Ensemble data assimilation using optimal control in the Wasserstein metric

Xin Liu [*],[1], Jason Frank

Mathematical Institute, Utrecht University, P.O. Box 80010, 3508 TA Utrecht, The Netherlands

ARTICLE INFO

ABSTRACT

An ensemble data assimilation method is proposed that is based on optimal control minimizing the cost of mismatch in the Wasserstein metric on the observation space. The new method achieved the optimal state without calculating the posterior distribution of the particle state and the particle states are evolved deterministically, which is easy to be implemented. The method is appropriate for systems in which multiple, noisy, partial observations are available (e.g. citizen weather stations or smart phones). The method is demonstrated for: (i) deterministic dynamics with uncertain initial conditions, (ii) multiple noisy observations of a randomly forced ordinary differential equation (ODE), (iii) observations from multiple sample paths from a stochastic differential equation (SDE). A bi-modal measure and a measure supported on a strange attractor are tested. The numerical results show that our method performs a relatively small Wasserstein distance which measures the approximation performance. But numerical implementation is a bit expensive due to the complexity of Wasserstein distance computation, especially with large set of particles.

1. Introduction

Data assimilation is a commonly used computational method for combining dynamic model simulations and observational data to estimate a state or trajectory of a dynamical system in fields as diverse as weather forecasting, computer vision, robotics and navigation. In uncertainty quantification, data assimilation may be used to approximate an evolving probability measure expressing uncertainty in the model, initial conditions or observations. Some references explaining the data assimilation include [1–4].

A common method for data assimilation is ensemble data assimilation, including the popular particle filter methods. Since particle filters were introduced in [5], they have become a very popular class of method that solve estimation problems in a recursive way depending on the observation data [6,7]. Particle filter algorithms use a set of particles to represent the posterior distribution of the stochastic process and they update their prediction in an approximate way. A common method of particle filters is sequential importance sampling (SIS), which relates all particles generated according to their importance weight at every stage [8]. However, The main difficulty with particle filter algorithm is that the particle weight would be unbalanced after a few steps, then the SIS algorithm will have a significant weight-degeneracy after a large number of iterations, i.e. all but one particle will be eliminated due to the low weight [9,10]. To avoid this problem, Sampling Importance Resampling (SIR) was introduced. The difference with SIS is that SIR resamples particles at every time stage, Specifically, it replicates the high-weighted particles and eliminates low-weighted particles. This approach is very useful and applied to solve many different kinds of problems. However, the disadvantage of SIR algorithm is that most resampled particles coincide, which leads to lower particle diversity especially for deterministic dynamics. Some other particle filter techniques have been developed to tackle the unbalanced weight problem. Zhao et al. [11] proposed an improved particle filter which depends on Pearson correlation coefficient to reduce the situation on particle degeneracy. Pearson correlation coefficient is applied to determine whether the particles are close to the true state or not, which is difficult to be applied in high dimensional systems. Zhu et al. [12] introduced an implicit equal-weights particle filter, which all particles are sampled with equal weight. However, the implicit equal weight is very complex to get. Frei et al. [13] tried to introduce a parameter $\gamma \in [0, 1]$ to bridge ensemble Kalman filter and particle filter by taking advantage of the ensemble kalman filter's nondegeneracy. But it is not very clear about how to determine the parameter γ. To obtain the estimation state, the methods proposed in [11–13] all need to calculate the posterior distribution.

In [14,15], the authors proposed a method to derive a particle filter by applying optimal control techniques. The method has a self-oriented formulation that provides a self-correcting feedback mechanism to stabilize the particles around the posterior. Inspired by [14,15], in this

---

paper, we introduce an alternative method to the construction of a particle filter by adding a control in the particle states. In contrast to [14,15], we evolve particle states deterministically. We obtain the optimal control without calculating the posterior distribution of the particle states. We employ a Wasserstein metric [16,17] in the cost function to measure the distance between probability distributions in the observation space.

Wasserstein distance arises from the optimal transport idea and now it is becoming more popular in machine learning and statistics. There are some references related to applying Wasserstein distance into data assimilation. Reich [18] introduced optimal transport into particle methods as a means of resampling. Sagar et al. [19] proposed an ensemble data assimilation over Riemannian manifold with Wasserstein distance. In paper [20], they proposed another new variation data assimilation combining with Wasserstein distance. [21] proposed a robust Kalman filter equipped with Wasserstein distance. Ning et al. [22] tried to reduce the forecast uncertainty caused from parameter estimation error in dissipative evolutionary equations through applying Wasserstein distance.

Although computationally complex, the Wasserstein distance is more robust than, e.g., the Kullback–Leibler divergence [23,24]. Since it relies on a metric equipped in a metric space, the Wasserstein distance can be employed for two measures even if their supports are mutually exclusive. As a result, the Wasserstein approach is applicable to alternative measures besides absolutely continuous ones, e.g. empirical measures or measures supported on strange attractors [25]. For instance, Wasserstein metric could be implemented to compute the distance between continuous distribution and discrete distribution. Unlike e.g. Euclidean distance and Helinger distance, Wasserstein distance has the advantage to capture the position errors, and the date shape is preserved during a morphing process [17,19,22].

This paper is organized as follows. Section 2 describes the ensemble data assimilation method which is optimal control-based data assimilation in detail. Lorenz 63 model and a bi-model system as numerical implementation are discussed in Section 3.

2. Optimal control-based data assimilation method

2.1. Data assimilation problem

In this paper we study an optimal control-based particle filter method for modeling uncertainty of a partially observed process. Our starting point is an ensemble of possibly noisy observations given in the form of K discrete time series

$$\hat{Z}_n^k = \hat{Z}^k(n\Delta t) \in \mathbf{R}^\ell, \quad n = 0, \dots, N, \quad k = 1, \dots, K,$$

where k denotes the ensemble index and n the time index over an interval $T = N\Delta t$.

We assume the underlying process $X(t)$, $t \in [0, T]$ is time-continuous and is described by either a deterministic (case $\sigma \equiv 0$) or stochastic differential equation

$$dX = a(X)\, dt + \Sigma\, dW, \tag{1}$$

where $X(t) \in \mathcal{D} \subset \mathbf{R}^d$ is the state at time t, $a(X) : \mathcal{D} \to \mathbf{R}^d$, $\Sigma \in \mathbf{R}^{d \times s}$ and $W(t)$ is an s-dimensional Wiener process.

Let $h(X) : \mathcal{D} \to \mathbf{R}^\ell$ be an observation function. Usually, one needs to deal with partial observations: $\ell < d$. An underlying assumption is that the state $X(t)$ is detectable by the observation function h. We assume the state $X(t)$ is unknown, due to uncertainty in initial condition, model error, or noise in the dynamics or measurements.

We distinguish three scenarios:

In the first scenario, we consider a deterministic system (i.e. $\Sigma \equiv 0$ in (1)) with uncertain initial condition and partial observations. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}^k(n\Delta t)), \quad n = 0, \dots, N, \quad k = 1, \dots, K, \tag{2}$$

where $\hat{X}^k(t)$, $k = 1, \dots, K$, denotes an ensemble of solutions of the deterministic differential equation (1), $\sigma \equiv 0$, with initial conditions $\hat{X}^k(0)$ drawn from a probability distribution.

In the second scenario, $\hat{X}(t)$ corresponds to a single sample path of the SDE (1) for which multiple noisy observations are available. This scenario models the case of weather measurements using a scattering of imperfect personal devices such as smart phones or private weather stations. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}(n\Delta t)) + \eta_n^k, \quad n = 0, \dots, N, \quad k = 1, \dots, K, \tag{3}$$

where the kth time series $\{\eta_n^k\}_{n=0}^N$ denotes the kth realization of the discrete noise process, and $\eta_n^k \sim \mathcal{N}(0, R)$, where $R \in \mathbf{R}^{\ell \times \ell}$ is the covariance matrix of the observational noise.

In the third scenario we assume we are given K sample paths of (1), i.e., $\hat{X}^k(t)$, $k = 1, \dots, K$, and the kth sequence $\{\hat{Z}_n^k\}_{n=0}^N$ is observed from $X^k(t)$, for $k = 1, \dots, K$. This scenario models the case of (possibly noisy) measurements of a repeated experiment with random forcing. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}^k(n\Delta t)) + \eta_n^k, \quad n = 0, \dots, N, \quad k = 1, \dots, K.$$

In all three scenarios, our objective is to estimate the uncertainty in our knowledge of the underlying process $\hat{X}(t)$ by approximating an evolving probability measure $\mu(x, t)$ such that for measurable $A \subset \mathcal{D}$,

$$\int_A \mu(x, n\Delta t)\, dx = \text{Prob}\{X(n\Delta t) \in A\}.$$

The measure μ will be approximated by an empirical measure $\nu_n(x)$ supported on an ensemble of J particles:

$$\nu_n(x) = \frac{1}{J} \sum_{j=1}^J \delta(x - X_n^j), \tag{4}$$

where δ denotes the Dirac distribution. The motion of the jth particle is governed by the drift vector field $a(X)$ and an optimal control via the differential equation

$$\frac{dX^j}{dt} = a(X^j) + Bu^j(t), \qquad j = 1, \dots, J,$$

$$Z^j(t) = h(X^j(t)), \qquad j = 1, \dots, J,$$

where $B \in \mathbf{R}^{d \times m}$ and $u^j(t) \in \mathbf{R}^m$ is the control input for jth particle at time t, chosen to minimize a cost function that penalizes mismatch (in the observation space) with respect to a Wasserstein metric. Of course, optimizing the mismatch does not guarantee the convergence of the measure ν to μ. Nevertheless such a strategy is common in variational data assimilation methods such as 4D-Var. The convergence question is related to concepts such as the synchronization of chaos, detectability, and Lyapunov stability theory [26–30]. By comparison with variational data assimilation, we can view the controls $u^j(t)$ as representing the unknown model error required to explain the observations.

The particle motion is discretized in time using Euler's method to obtain the discrete dynamics

$$X_{n+1}^j = X_n^j + \Delta t\, a(X_n^j) + \Delta t\, Bu_{n+1}^j, \quad n = 0, \dots, N-1, \quad j = 1, \dots, J, \tag{5}$$

$$Z_n^j = h(X_n^j), \qquad n = 0, \dots, N, \quad j = 1, \dots, J. \tag{6}$$

2.2. Wasserstein cost function

In this paper we study numerically the use of a Wasserstein metric to measure the mismatch in empirical distributions defined by the measurement ensemble and particle filter. The Wasserstein metric has found increased application in data assimilation, machine learning and data science in general, due to a number of attractive properties. For instance, the Wasserstein distance is well defined for singular measures and distributions, e.g. for measuring distance between empirical distributions or measures supported on strange attractors. Also, in the
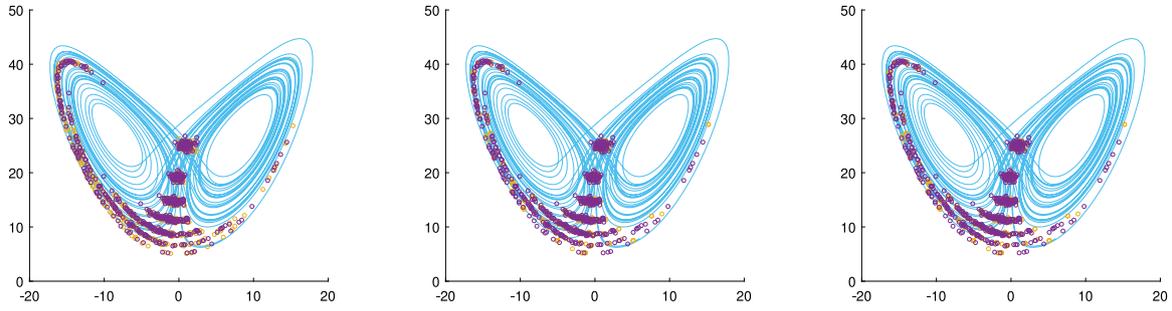
Fig. 1. Comparison of Wasserstein particle filter (left) and ensemble Kalman filter (middle) and SIR (right) for the Lorenz attractor at time $t = 0.6$. Partial observations (x-component only) were generated from a sample ensemble of trajectories whose final states are indicated by purple circles. The final states of the filters are indicated with yellow circles.
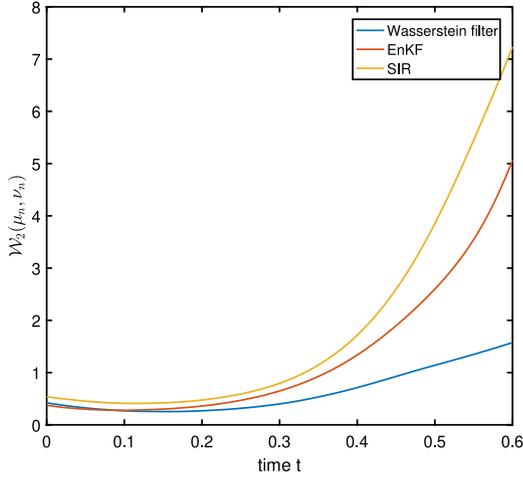
Fig. 2. The Wasserstein distance in full state (x, y, z).

Wasserstein metric, the geodesic path between two distributions is the optimal transport path, along which the deformation of a density is minimal. Consequently, in the context of data assimilation when observations may be sparse, the probability density function will deform in a minimal way between observation times.

Our goal is to choose the controls u_n^j in (5) so that the particle distribution ν_n approximates $\mu(x, n\Delta t)$. Given that we only have access to the sample observations $\{\hat{Z}_n^k\}$ we minimize a cost function that penalizes mismatch in the Wasserstein metric. Let

$$\hat{\zeta}_n(z) = \frac{1}{K} \sum_{k=1}^{K} \delta(z - \hat{Z}_n^k), \qquad \zeta_n(z) = \frac{1}{J} \sum_{j=1}^{J} \delta(z - Z_n^j) \qquad (7)$$

The cost function is defined as

$$C = \Delta t \sum_{n=0}^{N-1} \left[\sum_{j=1}^{J} \frac{1}{2} \|u_{n+1}^j\|^2 + \frac{\beta}{2} \mathcal{W}_2^2(\hat{\zeta}_{n+1}, \zeta_{n+1}) \right], \qquad (8)$$

where $\mathcal{W}_2$ denotes the 2-Wasserstein distance (see below) and the constant $\beta \geq 0$ is a weight parameter.

2.3. Calculation of Wasserstein distance

The Wasserstein distance is a metric on the space of probability measures. The p-Wasserstein distance between two probability measures μ and ν on a metric space $(\mathcal{X}, d)$ is given by

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}$$

where Π denotes the set of transport couplings of μ and ν, that is, $\Pi = \{\pi(x, y) | \int \pi(dx, y) = \mu(y), \int \pi(x, dy) = \nu(x)\}$. In general, the Wasserstein distance cannot be calculated analytically in most cases [23] and the

computational cost is higher than other distance, especially in high dimensions. Hence, efficient algorithms for computing Wasserstein distance are needed. The straightforward way to solve the Wasserstein distance is to use linear programming based algorithms such as network simplex.

For empirical measures such as (4), computing the Wasserstein distance reduces to solving an optimal transportation problem of weighted point sets, a special case of the minimum cost flow problem [31,32]. Given empirical measures

$$\mu(x) = \frac{1}{J} \sum_{j=1}^{J} \delta(x - \hat{X}^j), \quad \nu(x) = \frac{1}{K} \sum_{k=1}^{K} \delta(x - X^k),$$

consider the space of finite transport maps:

$$\mathcal{F} = \left\{ F = (f_{jk}) \in \mathbf{R}^{J \times K} \middle| f_{jk} \geq 0, \sum_j f_{jk} = \frac{1}{K}, \sum_k f_{jk} = \frac{1}{J} \right\}.$$

The 2-Wasserstein distance is equal to

$$\mathcal{W}_2(\mu, \nu) = \left(\min_{F \in \mathcal{F}} \sum_{jk} f_{jk} d_{jk}^2 \right)^{1/2}, \qquad (9)$$

where we use a weighted norm

$$d_{jk}^2 = \|\hat{X}^j - X^k\|_M^2 = (\hat{X}^j - X^k)^T M (\hat{X}^j - X^k). \qquad (10)$$

For instance, for noisy observations (3) with covariance matrix R, we choose $M = R^{-1}$ to reflect our confidence/uncertainty in the observations.

The minimization (9) constitutes a linear program. The Wasserstein distance can be efficiently estimated using the Sinkhorn algorithm [33].

2.4. Optimal control

To determine the optimal control $\{u_j^n\}$ in (5), we minimize the cost function (8) under constraints (5)–(6). We introduce Lagrange multipliers $\{\lambda_n^j\}$ and $\{\Lambda_n^j\}$ and define a discrete Lagrangian functional:

$$L = C + L_0 + L_\lambda,$$

where C is the cost function (8), L_0 enforces the constraint on the initial conditions, $X_0^j = \xi_0^j$, presumed known (or sampled from a known initial distribution),

$$L_0 = \sum_{j=1}^{J} \left[\lambda_0^j (X_0^j - \xi_0^j) + \Lambda_0^j (Z_0^j - h(\xi_0^j)) \right], \qquad (11)$$

and L_λ defines the constraint relations:

$$L_\lambda = \sum_{n=0}^{N-1} \sum_{j=1}^{J} \left[(\lambda_{n+1}^j)^T (X_{n+1}^j - X_n^j - \Delta t a(X_n^j) - \Delta t B u_{n+1}^j) \right.$$
$$\left. + (\Lambda_{n+1}^j)^T (Z_{n+1}^j - h(X_{n+1}^j)) \right]. \qquad (12)$$

Note that we include the observation function (6) as a constraint, as the observations appear implicitly in the cost function (8).
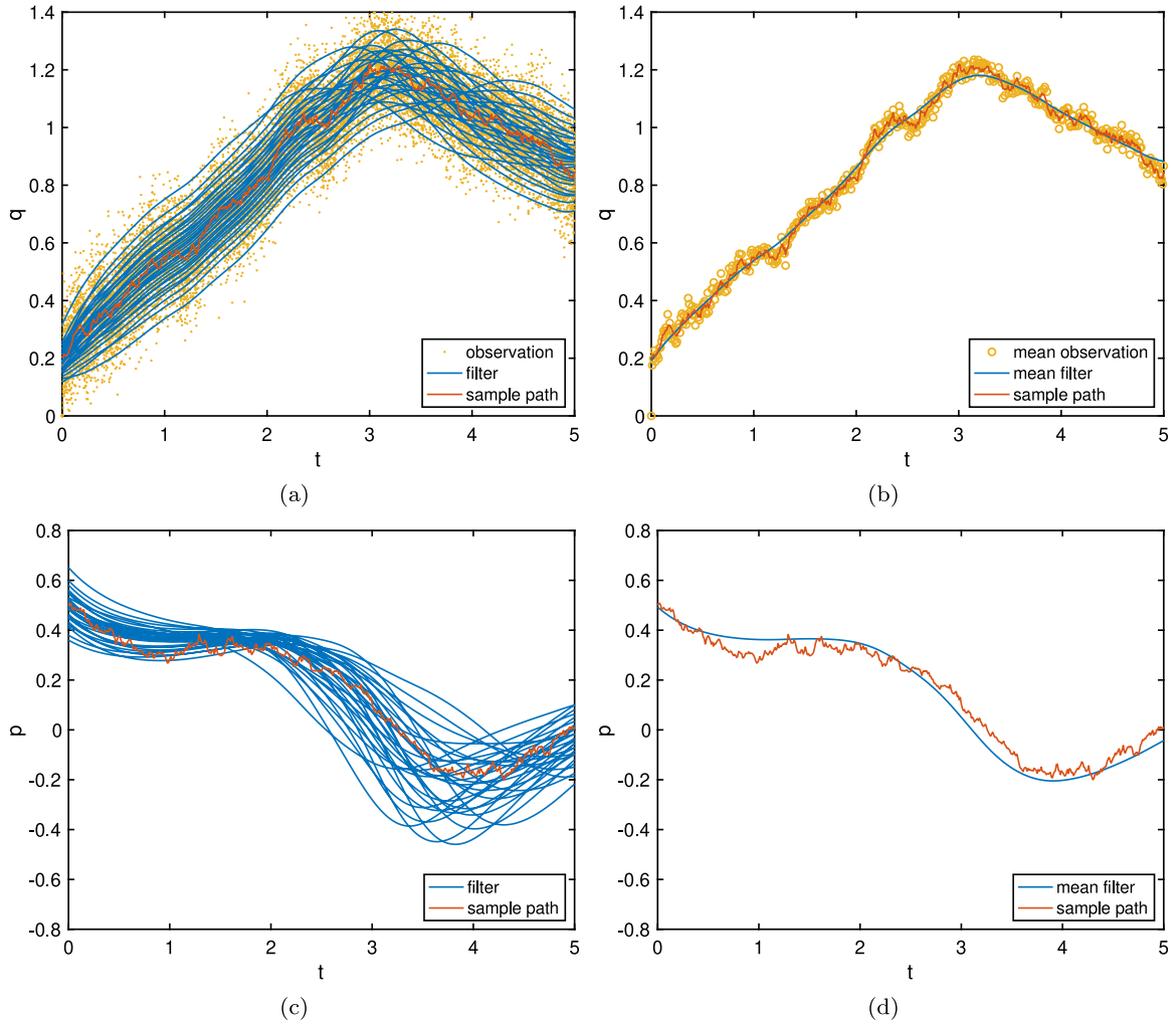
Fig. 3. Noisy observations of a single sample path, particle number $J = K = 30$. The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d). For this simulation $\beta = 4$ was used.

We demand that the Lagrangian be stationary under variations with respect to X_n^j, Z_n^j, λ_n^j and Λ_n^j. In addition we minimize L with respect to u_n^j. Assuming sufficient differentiability, we set derivatives of L with respect to these variables equal to zero. This approach is known to yield a variational integrator [34] that defines a symplectic map. In the context of optimal control, see for example [35,36].

Assuming the cost C is differentiable with respect to u at a (local) minimum, from $\partial L/\partial u_n^j = 0$ follows

$$u_n^j = B^T \lambda_n^j, \quad n = 1, \dots, N, \quad j = 1, \dots, J, \tag{13}$$

Enforcing $\partial L/\partial \lambda_n^j = 0$ and $\partial L/\partial \Lambda_n^j = 0$, and making use of (13), we obtain the filter relations (5)–(6):

$$X_{n+1}^j = X_n^j + \Delta t \, a(X_n^j) + \Delta t \, B B^T \lambda_{n+1}^j, \quad n = 0, \dots, N-1, \tag{14}$$

$$Z_n^j = h(X_n^j), \quad n = 0, \dots, N, \tag{15}$$

$$X_0^j = \xi_0^j. \tag{16}$$

From $\partial L/\partial Z_n^j = 0$, we obtain the definition

$$\Lambda_n^j = -\Delta t \, \frac{\beta}{2} \frac{\partial}{\partial Z_n^j} \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n), \quad n = 1, \dots, N, \quad j = 1, \dots, J, \tag{17}$$

where ζ_n and $\hat{\zeta}_n$ are given by (7).

Finally, from the condition $\partial L/\partial X_n^j = 0$, and making use of (17), we obtain the adjoint relations:

$$\lambda_1^j = \lambda_0^j - \Delta t \, \nabla a(X_0^j)^T \lambda_1^j, \tag{18}$$

$$\lambda_{n+1}^j = \lambda_n^j - \Delta t \, \nabla a(X_n^j)^T \lambda_{n+1}^j$$
$$+ \Delta t \frac{\beta}{2} \nabla h(X_n^j)^T \frac{\partial}{\partial Z_n^j} \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n), \quad n = 1, \dots, N-1, \tag{19}$$

$$\lambda_N^j = \Delta t \frac{\beta}{2} \nabla h(X_N^j)^T \frac{\partial}{\partial Z_N^j} \mathcal{W}_2^2(\zeta_N, \hat{\zeta}_N). \tag{20}$$

Numerical evaluation of the gradient of Wasserstein distance. To evaluate the second term on the right of (19), we represent $\nabla h(X_n^j)$ as a matrix of dimension $\ell \times d$. Denote the columns of this matrix by the vectors $\hat{h}_1, \dots, \hat{h}_d \in \mathbf{R}^\ell$. We approximate the Λ_n^j in (17) numerically using a finite difference formula:

$$\left(\nabla h(X_n^j)^T \frac{\partial \mathcal{W}_2^2}{\partial Z_n^j} \right)_i \approx \frac{1}{\varepsilon} \left[\mathcal{W}_2^2(\zeta_n^{(j,i)}(\varepsilon), \hat{\zeta}_n) - \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n) \right], \tag{21}$$

where

$$\zeta_n^{(j,i)}(\varepsilon) = \frac{1}{J} \left[\delta(z - (Z_n^j + \varepsilon \hat{h}_i)) + \sum_{k \neq j} \delta(z - Z_n^k) \right],$$

and ε can be chosen to be the square root of machine precision. Consequently, the second term on the right of (19) can be approximated using $d + 1$ evaluations of the Wasserstein distance.

The complete set of equations that define the filter can be expressed in terms of the variables X_n^j, Z_n^j and λ_n^j given by (14)–(16) and (18)–(20). Forward–backward sweep iteration proceeds by solving (14)–(16)
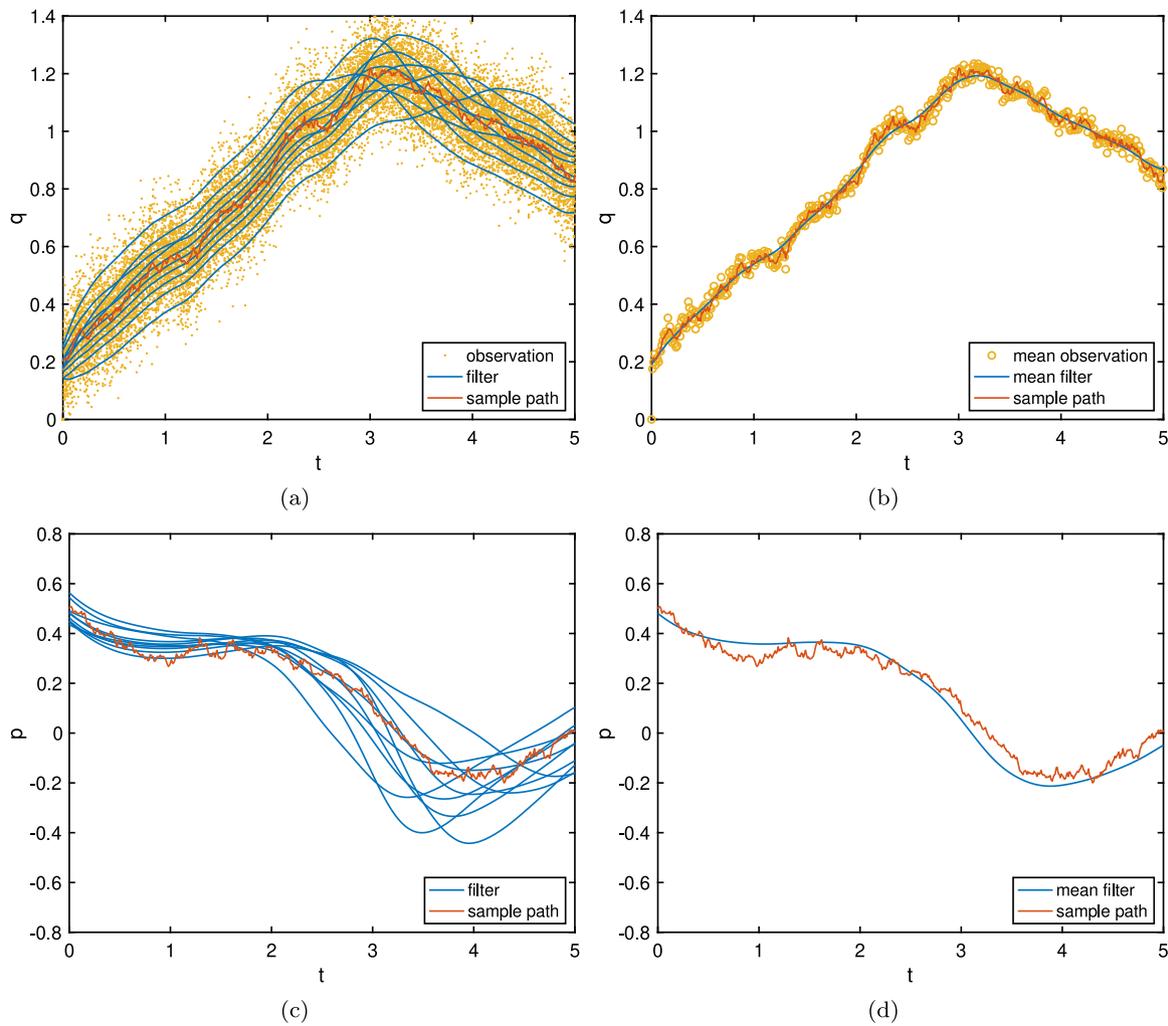
Fig. 4. Same as Fig. 3, but with particle number $J = 10$. The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d). For this simulation $\beta = 4$ was used in the cost function (8).

forward in time, followed by (18)–(20) backward in time, and repeating. However, such iteration is not convergent in general, especially for nonlinear dynamics.

Instead, the regularized forward–backward sweep method [37] proposed to augment the optimal control (13)

$$u_n^j = \frac{1}{1+\rho}\left[\lambda_n^j + \rho\left(\frac{X_{n+1}^j - X_n^j}{\Delta t} - a(X_n^j)\right)\right],$$

$$n = 1, \ldots, N, \quad j = 1, \ldots, J. \tag{22}$$

where $\rho > 0$ is the regularization parameter. Convergence of the resulting iteration for sufficiently large ρ is proven for continuous dynamics in [37]. The proof is confirmed for the discrete case with symplectic discretization in [37]. In practice the convergence can be greatly accelerated using Anderson acceleration with restart [38]. The Wasserstein distance is Lipschitz continuous with respect to the state variable [24,39]. Consequently, it satisfies the criterion for convergence of the regularized forward–backward sweep algorithm [37,40].

3. Numerical experiments

In this section, we study numerically the properties of the proposed filter for quantifying uncertainty in some simple differential equations. We first study the propagation of uncertainty in the initial condition of a deterministic differential equation, the Lorenz attractor model [41].

Subsequently, we consider stochastically forced motion in a double-well potential, for which the equilibrium distribution is bi-modal. We study both the case of a single sample path with noisy measurements and the case of multiple samples. In all numerical experiments we directly observe one dependent variable. Hence, the observations are partial ($\ell < d$), but the observation operator is linear (corresponding to a row of the identity).

For all experiments we computed the Wasserstein distance by solving the linear program (9), for which the complexity is unfavorable for large ensemble size [42,43]. Improved performance could possibly be achieved using the Sinkhorn iteration [33], especially given that the many Wasserstein distances that need to be computed via (21). Note that the transport paths in (21) are expected to be very similar, providing good starting values for the iterations. We have not investigated this further.

3.1. Uncertainty in initial condition: deterministic Lorenz 63 model

In this section we study the behavior of the particle filter to approximate a probability measure relaxing onto the attractor of the Lorenz 63 system [41]. The invariant measure of the Lorenz system is a Sinai–Ruelle–Bowen (SRB) measure, which is a kind of invariant measure [44], supported on a strange attractor of fractal dimension. The dynamics is deterministic, but we introduce uncertainty in the initial conditions by drawing an ensemble from a normal distribution.
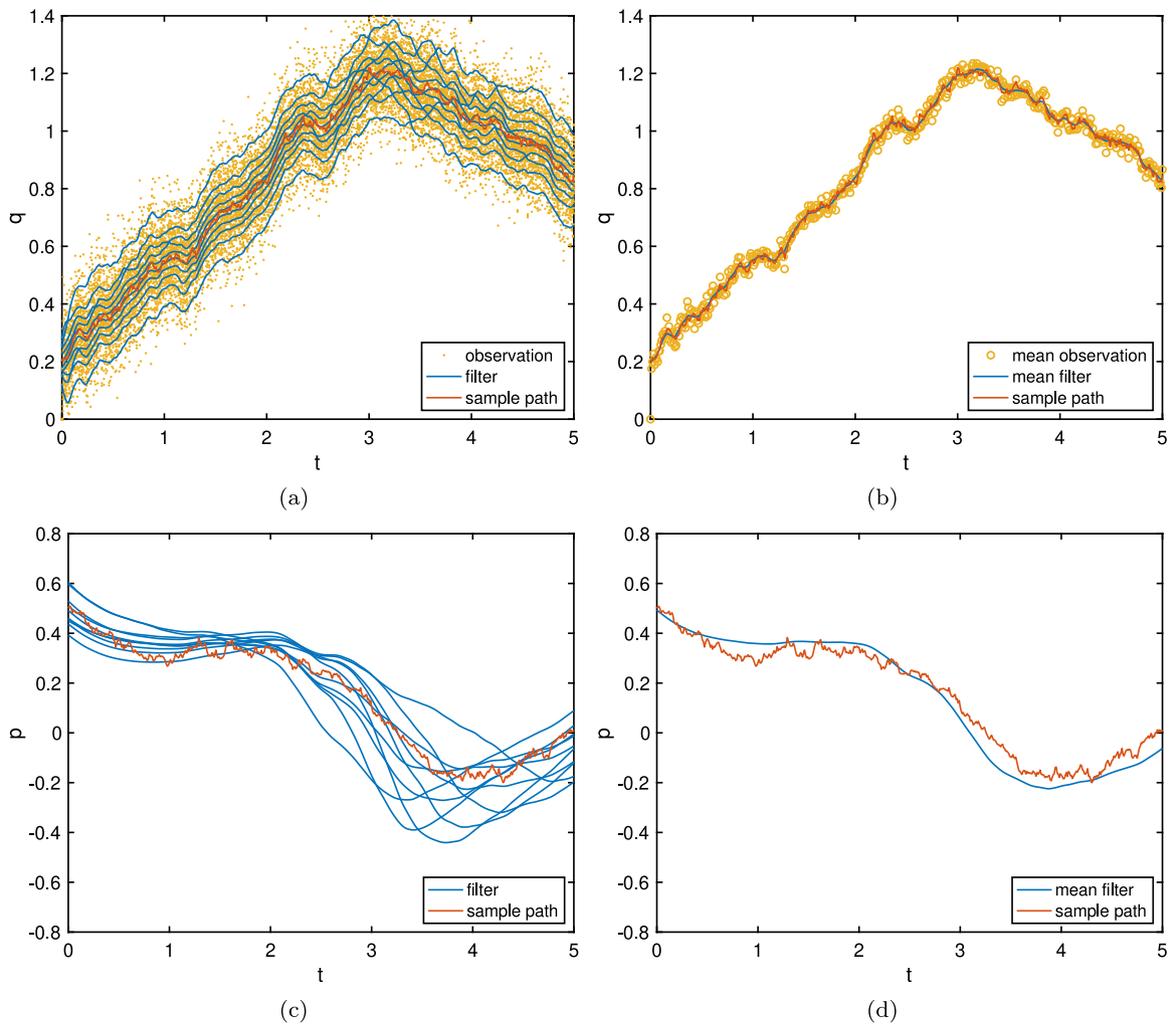
Fig. 5. Same as Fig. 4, but computed with $\beta = 200$ in the cost function (8). The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d).
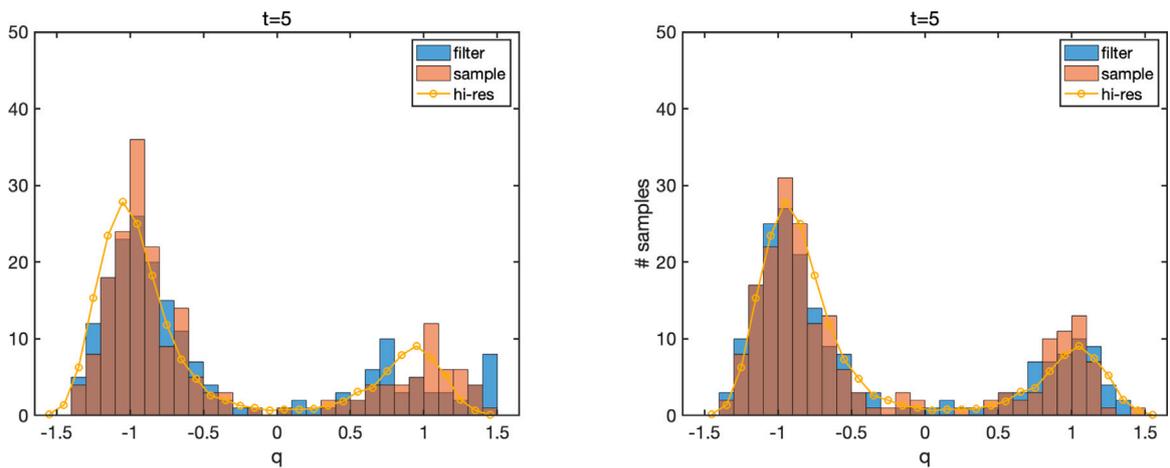
Fig. 6. Histograms of coordinate $q(t)$ of the sample ensemble (red) and filter (blue) at time $t = 5$ for weight parameters $\beta = 4$ (left) and $\beta = 200$ (right) for a 200-member ensemble. The yellow curve indicates the expected bin size based on a high-resolution sample of 20 000 members. The figure shows that the proposed method is effective at approximating a bi-modal probability density function.
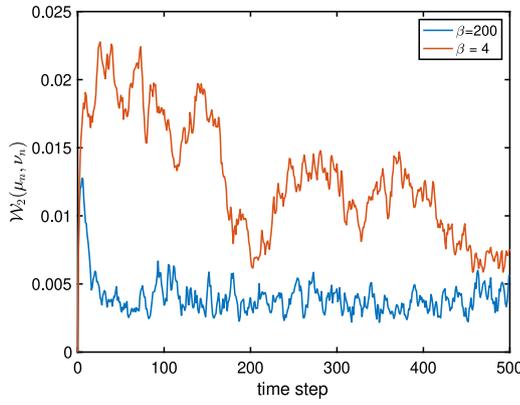
Fig. 7. Time-evolution of Wasserstein distance between the particle ensemble $\nu_n(q,p)$ and the sample ensemble $\mu_n(q,p)$ for $\beta = 4$ and $\beta = 200$.

We compare the particle filter to the SIR and another popular ensemble data assimilation: the ensemble Kalman filter (EnKF, [45]) to study the potential advantage of optimizing with respect to mismatch in the Wasserstein metric. The EnKF method focuses on properties of evolving Gaussian distributions (approximating the mean and covariance matrices), which are smooth absolutely continuous measures. The Wasserstein metric does not require differentiability of the evolving density. Consequently, it is useful for comparing measures that evolve on a strange attractor. Applying Euler's method, the discrete Lorenz system is given by

$$\hat{x}_{n+1} = \hat{x}_n + \Delta t\, c_1 (\hat{y}_n - \hat{x}_n), \tag{23}$$

$$\hat{y}_{n+1} = \hat{y}_n + \Delta t\, (\hat{x}_n (c_2 - \hat{z}_n) - \hat{y}_n), \tag{24}$$

$$\hat{z}_{n+1} = \hat{z}_n + \Delta t\, (\hat{x}_n \hat{y}_n - c_3 \hat{z}_n), \tag{25}$$

with the parameters $c_1 = 10, c_2 = 28, c_3 = 8/3$ as originally studied by Lorenz. We employ step-size $\Delta t = 0.001$.

To generate observations, we simulate an ensemble of $K = 100$ trajectories over the time interval $t \in [0,6]$, with initial conditions $X_0^k = (x_0^k, y_0^k, z_0^k)$ drawn from

$$x_0^k \sim \mathcal{N}(1, 0.5^2), \quad y_0^k \sim \mathcal{N}(-1, 0.5^2), \quad z_0^k \sim \mathcal{N}(25, 0.5^2).$$

This initial condition was chosen with a small variance but rapidly spreading ensemble that splits across the two lobes of the Lorenz attractor. In the experiment, as an example to illustrate our method, we assume the (partial) observable is the x-component

$$\hat{Z}_n^k = \hat{x}_n^k, \qquad k = 1, \dots, K, \quad n = 1, \dots, N.$$

and the control is applied only to the y-component. The particle dynamics satisfy

$$x_{n+1} = x_n + \Delta t\, c_1 (y_n - x_n), \tag{26}$$

$$y_{n+1} = y_n + \Delta t\, (x_n (c_2 - z_n) - y_n) + \Delta t u_{n+1}, \tag{27}$$

$$z_{n+1} = z_n + \Delta t\, (x_n y_n - c_3 z_n). \tag{28}$$

We select a particle filter ensemble size $J = 100$. For the Wasserstein metric (9)–(10) we choose $M = I$, and in the cost function $\beta = 60$. As for the ensemble Kalman filter algorithm (EnKF), ensemble size and initial error are same with particle filter and the observation is non-perturbed. The Same particle number is in SIR. Since the initial condition in our implemented example is unknown in state dynamic, all particles update their state through the filter at initial time.

The states of the three methods are shown in Fig. 1. Even it is not very clear to see the difference between the three approaches related to the state, it still can be seen that particles (yellow circles) under Wasserstein particle filter appear similar probability with ensemble of

trajectories in the right lobe of the attractor. This result is the best than the other two methods because particles under EnKF method and SIR method appear low probability with ensemble of trajectories.

The better approximation of the evolving measure by the Wasserstein particle filter is confirmed in Fig. 2, where we compare the Wasserstein distances between the sample ensemble μ_n based on the full states $\{\hat{X}_n^k\}$ and the filter ensembles $\nu_n(X)$ computed using the Wasserstein particle filter, EnKF and SIR. We see that Wasserstein distances of the empirical measures to that of the sample ensemble $\mathcal{W}_2(\nu_0, \mu_0) < 1$ for both filters, the final distance $\mathcal{W}_2(\nu_N, \mu_N)$ is approximately 0.5 for the particle filter and 5 for the EnKF.

3.2. Noisy observations: a randomly forced ODE

For the experiments in this and the next section we consider stochastically forced motion in a double-well potential:

$$dq = p\, dt + \sigma_q\, dW_q,$$

$$dp = (q - q^3 - rp)\, dt + \sigma_q\, dW_p,$$

where $r > 0$ is a damping parameter. For the numerical experiments we choose $r = 1, \sigma_q = \sigma_p = 0.1$. Probability distributions transported by this system converge to a bi-modal equilibrium state with peaks centered at the stable equilibria $(q^*, p^*) = (\pm 1, 0)$ of the drift vector field.

To generate samples of this system we discretize using the Euler–Maruyama method

$$\hat{q}_{n+1} = \hat{q}_n + \Delta t\, \hat{p}_n + \sigma_q \Delta W_{q,n}, \tag{29}$$

$$\hat{p}_{n+1} = \hat{p}_n + \Delta t\, (\hat{q}_n - \hat{q}_n^3 - r\hat{p}_n) + \sigma_p \Delta W_{p,n}, \tag{30}$$

where $\Delta W_{p,n}, \Delta W_{q,n} \sim \mathcal{N}(0, \Delta t)$ are independent and normally distributed. Noisy observations of the variable q are obtained from

$$\hat{Z}_n = \hat{q}_n + \eta_n, \quad n = 0, 1, \dots, N, \tag{31}$$

where $\eta_n \sim \mathcal{N}(0, \sigma_n^2)$.

For the particle filter, the motion of the jth controlled particle is given by

$$q_{n+1}^j = q_n^j + \Delta t\, p_n^j + \Delta t\, u_{q,n+1}^j, \tag{32}$$

$$p_{n+1}^j = p_n^j - \Delta t\, ((q_n^j)^3 - q_n^j + rp_n^j) + \Delta t\, u_{p,n+1}^j, \tag{33}$$

for $j = 1, \dots, J$, and the observation function applied to the jth particle yields

$$Z_n^j = q_n^j, \quad n = 0, \dots, N. \tag{34}$$

For all experiments we choose step size $\Delta t = 0.01$. In each computation, the regularization parameter ρ in (22) was experimentally determined as small as possible to still observe convergence of the forward–backward sweep iteration.

We first investigate the scenario of noisy observations of a single sample path of (29)–(30). We choose initial conditions $\hat{q}_0 = 0.2, \hat{p}_0 = 0.5$ and integrate to time $t = 5$. The particle filter positions were sampled from initial distribution $\nu_0(q, p)$ given by the product measure

$$q_0 \sim \mathcal{N}(0.2, 0.04^2), \qquad p_0 \sim \mathcal{N}(0.5, 0.06^2). \tag{35}$$

Observational noise was generated with standard deviation $\sigma_n = 0.1$, and for each time step we sample $K = 30$ noisy observations. Wasserstein metric is given by (9)–(10) with $M = R^{-1}$ and $R = \sigma_n^2 I$, where I is the identity matrix.

We apply the particle filter with particle number $J = 30$ and $J = 10$. The results for $J = 30$ are shown in Fig. 3. For these simulations, we chose $\beta = 4$ in the cost function (8). The red curves in 3 show the sample path $\hat{q}_n$ (upper plots) and $\hat{p}_n$ (lower plots). The noisy measurement data $\{\hat{Z}_n^k\}_{k=1}^K$ is plotted in yellow in 3(a), and the sample mean is plotted as yellow circles in 3(b). The particle trajectories are plotted as blue curves in Fig. 3(a) and (c). The particle mean trajectory is plotted in blue in Fig. 3(b) and (d). As expected, the pdf of the
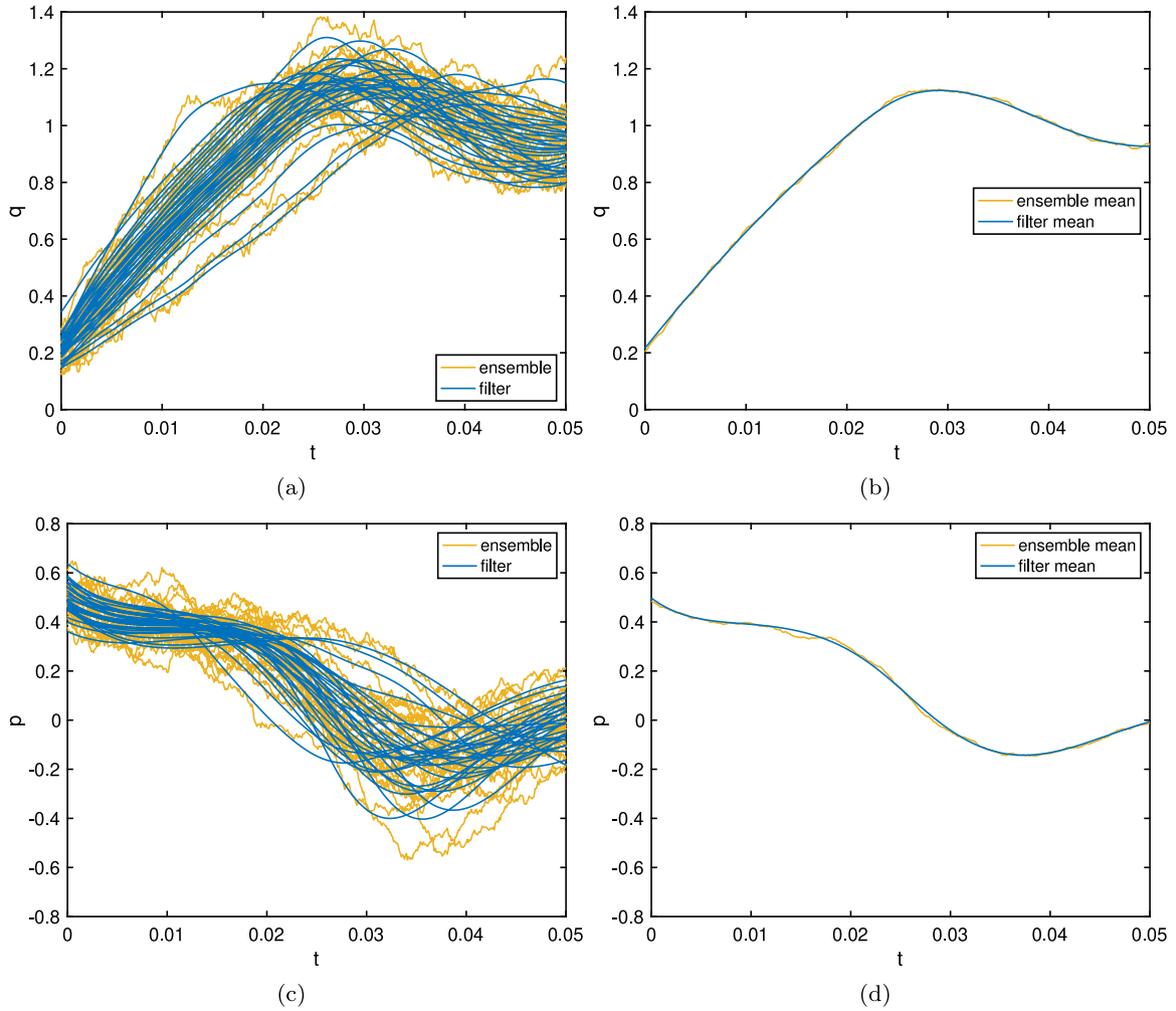
Fig. 8. Multiple sample paths of an SDE. An ensemble of $K = 30$ sample paths of the system (29)–(30) are plotted as yellow curves in (a) (q-component) and (c) (p-component). The corresponding particle filter paths ($J = 30$) are plotted as blue curves in (a) and (c). The ensemble means and particle filter means are compared in (b) and (d). For these simulations, $\beta = 4$.

observed q-component is approximately normally distributed about the sample path. This is not the case for the unobserved p-component, for which the marginal pdf is time-dependent. We see that the mean particle motion is much smoother than the sample path. It also appears as if the q-component, which is directly observed, is better estimated than the p-component.

In Fig. 4 we repeat the above experiment, but for a smaller particle size $J = 10$ for the particle filter. The conclusions are similar. The particle mean trajectory is of similar accuracy to the higher resolution simulation in Fig. 3.

The parameter β in the cost function (8) determines the relative weight of the observations compared of the cost of controlling the particle motion. In Fig. 5 we choose a much larger value $\beta = 200$ and repeat the experiment. We observe that the particle filter paths are much less smooth in the q-component in Fig. 5(a) and that the particle ensemble mean closely follows the sample path in Fig. 5(b). There is no noticeable improvement in the trajectories of the unobserved component p.

3.3. Multiple sample paths of a stochastic system

In this section we generate observations by simulating an ensemble of sample paths of the stochastic double well potential (29)–(30). All parameters are identical to those in the previous section unless stated otherwise.

We first study the approximation of the bimodal distribution at high resolution. For this example, we choose a deterministic (Dirac distribution) initial condition $q_0 = -1$, $p_0 = 0$ for both the samples and the filter particles. We generated a large number $K = 20000$ of sample paths to approximate the time evolving pdf, which is exhibited at time $t = 5$ by the yellow curve in Fig. 6. We then generated observations using an ensemble of size $K = 200$ without noise (i.e. $\sigma_n = 0$ in (31)) and applied the particle filter (14)–(20) with $J = 200$ particles. Histograms of the samples and particle filter pdfs are shown in Fig. 6 for parameter values $\beta = 4$ (left plot) and $\beta = 200$ (right plot). The bi-modality of the pdf is clearly noticeable, and the approximation more closely matches the observations for $\beta = 200$ as expected.

Fig. 7 shows the time evolution of the Wasserstein distance $\mathcal{W}_2$ over the full state empirical measures $\nu_n(q, p)$ $\mu_n(q, p)$ for $\beta = 4$ and $\beta = 200$. For $\beta = 200$, the Wasserstein distance is bounded below $\mathcal{W}_2 < 0.007$ for most of the interval. For $\beta = 4$ the distance is somewhat greater at around $\mathcal{W}_2 < 0.02$ but decreases over time.

In Figs. 8 and 9 we compare the particle filter approximation of an evolving measure with sample path ensemble size $K = 30$ and particle numbers $J = 30$ and $J = 10$. For both simulations we draw initial distributions for both sample paths and particles from (35). The same sample paths are used in both figures. We use $\beta = 4$ and $M = (0.1)^{-2} I$ in (10) (consistent with experiments in the previous section). We see that the sample path means of both the q- and p-components are well approximated.
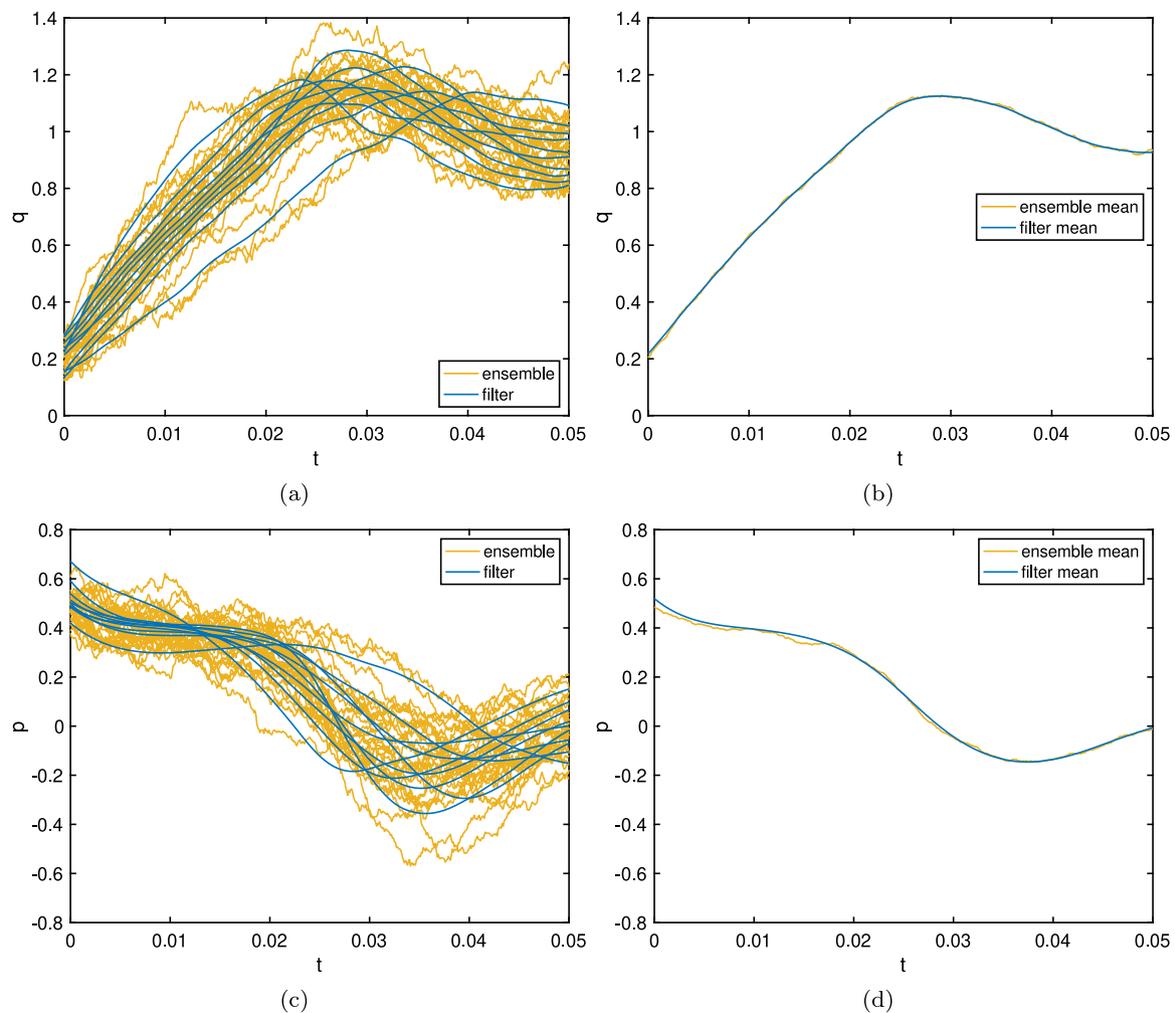
Fig. 9. Same as Fig. 9, but with $J = 10$. An ensemble of sample paths of the system (29)–(30) are plotted as yellow curves in (a) (q-component) and (c) (p-component). The corresponding particle filter paths with particle number $J = 20$ are plotted as blue curves in (a) and (c). The ensemble means and particle filter means are compared in (b) and (d). For these simulations, $\beta = 4$.

4. Conclusion

In this paper, we construct a particle filter in the form of an optimal control that minimizes mismatch in the Wasserstein distance on observation space. The particle states are evolved deterministically which is easy to apply. This new data assimilation algorithm avoid weight degeneracy, which is a main issue in conventional particle filter problems. Comparing with the feedback particle filter proposed in [14,15], we obtain the optimal control without calculating the posterior distribution of the particle states. Numerical examples show that the Wasserstein distance between the empirical measure on the whole state space is well bounded over the assimilation window. We compared scenarios with (i) deterministic (chaotic) dynamics with uncertainty in initial conditions, (ii) a single sample path of an SDE with multiple uncertain observations, and (iii) multiple sample paths of an SDE with accurate (partial) observations. The method was shown to recover bi-modal probability measures, compare favorably to the ensemble Kalman filter for a SRB measure, and accurately reproduce the sample path mean for latter scenario. The numerical computation is expensive due to the Wasserstein distance calculation and the results are suboptimal, as is a topic for future research. And the sinkhorn Iteration could be applied to computed Wasserstein distance in the further to improve performance with large ensemble size.

CRediT authorship contribution statement

Xin Liu: Design the method and modified the method, Mainly did the whole experiments and finished the original paper. **Jason Frank:** Proposed the conceptualization on applying the optimal control in data assimilation, Design the method and modified the method, Helped to modified the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] B. Wang, X. Zou, J. Zhu, Data assimilation and its applications, Proc. Natl. Acad. Sci. 97 (21) (2000) 11143–11144.

[2] F. Bouttier, P. Courtier, Data assimilation concepts and methods March 1999, in: Meteorological Training Course Lecture Series, vol. 718, ECMWF, 2002, p. 59.

[3] K. Law, A. Stuart, K. Zygalakis, Data Assimilation, Vol. 214, Springer, Cham, Switzerland, 2015.

[4] B.K.W. Lahoz, R. Menard, Data Assimilation, Springer, 2010.

[5] N.J. Gordon, D.J. Salmond, A.F. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proc. F (Radar Signal Process.) 140 (2) (1993) 107–113.

[6] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process. 50 (2) (2002) 174–188.

[7] A. Doucet, A.M. Johansen, A tutorial on particle filtering and smoothing: Fifteen years later, Handb. Nonlinear Filter. 12 (656–704) (2009) 3.

[8] A. Budhiraja, L. Chen, C. Lee, A survey of numerical methods for nonlinear filtering problems, Physica D 230 (1–2) (2007) 27–36.

[9] M. Speekenbrink, A tutorial on particle filters, J. Math. Psych. 73 (2016) 140–152.

[10] H.R. Künsch, Particle filters, Bernoulli 19 (4) (2013) 1391–1403.

[11] H. Zhou, Z. Deng, Y. Xia, M. Fu, A new sampling method in particle filter based on Pearson correlation coefficient, Neurocomputing 216 (2016) 208–215.

[12] M. Zhu, P.J. Van Leeuwen, J. Amezcua, Implicit equal-weights particle filter, Q. J. R. Meteorol. Soc. 142 (698) (2016) 1904–1919.

[13] M. Frei, H.R. Künsch, Bridging the ensemble Kalman and particle filters, Biometrika 100 (4) (2013) 781–800.

[14] T. Yang, P.G. Mehta, S.P. Meyn, A mean-field control-oriented approach to particle filtering, in: Proceedings of the 2011 American Control Conference, IEEE, 2011, pp. 2037–2043.

[15] T. Yang, P.G. Mehta, S.P. Meyn, Feedback particle filter, IEEE Trans. Automat. Control 58 (10) (2013) 2465–2480.

[16] S. Kolouri, S.R. Park, M. Thorpe, D. Slepcev, G.K. Rohde, Optimal mass transport: Signal processing and machine-learning applications, IEEE Signal Process. Mag. 34 (4) (2017) 43–59.

[17] C. Villani, Topics in Optimal Transportation, in: Graduate Studies in Mathematics, American Mathematical Soc., 2003, 58.

[18] S. Reich, A nonparametric ensemble transform method for Bayesian inference, SIAM J. Sci. Comput. 35 (4) (2013) A2013–A2024.

[19] S.K. Tamang, A. Ebtehaj, P.J. Van Leeuwen, D. Zou, G. Lerman, Ensemble Riemannian data assimilation over the Wasserstein space, Nonlinear Process. Geophys. 28 (3) (2021) 295–309.

[20] S.K. Tamang, A. Ebtehaj, D. Zou, G. Lerman, Regularized variational data assimilation for bias treatment using the Wasserstein metric, Q. J. R. Meteorol. Soc. 146 (730) (2020) 2332–2346.

[21] S. Shafieezadeh Abadeh, V.A. Nguyen, D. Kuhn, P.M. Mohajerin Esfahani, Wasserstein distributionally robust Kalman filtering, Adv. Neural Inf. Process. Syst. 31 (2018).

[22] N. Feyeux, A. Vidard, M. Nodet, Optimal transport for variational data assimilation, Nonlinear Process. Geophys. 25 (1) (2018) 55–66.

[23] L. Rüschendorf, The Wasserstein distance and approximation theorems, Probab. Theory Related Fields 70 (1) (1985) 117–129.

[24] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, 2017, arXiv preprint arXiv:1701.07875.

[25] E. del Barrio, E. Giné, C. Matrán, Central limit theorems for the Wasserstein distance between the empirical and the true distributions, Ann. Probab. 27 (2) (1999) 1009–1071.

[26] J. Frank, S. Zhuk, Symplectic Möbius integrators for LQ optimal control problems, in: Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on, IEEE, 2014, pp. 6377–6382.

[27] C. González-Tokman, B.R. Hunt, Ensemble data assimilation for hyperbolic systems, Physica D 243 (1) (2013) 128–142.

[28] B. De Leeuw, S. Dubinkina, J. Frank, A. Steyer, X. Tu, E.V. Vleck, Projected shadowing-based data assimilation, SIAM J. Appl. Dyn. Syst. 17 (4) (2018) 2446–2477.

[29] A. Trevisan, M. D'Isidoro, O. Talagrand, Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension, Q. J. R. Meteorol. Soc.: A J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr. 136 (647) (2010) 487–496.

[30] L. Palatella, A. Carrassi, A. Trevisan, Lyapunov vectors and assimilation in the unstable subspace: theory and applications, J. Phys. A 46 (25) (2013) 254020.

[31] D.P. Bertsekas, Linear Network Optimization: Algorithms and Codes, MIT Press, 1991.

[32] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, Network Flows: Theory Algorithms and Applications, Prentice Hall, 1993.

[33] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems, NIPS 26, 2013, pp. 2292–2300.

[34] J.E. Marsden, M. West, Discrete mechanics and variational integrators, Acta Numer. 10 (2001) 357–514.

[35] S. Ober-Blöbaum, Discrete mechanics and optimal control (Ph.D. thesis), Universtät Paderborn, 2004.

[36] J.M. Sanz-Serna, Symplectic runge–kutta schemes for adjoint equations, automatic differentiation, optimal control, and more, SIAM Rev. 58 (1) (2016) 3–33.

[37] X. Liu, J. Frank, Symplectic Runge–Kutta discretization of a regularized forward-backward sweep iteration for optimal control problems, J. Comput. Appl. Math. 383 (2021) 113133.

[38] N.C. Henderson, R. Varadhan, Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms, J. Comput. Graph. Statist. 28 (4) (2019) 834–846.

[39] W. E, J. Han, Q. Li, A mean-field optimal control formulation of deep learning, Res. Math. Sci. 6 (10) (2018).

[40] Q. Li, L. Chen, C. Tai, E. Weinan, Maximum principle based algorithms for deep learning, J. Mach. Learn. Res. 18 (1) (2018) 1–29.

[41] E.N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (2) (1963) 130–141.

[42] C. Tameling, M. Sommerfeld, A. Munk, Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications, Ann. Appl. Probab. 29 (5) (2019) 2744–2781.

[43] A.M. Oberman, Y. Ruan, An efficient linear programming method for optimal transportation, 2015, arXiv preprint arXiv:1509.03668.

[44] L.-S. Young, What are SRB measures, and which dynamical systems have them? J. Stat. Phys. 108 (5) (2002) 733–754.

[45] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, Springer, 2009.

Dr. Xin Liu, Phd graduated from Utrecht University in Netherlands. Research mostly focus on complex system, data assimilation, optimal control.

Prof. Jason Frank, head of the department of Mathematics, Utrecht University, Chair of Numerical Analysis, Mathematical institute Utrecht University, Editorial Board SIAM Journal on Scientific Computing, Board Member, National Mathematics Cluster, Board Member, Werkgemeenschap scientific Computing, Dutch scientific computing Steering committee member, complex systems studies, Utrecht University.

Main research on Numerical method for time dependent ODE and PDE. Data assimilation and parameter estimation Stochastic-dynamic methods for control/ correction of invariant densities and modeling of underresolved process Complex system science.