Transport methods for Bayesian computation: Part 2

Youssef Marzouk
joint work with Daniele Bigoni, Matthew Parno, Alessio Spantini, & Olivier Zahm

Department of Aeronautics and Astronautics
Center for Computational Engineering
Statistics and Data Science Center

Massachusetts Institute of Technology
http://uqgroup.mit.edu

Support from AFOSR, DARPA, DOE

25–26 September 2019

Preconditioning Monte Carlo with transport

What to do when $T_\sharp \eta \neq \pi$?

- Maybe close enough? Can evaluate variance diagnostic $\mathbb{V}\mathrm{ar}_\eta[\log(\eta / T_\sharp^{-1}\bar{\pi})]$, bound $\mathrm{Tr}(H_{T_\sharp^{-1}\pi})$, etc.
- Enrich T, e.g., add a layer or expand $\mathcal{T}_\triangle^h$ in the given layer
- **Sample the pullback:** treat $T_\sharp^{-1}\pi$ with an asymptotically exact scheme, e.g., Markov chain Monte Carlo

Preconditioning Monte Carlo with transport

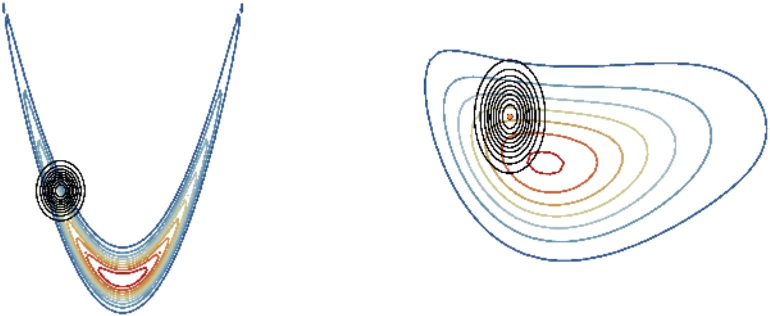
What to do when $T_\sharp \eta \neq \pi$?

- Maybe close enough? Can evaluate variance diagnostic $\mathbb{V}\mathrm{ar}_\eta[\log(\eta / T_\sharp^{-1}\bar{\pi})]$, bound $\mathrm{Tr}(H_{T_\sharp^{-1}\pi})$, etc.
- Enrich T, e.g., add a layer or expand $\mathcal{T}_\triangle^h$ in the given layer
- **Sample the pullback:** treat $T_\sharp^{-1}\pi$ with an asymptotically exact scheme, e.g., Markov chain Monte Carlo

One possible construction: **transport-accelerated MCMC**

- Transport map "preconditions" MCMC target; use MCMC iterates in maps-from-samples construction
- Can be understood in the framework of *adaptive MCMC*

Preconditioning MCMC



- Effective MCMC proposal = **adapted to the target**
  - Can we transform *proposals* or, equivalently, *targets* for better sampling?

Recall maps-from-samples construction

$$\min_{S \in \mathcal{S}_\triangle^h} \mathcal{D}_{KL}(\, S_\sharp \pi \,||\, \eta \,) = \min_{S \in \mathcal{S}_\triangle^h} \mathcal{D}_{KL}(\, \pi \,||\, S_\sharp^{-1} \eta \,)$$

- Suppose we have Monte Carlo samples $\{x_i\}_{i=1}^M \sim \pi$
- For standard Gaussian η, this problem is **convex** and **separable** for any π
- This is *density estimation via transport!* (cf. Tabak & Turner 2013)

Recall maps-from-samples construction

$$\min_{S \in \mathcal{S}_\triangle^h} \mathcal{D}_{KL}(S_\sharp \pi \,||\, \eta) = \min_{S \in \mathcal{S}_\triangle^h} \mathcal{D}_{KL}(\pi \,||\, S_\sharp^{-1} \eta)$$

- Suppose we have Monte Carlo samples $\{x_i\}_{i=1}^M \sim \pi$
- For standard Gaussian η, this problem is **convex** and **separable** for any π
- This is *density estimation via transport!* (cf. Tabak & Turner 2013)
- Equivalent to maximum likelihood estimation of S

$$\widehat{S} \in \arg \max_{S \in \mathcal{S}_\triangle^h} \frac{1}{M} \sum_{i=1}^M \log \underbrace{S_\sharp^{-1} \eta}_{\text{pullback}}(x_i), \qquad \eta = \mathcal{N}(0, \mathbf{I}_n),$$

- Each component $\widehat{S}^k$ of $\widehat{S}$ can be computed *separately*, via smooth convex optimization

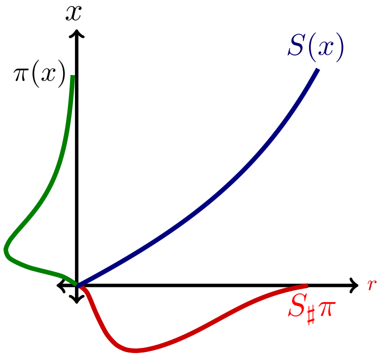
$$\widehat{S}^k \in \arg \min_{S^k \in \mathcal{S}_{\triangle,k}^h} \frac{1}{M} \sum_{i=1}^M \left(\frac{1}{2} S^k(x_i)^2 - \log \partial_k S^k(x_i) \right)$$

Transport + MCMC sampling

- View $\widehat{S}_\sharp\pi$ as the "preconditioned" target
  - In the MCMC setting, $\{x_i\}_{i=1}^M$ comprises *dependent* MCMC samples
  - $\widehat{S}_\sharp\pi$ may be far from standard Gaussian for small M and/or crude $\mathcal{S}_\triangle^h$

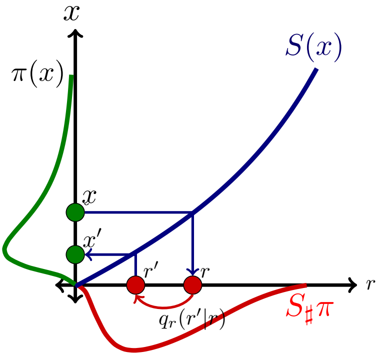
Map-accelerated MCMC

- **Ingredient #1: static map**
  - Perform MCMC in the reference space, on the “preconditioned” density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target



Map-accelerated MCMC

- **Ingredient #1: static map**
  - Perform MCMC in the reference space, on the "preconditioned" density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target

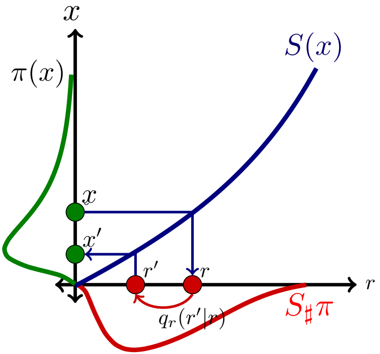


$$\alpha = \frac{\pi(S^{-1}(r')) \,|\, \nabla S^{-1} \,|_{r'} \; q_r(r \mid r')}{\pi(S^{-1}(r)) \,|\, \nabla S^{-1} \,|_{r} \; q_r(r' \mid r)}$$

simple proposal q_r on pushforward of target through map

Map-accelerated MCMC

- **Ingredient #1: static map**
  - Perform MCMC in the reference space, on the "preconditioned" density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target

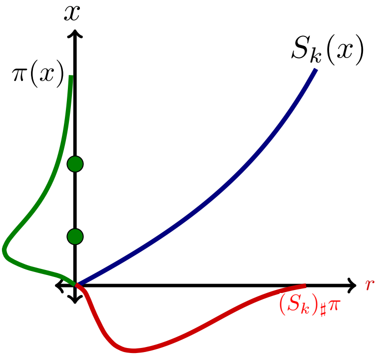


$$\alpha = \frac{\pi(S^{-1}(r'))\,|\,\nabla S^{-1}\,|_{r'}\; q_r(r\,|\,r')}{\pi(S^{-1}(r))\,|\,\nabla S^{-1}\,|_{r}\; q_r(r'\,|\,r)}$$

more complex proposal, directly on target distribution

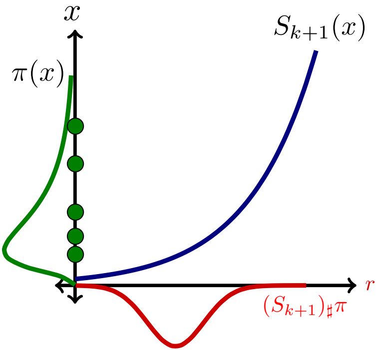
Map-accelerated MCMC

- **Ingredient #2: adaptive map**
  - Update the map with each MCMC iteration:
    more samples, more accurate $\mathbb{E}_\pi$, *better* S
  - Adaptive MCMC [Haario 2001, Andrieu 2006], but with nonlinear transformation to capture non-Gaussian structure



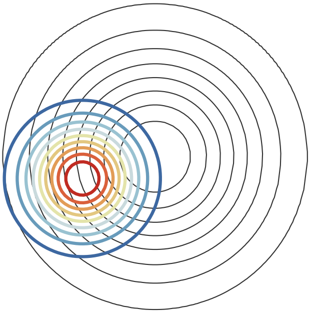
Map-accelerated MCMC

- **Ingredient #2: adaptive map**
  - Update the map with each MCMC iteration:
    more samples, more accurate $\mathbb{E}_\pi$*, better* S
  - Adaptive MCMC [Haario 2001, Andrieu 2006], but with nonlinear transformation to capture non-Gaussian structure

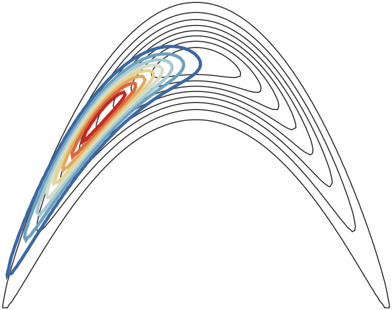


Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior



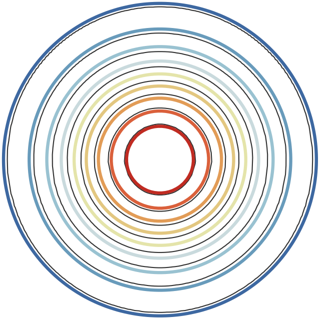
reference RW proposal



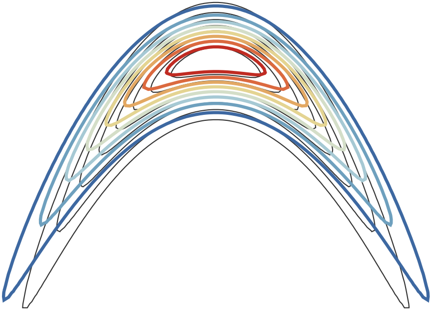
mapped RW proposal

Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior



reference independence proposal



mapped independence proposal

Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior
  - Solution: **delayed rejection** MCMC [Mira 2001]
  - First proposal = independent sample from η (global, more efficient); second proposal = random walk (local, more robust)

- Entire scheme is provably **ergodic** with respect to the exact posterior measure [Parno & M, *SIAM JUQ* 2018]
  - Requires enforcing some regularity conditions on maps, to preserve tail behavior of transformed target

Example: biological oxygen demand model

- Likelihood model:

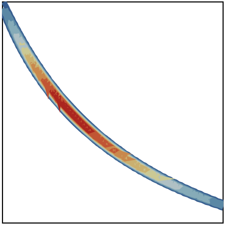
$$d = \theta_1(1 - \exp(-\theta_2 x)) + \epsilon$$
$$\epsilon \sim N(0, 2 \times 10^{-4})$$

- 20 noisy observations at

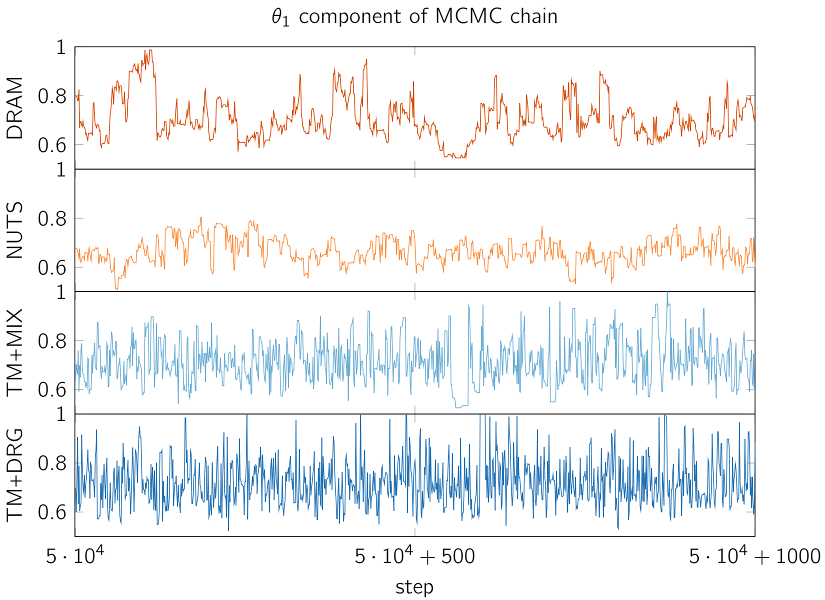
$$x = \left\{\frac{5}{5}, \frac{6}{5}, \ldots, \frac{25}{5}\right\}$$

- Degree-three polynomial map

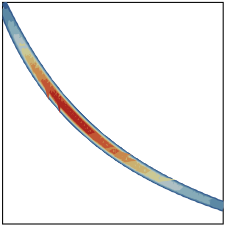
True posterior density



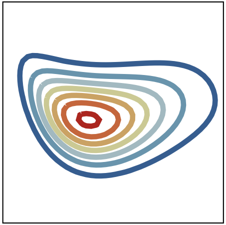
Results: MCMC chain



Transformed distribution

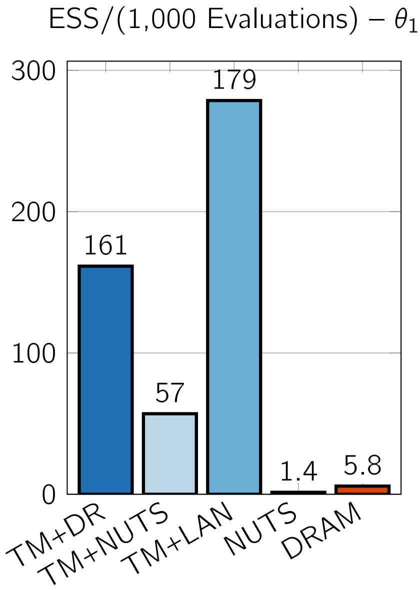


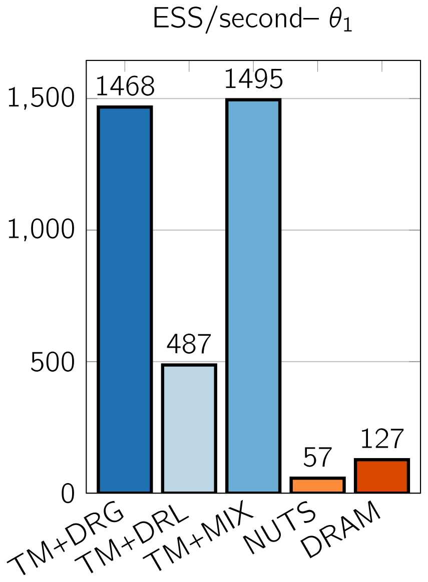
Original posterior π



Pushforward of posterior through *learned* map, $S_\sharp \pi$

Results: ESS per computational effort



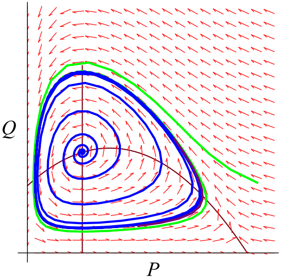


Example #2: predator-prey model

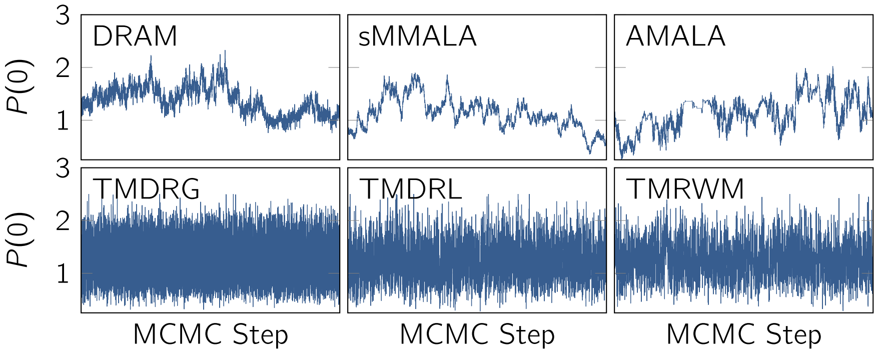
- Six parameter ODE population model

$$
\begin{aligned}
\frac{dP}{dt} &= rP\left(1 - \frac{P}{K}\right) - s\frac{PQ}{a+P} \\
\frac{dQ}{dt} &= u\frac{PQ}{a+P} - vQ
\end{aligned}
$$

- Five noisy observations of both populations
- Infer 6 parameters + 2 initial values; uniform priors



Predator-prey model: chains



Example: maple sap dynamics model

- Coupled PDE system for ice, water, and gas locations [Ceseri & Stockie 2013]
- Measure gas pressure in vessel
- Infer 10 physical model parameters
- Very challenging posterior!

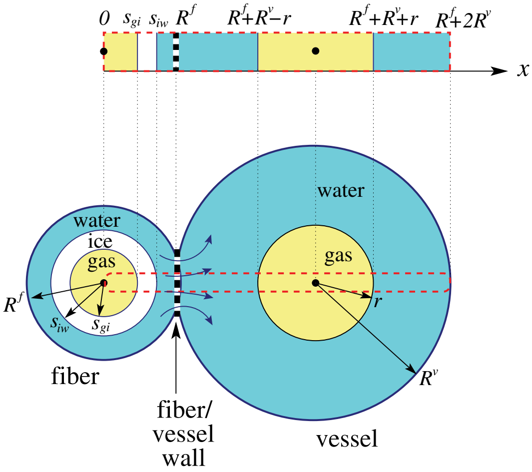
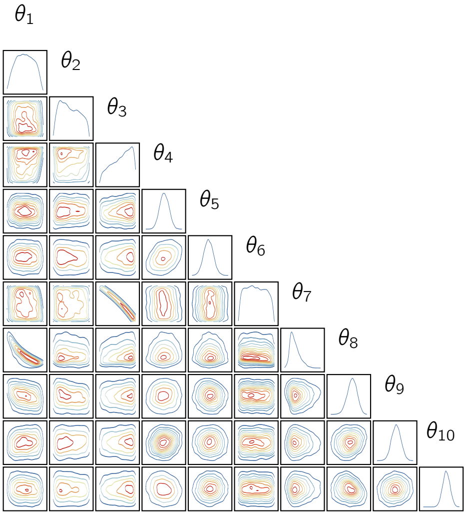
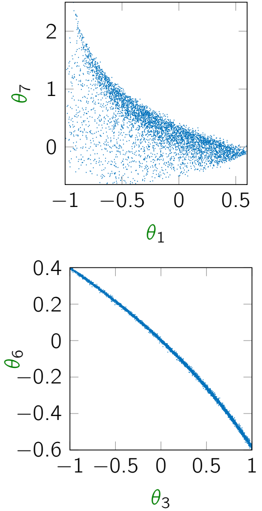


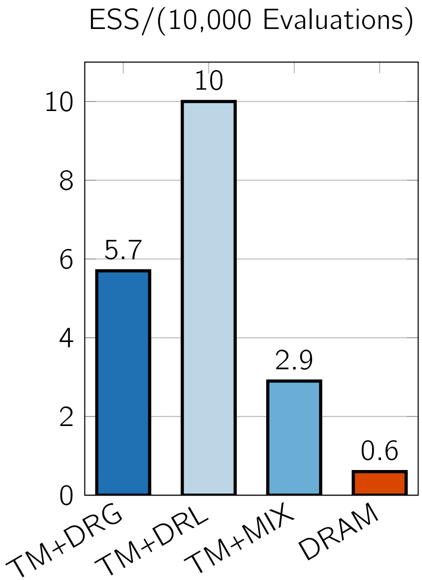
Image from *Ceseri and Stockie, 2013*

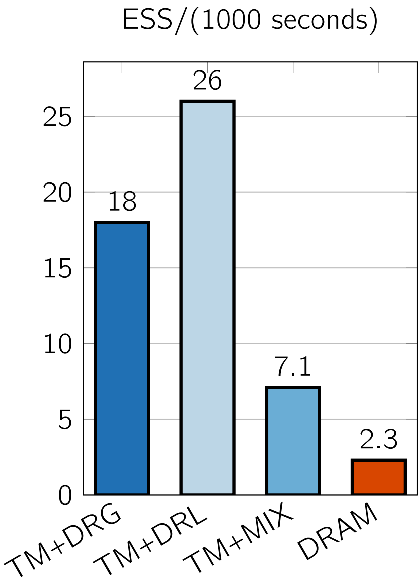
Maple posterior distribution





Results: ESS per computational effort





Comments on MCMC with transport maps

Useful characteristics of the algorithm:

- Map construction is easily parallelizable
- Requires no gradients from posterior density

Generalizes many current MCMC techniques:

- Adaptive Metropolis: map enables **non-Gaussian proposals** *and* a natural mixing between local and global moves
- Manifold MCMC [Girolami & Calderhead 2011]: map also defines a Riemannian metric

Looking to higher dimensions: regularized estimation of S

For simplicity, consider map components $S^k(\mathbf{x}) = \sum_j \beta_j \psi_j(x_{1:k-1}) + \alpha_k x_k$

$$\widehat{S}^k \in \arg \min_{S^k \in \mathcal{S}^h_{\triangle,k}} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2} S^k(\mathbf{x}_i)^2 - \log \partial_k S^k(\mathbf{x}_i) \right) + \lambda_N \|\boldsymbol{\beta}\|_1$$

Assume sub-Gaussian π and basis functions $\psi_j(\mathbf{x})$

Theorem [BZM]

For polynomial maps of degree m with sparsity s, with high probability

$$\mathbb{E}_\pi \left[D_{KL}\left(\pi(\mathbf{x}_k | \mathbf{x}_{1:k-1}) \,||\, \widehat{S}_k^\sharp \eta \right) \right] \lesssim \sqrt{\frac{s^2 m \log k}{N}}$$

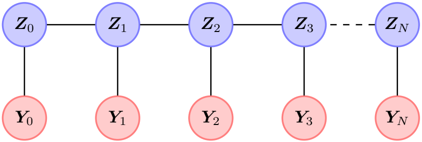
Takeaways

- Accurate estimation is feasible in high dimensions with $N \ll k$
- From factorization property of density, error in conditionals ensures

$$D_{KL}(\pi || \widehat{S}^\sharp \eta) \lesssim d \sqrt{\frac{s^2 m \log d}{N}}$$

Next topic: ensemble filtering via transport

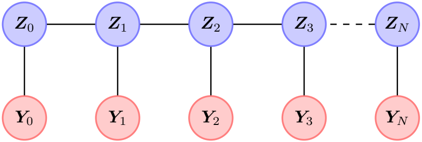
- **Nonlinear/non-Gaussian** state-space model:
  - Transition density $\pi_{\mathbf{Z}_k|\mathbf{Z}_{k-1}}$
  - Observation density (likelihood) $\pi_{\mathbf{Y}_k|\mathbf{Z}_k}$



- Interested in **recursively** updating the full Bayesian solution:
  $\pi_{\mathbf{Z}_{0:k}\,|\,\mathbf{y}_{0:k}} \to \pi_{\mathbf{Z}_{0:k+1}\,|\,\mathbf{y}_{0:k+1}}$ (smoothing)

Next topic: ensemble filtering via transport

- **Nonlinear/non-Gaussian** state-space model:
  - Transition density $\pi_{\mathbf{Z}_k|\mathbf{Z}_{k-1}}$
  - Observation density (likelihood) $\pi_{\mathbf{Y}_k|\mathbf{Z}_k}$



- Interested in **recursively** updating the full Bayesian solution:
  $\pi_{\mathbf{Z}_{0:k}\,|\,\mathbf{y}_{0:k}} \to \pi_{\mathbf{Z}_{0:k+1}\,|\,\mathbf{y}_{0:k+1}}$ (smoothing)
- Or focus on approximating the filtering distribution:
  $\pi_{\mathbf{Z}_k\,|\,\mathbf{y}_{0:k}} \to \pi_{\mathbf{Z}_{k+1}\,|\,\mathbf{y}_{0:k+1}}$ (marginals of the full Bayesian/smoothing solution)

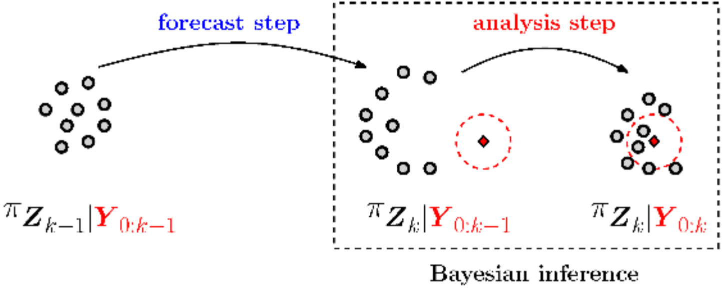
Problem setting

- Consider the filtering of state-space models with:
  1. High-dimensional states
  2. Challenging nonlinear dynamics (e.g., chaotic systems)
  3. Intractable transition kernels: can only obtain *forecast* samples, i.e., draws from $\pi_{\mathbf{Z}_{k+1} \mid \mathbf{z}_k}$
  4. Limited model evaluations, e.g., small ensemble sizes
  5. Sparse and local observations in space/time

- These constraints reflect typical challenges faced in numerical weather prediction, geophysical data assimilation

Ensemble Kalman filter

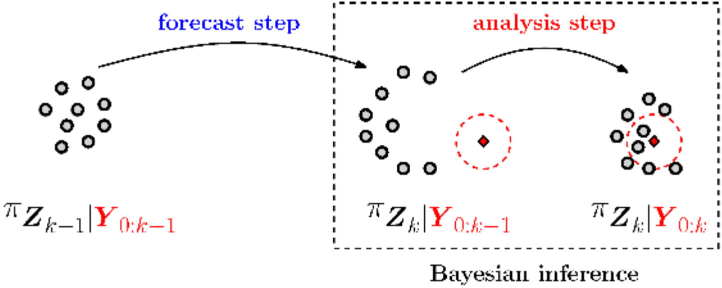
- State-of-the-art results (in terms of tracking) are typically obtained with the ensemble Kalman filter (EnKF)



- Move samples via an **affine** transformation; no weights or resampling!
- Yet ultimately **inconsistent**: does not converge to the true posterior

Ensemble Kalman filter

- State-of-the-art results (in terms of tracking) are typically obtained with the ensemble Kalman filter (EnKF)

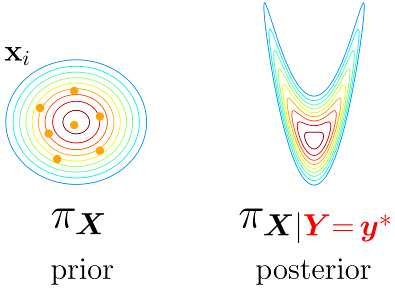


- Move samples via an **affine** transformation; no weights or resampling!
- Yet ultimately **inconsistent**: does not converge to the true posterior

Can we improve and *generalize* the EnKF, preserving scalability, via **nonlinear** transformations?

Assimilation step

At any assimilation time k, we have a Bayesian inference problem:



- $\pi_{\mathbf{X}}$ is the forecast distribution on $\mathbb{R}^n$
- $\pi_{\mathbf{Y}|\mathbf{X}}$ is the likelihood of the observations $\mathbf{Y} \in \mathbb{R}^d$
- $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$ is the filtering distribution for a realization $\mathbf{y}^*$ of the data

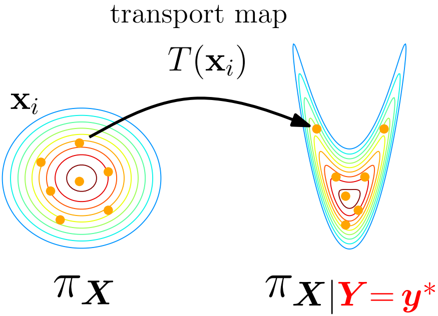
Goal: sample the posterior given only M prior samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$

Inference as a transportation of measures

- Seek a map T that pushes forward prior to posterior

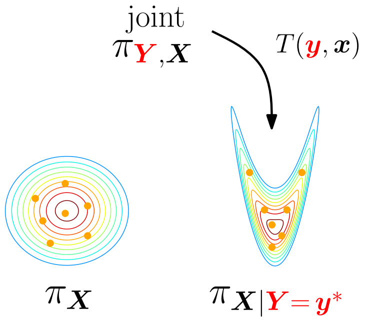
$$(\mathbf{x}_1, \ldots, \mathbf{x}_M) \sim \pi_{\mathbf{X}} \Longrightarrow (T(\mathbf{x}_1), \ldots, T(\mathbf{x}_M)) \sim \pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$$

- The map induces a coupling between prior and posterior measures



> How to construct a “good” coupling from very few prior samples?

Consider the joint distribution of state and observations



- Construct a map T from the joint distribution $\pi_{\mathbf{Y},\mathbf{X}}$ to the posterior
- T can be computed via convex optimization given samples from $\pi_{\mathbf{Y},\mathbf{X}}$
- Sample $\pi_{\mathbf{Y},\mathbf{X}}$ using the forecast ensemble and the likelihood

$$(\mathbf{y}_i, \mathbf{x}_i) \qquad \mathbf{y}_i \sim \pi_{\mathbf{Y}|\mathbf{X}=\mathbf{x}_i}$$

- **Intuition:** a generalization of the "perturbed observation" EnKF

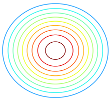
Couple the joint distribution with a standard normal

joint
$$\pi_{\boldsymbol{Y},\boldsymbol{X}} \xrightarrow{T : \mathbb{R}^{d+n} \to \mathbb{R}^n}$$

$\pi_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}^*}$

We can find T by computing a Knothe–Rosenblatt (KR) rearrangement S between $\pi_{\mathbf{Y},\mathbf{X}}$ and $\mathcal{N}(0, \mathbf{I}_{d+n})$

joint
$$\pi_{\boldsymbol{Y},\boldsymbol{X}} \xrightarrow{S : \mathbb{R}^{d+n} \to \mathbb{R}^{d+n}}$$



$\mathcal{N}(0, \mathbf{I}_{d+n})$

- We will show how to derive T from S...

Knothe–Rosenblatt (KR) rearrangement

- **Definition:** for any pair of absolutely continuous densities π, η on $\mathbb{R}^m$, there exists a unique triangular and monotone map $S : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$S_\sharp \pi = \eta$$

- Triangular function (nonlinear generalization of a triangular matrix):

$$S(x_1, \dots, x_m) = \begin{bmatrix} S^1(x_1) \\ S^2(x_1, x_2) \\ \vdots \\ S^m(x_1, x_2, \dots, x_m) \end{bmatrix}$$

- Existence stems from general factorization properties of a density,

$$\pi = \pi_{\mathbf{x}_1} \, \pi_{\mathbf{x}_2|\mathbf{x}_1} \cdots \pi_{\mathbf{x}_m|\mathbf{x}_1,\dots,\mathbf{x}_{m-1}}$$

Triangular maps enable conditional simulation

$$S(x_1, \ldots, x_m) = \begin{bmatrix} S^1(x_1) \\ S^2(x_1, x_2) \\ \vdots \\ S^m(x_1, x_2, \ldots, x_m) \end{bmatrix}$$

- Each component S^k links marginal conditionals of π and η
- For instance, if $\eta = \mathcal{N}(0, \mathbf{I})$, then for all $x_1, \ldots, x_{k-1} \in \mathbb{R}^{k-1}$

$$\xi \mapsto S^k(x_1, \ldots, x_{k-1}, \xi) \text{ pushes } \pi_{\mathbf{X}_k|\mathbf{X}_{1:k-1}}(\xi|\mathbf{x}_{1:k-1}) \text{ to } \mathcal{N}(0,1)$$

- Simulate the conditional $\pi_{\mathbf{X}_k|\mathbf{X}_{1:k-1}}$ by inverting a 1-D map $\xi \mapsto S^k(\mathbf{x}_{1:k-1}, \xi)$ at Gaussian samples (*need triangular structure*)

Filtering: the analysis map

- We are interested in the KR map S that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\mathcal{N}(0,\mathbf{I}_{d+n})$
- The KR map immediately has a block structure

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

which suggests two properties:

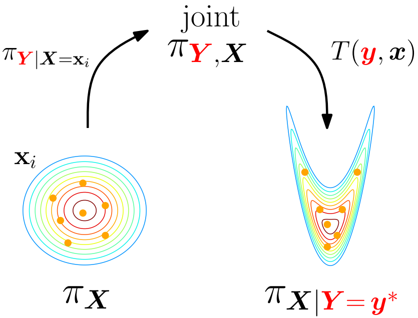
$$S^{\mathbf{X}} \text{ pushes } \pi_{\mathbf{Y},\mathbf{X}} \text{ to } \mathcal{N}(0,\mathbf{I}_n)$$

$$\boldsymbol{\xi} \mapsto S^{\mathbf{X}}(\mathbf{y}^*,\boldsymbol{\xi}) \text{ pushes } \pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*} \text{ to } \mathcal{N}(0,\mathbf{I}_n)$$

- The **analysis map** that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$ is then given by

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

A novel filtering algorithm with maps



Transport map ensemble filter

1. Compute forecast ensemble $\mathbf{x}_1, \ldots, \mathbf{x}_M$
2. Generate samples $(\mathbf{y}_i, \mathbf{x}_i)$ from $\pi_{\mathbf{Y},\mathbf{X}}$ with $\mathbf{y}_i \sim \pi_{\mathbf{Y}|\mathbf{X}=\mathbf{x}_i}$
3. Build an estimator $\widehat{T}$ of T
4. Compute analysis ensemble as $\mathbf{x}_i^{\mathrm{a}} = \widehat{T}(\mathbf{y}_i, \mathbf{x}_i)$ for $i = 1, \ldots, M$

Estimator for the analysis map

- Recall the form of S:

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S^{\mathsf{Y}}(\mathbf{y}) \\ S^{\mathsf{X}}(\mathbf{y},\mathbf{x}) \end{bmatrix}, \qquad S_\sharp \, \pi_{\mathsf{Y},\mathsf{X}} = \mathcal{N}(0, \mathbf{I}_{d+n}).$$

- We propose the following estimator $\widehat{T}$ of T:

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

where $\widehat{S}$ is a **maximum likelihood estimator** of S

Estimating the KR rearrangement from samples

Given samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ from a distribution π on $\mathbb{R}^{d+n}$, estimate the KR rearrangement S that pushes forward π to $\mathcal{N}(0, \mathbf{I}_{d+n})$

- Constrained MLE for S

$$\widehat{S} \in \arg\max_{S \in \mathcal{S}_\triangle^h} \frac{1}{M} \sum_{i=1}^{M} \log \underbrace{S_\sharp^{-1}\eta}_{\text{pullback}}(\mathbf{x}_i), \qquad \eta = \mathcal{N}(0, \mathbf{I}_n),$$

  where $\mathcal{S}_\triangle^h$ is an approximation space for the rearrangement

- Each component $\widehat{S}^k$ of $\widehat{S}$ can be computed **separately**, via smooth **convex optimization**

$$\boxed{\widehat{S}^k \in \arg\min_{S^k \in \mathcal{S}_{\triangle,k}^h} \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{2} S^k(\mathbf{x}_i)^2 - \log \partial_k S^k(\mathbf{x}_i) \right)}$$

Map parameterizations

$$\widehat{S}^k \in \arg \min_{S^k \in \mathcal{S}^h_{\triangle,k}} \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{2} S^k(\mathbf{x}_i)^2 - \log \partial_k S^k(\mathbf{x}_i) \right)$$

- In general, convex optimization
- Optimization is not needed for nonlinear separable parameterizations of the form $\widehat{S}^k(x_{1:k}) = \alpha x_k + g(x_{1:k-1})$ (just *linear regression*)
- **Connection to EnKF:** a linear parameterization of $\widehat{S}^k$ yields a particular form of EnKF with "perturbed observations"

- Choice of approximation space allows control of the bias and variance of $\widehat{S}$
- Richer parameterizations yield less bias, but potentially higher variance

Strategy: depart gradually from the linear ansatz by introducing local nonlinearities + regularization

Example: Lorenz-63

Simple example: three-dimensional Lorenz-63 system

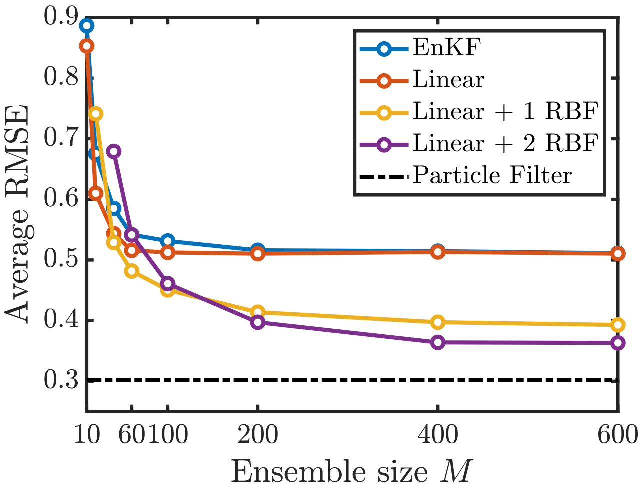
$$\begin{aligned}\frac{\mathrm{d}X_1}{\mathrm{d}t} &= \sigma(X_2 - X_1),\\ \frac{\mathrm{d}X_2}{\mathrm{d}t} &= X_1(\rho - X_3) - X_2\\ \frac{\mathrm{d}X_3}{\mathrm{d}t} &= X_1X_2 - \beta X_3\end{aligned}$$

- Chaotic setting: $\rho = 28$, $\sigma = 10$, $\beta = 8/3$
- Fully observed, with additive Gaussian observation noise $\mathcal{E}_j \sim \mathcal{N}(0, 2^2)$
- Assimilation interval $\Delta t = 0.1$
- Results computed over 2000 assimilation cycles, following spin-up

- **Map parameterizations:** $S^k(x_{1:k}) = \sum_{i \leq k} \Psi_i(x_i)$, with $\Psi_i =$ linear $+ \{$RBFs or sigmoids $\}$

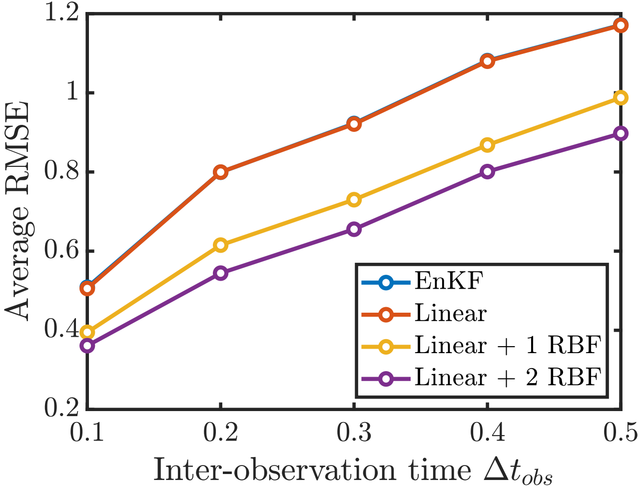
Example: Lorenz-63

Mean "tracking" error vs. ensemble size and choice of map

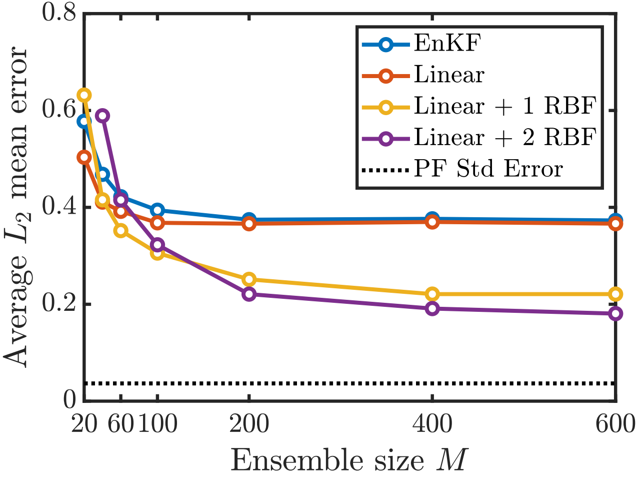


Example: Lorenz-63

How do $M \to \infty$ "plateaus" depend on assimilation interval?

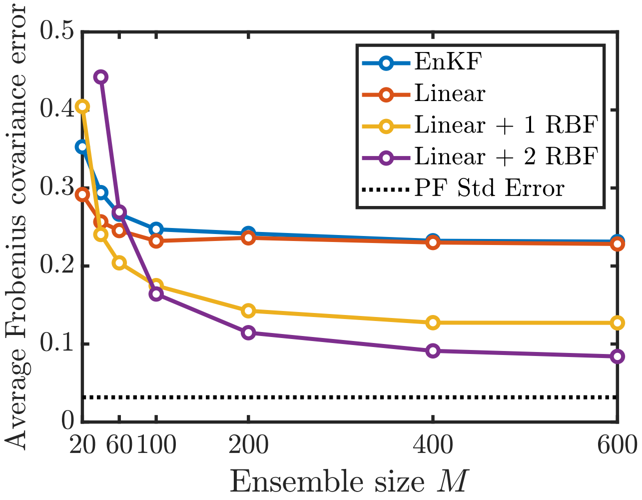


What about comparison to the *true Bayesian solution*?



Example: Lorenz-63

What about comparison to the *true Bayesian solution*?



“Localize” the map in high dimensions

- Regularize the estimator $\widehat{S}$ of S by imposing **sparsity**, e.g.,

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

- The sparsity of the kth component of S depends on the sparsity of the marginal conditional function $\pi_{\mathbf{X}_k|\mathbf{X}_{1:k-1}}(x_k|\mathbf{x}_{1:k-1})$
- **Localization heuristic:** let each $\widehat{S}^k$ depend on variables $(x_j)_{j<k}$ that are within a distance ℓ from x_k in state space. Estimate optimal ℓ offline
- Explicit link between sparsity of S and conditional independence in non-Gaussian graphical models

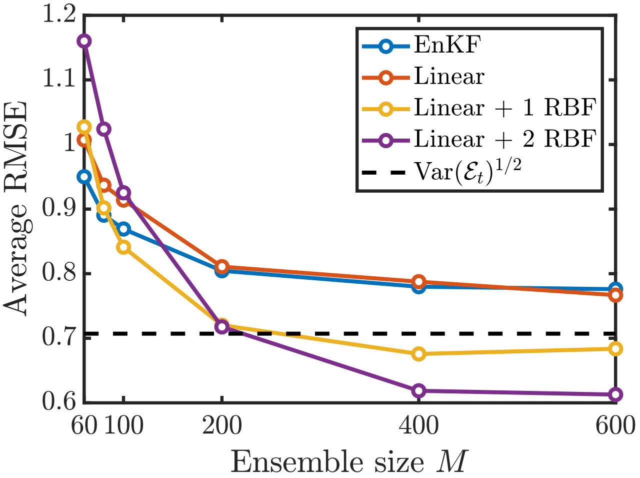
Lorenz-96 in chaotic regime (40-dimensional state)

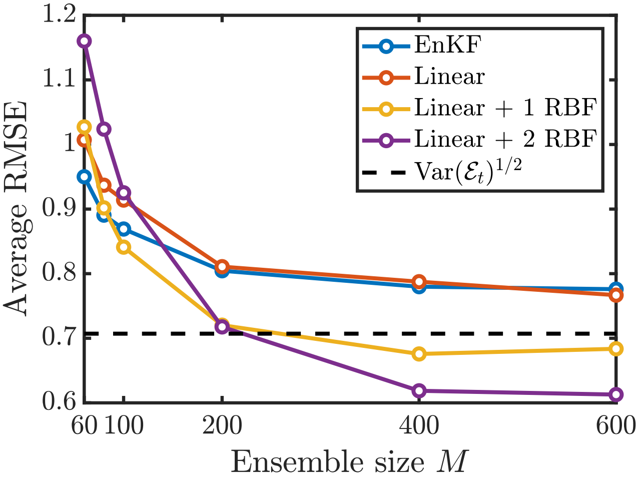
- A **hard** test-case configuration [Bengtsson et al. 2003]:

$$
\begin{aligned}
\frac{\mathrm{d}\mathbf{X}_j}{\mathrm{d}t} &= (\mathbf{X}_{j+1} - \mathbf{X}_{j-2})\mathbf{X}_{j-1} - \mathbf{X}_j + F, \qquad j = 1, \ldots, 40 \\
\mathbf{Y}_j &= \mathbf{X}_j + \mathcal{E}_j, \qquad j = 1, 3, 5 \ldots, 39
\end{aligned}
$$

  - $F = 8$ (chaotic) and $\mathcal{E}_j \sim \mathcal{N}(0, 0.5)$ (**small noise for PF**)
  - Time between observations: $\Delta_{\text{obs}} = 0.4$ (**large**)
  - Results computed over 2000 assimilation cycles, following spin-up

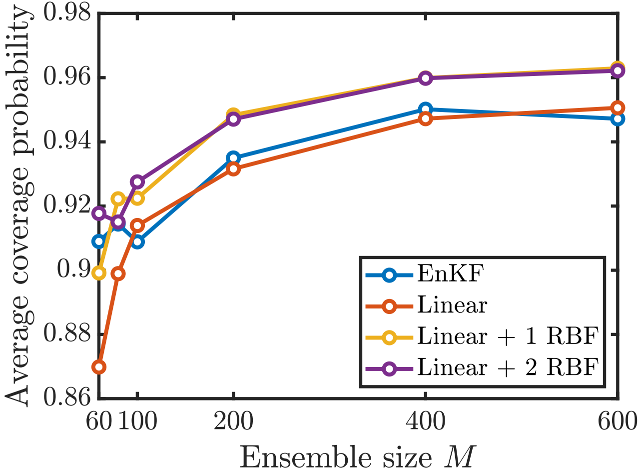
Lorenz-96: "hard" case





- The nonlinear filter is ≈ 25% more accurate in RMSE than EnKF

Lorenz-96: “hard” case



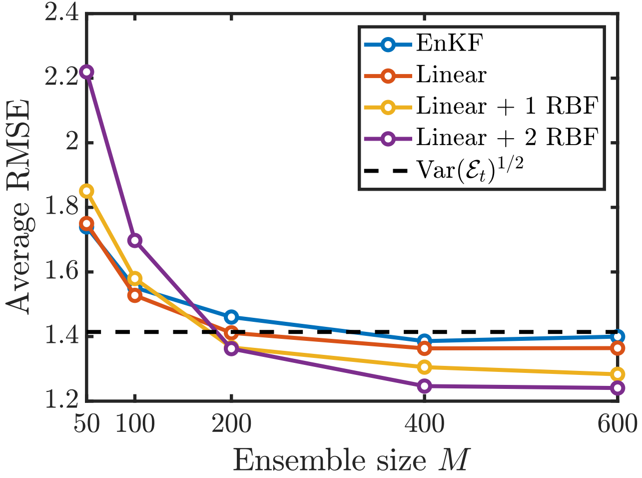
Lorenz-96: non-Gaussian noise

- A heavy-tailed noise configuration:

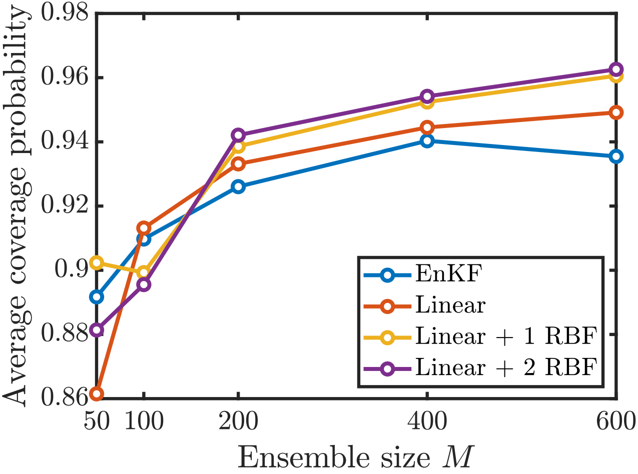
$$\begin{aligned}\frac{\mathrm{d}\mathbf{X}_j}{\mathrm{d}t} &= (\mathbf{X}_{j+1} - \mathbf{X}_{j-2})\mathbf{X}_{j-1} - \mathbf{X}_j + F, \qquad j = 1, \ldots, 40 \\ \mathbf{Y}_j &= \mathbf{X}_j + \mathcal{E}_j, \qquad j = 1, 5, 9, 13, \ldots, 37\end{aligned}$$

  - $F = 8$ (chaotic) and $\mathcal{E}_j \sim \text{Laplace}(\lambda = 1)$
  - Time between observations: $\Delta_{\text{obs}} = 0.1$
  - Results computed over 2000 assimilation cycles, following spin-up

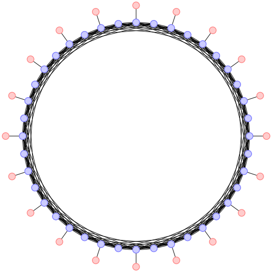
Lorenz-96: non-Gaussian noise



Lorenz-96: non-Gaussian noise



Lorenz-96: details on the filtering approximation



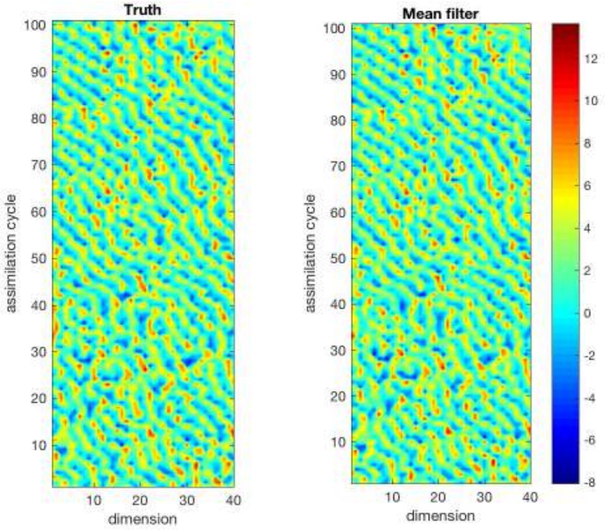
- Observations were assimilated one at a time
- Impose sparsity of the map with a 5-way interaction model *(above)*
- Separable and nonlinear parameterization of each component

$$\widehat{S}^k(x_{j_1}, \ldots, x_{j_p}, x_k) = \psi(x_{j_1}) + \ldots + \psi(x_{j_p}) + \widetilde{\psi}(x_k),$$

where $\psi(x) = a_0 + a_1 \cdot x + \sum_{i>1} a_i \exp(-(x - c_i)^2/\sigma)$.

- Much **more general** parameterizations are of course possible

Lorenz-96: tracking performance of the filter



- Simple and and localized nonlinearities have significant impact!

Remarks and questions

- Nonlinear generalization of the EnKF: move the ensemble members via local nonlinear transport maps, *no weights or degeneracy*
- Learn non-Gaussian features via nonlinear continuous transport and *convex optimization*
- Choice of map basis and **sparsity** provide regularization (e.g., *localization*)

Remarks and questions

- Nonlinear generalization of the EnKF: move the ensemble members via local nonlinear transport maps, *no weights or degeneracy*
- Learn non-Gaussian features via nonlinear continuous transport and *convex optimization*
- Choice of map basis and **sparsity** provide regularization (e.g., *localization*)

- In principle, filter is consistent as $\mathcal{S}^h_\triangle$ is enriched and $M \to \infty$, but a careful *error analysis* is needed!
- What is a good or even optimal choice of $\mathcal{S}^h_\triangle$ for any fixed ensemble size M?
- Can regularization penalties (e.g., ℓ_1) help identify sparse structure, and/or learn sparse maps from few samples?

Regularized estimation of S

For simplicity, consider map components $S^k(\mathbf{x}) = \sum_j \beta_j \psi_j(x_{1:k-1}) + \alpha_k x_k$

$$\widehat{S}^k \in \arg \min_{S^k \in \mathcal{S}^h_{\triangle,k}} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2} S^k(\mathbf{x}_i)^2 - \log \partial_k S^k(\mathbf{x}_i) \right) + \lambda_N \|\boldsymbol{\beta}\|_1$$

Assume sub-Gaussian π and basis functions $\psi_j(\mathbf{x})$

Theorem [BZM]

For polynomial maps of degree m with sparsity s, with high probability

$$\mathbb{E}_\pi \left[D_{KL}\left(\pi(\mathbf{x}_k | \mathbf{x}_{1:k-1}) \,||\, \widehat{S}^\sharp_k \eta \right) \right] \lesssim \sqrt{\frac{s^2 m \log k}{N}}$$

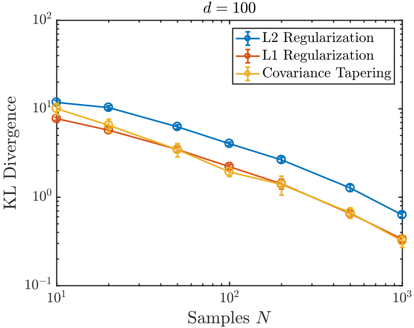
Takeaways

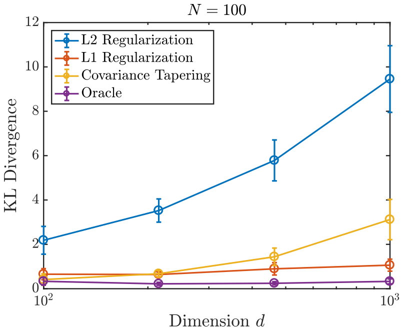
- Accurate estimation is feasible in high dimensions with $N \ll k$
- From factorization property of density, error in conditionals ensures

$$D_{KL}(\pi || \widehat{S}^\sharp \eta) \lesssim d \sqrt{\frac{s^2 m \log d}{N}}$$

Linear–Gaussian problem

- *Prior*: $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma_{pr})$ with exponential covariance
- *Likelihood*: Local observations $\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathcal{E}$ with $\mathcal{E} \sim \mathcal{N}(0, I)$

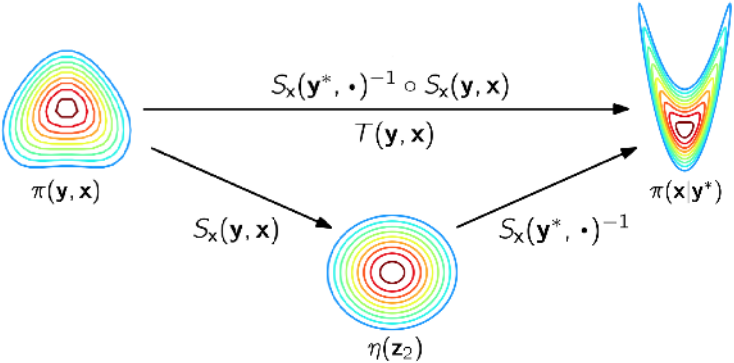




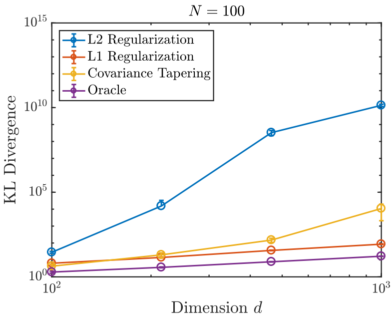
Takeaway

- Learning sparse prior-to-posterior map matches oracle scaling

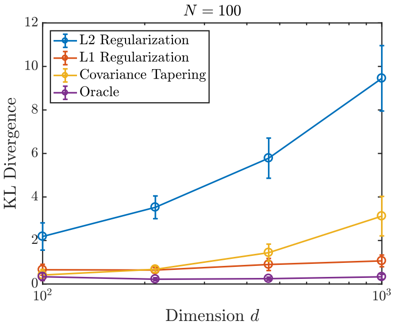
Compare two approaches for posterior sampling



Compare two approaches for posterior sampling



$$\mathbf{X}|\mathbf{y}^* \sim \widehat{S}_{\mathsf{X}}(\mathbf{y}^*, \cdot)_\sharp^{-1}\eta$$



$$\mathbf{X}|\mathbf{y}^* \sim \widehat{T}_\sharp \pi_{\mathbf{y},\mathbf{x}} \text{ for}$$
$$\widehat{T} = \widehat{S}_{\mathsf{X}}(\mathbf{y}^*, \cdot)^{-1} \circ \widehat{S}_{\mathsf{X}}(\cdot, \cdot)$$

- Propagating forecast through composed maps has lower error
- This is in fact a **general** approach to likelihood-free inference/ABC

Comments on transport-based ensemble filtering

- There is also a "square root" version of the nonlinear ensemble filter
- Continuous-time formulations?
- Nonlinear ensemble smoothers
- Open questions about estimation and regularization of continuous transport maps:
  - How to choose and *adapt* approximation space/basis to the forecast ensemble?
  - Properties of the estimator $\widehat{T}$, e.g., consistency, sample size requirements and scaling
  - Other forms of low-dimensional structure and regularization
- Applications to inference in "likelihood-free" settings

References

- A. Spantini, R. Baptista, Y. Marzouk. "Coupling techniques for nonlinear ensemble filtering." arXiv:1907.00389.
- D. Bigoni, O. Zahm, A. Spantini, Y. Marzouk. "Greedy inference with layers of lazy maps." arXiv:1906.00031.
- O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk. "Certified dimension reduction in nonlinear Bayesian inverse problems." arXiv:1807.03712.
- A. Spantini, D. Bigoni, Y. Marzouk. "Inference via low-dimensional couplings." *JMLR* 19(66): 1–71, 2018.
- M. Parno, Y. Marzouk, "Transport map accelerated Markov chain Monte Carlo." *SIAM JUQ* 6: 645–682, 2018.
- G. Detomasso, T. Cui, A. Spantini, Y. Marzouk, R. Scheichl, "A Stein variational Newton method." NeurIPS 2018.
- R. Morrison, R. Baptista, Y. Marzouk. "Beyond normality: learning sparse probabilistic graphical models in the non-Gaussian setting." NeurIPS 2017.
- Y. Marzouk, T. Moselhy, M. Parno, A. Spantini, "An introduction to sampling via measure transport." *Handbook of Uncertainty Quantification*, R. Ghanem, D. Higdon, H. Owhadi, eds. Springer (2016). arXiv:1602.05023.

- **General python code at** `http://transportmaps.mit.edu`