An Automated Measure Transport Framework for Online Nonlinear Filtering and Smoothing

https://transportmaps.mit.edu

D. Bigoni*, A. Spantini, Y. Marzouk
Massachusetts Institute of Technology

Past and present contributors:
T. Moselhy, M. Parno, R. Morrison,
R.M. Baptista, X. Huan, B. Zhang, Z. Wang

SIAM Annual Meeting – July 12, 2017

Inference via low-dimensional couplings

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk

Massachusetts Institute of Technology
Cambridge, MA 02139 USA
e-mail: spantini@mit.edu, dabi@mit.edu, ymarz@mit.edu

Abstract: Integration against an intractable probability measure is among the fundamental challenges of statistical inference, particularly in the Bayesian setting. A principled approach to this problem seeks a deterministic coupling of the measure of interest with a tractable "reference" measure (e.g., a standard Gaussian). This coupling is induced by a transport map, and enables direct simulation from the desired measure simply by evaluating the transport map at samples from the reference. Yet characterizing such a map—e.g., representing and evaluating it—grows challenging in high dimensions. The central contribution of this paper is to establish a link between the Markov properties of the target measure and the existence of certain *low-dimensional* couplings, induced by transport maps that are *sparse* or *decomposable*. Our analysis not only facilitates the construction of couplings in high-dimensional settings, but also suggests new inference methodologies. For instance, in the context of nonlinear and non-Gaussian state space models, we describe new variational algorithms for online filtering, smoothing, and parameter estimation. These algorithms implicitly characterize—via a transport map—the full posterior distribution of the sequential inference problem using *local* operations only incrementally more complex than regular filtering, while avoiding importance sampling or resampling.

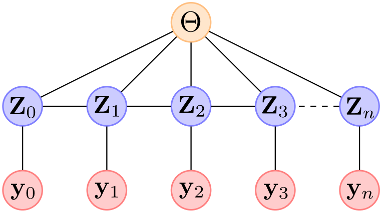
arXiv: 1703.06131

State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) , & \mathbf{w}_k &\sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) , \quad k \in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) , & \mathbf{v}_k &\sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) , \quad k \in \Xi \subset \Lambda \\
\mathbf{Z}_0 &\sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) , \quad \boldsymbol{\Theta} \sim \nu_\theta
\end{aligned}
$$

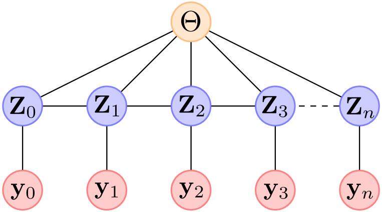
State-space model with hyperparameters

$$\mathbf{Z}_{k+1} = G(\mathbf{Z}_k, \mathbf{w}_k, \mathbf{\Theta}) \ , \qquad \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\mathbf{\Theta}) \ , \quad k \in \Lambda = \{0, \ldots, n\}$$
$$\mathbf{Y}_k = H(\mathbf{Z}_k, \mathbf{v}_k, \mathbf{\Theta}) \ , \qquad \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\mathbf{\Theta}) \ , \quad k \in \Xi \subset \Lambda$$
$$\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\mathbf{\Theta}) \ , \qquad \mathbf{\Theta} \sim \nu_\theta$$



State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \;, & \mathbf{w}_k &\sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \;, & k &\in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \;, & \mathbf{v}_k &\sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \;, & k &\in \Xi \subset \Lambda \\
\mathbf{Z}_0 &\sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \;, & \boldsymbol{\Theta} &\sim \nu_\theta
\end{aligned}
$$



Full Bayesian solution

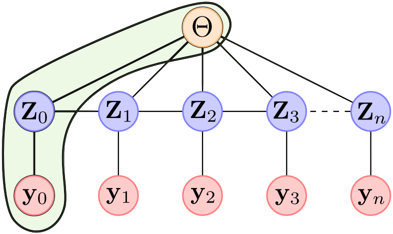
$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

State-space model with hyperparameters

$$\mathbf{Z}_{k+1} = G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \ , \qquad \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \ , \quad k \in \Lambda = \{0, \ldots, n\}$$
$$\mathbf{Y}_k = H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \ , \qquad \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \ , \quad k \in \Xi \subset \Lambda$$
$$\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \ , \qquad \boldsymbol{\Theta} \sim \nu_\theta$$



Filtering

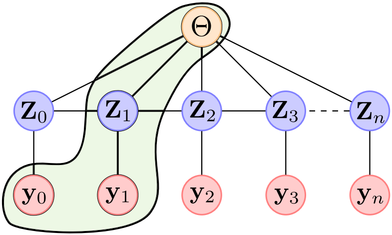
$$\pi\left(\Theta, \mathbf{Z}_0 | \mathbf{y}_{\Xi_{\leq 0}}\right)$$

Full Bayesian solution

$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$
$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$
$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \ , \qquad && \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \ , \quad && k \in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \ , \qquad && \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \ , \quad && k \in \Xi \subset \Lambda \\
\mathbf{Z}_0 &\sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \ , \qquad && \boldsymbol{\Theta} \sim \nu_\theta
\end{aligned}
$$



Filtering

$$\pi\left(\Theta, \mathbf{Z}_1 | \mathbf{y}_{\Xi_{\leq 1}}\right)$$

Full Bayesian solution

$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

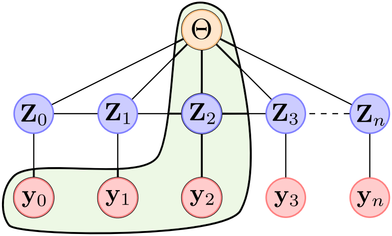
$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \ , & \mathbf{w}_k &\sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \ , & k &\in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \ , & \mathbf{v}_k &\sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \ , & k &\in \Xi \subset \Lambda
\end{aligned}
$$

$$
\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \ , \qquad \boldsymbol{\Theta} \sim \nu_\theta
$$



Filtering

$$
\pi\left(\Theta, \mathbf{Z}_2 | \mathbf{y}_{\Xi_{\leq 2}}\right)
$$

Full Bayesian solution

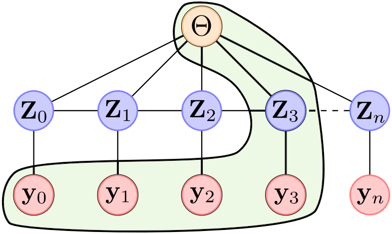
$$
\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)
$$

$$
\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)
$$

$$
\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)
$$

State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \,, && \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \,, \quad k \in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \,, && \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \,, \quad k \in \Xi \subset \Lambda \\
\mathbf{Z}_0 &\sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \,, \quad \boldsymbol{\Theta} \sim \nu_\theta
\end{aligned}
$$



Filtering

$$\pi\left(\Theta, \mathbf{Z}_3 | \mathbf{y}_{\Xi_{\leq 3}}\right)$$

Full Bayesian solution

$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

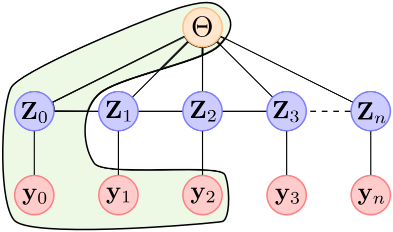
$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \ , \qquad & \mathbf{w}_k &\sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \ , \quad k \in \Lambda = \{0, \dots, n\} \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \ , \qquad & \mathbf{v}_k &\sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \ , \quad k \in \Xi \subset \Lambda
\end{aligned}
$$

$$
\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \ , \qquad \boldsymbol{\Theta} \sim \nu_\theta
$$



Fixed lag smoothing

$$
\pi\left(\Theta, \mathbf{Z}_0 | \mathbf{y}_{\Xi_{\leq 2}}\right)
$$

Full Bayesian solution

$$
\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)
$$

$$
\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)
$$

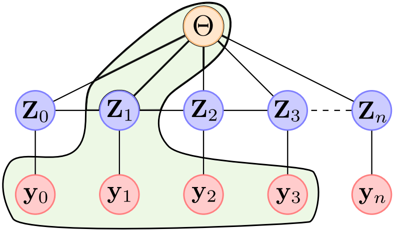
$$
\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)
$$

State-space model with hyperparameters

$$\mathbf{Z}_{k+1} = G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \ , \qquad \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \ , \quad k \in \Lambda = \{0, \dots, n\}$$

$$\mathbf{Y}_k = H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \ , \qquad \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \ , \quad k \in \Xi \subset \Lambda$$

$$\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \ , \qquad \boldsymbol{\Theta} \sim \nu_\theta$$



Fixed lag smoothing

$$\pi\left(\Theta, \mathbf{Z}_1 | \mathbf{y}_{\Xi_{\leq 3}}\right)$$

Full Bayesian solution

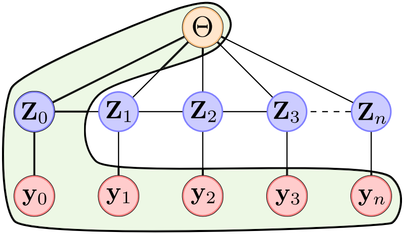
$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$

$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$

$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

State-space model with hyperparameters

$$\mathbf{Z}_{k+1} = G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) , \qquad \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) , \quad k \in \Lambda = \{0, \dots, n\}$$
$$\mathbf{Y}_k = H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) , \qquad \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) , \quad k \in \Xi \subset \Lambda$$
$$\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) , \qquad \boldsymbol{\Theta} \sim \nu_\theta$$



Fixed point smoothing

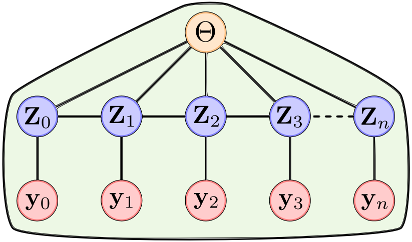
$$\pi \left(\Theta, \mathbf{Z}_0 | \mathbf{y}_\Xi \right)$$

Full Bayesian solution

$$\pi \left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi \right) \propto \mathcal{L} \left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda \right) \pi \left(\Theta, \mathbf{Z}_\Lambda \right)$$
$$\mathcal{L} \left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda \right) = \prod_{k \in \Xi} \mathcal{L} \left(\mathbf{y}_k | \Theta, \mathbf{Z}_k \right)$$
$$\pi \left(\Theta, \mathbf{Z}_\Lambda \right) = \pi \left(\Theta \right) \pi \left(\mathbf{Z}_0 | \Theta \right) \prod_{k \in \Lambda} \pi \left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta \right)$$

State-space model with hyperparameters

$$\mathbf{Z}_{k+1} = G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \;, \qquad \mathbf{w}_k \sim \boldsymbol{\nu}_{w_k}(\boldsymbol{\Theta}) \;, \quad k \in \Lambda = \{0, \ldots, n\}$$
$$\mathbf{Y}_k = H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \;, \qquad \mathbf{v}_k \sim \boldsymbol{\nu}_{v_k}(\boldsymbol{\Theta}) \;, \quad k \in \Xi \subset \Lambda$$
$$\mathbf{Z}_0 \sim \boldsymbol{\nu}_0(\boldsymbol{\Theta}) \;, \qquad \boldsymbol{\Theta} \sim \nu_\theta$$



Full Bayesian solution

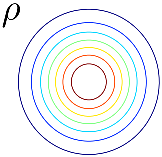
$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right)$$

Full Bayesian solution

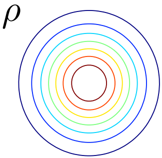
$$\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)$$
$$\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k\in\Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)$$
$$\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k\in\Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)$$

Measure transport – pullbacks [PB] and pushforwards [PF]

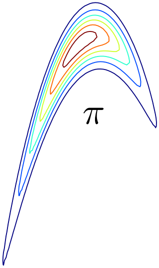
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$



Measure transport – pullbacks [PB] and pushforwards [PF]



- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$

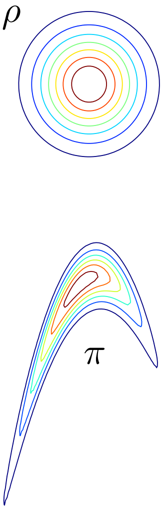


Measure transport – pullbacks [PB] and pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$



Measure transport – pullbacks [PB] and pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

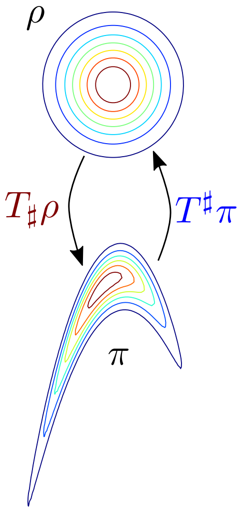
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

- We want T such that

  PF $\quad T_\sharp \rho = \pi$

  PB $\quad T^\sharp \pi = \rho$



Measure transport – pullbacks [PB] and pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

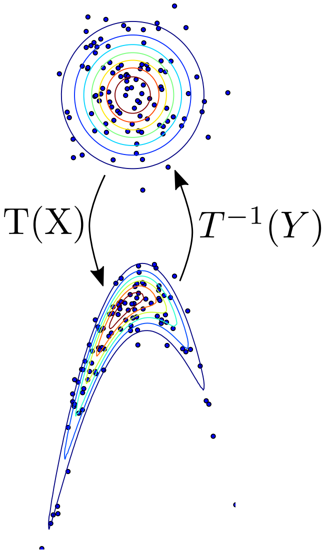
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

- We want T such that

  PF $\quad$ For $X \sim \boldsymbol{\nu}_\rho$, $T(X) \sim \boldsymbol{\nu}_\pi$

  PB $\quad$ For $Y \sim \boldsymbol{\nu}_\pi$, $T^{-1}(Y) \sim \boldsymbol{\nu}_\rho$



Measure transport – pullbacks [PB] and pushforwards [PF]

- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

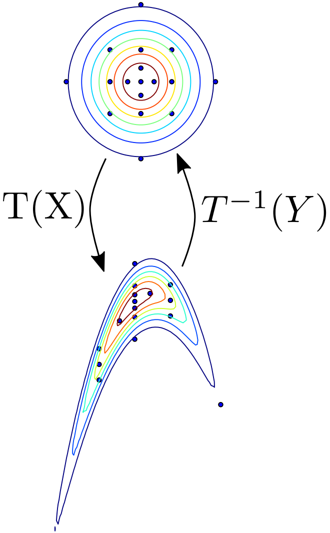
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  PB $\quad T^\sharp \pi = \pi \circ T |\nabla T|$

- We want T such that

  PF $\quad$ For $X \sim \boldsymbol{\nu}_\rho$, $T(X) \sim \boldsymbol{\nu}_\pi$

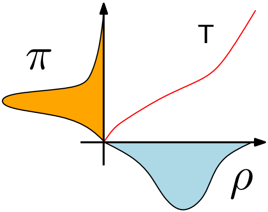
  PB $\quad$ For $Y \sim \boldsymbol{\nu}_\pi$, $T^{-1}(Y) \sim \boldsymbol{\nu}_\rho$



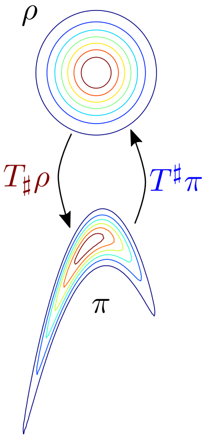
Knothe-Rosenblatt rearrangement

For any $\boldsymbol{\nu}_\rho, \boldsymbol{\nu}_\pi$ Lebesgue absolutely continuous there exists a **triangular monotone** map $T \in \mathcal{T}_>$ s.t.

$$T_\sharp \rho = \pi$$



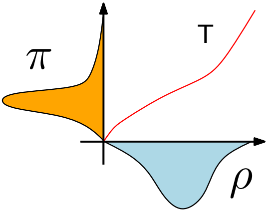
$$T(\mathbf{x}) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, \dots, x_d) \end{bmatrix}$$



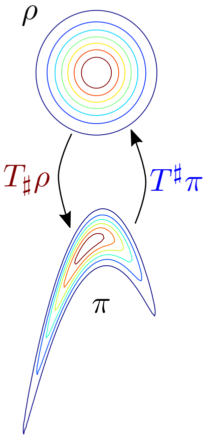
Knothe-Rosenblatt rearrangement

For any $\boldsymbol{\nu}_\rho, \boldsymbol{\nu}_\pi$ Lebesgue absolutely continuous there exists a **triangular monotone** map $T \in \mathcal{T}_>$ s.t.

$$T_\sharp \rho = \pi$$



$$T(\mathbf{x}) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, \dots, x_d) \end{bmatrix}$$



How to find the map $T \in \mathcal{T}_>$
such that $T_\sharp \rho = \pi$?

Minimize KL-divergence to find optimal map

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

Minimize KL-divergence to find optimal map

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T^\star_\sharp \rho \propto \pi$ **in parallel**

Minimize KL-divergence to find optimal map

$$T^\star = \underset{T \in \mathcal{T}_>}{\arg\min}\, D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) = \underset{T \in \mathcal{T}_>}{\arg\min}\, \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T_\sharp^\star \rho \propto \pi$ **in parallel**

We are working on $\mathcal{T}_>^n \subset \mathcal{T}_>$, so
how can we **evaluate the quality of the approximation**?

Convergence criterion

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \widetilde{\pi}} \right] + \log \int \widetilde{\pi}$$

Optimal $T^\star \in \mathcal{T}_>$ and $\int \widetilde{\pi} = 1 \quad \Rightarrow \quad \mathbb{E}_\rho \left[\log \frac{\rho}{(T^\star)^\sharp \widetilde{\pi}} \right] = 0$

But, optimal $\widetilde{T}^\star \in \mathcal{T}_>^n$ or $\int \widetilde{\pi} \neq 1 \quad \Rightarrow \quad \mathbb{E}_\rho \left[\log \frac{\rho}{(\widetilde{T}^\star)^\sharp \widetilde{\pi}} \right] \neq 0$

Convergence criterion

$$T^\star = \arg\min_{T\in\mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) = \arg\min_{T\in\mathcal{T}_>} \mathbb{E}_\rho\left[\log \frac{\rho}{T^\sharp \widetilde{\pi}}\right] + \log \int \widetilde{\pi}$$

Optimal $T^\star \in \mathcal{T}_>$ and $\int \widetilde{\pi} = 1 \quad \Rightarrow \quad \mathbb{E}_\rho\left[\log \frac{\rho}{(T^\star)^\sharp \widetilde{\pi}}\right] = 0$

But, optimal $\widetilde{T}^\star \in \mathcal{T}_>^n$ or $\int \widetilde{\pi} \neq 1 \quad \Rightarrow \quad \mathbb{E}_\rho\left[\log \frac{\rho}{(\widetilde{T}^\star)^\sharp \widetilde{\pi}}\right] \neq 0$

$D_{\mathrm{KL}}(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi) \approx \frac{1}{2}\mathbb{V}\left[\log \frac{\rho}{T^\sharp \widetilde{\pi}}\right]$ as $T \to T^\star$

Pros & cons

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \rho \| \pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T^\star_\sharp \rho \propto \pi$ **in parallel**
+ We can **assess convergence**!

Pros & cons

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \rho \| \pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T_\sharp^\star \rho \propto \pi$ **in parallel**
+ We can **assess convergence**!
− We need to **approximate d functions of up to d variables!**

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

Pros & cons

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}(T_\sharp \rho \| \pi) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi} \right]$$

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $T^\star_\sharp \rho \propto \pi$ **in parallel**
+ We can **assess convergence**!
− We need to **approximate d functions of up to d variables!**

Sources of low-dimensional structure

- Conditional independence
- Low-rank structure
- Smoothness
- Marginal independence

Decomposable transports

Factorization in terms of the I-map $\mathcal{G}$

For $\mathbf{Z} \sim \boldsymbol{\nu}_\pi$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an I-map for $\boldsymbol{\nu}_\pi$

if

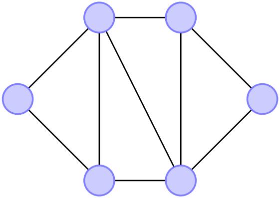
for all A, S, B (partition, S separator), $\mathbf{Z}_A \perp\!\!\!\perp \mathbf{Z}_B | \mathbf{Z}_S$

Factorization in terms of the I-map $\mathcal{G}$

For $\mathbf{Z} \sim \boldsymbol{\nu}_\pi$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an I-map for $\boldsymbol{\nu}_\pi$

if

for all A, S, B (partition, S separator), $\mathbf{Z}_A \perp\!\!\!\perp \mathbf{Z}_B | \mathbf{Z}_S$

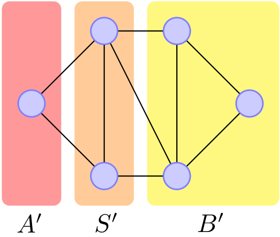


Factorization in terms of the I-map $\mathcal{G}$

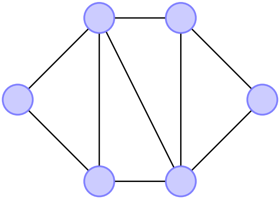
For $\mathbf{Z} \sim \boldsymbol{\nu}_\pi$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an I-map for $\boldsymbol{\nu}_\pi$

if

for all A, S, B (partition, S separator), $\mathbf{Z}_A \perp\!\!\!\perp \mathbf{Z}_B | \mathbf{Z}_S$



Factorization in terms of the I-map $\mathcal{G}$



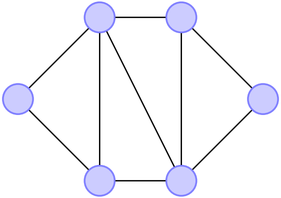
If $\mathcal{G}$ is an I-map of $\boldsymbol{\nu}_\pi$ and $\boldsymbol{\nu}_\pi$ admits a positive continuous density π

then

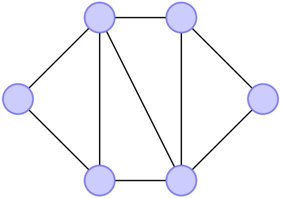
π factorizes with respect to $\mathcal{G}$, i.e. there exist ψ_C s.t.

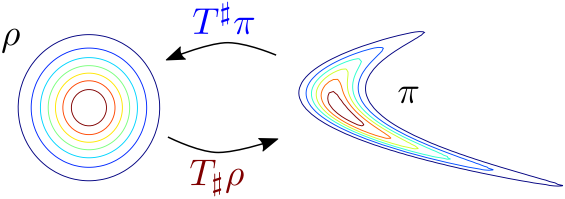
$$\pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \prod_{C \in \mathbf{C}} \psi_C(\boldsymbol{z}_C) \,, \qquad \mathfrak{c} < \infty$$

How to remove conditional dependencies?

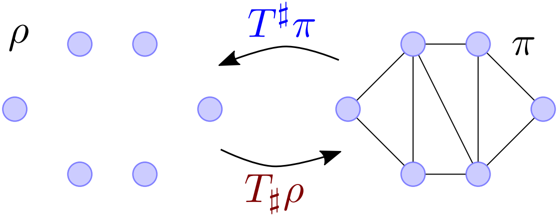


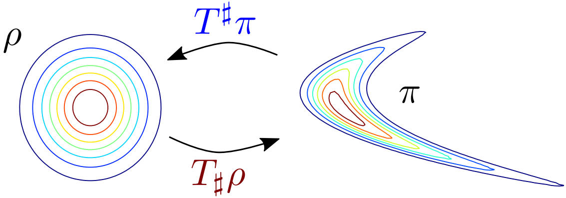
How to remove conditional dependencies?



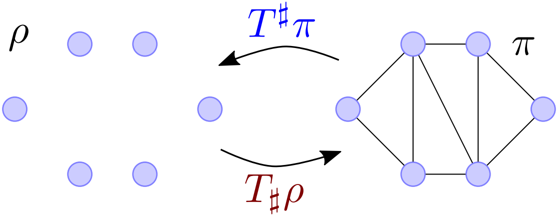


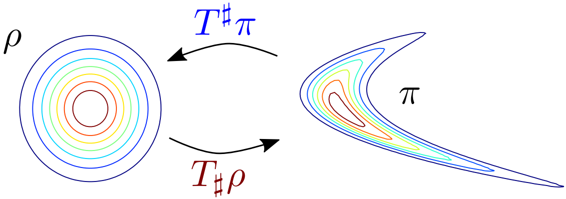
How to remove conditional dependencies?





How to remove conditional dependencies?

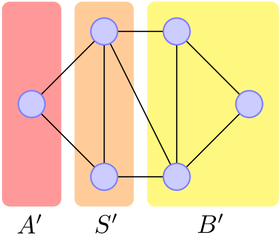




Removing the dependencies **all at once** may be **too expensive!**

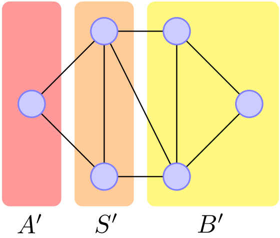
Remove conditional dependencies sequentially

π :



Remove conditional dependencies sequentially

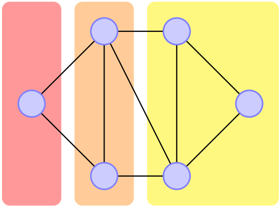
π :



$$\pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A' \cup S'}(\boldsymbol{z}_{A'}, \boldsymbol{z}_{S'}) \psi_{S' \cup B'}(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'})$$

Remove conditional dependencies sequentially

π :



$$\pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A' \cup S'}(\boldsymbol{z}_{A'}, \boldsymbol{z}_{S'}) \psi_{S' \cup B'}(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'})$$

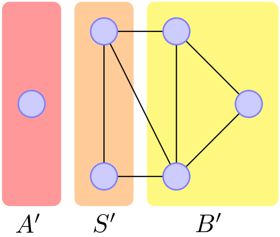
There exists $T_1 = L_1 \circ R_1$ that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\boldsymbol{\nu}_\pi$, where

$$L_1(\boldsymbol{z}) = \begin{bmatrix} L_1^A(\boldsymbol{z}_{S'}, \boldsymbol{z}_{A'}) \\ L_1^S(\boldsymbol{z}_{S'}) \\ \boldsymbol{z}_{B'} \end{bmatrix} \quad \text{and} \quad R_1(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A'} \\ R_1^S(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'}) \\ R_1^B(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'}) \end{bmatrix}$$

and L_1 pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\psi_{A' \cup S'} \cdot \rho_{B'}$, with $\rho_{B'} \sim \mathcal{N}(0, \mathbf{I})$

Remove conditional dependencies sequentially

$L_1^\sharp \pi$:



$$\pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A' \cup S'}(\boldsymbol{z}_{A'}, \boldsymbol{z}_{S'}) \psi_{S' \cup B'}(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'})$$

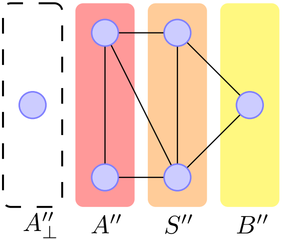
There exists $T_1 = L_1 \circ R_1$ that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\boldsymbol{\nu}_\pi$, where

$$L_1(\boldsymbol{z}) = \begin{bmatrix} L_1^A(\boldsymbol{z}_{S'}, \boldsymbol{z}_{A'}) \\ L_1^S(\boldsymbol{z}_{S'}) \\ \boldsymbol{z}_{B'} \end{bmatrix} \quad \text{and} \quad R_1(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A'} \\ R_1^S(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'}) \\ R_1^B(\boldsymbol{z}_{S'}, \boldsymbol{z}_{B'}) \end{bmatrix}$$

and L_1 pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\psi_{A' \cup S'} \cdot \rho_{B'}$, with $\rho_{B'} \sim \mathcal{N}(0, \mathbf{I})$

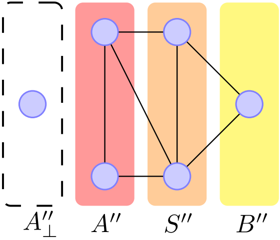
Remove conditional dependencies sequentially

$L_1^\sharp \pi :$



Remove conditional dependencies sequentially

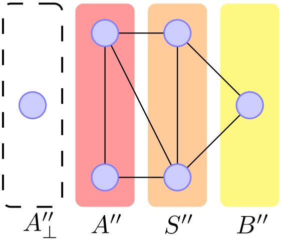
$L_1^\sharp \pi :$



$$L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A''_\perp \cup A'' \cup S''}(\boldsymbol{z}_{A''_\perp}, \boldsymbol{z}_{A''}, \boldsymbol{z}_{S''}) \psi_{S'' \cup B''}(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''})$$

Remove conditional dependencies sequentially

$L_1^\sharp \pi :$



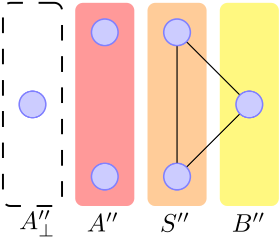
$$L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A''_\perp \cup A'' \cup S''}(\boldsymbol{z}_{A''_\perp}, \boldsymbol{z}_{A''}, \boldsymbol{z}_{S''}) \psi_{S'' \cup B''}(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''})$$

There exists $T_2 = L_2 \circ R_2$ that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $L_1^\sharp \pi$, where

$$L_2(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A''_\perp} \\ L_2^A(\boldsymbol{z}_{S''}, \boldsymbol{z}_{A''}) \\ L_2^S(\boldsymbol{z}_{S''}) \\ \boldsymbol{z}_{B''} \end{bmatrix} \quad \text{and} \quad R_2(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A''_\perp} \\ \boldsymbol{z}_{A''} \\ R_2^S(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''}) \\ R_2^B(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''}) \end{bmatrix}$$

and L_2 pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\psi_{A''_\perp \cup A'' \cup S''} \cdot \rho_{B''}$, with $\rho_{B''} \sim \mathcal{N}(0, \mathbf{I})$

Remove conditional dependencies sequentially

$L_2^\sharp L_1^\sharp \pi :$ 

$$L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A''_\perp \cup A'' \cup S''}(\boldsymbol{z}_{A''_\perp}, \boldsymbol{z}_{A''}, \boldsymbol{z}_{S''}) \psi_{S'' \cup B''}(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''})$$

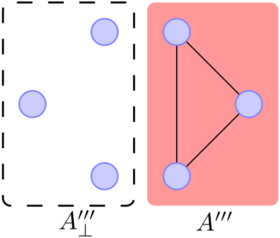
There exists $T_2 = L_2 \circ R_2$ that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $L_1^\sharp \pi$, where

$$L_2(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A''_\perp} \\ L_2^A(\boldsymbol{z}_{S''}, \boldsymbol{z}_{A''}) \\ L_2^S(\boldsymbol{z}_{S''}) \\ \boldsymbol{z}_{B''} \end{bmatrix} \quad \text{and} \quad R_2(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A''_\perp} \\ \boldsymbol{z}_{A''} \\ R_2^S(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''}) \\ R_2^B(\boldsymbol{z}_{S''}, \boldsymbol{z}_{B''}) \end{bmatrix}$$

and L_2 pushes forward $\mathcal{N}(0, \mathbf{I})$ to $\psi_{A''_\perp \cup A'' \cup S''} \cdot \rho_{B''}$, with $\rho_{B''} \sim \mathcal{N}(0, \mathbf{I})$

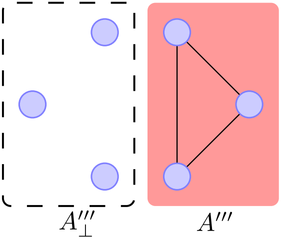
Remove conditional dependencies sequentially

$L_2^\sharp L_1^\sharp \pi :$



Remove conditional dependencies sequentially

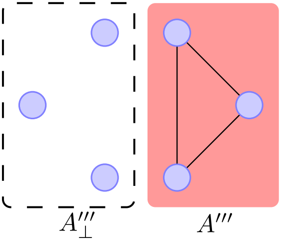
$L_2^\sharp L_1^\sharp \pi :$



$$L_2^\sharp L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A_\perp''' \cup A'''}(\boldsymbol{z}_{A_\perp'''}, \boldsymbol{z}_{A'''})$$

Remove conditional dependencies sequentially

$L_2^\sharp L_1^\sharp \pi :$



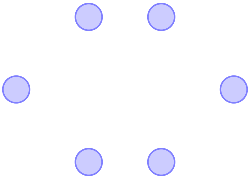
$$L_2^\sharp L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A_\perp''' \cup A'''}(\boldsymbol{z}_{A_\perp'''}, \boldsymbol{z}_{A'''})$$

There exists L_3 that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $L_2^\sharp L_1^\sharp \pi$, where

$$L_3(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_{A_\perp'''} \\ L_3^A(\boldsymbol{z}_{A'''}) \end{bmatrix}$$

Remove conditional dependencies sequentially

$L_3^\sharp L_2^\sharp L_1^\sharp \pi :$



$= \mathcal{N}(0, \mathbf{I})$

$$L_2^\sharp L_1^\sharp \pi(\boldsymbol{z}) = \frac{1}{\mathfrak{c}} \psi_{A_\perp''' \cup A'''}(\boldsymbol{z}_{A_\perp'''}, \boldsymbol{z}_{A'''})$$

There exists L_3 that pushes forward $\mathcal{N}(0, \mathbf{I})$ to $L_2^\sharp L_1^\sharp \pi$, where

$$L_3(\boldsymbol{z}) = \left[\begin{array}{l} \boldsymbol{z}_{A_\perp'''} \\ L_3^{A}(\boldsymbol{z}_{A'''}) \end{array} \right]$$

Stochastic volatility

- Latent log-volatilities modeled with an AR(1) process for $t = 1, \ldots, N$

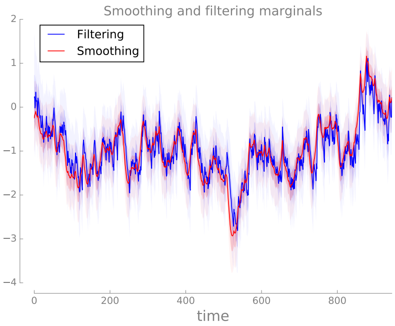
$$X_{t+1} = \mu + \phi(X_t - \mu) + \eta_t \ , \quad \eta_t \sim \mathcal{N}(0,1) \ , \quad X_1 \sim \mathcal{N}\left(0, 1/\left(1-\phi^2\right)\right)$$

$$\mu \sim \mathcal{N}(0,1) \ , \qquad \phi = 2\frac{\exp(\phi^\star)}{1+\exp(\phi^\star)} - 1 \ , \qquad \phi^\star \sim \mathcal{N}(3,1) \ .$$

- Observe the mean return for holding the asset at time t

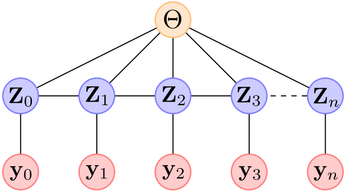
$$Y_t = \varepsilon_t \exp(X_t/2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1)$$

- We want to characterize $\pi \sim \mu, \phi, \mathbf{X}_{1:N} | \mathbf{Y}_{1:N}$



State-space model with hyperparameters

$$
\begin{aligned}
\mathbf{Z}_{k+1} &= G(\mathbf{Z}_k, \mathbf{w}_k, \boldsymbol{\Theta}) \,, \quad \mathbf{w}_k \sim \nu_{w_k}(\boldsymbol{\Theta}) \,, \quad k \in \Lambda = 0, \ldots, n \\
\mathbf{Y}_k &= H(\mathbf{Z}_k, \mathbf{v}_k, \boldsymbol{\Theta}) \,, \quad \mathbf{v}_k \sim \nu_{v_k}(\boldsymbol{\Theta}) \,, \quad k \in \Xi \subset \Lambda \\
\mathbf{Z}_0 &\sim \nu_0(\boldsymbol{\Theta})
\end{aligned}
$$



Full Bayesian solution

$$
\pi\left(\Theta, \mathbf{Z}_\Lambda | \mathbf{y}_\Xi\right) \propto \mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) \pi\left(\Theta, \mathbf{Z}_\Lambda\right)
$$

$$
\mathcal{L}\left(\mathbf{y}_\Xi | \Theta, \mathbf{Z}_\Lambda\right) = \prod_{k \in \Xi} \mathcal{L}\left(\mathbf{y}_k | \Theta, \mathbf{Z}_k\right)
$$

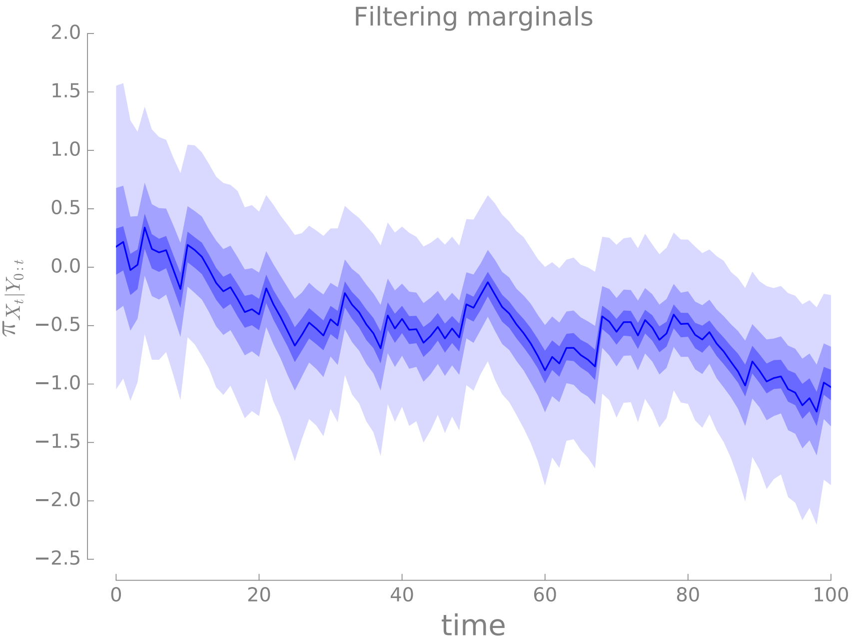
$$
\pi\left(\Theta, \mathbf{Z}_\Lambda\right) = \pi\left(\Theta\right) \pi\left(\mathbf{Z}_0 | \Theta\right) \prod_{k \in \Lambda} \pi\left(\mathbf{Z}_k | \mathbf{Z}_{k-1}, \Theta\right)
$$

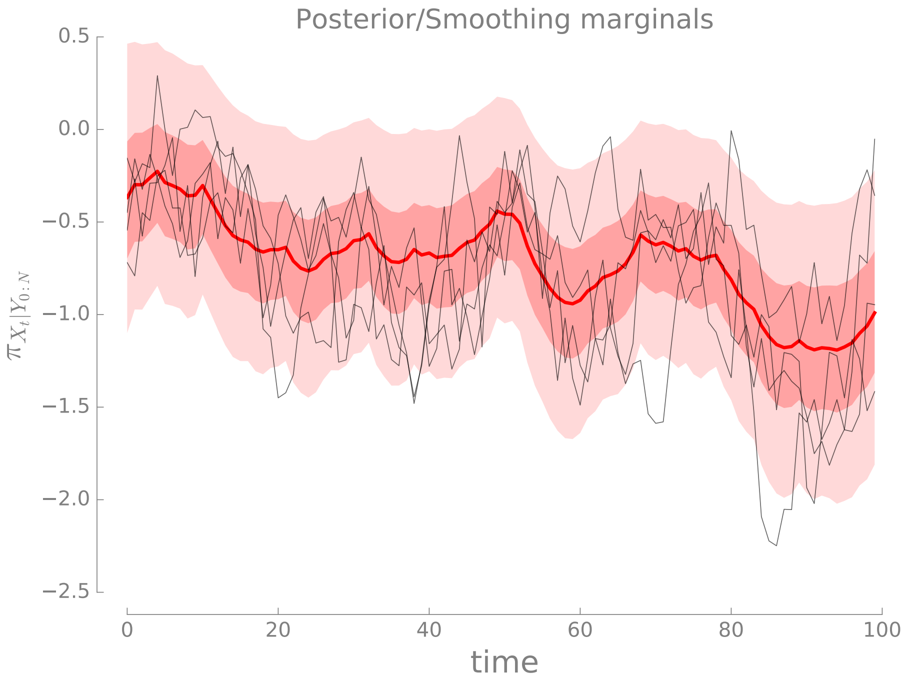
Stochastic volatility

Step-by-step 6–dimensional example

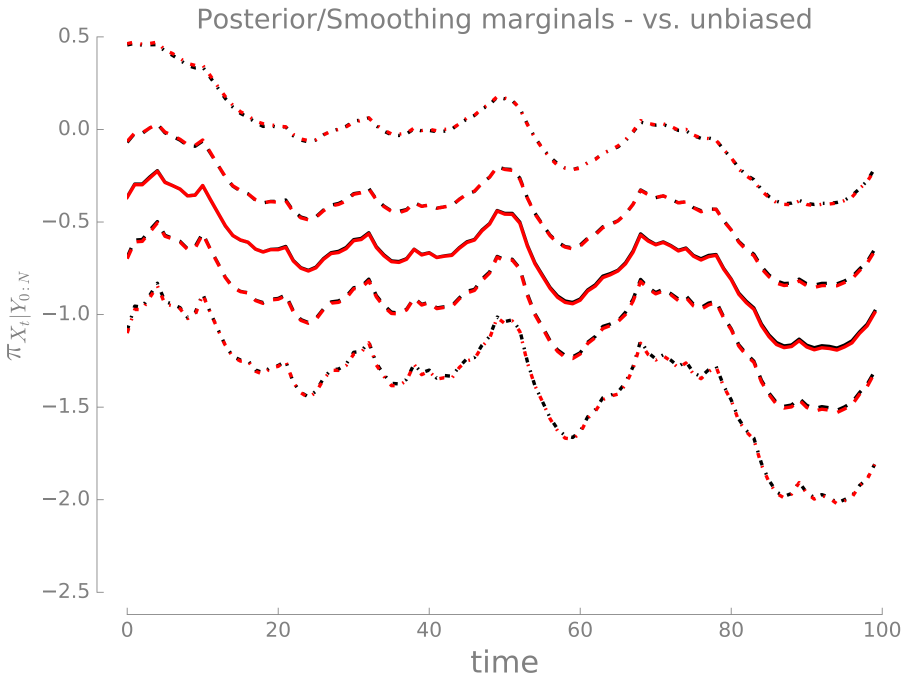
Stochastic volatility

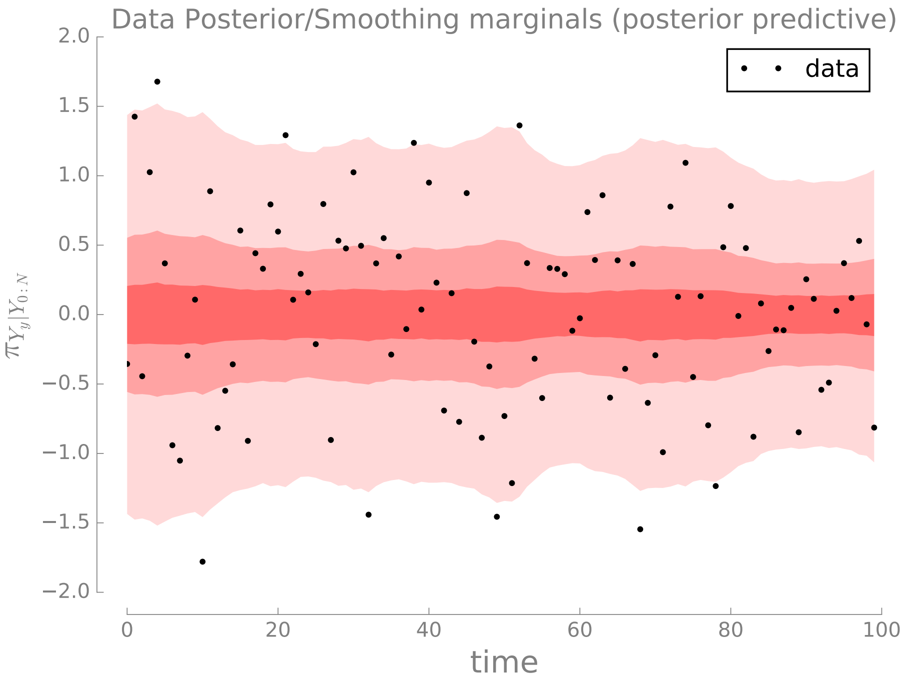
On-line 102–dimensional example


Filtering marginals
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
$\pi_{X_t|Y_{0:t}}$
0
20
40
60
80
100
time

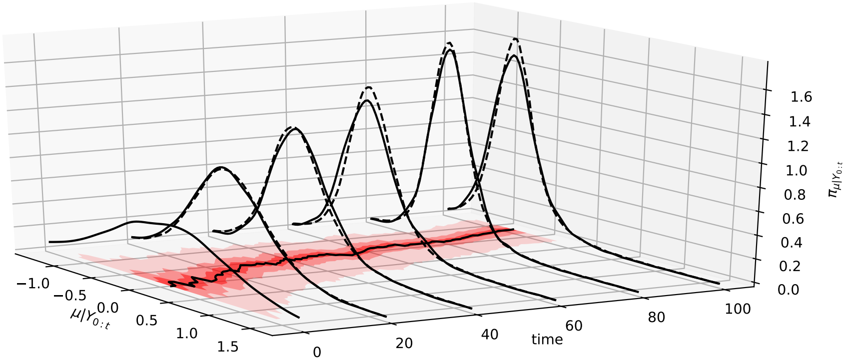


Posterior/Smoothing marginals
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
$\pi_{X_t|Y_{0:N}}$
0
20
40
60
80
100
time


Posterior/Smoothing marginals - vs. unbiased
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
$\pi_{X_t|Y_{0:N}}$
0
20
40
60
80
100
time


Data Posterior/Smoothing marginals (posterior predictive)
data
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
$\pi_{Y_y|Y_{0:N}}$
0
20
40
60
80
100
time

Filtering marginals of the hyperparameter μ



Key contributions

Robust **on-line** algorithm for nonlinear and non-Gaussian **filtering**, **smoothing** and **joint parameter/state estimation** via deterministic couplings and optimization.

Key contributions

Robust **on-line** algorithm for nonlinear and non-Gaussian **filtering**, **smoothing** and **joint parameter/state estimation** via deterministic couplings and optimization.

Ongoing works

- **Rao-blackwellized** version for linear dynamics with nonlinear hyperparameters
- Learning of **non-Gaussian graphical models**
- **Adaptive** construction of maps
- **Low-rank** transports (likelihood informed/active subspaces)

References:

El Moselhy et al. "Bayesian inference with optimal maps"
Parno et al. "Transport map accelerated Markov chain Monte Carlo"
Marzouk et al. "An introduction to sampling via measure transport"
Spantini et al. "Inference via low-dimensional couplings"
Bigoni et al. "On the computation of monotone transports"

Software: `https://transportmaps.mit.edu`

Contacts:

Daniele Bigoni – **dabi@mit.edu**
Alessio Spantini – **spantini@mit.edu**
Youssef Marzouk – **ymarz@mit.edu**

Thanks to:

U.S. DEPARTMENT OF ENERGY

Additional material

Triangular monotone maps

$$\mathcal{T}_{>} = \Big\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \ldots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \Big\}$$

Triangular monotone maps

$$\mathcal{T}_{>} = \Big\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \dots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \Big\}$$

Integrated squared representation – $\varepsilon > 0$

$$T^{(k)}(x_{1:k}) = c_k(x_{1:k-1}) + \int_0^{x_k} \Big(h_k(x_{1:k-1}, t) \Big)^2 + \varepsilon \, dt$$

Triangular monotone maps

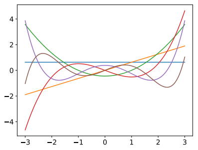
$$\boxed{\mathcal{T}_>^n} = \Big\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \dots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \Big\}$$

Integrated squared representation – $\varepsilon > 0$

$$T^{(k)}(x_{1:k}) = \boxed{c_k(x_{1:k-1})} + \int_0^{x_k} \left(\boxed{h_k(x_{1:k-1}, t)}\right)^2 + \varepsilon \, dt$$

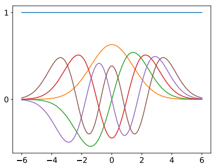
Constant part

$c_k(x_{1:k-1}) = \sum_{\mathbf{i} \in \mathcal{I}_k} \mathbf{a}_\mathbf{i} \Phi_\mathbf{i}(x_{1:k-1})$



Squared part

$h_k(x_{1:k-1}, t) = \sum_{\mathbf{j} \in \mathcal{J}_k} \mathbf{b}_\mathbf{j} \Psi_\mathbf{j}(x_{1:k-1}, t)$



Adaptivity

Adaptivity

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}\left(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi\right) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho\left[\log \frac{\rho}{T^\sharp \pi}\right]$$

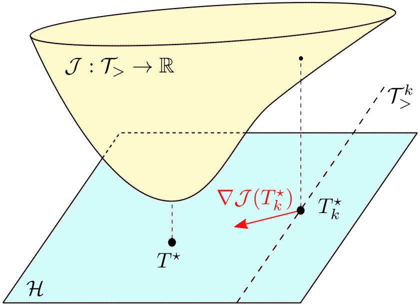
Adaptivity

$$T^\star = \arg\min_{T \in \mathcal{T}_>} D_{\mathrm{KL}}\left(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi\right) = \arg\min_{T \in \mathcal{T}_>} \mathbb{E}_\rho\left[\log \frac{\rho}{T^\sharp \pi}\right]$$

How to find the **best subset** $\mathcal{T}_>^n \subset \mathcal{T}_>$?

Refinement criterion

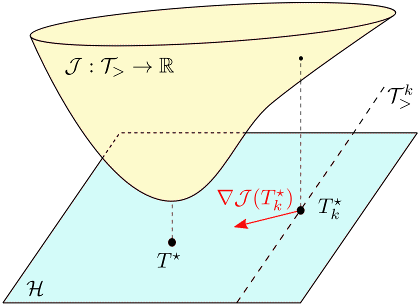
$$T^\star = \arg\min_{T \in \mathcal{T}_>} \underbrace{\mathbb{E}\left[\log \frac{\rho}{T^\sharp \widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$



Refinement criterion

$$T^\star = \arg\min_{T \in \mathcal{T}_>} \underbrace{\mathbb{E}\left[\log \frac{\rho}{T^\sharp \widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$

$$T_0^\star = \arg\min_{T \in \mathcal{T}_>^0} \mathcal{J}(T)$$

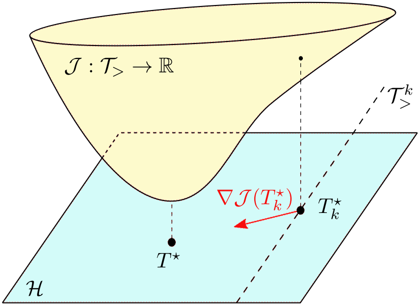


Refinement criterion

$$T^\star = \arg\min_{T\in\mathcal{T}_>} \underbrace{\mathbb{E}\left[\log \frac{\rho}{T^\sharp\widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$

$$T_0^\star = \arg\min_{T\in\mathcal{T}_>^0} \mathcal{J}(T)$$

$$\mathbf{a}_0^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^{n_0}} \mathcal{J}(T[\mathbf{a}])$$



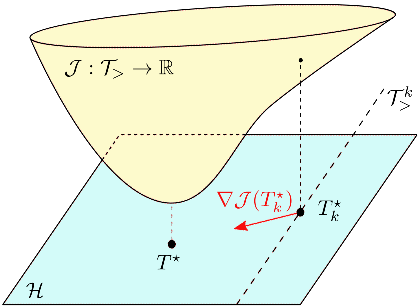
Refinement criterion

$$T^\star = \arg\min_{T\in\mathcal{T}_>} \underbrace{\mathbb{E}\left[\log\frac{\rho}{T^\sharp\widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$

$$T_0^\star = \arg\min_{T\in\mathcal{T}_>^0} \mathcal{J}(T)$$

$$\mathbf{a}_0^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^{n_0}} \mathcal{J}(T[\mathbf{a}])$$

$$\nabla_{\mathbf{a}}\mathcal{J}(T[\mathbf{a}_0^\star]) = 0$$



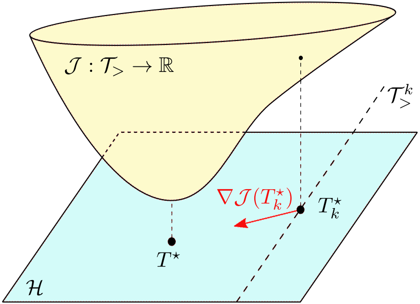
Refinement criterion

$$T^\star = \arg\min_{T\in\mathcal{T}_>} \underbrace{\mathbb{E}\left[\log\frac{\rho}{T^\sharp\widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$

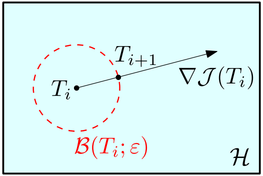
$$T_0^\star = \arg\min_{T\in\mathcal{T}_>^0} \mathcal{J}(T)$$

$$\mathbf{a}_0^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^{n_0}} \mathcal{J}(T[\mathbf{a}])$$

$$\nabla_{\mathbf{a}}\mathcal{J}(T[\mathbf{a}_0^\star]) = 0$$



$$\mathcal{J}(T_{i+1}) < \mathcal{J}(T_i)$$



The **first variation** $\nabla\mathcal{J}\left(T[\mathbf{a}_0^\star]\right) \neq 0$

There exists $\varepsilon > 0$ such that

$$\mathcal{J}\left(T[\mathbf{a}_0^\star] - \varepsilon\nabla\mathcal{J}\left(T[\mathbf{a}_0^\star]\right)\right) < \mathcal{J}\left(T[\mathbf{a}_0^\star]\right)$$

Use the first variation to enrich the approximation space

$$\boxed{\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) = \left(\nabla_{\mathbf{x}} T\right)^{-1} \left(\nabla_{\mathbf{x}} \log \frac{\rho}{T[\mathbf{a}_0^\star]^\sharp \pi} \right)}$$

Use the first variation to enrich the approximation space

$$\boxed{\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) = (\nabla_{\mathbf{x}} T)^{-1} \left(\nabla_{\mathbf{x}} \log \frac{\rho}{T[\mathbf{a}_0^\star]^\sharp \pi} \right)}$$

- $\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) : \mathbb{R}^d \to \mathbb{R}^d$ **is a map** in $\mathcal{H} \supset \mathcal{T}_>$

Use the first variation to enrich the approximation space

$$\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) = \left(\nabla_{\mathbf{x}} T\right)^{-1} \left(\nabla_{\mathbf{x}} \log \frac{\rho}{T[\mathbf{a}_0^\star]^\sharp \pi} \right)$$

- $\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) : \mathbb{R}^d \to \mathbb{R}^d$ **is a map** in $\mathcal{H} \supset \mathcal{T}_>$

Projection on $\mathcal{T}_>^1 \supset \mathcal{T}_>^0$

$$\mathbf{b}_1^\star = \arg\min_{\mathbf{b} \in \mathbb{R}^{n_1}} \left\| U[\mathbf{b}] - \nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) \right\|_{L^2_\rho}$$

Use the first variation to enrich the approximation space

$$\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) = \left(\nabla_{\mathbf{x}} T\right)^{-1} \left(\nabla_{\mathbf{x}} \log \frac{\rho}{T[\mathbf{a}_0^\star]^\sharp \pi} \right)$$

- $\nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) : \mathbb{R}^d \to \mathbb{R}^d$ **is a map** in $\mathcal{H} \supset \mathcal{T}_>$

Projection on $\mathcal{T}_>^1 \supset \mathcal{T}_>^0$

$$\mathbf{b}_1^\star = \arg\min_{\mathbf{b} \in \mathbb{R}^{n_1}} \left\| U[\mathbf{b}] - \nabla \mathcal{J}\left(T[\mathbf{a}_0^\star]\right) \right\|_{L^2_\rho}$$

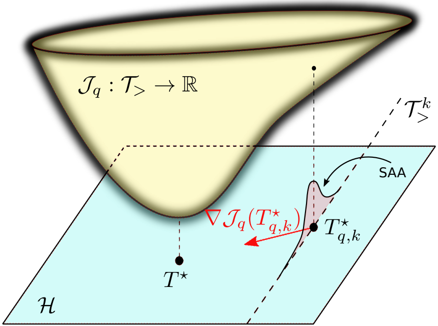
- **No new evaluation** of $\nabla_{\mathbf{x}} \log \pi$ is required
- $U[\mathbf{b}_1^\star]$ informs about **active variables** to be included
- $U[\mathbf{b}_1^\star]$ informs about **active coefficients** to be included

Controlling the sample average accuracy

$$T_k^\star = \underset{T\in\mathcal{T}_>^k}{\arg\min} -\mathbb{E}_\rho\left[\log T^\sharp\pi\right] \approx \underset{T\in\mathcal{T}_>^k}{\arg\min} \overbrace{-\sum_{1\leq i\leq q}\log T^\sharp\pi(\mathbf{x}_i)}^{\mathcal{J}_q(T)} =: T_{q,k}^\star$$

Controlling the sample average accuracy

$$T_k^\star = \underset{T \in \mathcal{T}_>^k}{\arg\min} -\mathbb{E}_\rho \left[\log T^\sharp \pi \right] \approx \underset{T \in \mathcal{T}_>^k}{\arg\min} \overbrace{- \sum_{1 \leq i \leq q} \log T^\sharp \pi(\mathbf{x}_i)}^{\mathcal{J}_q(T)} =: T_{q,k}^\star$$

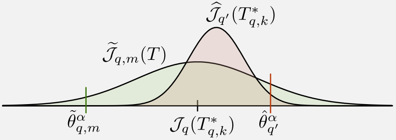


Controlling the sample average accuracy

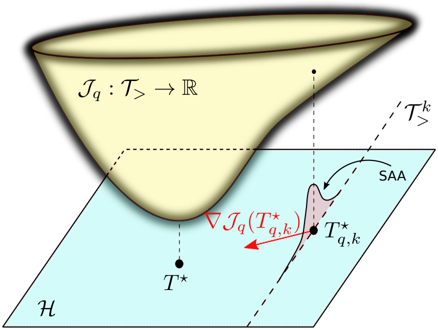
$$T_k^\star = \arg\min_{T\in\mathcal{T}_>^k} -\mathbb{E}_\rho\left[\log T^\sharp\pi\right] \approx \arg\min_{T\in\mathcal{T}_>^k} \overbrace{-\sum_{1\leq i\leq q} \log T^\sharp\pi(\mathbf{x}_i)}^{\mathcal{J}_q(T)} =: T_{q,k}^\star$$

Sample average approximation

$$\tilde{\theta}_{q,m} \leq \mathcal{J}(T_k^\star) \leq \hat{\theta}_{q'}$$

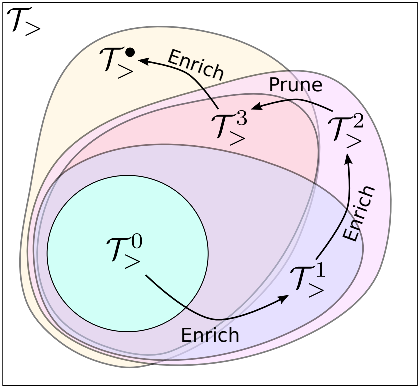


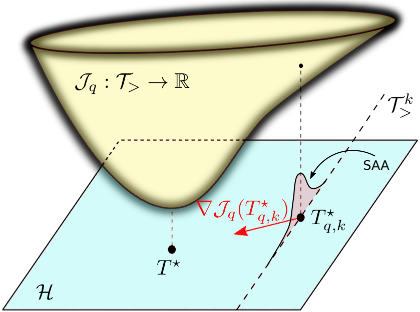
$$\widetilde{\mathcal{J}}_{q,m}(T) = \frac{1}{m}\sum_{i=1}^{m} \min_{T\in\mathcal{T}_>^k} \mathcal{J}_q(T)$$



Adaptivity ingredients

- **Convergence criterion – Variance diagnostic** : $\mathbb{V}\left[\log \frac{\rho}{T^\sharp \pi}\right]$
- **Refinement criterion – First variation** : $\nabla \mathcal{J}\left(T[\mathbf{a}^\star]\right)$
- **Stability criterion – Sample average approximation** : $\tilde{\theta}_{q,m} \leq \mathcal{J}(T_k^\star) \leq \hat{\theta}_{q'}$





The banana distribution – $d = 2$

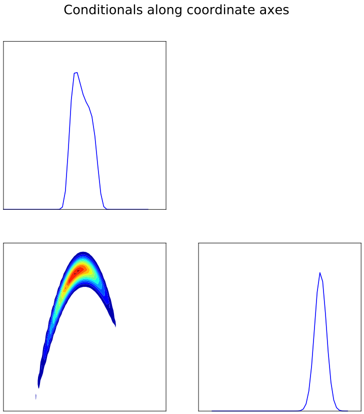
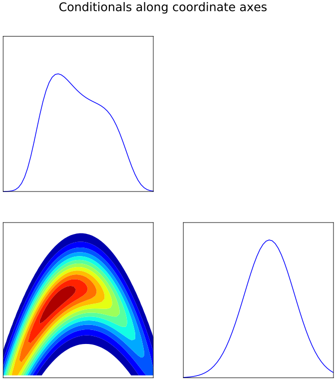


Figure: Target π

The banana distribution – $d = 2$

Iteration 1 – Pullback $T^\sharp \pi$

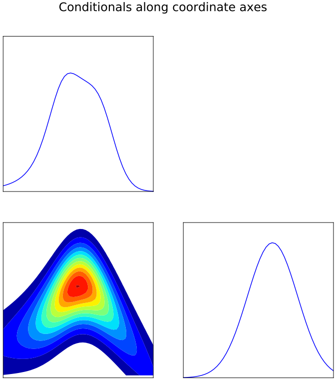


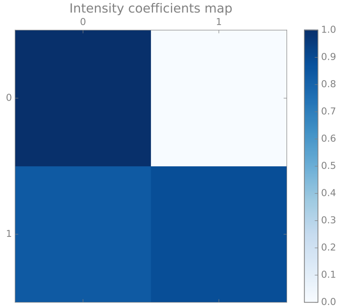
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

The banana distribution – $d = 2$

Iteration 2 – Pullback $T^\sharp \pi$



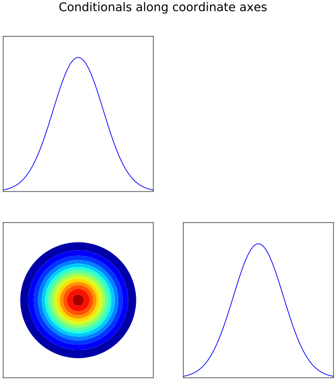


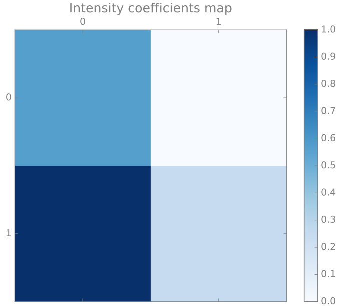
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

The banana distribution – $d = 2$

Iteration 3 – Pullback $T^\sharp \pi$

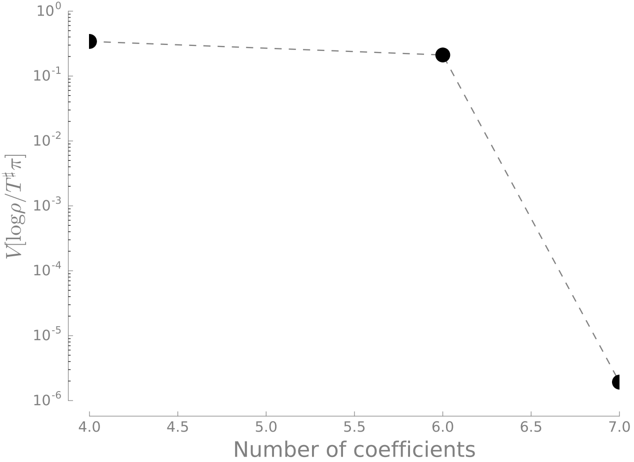




$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

The banana distribution – $d = 2$



Variance diagnostic

Stochastic volatility of financial assets – $d = 32$

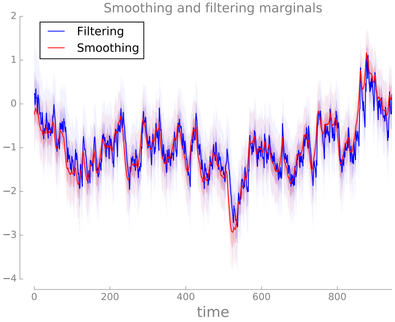
- Latent log-volatilities modeled with an AR(1) process for $t = 1, \dots, N$ $(N = 30)$

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

- Observe the mean return for holding the asset at time t

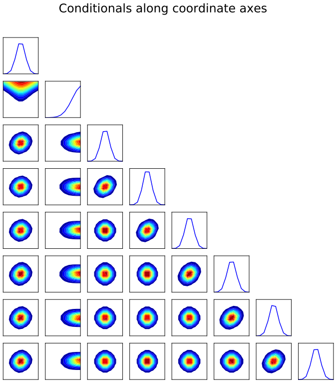
$$Y_t = \varepsilon_t \exp(X_t/2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1)$$

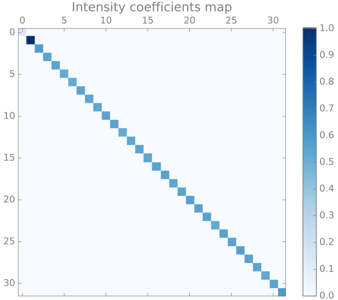
- We want to characterize $\pi \sim \mu, \phi, \mathbf{X}_{1:N} | \mathbf{Y}_{1:N}$



Stochastic volatility of financial assets – $d = 32$

Iteration 1 – Pullback $T^\sharp \pi$



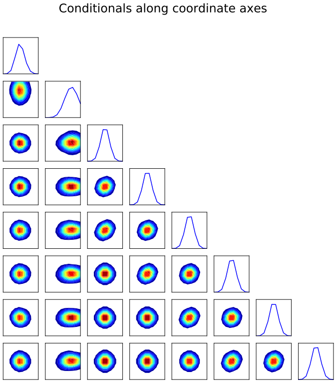


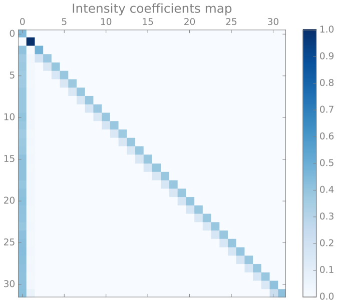
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 2 – Pullback $T^\sharp \pi$



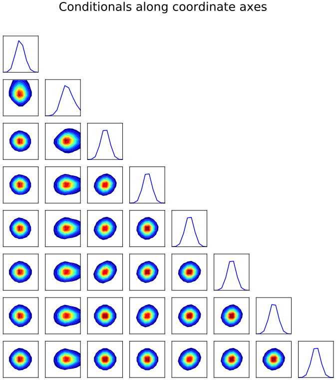


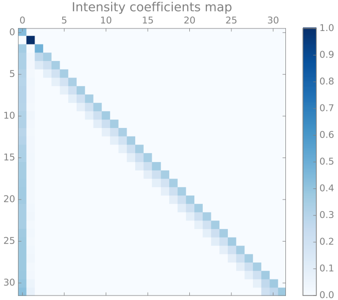
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 3 – Pullback $T^\sharp \pi$



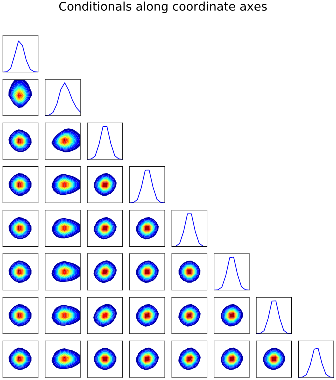


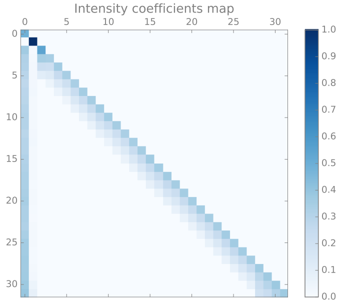
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 4 – Pullback $T^\sharp \pi$



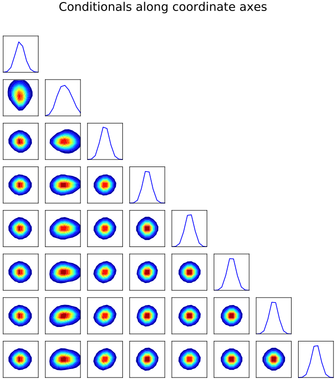


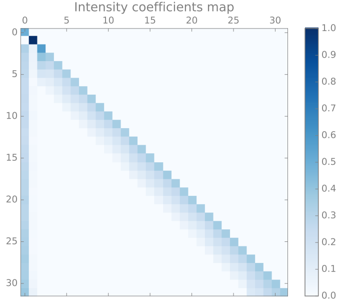
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 5 – Pullback $T^\sharp \pi$



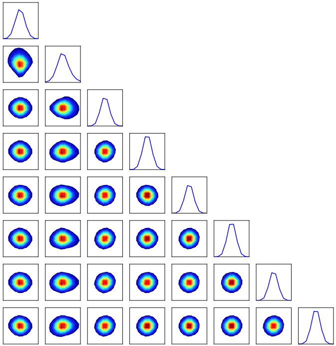


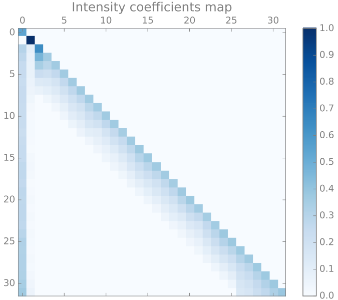
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 6 – Pullback $T^\sharp \pi$



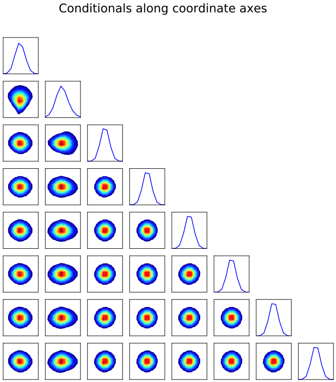


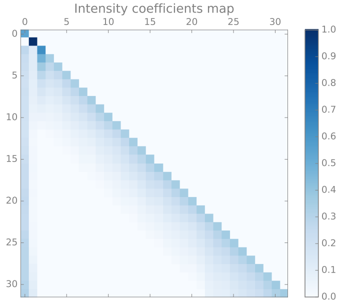
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 7 – Pullback $T^\sharp \pi$



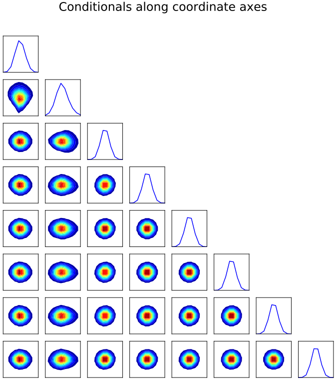


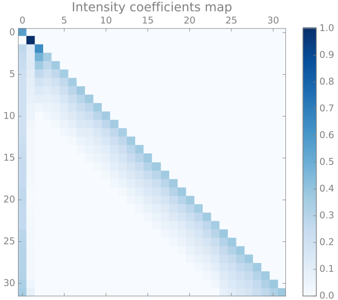
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 7 – Pullback $T^\sharp \pi$



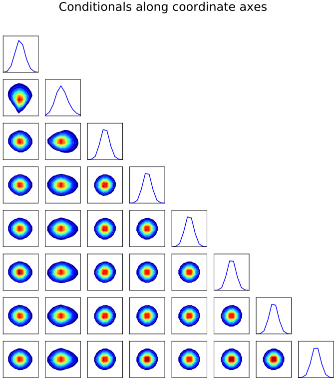


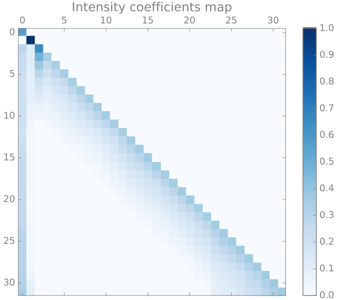
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 9 – Pullback $T^\sharp \pi$



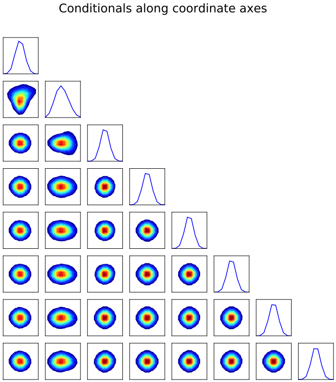


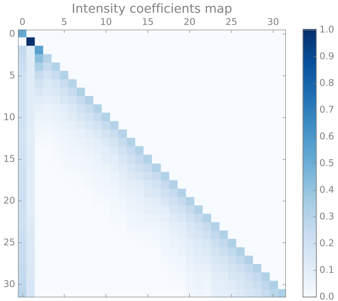
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 10 – Pullback $T^\sharp \pi$



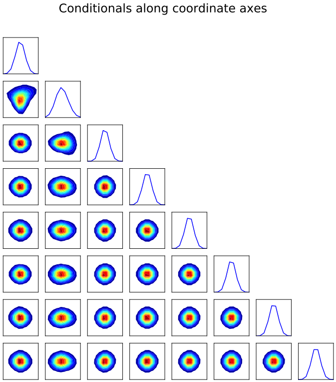


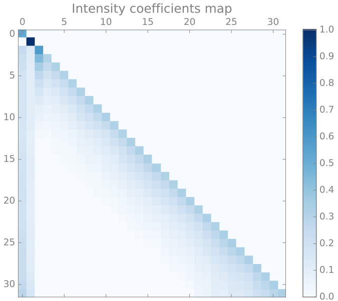
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 11 – Pullback $T^\sharp \pi$



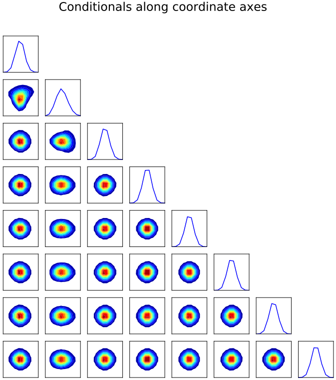


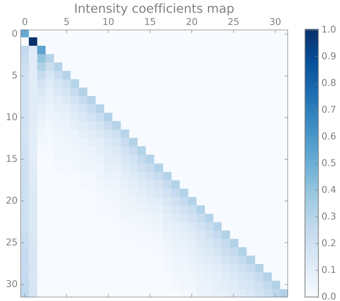
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 12 – Pullback $T^\sharp \pi$



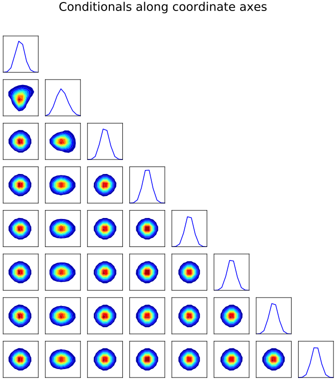


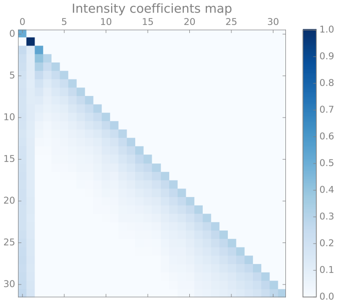
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

Iteration 13 – Pullback $T^\sharp \pi$

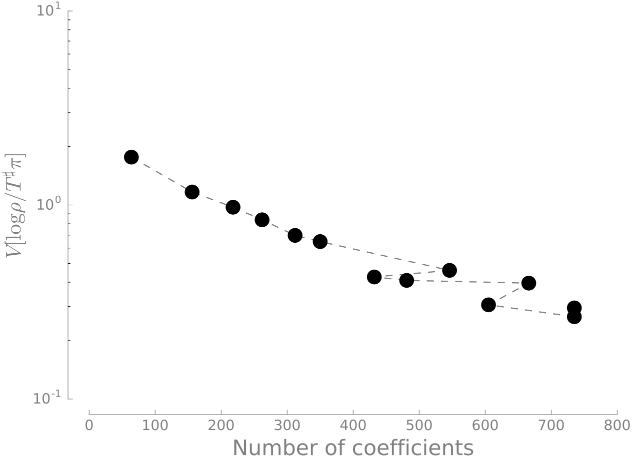




$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$



Variance diagnostic