Nonlinear Filtering with Brenier Optimal Transport Maps

Mohammad Al-Jarrah [1] **Niyizhen Jin** [2] **Bamdad Hosseini** [2] **Amirhossein Taghvaei** [1]

Abstract

This paper is concerned with the problem of nonlinear filtering, i.e., computing the conditional distribution of the state of a stochastic dynamical system given a history of noisy partial observations. Conventional sequential importance resampling (SIR) particle filters suffer from fundamental limitations, in scenarios involving degenerate likelihoods or high-dimensional states, due to the weight degeneracy issue. In this paper, we explore an alternative method, which is based on estimating the Brenier optimal transport (OT) map from the current prior distribution of the state to the posterior distribution at the next time step. Unlike SIR particle filters, the OT formulation does not require the analytical form of the likelihood. Moreover, it allows us to harness the approximation power of neural networks to model complex and multi-modal distributions and employ stochastic optimization algorithms to enhance scalability. Extensive numerical experiments are presented that compare the OT method to the SIR particle filter and the ensemble Kalman filter, evaluating the performance in terms of sample efficiency, high-dimensional scalability, and the ability to capture complex and multi-modal distributions.

1. Introduction

This paper is concerned with numerical methods for the solution of nonlinear filtering problems in discrete-time with a particular focus towards the settings where the distribution of interest is highly non-Gaussian and multi-modal. To formulate the filtering problem consider two stochastic processes: (i) a hidden Markov process, denoted by $\{X_t\}_{t=0}^{\infty}$, that represents the state of a dynamical system; (ii) an observed random process, denoted by $\{Y_t\}_{t=1}^{\infty}$, that represents sensory data. We assume the state and observation processes follow the probabilistic relationship

$$X_t \sim a(\cdot \mid X_{t-1}), \quad X_0 \sim \pi_0, \tag{1a}$$

$$Y_t \sim h(\cdot \mid X_t), \tag{1b}$$

where $a(\cdot \mid \cdot)$ and $h(\cdot \mid \cdot)$ denote the conditional probability kernels of X_{t+1} and Y_t, given X_t, respectively, and π_0 denotes the probability distribution of the initial state X_0. The filtering problem is to compute the conditional probability distribution of the hidden state X_t, given the history of observations $\mathcal{Y}_t = \{Y_1, Y_2, \dots, Y_t\}$, denoted by

$$\pi_t(\cdot) := \mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_t), \quad \text{for} \quad t = 1, 2, \dots \tag{1c}$$

The conditional distribution π_t is also referred to as the *posterior* or the *belief*.

In a general nonlinear and non-Gaussian setup, the posterior does not admit an explicit analytical solution. Therefore, it is necessary to design numerical methods to approximate it. Standard sequential importance re-sampling (SIR) particle filters (PF) approximate the posterior π_t with a weighted empirical distribution of particles $\sum_{i=1}^{N} w_t^i \delta_{X_t^i}$ (Gordon et al., 1993; Doucet & Johansen, 2009) with the particle locations and weights updated according to

$$\text{(propagation)} \quad X_t^i \overset{\text{i.i.d}}{\sim} \sum_{j=1}^{N} w_{t-1}^j a(\cdot \mid X_{t-1}^j), \tag{2a}$$

$$\text{(conditioning)} \quad w_t^i = \frac{h(Y_t \mid X_t^i)}{\sum_{j=1}^{N} h(Y_t \mid X_t^j)}. \tag{2b}$$

The propagation step involves simulation according to the dynamic model (1a), followed by a resampling procedure[1]. The conditioning step implements the Bayes's rule according to the observation model (1b).

SIR PF provides an exact solution to the filtering problem in the limit $N \to \infty$ (Del Moral & Guionnet, 2001). However, it suffers from the *curse of dimensionality (COD)*. Specifically, as the dimension of the state and observation grows, the likelihood becomes degenerate and concentrated on a small support, causing the majority of weights to become

[1]Department of Aeronautics & Astronautics, University of Washington, Seattle. [2]Department of Applied Mathematics, University of Washington, Seattle.. Correspondence to: Mohammad Al-Jarrah <mohd9485@uw.edu>, Niyizhen Jin <njin2@uw.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]In practice, resampling is usually done only when the variance of weights becomes large.

zero. This issue is known as *particle degeneracy* and can only be avoided when the number of particles grows exponentially with the dimension (Bickel et al., 2008; Bengtsson et al., 2008; Rebeschini & Van Handel, 2015); This is an active research area; see (Van Leeuwen et al., 2019) and references therein.

This curse of dimensionality has motivated the development of alternative algorithms that replace the conditioning step of SIR with a coupling or transport step; see (Daum et al., 2010; Crisan & Xiong, 2010; Reich, 2011; Yang et al., 2011; El Moselhy & Marzouk, 2012; Reich, 2013; De Melo et al., 2015; Yang et al., 2016; Marzouk et al., 2016; Mesa et al., 2019; Reich, 2019; Pathiraja et al., 2021; Taghvaei & Mehta, 2021; Calvello et al., 2022) as well as (Spantini et al., 2022) and (Taghvaei & Mehta, 2023) for a recent survey of these topics. A more detailed comparison is provided in Sec. 4. Here the posterior is approximated with a uniformly weighted distribution of particles $\frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$, with a general particle update law of the form

$$\text{(propagation)} \quad X_{t|t-1}^i \sim a(\cdot \mid X_{t-1}^i), \tag{3a}$$

$$\text{(conditioning)} \quad X_t^i = T_t(X_{t|t-1}^i, Y_t). \tag{3b}$$

Here, $T_t(\cdot, Y_t)$ represents a (possibly stochastic) transport map from the (prior) distribution $\mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_{t-1})$ to the posterior distribution $\mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_t)$.

This paper is concerned with the development of algorithms for approximating the above transport map. In particular, we utilize the recently introduced optimal transport (OT) approach in (Taghvaei & Hosseini, 2022; Al-Jarrah et al., 2023) to formulate the problem of finding the map T_t as a max-min stochastic optimization problem (see equations (11) and (10)) that only involves the particles $X_{t|t-1}^i$ and the associated observation $Y_{t|t-1}^i \sim h(\cdot \mid X_{t|t-1}^i)$, for $i = 1, \ldots, N$. The map T is represented by residual neural networks, as in Fig. 2, and trained with stochastic optimization algorithms (see Algorithm 3 for details). The OT approach is based on the combination of block-triangular transport, in the context of conditional simulation (Kovachki et al., 2023; Ray et al., 2022; Shi et al., 2022; Siahkoohi et al., 2021), with min-max formulations used to estimate OT maps as in (Makkuva et al., 2020; Fan et al., 2021; Rout et al., 2022). The OT methodology has two distinct features: (i) it can be implemented in a completely likelihood-free/simulation-based setting, because it only requires samples from the likelihood model, not its analytical form; (ii) and it allows for the application of neural networks to represent transport maps, that enables the capture of non-Gaussian and multi-modal distributions, as well as stochastic optimization algorithms that are scalable with the number of the particles. An illustrative example comparing the OT method with SIR particle filter is presented in Fig. 1. The contributions of the paper are twofold: (1) We present

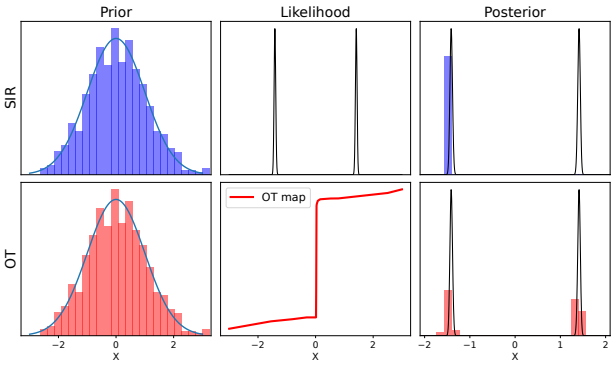


Figure 1: A comparison between the OT method and SIR for the static model of Sec. 3.1. The OT approach captures the bimodal posterior by pushing the prior through an OT map computed by solving equation (11). On the other hand, SIR only captures one mode due to the degeneracy of the likelihood leading to only a few weights that are order 1.

a theoretical analysis of the OT method focusing on consistency and stability results; these are outlined in Prop 2.3 and 2.4. We also provide a preliminary error analysis in Prop. 2.7 that compares the OT approach to SIR for a single conditioning step of the filter; (2) We present an extensive suite of numerical experiments in Sec. 3, containing toy examples, benchmark Lorenz models, and a high-dimensional filtering problem where image in-painting is performed in a dynamic manner. We evaluate the OT method, in comparison to the SIR and ensemble Kalman filter (EnKF) algorithms, in terms of sample efficiency and high-dimensional scalability in conjunction with the ability to capture complex and multi-modal distributions.

2. OT methodology

2.1. Exact filter

Let us begin by recalling the exact filter before moving on to transport maps and the OT formulation. Henceforth we assume $X_t \in \mathbb{R}^n$ and $Y_t \in \mathbb{R}^m$. The posterior distribution (1c) admits a recursive update law that is essential for the design of filtering algorithms. To present this recursive update, we introduce the following operators:

$$\text{(prop.)} \quad \pi \mapsto \mathcal{A}\pi := \int_{\mathbb{R}^n} a(\cdot \mid x)\pi(x)dx \tag{4a}$$

$$\text{(cond.)} \quad \pi \mapsto \mathcal{B}_y(\pi) := \frac{h(y \mid \cdot)\pi(\cdot)}{\int_{\mathbb{R}^n} h(y \mid x)\pi(x)dx} \tag{4b}$$

The first operator represents the update for the distribution of the state according to the dynamic model (1a). The second operator represents Bayes' rule that carries out the conditioning according to the observation model (1b). The exact posterior distribution $\pi_t := \mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_t)$ then

follows the sequential update law (Cappé et al., 2009)

$$\pi_{t|t-1} = \mathcal{A}\pi_{t-1}, \quad \pi_t = \mathcal{B}_{Y_t}(\pi_{t|t-1}), \tag{4c}$$

where the notation $\pi_{t|t-1}(\cdot) := \mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_{t-1})$ is used for the distribution of X_t before applying the conditioning with respect to the current observation Y_t.

2.2. Conditioning with transport maps

Transport or coupling-based approaches approximate the posterior π_t with an empirical distribution of a set of particles $\{X_t^i\}_{i=1}^N$, where the particles are simulated according to the equations (3a) and (3b). Although not explicitly stated, the transport map T_t, that is used in the conditioning step (3b), depends on the empirical distribution of the particles. As a result, the equation (3) represents an *interacting particle system*.

The problem of designing the transport map T_t is studied in the *mean-field* limit ($N = \infty$), which assumes that all particles are independent and identically distributed according to a single mean-field distribution, denoted by $\overline{\pi}_t$. Under this assumption, the update law for $\overline{\pi}_t$ may be expressed as

$$\overline{\pi}_{t|t-1} = \mathcal{A}\overline{\pi}_{t-1}, \quad \overline{\pi}_t = T_t(\cdot, Y_t)_{\#}\overline{\pi}_{t|t-1}, \tag{5}$$

where $\#$ denotes the push-forward operator. The transport map T_t should be designed such that $\overline{\pi}_t = \pi_t$ for all $t \geq 0$. Comparing equation (5) with (4c) concludes the following general design problem:

Problem: For all probability distributions π, find a map T such that

$$T(\cdot, y)_{\#}\pi = \mathcal{B}_y(\pi), \quad \forall y. \tag{6}$$

We call this procedure *conditioning with transport maps*. The following examples are illustrative.

Example 2.1 (Noiseless observation). Assume $X \sim \pi$ and $Y = h(X)$ where h is an invertible map. Then, the conditional distribution $\mathcal{B}_y(\pi) = \delta_{h^{-1}(y)}$ can simply be represented via a transport map $T(x, y) = h^{-1}(y)$.

Example 2.2 (Gaussian). Assume $X \sim \pi$ is Gaussian and the observation $Y \sim h(\cdot \mid X)$ corresponds to a linear observation model with additive Gaussian noise, meaning X and Y are jointly Gaussian. Then, the conditional distribution $\mathcal{B}_y(\pi)$ is Gaussian $N(\mu, \Sigma)$

$$\mu = \mathbb{E}[X] + K(y - \mathbb{E}[Y]),$$
$$\Sigma = \mathrm{Cov}(X) - \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}\mathrm{Cov}(Y, X),$$

and $K = \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}$. The stochastic map

$$X \mapsto X + K(y - Y), \tag{7}$$

transports π to $\mathcal{B}_y(\pi)$ for any value of the observation y. This can be checked by noting that $X + K(y - Y)$ is a Gaussian random variable with the same mean and covariance as $N(\mu, \Sigma)$. The map in equation (7) is the basis for the ensemble Kalman filter (EnKF) algorithm; see (Evensen, 2009; Bergemann & Reich, 2012; Calvello et al., 2022).

The consistency condition (6) does not specify a unique map T. In the following subsection, we formulate the problem of finding a valid map T as an OT problem alongside an stochastic optimization approach on which our numerical procedure is based on.

2.3. OT formulation

In order to present the OT formulation, consider $X \sim \pi$, $Y \sim h(\cdot \mid X)$, and let $\overline{X} \sim \pi$ be an independent copy of X. Also, let $P_{X,Y}$ denote the joint distribution of (X, Y), with marginals P_X and P_Y. First observe that the condition (6), which may be expressed as $T(\cdot, y)_{\#}P_X = P_{X|Y}(\cdot|y)$ a.e., is equivalent to

$$(T(\overline{X}, Y), Y) \sim P_{X,Y}. \tag{8a}$$

A justification for this result appears in Appendix B.1 and (Kovachki et al., 2023, Thm. 2.4). In order to select a unique map that satisfies the condition (8a), we formulate the (conditional) Monge problem:

$$\min_{T \in \mathcal{M}(P_X \otimes P_Y)} \mathbb{E}\left[c(T(\overline{X}, Y), \overline{X})\right], \quad \text{s.t. (8a) holds.} \tag{8b}$$

where $c(x, x') = \frac{1}{2}\|x - x'\|^2$ and $\mathcal{M}(P_X \otimes P_Y)$ is the set of maps $\mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^n$ that are $P_X \otimes P_Y$-measurable. The optimization (8b) is viewed as the Monge problem between the independent coupling $(\overline{X}, Y) \sim P_X \otimes P_Y$ and the joint distribution $(X, Y) \sim P_{X,Y}$ with transport maps that are constrained to be block-triangular $(x, y) \mapsto (T(x, y), y)$. Upon using the Kantorovich duality and the definition of c-concave functions[2] the Monge problem (8b) becomes

$$\max_{f \in c\text{-Concave}_x} \min_{T \in \mathcal{M}(P_X \otimes P_Y)} J(f, T), \tag{8c}$$

where the objective function $J(f, T)$ is equal to

$$\mathbb{E}\left[f(X, Y) - f(T(\overline{X}, Y), Y) + c(T(\overline{X}, Y), \overline{X})\right],$$

and the set c-Concave$_x$ denotes the set of functions on $\mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$ that are c-concave in their first variable everywhere. A rigorous justification of the max-min formulation, in the standard OT setting, appears in Appendix A.3 and (Makkuva et al., 2020; Rout et al., 2022). The following result is the extension to the conditional setting.

Proposition 2.3. *Assume π is absolutely continuous with respect to the Lebesgue measure with a convex support set $\mathcal{X}$, $\mathcal{B}_y(\pi)$ admits a density with respect to the Lebesgue measure $\forall y$, and $\mathbb{E}[\|X\|^2] < \infty$. Then, there exists a unique*

[2] with a squared loss c, f is c-concave iff $\frac{1}{2}\|\cdot\|^2 - f$ is convex.

pair $(\overline{f}, \overline{T})$, modulo an additive constant for $\overline{f}$, that solves the optimization problem (8c) and the map $\overline{T}(\cdot, y)$ is the OT map from π to $\mathcal{B}_y(\pi)$ for a.e. y.

Proof sketch. The proof is a direct consequence of applying Theorem 2.3 (iii) in (Carlier et al., 2016) and Porp. A.3 for the max-min formulation in the standard OT setting; see Appendix B.2 for details. □

The next proposition is concerned with the case that the max-min optimization problem (8c) is not solved exactly and provides an error bound for the computed OT maps in terms of the pertinent optimality gap. More precisely, for a pair (f, T), let

$$\begin{aligned}\epsilon(f,T) :=& J(f,T) - \min_S J(f,S) \\ &+ \max_g \min_S J(g,S) - \min_S J(f,S)\end{aligned} \tag{9}$$

be the total optimality gap for the max-min problem. We then have the following result which can be viewed as an extension of (Rout et al., 2022, Thm. 4.3) and (Makkuva et al., 2020, Thm. 3.6) by adding the observation variable Y. The proof appears in Appendix B.3.

Proposition 2.4. *Consider the setting of Prop. 2.3 with the optimal pair $(\overline{f}, \overline{T})$. Let (f, T) be a possibly non-optimal pair with an optimality gap $\epsilon(f, T)$. Assume $x \mapsto \frac{1}{2}\|x\|^2 - f(x, y)$ is α-strongly convex in x for all y. Then,*

$$\mathbb{E}\left[\|T(\overline{X}, Y) - \overline{T}(\overline{X}, Y)\|^2\right] \leq \frac{4}{\alpha}\epsilon(f, T).$$

2.4. Empirical approximation

In order to numerically solve the optimization problem (8c), the objective function $J(f, T)$ is approximated empirically in terms of samples according to

$$\begin{aligned}J^{(N)}(f,T) := \frac{1}{N}\sum_{i=1}^{N}\big[f(X^i, Y^i) \quad + \\ \frac{1}{2}\|T(\overline{X}^i, Y^i) - X^i\|^2 - f(T(\overline{X}^i, Y^i), Y^i)\big]\end{aligned} \tag{10}$$

where $X^i, \overline{X}^i \overset{\text{i.i.d}}{\sim} \pi$ and $Y^i \sim h(\cdot|X^i)$. In the context of filtering problem, $\{X^i\}_{i=1}^N$ is constructed from the particles $\{X^i_{t|t-1}\}_{i=1}^N$ and $\{\overline{X}^i\}_{i=1}^N$ may be constructed by an independent random shuffling of the same set, resulting in the filtering algorithm presented in Algorithm 3 in Appendix C. The function f and the map T are represented with a parametric class of functions, denoted by $\mathcal{F}$ and $\mathcal{T}$ respectively. This concludes the optimization problem

$$\max_{f\in\mathcal{F}} \min_{T\in\mathcal{T}} J^{(N)}(f,T). \tag{11}$$

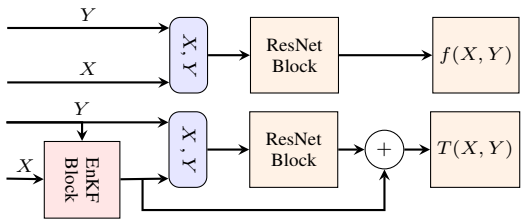


Figure 2: Neural net architectures for the function classes $\mathcal{F}$ and $\mathcal{T}$ within our proposed algorithm.

Here we propose to take both of these spaces to be neural network classes with architectures that are summarized in Fig. 2; further details about these architectures can be found in the Appendix C.

Remark 2.5. Note that our choice of $\mathcal{F}$ does not impose the constraint that $f(x, y)$ is c-Concave$_x$. We make this choice due to practical limitations of imposing convexity constraints on neural nets using, for example, input-convex networks (Amos et al., 2017; Bunne et al., 2022). However, note that if the computed f happens to be c-Concave$_x$ (which one can check a posteriori) then Prop. 2.4 remains applicable. We provide qualitative and quantitative results, in Sec. 3.1, that measure the convexity of $\frac{1}{2}\|x\|^2 - f(x, y)$ and monotonicity of the map T for a particular example.

Remark 2.6. The proposed computational procedure may be extended to the Riemannian manifold setting by using the square of the geodesic distance as the cost function c and modelling the map T as exponential of a parameterized vector-field; see (Grange et al., 2023).

2.5. Comparison with SIR

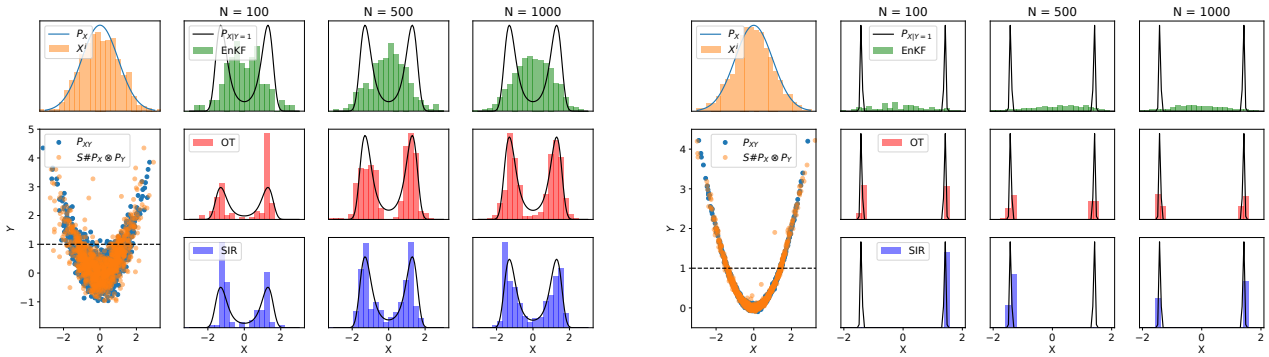
In order to compare the OT and the SIR approaches, we use the *dual bounded-Lipschitz metric* $d(\mu, \nu) := \sup_{g\in\mathcal{G}} \sqrt{\mathbb{E}\left|\int g d\mu - \int g d\nu\right|^2}$ on (possibly random) probability measures μ, ν, where $\mathcal{G}$ is the space of functions that are bounded by one and Lipschitz with a constant smaller than one.

For any probability distribution π, the SIR and OT approaches approximate the conditional distribution $\mathcal{B}_y(\pi)$ in two different ways. The SIR approach approximates the posterior with

$$\pi^{(\text{SIR})} := \frac{\sum_{i=1}^N h(y|X^i)\delta_{X^i}}{\sum_{i=1}^N h(y|X^i)} = \mathcal{B}_y(S^{(N)}(\pi)) \tag{12a}$$

where $X^i \overset{\text{i.i.d.}}{\sim} \pi$ and $\pi \mapsto S^{(N)}\pi := \frac{1}{N}\sum_{i=1}^N \delta_{X^i}$ is the sampling operator. On the other hand, the OT approach approximates the posterior by numerically solving the optimization problem (11). Assuming the pair $(\widehat{f}, \widehat{T})$ is a solution of this problem, the approximated posterior is

$$\pi^{(OT)} := \widehat{T}(\cdot, y)_{\#}\pi. \tag{12b}$$



(a) $\lambda_w = 0.4$. (b) $\lambda_w = 0.04$.

Figure 3: Numerical results for the static example in Sec. 3.1. (a) top-left: Samples $\{X^i\}_{i=1}^N$ from the prior P_X; bottom-left: samples $\{(X^i, Y^i)\}_{i=1}^N$ from the joint distribution P_{XY} in comparison with the transported samples $\{(T(X^{\sigma_i}, Y^i), Y^i)\}_{i=1}^N$; rest of the panels: transported samples for $Y = 1$ for different values of N and three different algorithms. (b) Similar results to panel (a) but for a smaller λ_w.

The following proposition provides an analysis of the approximation errors.

Proposition 2.7. *Consider $X \sim \pi$, $Y \sim h(\cdot|X)$, and the SIR and OT approaches explained above in order to approximate the posterior $\mathcal{B}_Y(\pi)$. Then,*

$$\liminf_{N\to\infty} \sqrt{N} d(\pi^{(SIR)}, \mathcal{B}_Y(\pi)) \geq \sup_{g\in\mathcal{G}} \sqrt{V_h(g)}, \tag{13a}$$

$$d(\pi^{(OT)}, \mathcal{B}_Y(\pi)) \leq \sqrt{\frac{4}{\alpha}\epsilon(\widehat{f}, \widehat{T})}, \tag{13b}$$

where $V_h(g) := \mathbb{E}[\overline{h}(Y|\overline{X})^2(g(\overline{X}) - \mathbb{E}[g(X)|Y])^2]$, $\overline{h}(x|y) := h(x|y)/\int h(x'|y)d\pi(x')$, $\overline{X}$ is an independent copy of X, α is the strong convexity constant for $x \mapsto \frac{1}{2}\|x\|^2 - \widehat{f}(x, y)$, and $\epsilon(\widehat{f}, \widehat{T})$ is the optimality gap (9) for the pair $(\widehat{f}, \widehat{T})$.

Proof sketch. The lower-bound (13a) follows from an application of the central limit results available for importance sampling in (Cappé et al., 2009, Thm. 9.1.8). The upper-bound (13a) follows from Prop. 2.4 and the definition of the metric d; see Appendix B.4 for details. □

The COD in SIR PF is demonstrated by considering the special case where $X = [X(1), \ldots, X(n)]$ and $Y = [Y(1), \ldots, Y(n)]$ are n-dimensional with independent and identically distributed components. Then, for N large enough, the SIR error satisfies the bound

$$d(\pi^{(\text{SIR})}, \mathcal{B}_Y(\pi)) \geq \frac{C\gamma^n}{\sqrt{N}}$$

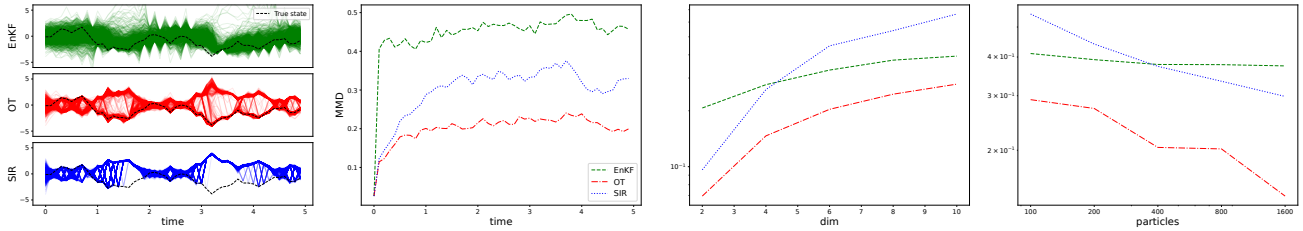
where C is a constant and $\gamma := \mathbb{E}[\overline{h}(Y(1)|\overline{X}(1))^2] > 1$. The COD is usually avoided by exploiting problem specific properties, such as spatial correlations or decay; see (Rebeschini & Van Handel, 2015; Van Leeuwen et al., 2019).

The OT upper-bound (13b) depends on the optimality gap $\epsilon(\widehat{f}, \widehat{T})$ which, in principle, decomposes to a bias and variance term. The bias term corresponds to the representation power of the function classes $\mathcal{F}$ and $\mathcal{T}$, in comparison with the complexity of the problem. The variance term corresponds to the statistical generalization errors due to the empirical approximation of the objective function. The variance term is expected to grow as $O(\frac{1}{\sqrt{N}})$ with a proportionality constant that depends on the complexity of the function classes, but independent of the dimension. In principle, the OT approach may also suffer from the COD under no additional assumptions on the problem. However, in comparison to SIR, it provides a more flexible design methodology that can exploit problem specific structure and regularity.

3. Numerical results

We perform several numerical experiments to study the performance of the OT approach in comparison with the EnKF and SIR algorithms. The OT algorithm 3 consists of solving (3) and (11) with f, T taken to be neural nets whose architectures were outlined in Fig. 2. The network weights are learned with a gradient ascent-descent procedure using the Adam optimization algorithm. To reduce the computational cost, the optimization iteration number decreases as the time grows because the OT map is not expected to change significantly from a time step to the next one. The computational cost and details of all three algorithms and additional numerical results appear in the Appendix C. The Python code for reproducing the numerical results is available online[3].

[3] https://github.com/Mohd9485/Filtering-with-Optimal-Transport



(a) Particles trajectory. (b) MMD vs time. (c) MMD vs the dimension. (d) MMD vs # of particles.

Figure 4: Numerical results for the dynamic example 15. The left panel shows the trajectory of the particles $\{X_t^1, \dots, X_t^N\}$ along with the trajectory of the true state X_t for EnKF, OT, and SIR algorithms, respectively. The second panel shows the MMD distance with respect to the exact conditional distribution. The last two panels show MMD variation with dimension and the number of particles.

3.1. A bimodal static example

Our first numerical experiment serves as a proof of concept and demonstrates the sample efficiency of the OT approach when the likelihood becomes degenerate. Specifically, we consider the task of computing the conditional distribution of a Gaussian hidden random variable $X \sim N(0, I_n)$ given the observation

$$Y = \frac{1}{2} X \odot X + \lambda_w W, \qquad W \sim N(0, I_n) \tag{14}$$

where $\odot$ denotes the element-wise (i.e., Hadamard) product. This model is specifically selected to produce a bimodal posterior. We only present the $n = 2$ case since the difference between OT and SIR was not significant when $n = 1$.

The first numerical results for this model are presented with a noise standard deviation of $\lambda_w = 0.4$ in Fig. 3a. The top left panel shows the initial particles as samples from the Gaussian prior distribution. The bottom left panel shows the pushforward of samples from $P_X \otimes P_Y$ via the block triangular map $S(x, y) = (T(x, y), y)$, in comparison to samples from P_{XY}, verifying the consistency condition (8a) for the map. Then, we pick a particular value for the observation $Y = 1$ (as shown by the dashed line) and present the histogram of (transported) particles in comparison with the exact conditional density. It is observed that both OT and SIR capture the bimodal posterior, while EnKF falls short since it always approximates the posterior with a Gaussian.

We repeat the procedure in Fig. 3b but for a smaller noise standard deviation $\lambda_w = 0.04$ which leads to a more degenerate posterior. Our results clearly demonstrate the weight degeneracy of SIR even in this low-dimensional setting, as all particles collapse into a single mode, while the OT approach still captures the bimodal posterior. Additional numerical results for this model are available in Appendix C.2.

Additionally, we present the learned function $\frac{1}{2}\|x\|_2^2 - \widehat{f}(x, y)$ and the learned map $\widehat{T}(x, y)$, as a function of x and for a $1D$ case with fixed value $y = 1$, in Fig. 5. It is observed that the function $\frac{1}{2}\|x\|_2^2 - \widehat{f}(x, 1)$ is not entirely convex, specially around the region with small posterior probability. Meanwhile, the transport map $\widehat{T}(x, 1)$ appears to be entirely monotone. In order to quantify the degree of monotonicity of $x \mapsto \widehat{T}(x, y)$, for all values of y, we employ a Monte-Carlo method with 10^5 samples to approximate

$$\mathbb{E}_{(X,X',Y)\sim P_X \otimes P_X \otimes P_Y}[\mathbf{1}_{\varrho(\widehat{T}(\cdot,Y),X,X')\leq 0}] \approx 0.0012,$$
$$\mathbb{E}_{(X,X',Y)\sim P_X \otimes P_X \otimes P_Y}[\varrho(\widehat{T}(\cdot, Y), X, X')] \approx 0.88$$

where $\varrho(S, x, x') := \frac{(S(x)-S(x')^\top(x-x')}{\|x-x'\|^2}$ for any map $S : \mathbb{R}^n \to \mathbb{R}^n$ and two points $x, x' \in \mathbb{R}^n$. If the map S is monotone, then $\varrho(S, x, x') \geq 0$ for all $x, x' \in \mathbb{R}^n$. If S is the gradient of a α-strongly convex function, then, $\rho(S, x, x') \geq \alpha$. The first number shows that the map is not monotone over a set of small probability, while the second number shows that the "degree" of monotinicity is positive on average. We perform a similar procedure to quantify the convexity of the function $\frac{1}{2}\|x\|_2^2 - \widehat{f}(x, y)$ by evaluating the monotonicity of its gradient: $F(x, y) := x - \nabla_x \widehat{f}(x, y)$[4]. With a similar Monte-Carlo method we obtained the respective quantities 0.033 and 1.57, indicating that $\frac{1}{2}\|x\|_2^2 - \widehat{f}(x, y)$ is non-convex over a set of small probability measure under the posterior. This empirical observation motivates future direction of theoretical research concerning the relaxation of the convexity assumption in Prop. 2.4 to convexity over sets of large measure under the posterior.

3.2. A bimodal dynamic example

We consider a dynamic version of the previous example according to the following model:

$$X_t = (1 - \alpha)X_{t-1} + 2\lambda V_t, \quad X_0 \sim \mathcal{N}(0, I_n) \tag{15a}$$
$$Y_t = X_t \odot X_t + \lambda W_t, \tag{15b}$$

[4]Checking the positive-definiteness of the Hessian matrix is not applicable in this case because $\widehat{f}(x, y)$ is not differentiable when we use ReLU activation functions.

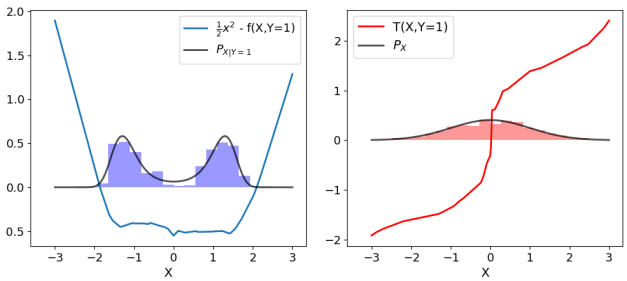


Figure 5: Numerical results for the bimodal static example in sec. (3.1). The left panel shows the function $\frac{1}{2}x^2 - \widehat{f}(x, 1)$ and the conditional distribution $P_{X|Y=1}$. The right panel shows the map $\widehat{T}(x, 1)$ and the prior distribution P_X.

where $\{V_t, W_t\}_{t=1}^{\infty}$ are i.i.d sequences of standard Gaussian random variables, $\alpha = 0.1$ and $\lambda = \sqrt{0.1}$. The choice of Y_t will once again lead to a bimodal posterior π_t at every time step.

The numerical results are depicted in Fig. 4: Panel (a) shows the trajectory of the particles for the three algorithms, along with the true state X_t denoted with a dashed black line. The OT approach produces a bimodal distribution of particles, while the EnKF gives a Gaussian approximation and the SIR approach exhibits the weight collapse and misses a mode for the time duration $t \in [1, 2] \cup [3, 4.5]$. Panel (b) presents a quantitative error analysis comparing the maximum-mean-discrepancy (MMD) between the particle distribution of each algorithm and the exact posterior; details such as the choice of the MMD kernel are presented in Appendix C.3. Since the exact posterior is not explicitly available it is approximated by simulating the SIR algorithm with $N = 10^5$ particles. This quantitative result affirms the qualitative observations of panel (a) that the OT posterior better captures the true posterior in time.

We also performed a numerical experiment to study the effect of the dimension n and the number of particles N on the performance of the three algorithms. The results are depicted in panels (c) and (d), respectively. It is observed that both EnKF and OT scale better with dimension compared to SIR. However, as the number of particles increases, the EnKF error remains constant, due to its Gaussian bias, while the approximation error for SIR and OT decreases.

3.3. The Lorenz 63 model

We present numerical results on the three-dimensional Lorenz 63 model which often serves as a benchmark for nonlinear filtering algorithms. The model is presented in Appendix C.4. The state X_t is 3-dimensional while the observation Y_t is 2-dimensional and consists of noisy measurements of the first and third components of the state.

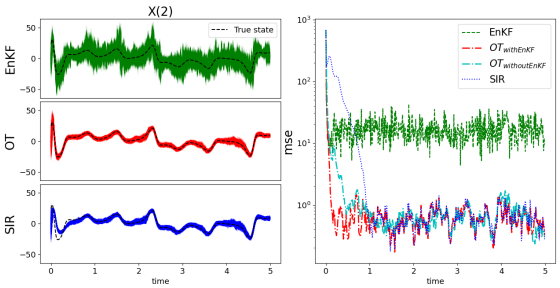


Figure 6: Numerical results for the Lorenz 63 example. The left panel shows the trajectory of the unobserved component of the true state and the particles. The right panel shows the MSE comparison.

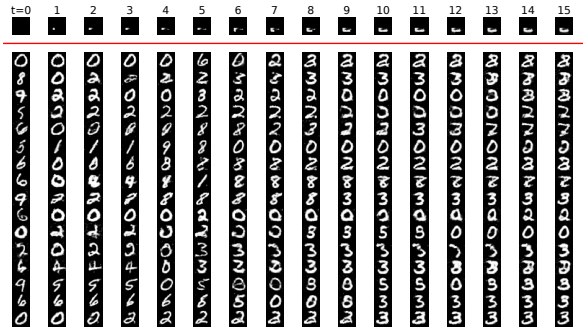
Figure 7: Numerical result for the static MNIST example. The left column shows the true image, the observed part, and the observation window (red square). The next three columns show 16 random particles from the final time-step of each algorithm.

The numerical results are presented in Fig. 6. The left panel shows the trajectory of the second component of the true state and the particles. The OT and EnKF are quicker in converging to the true state, with EnKF admitting larger variance. The right panel shows the mean-squared-error (MSE) in estimating the state confirming the qualitative observations. We present the MSE result for two variations of the OT method: either the EnKF layer in the architecture of Fig. 2 is implemented or not. The results show that the addition of the EnKF layer helps with the performance of the filter, while computationally, we observed more numerical stability when the EnKF layer is removed.

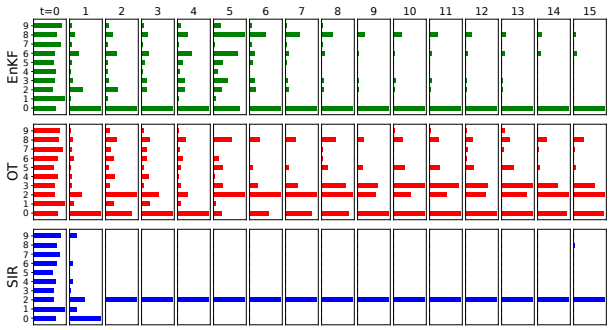
We have also performed numerical experiments on the Lorentz 96 model, another benchmark from the filtering literature, that appear in Appendix C.5. The results for that example are in line with Lorentz 63, i.e., EnKF and OT outperform SIR but are somewhat similar to each other.

3.4. Static image in-painting on MNIST

In order to evaluate the high-dimensional scalability of the OT approach, we consider the problem of computing conditional distributions on the 100-dimensional latent space of generative adversarial network (GAN) trained to represent the MNIST digits (Goodfellow et al., 2014). In particular,



(a) 16 particles from the OT method.



(b) Digits distribution.

Figure 8: Additional numerical results for the static MNIST example. (a) The first row shows the cumulative total observations up to each time step. The subsequent rows under the red line show 16 particles from the OT method (that correspond to the same 16 particles in Fig. 7; (b) The histogram of the the digits generated by the particles, from the three algorithms as a function of time, evaluated using an accurate MNIST classifier.

denoting the generator by $G : \mathbb{R}^{100} \to \mathbb{R}^{28\times 28}$, we consider the model:

$$Y_t = h(G(X), c_t) + \sigma W_t, \quad X \sim N(0, I_{100}),$$

where the observation function $(z, c) \in \mathbb{R}^{28\times 28} \times \mathbb{R}^2 \mapsto h(z, c) \in \mathbb{R}^{r\times r}$ is defined as the $r \times r$ window of pixels $z[c(1) : c(1) + r, c(2) : c(2) + r]$. The coordinates of the corner c_t moves from left to right and top to bottom scanning a partial portion of the image called the *observed part*. In order to make the problem more challenging, we fed a noisy measurement of the corner location to the algorithms. While the true image does not change, we included artificial Gaussian noise to the particles to avoid particle collapse. For for details, see Appendix C.6.

The numerical results are presented in Fig. 7. The left column presents the true image, the observed part of the image, and an instance of the 3×3 observation window with a red square. The filtering algorithms are implemented for 15 time steps, as the observation window moves, and 16 random particles from the final time-step of each algorithm are depicted in the rest of panels in Fig. 7. The results show that although the SIR particles are similar to the true image, they do not capture the true uncertainty corresponding to the observation, in contrast to the OT approach (the lower part of the digits $0, 3, 8$ are similar to the lower part of the digit 2).

In order to further demonstrate the performance of the OT approach, we present the trajectory of the particles in Fig. 8a. The top row shows the total observed part of the image up to that time-step, and the following 16 rows show the images generated from the particles that approximate the conditional distribution. Moreover, we used an accurate MNIST classifier to represent the histogram of the digits generated from the particles of each algorithm in Fig. (8b). According to these results, the SIR and EnKF methods are *overconfident* that the true digit is 2 and 0, respectively, while the OT method produces a more realistic and multi-modal posterior concentrated around the digits $\{0, 2, 3, 8\}$, all of which are natural in-paintings for the data.

3.5. Dynamic image in-painting on MNIST

We extend the previous example to the dynamic setting by introducing an update law for the latent variable of the GAN according to $X_{t+1} = (1 - \alpha)X_t + \lambda V_t$, where V_t is a standard Gaussian random variable. We consider the same observation model but we select a larger window size $r = 12$ and constrain the the vertical coordinate of the window to be at the bottom, while the horizontal coordinate is randomly selected.

A visual representation of the results is provided in Fig. 9. The top two rows display the true images and their observed portion at each time instant $t = 1, 2, \ldots, 20$. Subsequent rows show the images corresponding to four randomly selected particles from the three algorithms; a larger figure depicting more particles is provided in Appendix C.7. The results show the particle collapse of the SIR algorithm at all time steps, convergence of the EnKF algorithm to an incorrect distribution, and the capability of the OT approach in approximating the uncertainty associated with the observation at each time step. A figure similar to the Fig. 8b that shows the histogram of digits as a function of time is also provided in Appendix C.7 which further highlights these phenomenon.

4. Comparison with coupling-based methods

In this section, we provide a comprehensive comparison to other coupling-based methods for filtering. The particle flow method (Daum & Huang, 2012; De Melo et al., 2015) and feedback particle filter (FPF) (Yang et al., 2013; 2016)

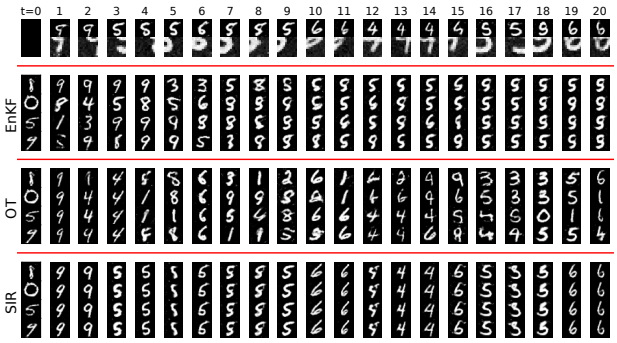


Figure 9: Numerical results for the dynamic MNIST example. The top two rows show the true image and the observations at each time step. The rest of the rows show the first 4 particles for each method, representing the posterior at that time step.

involve either an ordinary differential equation or stochastic differential equation that updates the locations of the particles so that the probability density of the particles follows a given PDE. The particles' equation involves an unknown vector-field that needs to be approximated by solving a certain partial differential equation (PDE). The main challenge in this type of algorithms is to approximate the aforementioned vector-field at each time-step. The time discretization for this type of equation often becomes unstable, especially for multi-modal posteriors or degenerate likelihoods. Our OT approach can be viewed as an exact time-discretization of the FPF algorithm, as shown in Prop. 2 in (Taghvaei & Hosseini, 2022), which resolves the time-discretization issues discussed above.

The ensemble transform particle filter (Reich, 2011) involves solving a linear program for the discrete OT problem from a uniform prior distribution to the weighted posterior distribution for the particular value of the observation. Solving the linear program becomes challenging as the number of particles N increases. Moreover, approximating the marginal with a weighted empirical distribution suffers from the same fundamental issue that importance sampling particle filter suffers from.

The coupling method proposed in (Spantini et al., 2022) is the closest method to our approach. It is also likelihood free and amenable to the neural net parameterizations. The main difference is in the form of the transport map. While in this paper we aim at finding the OT map from prior to the posterior, the approach in (Spantini et al., 2022) aims at finding the Knothe–Rosenblatt rearrangement. Thus our approach is more closely related to the semi-dual solutions to the OT problem and can utilize the existing theoretical results and computational methodologies.

Our work is based on the variational formulation of the Bayes' law presented in (Taghvaei & Hosseini, 2022) (TH). TH proposes $\min_f \max_g \mathbb{E}[f(X,Y) - f(\nabla_x g(\bar{X},Y),Y) + \bar{X}^\top \nabla_x g(\bar{X},Y)]$ where f and g are ICNNs. Our initial numerical study using this method did not provide compelling results due to the limitations of the ICNN architecture (please see Remark 2.5). To resolve this issue, we made two changes: we represent $f(x,y)$ as $\frac{1}{2}\|x\|^2 - \tilde{f}(x,y)$ and $\nabla_x g(x,y)$ as $T(x,y)$, where T and $\tilde{f}$ are ResNets. We also include an EnKF block in the architecture for T so that our method performs as well as EnKF if no training is performed. With these improvements, we were able to produce numerical results on toy, benchmark, and high-dimensional problems that are consistent and perform better than conventional baselines. The clear computational improvement can be observed by comparing Fig. 1 in the TH paper with Fig. 3 and Fig. 11 in this paper.

5. Concluding remarks and limitations

We presented theoretical and numerical results demonstrating the performance of an OT-based nonlinear filtering algorithm. A complete characterization of the bias-variance terms in the optimality gap of (11) is subject of ongoing work using the existing tools for convergence of empirical measures in Wasserstein metric (Fournier & Guillin, 2015), OT map estimation (Hütter & Rigollet, 2021), and Bayesian posterior perturbation analysis (Garbuno-Inigo et al., 2023).

Numerical experiments show the better performance of the OT approach for examples designed to have multi-modal distributions, while it is noted that the raw computational time of the OT approach is higher, and for nonlinear filtering examples that admit unimodal posterior, such as Lorentz-96, the EnKF provides a fast and reasonable approximation. A computational feature of the OT method is that it provides the user with the flexibility to set the computational budget: without any training, OT algorithm implements EnKF; with additional budget (increasing training iterations and complexity of the neural net), the accuracy is increased, see Appendix C.1. The computational efficiency of the OT approach can be improved by fine-tuning the neural network architectures, optimizing the hyper-parameters, and including an offline training stage for the first time step, which will be used as a warm-start for training at future time steps in the online implementation. This line of research is pursued in our recent work (Al-Jarrah et al., 2024), where a new data-driven nonlinear filtering algorithm was introduced aimed at ergodic state and observation dynamics. The algorithm consists of offline and online stages: The offline stage is expensive to train and learns a static conditioning transport map; The online stage is computationally cheap and uses the learned conditioning map without any further training, providing a competitive computational time compared with traditional methods during online inference.

Acknowledgements

The authors thank the reviewers for their constructive comments and feedback, which helped improve the paper. Mohammad Al-Jarrah and Amirhossein Taghvaei are supported by the National Science Foundation (NSF) award EPCN-2318977, Niyizhen Jin and Bamdad Hosseini are supported by the NSF award DMS-2208535. Both supports are gratefully acknowledged.

Impact Statement

This paper impacts the field of nonlinear filtering, data assimilation, and uncertainty quantification. It facilitates effective quantification and evaluation of uncertainty, allowing safe and reliable operation of autonomous systems.

References

Al-Jarrah, M., Hosseini, B., and Taghvaei, A. Optimal transport particle filters. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 6798–6805. IEEE, 2023.

Al-Jarrah, M., Hosseini, B., and Taghvaei, A. Data-driven approximation of stationary nonlinear filters with optimal transport maps. *arXiv preprint arXiv:2403.15630*, 2024.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Bengtsson, T., Bickel, P., and Li, B. Curse of dimensionality revisited: Collapse of the particle filter in very large scale systems. In *IMS Lecture Notes - Monograph Series in Probability and Statistics: Essays in Honor of David F. Freedman*, volume 2, pp. 316–334. Institute of Mathematical Sciences, 2008.

Bergemann, K. and Reich, S. An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3):213, 2012.

Bickel, P., Li, B., Bengtsson, T., et al. Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, pp. 318–329. Institute of Mathematical Statistics, 2008.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional Monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Calvello, E., Reich, S., and Stuart, A. M. Ensemble Kalman methods: a mean field perspective. *arXiv preprint arXiv:2209.11371*, 2022.

Cappé, O., Moulines, E., and Rydén, T. Inference in hidden Markov models. In *Proceedings of EUSFLAT Conference*, pp. 14–16, 2009.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Crisan, D. and Xiong, J. Approximate McKean-Vlasov representations for a class of SPDEs. *Stochastics An International Journal of Probability and Stochastics Processes*, 82(1):53–68, 2010. doi: 10.1080/17442500902723575.

Daum, F. and Huang, J. Particle flow and monge-kantorovich transport. In *2012 15th International Conference on Information Fusion*, pp. 135–142. IEEE, 2012.

Daum, F., Huang, J., and Noushin, A. Exact particle flow for nonlinear filters. In *Signal processing, sensor fusion, and target recognition XIX*, volume 7697, pp. 92–110. SPIE, 2010.

De Melo, F. E., Maskell, S., Fasiolo, M., and Daum, F. Stochastic particle flow for nonlinear high-dimensional filtering problems. *arXiv preprint arXiv:1511.01448*, 2015.

Del Moral, P. and Guionnet, A. 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, pp. 155–194. Elsevier, 2001.

Doucet, A. and Johansen, A. M. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(3):656–704, 2009.

El Moselhy, T. A. and Marzouk, Y. M. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Evensen, G. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003. doi: 10.1007/s10236-003-0036-9.

Evensen, G. *Data Assimilation: The Ensemble Kalman Filter*, volume 2. Springer, 2009.

Fan, J., Taghvaei, A., and Chen, Y. 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*, pp. 1571–1581. PMLR, 18–24 Jul 2021.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

Garbuno-Inigo, A., Helin, T., Hoffmann, F., and Hosseini, B. Bayesian posterior perturbation analysis with integral probability metrics. *arXiv preprint arXiv:2303.01512*, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gordon, N. J., Salmond, D. J., and Smith, A. F. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, volume 140, pp. 107–113. IET, 1993.

Grange, D., Al-Jarrah, M., Baptista, R., Taghvaei, A., Georgiou, T. T., Phillips, S., and Tannenbaum, A. Computational optimal transport and filtering on Riemannian manifolds. *IEEE Control Systems Letters*, 2023.

Hütter, J.-C. and Rigollet, P. Minimax rates of estimation for smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

Kakade, S., Shalev-Shwartz, S., Tewari, A., et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1):35, 2009.

Kovachki, N., Baptista, R., Hosseini, B., and Marzouk, Y. Conditional sampling with monotone GANs: from generative models to likelihood-free inference. *arXiv preprint arXiv:2006.06755*, 2023.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. Sampling via measure transport: An introduction, in Handbook of Uncertainty Quantification. *Springer*, pp. 1–41, 2016.

Mesa, D. A., Tantiongloc, J., Mendoza, M., Kim, S., and P. Coleman, T. A distributed framework for the construction of transport maps. *Neural computation*, 31(4): 613–652, 2019.

Moreau, J.-J. Proximité and dualityé in a Hilbertian space. *Bulletin of the Mathematical Society of France*, 93:273–299, 1965.

Pathiraja, S., Reich, S., and Stannat, W. Mckean–Vlasov SDEs in nonlinear filtering. *SIAM Journal on Control and Optimization*, 59(6):4188–4215, 2021.

Ray, D., Ramaswamy, H., Patel, D. V., and Oberai, A. A. The efficacy and generalizability of conditional GANs for posterior inference in physics-based inverse problems. *Numerical Algebra, Control and Optimization*, 2022.

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

Reich, S. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Analysis*, 51:235–249, 2011. doi: 10.1007/s10543-010-0302-4.

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

Reich, S. Data assimilation: The Schrödinger perspective. *Acta Numerica*, 28:635–711, 2019.

Rockafellar, R. T. *Convex analysis*, volume 11. Princeton university press, 1997.

Rout, L., Korotin, A., and Burnaev, E. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pp. 1792–1802. PMLR, 2022.

Siahkoohi, A., Rizzuti, G., Louboutin, M., Witte, P. A., and Herrmann, F. J. Preconditioned training of normalizing flows for variational inference in inverse problems. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2021.

Spantini, A., Baptista, R., and Marzouk, Y. Coupling techniques for nonlinear ensemble filtering. *SIAM Review*, 64 (4):921–953, 2022.

Taghvaei, A. and Hosseini, B. An optimal transport formulation of Bayes' law for nonlinear filtering algorithms. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 6608–6613. IEEE, 2022.

Taghvaei, A. and Mehta, P. G. Optimal transportation methods in nonlinear filtering. *IEEE Control Systems Magazine*, 41(4):34–49, 2021.

Taghvaei, A. and Mehta, P. G. A survey of feedback particle filter and related controlled interacting particle systems (CIPS). *Annual Reviews in Control*, 2023.

Van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S. Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, 145(723):2335–2365, 2019.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Yang, T., Mehta, P. G., and Meyn, S. P. A mean-field control-oriented approach to particle filtering. In *Proceedings of the 2011 American Control Conference*, pp. 2037–2043. IEEE, 2011.

Yang, T., Mehta, P. G., and Meyn, S. P. Feedback particle filter. *IEEE transactions on Automatic control*, 58(10): 2465–2480, 2013.

Yang, T., Laugesen, R. S., Mehta, P. G., and Meyn, S. P. Multivariable feedback particle filter. *Automatica*, 71: 10–23, 2016.

A. Preliminaries

We start by reviewing preliminaries on convex analysis and OT theory.

A.1. Preliminaries on convex analysis

We recall definitions from convex analysis that will be useful in the subsequent sections. These definitions appear in most expositions on convex analysis, e.g. see (Rockafellar, 1997).

- **Convex conjugate:** For a function $\phi : \mathbb{R}^n \mapsto \cup\{\pm\infty\}$, its convex conjugate, denoted as ϕ^*, is defined according to
$$\phi^*(y) = \sup_{x\in\mathbb{R}^n} (x^T y - \phi(x)). \tag{16}$$
Moreover, ϕ is convex iff there exists a function ψ such that $\phi = \psi^*$.
- **Strong convexity:** A function $\phi : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\pm\infty\}$ is α-strongly convex if for all $x, y, z \in \mathbb{R}^n$ and $t \in [0, 1]$:
$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y) - \frac{\alpha}{2}t(1-t)\|x-y\|^2.$$
- **Smoothness:** A function $\phi : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\pm\infty\}$ is β-smooth if it is differentiable and for all $x, y \in \mathbb{R}^n$:
$$\phi(y) \leq \phi(x) + \nabla\phi(x)^\top (y-x) + \frac{\beta}{2}\|x-y\|^2.$$

For a lower semi-continuous cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we recall the following definition.

- **inf-c convolution:** For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$, its inf-$c$ convolution, denoted by f^c, is defined according to
$$f^c(y) = \inf_{x\in\mathbb{R}^n} \left[c(x, y) - f(x)\right]. \tag{17}$$
Moreover, f is said to be c-concave if there exists a function g such that $f = g^c$.

Remark A.1. With a squared cost function $c(x, y) = \frac{1}{2}\|x-y\|_2^2$, the c-concave property is related to convexity. In particular,
$$f \text{ is } c\text{-concave if and only if } \phi = \tfrac{1}{2}\|\cdot\|^2 - f \text{ is convex.}$$
Moreover, the inf-c convolution and convex conjugate are related according to:
$$\frac{1}{2}\|\cdot\|^2 - f^c = (\frac{1}{2}\|\cdot\|^2 - f)^*.$$

We also recall the definition of sub-differential and gradient of a convex function.

- **Sub-differential and gradient:** The sub-differential of a convex function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ at $x \in \mathbb{R}^n$, denoted by $\partial\phi(x)$, is a set containing all vectors $y \in \mathbb{R}^n$ such that
$$\phi(z) \geq \phi(x) + y^\top (z-x), \quad \forall z \in \mathbb{R}^n.$$
Moreover, ϕ is differentiable at x, with derivative $\nabla\phi(x)$, iff $\partial\phi(x) = \{\nabla\phi(x)\}$.

The definition of the convex conjugate and sub-differential imply the relationship (Rockafellar, 1997, Thm. 23.5)
$$\phi(x) + \phi^*(y) = x^\top y \quad \Leftrightarrow \quad x \in \partial\phi^*(y) \quad \Leftrightarrow \quad y \in \partial\phi(x).$$
In particular, when ϕ^* is differentiable at $y \in \mathbb{R}^n$, we have the identity
$$\phi^*(y) = \nabla\phi^*(y)^\top y - \phi(\nabla\phi^*(y)).$$
The relationship between convexity and c-concave, when c is the squared cost, implies the similar identity when f^c is differentiable at $y \in \mathbb{R}^n$:
$$f^c(y) = c(y, y - \nabla f^c(y)) - f(y - \nabla f^c(y)).$$

Finally, we recall the duality relationship between strong convexity and smoothness (e.g. see (Kakade et al., 2009, Thm. 6)):
$$\phi \text{ is } \alpha\text{-strongly convex iff } \phi^* \text{ is } \tfrac{1}{\alpha}\text{-smooth.}$$

A.2. Preliminary on OT theory

In this subsection, we review the problem formulation and key results from OT theory. For a complete treatment of the subject, see (Villani, 2009).

Given two probability distributions P and Q on $\mathbb{R}^n$, the Monge optimal transportation problem aims to find a map $T\colon \mathbb{R}^n \to \mathbb{R}^n$ that solves the optimization problem

$$\inf_{T\in\mathcal{T}(P,Q)} \mathbb{E}_{Z\sim P}[c(Z,T(Z))], \tag{18}$$

where $\mathcal{T}(P,Q) := \{T\colon \mathbb{R}^n \to \mathbb{R}^n;\, T_{\#}P = Q\}$ is the set of all transport maps pushing forward P to Q, and $c\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a lower semi-continuous cost function that is bounded from below. The Monge problem is relaxed by replacing deterministic transport maps with stochastic couplings according to

$$\inf_{\pi\in\Pi(Q,P)} \mathbb{E}_{(Z',Z)\sim\pi}[c(Z',Z)], \tag{19}$$

where $\Pi(Q,P)$ denotes the set of all joint distributions on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals Q and P. This relaxation, due to Kantorovich, turns the Monge problem into a linear program, whose dual becomes

$$\sup_{(f,g)\in\text{Lip}_c} \{\mathcal{J}_{\text{dual}}(f,g) := \mathbb{E}_{Z'\sim Q}[f(Z')] + E_{Z\sim P}[g(Z)]\} \tag{20}$$

where Lip_c is the set of pairs of functions (f,g), from $\mathbb{R}^n \to \mathbb{R}$, that satisfy the constraint

$$f(z) + g(z') \le c(z,z'), \quad \forall z,z' \in R^n.$$

The well-known Brenier's result (Brenier, 1991) establishes the existence and uniqueness of the solution to the Monge problem by making a direct connection to the solution of the dual Kantorovich problem.

Theorem A.2 ((Brenier, 1991))**.** *Consider the Monge problem (18) with cost $c(z,z') = \|z-z'\|^2/2$. Assume P and Q have finite second-order moments and P is absolutely continuous with respect to the Lebesgue measure. Then, the Monge problem admits a unique minimizer $\bar{T}(z) = z - \nabla\bar{f}^c(z)$ where $\bar{f}$ is c-concave and the pair $(\bar{f},\bar{f}^c)$ are the unique (up to an additive constant) maximizer of the dual Kantorovich problem (20).*

A.3. Max-min formulation

In this subsection, we present the max-min formulation of the dual Kantorovich problem (20), which is useful for numerical approximation of OT maps (Makkuva et al., 2020; Rout et al., 2022). The presentation starts with a formal derivation followed by a proposition supported by rigorous arguments.

The first step is to use the result of Theorem A.2 to replace g with f^c, in the dual problem (20), resulting the following optimization problem

$$\sup_{f\in c\text{-Concave}} \mathcal{J}_{\text{dual}}(f,f^c),$$

where c-Concave is the set of functions on $\mathbb{R}^n$ that are c-concave. Next, we use the definition of f^c in (17) to conclude

$$\mathbb{E}_{Z\sim P}[f^c(Z)] = \mathbb{E}_{Z\sim P}[\inf_{z\in\mathbb{R}^n} c(z,Z) - f(z)] = \inf_{T\in\mathcal{M}(P)} \mathbb{E}_{Z\sim P}[c(T(Z),Z) - f(T(Z))]$$

where we assumed that

$$T(Z) = \arg\min_{z\in\mathbb{R}^n} c(Z,z) - f(z), \quad \text{a.e.}$$

for a measurable map $T : \mathbb{R}^n \to \mathbb{R}^n$. This assumption is valid for $T = \text{Id} - \nabla f^c$ whenever f^c is differentiable. Combining these two steps, the Kantorovich dual problem is expressed as the following max-min problem

$$\sup_{f\in c\text{-concave}} \inf_{T\in\mathcal{M}(P)} \{\mathcal{J}_{\text{max-min}}(f,T) := \mathbb{E}_{Z'\sim Q}[f(Z')] + \mathbb{E}_{Z\sim P}[c(T(Z),Z) - f(T(Z))]\} \tag{21}$$

The following proposition, which is an extension of the existing result in (Makkuva et al., 2020, Thm. 3.3), connects the max-min problem to the original Kantorovich problem.

Proposition A.3. *Consider the max-min problem (21) under the assumptions of Thm. A.2. Let $(\bar{f}, \bar{f}^c)$ be the optimal pair for the dual Kantorovich problem (20), and $\bar{T} = Id - \nabla f^c$ the OT map from P to Q. Then, $(\bar{f}, \bar{T})$ is the unique pair, up to an additive constant for $\bar{f}$, that solves the problem (21), i.e.*

$$\sup_{f \in c\text{-concave}} \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\textit{max-min}}(f, T) = \mathcal{J}_{\textit{max-min}}(\bar{f}, \bar{T}). \tag{22}$$

and if $\mathcal{J}_{\textit{max-min}}(\bar{f}, \bar{T}) = \mathcal{J}_{\textit{max-min}}(g, S)$, then $g = \bar{f}$ + constant and $S = \bar{T}$.

Proof. We present the proof in four steps.

Step 1: According to the definition of inf-c convolution (17)

$$f^c(Z) \leq c(T(Z), Z) - f(T(Z)), \quad \text{a.e.}$$

for all maps $T \in \mathcal{M}(P)$. Upon taking the expectation with respect to $Z \sim P$, adding $\mathbb{E}_{Z' \sim Q}[f(Z')]$, and taking the inf over $T \in \mathcal{M}(P)$,

$$\mathcal{J}_{\text{dual}}(f, f^c) \leq \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T).$$

Step 2: For any c-concave function f, introduce $f_\epsilon := f - \frac{\epsilon}{2}\|\cdot\|^2$. This transformation ensures that its inf-c convolution f_ϵ^c is differentiable a.e. ($f_\epsilon^c = \frac{1}{2}\|\cdot\|^2 - \phi_\epsilon^*$, with $\phi_\epsilon := \frac{1}{2}\|\cdot\|^2 - f_\epsilon$, is $\frac{1}{\epsilon}$-smooth because ϕ_ϵ is ϵ-strongly convex. This regularization process is known as *Moreau-Yosida approximation* (Moreau, 1965)). Because f_ϵ^c is differentiable, the minimum in the definition of $f_\epsilon^c(z)$ is achieved at a unique point $z - \nabla f_\epsilon^c(z)$, concluding the identity

$$f_\epsilon^c(Z) = c(Z - \nabla f_\epsilon^c(Z), Z) - f_\epsilon(Z - \nabla f_\epsilon^c(Z)), \quad \text{a.e.}$$

Using the fact that $f_\epsilon(z) \leq f(z)$ on the right-hand-side of this identity, we arrive at the inequality

$$f_\epsilon^c(Z) \geq c(Z - \nabla f_\epsilon^c(Z), Z) - f(Z - \nabla f_\epsilon^c(Z)), \quad \text{a.e.}$$

Taking the expectation over $Z \sim P$ and adding $\mathbb{E}_{Z' \sim Q}[f(Z')]$ yields

$$\mathcal{J}_{\text{dual}}(f, f_\epsilon^c) \geq \mathcal{J}_{\text{max-min}}(f, \text{Id} - \nabla f_\epsilon^c) \geq \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T),$$

where the last inequality follows because $\text{Id} - \nabla f_\epsilon^c \in \mathcal{M}(P)$. Finally, using $\mathbb{E}_{Z' \sim Q}[f(Z')] = \mathbb{E}_{Z' \sim Q}[f_\epsilon(Z')] + \epsilon M$, where $M := \mathbb{E}_{Z' \sim Q}[\frac{1}{2}\|Z'\|^2] < \infty$, concludes

$$\mathcal{J}_{\text{dual}}(f_\epsilon, f_\epsilon^c) + \epsilon M \geq \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T).$$

Step 3: Combining the results of step 1 and 2 gives

$$\mathcal{J}_{\text{dual}}(f, f^c) \leq \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T) \leq \mathcal{J}_{\text{dual}}(f_\epsilon, f_\epsilon^c) + \epsilon M$$

Taking the sup over $f \in c$-Concave and the fact that $(\bar{f}, \bar{f}^c)$ are the maximizer of the dual problem, concludes

$$\mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c) \leq \sup_{f \in c\text{-concave}} \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T) \leq \mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c) + \epsilon M$$

Taking the limit as $\epsilon \to 0$ concludes the identity

$$\sup_{f \in c\text{-concave}} \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T) = \mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c)$$

The proof of (22) is concluded by noting that $\mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c) = J_{\text{max-min}}(\bar{f}, \bar{T})$ for $\bar{T} = \text{Id} - \nabla \bar{f}^c$ because $\bar{f}^c$ is differentiable a.e..

Step 4: To prove uniqueness, let (g, S) be another optimal pair for $J_{\text{max-min}}$. Therefore,

$$\mathcal{J}_{\text{max-min}}(g, S) = \sup_{f \in c\text{-concave}} \inf_{T \in \mathcal{M}(P)} \mathcal{J}_{\text{max-min}}(f, T) = \mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c). \tag{23}$$

On the other hand, from the inequalities obtained in the previous steps

$$\mathcal{J}_{\text{dual}}(g, g^c) \leq \mathcal{J}_{\text{max-min}}(g, S) \leq \mathcal{J}_{\text{dual}}(g, g^c_\epsilon)$$

Taking the limit as $\epsilon \to 0$ and the application of monotone convergence theorem implies $\mathcal{J}_{\text{dual}}(g, g^c) = \mathcal{J}_{\text{max-min}}(g, S)$, which with the optimality condition (23), concludes $\mathcal{J}_{\text{dual}}(g, g^c) = \mathcal{J}_{\text{dual}}(\bar{f}, \bar{f}^c)$. The uniqueness of the solution $(\bar{f}, \bar{f}^c)$ to the dual problem (20) and the differentiability of $\bar{f}^c$ concludes that $g = \bar{f} + \text{const.}$ and $S = \text{Id} - \nabla \bar{f}^c = \bar{T}$. □

B. Proofs of the theoretical results

B.1. Justification for the consistency condition (8a)

The consistency condition (8a) implies that

$$\mathbb{E}[F(T(\overline{X}, Y), Y)] = \mathbb{E}[F(X, Y)]$$

for all measurable and bounded functions $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. In particular, with $F(x, y) = f(x)g(y)$, we have

$$\mathbb{E}[f(T(\overline{X}, Y))g(Y)] = \mathbb{E}[f(X)g(Y)]$$

for all measurable and bounded functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$. The definition of the conditional expectation implies

$$\mathbb{E}[f(T(\overline{X}, Y))|Y] = \mathbb{E}[f(X)|Y].$$

Finally, the independence of $\bar{X}$ and Y yields

$$\mathbb{E}[f(T(\overline{X}, y)] = \mathbb{E}[f(X)|Y = y], \quad \text{a.e. } y$$

for all measurable and bounded functions f, concluding $T(\cdot, y)_\# P_X = P_{X|Y}(\cdot|y)$ a.e. y.

B.2. Proof of Proposition 2.3

Express the objective function in (8c) as

$$J(f, T) = \mathbb{E}[\mathcal{J}^Y_{\text{max-min}}(f(\cdot, Y), T(\cdot, Y))]$$

where

$$\mathcal{J}^y_{\text{max-min}}(g, S) = \mathbb{E}_{X \sim P_{X|Y=y}}[g(X)] + \mathbb{E}_{\overline{X} \sim P_X}[c(S(\overline{X}), \overline{X}) - g(S(\overline{X})]$$

is the max-min objective function in (21) with the marginal distributions $Q = P_{X|Y=y}$ and $P = P_X$. Then, according to the Prop. A.3,

$$\sup_{g \in c\text{-Concave}} \inf_{S \in \mathcal{M}(P_X)} \mathcal{J}^y_{\text{max-min}}(g, S) = \mathcal{J}^y_{\text{max-min}}(\overline{g}_y, \overline{S}_y)$$

for a unique pair $(\overline{g}_y, \overline{S}_y)$ where $\overline{S}_y = \text{Id} - \nabla \overline{g}^c_y$ is the OT map from P_X to $P_{X|Y=y}$. Theorem 2.3 of (Carlier et al., 2016)[5] verifies that there is a measurable concatenations of the functions $\overline{g}_y$ to construct the pair $(\overline{f}, \overline{T})$ according to

$$\overline{f}(x, y) = \overline{g}_y(x), \quad \overline{T}(x, y) = x - \nabla \overline{g}^c_y(x).$$

The concatenated functions solve the optimization problem (8c) and $T(\cdot, y)$ serves the OT map from P_X to $P_{X|Y=y}$, i.e. $\mathcal{B}_y(\pi) = \overline{T}(\cdot, y)_\# \pi$.

[5]The dual function f is related to the dual function ϕ in (Carlier et al., 2016) with the transformation $\phi(x, y) = \frac{1}{2}\|x\|^2 - f(x, y)$.

B.3. Proof of Proposition 2.4

We present the proof under the change of variable $\phi(x,y) = \frac{1}{2}\|x\|^2 - f(x,y)$. Under this change of variable, the condition $f \in c\text{-Concave}_x$ translates to $\phi \in \text{CVX}_x$, i.e. ϕ is convex in the first variable x, and the objective function in (8c) becomes

$$J(f,T) = \mathbb{E}_{X\sim P_X}[\|X\|^2] - L(\phi,T) \tag{24}$$

where

$$L(\phi,T) := \mathbb{E}_{(X,Y)\sim P_{XY}}[\phi(X,Y)] + \mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\overline{X}^\top T(\overline{X},Y) - \phi(T(\overline{X},Y),Y)].$$

Next, we decompose the optimization gap $\epsilon(f,T) = \tilde{\epsilon}(\phi) + \tilde{\delta}(\phi,T)$ where,

$$J(\overline{f},\overline{T}) - \min_S J(f,S) = \max_S L(\phi,S) - L(\overline{\phi},\overline{T}) =: \tilde{\epsilon}(\phi)$$
$$J(f,T) - \min_S J(f,S) = \max_S L(\phi,S) - L(\phi,T) =: \tilde{\delta}(\phi,T)$$

and $\phi(x,y) = \frac{1}{2}\|x\|^2 - f(x,y)$ and similarly $\overline{\phi}(x,y) = \frac{1}{2}\|x\|^2 - \overline{f}(x,y)$ and we used (24).

The proposition assumption states that $\phi(x,y) = \frac{1}{2}\|x\|^2 - f(x,y)$ is α-strongly convex in x. The duality between strong convexity and smoothness implies that $\phi^*(x,y)$ is $\frac{1}{\alpha}$-smooth in x, where $\phi^*(\cdot,y)$ is the convex conjugate of $\phi(\cdot,y)$ for all y. As a result,

$$\phi^*(w,y) \;\le\; \phi^*(x,y) + \nabla_x\phi^*(x,y)^\top(w-x) + \frac{1}{2\alpha}\|w-x\|^2 := h(w,y).$$

The inequality $\phi^*(w,y) \;\le\; h(w,y)$ implies $\phi(w,y) \;\ge\; h^*(w,y)$. Consequently, we have the inequality:

$$\phi(w,y) \;\ge\; -\phi^*(x,y) + w^\top x + \frac{\alpha}{2}\|w - \nabla_x\phi^*(x,y)\|^2. \tag{25}$$

This inequality plays a crucial role in deriving lower bounds for the quantities $\tilde{\delta}(\phi,T)$ and $\tilde{\epsilon}(\phi)$. We first establish a bound for $\tilde{\delta}(\phi,T)$ using (25):

$$\begin{aligned}
\tilde{\delta}(\phi,T) &= \max_S L(\phi,S) - L(\phi,T)\\
&= L(\phi,\nabla_x\phi^*) - L(\phi,T)\\
&= \mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\phi^*(\overline{X},Y) - \overline{X}^T T(\overline{X},Y) + \phi(T(\overline{X},Y),Y)]\\
&\ge \frac{\alpha}{2}\mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\|T(\overline{X},Y) - \nabla_x\phi^*(\overline{X},Y)\|^2]
\end{aligned}$$

where the second and third equality follows from the fact that

$$\sup_{z\in\mathbb{R}^n}\{z^\top x - \phi(z,y)\} = \nabla\phi_x^*(x,y)^\top x - \phi(\nabla_x\phi^*(x,y),y) = \phi^*(x,y). \tag{26}$$

Next, we obtain the bound for $\tilde{\epsilon}(\phi)$ as follows:

$$\begin{aligned}
\tilde{\epsilon}(\phi) &\overset{(1)}{=} \max_S L(\phi,S) - L(\overline{\phi},\overline{T})\\
&\overset{(2)}{=} L(\phi,\nabla\phi_x^*) - L(\overline{\phi},\overline{T})\\
&\overset{(3)}{=} \mathbb{E}_{(X,Y)\sim P_{XY}}[\phi(X,Y) - \overline{\phi}(X,Y)] + \mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\phi^*(\overline{X},Y) - \overline{X}^\top\overline{T}(\overline{X},Y) + \overline{\phi}(\overline{T}(\overline{X},Y),Y)]\\
&\overset{(4)}{=} \mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\phi(\overline{T}(\overline{X},Y),Y) + \phi^*(\overline{X},Y) - \overline{X}^\top\overline{T}(\overline{X},Y)]\\
&\overset{(5)}{\ge} \frac{\alpha}{2}\mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\|\overline{T}(\overline{X},Y) - \nabla_x\phi^*(\overline{X},Y)\|^2]
\end{aligned}$$

where we used (26) in the second and third steps, $(\overline{T}(\overline{X},Y),Y) \sim P_{XY}$ in the fourth step, and (25) in the fifth step. Finally, we use the inequality $(a+b)^2 \le 2a^2 + 2b^2$ to obtain:

$$\begin{aligned}
\mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\|T(\overline{X},Y) - \overline{T}(\overline{X},Y)\|^2] &\le 2\mathbb{E}_{(\overline{X},Y)\sim P_X\otimes P_Y}[\|T(\overline{X},Y) - \nabla_x\phi^*(\overline{X},Y)\|^2]\\
&\quad + \|\nabla_x\phi^*(\overline{X},Y) - \overline{T}(\overline{X},Y)\|^2] \le \frac{4}{\alpha}(\tilde{\delta}(\phi,T) + \tilde{\epsilon}(\phi)) = \frac{4}{\alpha}\epsilon(f,T)
\end{aligned}$$

B.4. Proof of the Prop. 2.7

B.4.1. Proof of the lower-bound (13a)

The bound is based on the application of the central limit theorem to $\int g d\pi^{(SIR)} - \int g d(\mathcal{B}_Y(\pi))$ for any uniformly bounded function g (e.g. see (Cappé et al., 2009, Thm. 9.1.8)). In particular, the definition of $\pi^{(SIR)}$ and $\mathcal{B}_y(\pi)$ imply

$$\int g d\pi^{(SIR)} = \frac{\frac{1}{N}\sum_{i=1}^N h(Y|X^i) g(X^i)}{\frac{1}{N}\sum_{i=1}^N h(Y|X^i)} = \frac{\frac{1}{N}\sum_{i=1}^N \bar{h}(Y|X^i) g(X^i)}{\frac{1}{N}\sum_{i=1}^N \bar{h}(Y|X^i)}$$
$$\int g d\mathcal{B}_Y(\pi) = \mathbb{E}[g(X)|Y]$$

where $\bar{h}(y|x) = h(y|x)/\int h(y|x')d\pi(x')$ is the normalized likelihood. Then, subtracting the two expressions and multiplication by $\sqrt{N}$ yields

$$\sqrt{N}\left(\int g d\pi^{(SIR)} - \int g d\mathcal{B}_Y(\pi)\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \bar{h}(Y|X^i)\left(g(X^i) - \mathbb{E}[g(X)|Y]\right)}{\frac{1}{N}\sum_{i=1}^N \bar{h}(Y|X^i)}.$$

Application of the central limit theorem to the numerator, and law of the large numbers to the denominator, concludes the convergence

$$\sqrt{N}\left(\int g d\pi^{(SIR)} - \int g d\mathcal{B}_Y(\pi)\right) \longrightarrow Z \sim N(0, V_h(g)))$$

where

$$V_h(g) = \mathbb{E}\left[\bar{h}(Y|\bar{X})^2\left(g(\bar{X}) - \mathbb{E}[g(X)|Y]\right)^2\right]$$

and $\bar{X} \sim \pi$ is an independent copy of X. As a result,

$$\lim_{N\to\infty} \sqrt{N}\,\mathbb{E}\left[\left(\int g d\pi^{(SIR)} - \int g d\mathcal{B}_Y(\pi)\right)^2\right]^{\frac{1}{2}} = \sqrt{V_h(g)}.$$

Finally, upon using the definition of the metric,

$$\begin{aligned}\liminf_{N\to\infty} \sqrt{N} d(\pi^{(SIR)}, \mathcal{B}_Y(\pi)) &= \liminf_{N\to\infty} \sup_{g\in\mathcal{G}} \sqrt{N}\,\mathbb{E}\left[\left(\int g d\pi^{(SIR)} - \int g d\mathcal{B}_y(\pi)\right)^2\right]^{\frac{1}{2}} \\ &\geq \sup_{g\in\mathcal{G}} \liminf_{N\to\infty} \sqrt{N} E\left[\left(\int g d\pi^{(SIR)} - \int g d\mathcal{B}_y(\pi)\right)^2\right]^{\frac{1}{2}} \\ &= \sup_{g\in\mathcal{G}} \sqrt{V_h(g)}.\end{aligned}$$

B.4.2. Proof of the upper-bound (13b)

According to Prop. 2.3 and the definition (12b), we have $\mathcal{B}_Y(\pi) = \overline{T}(\cdot, Y)_\# \pi$ and $\pi^{(OT)} = \widehat{T}(\cdot, Y)_\# \pi$, respectively. As a result,

$$\begin{aligned}d(\pi^{(OT)}, \mathcal{B}_Y(\pi)) &= d(\widehat{T}(\cdot, Y)_\# \pi, \overline{T}(\cdot, Y)_\# \pi) \\ &= \sup_{g\in\mathcal{G}} \mathbb{E}\left[\left(\int g d(\widehat{T}(\cdot, Y)_\# \pi) - \int g d(\overline{T}(\cdot, Y)_\# \pi)\right)^2\right]^{\frac{1}{2}}\end{aligned}$$

By expressing

$$\int g d(\widehat{T}(\cdot, Y)_\# \pi) = \mathbb{E}[g(\widehat{T}(\bar{X}, Y)|Y] \quad \text{and}$$
$$\int g d(\overline{T}(\cdot, Y)_\# \pi) = \mathbb{E}[g(\overline{T}(\bar{X}, Y)|Y],$$

where $\bar{X} \sim \pi$ is an independent copy of X, we arrive at the identity

$$d(\pi^{(OT)}, \mathcal{B}_Y(\pi)) = \sup_{g \in \mathcal{G}} \mathbb{E}\left[\left(\mathbb{E}[g(\widehat{T}(\bar{X}, Y)|Y] - \mathbb{E}[g(\overline{T}(\bar{X}, Y)|Y]\right)^2\right]^{\frac{1}{2}}.$$

Upon application of the Jensen's inequality and the fact that g is Lipschitz with constant 1,

$$\begin{aligned} d(\pi^{(OT)}, \mathcal{B}_Y(\pi)) &\leq \sup_{g \in \mathcal{G}} \mathbb{E}\left[\left(g(\widehat{T}(\bar{X}, Y) - g(\overline{T}(\bar{X}, Y)\right)^2\right]^{\frac{1}{2}} \\ &\leq \mathbb{E}\left[\|\widehat{T}(\bar{X}, Y) - \overline{T}(\bar{X}, Y)\|^2\right]^{\frac{1}{2}}. \end{aligned}$$

Finally, according to Prop. 2.4, the right-hand-side of this inequality is bounded by $\sqrt{\frac{4}{\alpha}\epsilon(\widehat{f}, \widehat{T})}$, concluding the upper-bound (13b).

C. Numerical details and additional results

Our numerical results involve simulation of three filtering algorithms: (1) Ensemble Kalman filter (EnKF) (Evensen, 2003; Calvello et al., 2022), presented in Algorithm 1; (2) Sequential importance resampling (SIR) (Doucet & Johansen, 2009), presented in Algorithm 2; (3) OT particle filter (OTPF), presented in Algorithm 3.

In order to implement the OTPF, we used the ResNet neural network architecture to represent both f and T as they appear in Fig. 2. We used the Adam optimizer with different learning rates for each example to solve the max-min problem (10) with inner-loop iteration number always equal to 10. The number of particles, N, and the initial particles, is the same for all three algorithms. The EnKF algorithm is simulated with the additional regularization $\Gamma = \sigma_w^2 I$ where σ_w is the noise level in the observation signal, and I is the identity matrix. The numerical results are produced using the following two machines:

1. MACBOOK M1 Pro with 8-core CPU, 14-core GPU, and 16GB unified memory
2. MAC STUDIO M2 Max with 12-core CPU, 30-core GPU, and 64GB unified memory

Algorithm 1 Ensemble Kalman filter (EnKF)

Input: Initial particles $\{X_0^i\}_{i=1}^N \sim \pi_0$, observation signal $\{Y_t\}_{t=1}^{t_f}$, $\Gamma \succ 0$
dynamic model $a(x \mid x')$, observation model $h(y \mid x)$.
for $t = 1$ to t_f **do**
  Propagation:
  for $i = 1$ to N **do**
    $X_{t|t-1}^i \sim a(. \mid X_{t-1}^i)$
    $Y_t^i \sim h(. \mid X_{t|t-1}^i)$
  end for
  Conditioning:
  $\widehat{X}_t = \frac{1}{N}\sum_{i=1}^N X_{t|t-1}^i$
  $\widehat{Y}_t = \frac{1}{N}\sum_{i=1}^N Y_t^i$
  $C_t^{xy} = \frac{1}{N}\sum_{i=1}^N (X_{t|t-1}^i - \widehat{X}_t) \otimes (Y_t^i - \widehat{Y}_t)$
  $C_t^{yy} = \frac{1}{N}\sum_{i=1}^N (Y_t^i - \widehat{Y}_t) \otimes (Y_t^i - \widehat{Y}_t)$
  $K_t = C_t^{xy}(C_t^{yy} + \Gamma)^{-1}$
  for $i = 1$ to N **do**
    $X_t^i = X_{t|t-1}^i + K_t(Y_t - Y_t^i)$.
  end for
end for
Output: Particles $\{X_t^i\}_{i=1}^N$ for $t = 0, \ldots, t_f$.

Algorithm 2 Sequential importance resampling (SIR)

Input: Initial particles $\{X_0^i\}_{i=1}^N \sim \pi_0$, observation signal $\{Y_t\}_{t=1}^{t_f}$,
dynamic model $a(x \mid x')$, observation model $h(y \mid x)$.
for $t = 1$ to t_f **do**
  Propagation:
  for $i = 1$ to N **do**
    $X_{t|t-1}^i \sim a(. \mid X_{t-1}^i)$.
  end for
  Conditioning:
  $w_t^i = \frac{h(Y_t|X_{t|t-1}^i)}{\sum_{i=1}^N h(Y_t|X_{t|t-1}^i)}$
  $X_t^i \sim \sum_{i=1}^N w_t^i \delta_{X_{t|t-1}^i}^i$
end for
Output: Particles $\{X_t^i\}_{i=1}^N$ for $t = 0, \ldots, t_f$.

C.1. Computational time

The raw computational time of the algorithms, across all of the main examples, is presented in Table 1. It is observed that the computational time of the OT method is higher than the other two algorithms. These results are included for comprehensive understanding without an intensive focus on optimizing computational time, which may be achieved by careful selection of architectures for f and T, tuning the algorithm hyper-parameters, and including an offline training stage that provides a warm start for training in the online implementation. This is subject of ongoing research. In our preliminary efforts in this direction, we were able to reduce the computational time of the OT method for the Bimodal dynamic example to 2.9151s with only 1.3% loss in accuracy. This is achieved by performing the training for the initial step offline for 10^4 iterations. The resulting f and T are then used as warm start for the online implementation of the filter, with training iterations reduced to 8.

In order to further study the computational efficiency, we numerically evaluated the relationship/trade-off between computational time and filtering accuracy, for the bimodal dynamic example in section 3.2, for the OT and SIR methods. In the implementation of the OT method, we start with the outer-loop iteration number $K_{(1)}^{(out)} = 2^{12} = 4096$, as it appears in algorithm 3, and decrease it according to the rule $K_{(t+1)}^{(out)} = K_{(t)}^{(out)}/2$ until we reach a predefined final iteration number K_s.

The result is presented in Fig. 10. The filtering accuracy is evaluated in terms of the MMD distance with respect to the exact posterior distribution. Panel (a) shows the relationship between the MMD distance and computational time, as the problem dimension n increases. Panel (b) shows the same relationship, but as the number of particles N change. The result for the OT method is presented for three different values of the final iteration number K_s. The result of panel (a) shows that, although the absolute computational time of OT is larger than SIR, it scales better with the problem dimension. The results of panel (b) shows that the accuracy of the OT can be improved by increasing the number of particles, without significant increase in the computational time, while the SIR method should use a much larger number of particles to reach the same accuracy.

Table 1: Computational time in seconds for EnKF, SIR and OTPF.

EXAMPLE	ENKF	SIR	OTPF
BIMODEL DYNAMIC EXAMPLE	0.0095	0.0108	34.4555
LORENZ 63 EXAMPLE (200 STEPS)	0.0542	0.0605	50.9799
LORENZ 96 EXAMPLE (200 STEPS)	0.3029	0.3726	79.7705
STATIC IMAGE IN-PAINTING ON MNIST	0.4577	0.3880	529.8863
DYNAMIC IMAGE IN-PAINTING ON MNIST	0.5373	0.3128	728.8051

Algorithm 3 OT particle filter (OTPF)

Input: Initial particles $\{X_0^i\}_{i=1}^N \sim \pi_0$, observation signal $\{Y_t\}_{t=1}^{t_f}$,
dynamic and observation models $a(x \mid x'), h(y \mid x)$
batch size bs, optimizer and learning rates for f, T, inner loop iterations $K^{(\text{in})}$, outer iteration sequence $K_t^{(\text{out})}$.
Initialize: initialize neural net f, T according to the architecture in Fig. 2, and their weights θ_f, θ_T.
for $t = 1$ to t_f **do**
  Propagation:
  for $i = 1$ to N **do**
    $X_{t|t-1}^i \sim a(. \mid X_{t-1}^i)$
    $Y_t^i \sim h(. \mid X_{t|t-1}^i)$
  end for
  Conditioning:
  Create a random permutation $\{\sigma_i\}_{i=1}^N$
  for $k = 1$ to $K_t^{(\text{out})}$ **do**
    sample bs particles from $(X_{t|t-1}^i, X_{t|t-1}^{\sigma_i}, Y_t^i)$
    for $j = 1$ to $K^{(\text{in})}$ **do**
      Update θ_T to minimize $\frac{1}{bs}\sum_{i=1}^{bs}\left[\frac{1}{2}\|T(X_{t|t-1}^{\sigma_i}, Y_t^i) - X_{t|t-1}^{\sigma_i}\|^2 - f(T(X_{t|t-1}^{\sigma_i}, Y_t^i), Y_t^i)\right]$
    end for
    Update θ_f to minimize $\frac{1}{bs}\sum_{i=1}^{bs}\left[-f(X_{t|t-1}^i, Y_t^i) + f(T(X_{t|t-1}^{\sigma_i}, Y_t^i), Y_t^i)\right]$
  end for
  for $i = 1$ to N **do**
    $X_t^i = T(X_{t|t-1}^i, Y_t)$.
  end for
end for
Output: Particles $\{X_t^i\}_{i=1}^N$ for $t = 0, \ldots, t_f$.

C.2. Additional static example

Consider the following observation model:

$$Y = X + \sigma_w W, \qquad X \sim \frac{1}{2}N(-1, \sigma^2 I_2) + \frac{1}{2}N(+1, \sigma^2 I_2) \tag{27}$$

where $W \sim N(0, I_2)$ and $\sigma_w = 0.4$. In this example, the prior distribution is bimodal, while the observation model is linear. The numerical result for a varying number of particles is presented in Fig. 11, where it is observed that the OT algorithm is also capable of pushing a bimodal to an unimodal distribution. At the same time, EnKF underperforms compared to the other two methods due to its Gaussian approximation of the prior.

In this example, and also the examples in Sec. 3.1, we did not use the EnKF layer in modeling T. We used a ResNet with a single-block of size 32 and ReLU-activation function. The learning rate of the Adam optimizer is set to 10^{-3}, with the total number of iterations 2×10^4, and batch-size is 128.

C.3. Additional dynamic example

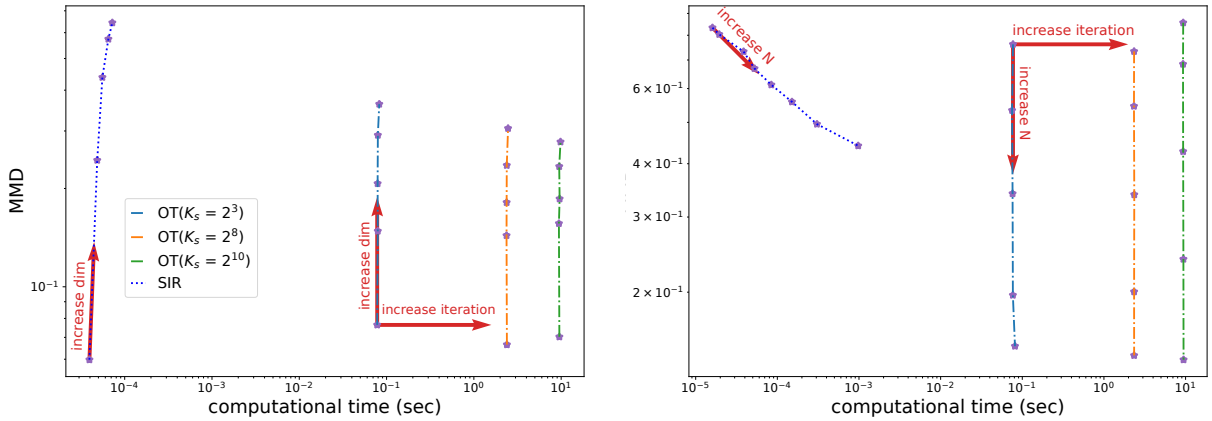
Consider the following model:

$$X_t = (1-\alpha)X_{t-1} + 2\lambda V_t, \quad X_0 \sim \mathcal{N}(0, I_n), \tag{28a}$$

$$Y_t = h(X_t) + \lambda W_t, \tag{28b}$$

where $\{V_t, W_t\}_{t=1}^{\infty}$ are i.i.d sequences of standard Gaussian random variables, $\alpha = 0.1$ and $\lambda = \sqrt{0.1}$. This is the generalization of the example 15 where now, instead of $h(x) = x \odot x$, the observation function is selected to be linear $h(x) = x$ or cubic $h(x) = x \odot x \odot x$.

We parameterize f and T as before, except that we included the EnKF layer in T and used two ResNet blocks. The learning rate for the optimizers of f and T are 10^{-3} and 2×10^{-3}, respectively. The total number of iterations is 1024, which is



(a) MMD vs computational time as dimension n varies.

(b) MMD vs computational time as particle size N varies.

Figure 10: Evaluation of the computational time of the OT and SIR algorithms, for the example in Sec. 3.2. The left panel shows the relationship between the MMD distance and computational time, as the problem dimension n increases. The computational time of the OT approach is evaluated for three different values of final number of iteration K_s, as explained in Sec. C.1. The right panel shows the same relationship, but as the number of particles N change.

divided by 2 after each time step (of the filtering problem) until it reaches 64. The batch-size is 64 and the number of particles $N = 1000$.

We quantify the performance of the algorithms by computing the maximum mean discrepancy (MMD) distance between the exact posterior distribution and the approximated distribution formed by the particles. In order to give an accurate approximation of the true posterior, we use the SIR method with $N = 10^5$ particles. The MMD distance between two empirical distributions $\mu = \frac{1}{N_1}\sum_{i=1}^{N_1}\delta_{U^i}$ and $\nu = \frac{1}{N_2}\sum_{i=1}^{N_2}\delta_{V^i}$ is

$$\text{MMD}_{\text{emp}}(\mu,\nu) = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}k(U^i,U^j) + \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}k(V^i,V^j) - \frac{2}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}k(U^i,V^j) \tag{29}$$

where $k(\cdot,\cdot)$ denotes the RBF kernel. The kernel bandwidth is selected individually for each example

The numerical results for the linear and cubic observation models are depicted in Fig. 12 and 13, respectively. EnKF performs better than the other two algorithms in the linear Gaussian setting, while it does not perform as well with the cubic observation model. It is noted that the OT algorithm's performance can be enhanced through fine-tuning and additional training iterations.

C.4. The details of the Lorenz 63 model and additional results

We consider the following Lorenz 63 model:

$$\begin{aligned}\begin{bmatrix}\dot{X}(1)\\ \dot{X}(2)\\ \dot{X}(3)\end{bmatrix} &= \begin{bmatrix}\sigma(X(2)-X(1))\\ X(1)(\rho-X(3))-X(2)\\ X(1)X(2)-\beta X(3)\end{bmatrix}, \quad X_0 \sim \mathcal{N}(\mu_0,\sigma_0^2 I_3),\\ Y_t &= \begin{bmatrix}X_t(1)\\ X_t(3)\end{bmatrix} + \sigma_{obs}W_t,\end{aligned} \tag{30}$$

where $[X(1),X(2),X(3)]^\top$ are the variables representing the hidden states of the system, and σ, ρ, and β are the model parameters. We choose $\sigma = 10$, $\rho = 28$, $\beta = 8/3$, $\mu_0 = [25,25,25]^\top$, and $\sigma_0^2 = 10$. The observed noise W is a 2-dimensional standard Gaussian random variable with $\sigma_{obs}^2 = 10$.

We simulated the Lorenz 63 model without any noise in the dynamics; however, we included an artificial noise

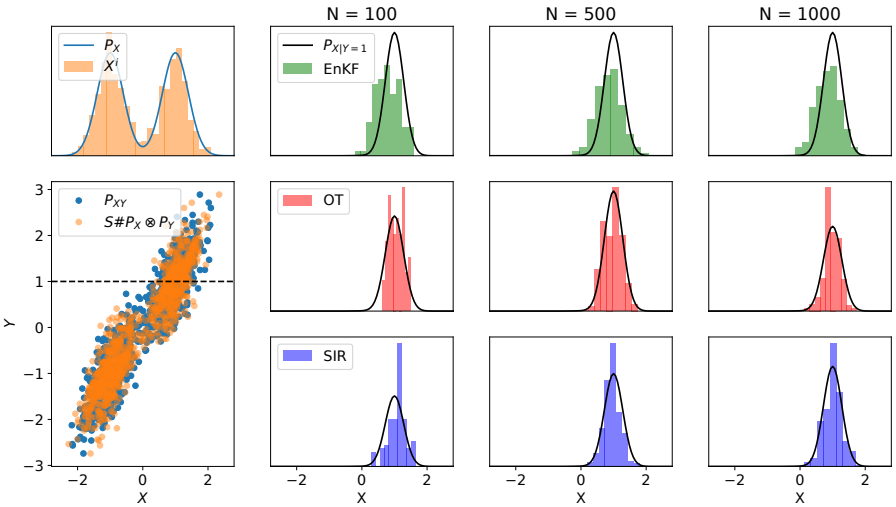


Figure 11: Numerical results for the static example in equation (27). Top-left: Samples $\{X^i\}_{i=1}^N$ from the prior P_X; bottom-left: samples $\{(X^i, Y^i)\}_{i=1}^N$ from the joint distribution P_{XY} in comparison with the transported samples $\{(T(X^{\sigma_i}, Y^i), Y^i)\}_{i=1}^N$; rest of the panels: transported samples for $Y = 1$ for different values of N and three different algorithms.

$\mathcal{N}(0, \sigma^2_{added} I_3)$, $\sigma^2_{added} = 1$ to the dynamic update step of the algorithms. In order to better study the difference between the three filters, we initialized the particles from a Gaussian distribution with mean $[0, 0, 0]$, while the true state is initialized with a Gaussian with mean $[25, 25, 25]$.

The neural network for T is a two-block residual network of size 64 with or without EnKF layer. The learning rate for the optimizers of f and T are 5×10^{-2} and 10^{-2}, respectively. The total number of iterations is 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 64 from the total of $N = 1000$ particles.

In the left panel in Fig. 14, we present the trajectories of the true states and particles for all three algorithms. In the right panel of Fig. 14, we show the mean-squared-error (MSE) of estimating the state by taking the average over 10 independent simulations. It is observed that both OT settings, with and without the EnKF layer, are performing better than EnKF and SIR overall. The EnKF layer helps with the performance at the initial stage; however, it requires careful tuning for the learning rate.

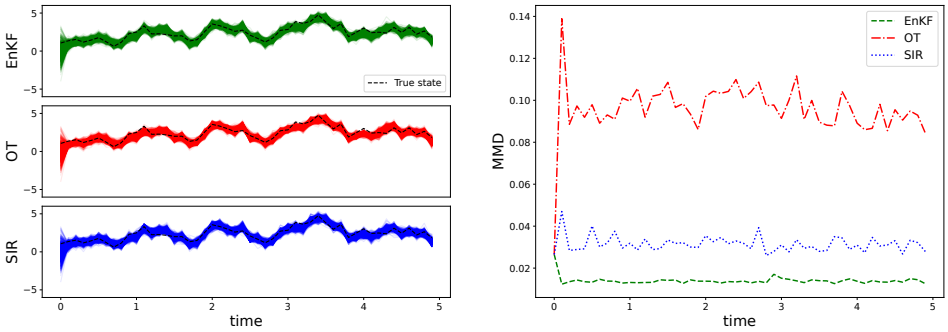
C.5. The Lorenz 96 model

Consider the following Lorenz-96 model:

$$\dot{X}(k) = (X(k+1) - X(k-2))X(k-1) - X(k) + F + \sigma V, \quad \text{for} \quad k = 1, \ldots, n$$
$$Y_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} X_t + \sigma W_t \tag{31}$$

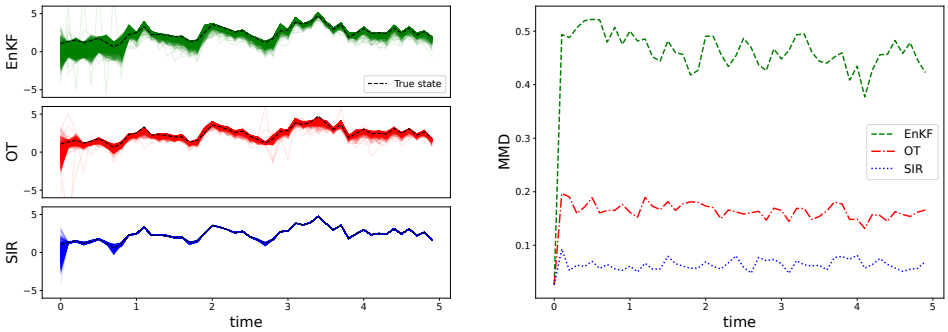
for $n = 9$ where $X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2 I_n)$ and we choose the convention that $X(-1) = X(n-1)$, $X(0) = X(n)$, and $X(n+1) = X(1)$, and $F = 2$ is a forcing constant. We choose the model parameters $\mu_0 = 25 \cdot \mathbf{1}_n$, and $\sigma_0^2 = 10^2$. The observed noise W is a n-dimensional standard Gaussian random variable with variance equal to 1.

For this experiment, we represent T as a two-block residual network of size 32 with EnKF layer. The learning rate for the optimizers of f and T are 10^{-3} and 10^{-1}, respectively. The total number of iterations is 1024, which is divided by 2 after each time step (of the filtering problem) until it reaches 64. The batch-size is 128, and the total of particles $N = 1000$.



(a) Particles trajectory. (b) MMD vs time.

Figure 12: Numerical results for the dynamic example (28) where $h(X_t) = X_t$. The left panel shows the trajectory of the particles $\{X_t^1, \ldots, X_t^N\}$ along with the trajectory of the true state X_t for EnKF, OT, and SIR algorithms, respectively. The second panel shows the MMD distance with respect to the exact conditional distribution.



(a) Particles trajectory. (b) MMD vs time.

Figure 13: Numerical results for the dynamic example (28) where $h(X_t) = X_t \odot X_t \odot X_t$. The left panel shows the trajectory of the particles $\{X_t^1, \ldots, X_t^N\}$ along with the trajectory of the true state X_t for EnKF, OT, and SIR algorithms, respectively. The second panel shows the MMD distance with respect to the exact conditional distribution.

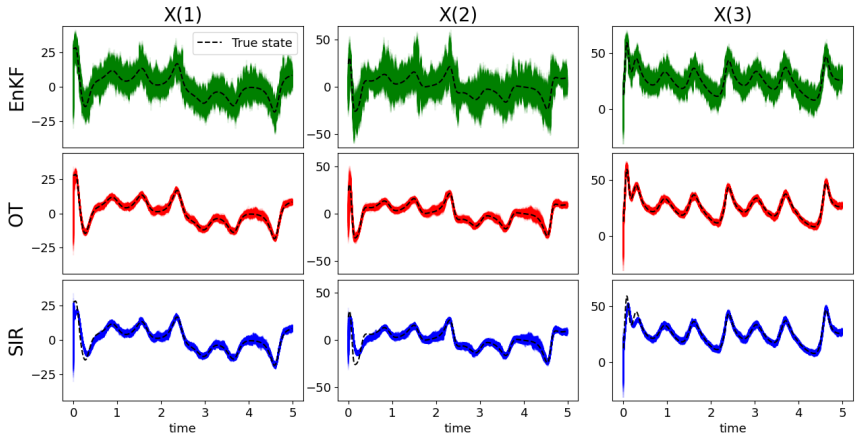
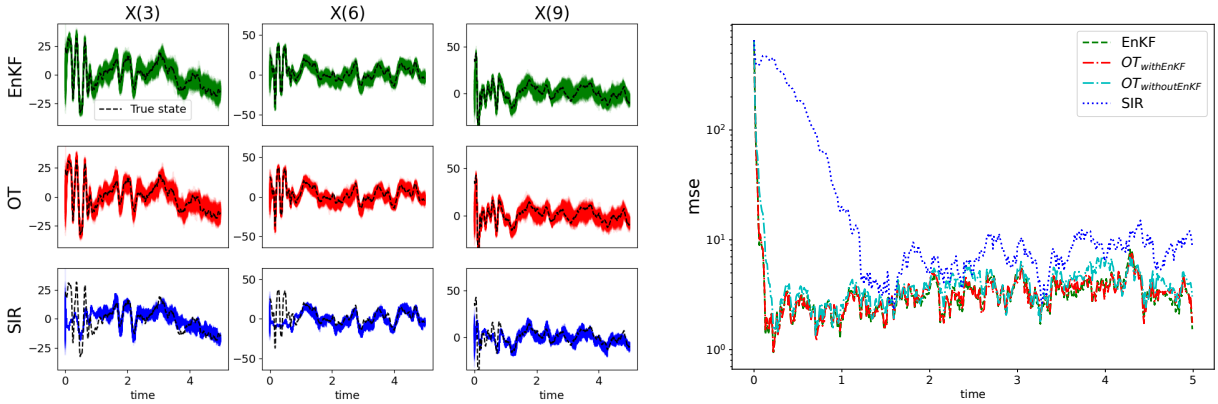


Figure 14: Numerical results for the Lorenz 63 example (30). The figure shows the trajectory of the particles $\{X^i\}$ along with the trajectory of the true state.Each column represents one state, and each row represents one algorithm.

The numerical results are shown in Fig. 15. The columns in panel (a) represent unobserved states, and the rows represent the results for each algorithm. Panel (b) shows the MSE in estimating the state, averaged over 10 independent simulations.

The results indicate that the OT algorithm performs as well as the EnKF algorithm, which appears to perform well for this problem.



(a) Particles trajectory.

(b) MSE vs time.

Figure 15: Numerical results for the Lorenz 96 example (31). (a) The trajectory of the particles $\{X^i\}$ along with the trajectory of the true state; each column represents one of the unobserved state, and each row represents one algorithm; (b) Comparison of the MSE in estimating the state as a function of time.

C.6. Additional details and results for static image in-painting on MNIST

In this example, we consider the problem of computing conditional distributions on the 100-dimensional latent space of generative adversarial network (GAN) trained to represent the MNIST digits (Goodfellow et al., 2014). In particular, denoting the generator by $G : \mathbb{R}^{100} \to \mathbb{R}^{28\times 28}$, we consider the model:

$$Y_t = h(G(X), c_t) + \sigma W_t, \quad X \sim N(0, I_{100}),$$

where the observation function $(z, c) \in \mathbb{R}^{28\times 28} \times \mathbb{R}^2 \mapsto h(z, c) \in \mathbb{R}^{r\times r}$ is defined as the $r \times r$ window of pixels $z[c(1) : c(1) + r, c(2) : c(2) + r]$. The coordinates of the corner c_t move from left to right and top to bottom scanning a partial portion of the image called the *observed part*. In order to make the problem more challenging, we fed a noisy measurement of the corner location to the algorithms by adding a uniform random integer between -2 and 2 for each axis. The observational noise set to be $\sigma = 10^{-1}$, with every component of $W_t \sim N(0, 1)$. While the true image does not change, we included artificial Gaussian noise $N(0, \sigma^2 I_{100})$ to the particles to avoid particle collapse.

For this example, we used the publicly available codes for a GAN model on MNIST [6] and a classifier achieving an accuracy exceeding 98% on the MNIST test dataset [7].

The neural network for T is a residual network with size 320 without the EnKF layer. The learning rate for both f and T is 10^{-3} with $2^{12} = 4096$ iterations. Batch-size is 64 and total number of particles $N = 1000$.

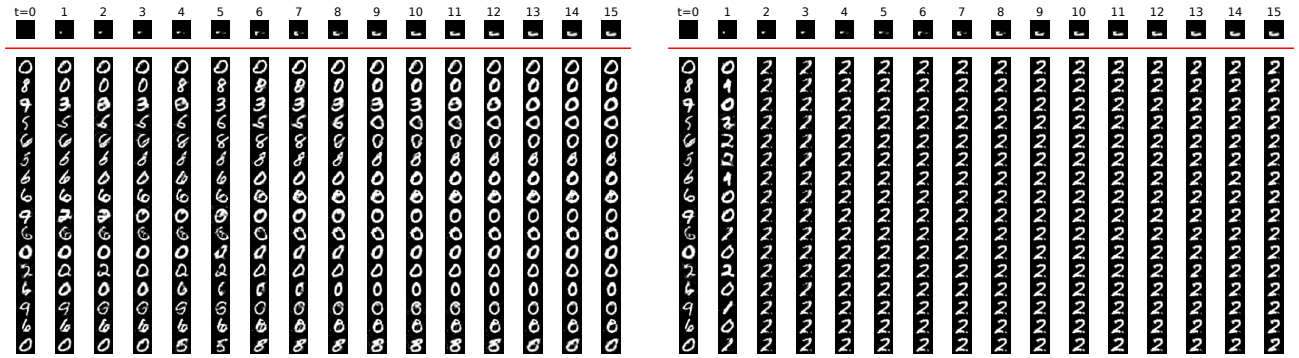
Additional numerical results for this example are presented in Fig. 16. The top row shows the total observed part of the image up to that time step, and the following 16 rows show the images generated from the particles that approximate the conditional distribution.

C.7. Additional details and results for dynamic image in-painting on MNIST

We extended the previous static example by adding a dynamic update for the latent variable X as follows:

$$X_{t+1} = (1 - \alpha)X_t + \sigma V_t,$$
$$Y_{t+1} = h(G(X_{t+1}), c_{t+1}) + \sigma_w W_{t+1},$$

[6] https://github.com/lyeoni/pytorch-mnist-GAN/tree/master
[7] https://nextjournal.com/gkoehler/pytorch-mnist



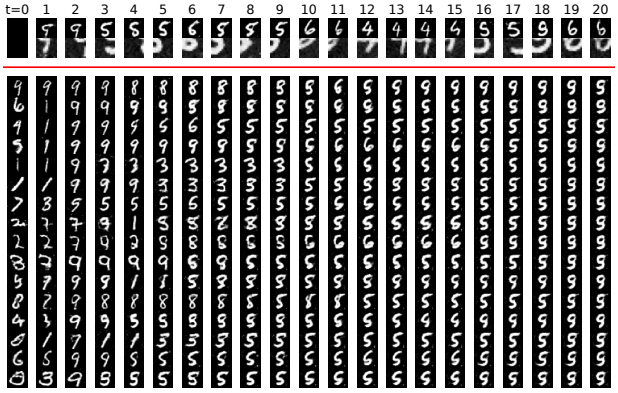
(a) EnKF particles.

(b) SIR particles.

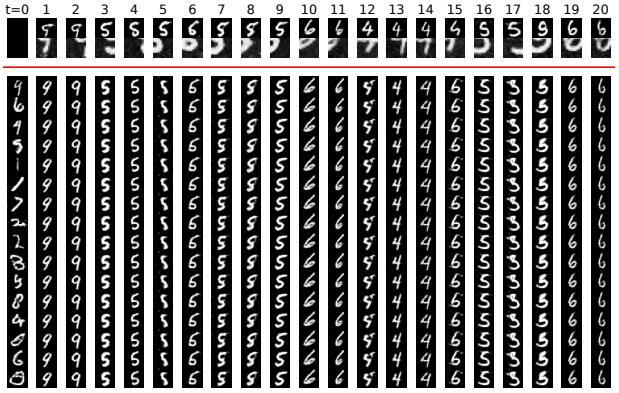
Figure 16: Additional numerical results for the static image in-painting on MNIST. The first row shows the cumulative total observations up to each time step. The subsequent rows under the red line show 16 particles from the EnKF and SIR methods, respectively.

where V_t, W_t are standard Gaussian with the appropriate dimension, $\sigma = \sqrt{2\alpha - \alpha^2}$, $\sigma_w = 10^{-1}$, and $\alpha = 0.2$. The observation function $h_t(G(X_t), c_t) = G(X_t)[28 - r : 28, c_t : c_t + r]$, $c_t \sim \mathit{Unif}_{\mathit{integer}}(1, 28 - r)$, and $r = 12$. The OT algorithm parameters are similar to the static case.

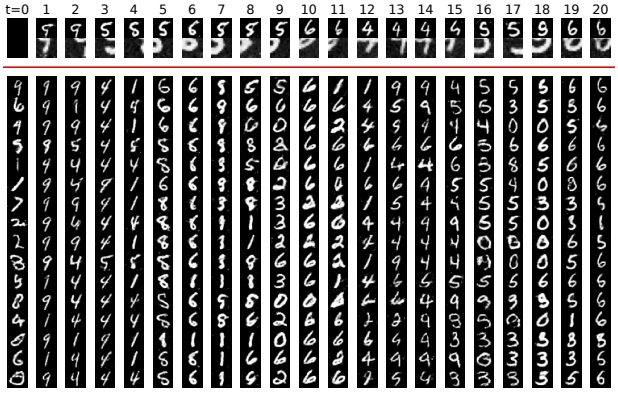
We present the trajectory of the particles in Fig. 17 (a),(b), and (c). The top row shows the true image, and the second row shows the observation at each time step. The top row shows the total observed part of the image up to that timestep, and the following 16 rows show the images generated from the particles that approximate the conditional distribution. In Fig. 17(d), we used an accurate MNIST classifier to represent the histogram of the digits generated from the particles of each algorithm.



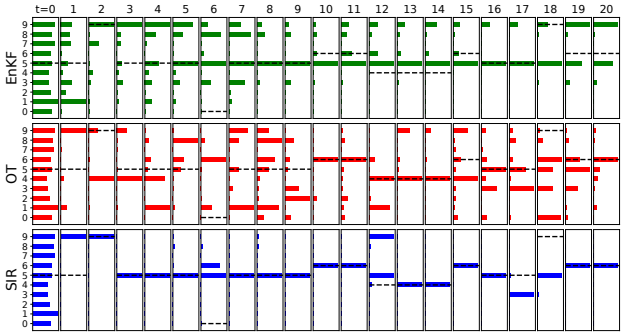
(a) EnKF particles.



(b) SIR particles.



(c) OT particles.



(d) Particles distribution.

Figure 17: Additional numerical results for the dynamic image in-painting on MNIST. (a-b-c) The first row shows the true image, and the second row shows the observation at each time step. The subsequent rows under the red line show 16 particles from the EnKF, SIR, and OT methods, respectively. (d) The histogram of the digits generated by the particles from the three algorithms as a function of time, evaluated using an accurate MNIST classifier.