Scalable inference with Transport Maps

D. Bigoni (**dabi@mit.edu**), A. Spantini, Y.M. Marzouk

Massachusetts Institute of Technology

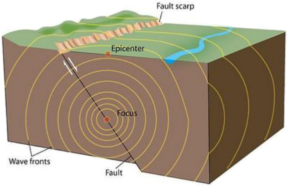
Past and present contributors:

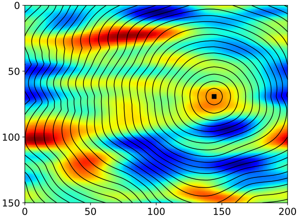
Tarek El Moselhy, Matthew Parno, Xun Huan, Rebecca Morrison,
Ricardo M. Batista, Benjamin Zhang, Zheng Wang

SIAM UQ 2018

Los Angeles – 4/17/2018

Bayesian inference – an oversimplified example

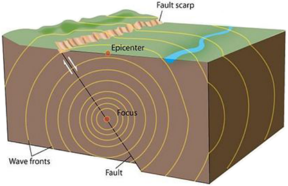


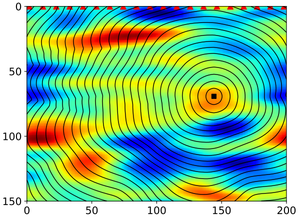


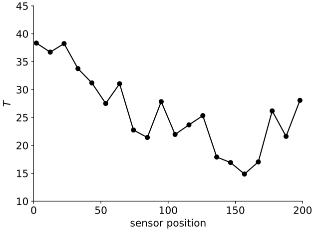
Mathematical model

$$|\nabla \underbrace{G(\mathbf{x})}_{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}^{-1}$$

Bayesian inference – an oversimplified example







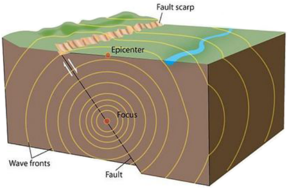
Mathematical model

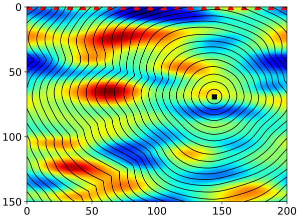
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

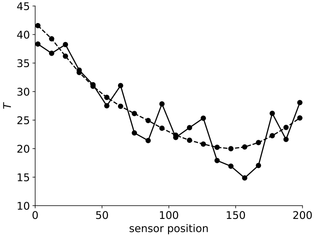
Observational model

$$\overbrace{\mathbf{d}}^{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

Bayesian inference – an oversimplified example







Mathematical model

travel time

$$|\nabla G(\mathbf{x})| = v(\mathbf{x})^{-1}$$

velocity field

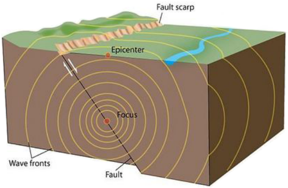
Observational model

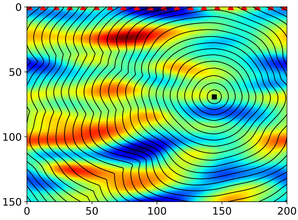
data

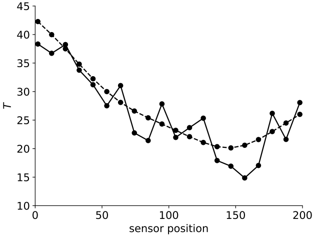
$$\mathbf{d} = \mathbf{G}(\mathbf{v}) + \varepsilon$$

noise

Bayesian inference – an oversimplified example







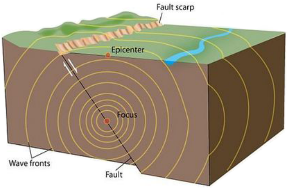
Mathematical model

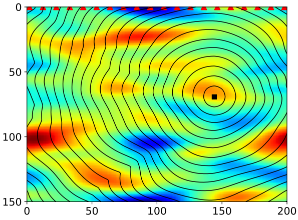
$$|\nabla \underbrace{G(\mathbf{x})}_{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}^{-1}$$

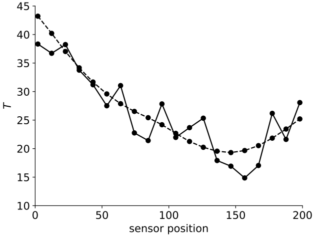
Observational model

$$\underbrace{\mathbf{d}}_{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

Bayesian inference – an oversimplified example







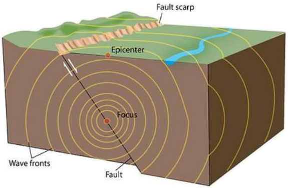
Mathematical model

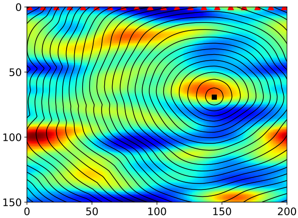
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

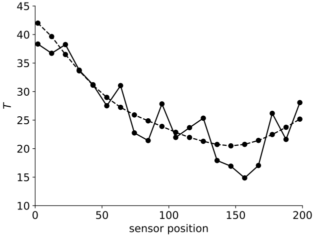
Observational model

$$\overbrace{\mathbf{d}}^{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

Bayesian inference – an oversimplified example







Mathematical model

travel time

$$|\nabla G(\mathbf{x})| = v(\mathbf{x})^{-1}$$

velocity field

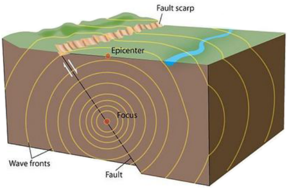
Observational model

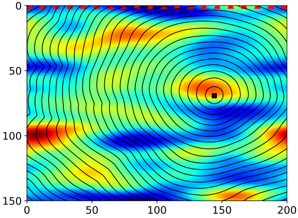
data

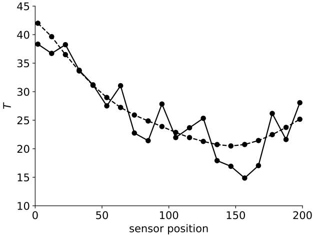
$$\mathbf{d} = \mathbf{G}(\mathbf{v}) + \varepsilon$$

noise

Bayesian inference – an oversimplified example







Mathematical model

$$|\nabla \underbrace{G(\mathbf{x})}_{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

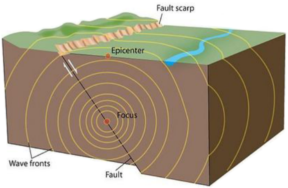
Observational model

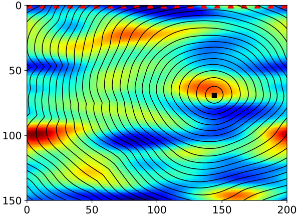
$$\overbrace{\mathbf{d}}^{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

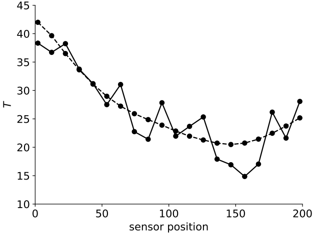
Bayesian inference model

$$\underbrace{\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})}_{\text{posterior}} \propto \overbrace{\mathcal{L}_{\mathbf{d}}(\mathbf{v})}^{\text{likelihood}} \underbrace{\pi_{\text{pr}}(\mathbf{v})}_{\text{prior}} = \pi_{\varepsilon}(\mathbf{d} - \mathbf{G}(\mathbf{v}))\pi_{\text{pr}}(\mathbf{v})$$

Bayesian inference – an oversimplified example







Bayesian inference model

$$\underbrace{\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})}_{\text{posterior}} \propto \overbrace{\mathcal{L}_{\mathbf{d}}(\mathbf{v})}^{\text{likelihood}} \underbrace{\pi_{\text{pr}}(\mathbf{v})}_{\text{prior}} = \pi_{\boldsymbol{\varepsilon}}(\mathbf{d} - \mathbf{G}(\mathbf{v}))\pi_{\text{pr}}(\mathbf{v})$$

Decisions under uncertainty

$$\min_{\delta} \int L(\mathbf{v}, \delta)\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})d\mathbf{v}$$

Goal: characterize $\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})$, i.e.

- construct approximations

$$\int f(\mathbf{v})\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})d\mathbf{v} \approx \int f(\mathbf{v})\tilde{\pi}_{\text{pos}}(\mathbf{v}|\mathbf{d})d\mathbf{v} \approx \sum_{i=1}^{n} f(\mathbf{v}^{(i)})\mathbf{w}^{(i)}$$

- control the error between $\pi_{\text{pos}}(\mathbf{v}|\mathbf{d})$ and $\tilde{\pi}_{\text{pos}}(\mathbf{v}|\mathbf{d})$

Difficulties:

- $\mathbf{v} \in \mathbb{R}^d$ where $d \gg 1$
- The model $\mathbf{G}(\mathbf{v})$ is non-linear
- Evaluation of the model $\mathbf{G}(\mathbf{v})$ is expensive

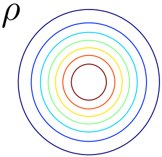
Outline

Transport maps

Adaptivity

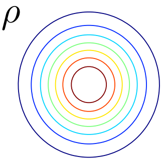
Transport maps – Pullbacks [PB] and Pushforwards [PF]

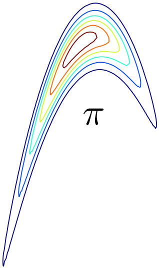
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$



Transport maps – 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 $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$



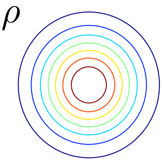


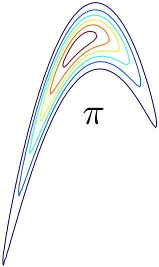
Transport maps – Pullbacks [PB] and Pushforwards [PF]

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





Transport maps – 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

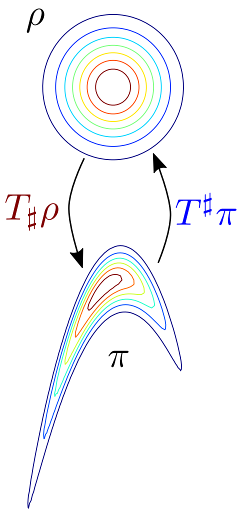
  PF $T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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

- We want T such that

  PF $T_\sharp \rho = \pi$

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



Transport maps – 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

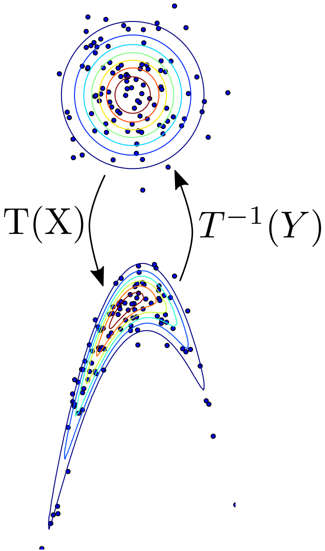
  PF $T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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

- We want T such that

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

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



Transport maps – 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

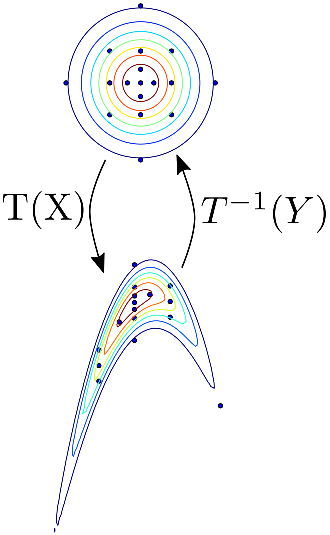
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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



Transport maps – Pullbacks [PB] and Pushforwards [PF]

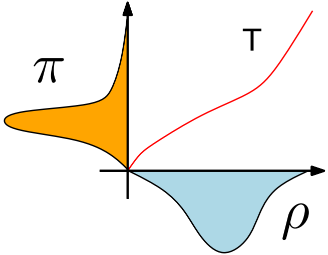
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

  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

$\forall\ \boldsymbol{\nu}_\rho, \boldsymbol{\nu}_\pi$ Lebesgue absolutely continuous
$\exists$ a **triangular monotone** map s.t. $T_\sharp \rho = \pi$

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

Triangular monotone maps

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

Triangular monotone maps

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

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

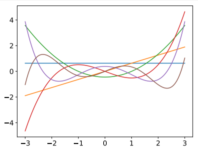
$$\boxed{\mathcal{T}_{>}^{n}} = \Big\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \ldots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \Big\}$$

Integrated squared representation – $\varepsilon > 0$

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

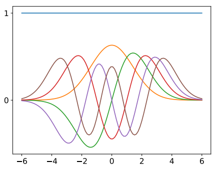
Constant part

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



Squared part

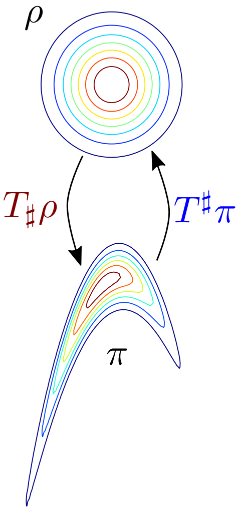
$$h_k(x_{1:k-1}, t) = \sum_{\mathbf{j} \in \mathcal{J}_k} \mathbf{b}_\mathbf{j} \Psi_\mathbf{j}(x_{1:k-1}, t)$$



Knothe-Rosenblatt rearrangement

$\forall\ \boldsymbol{\nu}_\rho, \boldsymbol{\nu}_\pi$ Lebesgue absolutely continuous
$\exists$ a **triangular monotone** map s.t. $T_\sharp \rho = \pi$

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

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

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

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

Convergence criterion – Variance diagnostic

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

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 $\int \widetilde{\pi} \neq 1 \quad \Rightarrow \quad \mathbb{E}_\rho \left[\log \frac{\rho}{\left(\widetilde{T}^\star\right)^\sharp \widetilde{\pi}} \right] \neq 0$

Convergence criterion – Variance diagnostic

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

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 $\int \widetilde{\pi} \neq 1 \quad \Rightarrow \quad \mathbb{E}_\rho \left[\log \frac{\rho}{\left(\widetilde{T}^\star\right)^\sharp \widetilde{\pi}} \right] \neq 0$

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

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 $T_\sharp^\star \rho \propto \pi$ **in parallel**
+ We can **assess convergence**!
+ The map can be used as a **preconditioner** for other unbiased methods

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**!
- \+ The map can be used as a **preconditioner** for other unbiased methods
- − 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**!
+ The map can be used as a **preconditioner** for other unbiased methods

− We need to **approximate d functions of up to d variables!**

Sources of low-dimensional structure

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

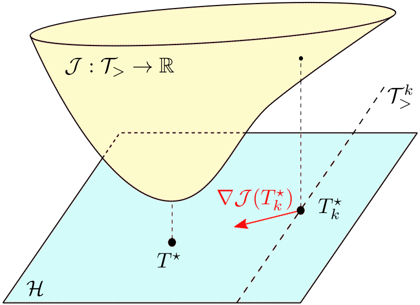
Adaptivity

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

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

Refinement criterion

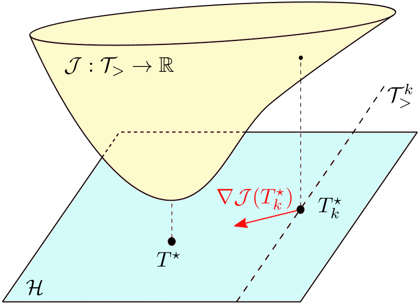
$$T^\star = \arg\min_{T \in \mathcal{T}_>} \underbrace{\mathbb{E}\left[\log \frac{\rho}{T^\sharp \widetilde{\pi}}\right]}_{\mathcal{J}(T)}$$



Refinement criterion

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

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

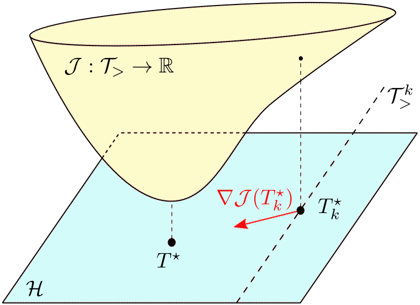


Refinement criterion

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

$$T_k^\star = \underset{T\in\mathcal{T}_>^k}{\arg\min}\, \mathcal{J}(T)$$

$$\mathbf{a}_k^\star = \underset{\mathbf{a}\in\mathbb{R}^{n_k}}{\arg\min}\, \mathcal{J}(T[\mathbf{a}])$$



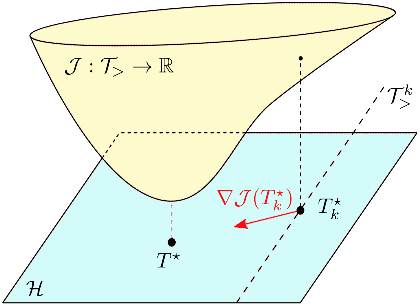
Refinement criterion

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

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

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

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



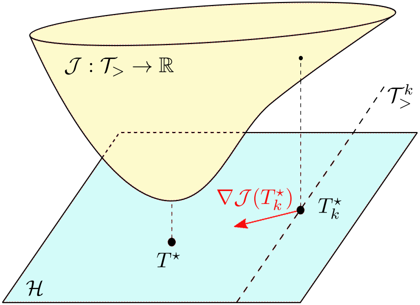
Refinement criterion

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

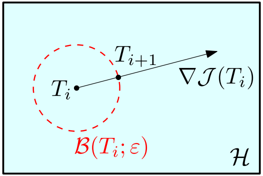
$$T_k^\star = \arg\min_{T \in \mathcal{T}_>^k} \mathcal{J}(T)$$

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

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



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



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

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

Use the first variation to enrich the approximation space

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

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

Use the first variation to enrich the approximation space

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

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

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

$$\mathbf{b}_{k+1}^\star = \underset{\mathbf{b} \in \mathbb{R}^{n_{k+1}}}{\arg\min} \left\| U[\mathbf{b}] - \nabla \mathcal{J}\left(T[\mathbf{a}_k^\star]\right) \right\|_{L^2_\rho}$$

Use the first variation to enrich the approximation space

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

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

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

$$\mathbf{b}_{k+1}^\star = \arg\min_{\mathbf{b} \in \mathbb{R}^{n_{k+1}}} \left\| U[\mathbf{b}] - \nabla \mathcal{J}\left(T[\mathbf{a}_k^\star]\right) \right\|_{L^2_\rho}$$

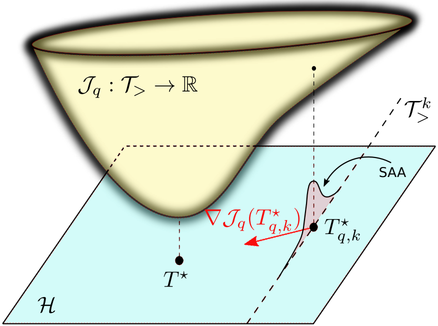
- **No new evaluation** of $\nabla_{\mathbf{x}} \log \pi$ is required
- $U[\mathbf{b}_{k+1}^\star]$ informs about **active variables** to be included
- $U[\mathbf{b}_{k+1}^\star]$ informs about **active coefficients** to be included

Controlling the sample average accuracy

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

Controlling the sample average accuracy

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

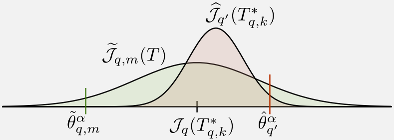


Controlling the sample average accuracy

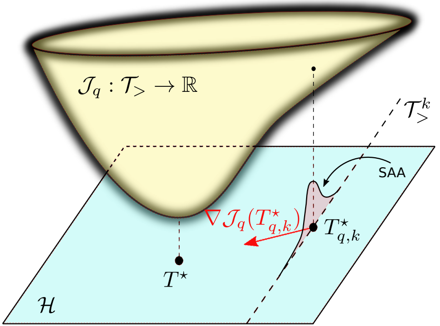
$$T_k^\star = \underset{T \in \mathcal{T}_>^k}{\arg\min} -\mathbb{E}_\rho \left[\log T^\sharp \pi \right] \approx \underset{T \in \mathcal{T}_>^k}{\arg\min} \overbrace{- \sum_{1 \le i \le q} \log T^\sharp \pi(\mathbf{x}_i)\ w_i}^{\mathcal{J}_q(T)} =: T_{q,k}^\star$$

Sample average approximation

$$\tilde\theta_{q,m} \le \mathcal{J}(T_k^\star) \le \hat\theta_{q'}$$

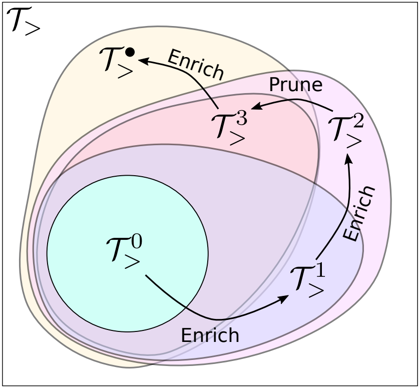


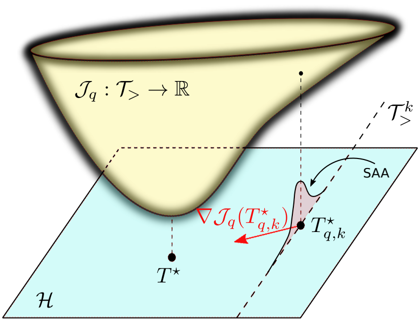
$$\widetilde{J}_{q,m}(T) = \frac{1}{m} \sum_{i=1}^{m} \min_{T \in \mathcal{T}_>^k} \mathcal{J}_q(T)$$



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





Stochastic volatility of financial assets – $d = 32$

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

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

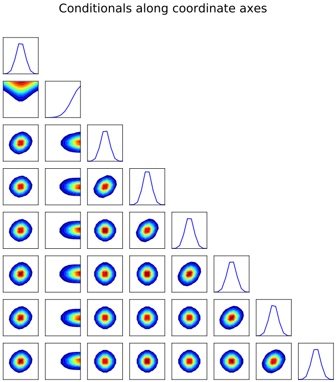
- Observe the mean return for holding the asset at time t

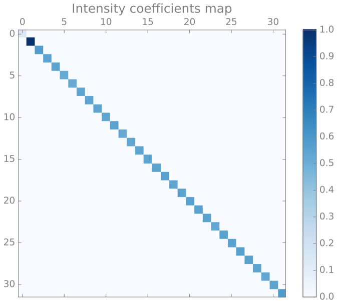
$$Y_t = \varepsilon_t \exp(X_t/2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1)$$

- 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 $T^{\sharp}\pi$



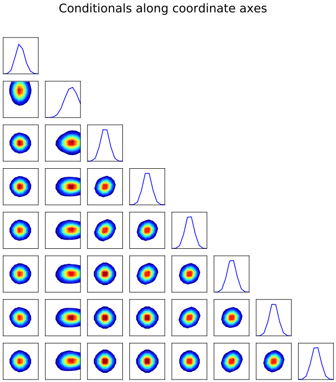


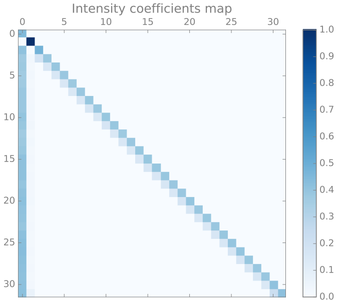
$\nabla_{\mathbf{x}} T$

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 $T^{\sharp}\pi$



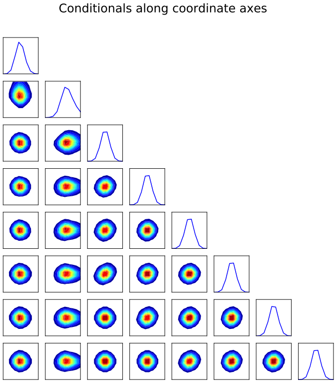


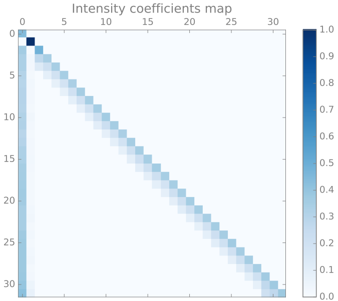
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



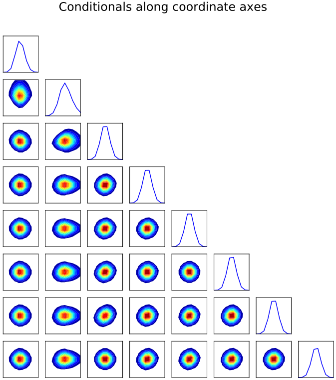


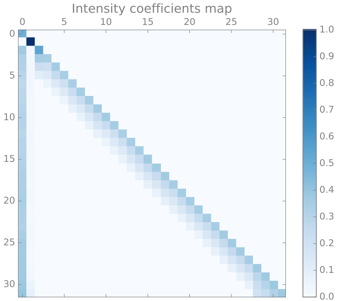
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



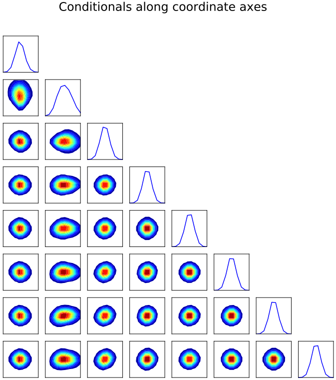


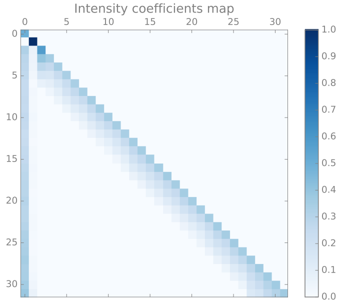
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



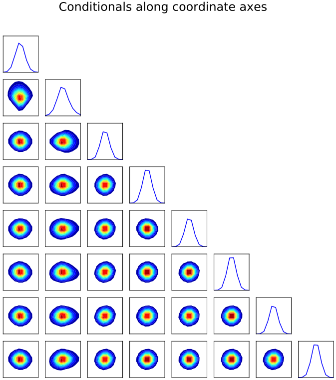


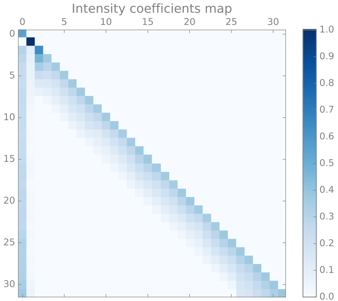
$\nabla_{\mathbf{x}} T$

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 $T^{\sharp}\pi$



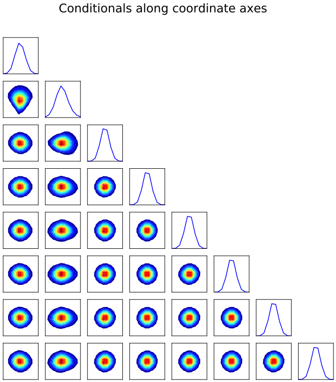


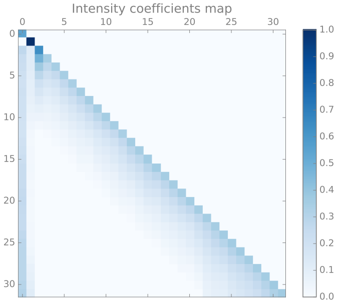
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



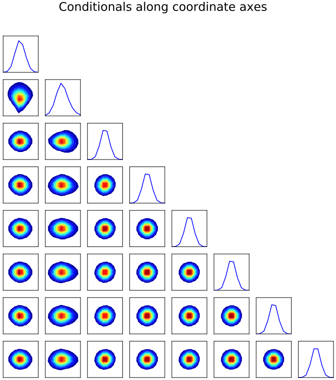


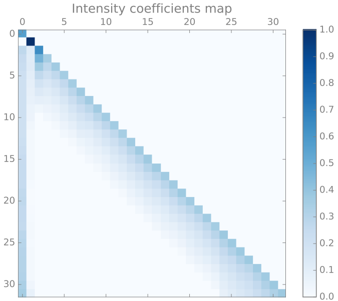
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



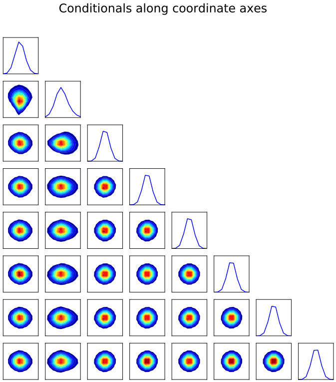


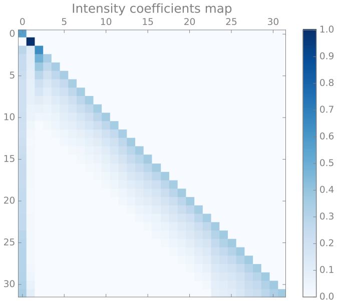
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$





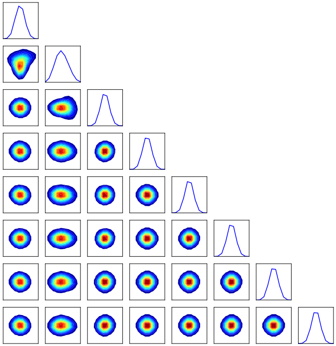
$\nabla_{\mathbf{x}} T$

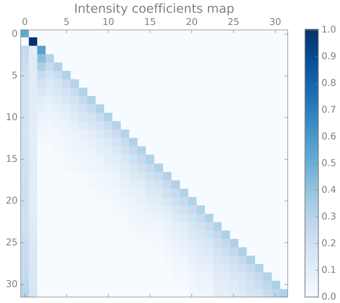
Reminder: $T^\sharp \pi \approx \rho$, where ρ is the density of $\mathcal{N}(0, \mathbf{I})$

Stochastic volatility of financial assets – $d = 32$

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

Conditionals along coordinate axes



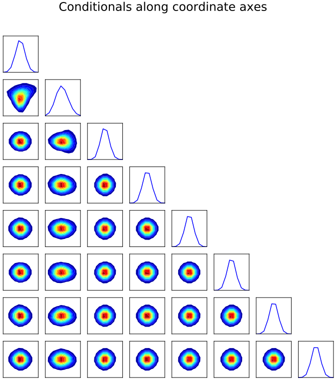


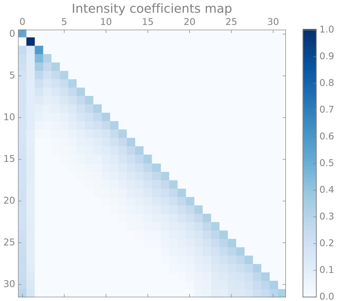
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



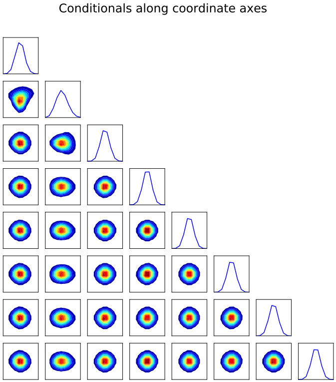


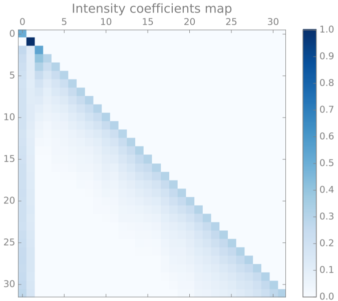
$\nabla_{\mathbf{x}} T$

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 $T^\sharp \pi$



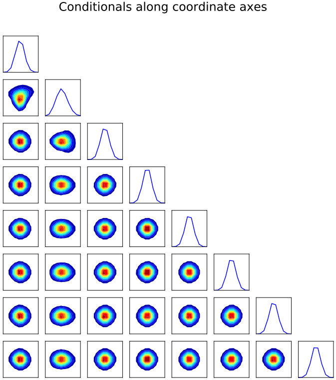


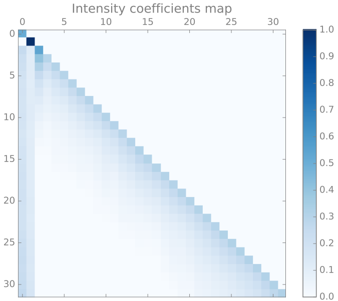
$\nabla_{\mathbf{x}} T$

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 $T^{\sharp}\pi$

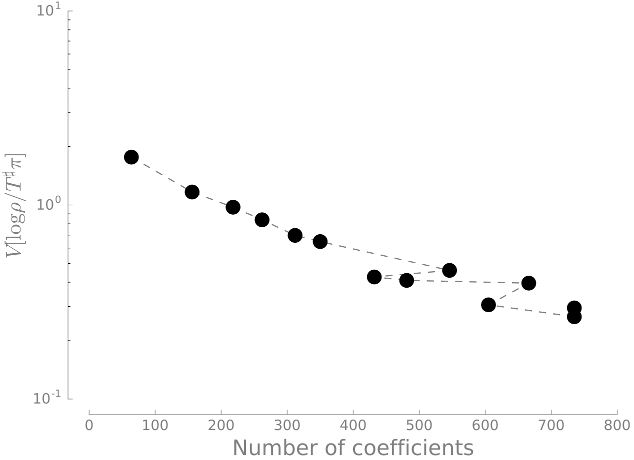




$\nabla_{\mathbf{x}} T$

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

Key contributions

Algorithms for characterizing probability measures
via **deterministic couplings** and **optimization**,
exploiting **smoothness** and **marginal independence**

Contact: Daniele Bigoni – **dabi@mit.edu**

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

References:

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

Thanks to: U.S. DEPARTMENT OF ENERGY, DARPA