Density estimation and conditional simulation using triangular transport

Youssef Marzouk[1]
joint work with Ricardo Baptista,[1] Olivier Zahm,[2] and Jakob Zech[3]

[1]Massachusetts Institute of Technology
http://uqgroup.mit.edu

[2]INRIA and Université Grenoble Alpes

[3]Universität Heidelberg

Support from AFOSR, DOE, NSF, ONR

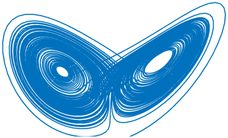
26 July 2021

Setting: Generative model with intractable prior and likelihood

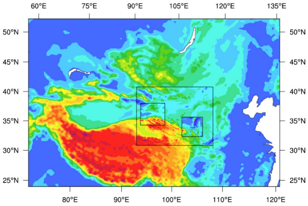
- Parameters $\mathbf{x} \sim \pi_{\mathbf{X}}$
- Data $\mathbf{y} \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\mathbf{x})$
- We can easily simulate $(\mathbf{x}^i, \mathbf{y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$

Goal: Sample from the posterior $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$ for any $\mathbf{y}^*$

Applications: Geophysical data assimilation (ensemble filtering), parameter inference in stochastic models



Lorenz-63 system



Numerical weather prediction

Motivation: likelihood-free Bayesian inference

Setting: Generative model with intractable prior and likelihood

- Parameters $\mathbf{x} \sim \pi_{\mathbf{X}}$
- Data $\mathbf{y} \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\mathbf{x})$
- We can easily simulate $(\mathbf{x}^i, \mathbf{y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$

Goal: Estimate mutual information $I(X;Y)$

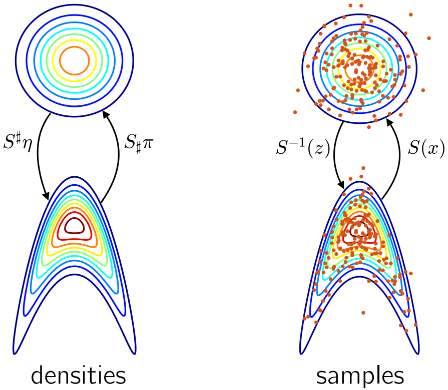
Application: Bayesian optimal experimental design

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}_{\boldsymbol{Y}}\left[\mathrm{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}} \,||\, \pi_{\mathbf{X}})\right] \\
&= \mathbb{E}_{\boldsymbol{Y},\boldsymbol{X}}[\log \pi(\mathbf{x}|\boldsymbol{y}) - \log \pi(\mathbf{x})] = \mathbb{E}_{\boldsymbol{Y},\boldsymbol{X}}[\log \pi(\boldsymbol{y}|\mathbf{x}) - \log \pi(\boldsymbol{y})]
\end{aligned}
$$

$\Rightarrow$ Need to estimate conditional *and* marginal densities over a range of values of $\mathbf{X}$ and $\mathbf{Y}$

Link these goals to transport

- A **transport map** S induces a *deterministic coupling* between a target distribution π and a reference distribution η
  - Generate cheap and independent samples: $\mathbf{z} \sim \eta \;\Leftrightarrow\; S^{-1}(\mathbf{z}) \sim \pi$
  - Estimate the target density: $\pi(\mathbf{x}) = S^{\sharp}\eta(\mathbf{x}) := \eta \circ S(\mathbf{x})\, |\det \nabla S(\mathbf{x})|$



Monotone triangular transport maps

Specifically, consider the **Knothe–Rosenblatt (KR) rearrangement**

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

1. Monotone ($\partial_k S^k > 0$) triangular map S satisfying $S_\sharp \pi = \eta$; exists and is unique under mild conditions on π and η
2. Easily invertible, with $\det \nabla S(\mathbf{x})$ is tractable
3. Each component S^k characterizes one marginal conditional of π

$$\pi_{\mathbf{X}} = \pi_{X_1} \pi_{X_2|X_1} \cdots \pi_{X_d|X_1,\ldots,X_{d-1}}$$

4. The KR map is a *limit* of optimal transport maps obtained under anisotropic quadratic cost, e.g., $c_t(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^d t^{i-1}(x_i - z_i)^2$ as $t \to 0$ [Carlier et al. 2009]

Conditional density estimation and simulation

- Given joint prior model $\pi_{\mathbf{Y},\mathbf{X}}$ for parameters $\mathbf{X} \in \mathbb{R}^n$, data $\mathbf{Y} \in \mathbb{R}^m$: seek the KR map S that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2} := \mathcal{N}(0, \mathbf{I}_{m+n})$
- The KR map immediately has a block structure

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

which suggests two properties of the lower block:

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

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

Conditional density estimation and simulation

- Given joint prior model $\pi_{\mathbf{Y},\mathbf{X}}$ for parameters $\mathbf{X} \in \mathbb{R}^n$, data $\mathbf{Y} \in \mathbb{R}^m$: seek the KR map S that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2} := \mathcal{N}(0, \mathbf{I}_{m+n})$
- The KR map immediately has a block structure

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

which suggests two properties of the lower block:

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

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

1. Approximate the conditional **density**:

$$\boxed{\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*} = S^{\mathcal{X}}(\mathbf{y}^*, \cdot)^{\sharp}\mathcal{N}(0, \mathbf{I}_n)}$$

Conditional density estimation and simulation

- Given joint prior model $\pi_{\mathbf{Y},\mathbf{X}}$ for parameters $\mathbf{X} \in \mathbb{R}^n$, data $\mathbf{Y} \in \mathbb{R}^m$: seek the KR map S that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2} := \mathcal{N}(0, \mathbf{I}_{m+n})$
- The KR map immediately has a block structure

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

which suggests two properties of the lower block:

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

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

2. **Sample** the conditional distribution $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$ with a *single* map:

$$\boxed{\text{Solve } S^{\mathcal{X}}(\mathbf{y}^*, \mathbf{x}^i) = \boldsymbol{\xi}^i \text{ for } \mathbf{x}^i \text{ given } \boldsymbol{\xi}^i \sim \mathcal{N}(0, \mathbf{I}_n)}$$

Conditional density estimation and simulation

- Given joint prior model $\pi_{\mathbf{Y},\mathbf{X}}$ for parameters $\mathbf{X} \in \mathbb{R}^n$, data $\mathbf{Y} \in \mathbb{R}^m$: seek the KR map S that pushes $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_1,\mathbf{Z}_2} := \mathcal{N}(0, \mathbf{I}_{m+n})$
- The KR map immediately has a block structure

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

which suggests two properties of the lower block:

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

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

3. **Sample** the conditional via a *composed map* T that pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$:

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

A general recipe

- **Estimate** the triangular map S (e.g., in some *parameterized family*) from $(\mathbf{y}^i, \mathbf{x}^i)_{i=1}^n \sim \pi_{\mathbf{Y},\mathbf{X}}$
- Use relevant parts of the estimated map to *generate* conditional samples or to *approximate* relevant conditional (or marginal) densities

A general recipe

- **Estimate** the triangular map S (e.g., in some *parameterized family*) from $(\mathbf{y}^i, \mathbf{x}^i)_{i=1}^n \sim \pi_{\mathbf{Y},\mathbf{X}}$
- Use relevant parts of the estimated map to *generate* conditional samples or to *approximate* relevant conditional (or marginal) densities

Many applications of this approach:

- Likelihood-free/simulation-based inference
- Optimal experimental design and MI estimation
- Nonlinear filtering (ensemble Kalman filter $\Leftrightarrow$ linear $S(\mathbf{y}, \mathbf{x})$; see generalizations in [Spantini et al. arXiv:1907.00389])
- Triangular maps are the building block of autoregressive *normalizing flows* in machine learning...

Plan for the rest of talk

Some underlying methodological questions:

1. How to **approximate** triangular transport maps?
2. Properties of the **optimization** problem arising in transport map estimation
3. The unreasonable effectiveness of **composed maps** for conditional simulation

Topic #1: approximation of triangular maps

- Consider triangular maps on bounded domains (e.g., $[0, 1]^d$)
- **Main results:**
  - If both the reference and target densities f_η, f_π are **analytic**, the **Knothe–Rosenblatt map T is analytic**
  - T can be approximated with rational functions or deep ReLU networks, via constructions that guarantee *monotonicity* and *bijectivity*
  - Explicit *a priori* descriptions of ansatz spaces
  - Exponential convergence rates

Regularity of T

$\{z \in \mathbb{C} : \mathrm{dist}(z, [-1,1]) < r\}$

$\mathcal{B}_r :=$  $\subseteq \quad \mathbb{C}$

Theorem (informal, [ZM20])

Let f_η, $f_\pi : \times_{j=1}^{d} \mathcal{B}_{r_j} \to \mathbb{C}$ be analytic and bounded for $(r_j)_{j=1}^{d}$ monotonically increasing. Then

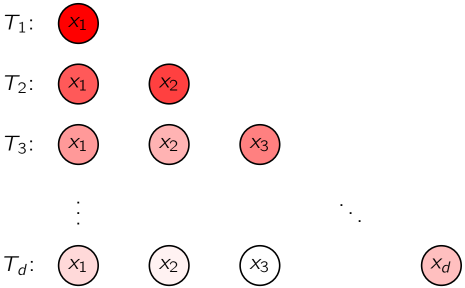
- *$T_k : \times_{j=1}^{k} \mathcal{B}_{Cr_j} \to \mathbb{C}$ is analytic for some $C > 0$,*
- *if $r_k \gg 1$ then $T_k(\mathbf{x}) \sim x_k$.*

[ZM20] J. Zech and Y. Marzouk, arXiv:2006.06994, 2020.

Where/how should we invest degrees of freedom to approximate T?

$$
\begin{array}{lccccc}
T_1: & x_1 & & & & \\
T_2: & x_1 & x_2 & & & \\
T_3: & x_1 & x_2 & x_3 & & \\
 & \vdots & & & \ddots & \\
T_d: & x_1 & x_2 & x_3 & & x_d
\end{array}
$$

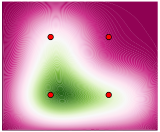
Where/how should we invest degrees of freedom to approximate T?



$$\mathbb{P}_{\Lambda_{\varepsilon,k}} := \operatorname{span}\left\{ \prod_{j=1}^{k} x_j^{\nu_j} \ : \ \boldsymbol{\nu} \in \Lambda_{\varepsilon,k} \right\},$$

$$\Lambda_{\varepsilon,k} := \left\{ \boldsymbol{\nu} \in \mathbb{N}_0^k \ : \ (1 + r_k)^{-\max\{1,\nu_k\}} \prod_{j=1}^{k-1} (1 + r_j)^{-\nu_j} > \varepsilon \right\}$$

Convergence rates in finite dimension



Example: **PDE inverse problem**

$-\mathrm{div}(a\nabla u) = f$

$a(s) = 1 + \sum_{j=1}^{d} x_j \, \psi_j(s)$

Reference and target on $[-1,1]^d$:

- $\eta = \otimes_{j=1}^{d} \frac{\lambda}{2}$
- π = posterior, i.e., $\pi_{\mathbf{x}\,|\{u(s_i)\}}$

Theorem (informal, [ZM20])

There exist (a priori) ansatz spaces A_ε employing $N_\varepsilon = \sum_{k=1}^{d} |\Lambda_{\varepsilon,k}| \in \mathbb{N}$ degrees of freedom and $\widetilde{T} \in A_\varepsilon$ s.t.

- *A_ε of* **rational fcts***:* $\mathrm{dist}(\widetilde{T}_\sharp \eta, \pi) \lesssim \exp(-\beta N_\varepsilon^{\frac{1}{d}})$
- *A_ε of* **ReLU NNs***:* $\mathrm{dist}(\widetilde{T}_\sharp \eta, \pi) \lesssim \exp(-\beta N_\varepsilon^{\frac{1}{d+1}})$

with $\mathrm{dist} \in \{$*Hellinger, TV, KL,* $W_p\}$.

[ZM20] J. Zech and Y. Marzouk, arXiv:2006.06994, 2020.

Significance:

- Many recent ML approaches employ triangular maps (neural autoregressive flows, sum-of-squares polynomial flow, neural spline flow, etc.)
- Few results on universality; fewer still on convergence rates!
- Additionally: *dimension-independent* higher-order convergence rates for certain inference problems in PDEs (see [ZM20])

Next steps: less smoothness, unbounded domains

Topic #2: estimating monotone triangular maps

Many *special* cases of triangular maps are in practical use:

- Example: masked autoregressive flow [Papamakarios et al. 2017]

$$S^k(x_1, \dots, x_k) = \mu_k(\mathbf{x}_{i<k}) + x_k \exp(\alpha_k(\mathbf{x}_{i<k}))$$

- **Numerous** others [Jaini et al. 2019, Wehenkel & Louppe 2019, etc.]
- *Compose* these transformations, interleaved with *permutations*:
  - Universal approximators [Teshima et al. 2020] but *no longer triangular*

Topic #2: estimating monotone triangular maps

Many *special* cases of triangular maps are in practical use:

- Example: masked autoregressive flow [Papamakarios et al. 2017]

$$S^k(x_1, \ldots, x_k) = \mu_k(\mathbf{x}_{i<k}) + x_k \exp(\alpha_k(\mathbf{x}_{i<k}))$$

- **Numerous** others [Jaini et al. 2019, Wehenkel & Louppe 2019, etc.]
- *Compose* these transformations, interleaved with *permutations*:
  - Universal approximators [Teshima et al. 2020] but *no longer triangular*
- In general, maximum likelihood estimation in these models is a **challenging optimization problem**:

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

Topic #2: estimating monotone triangular maps

Goal: seek a *general* representation of monotone triangular functions that is "easy" to estimate...

Existing methods for enforcing monotonicity:

- Enforce $\partial_k S^k(\mathbf{x}^i_{1:k}) > 0$ at finite training samples $i = 1, \ldots, n$
- Or enforce by construction: e.g., SOS polynomial flows [Jaini et al. 2019]

$$S^k(\mathbf{x}_{1:k}) = a_k(\mathbf{x}_{<k}) + \int_0^{x_k} b_k(\mathbf{x}_{<k}, t)^2 dt$$

Improved idea: Represent S^k via an **invertible** "rectifier"

$$S^k(\mathbf{x}_{1:k}) = \mathcal{R}_k(f)(\mathbf{x}_{1:k}) \coloneqq f(\mathbf{x}_{<k}, 0) + \int_0^{x_k} g(\partial_k f(\mathbf{x}_{<k}, t))dt,$$

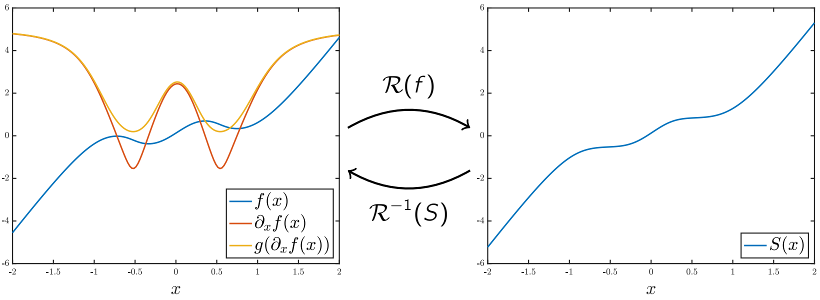
where $g\colon \mathbb{R} \to \mathbb{R}_{>0}$ is bijective & smooth and $f\colon \mathbb{R}^k \to \mathbb{R}$ is unconstrained

Parameterizing monotone maps

Rectification of f (1-D example)

For smooth f and bijective $g \colon \mathbb{R} \to \mathbb{R}_{>0}$ (e.g., $g(x) = \log(1 + e^x)$)

$$S(x) = \mathcal{R}(f)(x) := f(0) + \int_0^x g(\partial_x f(t))dt,$$



Approximating monotone maps

Convert constrained minimization to an unconstrained problem:

$$\min_{\{S:\partial_k S>0\}} \underbrace{\mathbb{E}_\pi\left[\frac{1}{2}S(\mathbf{x}_{1:k})^2 - \log|\partial_k S(\mathbf{x}_{1:k})|\right]}_{\mathcal{J}_k(S),\ \text{convex in } S} \quad \Leftrightarrow \quad \min_f \underbrace{\mathcal{J}_k \circ \mathcal{R}_k(f)}_{\mathcal{L}_k(f)}$$

- With this reparameterization, we **lose** convexity!
- When will the objective still have “nice” properties?

Approximating monotone maps

Convert constrained minimization to an unconstrained problem:

$$\min_{\{S:\partial_k S>0\}} \underbrace{\mathbb{E}_\pi\left[\frac{1}{2}S(\mathbf{x}_{1:k})^2 - \log|\partial_k S(\mathbf{x}_{1:k})|\right]}_{\mathcal{J}_k(S),\ \text{convex in } S} \quad \Leftrightarrow \quad \min_f \underbrace{\mathcal{J}_k \circ \mathcal{R}_k(f)}_{\mathcal{L}_k(f)}$$

- With this reparameterization, we **lose** convexity!
- When will the objective still have "nice" properties?

One example: consider the space of functions

$$H^{1,k}_\pi(\mathbb{R}^k) := \left\{f : \mathbb{R}^k \to \mathbb{R} \text{ s.t. } \int |f(\mathbf{x})|^2 + |\partial_k f(\mathbf{x})|^2 d\pi(\mathbf{x}) < \infty\right\}$$

Some current results [BZM20]:

Let $\pi(\mathbf{x}) \leq C\eta(\alpha\mathbf{x})$ for some $C < \infty$, $\alpha > 0$, and η standard Gaussian. Then, for smooth, bijective, and positive g, $\mathcal{L}_k : H^{1,k}_\pi \to \mathbb{R}$ is continuous and bounded.

Approximating monotone maps

Convert constrained minimization to an unconstrained problem:

$$\min_{\{S:\partial_k S>0\}} \underbrace{\mathbb{E}_\pi\left[\frac{1}{2}S(\mathbf{x}_{1:k})^2 - \log|\partial_k S(\mathbf{x}_{1:k})|\right]}_{\mathcal{J}_k(S),\ \text{convex in } S} \quad \Leftrightarrow \quad \min_f \underbrace{\mathcal{J}_k \circ \mathcal{R}_k(f)}_{\mathcal{L}_k(f)}$$

- With this reparameterization, we **lose** convexity!
- When will the objective still have "nice" properties?

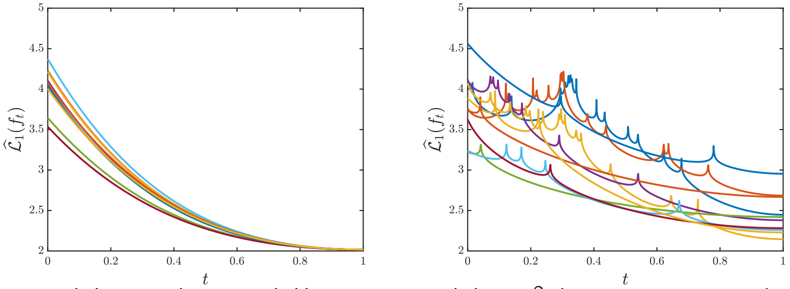
Consider the space of functions $\tilde{H}^{1,k}_\pi(\mathbb{R}^k) :=$
$\left\{f : \mathbb{R}^k \to \mathbb{R} \text{ s.t. } \int |f(\mathbf{x})|^2 + |\partial_k f(\mathbf{x})|^2 d\pi(\mathbf{x}) < \infty,\ \partial_k f(\mathbf{x}) \geq M > -\infty\right\}$

A *conjecture*:

Let $\pi(\mathbf{x}) \leq C\eta(\alpha\mathbf{x})$ for some $0 < C, \alpha < \infty$ and η standard Gaussian. Then, for smooth, bijective, and positive g satisfying certain additional assumptions, every local minimum of $\mathcal{L}_k : \tilde{H}^{1,k}_\pi \to \mathbb{R}$ is a global minimum.

Numerical results: approximating monotone maps

- Mixture of Gaussians target density π
- Approximate objective as $\widehat{\mathcal{L}}_k$ using $n = 50$ samples
- Evaluate $\widehat{\mathcal{L}}_k$ along segments connecting random initial maps ($t = 0$) to critical points of gradient-based optimizer ($t = 1$)



$g(x) = \log(1 + \exp(x))$ $\quad$ $g(x) = x^2$ (cf. SOS poly flow)

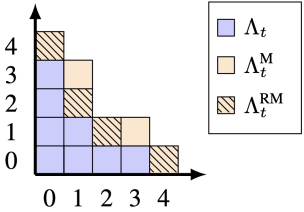
Smooth objective with a single minimizer = **fast and reliable training!**

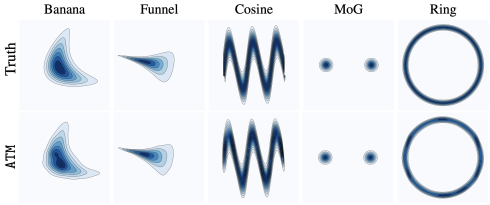
Adaptive transport map (ATM) algorithm

Approach: Use any linear parameterization of $f(\mathbf{x})$ (e.g., Hermite functions, Hermite polynomials, wavelets) + greedy enrichment

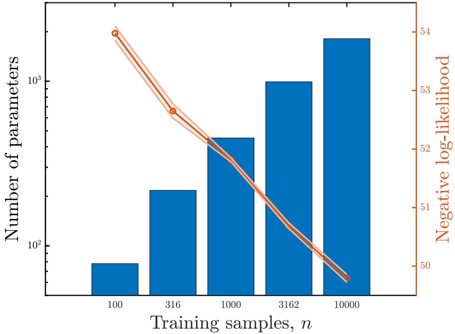
Greedy adaptation

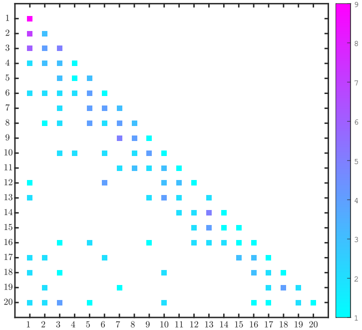
- Look for a **sparse** expansion $f(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\Lambda} c_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\mathbf{x})$
- Add one element at a time to set of active multi-indices Λ_t
- Restrict Λ_t to be downward closed
- Search for new features in the reduced margin of Λ_t
- Stopping the search (via cross-validation) tailors the map representation to the sample size n





Some ATM results





Density estimation for state of chaotic Lorenz-96 system ($d = 20$) with increasing sample size n:

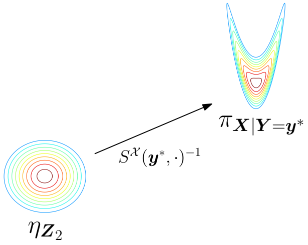
- Greedy approach identifies **sparsity** in triangular map, which reflects *conditional independence* in the target distribution [Spantini et al. 2018]

Another approach to simulating $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$

Recall: target $\pi_{\mathbf{Y},\mathbf{X}}$, reference $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, and the triangular map

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

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



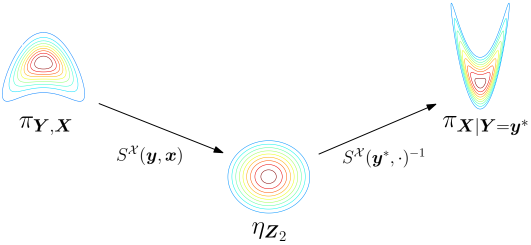
Topic #3: single versus composed maps

Another approach to simulating $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$

Recall: target $\pi_{\mathbf{Y},\mathbf{X}}$, reference $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, and the triangular map

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

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

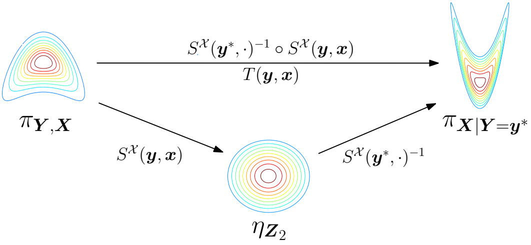


Another approach to simulating $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$

Recall: target $\pi_{\mathbf{Y},\mathbf{X}}$, reference $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, and the triangular map

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

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



Topic #3: single versus composed maps

Another approach to simulating $\pi_{\mathbf{X}|\mathbf{Y}=\mathbf{y}^*}$

Recall: target $\pi_{\mathbf{Y},\mathbf{X}}$, reference $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$, and the triangular map

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

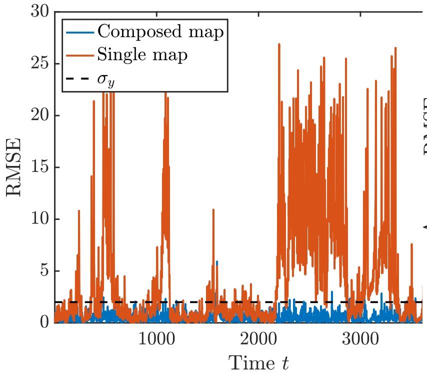
- $S^{\mathcal{X}}(\mathbf{y},\cdot)$ pulls back $\eta_{\mathbf{Z}_2}$ to $\pi_{\mathbf{X}|\mathbf{y}}$ for any $\mathbf{y}$
- $S^{\mathcal{X}}(\mathbf{y},\mathbf{x})$ pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\eta_{\mathbf{Z}_2}$

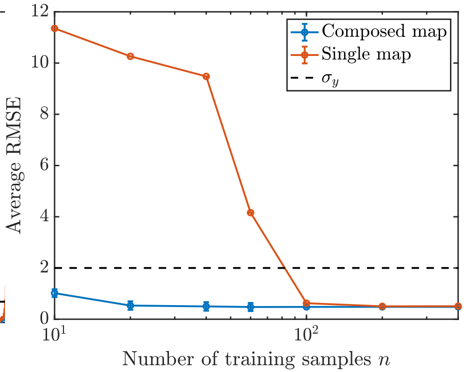
A "composed map" that pushes forward $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\mathbf{y}^*}$ is

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

Nonlinear filtering in the Lorenz-63 model:

- Error in filtering with linear maps: composed maps have smaller RMSE on average



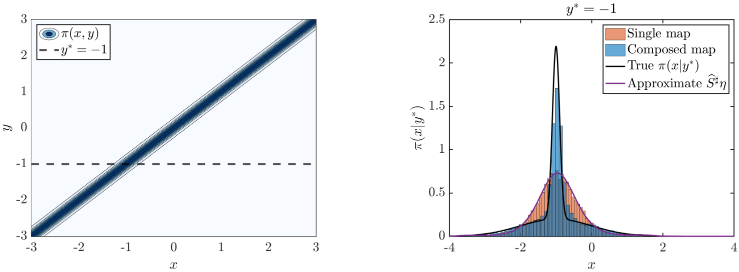


- When S is *linear*, the composed map T is exactly the update/analysis step of the *ensemble Kalman filter* [Spantini, Baptista, M 2019].

Advantages of composed maps

Gaussian mixture [Sisson et al. 2007]

- Prior $\pi_X = \mathcal{U}(-10, 10)$
- Likelihood $\pi_{Y|x} = 0.5\mathcal{N}(x, 1) + 0.5\mathcal{N}(x, 0.01)$
- Approximate $S^{\mathcal{X}}$ using degree 5 polynomials ($\pi_{\mathbf{X}|\mathbf{Y}}$ not in-class)
- Compare samples from composed map T to single map $S^{\mathcal{X}}$

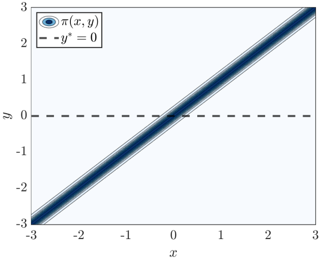


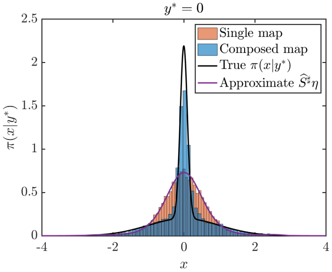
Takeaway: Posterior estimate from composed map has smaller bias

Advantages of composed maps

Gaussian mixture [Sisson et al. 2007]

- Prior $\pi_X = \mathcal{U}(-10, 10)$
- Likelihood $\pi_{Y|x} = 0.5\mathcal{N}(x, 1) + 0.5\mathcal{N}(x, 0.01)$
- Approximate $S^{\mathcal{X}}$ using degree 5 polynomials ($\pi_{\mathbf{X}|\mathbf{Y}}$ not in-class)
- Compare samples from composed map T to single map $S^{\mathcal{X}}$



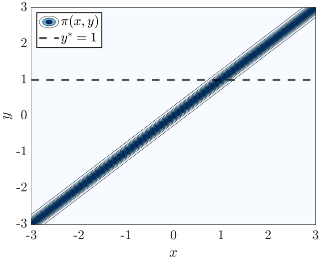


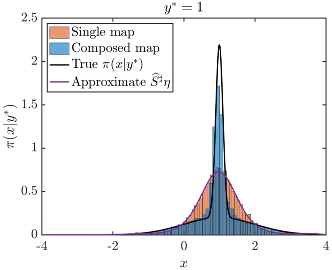
Takeaway: Posterior estimate from composed map has smaller bias

Advantages of composed maps

Gaussian mixture [Sisson et al. 2007]

- Prior $\pi_X = \mathcal{U}(-10, 10)$
- Likelihood $\pi_{Y|x} = 0.5\mathcal{N}(x, 1) + 0.5\mathcal{N}(x, 0.01)$
- Approximate $S^{\mathcal{X}}$ using degree 5 polynomials ($\pi_{\mathbf{X}|\mathbf{Y}}$ not in-class)
- Compare samples from composed map T to single map $S^{\mathcal{X}}$





Takeaway: Posterior estimate from composed map has smaller bias

Analyzing the difference between T and $S^{\mathcal{X}}$

- The distribution $\pi_{\widehat{T},\mathbf{y}^*}$ of the "analysis" random variable $\widehat{T}(\mathbf{Y},\mathbf{X})$ is

$$\pi_{\widehat{T},\mathbf{y}^*} = \widehat{S}^{\mathcal{X}}(\mathbf{y}^*,\cdot)^\sharp\left(\widehat{S}^{\mathcal{X}}_\sharp \pi_{\mathbf{Y},\mathbf{X}}\right)$$

Main idea: T uses information from neighboring *true* conditional densities

Theorem

If the conditionals $\pi_{\mathbf{X}|\mathbf{y}}$ depend continuously on $\mathbf{y}$, then

$$D_{KL}(\pi_{\mathbf{X}|\mathbf{y}^*}||\pi_{\widehat{T},\mathbf{y}^*}) \leq D_{KL}(\pi_{\mathbf{X}|\mathbf{y}^*}||\widehat{S}^{\mathcal{X}}(\mathbf{y}^*,\cdot)^\sharp\eta).$$

The inequality is strict when $\widehat{S}^{\mathcal{X}}$ does not perfectly pull back η to $\pi_{\mathbf{X}|\mathbf{Y}}$

Takeaway: Composed map will yield smaller bias than single map

Analyzing the difference between T and $S^{\mathcal{X}}$

- The distribution $\pi_{\widehat{T},\mathbf{y}^*}$ of the "analysis" random variable $\widehat{T}(\mathbf{Y},\mathbf{X})$ is

$$\pi_{\widehat{T},\mathbf{y}^*} = \widehat{S}^{\mathcal{X}}(\mathbf{y}^*,\cdot)^\sharp \int \widehat{S}^{\mathcal{X}}(\mathbf{y},\cdot)_\sharp\, \pi_{\mathbf{X}|\mathbf{y}}\, \pi_{\mathbf{Y}}(\mathbf{y})\mathrm{d}\mathbf{y}$$

Main idea: T uses information from neighboring *true* conditional densities

Theorem

If the conditionals $\pi_{\mathbf{X}|\mathbf{y}}$ depend continuously on $\mathbf{y}$, then

$$D_{KL}(\pi_{\mathbf{X}|\mathbf{y}^*}||\pi_{\widehat{T},\mathbf{y}^*}) \leq D_{KL}(\pi_{\mathbf{X}|\mathbf{y}^*}||\widehat{S}^{\mathcal{X}}(\mathbf{y}^*,\cdot)^\sharp\eta).$$

The inequality is strict when $\widehat{S}^{\mathcal{X}}$ does not perfectly pull back η to $\pi_{\mathbf{X}|\mathbf{Y}}$

Takeaway: Composed map will yield smaller bias than single map

Exploiting structure in composed maps

Main result [BM21]

For any map $\mathbf{Z} = S^{\mathcal{X}}(\mathbf{Y}, \mathbf{X})$ such that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}$, the map $T(\mathbf{y}, \mathbf{x})$ will sample exactly from the posterior density $\pi_{\mathbf{X}|\mathbf{y}^*}$

Is a **different objective function** then more suitable for finding T?

- Finding $S^{\mathcal{X}}$ such that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}$ *and* $\mathbf{Z} \sim \eta = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is one option
- Can we instead use a reference η that is closer to $\pi_{\mathbf{X}|\mathbf{Y}}$?
- In practice this results in $S^{\mathcal{X}}$ being a simpler map

Exploiting structure in composed maps

Main result [BM21]

For any map $\mathbf{Z} = S^{\mathcal{X}}(\mathbf{Y}, \mathbf{X})$ such that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}$, the map $T(\mathbf{y}, \mathbf{x})$ will sample exactly from the posterior density $\pi_{\mathbf{X}|\mathbf{y}^*}$

Is a **different objective function** then more suitable for finding T?

- Finding $S^{\mathcal{X}}$ such that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}$ *and* $\mathbf{Z} \sim \eta = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is one option
- Can we instead use a reference η that is closer to $\pi_{\mathbf{X}|\mathbf{Y}}$?
- In practice this results in $S^{\mathcal{X}}$ being a simpler map

Approach: Find $S^{\mathcal{X}}$ by minimizing mutual information:

$$
\begin{aligned}
I(\mathbf{Z}, \mathbf{Y}) = \mathbb{E}_{\mathbf{Y}}[D_{KL}(\pi_{\mathbf{Z}|\mathbf{y}} || \pi_{\mathbf{Z}})] &= \mathbb{E}_{\mathbf{Y}}[D_{KL}(\pi_{\mathbf{X}|\mathbf{y}} || S^{\mathcal{X}}(\mathbf{y}, \cdot)^{\sharp} \pi_{\mathbf{Z}})] \\
&= \mathbb{E}_{\mathbf{Y}}[D_{KL}(\pi_{\mathbf{X}|\mathbf{y}} || \pi_{\widehat{T}, \mathbf{y}^*})]
\end{aligned}
$$

Related work: [Tabak et al. 2020] based on optimal transport

Conclusions

- **Central idea:** density estimation and conditional simulation using triangular transport
  - Broad range of applications, including *data assimilation* and other instances of *likelihood-free inference*, as well as *normalizing flows*
  - Approximation results for triangular maps
  - Map estimation: optimization and adaptive parameterizations
  - *Composed map* approach to conditional simulation

Conclusions

- **Central idea:** density estimation and conditional simulation using triangular transport
  - Broad range of applications, including *data assimilation* and other instances of *likelihood-free inference*, as well as *normalizing flows*
  - Approximation results for triangular maps
  - Map estimation: optimization and adaptive parameterizations
  - *Composed map* approach to conditional simulation

- **Additional ongoing work**
  - Approximation of triangular maps in *infinite dimensions* (see [ZM20])
  - *Statistical consistency* of transport map density estimation
  - Low-rank structure in transport maps
  - Block-triangular maps and links to optimal transport (with R. Baptista, B. Hosseini, N. Kovachki)
  - MM approaches to minimizing mutual information

Conclusions

Thanks for your attention!

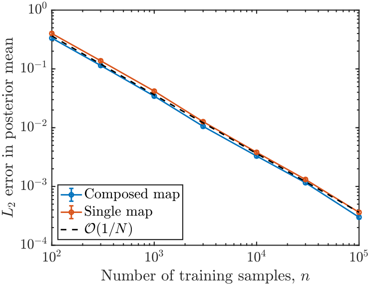
References

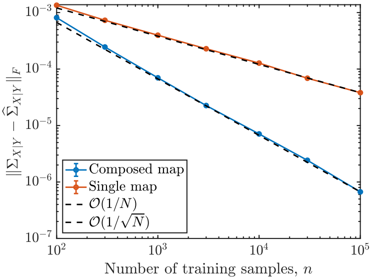
- R. Baptista, Y. Marzouk, R. Morrison, O. Zahm. "Learning non-Gaussian graphical models via Hessian scores and triangular transport." arXiv:2101:03093, 2021.
- J. Zech, Y. Marzouk. "Sparse approximation of triangular transports on bounded domains." arXiv:2006.06994, 2021.
- M. Brennan, D. Bigoni, O. Zahm, A. Spantini, Y. Marzouk. "Greedy inference with structure-exploiting lazy maps." *NeurIPS 2020*, arXiv:1906.00031.
- R. Baptista, O. Zahm, Y. Marzouk. " An adaptive transport framework for joint and conditional density estimation." arXiv:2009.10303, 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.
- O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk. "Certified dimension reduction in nonlinear Bayesian inverse problems." arXiv:1807.03712, 2021.
- A. Spantini, D. Bigoni, Y. Marzouk. "Inference via low-dimensional couplings." *JMLR* 19(66): 1–71, 2018.

Advantages of composed map

Bayesian linear regression model [Papamakarios & Murray 2016]

- Prior $\pi_{\mathbf{X}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for $d = 10$
- Likelihood $\pi_{\mathbf{Y}|\mathbf{X}} = \prod_{i=1}^{m} \mathcal{N}(\mathbf{x}^T \mathbf{u}_i, \sigma^2)$ for $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ and $m = 6$
- Gaussian posterior $\pi_{\mathbf{X}|\mathbf{y}^*}$ available in closed form
- Evaluate convergence of posterior mean and covariance





Takeaway: Posterior covariance estimate from composed map converges more quickly

Under the hood of the linear–Gaussian example

- Map S is allowed to be any affine function; posterior $\pi_{\mathbf{X}|\mathbf{y}^*}$ is *in-class* for both single and composed maps
  - Approximation error entirely due to **variance** of map estimate
- For composed map, approximate posterior covariance is a *squared* perturbation of the *true* posterior covariance:

$$\widehat{\Sigma}^{\text{comp}}_{\mathbf{X}|\mathbf{Y}} = \Sigma_{\mathbf{X}|\mathbf{Y}} + \Delta, \text{ where}$$

$$\Delta = \left(\widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}^{-1}_{\mathbf{YY}}\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{XY}}\Sigma^{-1/2}_{\mathbf{YY}}\right)\left(\widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}^{-1}_{\mathbf{YY}}\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{XY}}\Sigma^{-1/2}_{\mathbf{YY}}\right)^\top$$

- Single map must instead re-capture all terms:

$$\widehat{\Sigma}^{\text{sing}}_{\mathbf{X}|\mathbf{Y}} = \widehat{\Sigma}_{\mathbf{XX}} - \widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}^{-1}_{\mathbf{YY}}\widehat{\Sigma}^\top_{\mathbf{XY}}$$