Transport methods for nonlinear filtering and likelihood-free inference

Youssef Marzouk, joint work with Ricardo Baptista,
Alessio Spantini, & Olivier Zahm

Department of Aeronautics and Astronautics
Center for Computational Science and Engineering
Statistics and Data Science Center
Massachusetts Institute of Technology
http://uqgroup.mit.edu

Sea Ice Modeling and Data Assimilation (SIMDA) Seminar

6 October 2020

Main ideas #1

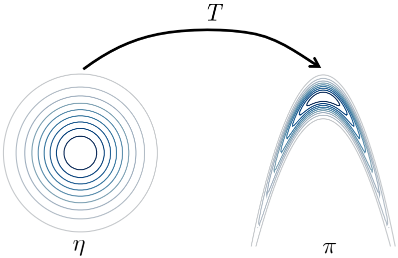
- *Online* **Bayesian inference** in *dynamical* models with *sequential data* is central to our proposed work
  - Data assimilation, in the broadest sense
- New nonlinear ensemble schemes, based on transport, can provide a *consistent* approach to inference in general *non-Gaussian* settings
  - Generalizations of the ensemble Kalman filter (EnKF)
- These schemes are methods for **likelihood-free inference** (LFI) or **approximate Bayesian computation** (ABC), and have broader utility!

Main ideas #2

- Application to high-dimensional and disparate data requires **detecting** and **exploiting low-dimensional structure**
  - Many varieties of structure, e.g.: sparsity and conditional independence (especially given disparate *data sources*), smoothness/low rank, piecewise smoothness, hierarchical low rank, multiscale structure
  - Structure in parameters/state *and* in data!
  - Relate also to *data-driven discretizations* of inverse problems

- These new inference methods can also facilitate optimal experimental design

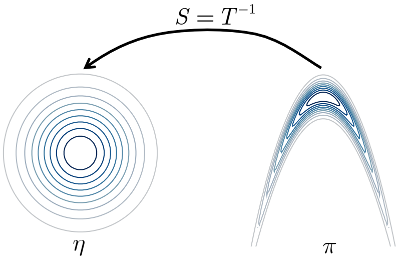
Deterministic couplings of probability measures



Core idea

- Choose a *reference distribution* η (e.g., standard Gaussian)
- Seek a transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_\sharp \eta = \pi$

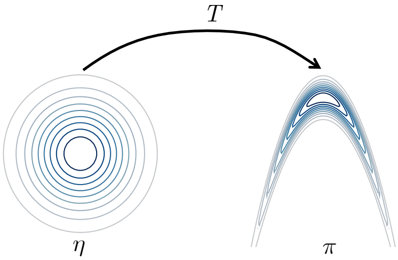
Deterministic couplings of probability measures



Core idea

- Choose a *reference distribution* η (e.g., standard Gaussian)
- Seek a transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_\sharp \eta = \pi$
- Equivalently, find $S = T^{-1}$ such that $S_\sharp \pi = \eta$

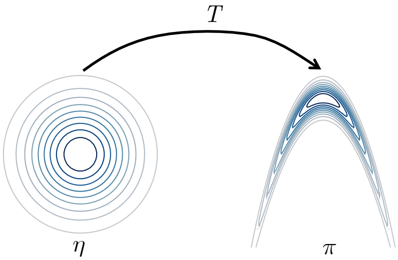
Deterministic couplings of probability measures



Core idea

- Choose a *reference distribution* η (e.g., standard Gaussian)
- Seek a transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_\sharp \eta = \pi$
- Equivalently, find $S = T^{-1}$ such that $S_\sharp \pi = \eta$
- In principle, enables *exact* (independent, unweighted) sampling!

Deterministic couplings of probability measures



Core idea

- Choose a *reference distribution* η (e.g., standard Gaussian)
- Seek a transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_\sharp \eta = \pi$
- Equivalently, find $S = T^{-1}$ such that $S_\sharp \pi = \eta$
- Satisfying these conditions only **approximately** can still be useful!

Choice of transport map

Consider the triangular **Knothe-Rosenblatt rearrangement** on $\mathbb{R}^d$

$$S(\mathbf{x}) = \begin{bmatrix} S^1(x_1) \\ S^2(x_1, x_2) \\ \vdots \\ S^d(x_1, x_2, \dots, x_d) \end{bmatrix}$$

1. Unique S s.t. $S_\sharp \pi = \eta$ exists under mild conditions on π and η
2. Map is easily invertible and Jacobian ∇S is simple to evaluate
3. Monotonicity is essentially one-dimensional: $\partial_{x_k} S^k > 0$
4. Each component S^k characterizes one marginal conditional

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2|x_1)\cdots\pi(x_d|x_1,\dots,x_{d-1})$$

Ubiquity of triangular maps

Many “flows” proposed in ML are special cases of triangular maps, e.g.,

- NICE: Nonlinear independent component estimation [Dinh et al. 2015]
$$S^k(x_1,\ldots,x_k) = \mu_k(\mathbf{x}_{x<k}) + x_k$$
- Inverse autoregressive flow [Dinh et al. 2017]
$$S^k(x_1,\ldots,x_k) = (1-\sigma_k(\mathbf{x}_{x<k}))\mu_k(\mathbf{x}_{x<k}) + x_k\sigma_k(\mathbf{x}_{x<k})$$
- Masked autoregressive flow [Papamakarios et al. 2017]
$$S^k(x_1,\ldots,x_k) = \mu_k(\mathbf{x}_{x<k}) + x_k \exp(\alpha_k(\mathbf{x}_{x<k}))$$
- Sum-of-squares polynomial flow [Jaini et al. 2019]
$$S^k(x_1,\ldots,x_k) = a_k(\mathbf{x}_{x<k}) + \int_0^{x_k} \sum_{\kappa=1}^{p} (\mathrm{poly}(t;\mathbf{a}_{\kappa,k}(\mathbf{x}_{x<k}))^2 dt$$

Ubiquity of triangular maps

Many "flows" proposed in ML are special cases of triangular maps, e.g.,

- NICE: Nonlinear independent component estimation [Dinh et al. 2015]
$$S^k(x_1, \ldots, x_k) = \mu_k(\mathbf{x}_{x<k}) + x_k$$
- Inverse autoregressive flow [Dinh et al. 2017]
$$S^k(x_1, \ldots, x_k) = (1 - \sigma_k(\mathbf{x}_{x<k}))\mu_k(\mathbf{x}_{x<k}) + x_k \sigma_k(\mathbf{x}_{x<k})$$
- Masked autoregressive flow [Papamakarios et al. 2017]
$$S^k(x_1, \ldots, x_k) = \mu_k(\mathbf{x}_{x<k}) + x_k \exp(\alpha_k(\mathbf{x}_{x<k}))$$
- Sum-of-squares polynomial flow [Jaini et al. 2019]
$$S^k(x_1, \ldots, x_k) = a_k(\mathbf{x}_{x<k}) + \int_0^{x_k} \sum_{\kappa=1}^{p} (\text{poly}(t; \mathbf{a}_{\kappa,k}(\mathbf{x}_{x<k}))^2 dt$$
- Many **ad hoc choices** and **challenging optimization problems . . .**

How to construct triangular maps?

Some past work: "maps from densities," i.e., *variational inference* with the direct map T [Moselhy & M 2012]

How to construct triangular maps?

Some past work: "maps from densities," i.e., *variational inference* with the direct map T [Moselhy & M 2012]

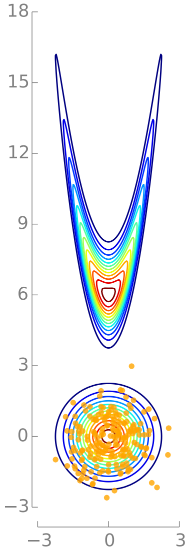
$$\min_{T \in \mathcal{T}_{\triangle}^{h}} \mathcal{D}_{KL}(\, T_{\sharp}\, \eta \,||\, \pi \,) = \min_{T \in \mathcal{T}_{\triangle}^{h}} \mathcal{D}_{KL}(\, \eta \,||\, T_{\sharp}^{-1}\, \pi \,)$$

- π is the "target" density on $\mathbb{R}^n$; η is, e.g., $\mathcal{N}(0, \mathbf{I}_n)$
- $\mathcal{T}_{\triangle}^{h}$ is a set of monotone lower triangular maps
  - $\mathcal{T}_{\triangle}^{h \to \infty}$ contains the *Knothe–Rosenblatt* rearrangement
- Expectation is with respect to the *reference* measure η
  - Compute via, e.g., Monte Carlo, sparse quadrature
- Use unnormalized evaluations of π and its gradients
- No MCMC or importance sampling
- In general non-convex, unless π is log-concave

Illustrative example

$$\min_{T} \mathbb{E}_{\eta}[-\log \pi \circ T - \sum_{k} \log \partial_{x_k} T^k]$$

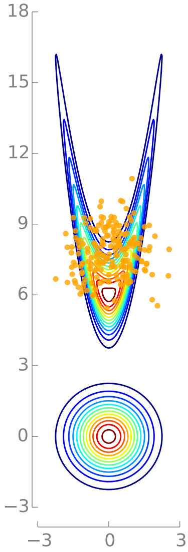
- Parameterized map $T \in \mathcal{T}_{\triangle}^{h} \subset \mathcal{T}_{\triangle}$
- Optimize over coefficients of parameterization
- Use gradient-based optimization
- The posterior is in the tail of the reference



Illustrative example

$$\min_T \mathbb{E}_\eta [-\log \pi \circ T \; - \sum_k \log \partial_{x_k} T^k]$$

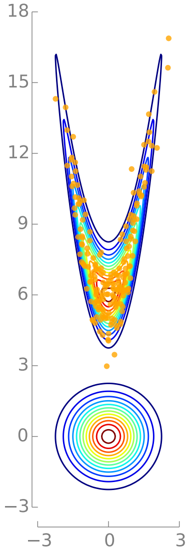
- Parameterized map $T \in \mathcal{T}_\triangle^h \subset \mathcal{T}_\triangle$
- Optimize over coefficients of parameterization
- Use gradient-based optimization
- The posterior is in the tail of the reference



Illustrative example

$$\min_T \mathbb{E}_\eta[-\log \pi \circ T - \sum_k \log \partial_{x_k} T^k]$$

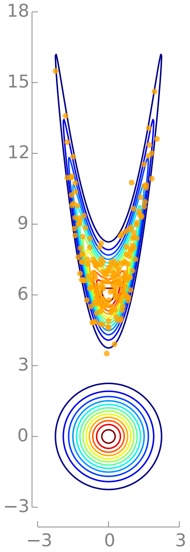
- Parameterized map $T \in \mathcal{T}_\triangle^h \subset \mathcal{T}_\triangle$
- Optimize over coefficients of parameterization
- Use gradient-based optimization
- The posterior is in the tail of the reference



Illustrative example

$$\min_T \mathbb{E}_\eta[-\log \pi \circ T \;-\sum_k \log \partial_{x_k} T^k\,]$$

- Parameterized map $T \in \mathcal{T}_\triangle^h \subset \mathcal{T}_\triangle$
- Optimize over coefficients of parameterization
- Use gradient-based optimization
- The posterior is in the tail of the reference



How to construct triangular maps?

Alternative formulation: **“maps from samples”** [Parno PhD thesis, 2014]

- **Given samples $(\mathbf{x}^i)_{i=1}^M \sim \pi$:** find components via convex (wrt S^k) constrained minimization:

$$\min_S \, D_{KL}(\pi || S^\sharp \eta) \quad \Leftrightarrow \quad \min_{S^k \,:\, \partial_k S^k > 0} \mathbb{E}_\pi \left[\frac{1}{2} S^k(\mathbf{x}_{1:k})^2 - \log \partial_k S^k(\mathbf{x}_{1:k}) \right] \; \forall k$$

- Approximate $\mathbb{E}_\pi$ given i.i.d. samples from π: KL minimization equivalent to maximum likelihood estimation

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

An underlying challenge: maps in high dimensions

- Major bottleneck: representation of the map, e.g., cardinality of the map basis
- How to make the construction/representation of high-dimensional transports tractable?

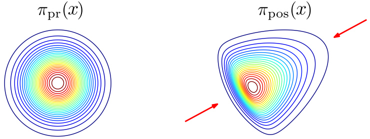
Low-dimensional structure of transport maps

An underlying challenge: maps in high dimensions

- Major bottleneck: representation of the map, e.g., cardinality of the map basis
- How to make the construction/representation of high-dimensional transports tractable?

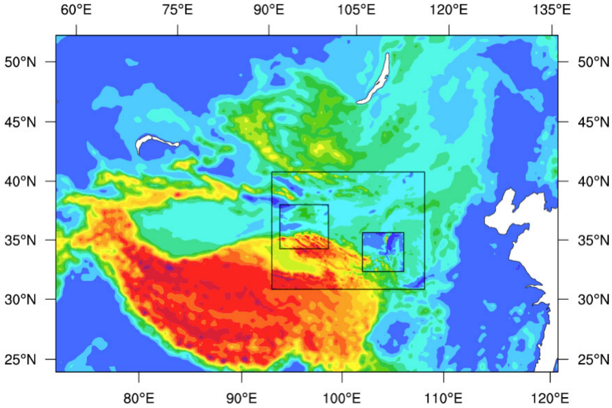
Main ideas:

1. Exploit **Markov structure** of the target distribution
   - Leads to sparsity and/or decomposability of transport maps [Spantini, Bigoni, & M JMLR 2018]
2. Exploit **low rank** structure
   - Common in *inverse problems*. Near-identity or "lazy" maps [Brennan et al. NeurIPS 2020, Zahm et al. 2018]



Remainder of this talk: sequential and "likelihood-free"

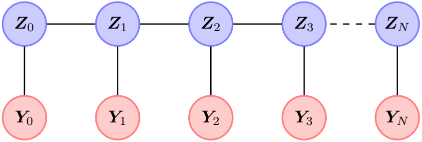
Can transport help solve **sequential inference** problems where key **density functions cannot be evaluated?**



[image: NCAR]

Formalize as Bayesian filtering

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



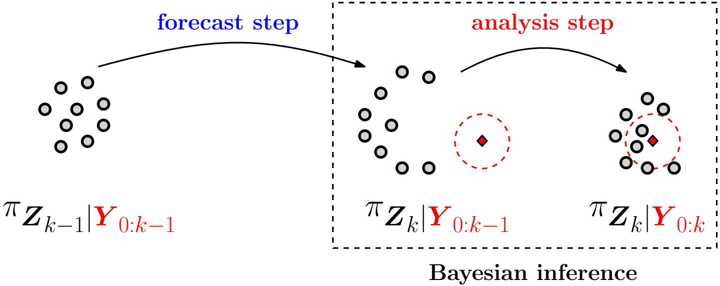
- Focus on recursively approximating the filtering distribution: $\pi_{\mathbf{Z}_k \mid \mathbf{y}_{0:k}} \rightarrow \pi_{\mathbf{Z}_{k+1} \mid \mathbf{y}_{0:k+1}}$ (marginals of the full Bayesian solution)

Problem setting

- Consider the filtering of state-space models with:
  1. High-dimensional states
  2. Challenging nonlinear dynamics
  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
- These constraints reflect challenges faced in the **sea ice prediction** problem. . .

Ensemble Kalman filter

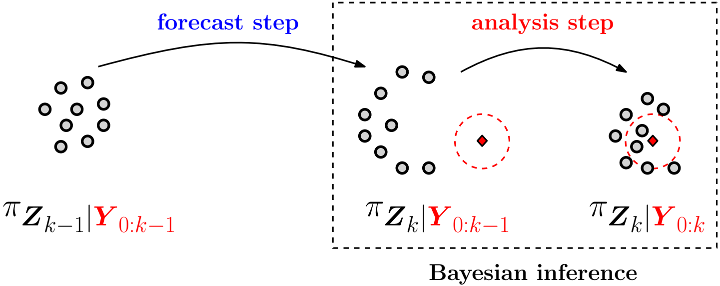
- State-of-the-art results (in terms of tracking) are often obtained with the ensemble Kalman filter (EnKF)



- 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 often obtained with the ensemble Kalman filter (EnKF)

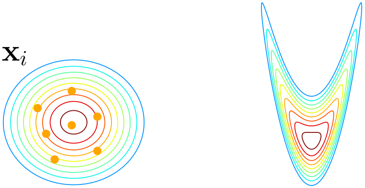


- Move samples via an **affine** transformation; no weights or resampling!
- Yet ultimately **inconsistent**: does not converge to the true posterior

Can we *improve* and *generalize* the EnKF while preserving scalability?

Assimilation step

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



$\pi_{\boldsymbol{X}}$ prior

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

- $\pi_{\mathbf{X}}$ 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

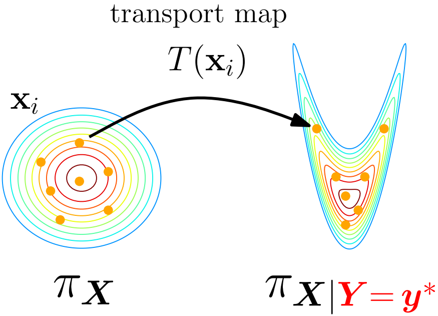
Goal: sample the posterior given only prior samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ and the ability to simulate data $\mathbf{y}_i|\mathbf{x}_i$

Inference as transportation of measure

- Seek a map T that pushes forward prior to posterior

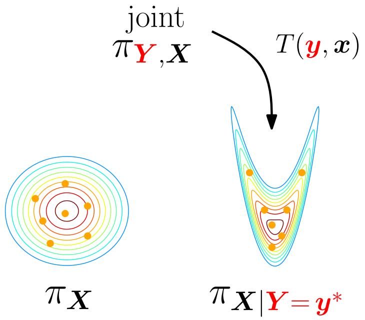
$$(\mathbf{x}_1, \ldots, \mathbf{x}_M) \sim \pi_{\mathbf{X}} \Longrightarrow (T(\mathbf{x}_1), \ldots, T(\mathbf{x}_M)) \sim \pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$$

- 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

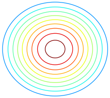
Couple the joint distribution with a standard normal

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...

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:

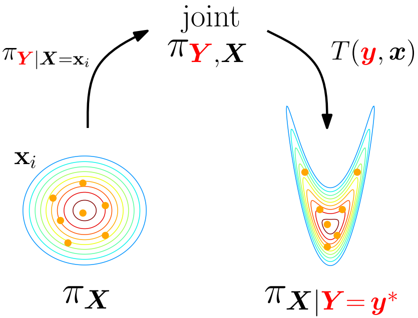
$$S^{\mathbf{X}} \text{ pushes } \pi_{\mathbf{Y},\mathbf{X}} \text{ to } \mathcal{N}(0, \mathbf{I}_n)$$

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

- 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 likelihood-free inference 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:

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

- We propose a simple estimator $\widehat{T}$ of T:

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

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

- This is simply the earlier "maps from samples" approach!

Map parameterizations

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

- 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

Example: Lorenz-63

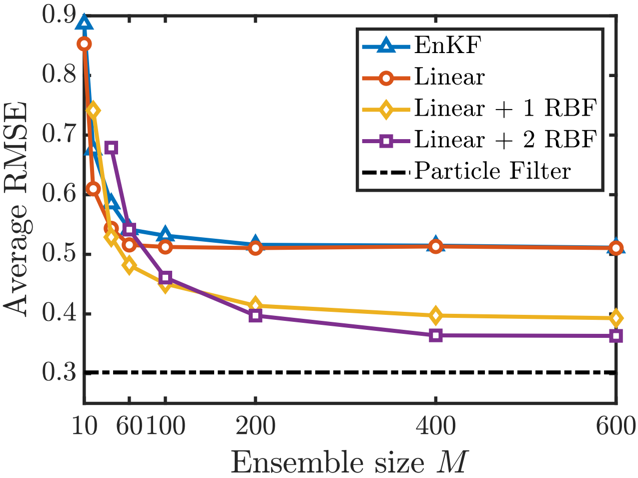
Simple example: three-dimensional Lorenz-63 system

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

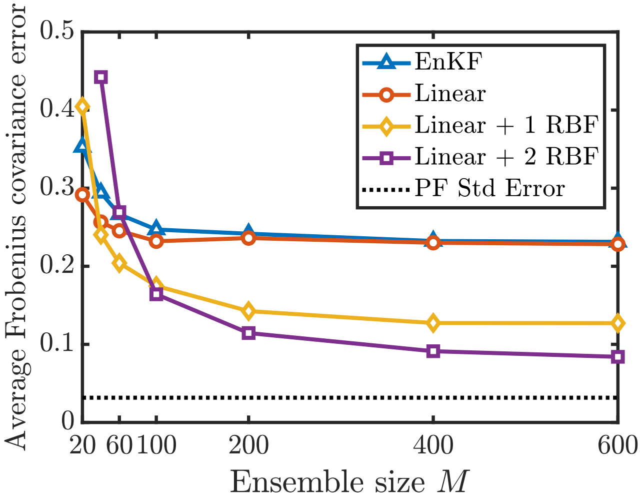
- Chaotic setting: $\rho = 28$, $\sigma = 10$, $\beta = 8/3$
- Fully observed, with additive Gaussian observation noise $\mathcal{E}_j \sim \mathcal{N}(0, 2^2)$
- Assimilation interval $\Delta t = 0.1$
- Results computed over 2000 assimilation cycles, following spin-up
- **Map parameterizations:** $S^k(x_{1:k}) = \sum_{i \leq k} \Psi_i(x_i)$, with $\Psi_i =$ linear + $\{$RBFs or sigmoids $\}$

Example: Lorenz-63

Mean “tracking” error vs. ensemble size and choice of map



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.,

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

- 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 described in

  [Inference via low-dimensional couplings, Spantini/Bigoni/M JMLR 2018]

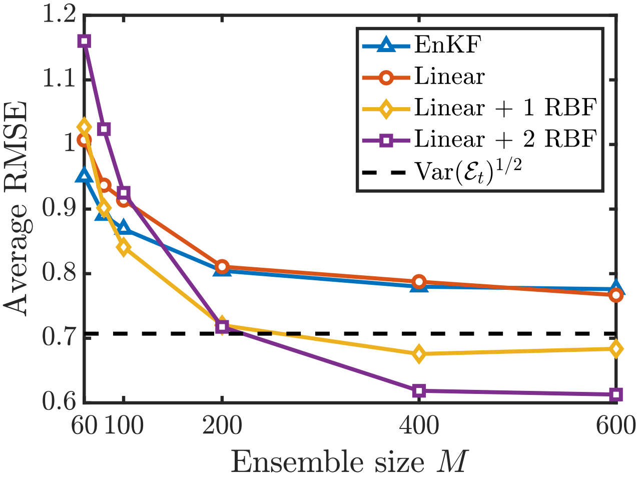
Lorenz-96 in chaotic regime (40-dimensional state)

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

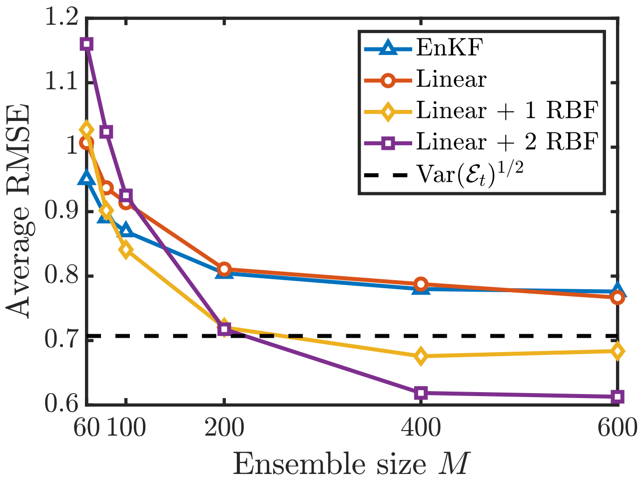
$$\begin{aligned}\frac{\mathrm{d}\mathbf{X}_j}{\mathrm{d}t} &= (\mathbf{X}_{j+1} - \mathbf{X}_{j-2})\mathbf{X}_{j-1} - \mathbf{X}_j + F, \qquad j = 1, \dots, 40 \\ \mathbf{Y}_j &= \mathbf{X}_j + \mathcal{E}_j, \qquad j = 1, 3, 5 \dots, 39\end{aligned}$$

  - $F = 8$ (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

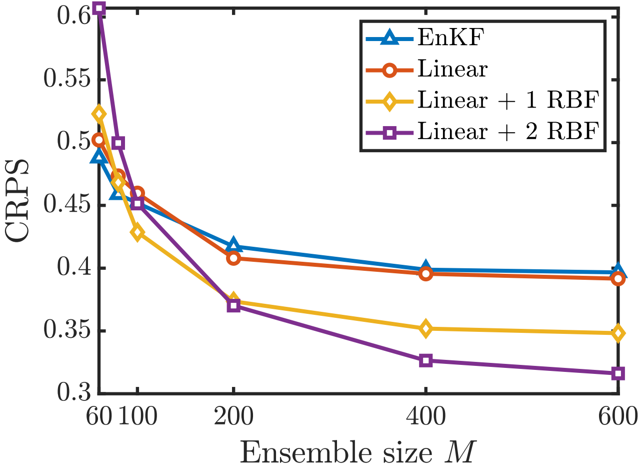


Lorenz-96: “hard” case



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

Lorenz-96: “hard” case



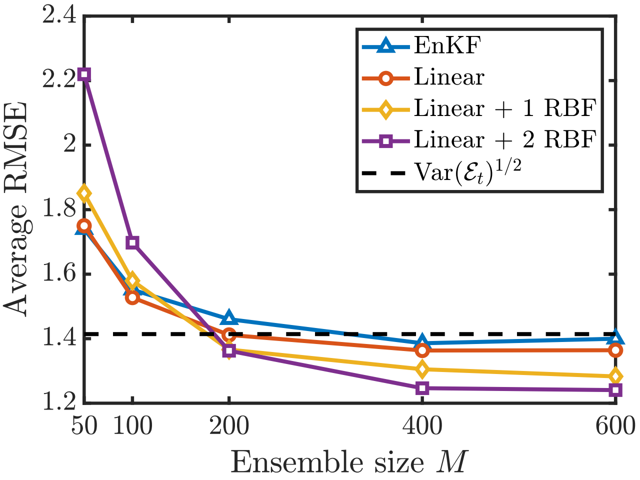
Lorenz-96: non-Gaussian noise

- A heavy-tailed noise configuration:

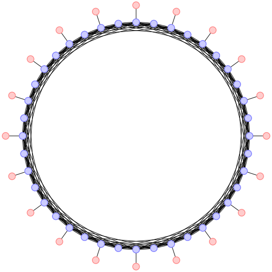
$$\begin{aligned} \frac{\mathrm{d}\mathbf{X}_j}{\mathrm{d}t} &= (\mathbf{X}_{j+1} - \mathbf{X}_{j-2})\mathbf{X}_{j-1} - \mathbf{X}_j + F, \qquad j = 1, \dots, 40 \\ \mathbf{Y}_j &= \mathbf{X}_j + \mathcal{E}_j, \qquad j = 1, 5, 9, 13, \dots, 37 \end{aligned}$$

  - $F = 8$ (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: details on the filtering approximation



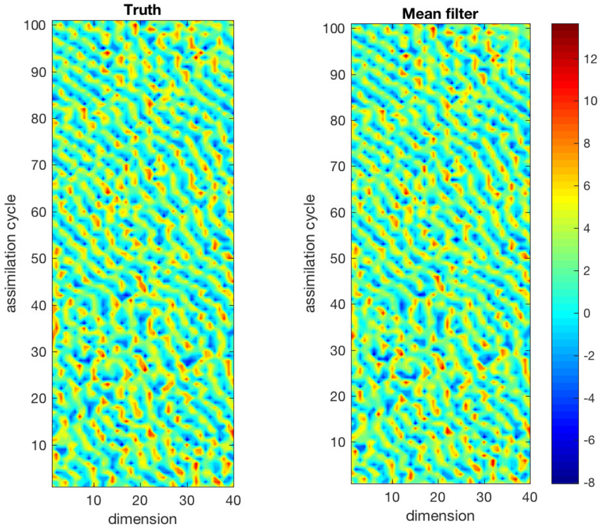
- Impose sparsity of the map with a 5-way interaction model *(above)*
- Separable and nonlinear parameterization of each component

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

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

- **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*)

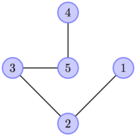
Remarks and questions

- **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*)

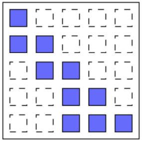
- In principle, filter is consistent as $\mathcal{S}_{\triangle}^{h}$ is enriched and $M \to \infty$. But what is a good choice of $\mathcal{S}_{\triangle}^{h}$ for any fixed ensemble size M?
- Are there better estimators than maximum likelihood? What are the finite-sample *statistical properties* of candidate estimators, and properties of the associated optimization problems?
- How to relate map structure/parameterization to the underlying dynamics, observation operators, and data?

Theorem [Spantini et al. 2018]

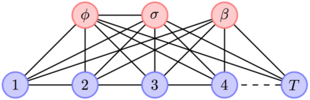
Conditional independence properties of π (encoded by graph $\mathcal{G}$) define a lower bound on the sparsity of S such that $S^\sharp \eta = \pi$



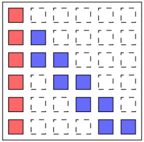
$\mathcal{G}$ of π



Sparsity of S



State-space model



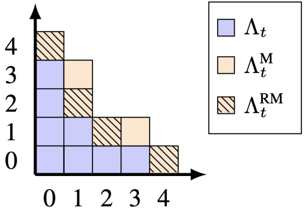
Sparsity of S

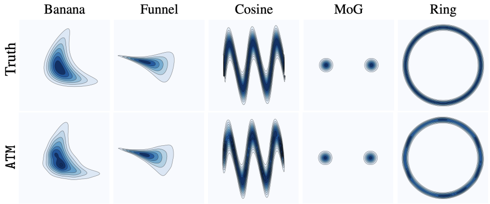
Main idea: Discover sparse structure using `ATM` algorithm

Goal: Approximate map given n i.i.d. samples from π

Greedy enrichment procedure

- Look for sparse expansion $f(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\Lambda} c_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\mathbf{x})$
- Use tensor-product Hermite functions $\psi_{\boldsymbol{\alpha}}(\mathbf{x}) = P_{\alpha_j}(\mathbf{x})\exp(-\|\mathbf{x}\|^2/2)$
- Add one element to set of active multi-indices Λ_t at a time
- Restrict Λ_t to be downward closed
- Search for new features in the reduced margin of Λ_t





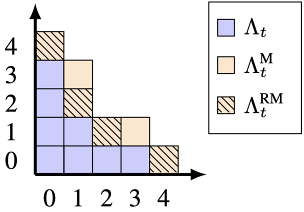
Adaptive transport map (ATM) algorithm (Baptista et al. 2020)

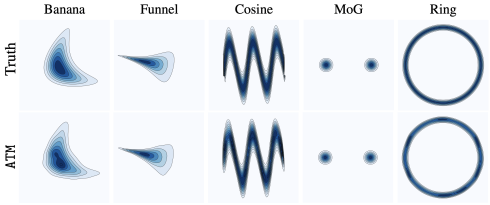
Goal: Approximate map given n i.i.d. samples from π

Initialize $\Lambda_t = \emptyset$ (i.e., $f_0 = 0$)
For $t = 0, \ldots, m$

1. Find reduced margin Λ_t^{RM} of Λ_t
2. Add new feature $\Lambda_{t+1} = \Lambda_t \cup \alpha_{t+1}^*$ for $\alpha_{t+1}^* \in \Lambda_t^{RM}$
3. Update approximation $f_{t+1} = \mathrm{argmin}_{f \in \mathrm{span}(\psi_{\Lambda_{t+1}})} \mathcal{L}_k(f)$

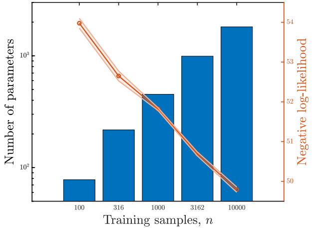
In practice, we can choose m (# of features) via cross-validation



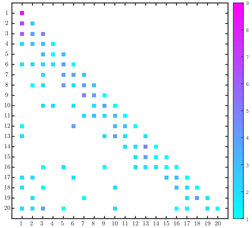


Numerical example: Lorenz-96 data

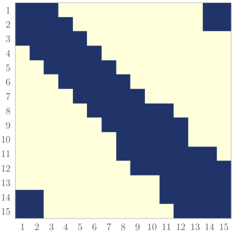
- Distribution of state at a fixed time starting from a Gaussian initial condition



Log-like and m vs n



Sparsity of S: $n = 316$



Conditional independence

Takeaways:

- `ATM` discovers conditional independence structure in the state
- Natural semi-parametric method that gradually increases m with n

Back to conditional maps: link to ABC

"Analysis" step of the ensemble filtering scheme is an instance of **likelihood-free inference** or **approximate Bayesian computation** (ABC):

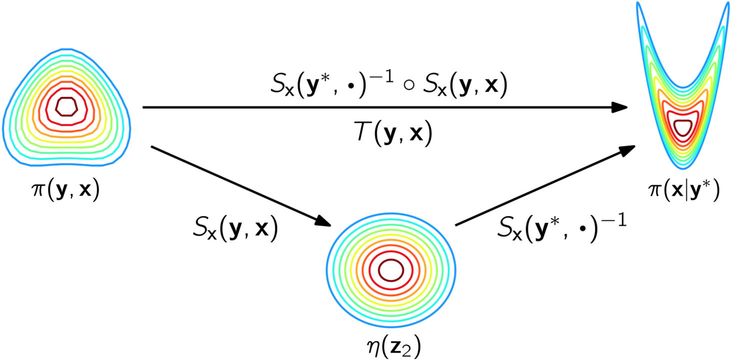
- Central idea: only need to simulate from $\pi_{\mathbf{Y},\mathbf{X}}$ in order to construct $\widehat{T}$ and thus draw samples from $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$

Back to conditional maps: link to ABC

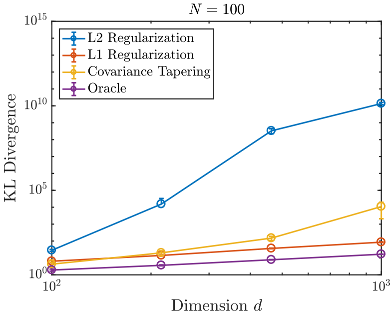
"Analysis" step of the ensemble filtering scheme is an instance of **likelihood-free inference** or **approximate Bayesian computation** (ABC):

- Central idea: only need to simulate from $\pi_{\mathbf{Y},\mathbf{X}}$ in order to construct $\widehat{T}$ and thus draw samples from $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$
- The map $\widehat{T}$ has some remarkable properties. . .

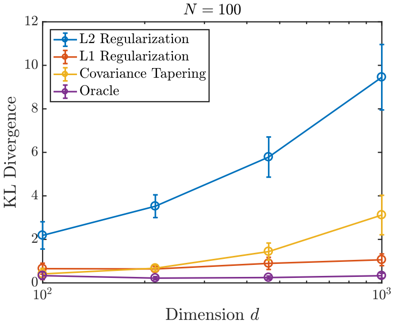
Compare two approaches for posterior sampling



Compare two approaches for posterior sampling



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



$\mathbf{X}|\mathbf{y}^* \sim \widehat{T}_\sharp \pi_{\mathbf{y},\mathbf{x}}$ for

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

- Propagating the joint prior through composed maps has lower error!

Composed maps for ABC

- Simple and incomplete examples:
  - Remove dependence of the *mean* on $\mathbf{y}$: $\widehat{S}^X(\mathbf{y}, \mathbf{x}) = \mathbf{x} - \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}]$
    - If $\mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] \approx \mathbf{c} + \boldsymbol{\beta}\mathbf{y}$, we have the EnKF
  - Remove dependence of the *mean* and *variance* on $\mathbf{y}$: $\widehat{S}^X(\mathbf{y}, \mathbf{x}) = \mathbb{V}\mathrm{ar}(\mathbf{X}|\mathbf{Y} = \mathbf{y})^{-1/2}(\mathbf{x} - \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}])$

Composed maps for ABC

- Simple and incomplete examples:
  - Remove dependence of the *mean* on $\mathbf{y}$: $\widehat{S}^X(\mathbf{y}, \mathbf{x}) = \mathbf{x} - \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}]$
    - If $\mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] \approx \mathbf{c} + \boldsymbol{\beta}\mathbf{y}$, we have the EnKF
  - Remove dependence of the *mean* and *variance* on $\mathbf{y}$: $\widehat{S}^X(\mathbf{y}, \mathbf{x}) = \mathbb{V}\mathrm{ar}(\mathbf{X}|\mathbf{Y} = \mathbf{y})^{-1/2}(\mathbf{x} - \mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}])$
- The transport map framework offers a much more general set of possibilities, which in principle includes the exact transformation
- **Current work:** Alternative optimization objectives (hence map estimators $\widehat{T}$) and tailored parameterizations
- Also useful for *optimal Bayesian experimental design!*

SIMDA plans, connections, and discussion

- Nonlinear ensemble algorithms for filtering, smoothing, and joint state-parameter inference (quite general)
- Generative modeling with transport maps: will we need to build priors—spatiotemporal statistical models for sea ice thickness—from data?
- Likelihood-free inference problems:
  - Are there sea ice problems where we will need to extract *conditional relationships* from large data sets? Conditional density estimation, conditional simulation.
  - High-dimensional and disparate sources of data
- Are there strong/essential non-Gaussianities in the sea ice problem? Tail behaviors?
- Explore cost-accuracy tradeoffs: balance posterior fidelity with computational effort
- Use conditional density estimates in optimal experimental design

Thanks for your attention!

References

- R. Baptista, O. Zahm, Y. Marzouk. " An adaptive transport framework for joint and conditional density estimation." arXiv:2009.10303, 2020.
- J. Zech, Y. Marzouk. "Sparse approximation of triangular transports on bounded domains." arXiv:2006.06994, 2020.
- N. Kovachki, R. Baptista, B. Hosseini and Y. Marzouk, "Conditional sampling with monotone GANs," arXiv:2006.06755, 2020.
- A. Spantini, R. Baptista, Y. Marzouk. "Coupling techniques for nonlinear ensemble filtering." arXiv:1907.00389, 2020.
- M. Brennan, D. Bigoni, O. Zahm, A. Spantini, Y. Marzouk. "Greedy inference with structure-exploiting lazy maps." *NeurIPS 2020*, arXiv:1906.00031.
- O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk. "Certified dimension reduction in nonlinear Bayesian inverse problems." arXiv:1807.03712, 2019.
- 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.
- R. Morrison, R. Baptista, Y. Marzouk. "Beyond normality: learning sparse probabilistic graphical models in the non-Gaussian setting." NeurIPS 2017.