Toward consistent nonlinear filtering and smoothing via measure transport

Ricardo Baptista[1]

Joint work with Alessio Spantini[2], Youssef Marzouk[2], Max Ramgraber[2], Mathieu Le Provost[3]

[1]Computing + Mathematical Sciences
California Institute of Technology

[2]Center for Computational Science and Engineering
Massachusetts Institute of Technology

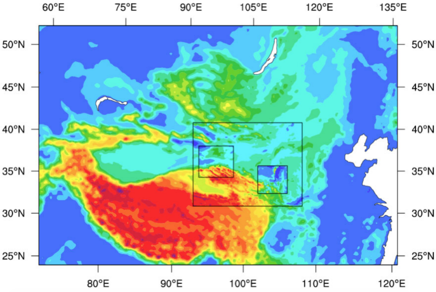
[3]Department of Computer Science
Long Island University

NCAR CISL Seminar

July 6, 2023

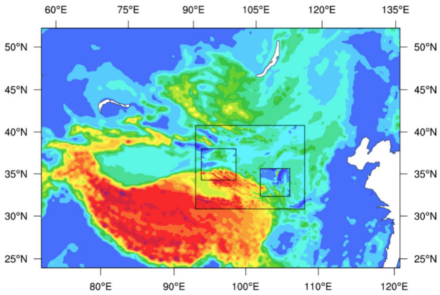
Sequential inference is ubiquitous

- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.

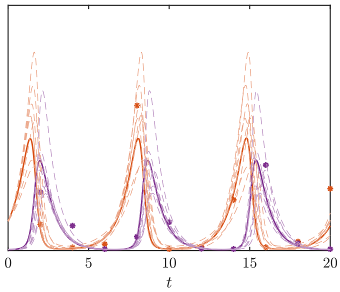


Wind forecast

- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



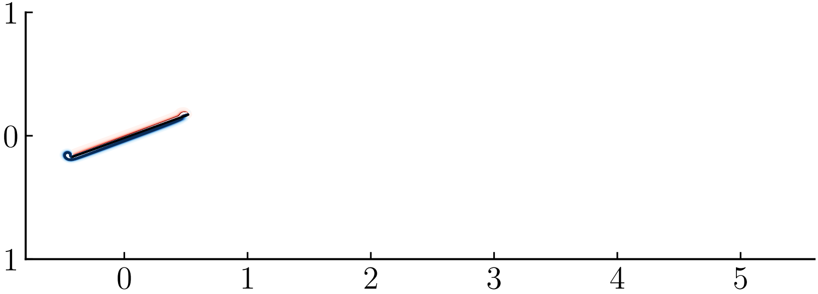
Wind forecast



Epidemiological forecast

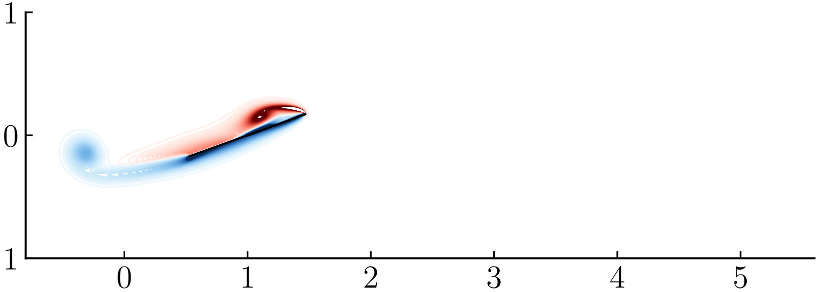
Sequential inference is ubiquitous

- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

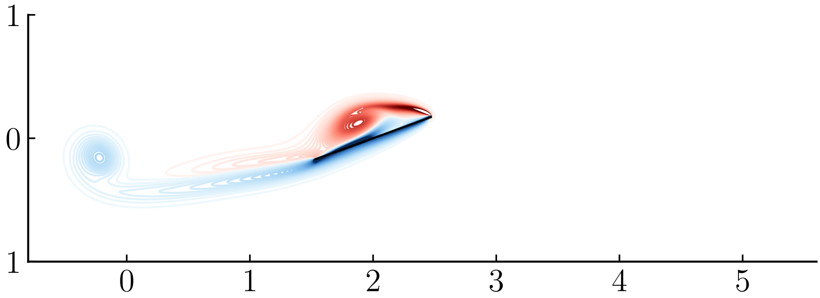
- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

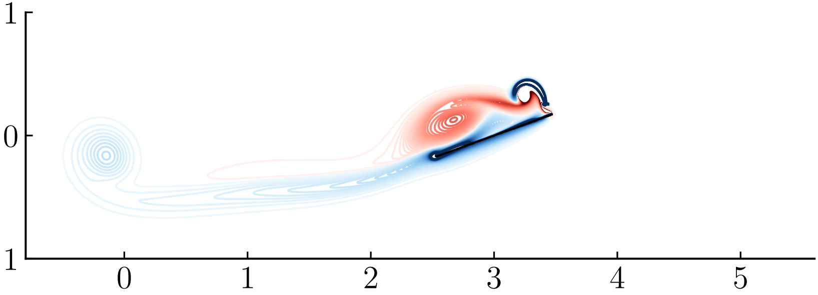
Sequential inference is ubiquitous

- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

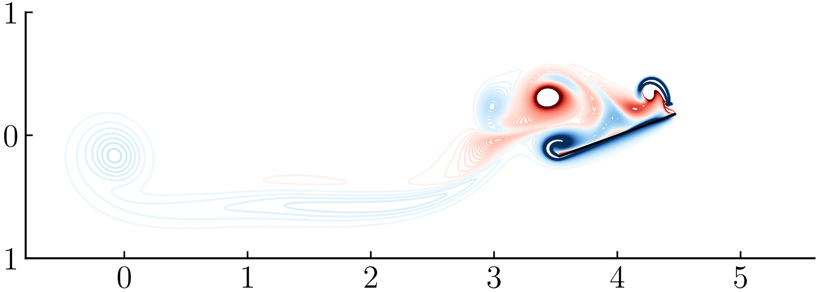
- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

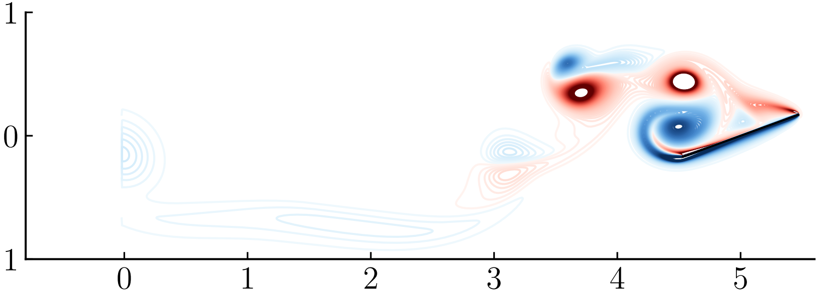
Sequential inference is ubiquitous

- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

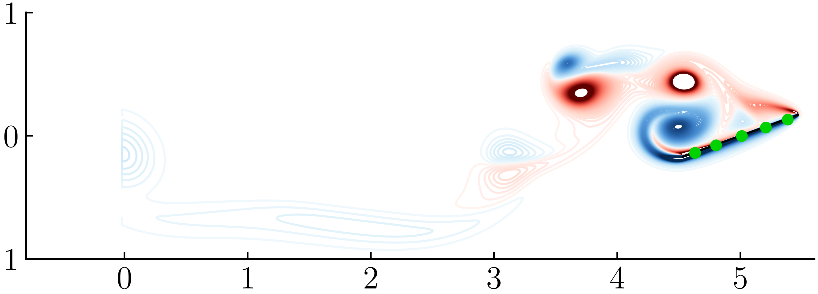
- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



Vortex shedding around an aircraft wing

Sequential inference is ubiquitous

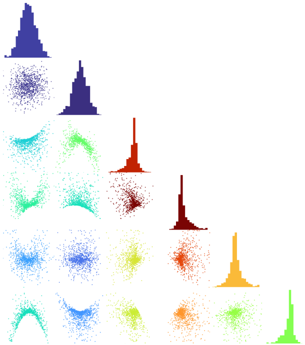
- **Goal**: Sequential state estimation in a Bayesian setting
- **Applications**: Weather prediction, oceanography, finance, population dynamics, pharmacology, robotics, aerodynamics, etc.



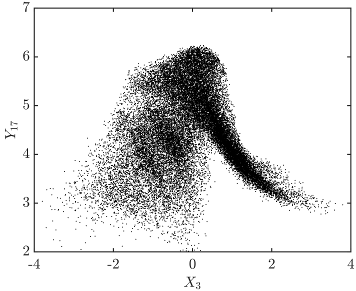
Vortex shedding around an aircraft wing

Non-Gaussianity is ubiquitous

- Non-Gaussianity can include multi-modality and/or tail-heaviness



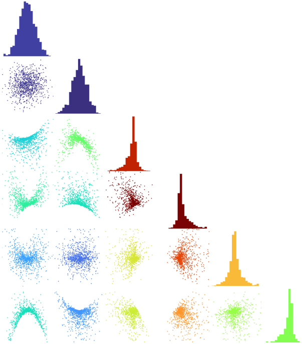
Lorenz-63 smoothing ensemble



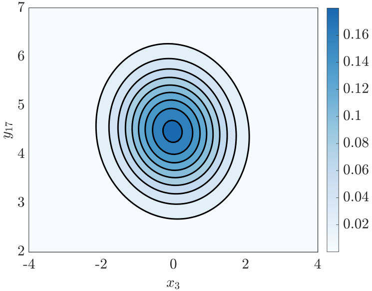
$(\mathbf{X}, \mathbf{Y})$ distribution in additive manufacturing model [B et al., 2022]

Non-Gaussianity is ubiquitous

- Non-Gaussianity can include multi-modality and/or tail-heaviness



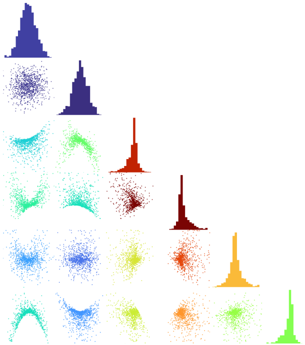
Lorenz-63 smoothing ensemble



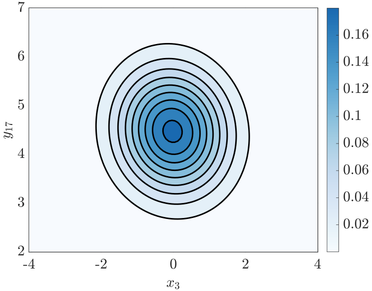
$(\mathbf{X}, \mathbf{Y})$ distribution in additive manufacturing model [B et al., 2022]

Non-Gaussianity is ubiquitous

- Non-Gaussianity can include multi-modality and/or tail-heaviness



Lorenz-63 smoothing ensemble

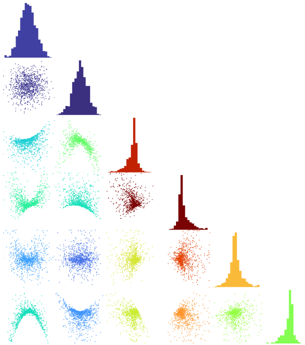


$(\mathbf{X}, \mathbf{Y})$ distribution in additive manufacturing model [B et al., 2022]

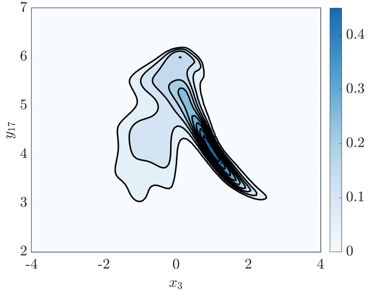
Takeaway: Gaussian approximations under-predict data informativeness

Non-Gaussianity is ubiquitous

- Non-Gaussianity can include multi-modality and/or tail-heaviness



Lorenz-63 smoothing ensemble



$(\mathbf{X}, \mathbf{Y})$ distribution in additive manufacturing model [B et al., 2022]

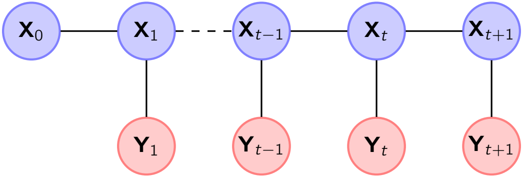
Takeaway: Gaussian approximations under-predict data informativeness

Goal: Develop consistent inference methods for non-Gaussian problems

Sequential Bayesian inference

State-space models

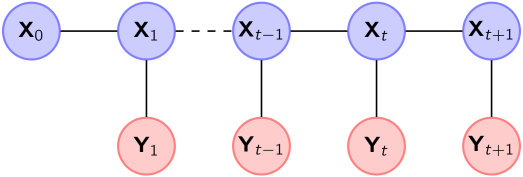
- States follow model dynamics $\pi_{\mathbf{X}_t|\mathbf{X}_{t-1}}$
- Observations follow likelihood function $\pi_{\mathbf{Y}_t|\mathbf{X}_t}$



Goal: Recursively sample distributions $\pi_{\mathbf{X}_t|\mathbf{y}_1^*,\dots,\mathbf{y}_t^*}$ or $\pi_{\mathbf{X}_{1:t}|\mathbf{y}_1^*,\dots,\mathbf{y}_t^*}$

State-space models

- States follow model dynamics $\pi_{\mathbf{X}_t|\mathbf{X}_{t-1}}$
- Observations follow likelihood function $\pi_{\mathbf{Y}_t|\mathbf{X}_t}$



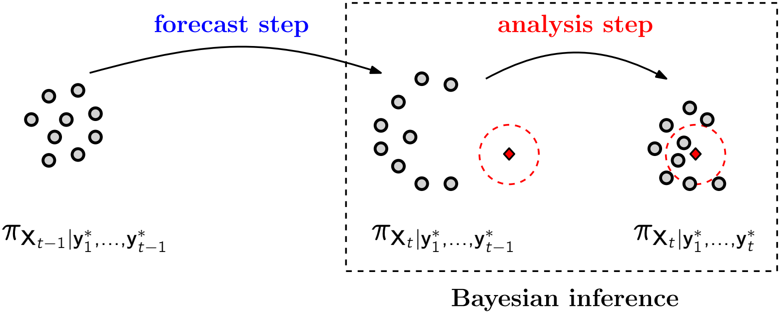
Goal: Recursively sample distributions $\pi_{\mathbf{X}_t|\mathbf{y}_1^*,\dots,\mathbf{y}_t^*}$ or $\pi_{\mathbf{X}_{1:t}|\mathbf{y}_1^*,\dots,\mathbf{y}_t^*}$

Common challenges leading to non-Gaussianity

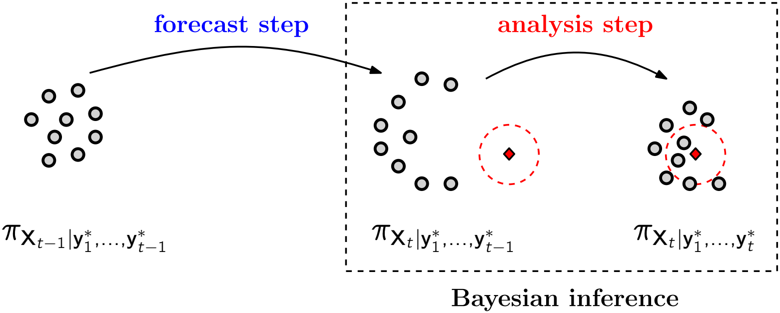
- Nonlinear dynamical models or observation operators
- Sparse observations in space and time

Ensemble filtering and smoothing

Approach: Approximate distributions using limited samples

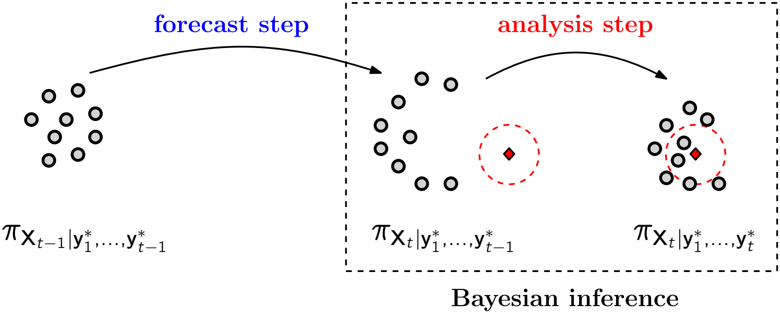


Approach: Approximate distributions using limited samples



Ensemble filtering and smoothing

Approach: Approximate distributions using limited samples

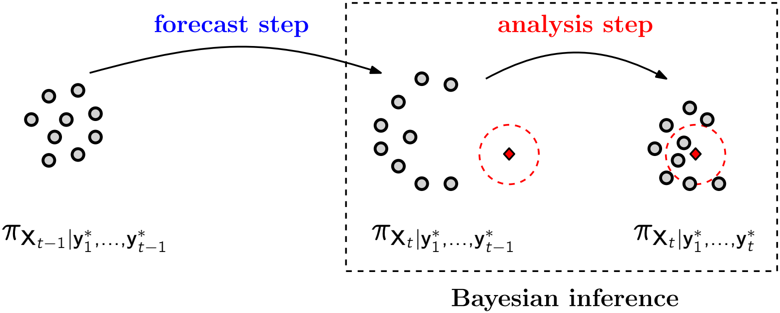


Ensemble Kalman filters and smoothers

- Analysis step updates particles by estimating a linear transformation
- Inconsistent for capturing Bayesian solution

Ensemble filtering and smoothing

Approach: Approximate distributions using limited samples



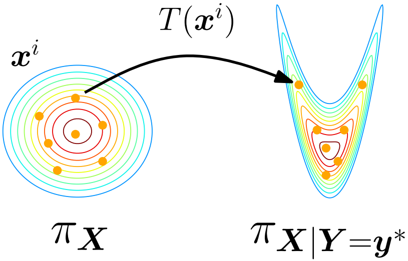
Ensemble Kalman filters and smoothers

- Analysis step updates particles by estimating a linear transformation
- Inconsistent for capturing Bayesian solution

Goal: Perform analysis consistently and robustly in non-Gaussian settings

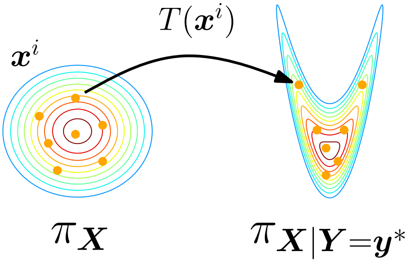
Prior-to-posterior transformations

Idea: Find map T that take samples from prior $\pi_{\mathbf{X}}$ to posterior $\pi_{\mathbf{X}|\mathbf{Y}}$



Prior-to-posterior transformations

Idea: Find map T that take samples from prior $\pi_{\mathbf{X}}$ to posterior $\pi_{\mathbf{X}|\mathbf{Y}}$

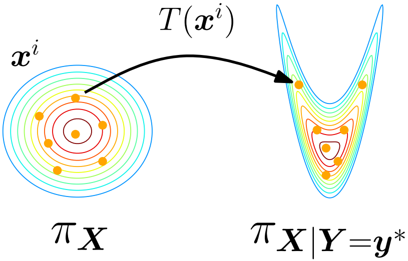


Plan for this talk:

1. Maps for filtering $\mathbf{X} = \mathbf{X}_t$?

Prior-to-posterior transformations

Idea: Find map T that take samples from prior $\pi_{\mathbf{X}}$ to posterior $\pi_{\mathbf{X}|\mathbf{Y}}$

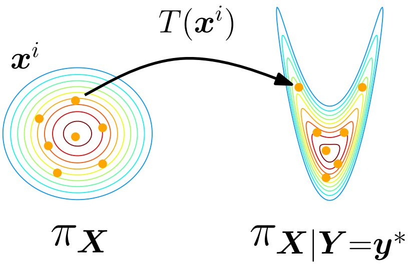


Plan for this talk:

1. Maps for filtering $\mathbf{X} = \mathbf{X}_t$?
2. Maps for smoothing $\mathbf{X} = \mathbf{X}_{1:t}$?

Prior-to-posterior transformations

Idea: Find map T that take samples from prior $\pi_{\mathbf{X}}$ to posterior $\pi_{\mathbf{X}|\mathbf{Y}}$

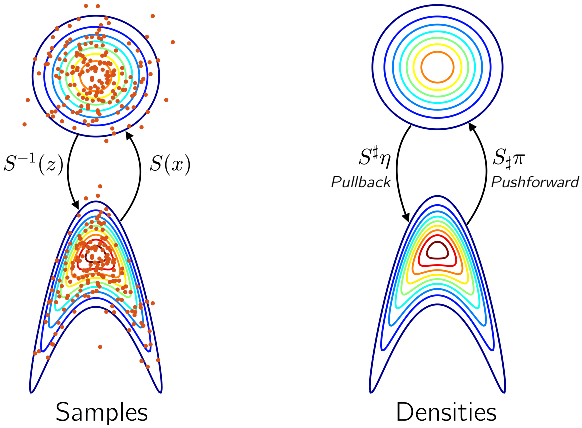


Plan for this talk:

1. Maps for filtering $\mathbf{X} = \mathbf{X}_t$?
2. Maps for smoothing $\mathbf{X} = \mathbf{X}_{1:t}$?
3. Leveraging structure in T to tackle high-dimensional problems

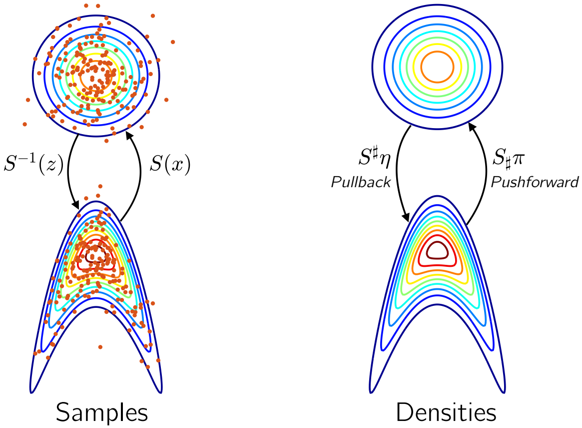
Transport maps characterize distributions

- Transport map S induces a deterministic coupling between a target density π and a reference density η (e.g., standard normal)
  - Generate cheap and independent samples: $\mathbf{x} \sim \pi \Leftrightarrow S(\mathbf{x}) \sim \eta$
  - Evaluate the target density: $\pi(\mathbf{x}) = S^\sharp \eta(\mathbf{x}) := \eta \circ S(\mathbf{x}) |\det \nabla S(\mathbf{x})|$



Transport maps characterize distributions

- Transport map S induces a deterministic coupling between a target density π and a reference density η (e.g., standard normal)
  - Generate cheap and independent samples: $\mathbf{z} \sim \eta \Leftrightarrow S^{-1}(\mathbf{z}) \sim \pi$
  - Evaluate the target density: $\pi(\mathbf{x}) = S^\sharp \eta(\mathbf{x}) := \eta \circ S(\mathbf{x}) |\det \nabla S(\mathbf{x})|$



Monotone triangular maps

As a building block, consider the **Knothe-Rosenblatt rearrangement**

$$S(\mathbf{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, x_2, \dots, x_d) \end{bmatrix}$$

Monotone triangular maps

As a building block, consider the **Knothe-Rosenblatt rearrangement**

$$S(\mathbf{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, x_2, \dots, x_d) \end{bmatrix}$$

1. Unique under mild assumptions on π and η
2. Invertibility is guaranteed by one-dimensional monotonicity $\partial_k S_k > 0$
3. $S^{-1}(\mathbf{z})$ and $\det \nabla S(\mathbf{x})$ are simple to evaluate
4. Each component S_k characterizes one marginal conditional

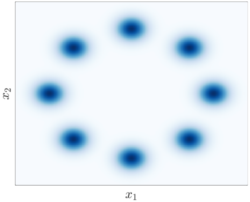
$$\pi_{\mathbf{X}} = \pi_{\mathbf{X}_1} \pi_{\mathbf{X}_2|\mathbf{X}_1} \cdots \pi_{\mathbf{X}_d|\mathbf{X}_{1,\dots,}\mathbf{X}_{d-1}}$$

Learning expressive triangular maps from samples

Given target density π and standard Gaussian η,

$$\min_{S} \; D_{KL}(\pi || S^{\sharp}\eta)$$

$$\Leftrightarrow \min_{\{s : \partial_k s > 0\}} \mathbb{E}_{\pi}\left[\frac{1}{2} s(\mathbf{x}_{1:k})^2 - \log |\partial_k s(\mathbf{x}_{1:k})|\right] \; \forall k$$

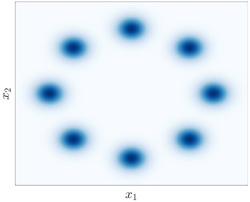


Learning expressive triangular maps from samples

Given target density π and standard Gaussian η,

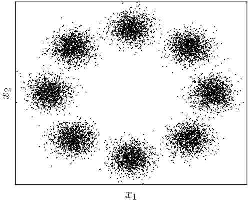
$$\min_{S} \; \mathrm{D}_{\mathrm{KL}}(\pi || S^{\sharp}\eta)$$

$$\Leftrightarrow \min_{\{s:\partial_k s>0\}} \mathbb{E}_{\pi}\left[\frac{1}{2}s(\mathbf{x}_{1:k})^2 - \log|\partial_k s(\mathbf{x}_{1:k})|\right] \; \forall k$$



Given samples $\{\mathbf{x}^i\}_{i=1}^n \sim \pi$, find $\widehat{S}_k$ via

$$\underset{\{s:\partial_k s>0\}}{\arg\min} \; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}s(\mathbf{x}^i_{1:k})^2 - \log|\partial_k s(\mathbf{x}^i_{1:k})|\right]$$



Target density approximation: $\widehat{\pi}(\mathbf{x}) := \widehat{S}^{\sharp}\eta(\mathbf{x})$

Triangular maps enable conditional sampling

Consider the triangular map pushing forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$:

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}) \\ S^{\mathcal{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- $S^{\mathcal{Y}}$ pushes forward $\pi_{\mathbf{Y}}$ to $\eta_{\mathbf{Z}_1}$
- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pushes forward $\pi_{\mathbf{X}|\mathbf{y}}$ to $\eta_{\mathbf{Z}_2}$ for any $\mathbf{y}$

Triangular maps enable conditional sampling

Consider the triangular map pushing forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$:

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}) \\ S^{\mathcal{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- $S^{\mathcal{Y}}$ pushes forward $\pi_{\mathbf{Y}}$ to $\eta_{\mathbf{Z}_1}$
- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pushes forward $\pi_{\mathbf{X}|\mathbf{y}}$ to $\eta_{\mathbf{Z}_2}$ for any $\mathbf{y}$

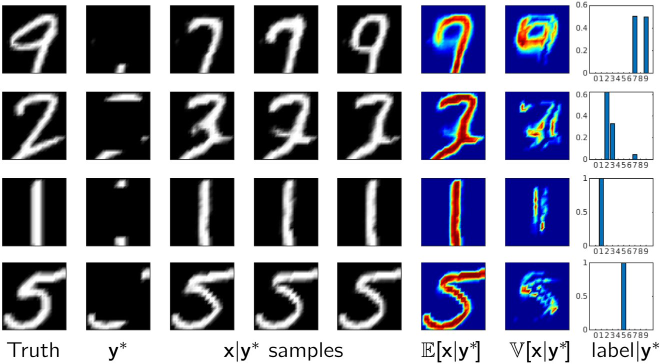
Recipe for amortized inference:

To characterize posterior $\pi_{\mathbf{X}|\mathbf{y}^*} \propto \pi_{\mathbf{y}^*|\mathbf{X}}\pi_{\mathbf{X}}$ given an observation $\mathbf{y}^*$:

- Simulate from the model: $\mathbf{x}^i \sim \pi_{\mathbf{X}}$, $\mathbf{y}^i \sim \pi_{\mathbf{Y}|\mathbf{x}^i}$
- Estimate $S^{\mathcal{X}}$ from joint samples $(\mathbf{x}^i,\mathbf{y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$
- Simulate $\widehat{S}^{\mathcal{X}}(\mathbf{y}^*,\cdot)^{-1}\Big|_{\mathbf{z}^i} \sim \pi_{\mathbf{X}|\mathbf{y}^*}$ for $\mathbf{z}^i \sim \eta_{\mathbf{Z}_2}$

Related Work: Simulation-based or likelihood-free inference [Papamakarios & Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019]

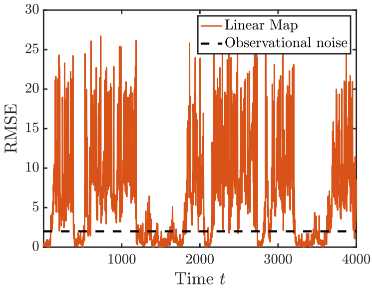
- **Goal**: Reconstruct image after removing its center section
- Use map to sample from the conditional distribution for the 14×14 center pixels of a 28×28 MNIST handwritten digit
- Estimate conditional mean and variance and classify digit probability

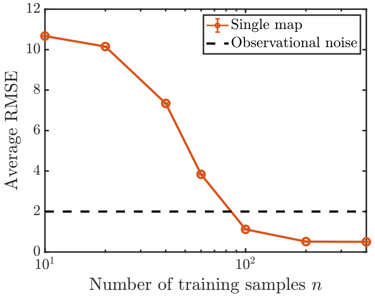


Note: Prior distributions in imaging problems have no analytic form

Lorenz-63 model

- Infer the hidden state given noisy point-wise observations
- With $N = 50$ samples, we can *at best* estimate linear maps
- Measure root-mean-squared error (RMSE) of ensemble mean

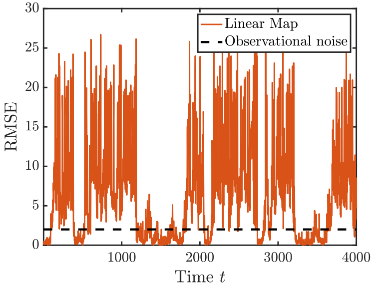


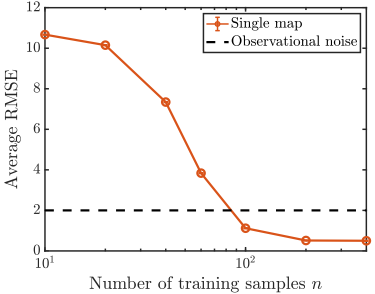


Will this always work well?

Lorenz-63 model

- Infer the hidden state given noisy point-wise observations
- With $N = 50$ samples, we can *at best* estimate linear maps
- Measure root-mean-squared error (RMSE) of ensemble mean





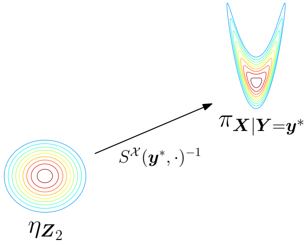
Takeaway: This approach yields large errors with limited samples

Another approach: compose maps for sampling

For $\pi_{\mathbf{Y},\mathbf{X}}$ and $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}) \\ S^{\mathcal{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pushes forward $\pi_{\mathbf{X}|\mathbf{y}}$ to $\eta_{\mathbf{Z}_2}$ *for any* $\mathbf{y}$
- $S^{\mathcal{X}}(\mathbf{y},\mathbf{x})$ pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_2}$

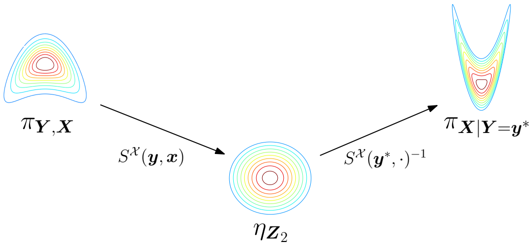


Another approach: compose maps for sampling

For $\pi_{\mathbf{Y},\mathbf{X}}$ and $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}) \\ S^{\mathcal{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pushes forward $\pi_{\mathbf{X}|\mathbf{y}}$ to $\eta_{\mathbf{Z}_2}$ *for any* $\mathbf{y}$
- $S^{\mathcal{X}}(\mathbf{y},\mathbf{x})$ pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_2}$

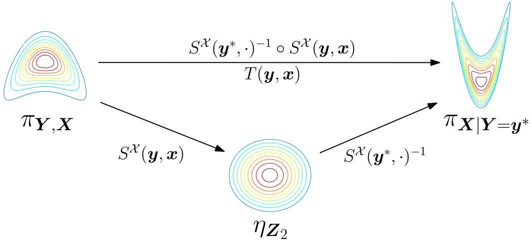


Another approach: compose maps for sampling

For $\pi_{\mathbf{Y},\mathbf{X}}$ and $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}) \\ S^{\mathcal{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pushes forward $\pi_{\mathbf{X}|\mathbf{y}}$ to $\eta_{\mathbf{Z}_2}$ *for any* $\mathbf{y}$
- $S^{\mathcal{X}}(\mathbf{y},\mathbf{x})$ pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_2}$



Another approach: compose maps for sampling

The prior-to-posterior map that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\mathbf{y}^*}$ is

$$\boxed{T_{\mathbf{y}^*}(\mathbf{y},\mathbf{x}) = S^{\mathcal{X}}(\mathbf{y}^*,\cdot)^{-1} \circ S^{\mathcal{X}}(\mathbf{y},\mathbf{x})}$$

Another approach: compose maps for sampling

The prior-to-posterior map that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\mathbf{y}^*}$ is

$$\boxed{T_{\mathbf{y}^*}(\mathbf{y},\mathbf{x}) = S^{\mathcal{X}}(\mathbf{y}^*,\cdot)^{-1} \circ S^{\mathcal{X}}(\mathbf{y},\mathbf{x})}$$

Stochastic map algorithm:

1. Estimate $S^{\mathcal{X}}$ using $(\mathbf{y}^i,\mathbf{x}^i) \sim \pi_{\mathbf{Y},\mathbf{X}}$
2. Evaluate composed map $T_{\mathbf{y}^*}(\mathbf{y},\mathbf{x})$ to approximately sample posterior

Stochastic map algorithm for filtering

Forecast step

1. Apply dynamics to generate forecast ensemble $(\mathbf{x}_t^f)^i \sim \pi_{\mathbf{X}_t|\mathbf{x}_{t-1}^i}$

Analysis step

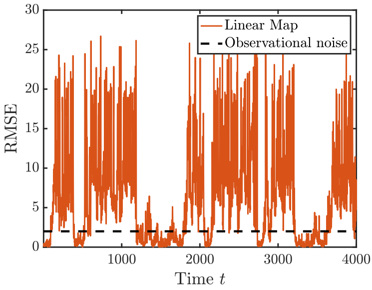
1. Sample observations $\mathbf{y}_t^i \sim \pi_{\mathbf{Y}_t|(\mathbf{x}_t^f)^i}$ using forecast samples
2. Estimate lower-triangular map S that couples $\pi_{\mathbf{Y}_t,\mathbf{X}_t}$ and $\mathcal{N}(\mathbf{0},\mathbf{I})$

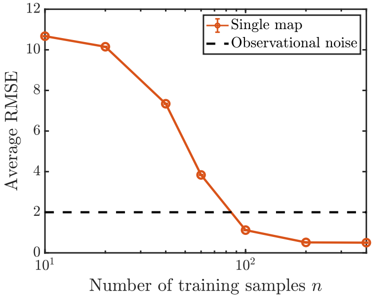
$$S(\mathbf{y}_t, \mathbf{x}_t) = \begin{bmatrix} S^{\mathcal{Y}}(\mathbf{y}_t) \\ S^{\mathcal{X}}(\mathbf{y}_t, \mathbf{x}_t) \end{bmatrix}$$

3. Compose maps $T_{\mathbf{y}_t^*}(\mathbf{y}_t, \mathbf{x}_t) = S^{\mathcal{X}}(\mathbf{y}_t^*, \cdot)^{-1} \circ S^{\mathcal{X}}(\mathbf{y}_t, \mathbf{x}_t)$
4. Generate analysis ensemble $\mathbf{x}_t^i = T_{\mathbf{y}_t^*}(\mathbf{y}_t^i, \mathbf{x}_t^i)$ for $i = 1, \ldots, N$

Lorenz-63 model

- Infer the hidden state given noisy point-wise observations
- With $N = 50$ samples, we can *at best* estimate linear maps
- Measure root-mean-squared error (RMSE) of ensemble mean

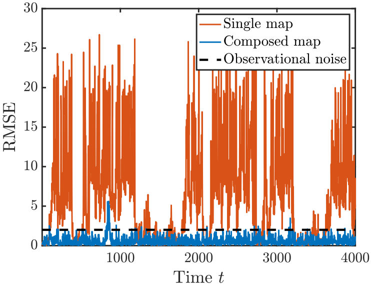


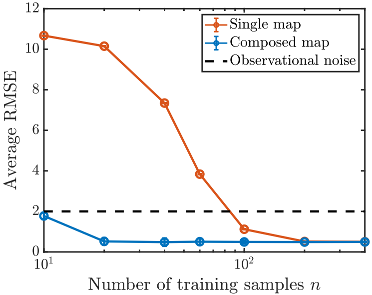


Takeaway: Composed maps have stable RMSE with limited samples

Lorenz-63 model

- Infer the hidden state given noisy point-wise observations
- With $N = 50$ samples, we can *at best* estimate linear maps
- Measure root-mean-squared error (RMSE) of ensemble mean





Takeaway: Composed maps have stable RMSE with limited samples

Numerical details of the stochastic map algorithm

Generalization of the EnKF

- Restricting $S^{\mathcal{X}}$ to be affine in $\mathbf{x}_t, \mathbf{y}_t$, we recover the transformation

$$T_{\mathbf{y}_t^*}(\mathbf{y}_t, \mathbf{x}_t) = \mathbf{x}_t - \Sigma_{\mathbf{x}_t,\mathbf{y}_t}\Sigma_{\mathbf{y}_t}^{-1}(\mathbf{y}_t - \mathbf{y}_t^*),$$

- Transport maps allow for the gradual introduction of nonlinear terms
- Nonlinear maps $T_{\mathbf{y}_t^*}$ capture non-Gaussian structure of $\pi_{\mathbf{Y}_t,\mathbf{X}_t}$

Generalization of the EnKF

- Restricting $S^{\mathcal{X}}$ to be affine in $\mathbf{x}_t, \mathbf{y}_t$, we recover the transformation

$$T_{\mathbf{y}_t^*}(\mathbf{y}_t, \mathbf{x}_t) = \mathbf{x}_t - \Sigma_{\mathbf{x}_t,\mathbf{y}_t}\Sigma_{\mathbf{y}_t}^{-1}(\mathbf{y}_t - \mathbf{y}_t^*),$$

- Transport maps allow for the gradual introduction of nonlinear terms
- Nonlinear maps $T_{\mathbf{y}_t^*}$ capture non-Gaussian structure of $\pi_{\mathbf{Y}_t,\mathbf{X}_t}$

Example map parameterization

- Each component is the sum of nonlinear univariate functions

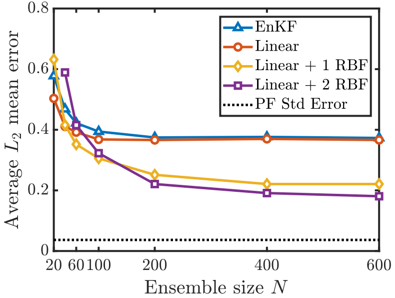
$$S_k(z_1, \dots, z_k) = \mathbf{u}_1(z_1) + \cdots + \mathbf{u}_k(z_k),$$

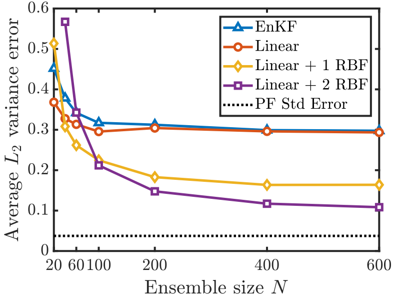
where $\mathbf{u}_i(z) = u_{i,0}z + \sum_{j=1}^{p} u_{ij}\,\mathcal{N}(z; \xi_j, \sigma_j^2)$ and $\mathbf{u}_k(z_k)$ is monotone

Nonlinear maps capture filtering distribution

Lorenz-63 model

- $d = 3$ with $\Delta t_{obs} = 0.1$ and fully-observed state
- Observations follow $\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\eta}_t$ with $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I})$
- Measure root-mean-squared-error $\text{RMSE}(t) = \|\mathbf{x}_t^* - \mathbb{E}[\mathbf{x}_t|\mathbf{y}_{1:t}^*]\|_2/\sqrt{d}$
- Compare statistics to a particle filter (PF) with 1M samples



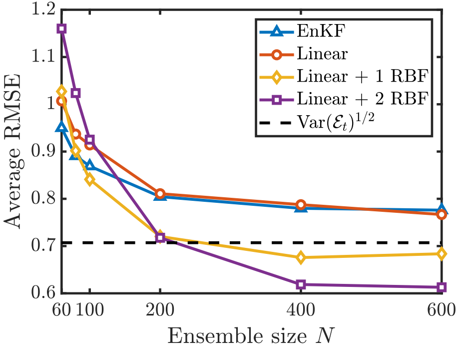


Improved posterior estimates is also stable with increasing Δt_{obs}

Nonlinear maps improve tracking

Lorenz-96 model: chaotic dynamics

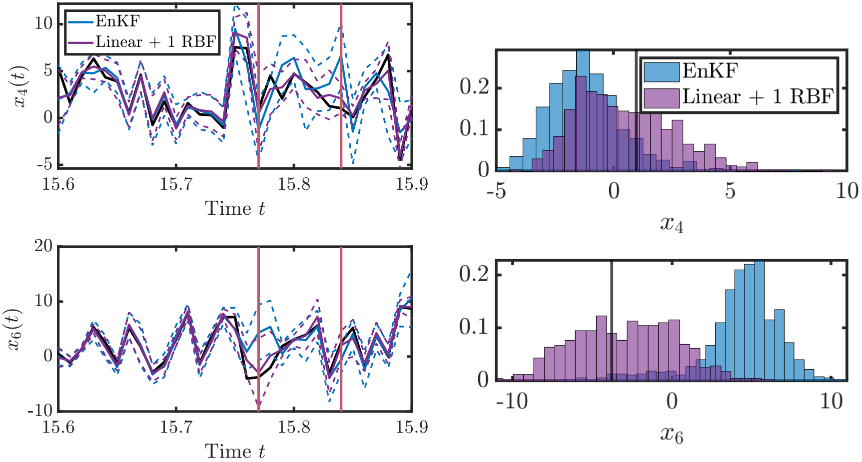
- 40 states, 20 observations, and $\Delta t_{obs} = 0.4$ (**large!**)
- Measure average RMSE (*left*) over 2000 assimilation cycles
- Parametrize maps with increasing nonlinearity using RBFs



Nonlinear maps also improve estimates of posterior moments

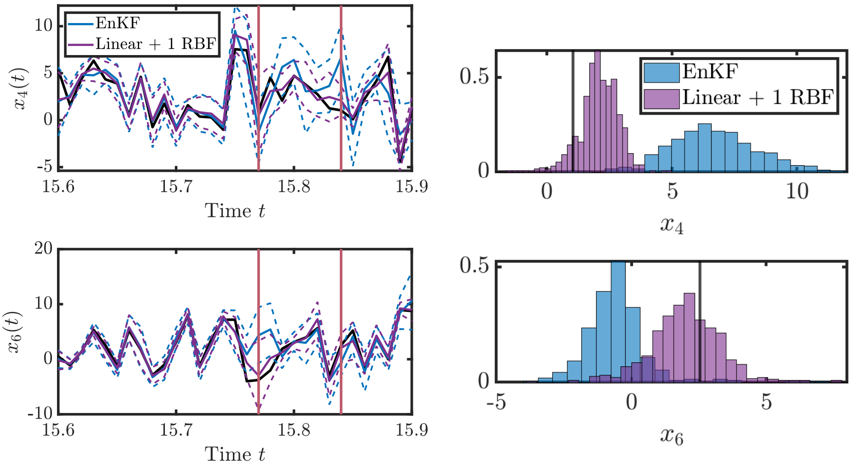
Nonlinear maps better capture uncertainty in true state

- Tracking two marginals of Lorenz-96 system at two assimilation times
- Compare ensemble distribution from EnKF and nonlinear maps



Nonlinear maps better capture uncertainty in true state

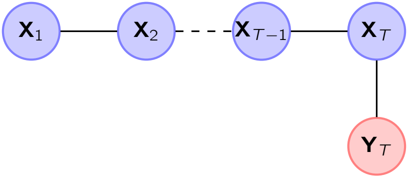
- Tracking two marginals of Lorenz-96 system at two assimilation times
- Compare ensemble distribution from EnKF and nonlinear maps



Extension to smoothing

Goal: Characterize full smoothing distribution $\pi_{\mathbf{X}_{1:T}|\mathbf{y}_{1:T}}$ or a marginal

- Consider update for all states given a single observation at time T



Ensemble Transport Smoother: Apply stochastic map algorithm on joint states over time:

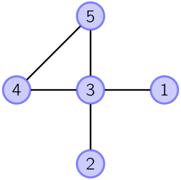
$$T_{\mathbf{y}_T^*}(\mathbf{y}_T, \mathbf{x}_{1:T}) = S^{\mathcal{X}}(\mathbf{y}_T^*, \cdot)^{-1} \circ S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T})$$

- Ordering of states in $S^{\mathcal{X}}$ defines different smoothing algorithms
- Exploiting the Markov structure of the states yields sparse maps

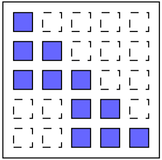
Transport maps exploit conditional independence

Theorem: Sparsity of triangular maps [Spantini et al., 2018]

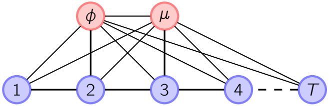
Conditional independence of target distribution π (encoded by graph) defines functional dependence of S such that $S^\sharp \eta = \pi$



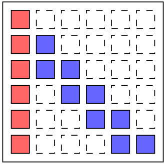
Markov structure of 5-dimensional distribution



Sparsity of $\partial_j S_k$



Markov structure of hidden Markov model



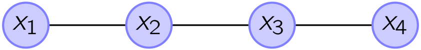
Sparsity of $\partial_j S_k$

Transport maps exploit conditional independence

Theorem: Sparsity of triangular maps [Spantini et al., 2018]

Conditional independence of target distribution π (encoded by graph) defines functional dependence of S such that $S^\sharp \eta = \pi$

$$\begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ S_3({\color{lightgray}x_1}, x_2, x_3) \\ S_4({\color{lightgray}x_1}, {\color{lightgray}x_2}, x_3, x_4) \end{bmatrix} \begin{array}{lll} \to \pi(x_1) & & \\ \to \pi(x_2|x_1) & & \\ \to \pi(x_3|x_1, x_2) & = \pi(x_3|x_2) & X_3 \perp\!\!\!\perp X_1 | X_2 \\ \to \pi(x_4|x_1, x_2, x_3) & = \pi(x_4|x_3) & X_4 \perp\!\!\!\perp (X_1, X_2) | X_3 \end{array}$$



Two new classes of smoothers [Ramgraber, B et al., 2022]

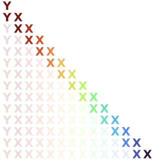
Backwards-in-time: uses the ordering $\mathbf{x}_T, \ldots, \mathbf{x}_1$

$$S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T}) \overset{CI}{=} \begin{bmatrix} S_T(\mathbf{y}_T, \mathbf{x}_T) \\ S_{T-1}(\mathbf{x}_T, \mathbf{x}_{T-1}) \\ \vdots \\ S_1(\mathbf{x}_2, \mathbf{x}_1) \end{bmatrix}$$

(CI) exploits chain structure: $\mathbf{x}_{1:T-1} \perp\!\!\!\perp \mathbf{y}_T | \mathbf{x}_T$ and $\mathbf{x}_{1:s-1} \perp\!\!\!\perp \mathbf{x}_{s+1:T} | \mathbf{x}_s$

Backwards-in-time: uses the ordering $\mathbf{x}_T, \ldots, \mathbf{x}_1$

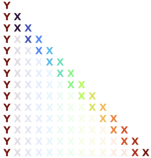
$$S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T}) \overset{CI}{=} \begin{bmatrix} S_T(\mathbf{y}_T, \mathbf{x}_T) \\ S_{T-1}(\mathbf{x}_T, \mathbf{x}_{T-1}) \\ \vdots \\ S_1(\mathbf{x}_2, \mathbf{x}_1) \end{bmatrix}$$



(CI) exploits chain structure: $\mathbf{x}_{1:T-1} \perp\!\!\!\perp \mathbf{y}_T | \mathbf{x}_T$ and $\mathbf{x}_{1:s-1} \perp\!\!\!\perp \mathbf{x}_{s+1:T} | \mathbf{x}_s$

Forwards-in-time: uses the ordering $\mathbf{x}_1, \ldots, \mathbf{x}_T$

$$S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T}) \overset{CI}{=} \begin{bmatrix} S_1(\mathbf{y}_T, \mathbf{x}_1) \\ S_2(\mathbf{y}_T, \mathbf{x}_1, \mathbf{x}_2) \\ \vdots \\ S_T(\mathbf{y}_T, \mathbf{x}_{T-1}, \mathbf{x}_T) \end{bmatrix}$$



(CI) exploits chain structure: $\mathbf{x}_s \perp\!\!\!\perp \mathbf{x}_{1:s-2} | \mathbf{x}_{s-1}$ for $s \geq 2$

Two new classes of smoothers [Ramgraber, B et al., 2022]

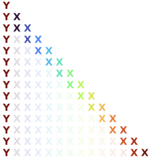
Backwards-in-time: uses the ordering $\mathbf{x}_T, \dots, \mathbf{x}_1$

$$S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T}) \stackrel{CI}{=} \begin{bmatrix} S_T(\mathbf{y}_T, \mathbf{x}_T) \\ S_{T-1}(\mathbf{x}_T, \mathbf{x}_{T-1}) \\ \vdots \\ S_1(\mathbf{x}_2, \mathbf{x}_1) \end{bmatrix}$$

(CI) exploits chain structure: $\mathbf{x}_{1:T-1} \perp\!\!\!\perp \mathbf{y}_T | \mathbf{x}_T$ and $\mathbf{x}_{1:s-1} \perp\!\!\!\perp \mathbf{x}_{s+1:T} | \mathbf{x}_s$

Forwards-in-time: uses the ordering $\mathbf{x}_1, \dots, \mathbf{x}_T$

$$S^{\mathcal{X}}(\mathbf{y}_T, \mathbf{x}_{1:T}) \stackrel{CI}{=} \begin{bmatrix} S_1(\mathbf{y}_T, \mathbf{x}_1) \\ S_2(\mathbf{y}_T, \mathbf{x}_1, \mathbf{x}_2) \\ \vdots \\ S_T(\mathbf{y}_T, \mathbf{x}_{T-1}, \mathbf{x}_T) \end{bmatrix}$$



(CI) exploits chain structure: $\mathbf{x}_s \perp\!\!\!\perp \mathbf{x}_{1:s-2} | \mathbf{x}_{s-1}$ for $s \geq 2$

- Empirical results suggest backward-in-time accumulates less errors
- Forwards smoother constrains state trajectories by dynamics

Focusing on backwards smoother

Sequential context: The joint decomposition simplifies

$$\pi(\mathbf{x}_{1:T}|\mathbf{y}^*_{1:T}) = \pi(\mathbf{x}_T|\mathbf{y}^*_{1:T}) \prod_{s=1}^{T-1} \pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:T})$$
$$= \pi(\mathbf{x}_T|\mathbf{y}^*_{1:T}) \prod_{s=1}^{T-1} \pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:s})$$

- Component S_s samples $\pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:s})$
- We estimate S_s using filtering ensemble $(\mathbf{x}^i_s, \mathbf{x}^i_{s+1}) \sim \pi(\mathbf{x}_s, \mathbf{x}_{s+1}|\mathbf{y}^*_{1:s})$

Focusing on backwards smoother

Sequential context: The joint decomposition simplifies

$$\begin{aligned}\pi(\mathbf{x}_{1:T}|\mathbf{y}^*_{1:T}) &= \pi(\mathbf{x}_T|\mathbf{y}^*_{1:T}) \prod_{s=1}^{T-1} \pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:T}) \\ &= \pi(\mathbf{x}_T|\mathbf{y}^*_{1:T}) \prod_{s=1}^{T-1} \pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:s})\end{aligned}$$

- Component S_s samples $\pi(\mathbf{x}_s|\mathbf{x}_{s+1}, \mathbf{y}^*_{1:s})$
- We estimate S_s using filtering ensemble $(\mathbf{x}^i_s, \mathbf{x}^i_{s+1}) \sim \pi(\mathbf{x}_s, \mathbf{x}_{s+1}|\mathbf{y}^*_{1:s})$

Generalization of the Ensemble RTS smoother

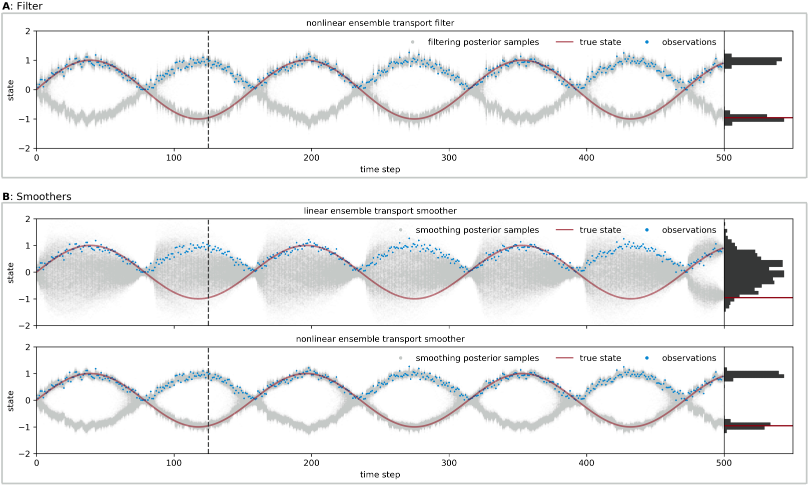
- Restricting $S^{\mathcal{X}}$ to be affine in $\mathbf{y}_t, \mathbf{x}_{1:t}$, we recover the transformation

$$\begin{aligned}T_{\mathbf{y}^*_T}(\mathbf{y}_T, \mathbf{x}_T) &= \mathbf{x}_T - \Sigma_{\mathbf{x}_T,\mathbf{y}_T}\Sigma^{-1}_{\mathbf{y}_T}(\mathbf{y}_T - \mathbf{y}^*_T) \\ T_{\mathbf{x}^*_{s+1}}(\mathbf{x}_s, \mathbf{x}_{s+1}) &= \mathbf{x}_s - \Sigma_{\mathbf{x}_s,\mathbf{x}_{s+1}}\Sigma^{-1}_{\mathbf{x}_{s+1}}(\mathbf{x}_{s+1} - \mathbf{x}^*_{s+1}), \qquad s < t\end{aligned}$$

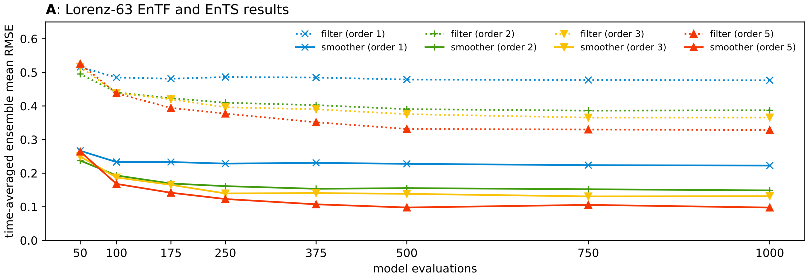
Takeaway: Non-linear transport maps generalize linear smoothers

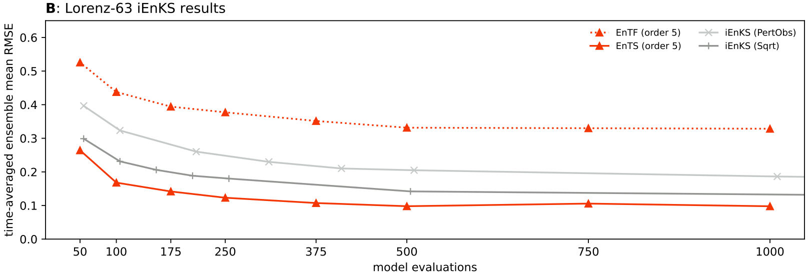
Nonlinear smoothers capture bimodal distributions

- Sinusoidal state x_t with observation $y_t = |x_t + \gamma|$ for $\gamma \sim \mathcal{N}(0, 0.1)$
- Infer state using random walk model without knowing true dynamics
- Backward smoother is initialized from nonlinear transport filter



Lorenz-63 model





Tackling high-dimensional inference problems

So far: Transport maps are consistent for sampling non-Gaussian filtering and smoothing distributions without requiring importance weights

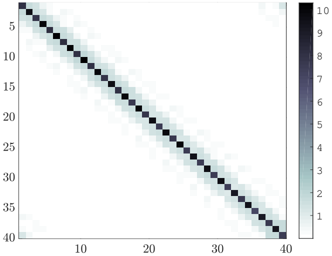
Tackling high-dimensional inference problems

So far: Transport maps are consistent for sampling non-Gaussian filtering and smoothing distributions without requiring importance weights

How do we compute transport maps given small ensemble sizes?

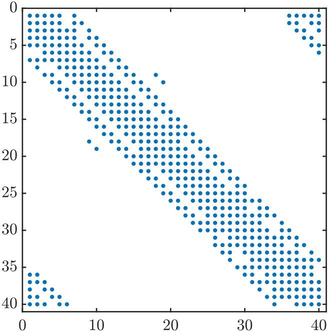
1. Localize estimators with approximate Markov structure
2. Targeted non-linearity using hybrid nonlinear+linear maps
3. Restrict inference to relevant low-dimensional subspaces

Many spatial fields satisfy approximate Markov properties



Inverse covariance matrix for
Lorenz-96 model forecast is **sparse**

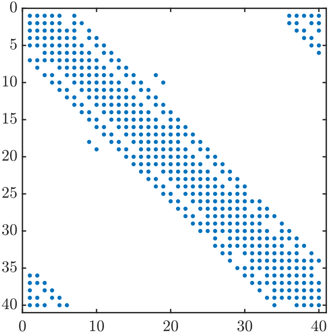
Many spatial fields satisfy approximate Markov properties



Inverse covariance matrix for
Lorenz-96 model forecast is **sparse**

1. Transport maps are easy to “localize" in high dimensions

Many spatial fields satisfy approximate Markov properties



Inverse covariance matrix for Lorenz-96 model forecast is **sparse**

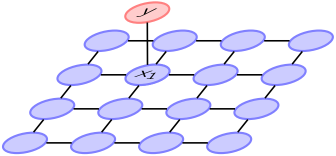
- **Idea**: Regularize the estimation of S by *imposing sparsity*:

$$\widehat{S}(\mathbf{x}) = \begin{bmatrix} \widehat{S}^1(x_1) \\ \widehat{S}^2(x_1, x_2) \\ \widehat{S}^3(\quad, x_2, x_3) \\ \widehat{S}^4(\qquad, x_3, x_4) \end{bmatrix}$$

- **Heuristic**: Let $\widehat{S}^k$ depend on neighboring variables $(x_j)_{j<k}$ that are physically close to x_k:

$$\widehat{S}^k(x_1, \dots, x_k) \approx \widehat{S}^k(x_{N(k)}, x_k)$$

Local-likelihood models: Scalar observation $y \sim \pi_{Y|X_1}$

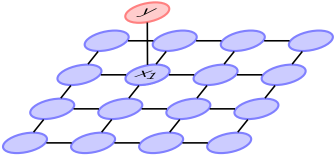


$$T(y, \mathbf{x}) = \begin{bmatrix} T_1(y, x_1) \\ \vdots \\ T_l(x_1, \dots, x_l) \\ L_{l+1}(x_1, \dots, x_{l+1}) \\ \vdots \\ L_d(x_1, \dots, x_d) \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} \text{Nonlinear maps} \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} \text{Affine maps: EnKF update} \end{matrix}$$

Idea: For conditionally Gaussian models, use nonlinear updates T_k for state variables $\mathbf{x}_{1:l}$ and use linear updates L_k for $\mathbf{x}_{l+1:d}$

2. Structured hybrid linear and nonlinear maps

Local-likelihood models: Scalar observation $y \sim \pi_{Y|X_1}$



$$T(y, \mathbf{x}) = \begin{bmatrix} T_1(y, x_1) \\ \vdots \\ T_l(x_1, \ldots, x_l) \\ L_{l+1}(x_1, \ldots, x_{l+1}) \\ \vdots \\ L_d(x_1, \ldots, x_d) \end{bmatrix} \begin{matrix} \left.\right\} \text{Nonlinear maps} \\ \left.\right\} \text{Affine maps: EnKF update} \end{matrix}$$

Idea: For conditionally Gaussian models, use nonlinear updates T_k for state variables $\mathbf{x}_{1:l}$ and use linear updates L_k for $\mathbf{x}_{l+1:d}$

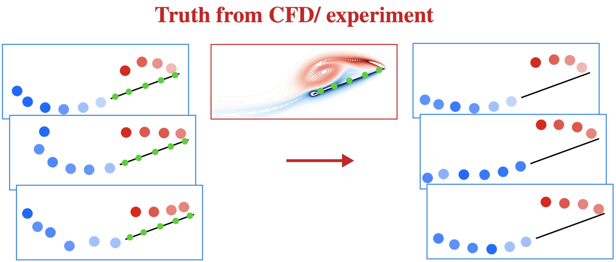
Special cases:

- $l = 1$: Nonlinear T_1 and keeping all other components affine is related to the rank histogram filter [Andersen 2010]
- With decay in correlation, $L_{l+1}, \ldots, L_d$ reverts to an identity map

3. Low-rank updates via an example in turbulent flows

Inference problem:

- States $\mathbf{x}_t$: Positions and strengths of point vortices
- Observations $\mathbf{y}_t$: Pressure observations along airfoil



Challenges:

- High-dimensional states and observations $d = 180$ and $m = 50$
- Observations are non-local: $\mathbf{y}_t$ is related to all $\mathbf{x}_t$ by Poisson equation
- Limited ensemble of size $N = \mathcal{O}(100)$

Low-rank stochastic map filter

Main ideas

- Only part of the state $\mathbf{x}_r = U_r^T \mathbf{x}$ is informed by the observations
- Only part of the observation $\mathbf{y}_s = V_s^T \mathbf{y}$ is relevant to the states

Main ideas

- Only part of the state $\mathbf{x}_r = U_r^T \mathbf{x}$ is informed by the observations
- Only part of the observation $\mathbf{y}_s = V_s^T \mathbf{y}$ is relevant to the states

- Consider the posterior approximation at each assimilation step

$$\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \widehat{\pi}_{\mathbf{X}_r|\mathbf{Y}_s}(\mathbf{x}_r|\mathbf{y}_s)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\mathbf{x}_\perp|\mathbf{x}_r)$$

- **Approach**: Find U_r, V_s with small r and s from prior ensemble and observation operator such that $\pi_{\mathbf{X}|\mathbf{Y}} \approx \widehat{\pi}_{\mathbf{X}|\mathbf{Y}}$ [B, Marzouk et al., 2022]

Low-rank stochastic map filter

Main ideas

- Only part of the state $\mathbf{x}_r = U_r^T \mathbf{x}$ is informed by the observations
- Only part of the observation $\mathbf{y}_s = V_s^T \mathbf{y}$ is relevant to the states

- Consider the posterior approximation at each assimilation step

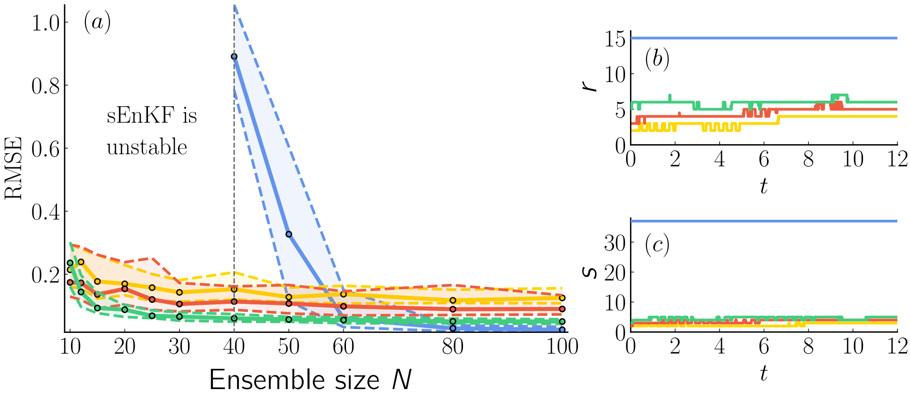
$$\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \widehat{\pi}_{\mathbf{X}_r|\mathbf{Y}_s}(\mathbf{x}_r|\mathbf{y}_s)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\mathbf{x}_\perp|\mathbf{x}_r)$$

- **Approach**: Find U_r, V_s with small r and s from prior ensemble and observation operator such that $\pi_{\mathbf{X}|\mathbf{Y}} \approx \widehat{\pi}_{\mathbf{X}|\mathbf{Y}}$ [B, Marzouk et al., 2022]
- **Result**: Prior-to-posterior map only acts on low-dimensional variables

$$T_{\mathbf{y}^*}(\mathbf{y}, \mathbf{x}) = U_r T^r_{\mathbf{y}^*_s}(V_s^T \mathbf{y}, U_r^T \mathbf{x}) + U_\perp U_\perp^T \mathbf{x}$$

- T_r can be linear [Le Provost, B et al., 2022] or non-linear

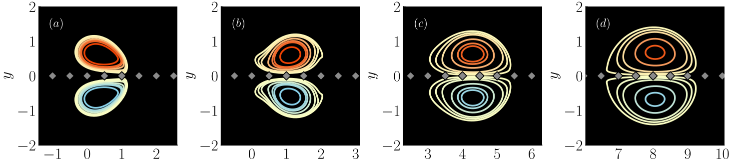
Low-rank filter is stable for small ensemble sizes



Observations:

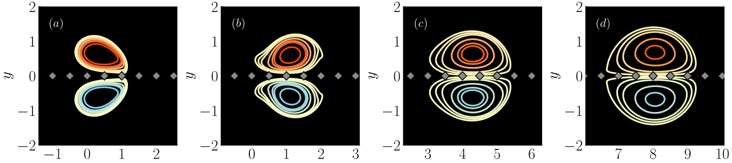
- RMSE is stable for small N for different energy ratios
- Adaptive reduced dimensions r, s do not increase over time

High-fidelity numerical simulation at Reynolds number 1000

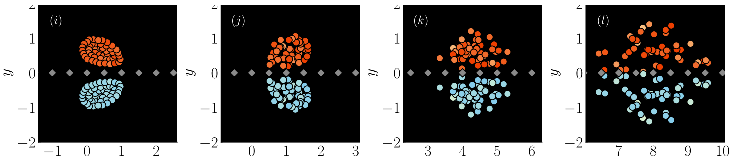


Low-rank EnKF is stable with model error

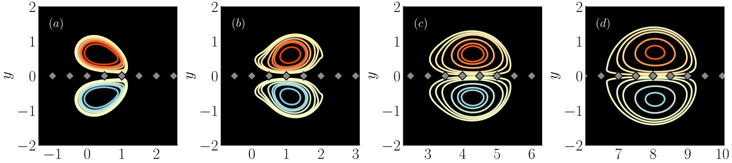
High-fidelity numerical simulation at Reynolds number 1000



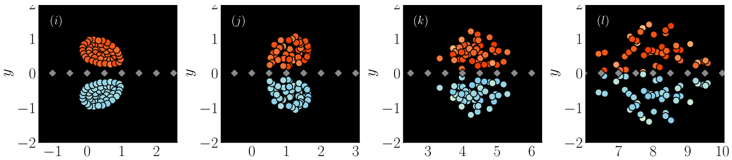
Inviscid vortex model with EnKF



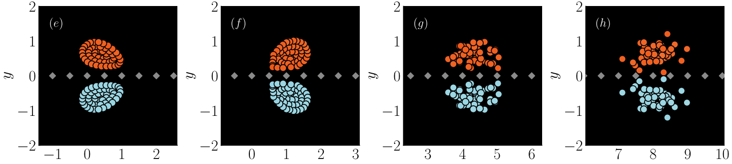
High-fidelity numerical simulation at Reynolds number 1000



Inviscid vortex model with EnKF



Inviscid vortex model with LR-EnKF



Conclusions and outlook

Conclusions and outlook

Central idea: consistent data assimilation using measure transport

- **Composed transport maps** generalize ensemble filters and smoothers
- Nonlinear maps **improve state estimation** for chaotic systems
- **Exploit (approximate) conditional independence structure** for scaling to high-dimensional inference problems

Conclusions and outlook

Central idea: consistent data assimilation using measure transport

- **Composed transport maps** generalize ensemble filters and smoothers
- Nonlinear maps **improve state estimation** for chaotic systems
- **Exploit (approximate) conditional independence structure** for scaling to high-dimensional inference problems

Ongoing work

- Square-root versions of nonlinear filters and smoothers
- Connections to other nonlinear filters, e.g., conjugate transform filter [Chipilski 2023]

Conclusions and outlook

Central idea: consistent data assimilation using measure transport

- **Composed transport maps** generalize ensemble filters and smoothers
- Nonlinear maps **improve state estimation** for chaotic systems
- **Exploit (approximate) conditional independence structure** for scaling to high-dimensional inference problems

Ongoing work

- Square-root versions of nonlinear filters and smoothers
- Connections to other nonlinear filters, e.g., conjugate transform filter [Chipilski 2023]

Main references: arXiv:1907.00389, arXiv:2203.05120, arXiv:2210.17000

Thank You

Supported by the U.S. Department of Energy and NSERC

References I

[1] Jeffrey L Anderson. "A non-Gaussian ensemble filter update for data assimilation". In: *Monthly Weather Review* 138.11 (2010), pp. 4186–4198.

[2] Ricardo Baptista, Lianghao Cao, Joshua Chen, et al. "Bayesian model calibration for block copolymer self-assembly: Likelihood-free inference and expected information gain computation via measure transport". In: *arXiv preprint arXiv:2206.11343* (2022).

[3] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. "Gradient-based data and parameter dimension reduction for Bayesian models: an information theoretic perspective". In: *arXiv preprint arXiv:2207.08670* (2022).

[4] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[5] Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. "Conditional Sampling With Monotone GANs". In: *arXiv:2006.06755* (2020).

[6] Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff D Eldredge. "A low-rank ensemble Kalman filter for elliptic observations". In: *Proceedings of the Royal Society A* 478.2266 (2022), p. 20220182.

[7] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. "Sampling via measure transport: An introduction". In: *Handbook of Uncertainty Quantification* (2016), pp. 1–41.

[8] George Papamakarios and Iain Murray. "Fast epsilon-free inference of simulation models with bayesian conditional density estimation". In: *arXiv preprint arXiv:1605.06376* (2016).

[9] Maximilian Ramgraber, Ricardo Baptista, Dennis McLaughlin, and Youssef Marzouk. "Ensemble transport smoothing–Part 1: unified framework". In: *arXiv preprint arXiv:2210.17000* (2022).

[10] Maximilian Ramgraber, Ricardo Baptista, Dennis McLaughlin, and Youssef Marzouk. "Ensemble transport smoothing–Part 2: nonlinear updates". In: *arXiv preprint arXiv:2210.17435* (2022).

[11] Alessio Spantini, Ricardo Baptista, and Youssef Marzouk. "Coupling techniques for nonlinear ensemble filtering". In: *SIAM Review* (2022, In press).

[12] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. "Inference via low-dimensional couplings". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2639–2709.