High-dimensional Bayesian filtering with nonlinear local couplings

Ricardo Baptista, Daniele Bigoni, Alessio Spantini, Youssef Marzouk

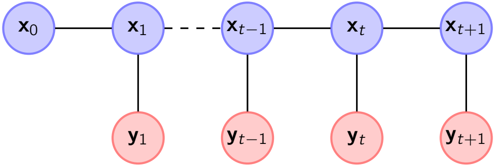
Massachusetts Institute of Technology
Department of Aeronautics & Astronautics

14th International EnKF Workshop
Voss, Norway

June 4, 2019

Non-Gaussian state-space model

- Model dynamics - transition kernel: $\mathbf{x}_t \sim f(\cdot|\mathbf{x}_{t-1})$
- Observations - likelihood model: $\mathbf{y}_t \sim g(\cdot|\mathbf{x}_t)$

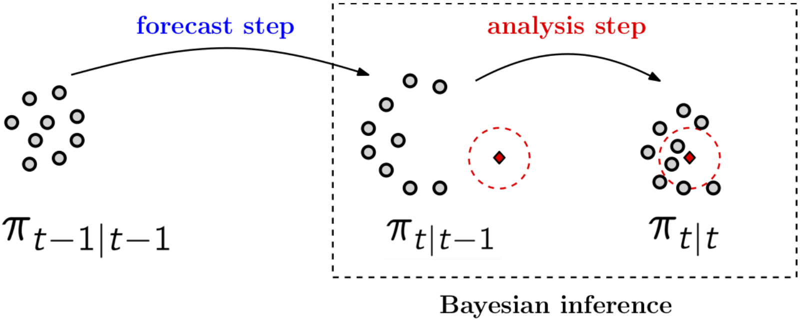


Goal: Recursively estimate filtering distributions $\pi_{t|t} := \pi(\mathbf{x}_t|\mathbf{y}_1, \ldots, \mathbf{y}_t)$

Challenges of nonlinear filtering

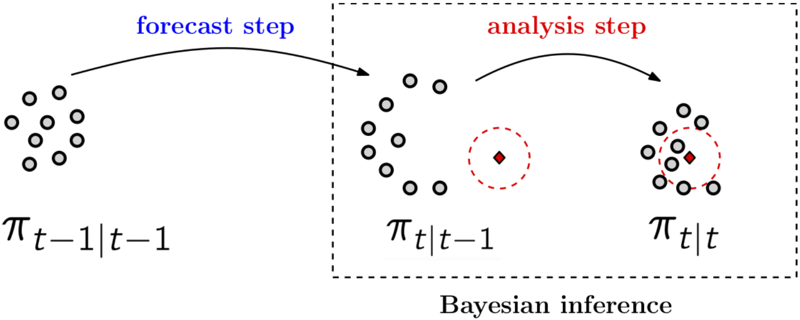
- Complex (e.g., chaotic) dynamics with intractable kernels
- High-dimensional states, $\mathbf{x}_t \in \mathbb{R}^d$ for $d \sim \mathcal{O}(10^6)$
- Sparse observations in space and time
- Limited model evaluations available (e.g., small ensemble sizes)

Ensemble filtering



State-of-the-art (tracking) results are typically found with the EnKF

Ensemble filtering



State-of-the-art (tracking) results are typically found with the EnKF

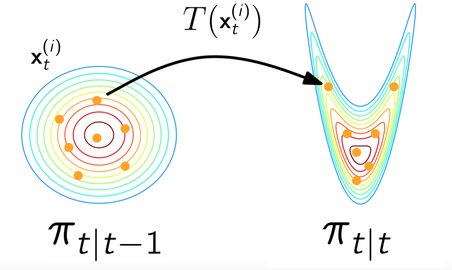
Drawbacks with the EnKF

- Particles are constrained to a linear prior-to-posterior update
- Inconsistent for capturing Bayesian solution
- Modern implementations require extensive tuning for stability

Stochastic map algorithm

Generalization of EnKF for inference step

Find a nonlinear map T that couples forecast $\pi_{t|t-1}$ and analysis $\pi_{t|t}$



Main Idea: Move samples without weights or resampling

- Learn T given $M \ll d$ forecast samples $\mathbf{x}_t^{(i)} \sim \pi_{t|t-1}$
- Generate analysis samples $T(\mathbf{x}_t^{(i)}) \sim \pi_{t|t}$ for $i = 1, \ldots, M$

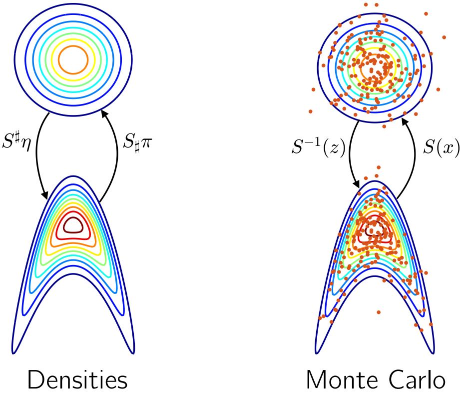
Building block of Stochastic Map

Transport Maps [Parno and Marzouk, 2018]

- Deterministic coupling between densities π, η on $\mathbb{R}^d$ such that

$$\pi(\mathbf{x}) = S^{\sharp}\eta(\mathbf{x}) := \eta \circ S(\mathbf{x})|\det(\nabla S(\mathbf{x}))|$$

- Generate cheap and independent samples $\mathbf{x} \sim \pi \;\Rightarrow\; S(\mathbf{x}) \sim \eta$



Triangular and monotone maps

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. Coupling exists and is unique under mild conditions on π and η
2. For Gaussian η, find S by solving decoupled convex MLE problems

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

- Given samples $\mathbf{x}^{(i)} \sim \pi$, find S^k via

$$\min_{S_k} \frac{1}{M} \sum_{i=1}^{M} \left[\frac{1}{2} S_k(\mathbf{x}^{(i)})^2 - \log |\partial_k S_k(\mathbf{x}^{(i)})| \right] \text{ s.t. } \partial_k S_k > 0$$

Triangular maps enable conditional sampling

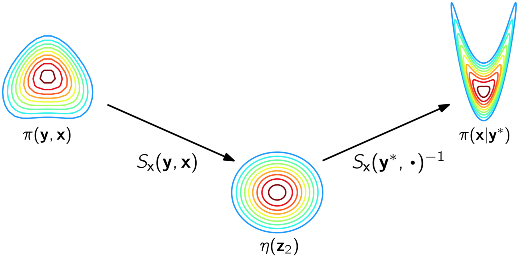
3. Each component S_k characterizes one marginal conditional of π

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

- For $\pi(\mathbf{y},\mathbf{x})$ and $\eta(\mathbf{z}_1,\mathbf{z}_2)$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S_{\mathrm{y}}(\mathbf{y}) \\ S_{\mathrm{x}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- The map $\mathbf{x} \mapsto S_{\mathrm{x}}(\mathbf{y}^*,\mathbf{x})$ pushes $\pi(\mathbf{x}|\mathbf{y}^*)$ to $\eta(\mathbf{z}_2)$
- $S_{\mathrm{x}}(\mathbf{y},\mathbf{x})$ pushes $\pi(\mathbf{x},\mathbf{y})$ to $\eta(\mathbf{z}_2)$



Triangular maps enable conditional sampling

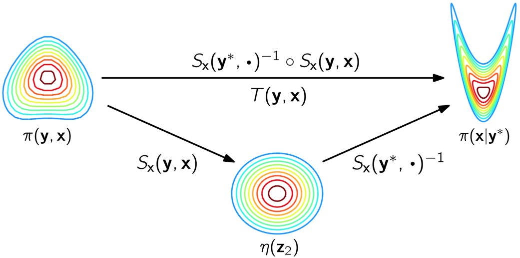
3. Each component S_k characterizes one marginal conditional of π

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

- For $\pi(\mathbf{y},\mathbf{x})$ and $\eta(\mathbf{z}_1,\mathbf{z}_2)$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S_{\mathrm{y}}(\mathbf{y}) \\ S_{\mathrm{x}}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- The map $\mathbf{x} \mapsto S_{\mathrm{x}}(\mathbf{y}^*,\mathbf{x})$ pushes $\pi(\mathbf{x}|\mathbf{y}^*)$ to $\eta(\mathbf{z}_2)$
- $S_{\mathrm{x}}(\mathbf{y},\mathbf{x})$ pushes $\pi(\mathbf{x},\mathbf{y})$ to $\eta(\mathbf{z}_2)$



Triangular maps enable conditional sampling

3. Each component S_k characterizes one marginal conditional of π

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

- For $\pi(\mathbf{y},\mathbf{x})$ and $\eta(\mathbf{z}_1,\mathbf{z}_2)$, consider the triangular map

$$S(\mathbf{y},\mathbf{x}) = \begin{bmatrix} S_\mathrm{y}(\mathbf{y}) \\ S_\mathrm{x}(\mathbf{y},\mathbf{x}) \end{bmatrix}$$

- The map $\mathbf{x} \mapsto S_\mathrm{x}(\mathbf{y}^*,\mathbf{x})$ pushes $\pi(\mathbf{x}|\mathbf{y}^*)$ to $\eta(\mathbf{z}_2)$
- $S_\mathrm{x}(\mathbf{y},\mathbf{x})$ pushes $\pi(\mathbf{x},\mathbf{y})$ to $\eta(\mathbf{z}_2)$

The analysis map that pushes $\pi(\mathbf{y},\mathbf{x})$ to $\pi(\mathbf{x}|\mathbf{y}^*)$ is given by

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

Stochastic Map algorithm

Forecast step

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

Analysis step

1. *Perturbed observations*: Sample $\mathbf{y}_t^{(i)} \sim g(\cdot|\mathbf{x}_t^{(i)})$ using forecast
2. Estimate lower-triangular map $\widehat{S}$ that couples $\pi(\mathbf{y}_t, \mathbf{x}_t)$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$

$$\widehat{S}(\mathbf{y}, \mathbf{x}) = \begin{bmatrix} \widehat{S}_{\mathsf{y}}(\mathbf{y}) \\ \widehat{S}_{\mathsf{x}}(\mathbf{y}, \mathbf{x}) \end{bmatrix}$$

3. Compose maps $\widehat{T}(\mathbf{y}, \mathbf{x}) = \widehat{S}_{\mathsf{x}}(\mathbf{y}^*, \cdot)^{-1} \circ \widehat{S}_{\mathsf{x}}(\mathbf{y}, \mathbf{x})$
4. Generate analysis ensemble $(\mathbf{x}_t^a)^{(i)} = \widehat{T}(\mathbf{y}_t^{(i)}, \mathbf{x}_t^{(i)})$ for $i = 1, \ldots, M$

Numerical details of the Stochastic Map algorithm

Connection with the EnKF

- When restricting $S_{\mathbf{x}}$ to be affine, the map is the EnKF transformation

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

- Transport maps allow for the gradual introduction of nonlinear terms
- Nonlinearities in T capture non-Gaussian structure of $\pi(\mathbf{y}_t, \mathbf{x}_t)$

Example map parameterization

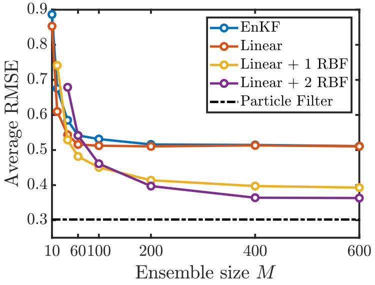
- Each component is the sum of nonlinear univariate functions

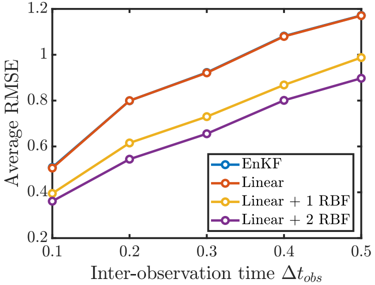
$$S_k(z_1, \ldots, z_k) = \mathbf{u}_1(z_1) + \cdots + \mathbf{u}_k(z_k),$$

  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
- Could also use polynomial expansions (more later...)

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 $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I})$
- Compare statistics to a particle filter (PF) with 1M samples

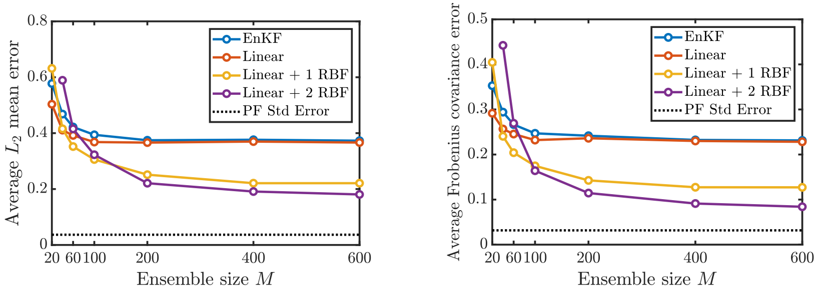




Takeaway: Nonlinearities improve tracking and are stable with Δt_{obs}

Lorenz-63 model

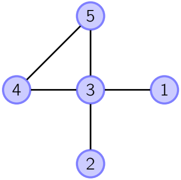
- $d = 3$ with $\Delta t_{obs} = 0.1$ and fully-observed state
- Observations follow $\mathbf{y}_t = \mathbf{x}_t + \boldsymbol{\eta}_t$ with $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I})$
- Compare statistics to a particle filter (PF) with 1M samples



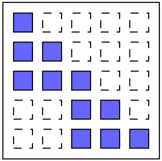
Takeaway: Nonlinearities improve posterior mean and variance estimates

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

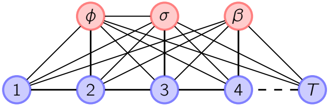
Conditional independence of π defines functional dependence of $S_k(\mathbf{x})$



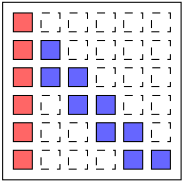
Markov structure of 5-dimensional distribution



Sparsity of $\partial_j S_k$



Markov structure of stochastic volatility problem



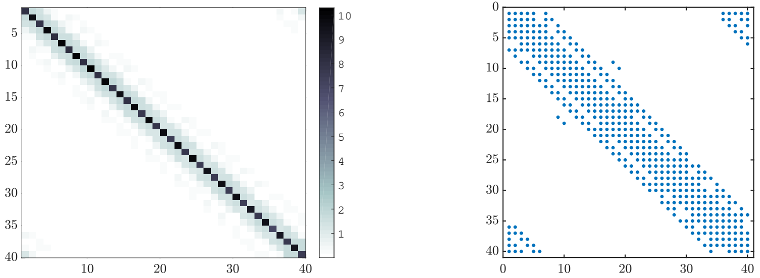
Sparsity of $\partial_j S_k$

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

Conditional independence of π defines functional dependence of $S_k(\mathbf{x})$

Lorenz-96 model

- Estimate forecast covariance $C_{t|t-1}$ over 1000 assimilation cycles



Average $C_{t|t-1}^{-1}$ — Sparsity of $C_{t|t-1}^{-1}$

In practice, distributions in filtering have ≈conditional independence

The map is easy to "localize" in high dimensions

- Regularize the estimation of S by *imposing sparsity* in $\widehat{S}$:

$$\widehat{S}(x_1, \ldots, x_4) = \begin{bmatrix} \widehat{S}_1(x_1) \\ \widehat{S}_2(x_1, x_2) \\ \widehat{S}_3(\quad , x_2, x_3) \\ \widehat{S}_4(\qquad , x_3, x_4) \end{bmatrix}$$

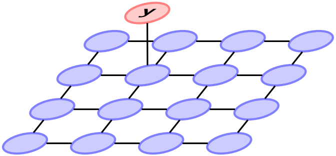
- **Heuristic**: Let $\widehat{S}_k$ depend on neighborhood variables $(x_j)_{j<k}$ that are within a distance r from x_k in state-space:

$$\widehat{S}_k(x_1, \ldots, x_k) \approx \widehat{S}_k(x_{N_r(k)}, x_k)$$

Approach: Parametrize sparsity with neighborhood size and tune parameters by minimizing RMSE over many assimilation cycles

Analysis map has another form of sparsity

- For local likelihood models T decays based on correlation length
- S^x also inherits decay and only needs to be partly estimated



$$S(\mathbf{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_l(x_1, \dots, x_l) \\ x_{l+1} \\ \vdots \\ x_d \end{bmatrix}$$

Approximate S_k as $S_k(x_{N_r(k)}, x_k)$.

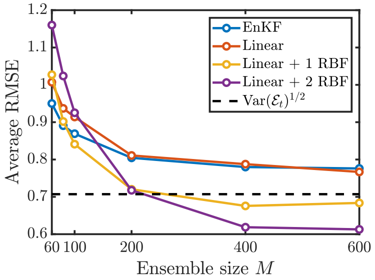
Reverts to identity from decay of correlation

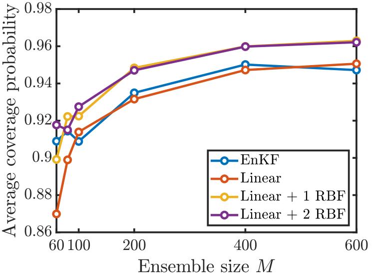
Approach: Parametrize sparsity with neighborhood size and # of non-Identity components and tune parameters by minimizing RMSE

Performance of Stochastic Maps

Lorenz-96 model

- $d = 40$ with $F = 8$, $\Delta t_{obs} = 0.4$ (**large!**) and 20 observations
- Measure RMSE (*left*) and the coverage probability of the empirical [2.5%,97.5%] quantiles (*right*) over 2000 assimilation cycles





Takeaway: Nonlinearities improve tracking given sufficient samples to reliably learn parameters

Linear Transport Maps

- Linear components: $S(\mathbf{x}) = \mathbf{L}\mathbf{x}$, with lower-triangular $\mathbf{L}$
- Approximating density: $\pi = S^\sharp \eta = \mathcal{N}(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}^{-1} = \mathbf{L}\mathbf{L}^T$

Connection to Linear Regression

- Normalize diagonal: $S_k(x) = L_{kk}(\beta_1 x_1 + \cdots + \beta_{k-1}x_{k-1} + x_k)$
- Rewrite MLE optimization problem for linear map parameters:

$$\min_{S_k} \; \mathbb{E}_\pi \left[\tfrac{1}{2} S_k(\mathbf{x})^2 - \log |\partial_k S_k(\mathbf{x})| \right]$$

- Using samples from π:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2M} \|\mathbf{x}_{1:k-1}\boldsymbol{\beta} + \mathbf{x}_k\|_2^2, \quad \widehat{L}_{kk} = \left(\tfrac{1}{M} \|\mathbf{x}_{1:k-1}\hat{\boldsymbol{\beta}} + \mathbf{x}_k\|_2^2 \right)^{-1/2}$$

Linear Transport Maps

- Linear components: $S(\mathbf{x}) = \mathbf{L}\mathbf{x}$, with lower-triangular $\mathbf{L}$
- Approximating density: $\pi = S^\sharp \eta = \mathcal{N}(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}^{-1} = \mathbf{L}\mathbf{L}^T$

Connection to Linear Regression

- Normalize diagonal: $S_k(x) = L_{kk}(\beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + x_k)$
- Rewrite MLE optimization problem for linear map parameters:

$$\min_{L_{kk}>0, \boldsymbol{\beta}} \mathbb{E}_\pi \left[\tfrac{1}{2} L_{kk}^2 (x_{1:k-1} \boldsymbol{\beta} + x_k)^2 - \log |L_{kk}| \right]$$

- Using samples from π:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2M} \|\mathbf{x}_{1:k-1} \boldsymbol{\beta} + \mathbf{x}_k\|_2^2, \quad \widehat{L}_{kk} = \left(\tfrac{1}{M} \|\mathbf{x}_{1:k-1} \hat{\boldsymbol{\beta}} + \mathbf{x}_k\|_2^2 \right)^{-1/2}$$

Linear Transport Maps

- Linear components: $S(\mathbf{x}) = \mathbf{L}\mathbf{x}$, with lower-triangular $\mathbf{L}$
- Approximating density: $\pi = S^{\sharp}\eta = \mathcal{N}(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}^{-1} = \mathbf{L}\mathbf{L}^T$

Connection to Linear Regression

- Normalize diagonal: $S_k(x) = L_{kk}(\beta_1 x_1 + \cdots + \beta_{k-1}x_{k-1} + x_k)$
- Rewrite MLE optimization problem for linear map parameters:

$$\min_{L_{kk}>0,\boldsymbol{\beta}} \mathbb{E}_\pi\left[\tfrac{1}{2}L_{kk}^2(x_{1:k-1}\boldsymbol{\beta} + x_k)^2 - \log|L_{kk}|\right]$$

- Using samples from π:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2M}\|\mathbf{x}_{1:k-1}\boldsymbol{\beta} + \mathbf{x}_k\|_2^2, \quad \widehat{L}_{kk} = \left(\tfrac{1}{M}\|\mathbf{x}_{1:k-1}\hat{\boldsymbol{\beta}} + \mathbf{x}_k\|_2^2\right)^{-1/2}$$

Proposed Approach: Add ℓ_1-penalty for sparse regression (LASSO):

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2M}\|\mathbf{x}_{1:k-1}\boldsymbol{\beta} + \mathbf{x}_k\|_2^2 + \lambda_n\|\boldsymbol{\beta}\|_1$$

Maps generalize to non-Gaussian densities

- E.g., Parametrize monotone nonlinear maps using:

$$S_k(\mathbf{x}_{1:k}) = \sum_j \beta_j \psi_j(\mathbf{x}_{1:k-1}) + \int_0^{x_k} h_{\boldsymbol{\alpha}}(\mathbf{x}_{1:k-1}, t)dt$$

- $h_\alpha > 0$ for strict monotonicity with respect to x_k
- Add ℓ_1-penalty to learn sparsity of $\boldsymbol{\beta}, \boldsymbol{\alpha}$ parameters

Parameterizations cases

1. Gaussian conditionals with constant variance: $h_{\boldsymbol{\alpha}} = \alpha_k$
   - $S_k(\mathbf{x}_{1:k}) = \sum_j \beta_j \psi_j + \alpha_k x_k$
2. Gaussian conditionals with variance depending on $\mathbf{x}_{1:k-1}$
   - $S_k(\mathbf{x}_{1:k}) = \sum_j \beta_j \psi_j + h_{\boldsymbol{\alpha}}(\mathbf{x}_{1:k-1}) x_k$
3. Fully general monotone case
   - $S_k(\mathbf{x}_{1:k}) = \sum_j \beta_j \psi_j + \int_0^{x_k} (\sum_j \alpha_j \phi_j(\mathbf{x}_{1:k-1}, t))^2 dt$

Theoretical performance

Assumptions: Gaussian conditionals with $h_{\boldsymbol{\alpha}} = \alpha_k$ and sub-Gaussian π

Result: Out-of-sample performance

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

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

Takeaways

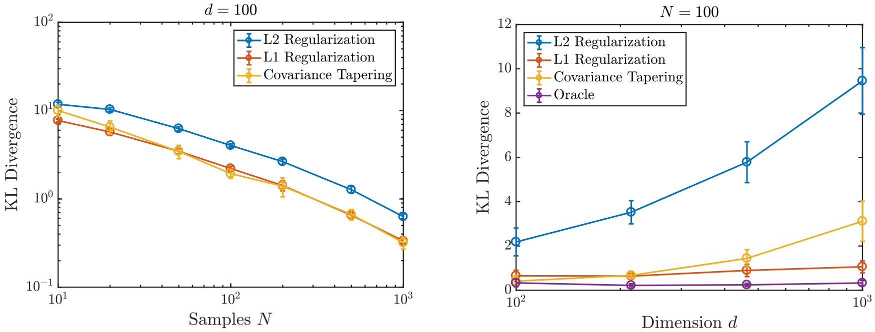
- Accurate estimation is feasible in high-dimensions with $N \ll k$
- From factorization property of density, error in conditionals ensures

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

- ℓ_2 regularization requires $N = \mathcal{O}(k)$ samples for each component

Linear Gaussian problem

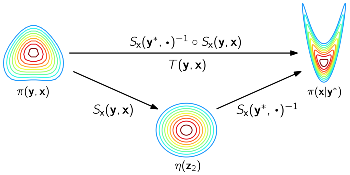
- *Prior*: $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma_{pr})$ with exponential covariance
- *Likelihood*: Local observations $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Gamma)$

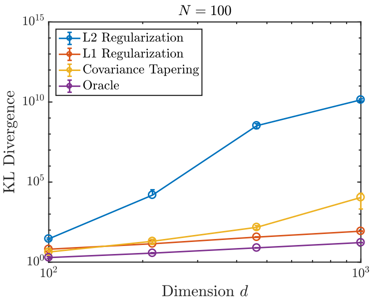


Takeaway

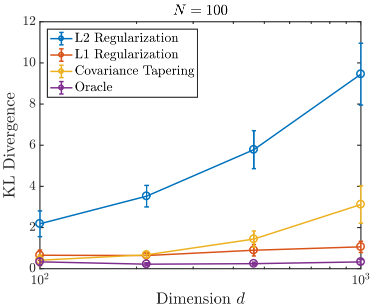
- Learning sparse prior-to-posterior map T matches oracle scaling

Two approaches for posterior sampling





$\mathbf{x}|\mathbf{y}^* \sim (\widehat{S}_x)^\sharp \eta$



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

Takeaway

- Propagating forecast through composed maps has lower error

Conclusion and Outlook

Summary

- Composed **couplings** to build nonlinear prior-to-posterior maps
- Demonstrated improved tracking and posterior moment statistics
- Regularized map estimation to learn sparse **high-dimensional maps**

Outlook on Future Work

- Explore *optimal estimators* for choosing nonlinearity given M samples
- Learn combination of **sparse and low-rank** structure in T

Preprint will be available soon!

Conclusion and Outlook

Summary

- Composed **couplings** to build nonlinear prior-to-posterior maps
- Demonstrated improved tracking and posterior moment statistics
- Regularized map estimation to learn sparse **high-dimensional maps**

Outlook on Future Work

- Explore *optimal estimators* for choosing nonlinearity given M samples
- Learn combination of **sparse and low-rank** structure in T

Preprint will be available soon!

Thank You

Supported by the Air Force Office of Scientific Research

References I

Parno, M. D. and Marzouk, Y. M. (2018).
Transport map accelerated markov chain monte carlo.
SIAM/ASA Journal on Uncertainty Quantification, 6(2):645–682.

Spantini, A., Baptista, R., and Marzouk, Y. (2019).
Coupling techniques in nonlinear ensemble filtering: non-Gaussian generalizations of the EnKF.
preprint.

Spantini, A., Bigoni, D., and Marzouk, Y. (2018).
Inference via low-dimensional couplings.
The Journal of Machine Learning Research, 19(1):2639–2709.