Adaptive construction of measure transports for Bayesian inference

Daniele Bigoni Alessio Spantini Youssef M. Marzouk

Aeronautics and Astronautics, Massachusetts Institute of Technology

✉: {dabi,spantini,ymarz}@mit.edu, 🏠: http://uqgroup.mit.edu

Introduction

The solution of many inference problems requires the evaluation of integrals with respect to a complex – ***target*** *– distribution P. Chosen an amenable* ***reference*** *distribution R, we solve the* ***transportation problem****, i.e., we seek the map T that pushes forward R to P – denoted $T_\sharp R = P$ – or in other words:*

$$T : \mathbb{R}^d \to \mathbb{R}^d \quad \text{s.t.} \quad R(A) = P(T(A)) \ , \quad \forall A \in \sigma(\mathbb{R}^d) \ . \tag{1}$$

This map turns challenging integration problems into tractable ones:

$$I[f] = \int f(\boldsymbol{x}) P(d\boldsymbol{x}) = \int f \circ T(\boldsymbol{x}) R(d\boldsymbol{x}) \ . \tag{2}$$

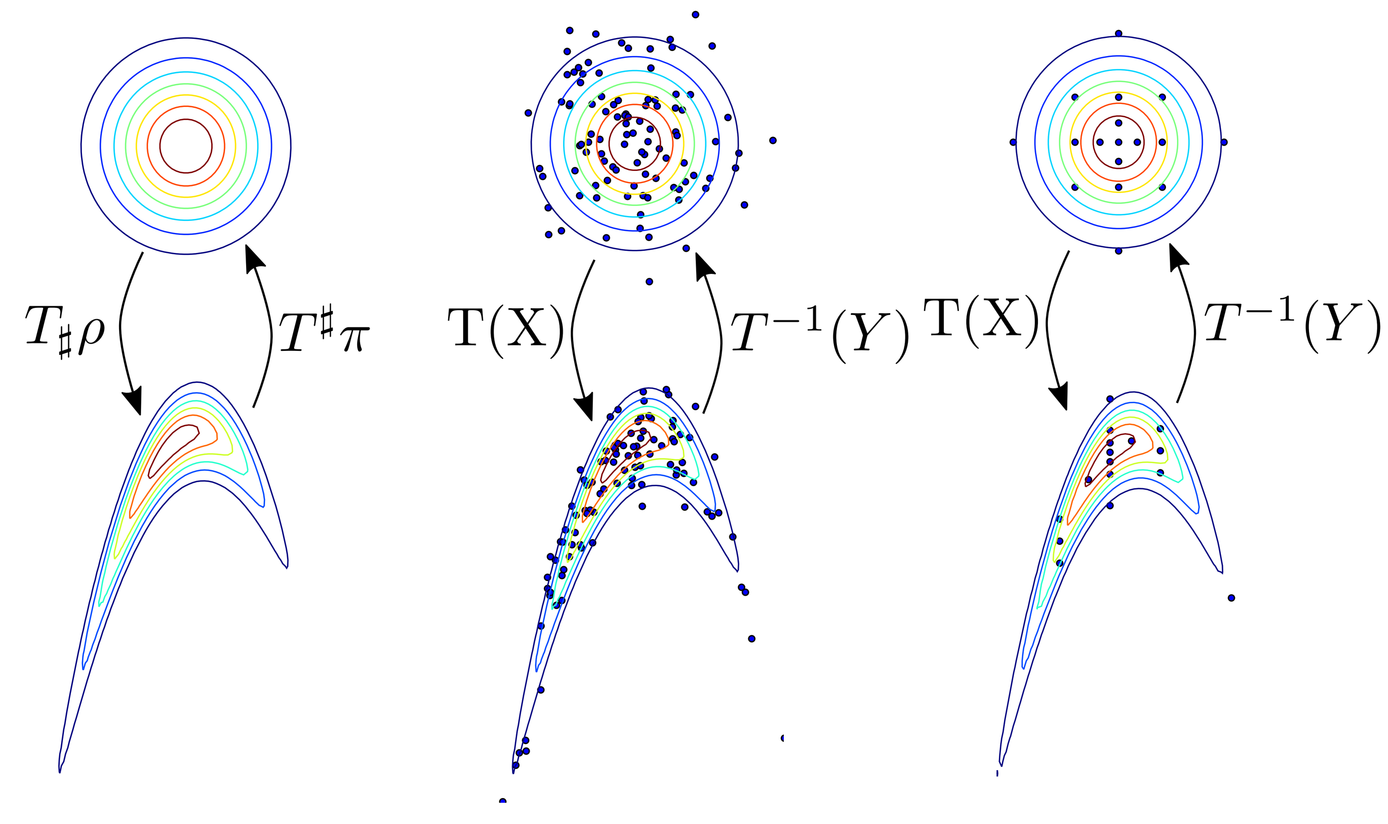
Measure transport

Let ρ be the density of R and $\tilde{\pi} = C\pi$ be the unnormalized density of P. The map T such that $T_\sharp R = P$ defines the following identities:

$$\textbf{Pushforward:} \quad T_\sharp \rho(\boldsymbol{x}) = \rho \circ T^{-1}(\boldsymbol{x}) \ |\nabla T^{-1}(\boldsymbol{x})| = \pi \ , \tag{3}$$

$$\textbf{Pullback:} \quad T^\sharp \pi(\boldsymbol{x}) = \pi \circ T(\boldsymbol{x}) \ |\nabla T(\boldsymbol{x})| = \rho \ , \tag{4}$$

and is such that, for $\mathbf{X} \sim R$, $T(\mathbf{X}) \sim P$.



(a) Density transformation (b) Monte Carlo sampling (c) Sparse grid quadrature

Knothe-Rosenblatt rearrangement [1]

Let the spaces of **triangular and monotone** maps be:

$$\begin{aligned} \mathcal{H}_\triangle &:= \{T : \mathbb{R}^d \to \mathbb{R}^d \mid [T(\boldsymbol{x})]_i := T^{(i)}(x_1, \dots, x_i)\} \\ \mathcal{T}_\triangle &:= \{T : \mathbb{R}^d \to \mathbb{R}^d \mid T \in \mathcal{H}_\triangle \text{ and } \partial_{x_i} T^{(i)} > 0\} \end{aligned} \tag{5}$$

For any two measures R and P there exists a unique $T^\star \in \mathcal{T}_\triangle$ such that $T^\star_\sharp R = P$, provided R has no atoms.

The transport map framework

The transportation problem is casted as a minimization problem in terms of the **Kullback-Leibler divergence** [2, 3]:

$$T^\star = \arg\min_{T \in \mathcal{T}_\triangle} \mathcal{D}_{\text{KL}}(T_\sharp \rho \| \tilde{\pi}) = \arg\min_{T \in \mathcal{T}_\triangle} \mathbb{E}_\rho \left[-\log T^\sharp \tilde{\pi} \right] \ . \tag{6}$$

We approximate $\mathbb{E}_\rho$ using the quadrature $\mathcal{Q}_q$:

$$T^\star \approx T^\star_q = \arg\min_{T \in \mathcal{T}_\triangle} \mathcal{Q}_q \left(-\log T^\sharp \tilde{\pi} \right) = \arg\min_{T \in \mathcal{T}_\triangle} - \sum_{i=1}^{q} \log T^\sharp \tilde{\pi}(\boldsymbol{x}_i) \ w_i \ . \tag{7}$$

We can express every element of $\mathcal{T}_\triangle$ by defining components $T^{(i)}$ as [4]:

$$T^{(i)}(x_{1:i}) = c^{(i)}(x_{1:i-1}) + \int_0^{x_i} \exp\left(h^{(i)}(x_{1:i-1}, t) \right) dt \ . \tag{8}$$

Let $\mathcal{T}^k_\triangle \subset \mathcal{T}_\triangle$, with $n_k = \dim(\mathcal{T}^k_\triangle) < \dim(\mathcal{T}_\triangle) = \infty$. The approximate transportation problem (7) can be further approximated by:

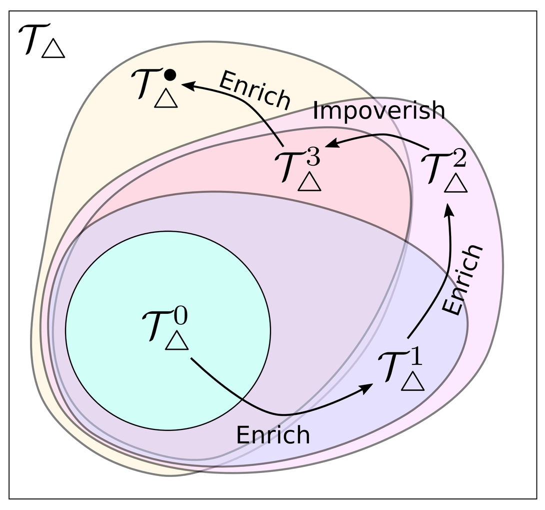
$$T^\star_q \approx T^\star_{q,k} = \arg\min_{T \in \mathcal{T}^k_\triangle} - \sum_{i=1}^{q} \log T^\sharp \tilde{\pi}(\boldsymbol{x}_i) \ w_i = \arg\min_{T \in \mathcal{T}^k_\triangle} \mathcal{J}_q(T) \ . \tag{9}$$

The following **variance diagnostic** is a global convergence criterion for the approximation of $T^\star$ by $T^\star_{q,k}$ in the space $\mathcal{T}_\triangle$, i.e. for $T \to T^\star$

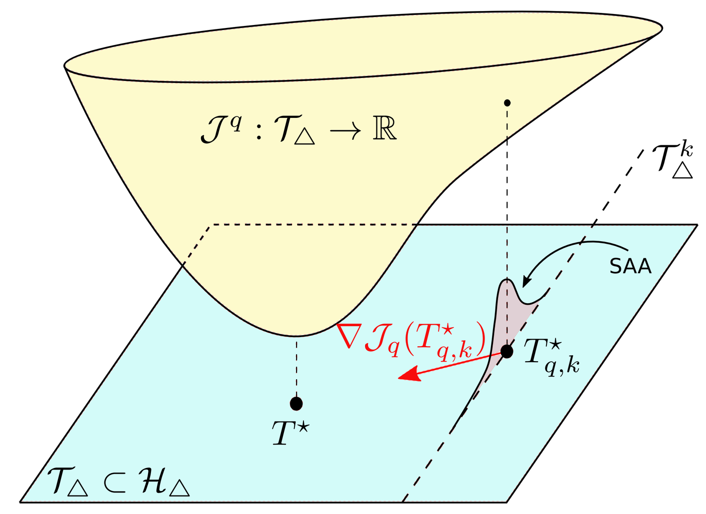
$$\mathbb{V}_\rho \left[\log \frac{\rho}{T^\sharp \tilde{\pi}} \right] \to 0 \quad \text{asymptotically as} \quad \frac{1}{2} \mathcal{D}_{\text{KL}}(T_\sharp \rho \| \tilde{\pi}) \to -\log C \ . \tag{10}$$

Adaptive enrichment/impoverishment

The goal of an adaptive scheme is the selection of $\mathcal{T}^\bullet_\triangle \subset \mathcal{T}_\triangle$ $(\dim(\mathcal{T}^\bullet_\triangle) = n_\bullet < \infty)$ and $q_\bullet \in \mathbb{N}$ such that $\mathbb{V}_\rho \left[\log \left(\rho / T^{\star,\sharp}_{q_\bullet,\bullet} \tilde{\pi} \right) \right] < \varepsilon_\bullet$.



(d) Adaptive procedure



(e) First variation

1 Sample average approximation [5]: $T^\star_{q,k} = \arg\min_{T \in \mathcal{T}^k_\triangle} \mathcal{J}_q(T)$

- If $\dim \mathcal{T}^k_\triangle$ good for q and $\mathbb{V}_\rho \left[\log \left(\rho / T^{\star,\sharp}_{q,k} \tilde{\pi} \right) \right] < \varepsilon_\bullet$, then `DONE`
- If $\dim \mathcal{T}^k_\triangle$ too big for q, **impoverish** to obtain $\mathcal{T}^{k+1}_\triangle \subset \mathcal{T}^k_\triangle$ (`goto` 1)

2 If $\mathcal{T}^k_\triangle \subsetneq \mathcal{T}^\bullet_\triangle$, then the **first variation** $\nabla \mathcal{J}_q \left(T^\star_{q,k} \right) \in \mathcal{H}_\triangle$ is

$$0 \neq \nabla \mathcal{J}_q \left(T^\star_{q,k} \right) = \left(\nabla_{\boldsymbol{x}} T^\star_{q,k} \right)^{-1} \left(\nabla_{\mathbf{x}} \log \frac{\rho}{T^{\star,\sharp}_{q,k} \tilde{\pi}} \right) \ . \tag{11}$$

Then $\exists \ \varepsilon > 0$ s.t. $\mathcal{J}_q \left(T^\star_{q,k} - \varepsilon \nabla \mathcal{J}_q \left(T^\star_{q,k} \right) \right) < \mathcal{J}_q \left(T^\star_{q,k} \right)$

3 Approximate $U^\star_{q,k+1} = \arg\min_{U \in \mathcal{H}^{k+1}_\triangle} \left\| U - \nabla \mathcal{J}_q \left(T^\star_{q,k} \right) \right\|_{L^2_\rho}$

4 Enrich space to obtain $\mathcal{T}^{k+1}_\triangle \supset \mathcal{T}^k_\triangle$ (`goto` 1).

Stochastic volatility: state-parameter estimation

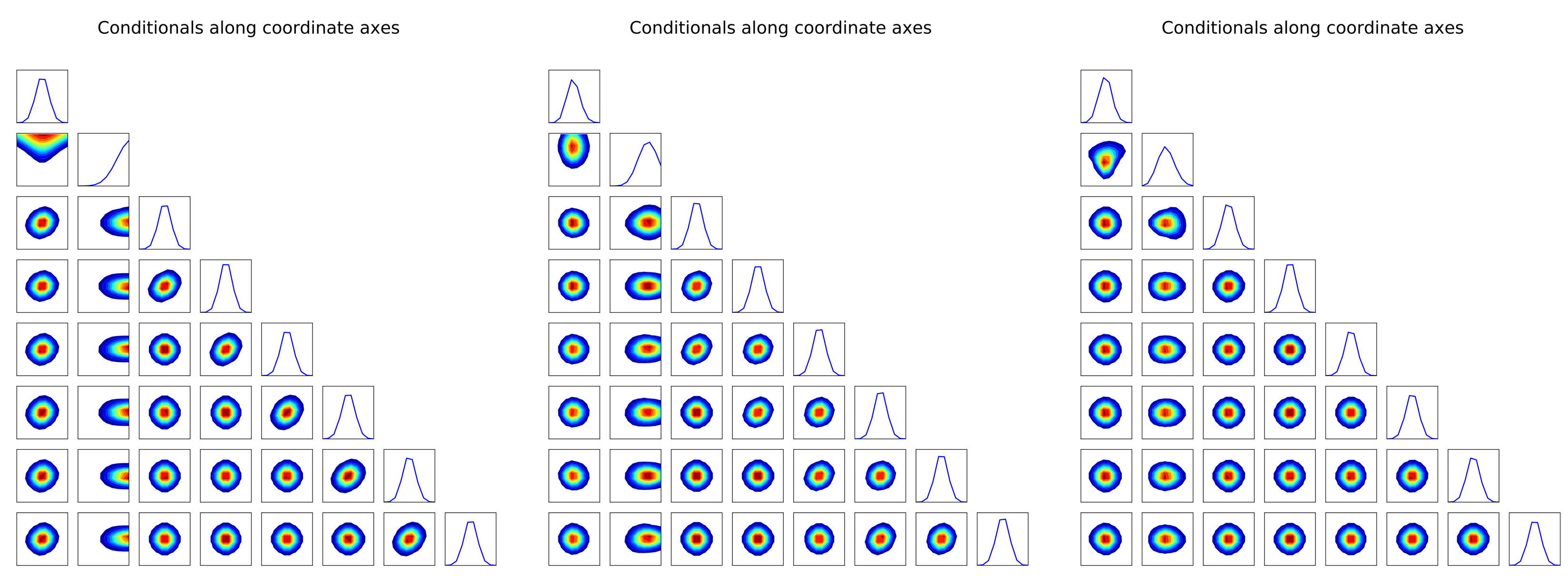
Auto-regressive process to model the log-volatility X_t of an asset at time t [6]:

$$X_{t+1} = \mu + \phi(X_t - \mu) + \eta_t \ , \quad \eta_t \sim \mathcal{N}(0,1) \ , \quad X_1 \sim \mathcal{N}\left(0, 1/\left(1 - \phi^2\right)\right) \ , \tag{12}$$

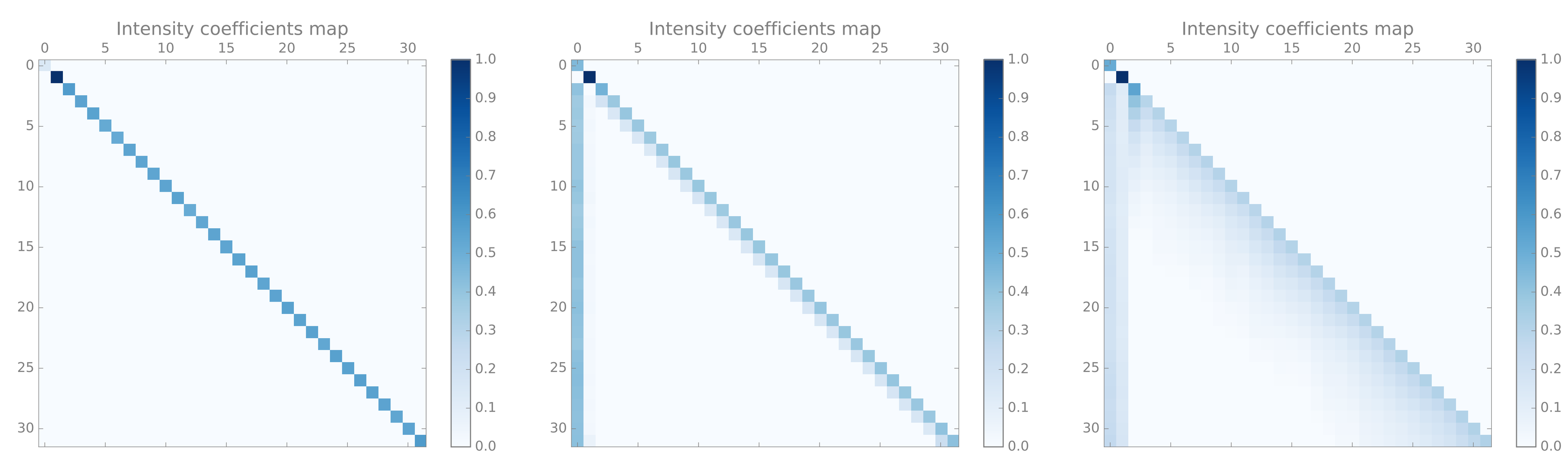
where μ, ϕ are hyper-parameters with $\mu \sim \mathcal{N}(0, \sigma^2_\mu)$ and $\frac{\phi+1}{2} \sim \text{Beta}(10,1)$.
The observed return Y_t follows the price evolution model

$$Y_t = \varepsilon_t \exp(X_t/2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1) \ . \tag{13}$$

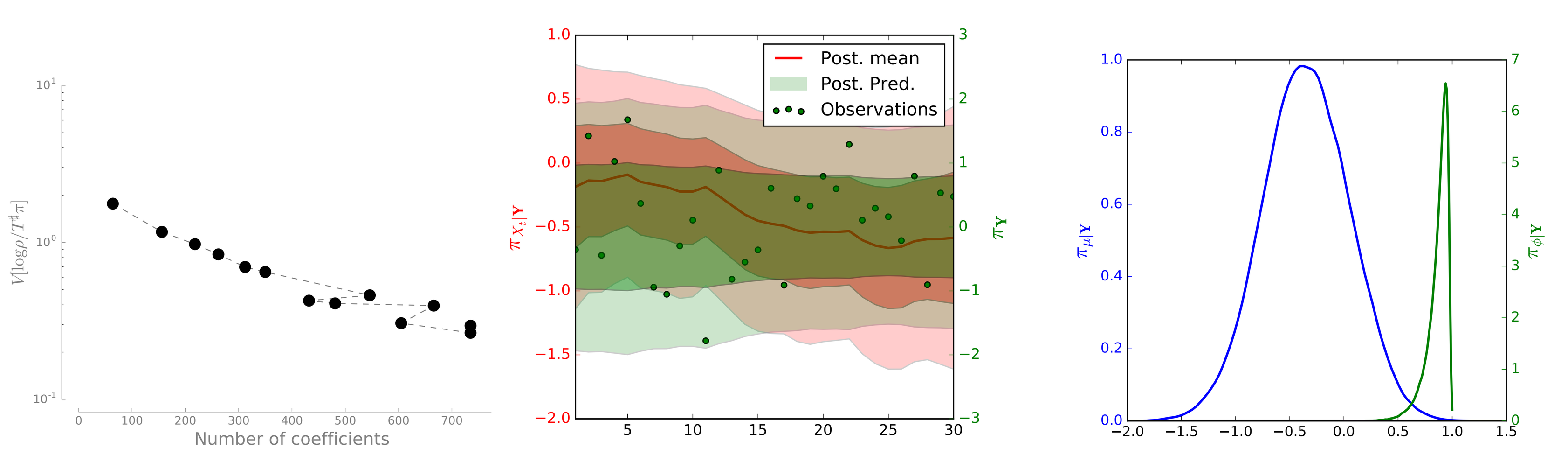
We characterize the **full posterior** $\pi \sim \mu, \phi, \mathbf{X}_{1:N} | \mathbf{Y}_{1:N}$ up to time $N = 30$.



(f) Iteration 0 - $T^\sharp\tilde{\pi}$ (g) Iteration 1 - $T^\sharp\tilde{\pi}$ (h) Iteration 12 - $T^\sharp\tilde{\pi}$



(i) Iteration 0 - $\nabla_{\boldsymbol{x}} T$ (j) Iteration 1 - $\nabla_{\boldsymbol{x}} T$ (k) Iteration 12 - $\nabla_{\boldsymbol{x}} T$



(l) Variance diagnostic (m) Posterior and post. predictive (n) Posterior hyper-parameters

Conclusions

- Approximate complex distribution P
- Cheap i.i.d. sampling from P
- Optimization based approach
- Asymptotic convergence criterion
- Robust adaptive algorithm exploiting marginal independence

Outlook

- Analysis of the interaction between tolerances and stopping criteria
- Embed in the approximation of decomposable transports [7]
- Applications on PDE driven inference problems

[1] J. V. Ryff. "Measure preserving transformations and rearrangements". In: *Journal of Mathematical Analysis and Applications* 31.2 (1970).

[2] T. A. El Moselhy et al. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* (2012).

[3] Y. Marzouk et al. "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification*. Ed. by R. G. Ghanem et al. Cham: Springer International Publishing, 2016.

[4] D. Bigoni et al. "On the computation of monotone transports". In: *Preprint* (2016).

[5] A. Shapiro. "Sample Average Approximation". In: *Encyclopedia of Operations Research and Management Science*. 3. Boston, MA: Springer US, 2013.

[6] J. Durbin et al. "Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.1 (2000).

[7] A. Spantini et al. "Measure transport, variational inference, and low-dimensional maps". In: *Preprint* (2016).

This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0010518 and DE-SC000929.7