Measure transport and optimization-based samplers for Bayesian computation

Youssef Marzouk
joint work with Daniele Bigoni, Alessio Spantini, Zheng Wang

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

Support from DOE Office of Advanced Scientific Computing Research

1 September 2016
Bayesian inference in large-scale models

Observations $y$  Parameters $x$

\[ \pi_{\text{pos}}(x) := \pi(x | y) \propto \pi(y | x) \pi_{\text{pr}}(x) \]

Bayes’ rule

- Need to characterize the posterior distribution (density $\pi_{\text{pos}}$)
- This is a challenging task since:
  - $x \in \mathbb{R}^n$ is typically **high-dimensional** (e.g., a discretized function)
  - $\pi_{\text{pos}}$ is **non-Gaussian**
  - evaluations of $\pi_{\text{pos}}$ may be **expensive**
- $\pi_{\text{pos}}$ can be evaluated up to a normalizing constant
Computational challenges

- Extract information from the posterior (means, covariances, event probabilities, predictions) by evaluating posterior expectations:

\[
\mathbb{E}_{\pi_{\text{pos}}}[h(x)] = \int h(x)\pi_{\text{pos}}(x)dx
\]

- Key strategies for making this computationally tractable
  - Approximations of the forward model, e.g., spectral expansions, local interpolants, reduced order models, multi-fidelity approaches
  - Efficient and structure-exploiting sampling schemes
Computational challenges

- Extract information from the posterior (means, covariances, event probabilities, predictions) by evaluating posterior expectations:

\[ \mathbb{E}_{\pi_{\text{pos}}}[h(x)] = \int h(x)\pi_{\text{pos}}(x)dx \]

- Key strategies for making this computationally tractable
  - Approximations of the forward model, e.g., spectral expansions, local interpolants, reduced order models, multi-fidelity approaches
  - Efficient and structure-exploiting sampling schemes
Core idea

- Choose $\pi_{\text{ref}}$ (e.g., Gaussian). Set $\pi_{\text{tar}} := \pi_{\text{pos}}$.
- Seek a map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_{\#}\pi_{\text{ref}} = \pi_{\text{tar}}$
Transport maps between probability distributions

Core idea

- Choose $\pi_{\text{ref}}$ (e.g., Gaussian). Set $\pi_{\text{tar}} := \pi_{\text{pos}}$.
- Equivalently, find $S = T^{-1}$ such that $S \# \pi_{\text{tar}} = \pi_{\text{ref}}$
Transport maps between probability distributions

Core idea

- Choose $\pi_{\text{ref}}$ (e.g., Gaussian). Set $\pi_{\text{tar}} := \pi_{\text{pos}}$.
- Seek a map $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $T_{\#} \pi_{\text{ref}} = \pi_{\text{tar}}$
- **Useful outcomes...**
  - *Independent* and *unweighted* samples from the target
  - “Precondition” other sampling or quadrature schemes
Various types of transport

- **Optimal transport:**
  
  \[ T^{\text{opt}} = \arg\min_T \int_{\mathbb{R}^n} c(x, T(x)) \, d\pi_{\text{ref}}(x) \]

  s.t. \( T\# \pi_{\text{ref}} = \pi_{\text{tar}} \)

  - Monge (1781) problem; many nice properties, but numerically challenging in general continuous cases...

- **Knothe-Rosenblatt rearrangement:**

  \[ T(x) = \begin{bmatrix}
  T^1(x_1) \\
  T^2(x_1, x_2) \\
  \vdots \\
  T^n(x_1, x_2, \ldots, x_n)
  \end{bmatrix} \]

  - Exists and is unique (up to ordering) under mild conditions
  - Jacobian determinant easy to evaluate
  - Monotonicity is essentially one-dimensional: \( \partial_{x_k} T^k > 0 \)
**Variational characterization** of the direct map $T$ [Moselhy & M 2012]:

$$\min_{T \in \mathcal{T}_\triangle} D_{KL}(T\# \pi_{\text{ref}} \mid\mid \pi_{\text{tar}})$$

- $\mathcal{T}_\triangle$ is the set of monotone lower **triangular** maps
  - Contains the *Knothe-Rosenblatt* rearrangement
- Expectation is with respect to *reference* measure
  - Compute via, e.g., Monte Carlo, QMC, quadrature
- Use evaluations of $\pi_{\text{tar}}$ (and its gradients) directly; **avoid** MCMC or importance sampling altogether!

- Parameterize $k$-th component map $T^k(x)$ with coefficients $f_k \in \mathbb{R}^{p_k}$
  - Monotone parameterization, $\partial_{x_k} T^k > 0$:

$$T^k(x_1, \ldots, x_k) = a_k(x_1, \ldots, x_{k-1}) + \int_0^{x_k} \exp(b_k(x_1, \ldots, x_{k-1}, w)) \, dw$$
\[
\begin{align*}
\min_{f_1, \ldots, f_n} & \mathbb{E}_{\pi_{\text{ref}}} \left[ -\log \pi_{\text{tar}} \circ T - \sum_k \log \partial_x T^k \right] \\
\text{Parameterized map} & \quad T(x; f_1, \ldots, f_n) \\
\text{Optimize over} & \quad f_1, \ldots, f_n \\
\text{Use gradient-based optimization} & \quad \text{(here, BFGS)} \\
\text{Approximate} & \quad \mathbb{E}_{\pi_{\text{ref}}}[g] \approx \sum_i w_i g(x_i) \\
\text{The posterior is in the tail of the reference!}
\end{align*}
\]
\[
\min_{f_1,\ldots,f_n} \mathbb{E}_{\pi_{\text{ref}}} \left[ -\log \pi_{\text{tar}} \circ T - \sum_k \log \partial x_k T^k \right]
\]

- Parameterized map \( T(x; f_1, \ldots, f_n) \)
- Optimize over \( f_1, \ldots, f_n \)
- Use gradient-based optimization (here, BFGS)
- Approximate \( \mathbb{E}_{\pi_{\text{ref}}} [g] \approx \sum_i w_i g(x_i) \)
- The posterior is in the tail of the reference!
\[
\min_{f_1, \ldots, f_n} \mathbb{E}_{\pi_{\text{ref}}} [-\log \pi_{\text{tar}} \circ T - \sum_k \log \partial_{x_k} T^k]
\]

- Parameterized map \( T(x; f_1, \ldots, f_n) \)
- Optimize over \( f_1, \ldots, f_n \)
- Use gradient-based optimization (here, BFGS)
- Approximate \( \mathbb{E}_{\pi_{\text{ref}}}[g] \approx \sum_i w_i g(x_i) \)
- The posterior is in the tail of the reference!
Simple example

\[
\min_{f_1, \ldots, f_n} \mathbb{E}_{\pi_{\text{ref}}} \left[ -\log \pi_{\text{tar}} \circ T - \sum_k \log \partial_{x_k} T^k \right]
\]

- Parameterized map \( T(x; f_1, \ldots, f_n) \)
- Optimize over \( f_1, \ldots, f_n \)
- Use gradient-based optimization (here, BFGS)
- Approximate \( \mathbb{E}_{\pi_{\text{ref}}}[g] \approx \sum_i w_i g(x_i) \)
- The posterior is in the tail of the reference!
Simple example

\[
\min_{f_1, \ldots, f_n} \mathbb{E}_{\pi_{\text{ref}}} \left[ -\log \pi_{\text{tar}} \circ T - \sum_k \log \partial x_k \, T^k \right]
\]

- Parameterized map \( T(x; f_1, \ldots, f_n) \)
- Optimize over \( f_1, \ldots, f_n \)
- Use gradient-based optimization (here, BFGS)
- Approximate \( \mathbb{E}_{\pi_{\text{ref}}}[g] \approx \sum_i w_i \, g(x_i) \)
- The posterior is in the tail of the reference!
Potential advantages

\[
\min_{f_1, \ldots, f_n} \mathbb{E}_{\pi_{\text{ref}}} \left[ -\log \pi_{\text{tar}} \circ T - \sum_{k} \log \partial x_k T^k \right]
\]

- **Move** samples; don’t just reweigh them!
- *Independent, unweighted, and cheap* samples from the target (or close to it): \( x_i \sim \pi_{\text{ref}} \Rightarrow T(x_i) \sim \pi_{\text{tar}} \)
- Clear convergence criterion, even with unnormalized target density:

\[
\mathcal{D}_{KL}( T_{\#} \pi_{\text{ref}} \big\| \pi_{\text{tar}} ) \approx \frac{1}{2} \text{Var}_{\pi_{\text{ref}}} \left[ \log \pi_{\text{ref}} - \log T_{\#}^{-1} \bar{\pi}_{\text{tar}} \right]
\]

- Key steps are embarrassingly parallel
Potential advantages

\[
\min_{f_1, \ldots, f_n} \mathbb{E}_{\pi_{\text{ref}}}[ - \log \pi_{\text{tar}} \circ T - \sum_k \log \partial_{x_k} T^k ]
\]

- **Move** samples; don’t just reweigh them!
- *Independent, unweighted, and cheap* samples from the target (or close to it: \( x_i \sim \pi_{\text{ref}} \Rightarrow T(x_i) \sim \pi_{\text{tar}} \))
- Clear convergence criterion, even with unnormalized target density:

\[
\mathcal{D}_{KL}( T_{\#} \pi_{\text{ref}} \mid \mid \pi_{\text{tar}} ) \approx \frac{1}{2} \text{Var}_{\pi_{\text{ref}}}[ \log \pi_{\text{ref}} - \log T_{\#}^{-1} \bar{\pi}_{\text{tar}} ]
\]

- Key steps are embarrassingly parallel
- Yet we exchange a high-dimensional sampling task for a **high-dimensional optimization problem**
  - **Major bottleneck:** dimension of the map basis \( f_1, \ldots, f_n \)
  - **KEY:** exploit Markov structure of the posterior
Markov properties and graphs

- Let $X_1, \ldots, X_n$ be random variables with joint density $\pi_{\text{pos}} > 0$
- The collection $X_1, \ldots, X_n$ satisfies the global Markov property relative to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ iff:
  $$X_A \perp \perp X_B \mid X_S$$  whenever $S$ is a separator set for $A, B \subset \mathcal{V}$

- $\mathcal{G}$ is an independence map (I-map) for $\pi_{\text{pos}}$
- $\mathcal{G}$ is a continuous and non-Gaussian Markov network for $\pi_{\text{pos}}$

Key message

Sparsity properties of $\mathcal{G}$ lead to efficient inference
∃ many I-maps for $\pi_{\text{pos}}$. Which one is \textbf{sparsest}?

**Theorem:** Define $G$ s.t. $(i, j) \notin \mathcal{E}$ if and only if $\partial_{x_i, x_j} \log \pi_{\text{pos}} = 0$. The resulting $G$ is the unique minimal independence map for $\pi_{\text{pos}}$.

$$(i, j) \notin \mathcal{E} \iff X_i \perp \perp X_j \mid \mathbf{X}_{\mathcal{V}\setminus\{i,j\}}$$

- Pairwise $\Rightarrow$ global Markov property if $\pi_{\text{pos}} > 0$
- Conditional independence is a second-order property for smooth $\pi_{\text{pos}}$
Example: stochastic volatility model. Latent log-volatilities obey an AR(1) process for $t = 1, \ldots, T$:

$$x_{t+1} = \phi x_t + \sigma^2 \eta_t, \quad \eta_t \sim \mathcal{N}(0, 1), \quad x_1 \sim \mathcal{N}(0, \frac{\sigma^2}{1 - \phi^2})$$

Observe the mean return for holding the asset at time $t$:

$$y_t = \epsilon_t \beta \exp(0.5 x_t), \quad \epsilon_t \sim \mathcal{N}(0, 1), \quad t = 1, \ldots, T$$

Consider the hyperparameters $(\phi, \sigma^2, \beta)$ fixed (for now):

$$\pi_{\text{pos}} \sim x_1, \ldots, x_T \mid \phi, \sigma^2, \beta, y_1, \ldots, y_T$$

The graphical model $G$ associated with $\pi_{\text{pos}}$ is a tree (chain)

Here, $G$ is found without computing Hessians of the posterior
Theorem [Sparsity of triangular transport]

If $G$ is an I-map for $\pi\text{pos}$, then we can determine *tight* lower bounds on the sparsity patterns of:

- **Direct** transport $T_\# \pi\text{ref} = \pi\text{pos}$
- **Inverse** transport $S_\# \pi\text{pos} = \pi\text{ref}$

only by performing operations on the graph $G$ (no need to evaluate $\pi\text{pos}$).

**Example:** Sparsity of inverse transport $S_\# \pi\text{pos} = \pi\text{ref}$

\[
P_{kj} = \partial_{x_j} S^k
\]
Sparsity of triangular transport

Theorem [Sparsity of triangular transport]

If $\mathcal{G}$ is an I-map for $\pi_{\text{pos}}$, then we can determine 
\textit{tight} lower bounds on the sparsity patterns of:

- **Direct** transport $T_{\|} \pi_{\text{ref}} = \pi_{\text{pos}}$
- **Inverse** transport $S_{\|} \pi_{\text{pos}} = \pi_{\text{ref}}$

only by performing operations on the graph $\mathcal{G}$ (no need to evaluate $\pi_{\text{pos}}$).

**Example:** Sparsity of inverse transport $S_{\|} \pi_{\text{pos}} = \pi_{\text{ref}}$

\[ \mathcal{G}^5 \]

\[ P_{kj} = \partial_{x_j} S^k \]
Theorem [Sparsity of triangular transport]

If \( G \) is an I-map for \( \pi_{\text{pos}} \), then we can determine tight lower bounds on the sparsity patterns of:

- **Direct** transport \( T_\# \pi_{\text{ref}} = \pi_{\text{pos}} \)
- **Inverse** transport \( S_\# \pi_{\text{pos}} = \pi_{\text{ref}} \)

only by performing operations on the graph \( G \) (no need to evaluate \( \pi_{\text{pos}} \)).

**Example:** Sparsity of inverse transport \( S_\# \pi_{\text{pos}} = \pi_{\text{ref}} \)
Sparsity of triangular transport

**Theorem [Sparsity of triangular transport]**

If $\mathcal{G}$ is an I-map for $\pi_{\text{pos}}$, then we can determine *tight* lower bounds on the sparsity patterns of:

- **Direct** transport $T_{\parallel} \pi_{\text{ref}} = \pi_{\text{pos}}$
- **Inverse** transport $S_{\parallel} \pi_{\text{pos}} = \pi_{\text{ref}}$

only by performing operations on the graph $\mathcal{G}$ (no need to evaluate $\pi_{\text{pos}}$).

**Example:** Sparsity of inverse transport $S_{\parallel} \pi_{\text{pos}} = \pi_{\text{ref}}$

\[ P_{kj} = \partial_{x_j} S^k \]
Theorem [Sparsity of triangular transport]

If $G$ is an l-map for $\pi_{pos}$, then we can determine \textit{tight} lower bounds on the sparsity patterns of:

- **Direct** transport $T_\pi \pi_{ref} = \pi_{pos}$
- **Inverse** transport $S_\pi \pi_{pos} = \pi_{ref}$

only by performing operations on the graph $G$ (no need to evaluate $\pi_{pos}$).

**Example:** Sparsity of inverse transport $S_\pi \pi_{pos} = \pi_{ref}$
Theorem [Sparsity of triangular transport]

If $\mathcal{G}$ is an I-map for $\pi_{pos}$, then we can determine \textit{tight} lower bounds on the sparsity patterns of:

- **Direct** transport $T_{\parallel} \pi_{ref} = \pi_{pos}$
- **Inverse** transport $S_{\parallel} \pi_{pos} = \pi_{ref}$

only by performing operations on the graph $\mathcal{G}$ (no need to evaluate $\pi_{pos}$).

**Example:** Sparsity of inverse transport $S_{\parallel} \pi_{pos} = \pi_{ref}$
Theorem [Sparsity of triangular transport]

If $\mathcal{G}$ is an l-map for $\pi_{\text{pos}}$, then we can determine tight lower bounds on the sparsity patterns of:

- **Direct** transport $T_{\#} \pi_{\text{ref}} = \pi_{\text{pos}}$
- **Inverse** transport $S_{\#} \pi_{\text{pos}} = \pi_{\text{ref}}$

only by performing operations on the graph $\mathcal{G}$ (no need to evaluate $\pi_{\text{pos}}$).

**Example:** Sparsity of inverse transport $S_{\#} \pi_{\text{pos}} = \pi_{\text{ref}}$
Inverse vs. direct transport

- Easy to find the inverse transport $S : \mathbb{R}^n \to \mathbb{R}^n$:

  $$\min_{S \in S_{\Delta}} D_{KL}(\pi_{\text{ref}} \parallel S\# \pi_{\text{pos}})$$

  $$P_{kj} = \partial_{x_j} S^k$$

- Trivial to invert a triangular function (sequence of 1D root findings)
- The sparsity pattern can be enforced in the approximation space $S_{\Delta}$
Inverse vs. direct transport

- Easy to find the inverse transport $S : \mathbb{R}^n \rightarrow \mathbb{R}^n$:
  $$\min_{S \in S_\Delta} \mathcal{D}_{KL}(\pi_{\text{ref}} \parallel S_{\#} \pi_{\text{pos}})$$

  $$P_{kj} = \partial_{x_j} S^k$$

- Trivial to invert a triangular function (sequence of 1D root findings)
- The sparsity pattern can be enforced in the approximation space $S_\Delta$

- What about the direct transport $T$?
- Direct transports tend to be extremely dense...
Compute the inverse transport!

Example: stochastic volatility model. $G$ is a chain.

Key message
Compute the inverse transport $S$ and evaluate $T(x) = S^{-1}(x)$ pointwise

- The inverse transport takes advantage of the sparse structure of $G$
- Same spirit as GMRF, but for general non-Gaussian distributions
The ordering problem

- Triangular transport $\implies$ need to fix ordering of input variables
- **Key observation:** the sparsity pattern of the inverse transport depends on the ordering of the input variables

$$\mathcal{G}$$  \hspace{1cm}  $P_{kj} = \partial_{x_j} S^k$

$$\mathcal{G} \circ \sigma$$  \hspace{1cm}  $P_{kj} = \partial_{x_j} S^k$

- How to find the best ordering?
- Standard question in graph theory (NP-hard problem)
- Use the same heuristics employed for **sparse Cholesky**!
The ordering problem

- **Example:** stochastic volatility model with hyperparameters
  \[ \pi_{\text{pos}} \sim x_1, \ldots, x_T, \phi, \sigma^2, \beta \mid y_1, \ldots, y_T \]
  - The graphical model \( \mathcal{G} \) associated with \( \pi_{\text{pos}} \) has now many cycles!

- A possible **best ordering** is: \( \phi, \sigma^2, \beta, x_1, \ldots, x_T \)
  - Complexity of inverse transport parameterization is linear in dimension.
The ordering problem

- **Example:** stochastic volatility model with hyperparameters
  \[ \pi_{\text{pos}} \sim x_1, \ldots, x_T, \phi, \sigma^2, \beta \mid y_1, \ldots, y_T \]

- The graphical model \( \mathcal{G} \) associated with \( \pi_{\text{pos}} \) has now many cycles!

- A possible **best ordering** is: \( \phi, \sigma^2, \beta, x_1, \ldots, x_T \)

- Complexity of inverse transport parameterization is linear in dimension.

- The direct transport is dense, but low dimensional structure might lie elsewhere...
Consider a simple I-map $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for a target density $\pi$

Fix a reference density $\eta = \prod_j \eta_{X_j}$, where $\eta_{X_j} = \int \eta \, dx_{\sim j}$

**Questions:**

- Is there a property of $\mathcal{G}$ that can induce low-dimensional structure in the direct transport $T$?
- *What is* this low-dimensional structure?
A triple \((A, S, B)\) of disjoint nonempty subsets of the vertex set \(\mathcal{V}\) forms a **decomposition** of \(G\) if the following hold:

1. \(\mathcal{V} = A \cup S \cup B\)
2. \(S\) separates \(A\) from \(B\) in \(G\)
3. \(S\) is a clique.
For a given decomposition \((A, S, B)\), consider the transport map:

\[
T_1(x) = \begin{bmatrix}
g(x_S) \\
h(x_S, x_A) \\
x_B
\end{bmatrix} = \begin{bmatrix}
T^{SUA}(x_{SUA}) \\
x_B
\end{bmatrix}
\]

such that the submap \(T^{SUA}\) satisfies \(\eta_{x_{SUA}} \xrightarrow{T^{SUA}} \pi x_{SUA}\).

What can we say about the pullback density \(T_1^\# \pi\)?
Figure: I-map for the pullback of $\pi$ through $T$

- Just remove any edge incident to any node in $A$
- $T_1$ is essentially a 3-D map
- Pulling back $\pi$ through $T_1$ makes $X_A$ independent of $X_{S \cup B}$!
Consider a new decomposition \((A, S, B)\)

Compute transport \(T_k\)

Pull back by \(T_k\)

Figure: I-map for the pullback of \(\pi\) through \(T\)

Recursion at step \(k\):

\[ T = T_1 \]
Graph decomposition

Figure: I-map for the pullback of $\pi$ through $T$

Recursion at step $k$:

1. Consider a new decomposition $(A, S, B)$
2. Compute transport $T_k$
3. Pull back by $T_k$
Graph decomposition

$T = T_1 \circ T_2$

- **Figure:** I-map for the pullback of $\pi$ through $T$
- $T_2$ is essentially a 4-D map.
- Each time we pull back by a new map we remove edges.
- **Intuition.** Continue the recursion until no edges are left...
Figure: I-map for the pullback of \( \pi \) through \( T \)

- \( T_2 \) is essentially a 4-D map.
- Each time we pull back by a new map we remove edges.
- Intuition. Continue the recursion until no edges are left...
Graph decomposition

Figure: I-map for the pullback of $\pi$ through $T$

- $T$ is still dense, but with a low-dimensional parameterization
- Decomposability of $\mathcal{G}$ induces a low-dimensional parameterization of the direct transport
- In fact, we can make this more general...

$T = T_1 \circ T_2 \circ T_3$
Theorem [Decomposition of transports]

Let $\mathcal{G}$ be an I-map for $\pi : \mathbb{R}^n \rightarrow \mathbb{R}_{++}$ and let $\eta = \prod_j \eta_{X_j}$ be a reference. If $(A, S, B)$ is a proper decomposition of $\mathcal{G}$ then:

1. ∃ a transport map:
   
   \[ T = T_1 \circ T_2 \]

   where
   
   - $T_1$ is a monotone triangular transport s.t. $\eta \xrightarrow{T_1} \pi_{X_{\text{AUS}}} \cdot (\prod_{j \in B} \eta_{X_j})$
   - $T_1$ is the identity map along components in $B$: $T_1^k(x) = x_k$ for $k \in B$
   - $T_2$ is any transport s.t. $\eta \xrightarrow{T_2} T_1^\# \pi$

2. $X_A$ is independent of $X_{\text{SUB}}$ w.r.t. the pullback density $T_1^\# \pi$
   
   - $T_2$ can be the identity along components in $A$: $T_2^k(x) = x_k$ for $k \in A$

Key point: can apply the theorem recursively to further decompose $T_2$

Decomposition of graph $\implies$ decomposition of transport
Decomposition of transports

$T = T_1 \circ T_2 \circ T_3$

- $G$ sparse implies \{S sparse and $T$ decomposable\} (dual perspective)
- Decomposability of $T$ can be predicted entirely by graph operations
- Then enforce low-dimensional structure in $\min D_{KL}(T \# \pi_{\text{ref}} \| \pi_{\text{pos}})$
- **Decouple** the nominal dimension of the problem from the dimension of each transport $T_k$
Decomposition of transports

\[ T = T_1 \circ T_2 \circ T_3 \]

- \( G \) sparse implies \{\( S \) sparse and \( T \) decomposable\} (dual perspective)
- Decomposability of \( T \) can be predicted entirely by graph operations
- Then enforce low-dimensional structure in \( \min D_{KL}( T_\# \pi_{ref} \| \pi_{pos} ) \)
- **Decouple** the nominal dimension of the problem from the dimension of each transport \( T_k \)
- In certain cases the transports \( (T_k)_k \) can also be computed sequentially . . .
Example: Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = \text{Id} \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

Start with the identity map.
**Example:** Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = \text{Id} \]

**Figure:** I-map for the pullback of \( \pi \) through \( T \)

Find a *good* first decomposition of \( \mathcal{G} \). **Warning:** \( S \) is not complete.
**Example:** Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = \text{Id} \]

**Figure:** I-map for the pullback of \( \pi \) through \( T \)

Make \( S \) complete. The resulting graph is still an I-map for \( \pi \).
Example: Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = T_1 \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

Compute \( T_1 \) and pull back \( \pi \)
Example: Stochastic volatility model with hyperparameters

- Build the decomposition recursively:
  
  \[ T = T_1 \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

- Find a decomposition of an I-map for the pullback of \( \pi \) through \( T_1 \)
Example: Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = T_1 \circ T_2 \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

Compute \( T_2 \) and pull back
**Example:** Stochastic volatility model with hyperparameters

**Build the decomposition recursively:**

\[ T = T_1 \circ T_2 \]

**Figure:** I-map for the pullback of \( \pi \) through \( T \)

**Continue the recursion until there are no edges left...**
**Example:** Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = T_1 \circ T_2 \circ T_3 \]

**Figure:** I-map for the pullback of \( \pi \) through \( T \)

Continue the recursion until there are no edges left...
Example: Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = T_1 \circ T_2 \circ T_3 \circ \cdots \circ T_{n-2} \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

Continue the recursion until there are no edges left...
Example: Stochastic volatility model with hyperparameters

Build the decomposition recursively:

\[ T = T_1 \circ T_2 \circ T_3 \circ \cdots \circ T_{n-2} \circ T_{n-1} \]

Figure: I-map for the pullback of \( \pi \) through \( T \)

Each map \( T_k \) is essentially 5-D regardless of \( n \)

We have solved a smoothing problem sequentially
Special case: Bayesian filtering and smoothing

- Given prior dynamic $\pi_{X_k|X_{k-1}}$, likelihood $\pi_{Y_k|X_k}$, reference dists $\eta_{X_k}$

- Define recursively $T_k(x_{k-1}, x_k) = [A_k(x_{k-1}, x_k); B_k(x_k)]$ s.t.

\[
\eta_{X_{k-1}} \cdot \eta_{X_k} \xrightarrow{T_k} \eta_{X_{k-1}} \cdot \pi_{Y_k|X_k} \cdot \pi_{X_k|X_{k-1}}(\cdot | B_{k-1}(\cdot))
\]

**Theorem** [Application of the decomposition theorem for trees]

1. $\eta_{X_k} \xrightarrow{B_k} \pi_{X_k|Y_1,\ldots,Y_k}$ (*filtering solution*)

2. Let $\mathcal{L}_k(x) := [x_{1:k-2}; T_k(x_{k-1}, x_k); x_{k+1:}\infty]$. Then

\[
\eta_{X_1} \cdots \eta_{X_k} \xrightarrow{(\mathcal{L}_1 \cdots \mathcal{L}_k)} \pi_{X_1,\ldots,X_k|Y_1,\ldots,Y_k} \text{ (*smoothing solution*)}
\]
Example: conditioned diffusion

Particle in a double-well potential, observed occasionally
Selected *conditionals* of the pullback $T^{-1}_{\#}\pi_{\text{tar}}$: linear map (Laplace) versus nonlinear map.
For certain graphs, the decomposition theorem does not imply decoupling between the nominal dimension of the problem and the dimension of each transport $T_i$

- Here, $G$ is an $n \times n$ grid graph
- $T^{SUA}$ acts on $2n$ dimensions at each stage

Nonetheless, the notion of composition of transports has still potential...
Beyond the decomposition theorem

- **Key idea:** seek near-identity structure instead of decompositions
- Example: fix target $\pi$ to be the posterior density of a Bayesian inference problem,

\[
\pi(x) := \pi_{\text{pos}}(x) \propto \pi_{Y|X}(y|x) \pi_X(x)
\]

- Let $T_{pr}$ push forward the reference $\eta$ to the prior $\pi_X$ (prior map)

\[
\hat{\pi}_{\text{pos}}(x) := T_{\#} \pi_{\text{pos}}(x) \propto \pi_{Y|X}(y|T_{pr}(x)) \eta(x)
\]

**Theorem [Graph decoupling]**

If $\eta = \prod_i \eta_{X_i}$ and

\[
\text{rank } \mathbb{E}_{\eta} [\nabla \log R \otimes \nabla \log R] = k, \quad R = \frac{\hat{\pi}_{\text{pos}}}{\eta} = \pi_{Y|X} \circ T_{pr}
\]

then there exists a rotation $Q$ such that:

\[
Q_{\#} \hat{\pi}_{\text{pos}}(x) = g(x_1, \ldots, x_k) \prod_{i>k} \eta_{X_i}(x_i)
\]
Changing the Markov structure...

The pullback has a different Markov structure:

\[ Q^\# \hat{\pi}_{\text{pos}}(\mathbf{x}) = g( x_1, \ldots, x_k ) \prod_{i>k}^{n} \eta X_i(x_i) \]

Corollary: There exists a transport \( T^\# \eta = Q^\# \hat{\pi}_{\text{pos}} \) of the form

\[ T(\mathbf{x}) = [ g(\mathbf{x}_{1:k}), x_{k+1}, \ldots, x_n ] \]

where \( g : \mathbb{R}^k \to \mathbb{R}^k \).

The composition \( T_{\text{pr}} \circ Q \circ T \) pushes forward \( \eta \) to \( \pi_{\text{pos}} \).

Why low rank structure? For example, few data-informed directions.
Log-Gaussian Cox process

(a) Prior sample
(b) Observations
(c) True field

- 4096-D **GMRF prior**, \( \mathbf{X} \sim \mathcal{N}(\mu, \Gamma) \), \( \Gamma^{-1} \) specified through \( \triangle + \kappa^2 \text{Id} \)
- 20 **sparse observations** at locations \( i \in \mathcal{I} \), \( \mathbf{Y}_i|\mathbf{X}_i \sim \text{Pois}(\exp \mathbf{X}_i) \)
- Posterior density \( \mathbf{X}|\mathbf{Y} \sim \pi_{\text{pos}} \) is:

\[
\pi_{\text{pos}}(\mathbf{x}) \propto \prod_{i \in \mathcal{I}} \exp[- \exp(x_i) + x_i \cdot y_i] \exp\left[-\frac{1}{2}(x - \mu)^\top \Gamma^{-1}(x - \mu)\right]
\]

- What is an independence map \( \mathcal{G} \) for \( \pi_{\text{pos}} \)?
Log-Gaussian Cox process

- 4096-D GMRF prior, $X \sim \mathcal{N}(\mu, \Gamma)$, $\Gamma^{-1}$ specified through $\triangle + \kappa^2 \text{Id}$
- 20 sparse observations at locations $i \in I$, $Y_i|X_i \sim \text{Pois}(\exp(X_i))$
- Posterior density $X|Y \sim \pi_{\text{pos}}$ is:

\[
\pi_{\text{pos}}(x) \propto \prod_{i \in I} \exp[-\exp(x_i) + x_i \cdot y_i] \exp\left[-\frac{1}{2}(x - \mu)^\top \Gamma^{-1}(x - \mu)\right]
\]

- What is an independence map $\mathcal{G}$ for $\pi_{\text{pos}}$? A $64 \times 64$ grid.
Fix $\pi_{\text{ref}} \sim \mathcal{N}(0, I)$ and let $T_{\text{pr}}$ push forward $\pi_{\text{ref}}$ to $\pi_{\text{pr}}$ (prior map).

Consider the pullback $\hat{\pi}_{\text{pos}} = T_{\text{pr}}^\# \pi_{\text{pos}}$ and find that

$$\text{rank } \mathbb{E}_{\pi_{\text{ref}}} \left[ \nabla \log R \otimes \nabla \log R \right] = 20 \ll n, \quad R = \hat{\pi}_{\text{pos}}/\pi_{\text{ref}}$$

Deflate the problem and compute a transport map in 20 dimensions.

Change from prior to posterior concentrated in a low-dimensional subspace ("likelihood-informed subspace"; Cui, Law, M 2015)
Given an approximate map \( \tilde{T} \), we can use the pushforward \( \tilde{T}^\# \pi_{\text{ref}} \) as a surrogate for \( \pi_{\text{tar}} \):

- KL divergence easily estimated via the pullback \( \tilde{T}^{-1} \pi_{\text{tar}} \)
- Biased expectations, but virtually no variance

Alternatively, we can directly evaluate the \textbf{pullback} density

\[
\tilde{T}^{-1} \pi_{\text{tar}} = (\pi_{\text{tar}} \circ \tilde{T}) \mid \det \nabla \tilde{T} \mid
\]

- Sample the pullback via MCMC [Parno & M 2015]
- Build another map...
- More generally: \textbf{precondition} any quadrature/sampling scheme
Other approaches to approximate transport

Thus far:

- We are solving optimization problems over spaces of maps.
- $T^h_\triangle$ and $S^h_\triangle$ inherit useful structure from $\pi_{\text{tar}}$.
- Spaces can be refined to approach an exact transport.
- Computable error bounds for approximate transports.
Other approaches to approximate transport

- **Thus far:**
  - We are solving optimization problems *over* spaces of maps
  - $\mathcal{T}_h^\Delta$ and $\mathcal{S}_h^\Delta$ inherit useful structure from $\pi_{\text{tar}}$
  - Spaces can be refined to approach an **exact** transport
  - Computable error bounds for *approximate* transports

- Can one *avoid* explicit parameterization of transports?
Other approaches to approximate transport

- **Thus far:**
  - We are solving optimization problems *over* spaces of maps
  - $T^h_\Delta$ and $S^h_\Delta$ inherit useful structure from $\pi_{\text{tar}}$
  - Spaces can be refined to approach an **exact** transport
  - Computable error bounds for *approximate* transports

- Can one *avoid* explicit parameterization of transports?

- **Optimization-based samplers:**
  - Implicit sampling [Chorin, Morzfeld, Tu 2009–2012]
  - Randomize-then-optimize (RTO) [Bardsley *et al.* 2014]
Other approaches to approximate transport

- **Thus far:**
  - We are solving optimization problems over spaces of maps
  - $\mathcal{T}_h^\Delta$ and $\mathcal{S}_h^\Delta$ inherit useful structure from $\pi_{\text{tar}}$
  - Spaces can be refined to approach an **exact** transport
  - Computable error bounds for **approximate** transports

- **Can one avoid** explicit parameterization of transports?

- **Optimization-based samplers:**
  - Implicit sampling [Chorin, Morzfeld, Tu 2009–2012]
  - Randomize-then-optimize (RTO) [Bardsley et al. 2014]

- **These samplers implement particular** transport maps
  - *Action* of the map is evaluated by solving an optimization problem
  - Necessarily *approximate*, but pushforward density is known; can then correct via Metropolization or importance weights
Central equation in implicit sampling:

$$\log \bar{\pi}_{\text{tar}}(x) - \mathcal{C} = \log \eta(\xi)$$

Underdetermined ($x, \xi \in \mathbb{R}^n$) but easily solved in certain settings...
Implicit sampling

- Central equation in implicit sampling:

\[
\log \bar{\pi}_{\text{tar}}(x) - c = \log \eta(\xi)
\]

- Underdetermined \((x, \xi \in \mathbb{R}^n)\) but easily solved in certain settings...

- Compare to optimality condition \((\mathcal{D}_{KL} = 0)\) for problem of finding a direct map \(T\), with \(x = T(\xi)\):

\[
\mathbb{E}_\eta[\log \bar{\pi}_{\text{tar}}(T(\xi)) + \log \det \nabla T(\xi) - \log \beta - \log \eta(\xi)] = 0
\]
Implicit sampling

- Central equation in implicit sampling:

\[ \log \tilde{\pi}_{\text{tar}}(x) - \mathcal{C} = \log \eta(\xi) \]

  - Underdetermined \((x, \xi \in \mathbb{R}^n)\) but easily solved in certain settings...

- Compare to optimality condition \((\mathcal{D}_{KL} = 0)\) for problem of finding a direct map \(T\), with \(x = T(\xi)\):

\[
\mathbb{E}_{\eta}[\log \tilde{\pi}_{\text{tar}}(T(\xi)) + \log \det \nabla T(\xi) - \log \beta - \log \eta(\xi)] = 0
\]

- **Key differences:**
  - Absence of Jacobian determinant; implicit samples *must* be endowed with weights
  - *Global statement* about a map \(T\) (appearing explicitly) versus a relationship between points in \(\mathbb{R}^n\)
Random-map implicit sampling [Morzfeld et al. 2012]

- Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + C$ have 
  *star-shaped* contours
- Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$
Random-map implicit sampling [Morzfeld et al. 2012]

- Let \( \log \pi_{\text{tar}}(x) = -\Phi(x) + \mathbf{c} \) have *star-shaped* contours
- Seek samples and weights \( \{(x^{(i)}, w^{(i)})\} \)
- Find posterior mode \( x_0 \)
Random-map implicit sampling [Morzfeld et al. 2012]

Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + c$ have 
*star-shaped* contours

Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$

1. Find posterior mode $x_0$
2. Find Hessian at mode 
$$[\nabla^2 \Phi]^{-1}_{x=x_0} = LL^T$$
Random-map implicit sampling [Morzfeld et al. 2012]

- Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + c$ have star-shaped contours
- Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$

1. Find posterior mode $x_0$
2. Find Hessian at mode $[\nabla^2 \Phi]^{-1}_{x=x_0} = LL^T$
3. For each sample:
   1. Draw $\xi^{(i)} \sim \eta = \mathcal{N}(0, I)$
Random-map implicit sampling [Morzfeld et al. 2012]

- Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + \mathcal{C}$ have *star-shaped* contours
- Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$

1. Find posterior mode $x_0$
2. Find Hessian at mode $[\nabla^2 \Phi]^{-1}_{x=x_0} = LL^T$
3. For each sample:
   1. Draw $\xi^{(i)} \sim \eta = \mathcal{N}(0, I)$
   2. Find $\hat{\lambda}^{(i)}$ such that
      \[
      \begin{cases}
      x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
      \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)\top} \xi^{(i)}
      \end{cases}
      \]
Random-map implicit sampling [Morzfeld et al. 2012]

- Let \( \log \pi_{\text{tar}}(x) = -\Phi(x) + \mathcal{C} \) have
  \textit{star-shaped} contours

- Seek samples and weights \( \{(x^{(i)}, w^{(i)})\} \)

1. Find posterior mode \( x_0 \)
2. Find Hessian at mode
\[
\left[ \nabla^2 \Phi \right]^{-1}_{x=x_0} = LL^\top
\]
3. For each sample:
   1. Draw \( \xi^{(i)} \sim \eta = \mathcal{N}(0, I) \)
   2. Find \( \hat{\lambda}^{(i)} \) such that
      \[
      \begin{cases}
        x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
        \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)\top} \xi^{(i)}
      \end{cases}
      \]
Random-map implicit sampling [Morzfeld et al. 2012]

Let \( \log \pi_{\text{tar}}(x) = -\Phi(x) + \mathcal{C} \) have \textit{star-shaped} contours

Seek samples and weights \( \{(x^{(i)}, w^{(i)})\} \)

1. Find posterior mode \( x_0 \)
2. Find Hessian at mode \( \left[ \nabla^2 \Phi \right]^{-1}_{x=x_0} = LL^\top \)
3. For each sample:
   1. Draw \( \xi^{(i)} \sim \eta = \mathcal{N}(0, I) \)
   2. Find \( \hat{\lambda}^{(i)} \) such that
      \[
      \begin{cases}
        x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
        \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)} \xi^{(i)}
      \end{cases}
      \]
Random-map implicit sampling [Morzfeld et al. 2012]

Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + \mathcal{C}$ have
star-shaped contours

Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$

1. Find posterior mode $x_0$
2. Find Hessian at mode
   \[ \left[ \nabla^2 \Phi \right]^{-1}_{x=x_0} = LL^T \]
3. For each sample:
   1. Draw $\xi^{(i)} \sim \eta = \mathcal{N}(0, I)$
   2. Find $\hat{\lambda}^{(i)}$ such that
      \[
      \begin{cases}
      x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
      \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)\top} \xi^{(i)}
      \end{cases}
      \]
   3. Calculate importance weight
      \[
      w^{(i)} \propto \frac{\pi_{\text{tar}}(x^{(i)})}{\mathcal{T} \# \eta(x^{(i)})}
      \]
Random-map implicit sampling [Morzfeld et al. 2012]

- Let $\log \pi_{\text{tar}}(x) = -\Phi(x) + \mathcal{C}$ have \textit{star-shaped} contours
- Seek samples and weights $\{(x^{(i)}, w^{(i)})\}$

1. Find posterior mode $x_0$
2. Find Hessian at mode $\left[ \nabla^2 \Phi \right]^{-1}_{x=x_0} = LL^T$
3. For each sample:
   - Draw $\xi^{(i)} \sim \eta = \mathcal{N}(0, I)$
   - Find $\hat{\lambda}^{(i)}$ such that
     $$\begin{cases} 
     x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
     \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)\top} \xi^{(i)}
     \end{cases}$$
   - Calculate importance weight
     $$w^{(i)} = \frac{\exp \left( -\Phi(x^{(i)}) \right)}{\left| \text{det} \frac{\partial x}{\partial \xi} \right|^{-1} \exp \left( -\frac{1}{2} \xi^{(i)\top} \xi^{(i)} \right)}$$
Random-map implicit sampling [Morzfeld et al. 2012]

- Let \( \log \pi_{\text{tar}}(x) = -\Phi(x) + C \) have *star-shaped* contours
- Seek samples and weights \( \{(x^{(i)}, w^{(i)})\} \)

1. Find posterior mode \( x_0 \)
2. Find Hessian at mode \( \left[ \nabla^2 \Phi \right]^{-1}_{x=x_0} = L L^\top \)
3. For each sample:
   1. Draw \( \xi^{(i)} \sim \eta = \mathcal{N}(0, I) \)
   2. Find \( \hat{\lambda}^{(i)} \) such that
      \[
      \begin{cases}
        x^{(i)} = x_0 + \hat{\lambda}^{(i)} L \xi^{(i)} \\
        \Phi(x^{(i)}) - \Phi(x_0) = \frac{1}{2} \xi^{(i)} \top \xi^{(i)}
      \end{cases}
      \]
   3. Calculate importance weight
      \[
      w^{(i)} = \frac{\exp \left( -\Phi(x^{(i)}) \right)}{\left| \det \frac{\partial x}{\partial \xi} \right|^{-1} \exp \left(-\frac{1}{2} \xi^{(i)} \top \xi^{(i)} \right)}
      \]
Let $\pi_{\text{tar}}$ have the special form:

$$\pi_{\text{tar}}(x) \propto \exp\left(-\frac{1}{2}\|F(x)\|^2\right)$$

for some $F : \mathbb{R}^n \to \mathbb{R}^m$ with $m \geq n$, such that there exists a $Q \in \mathbb{R}^{m \times n}$ for which $Q^\top \circ F : \mathbb{R}^n \to \mathbb{R}^n$ is invertible.

Put $\eta = \mathcal{N}(0, I)$ and $S = T^{-1} := Q^\top \circ F$. Then

$$T_\# \eta \propto \exp\left(-\frac{1}{2}\|Q^\top F(x)\|^2\right) |\det Q^\top \nabla F(x)|,$$

which can be close to $\pi_{\text{tar}}$.
Randomize-then-optimize (RTO) [Bardsley et al. 2014]

Let $\pi_{\text{tar}}$ have the special form:

$$
\pi_{\text{tar}}(x) \propto \exp\left(-\frac{1}{2}\|F(x)\|^2\right)
$$

for some $F : \mathbb{R}^n \to \mathbb{R}^m$ with $m \geq n$, such that there exists a $Q \in \mathbb{R}^{m \times n}$ for which $Q^\top \circ F : \mathbb{R}^n \to \mathbb{R}^n$ is invertible.

Put $\eta = \mathcal{N}(0, I)$ and $S = T^{-1} := Q^\top \circ F$. Then

$$
T_\# \eta \propto \exp\left(-\frac{1}{2}\|Q^\top F(x)\|^2\right) \det Q^\top \nabla F(x),
$$

which can be close to $\pi_{\text{tar}}$

Special form above corresponds to many Bayesian inverse problems:

$$
\pi_{\text{tar}}(x) = p(x|y) \propto \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} x \\ f(x) - y \\ F(x) \end{bmatrix}\right\|^2\right)
$$
$\pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right)$

Generate proposal samples $\{x_k\}$:
\[ \pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right) \]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
RTO: geometric interpretation

\[
\pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right)
\]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
2. Find \( Q \in \mathbb{R}^{m \times n} \) s.t. \( QR = \nabla F(x_0) \)
RTO: geometric interpretation

\[ \pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right) \]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
2. Find \( Q \in \mathbb{R}^{m \times n} \) s.t. \( QR = \nabla F(x_0) \)
3. For each sample:
   1. Sample \( \xi^{(i)} \) from standard Gaussian
RTO: geometric interpretation

\[ \pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right) \]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
2. Find \( Q \in \mathbb{R}^{m \times n} \) s.t. \( QR = \nabla F(x_0) \)
3. For each sample:
   1. Sample \( \xi^{(i)} \) from standard Gaussian
   2. Find \( x^{(i)} \) by minimizing \( \| Q^T (F(x) - \xi^{(i)}) \|^2 \)
RTO: geometric interpretation

\[ \pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right) \]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
2. Find \( Q \in \mathbb{R}^{m \times n} \) s.t. \( QR = \nabla F(x_0) \)
3. For each sample:
   1. Sample \( \xi^{(i)} \) from standard Gaussian
   2. Find \( x^{(i)} \) by minimizing \( \| Q^T (F(x) - \xi^{(i)}) \|^2 \)
   3. Evaluate proposal density:
      \[ q(x^{(i)}) \propto \left| \det Q^T \nabla F(x^{(i)}) \right| \exp \left( -\frac{1}{2} \| Q^T (F(x^{(i)})) \|^2 \right) \]
\[ \pi_{\text{tar}}(x) \propto \exp \left( -\frac{1}{2} \| F(x) \|^2 \right) \]

Generate proposal samples \( \{x_k\} \):

1. Find posterior mode \( x_0 \)
2. Find \( Q \in \mathbb{R}^{m \times n} \) s.t. \( QR = \nabla F(x_0) \)
3. For each sample:
   1. Sample \( \xi^{(i)} \) from standard Gaussian
   2. Find \( x^{(i)} \) by minimizing \( \| Q^\top (F(x) - \xi^{(i)}) \|^2 \)
   3. Evaluate proposal density:
      \[ q(x^{(i)}) \propto \left| \det Q^\top \nabla F(x^{(i)}) \right| \exp \left( -\frac{1}{2} \| Q^\top (F(x^{(i)})) \|^2 \right) \]
These optimization-based samplers implement certain transports:

- Inverse map $S$ is specified in closed form
- Repeatedly solve an optimization problem to evaluate $T(\xi) = S^{-1}(\xi)$
  - Versus explicit construction of $T$: solve a \textit{large} optimization problem to find the map, followed by cheap evaluations
- Pushforward density $T_\# \eta$ can also be evaluated

What can be gained from this perspective? Some generalizations:

- Adapt the available parameters of $S$
- Combine transport with general MCMC proposals
- Construct mixtures of transports...
MCMC with approximate transport: local vs. global proposals

Transport map accelerated MCMC

- Use any transition kernel $q(\xi \rightarrow \xi')$ for MCMC on $S^\#\pi_{\text{tar}}$ (left)
- Yields transformed proposal on the target distribution (right)
Transport map accelerated MCMC [Parno & M 2015]:
- Use any transition kernel \( q(\xi \rightarrow \xi') \) for MCMC on \( S_{\pi_{\text{tar}}} \) (left)
- Yields transformed proposal on the target distribution (right)
Transport map accelerated MCMC [Parno & M 2015]:

- Use any transition kernel $q(\xi \to \xi')$ for MCMC on $S_\# \pi_{\text{tar}}$ (left)
- Yields transformed proposal on the target distribution (right)
Example: local vs. global proposals

Target distributions $\pi_{\text{tar}}$:

(easy)

(hard)
Example: local vs. global proposals

Per 10000 MCMC steps:

<table>
<thead>
<tr>
<th>Implicit sampling</th>
<th>ESS (easy)</th>
<th>ESS (hard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>independence MH</td>
<td>5124</td>
<td>368</td>
</tr>
<tr>
<td>random walk MH</td>
<td>1276</td>
<td>749</td>
</tr>
</tbody>
</table>

Per 10000 function evaluations:

<table>
<thead>
<tr>
<th>Implicit sampling</th>
<th>ESS (easy)</th>
<th>ESS (hard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>independence MH</td>
<td>727</td>
<td>53</td>
</tr>
<tr>
<td>random walk MH</td>
<td>182</td>
<td>127</td>
</tr>
</tbody>
</table>
Conclusions

- Bayesian inference through the variational construction of transport maps
- Computation of transport maps in high dimensions, leveraging the Markov structure of the posterior:
  - Alternative: optimization-based samplers implement specific approximate transports
    - Correct via a rich variety of Metropolization techniques
- Much ongoing work...
  - Adaptive refinement of monotone maps
  - Preconditioning quadrature and QMC schemes
  - Approximately sparse Markov structures
  - Filtering and smoothing
References


- Map-accelerated MCMC implemented in MUQ (MIT Uncertainty Quantification library), [http://muq.mit.edu](http://muq.mit.edu)

- New papers and Python code on the way ...