linear feedback particle filter. In *2016 American Control Conference (ACC)*, pages 3614–3619. IEEE, 2016.

[45] Amirhossein Taghvaei and Prashant G Mehta. An optimal transport formulation of the ensemble Kalman filter. *IEEE Transactions on Automatic Control*, 66(7):3052–3067, 2020.

[46] Amirhossein Taghvaei and Prashant G Mehta. Optimal transportation methods in nonlinear filtering. *IEEE Control Systems Magazine*, 41(4):34–49, 2021.

[47] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[48] Ramon Van Handel. Observability and nonlinear filtering. *Probability theory and related fields*, 145:35–74, 2009.

[49] Ramon Van Handel. Uniform observability of hidden Markov models and filter stability for unstable signals. 2009.

[50] Ramon Van Handel. Nonlinear filtering and systems theory. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS semi-plenary paper)*, 2010.

[51] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[52] Tao Yang, Richard S Laugesen, Prashant G Mehta, and Sean P Meyn. Multivariable feedback particle filter. *Automatica*, 71:10–23, 2016.

[53] Tao Yang, Prashant G Mehta, and Sean P Meyn. Feedback particle filter. *IEEE transactions on Automatic control*, 58(10):2465–2480, 2013.

[54] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. *arXiv preprint arXiv:1711.02771*, 2017.