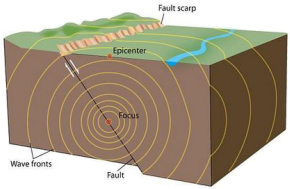
Adaptive construction of Transport Maps for efficient sampling

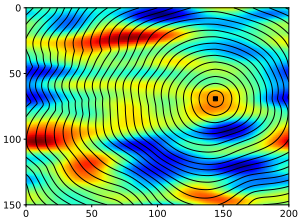
D. Bigoni (**dabi@mit.edu**), A. Spantini, Y.M. Marzouk
Massachusetts Institute of Technology

Past and present contributors:
Tarek El Moselhy, Matthew Parno, Xun Huan, Rebecca Morrison,
Ricardo M. Batista, Benjamin Zhang, Zheng Wang

MCQMC 2018
Rennes – July 5, 2018

Bayesian inference – an oversimplified example





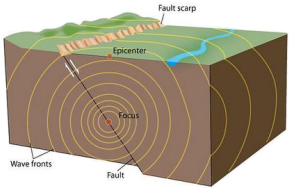
Mathematical model

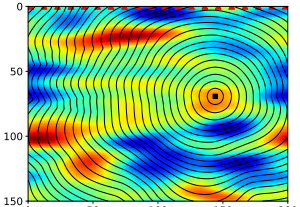
travel time

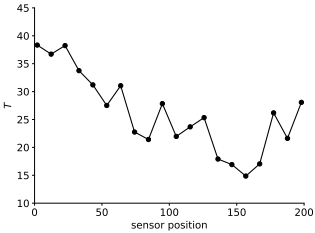
$$|\nabla G(\mathbf{x})| = v(\mathbf{x})^{-1}$$

velocity field

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

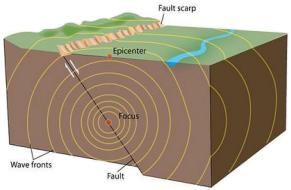
velocity field

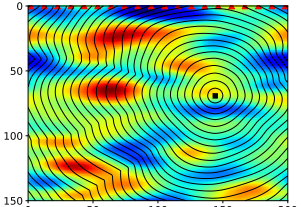
Observational model

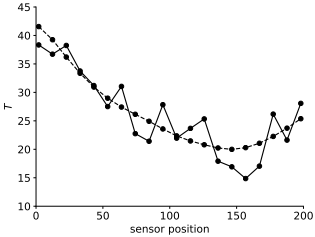
data

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

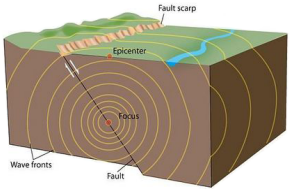
$$|\nabla \overbrace{G(\mathbf{x})}| = \underbrace{v(\mathbf{x})}^{-1}$$

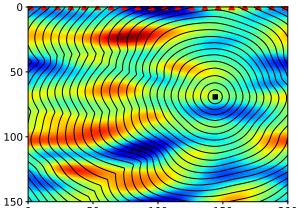
velocity field

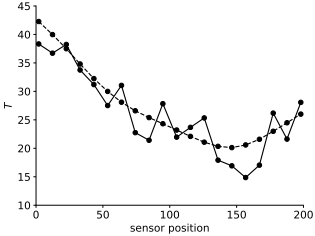
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

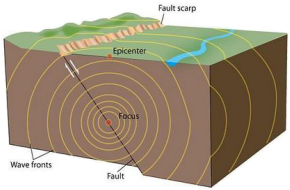
$$|\nabla \overbrace{G(\mathbf{x})}| = \underbrace{v(\mathbf{x})}^{-1}$$

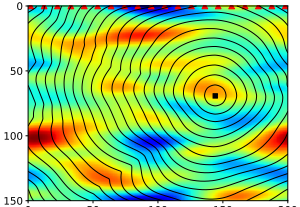
velocity field

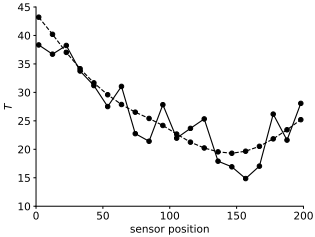
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

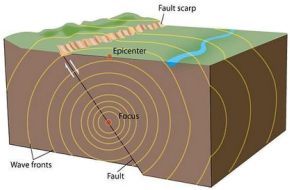
$$|\nabla \overbrace{G(\mathbf{x})}| = \underbrace{v(\mathbf{x})}^{-1}$$

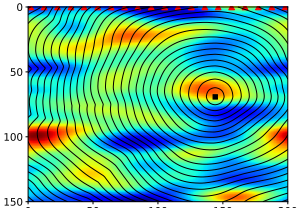
velocity field

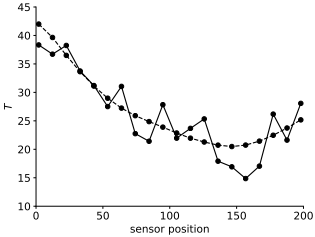
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

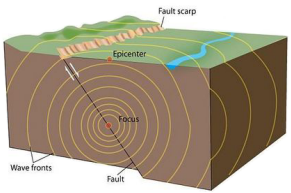
$$|\nabla \overbrace{G(\mathbf{x})}| = \underbrace{v(\mathbf{x})}^{-1}$$

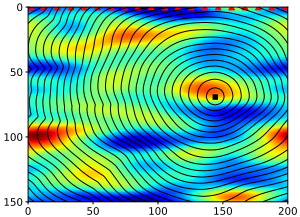
velocity field

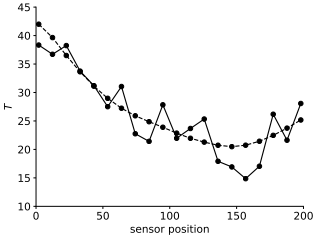
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

velocity field

Observational model

data

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

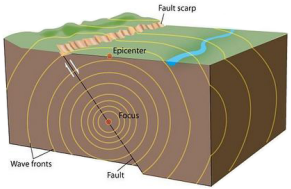
noise

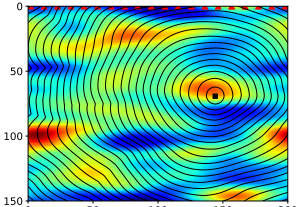
Bayesian inference model

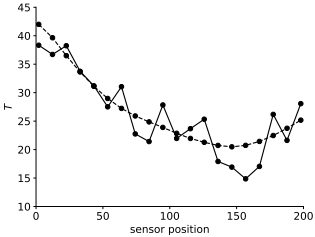
likelihood

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

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

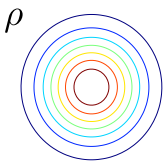
Transport maps

Adaptivity

Off-line learning / On-line inference

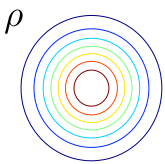
Transport maps – Pullbacks [PB] and Pushforwards [PF]

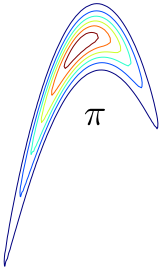
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$



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



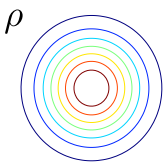


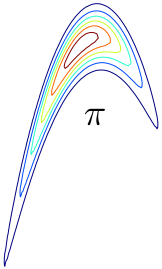
Transport maps – Pullbacks [PB] and Pushforwards [PF]

- 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

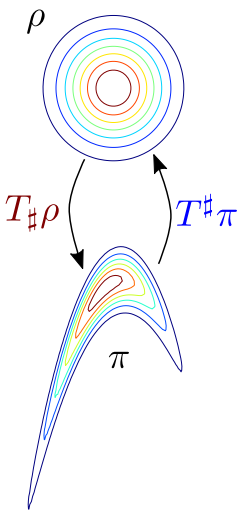
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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

- We want T such that

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

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



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

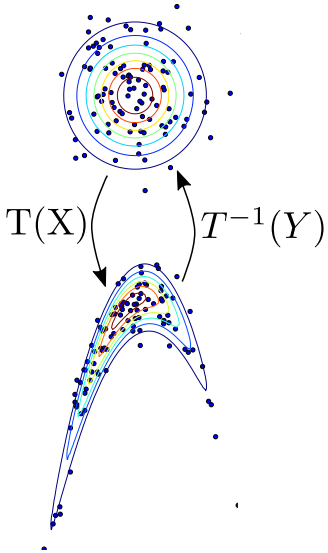
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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

- We want T such that

  PF 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

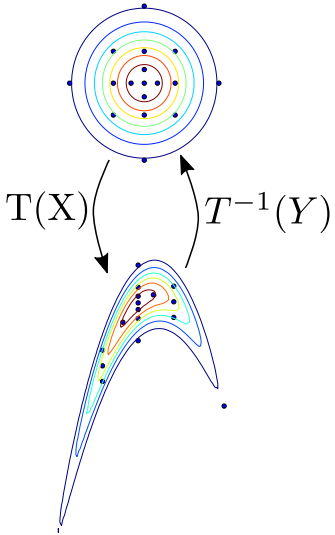
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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

- We want T such that

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

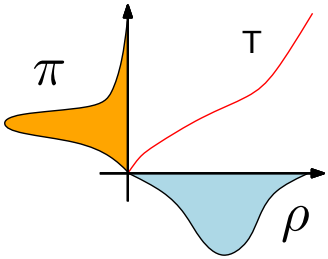
- We want T such that

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

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

Knothe-Rosenblatt rearrangement

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



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

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

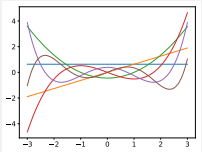
$$\boxed{\mathcal{T}_{>}^{n}} = \left\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \ldots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \right\}$$

Integrated squared representation – $\varepsilon > 0$

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

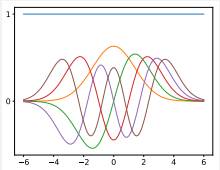
Constant part

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



Squared part

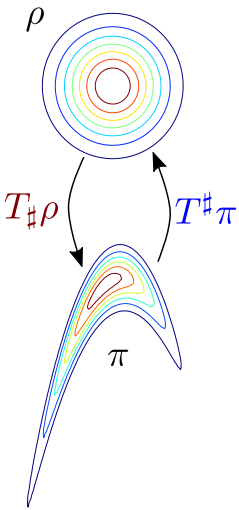
$h_k(x_{1:k-1}, t) = \sum_{\mathbf{j} \in \mathcal{J}_k} \mathbf{b_j} \Psi_{\mathbf{j}}(x_{1:k-1}, t)$



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

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

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 $T^\star_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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 $T^\star_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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

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

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 – Variance diagnostic

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

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

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

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

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 $T^\star_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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 $T_\sharp^\star \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**
+ We can **assess convergence**!
+ The map can be used as a **preconditioner** for other unbiased methods

$$\int f(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) \frac{\pi(\mathbf{x})}{T_\sharp \rho(\mathbf{x})} T_\sharp \rho(\mathbf{x}) d\mathbf{x} = \int f \circ T(\mathbf{x}) \frac{T^\sharp \pi(\mathbf{x})}{\rho(\mathbf{x})} \rho(\mathbf{x}) d\mathbf{x}$$

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 $T^\star_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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 $T_\sharp^\star \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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 [Spantini '18]
- Low-rank structure

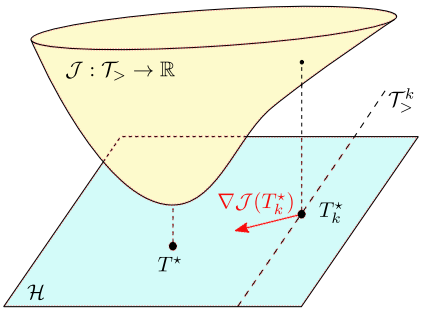
Adaptivity

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

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

Refinement criterion

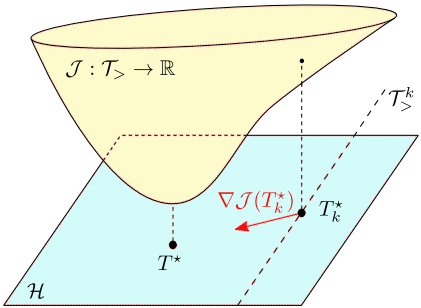
$$\mathbf{a}_k^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^{n_k}} \underbrace{\mathbb{E}\left[\log \frac{\rho}{T[\mathbf{a}]^\sharp \widetilde{\pi}}\right]}_{\mathcal{J}(T[\mathbf{a}])}$$



Refinement criterion

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

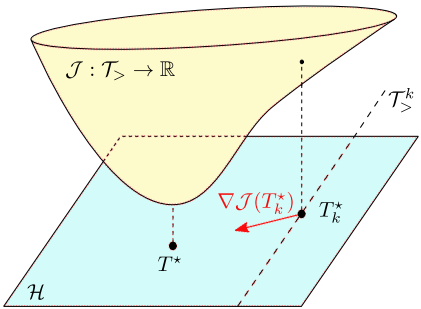
$$\nabla_{\mathbf{a}} \mathcal{J}(T[\mathbf{a}_k^\star]) = 0$$



Refinement criterion

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

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



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

There exists $\varepsilon > 0$ such that
$\mathcal{J}\left(T[\mathbf{a}_k^\star] - \varepsilon \nabla \mathcal{J}\left(T[\mathbf{a}_k^\star]\right)\right) < \mathcal{J}\left(T[\mathbf{a}_k^\star]\right)$

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

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

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

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 $\mathcal{H} \supset \mathcal{T}_>$
- We find the **"enriched"** space $\mathcal{T}_>^{k+1}$ ($\widetilde{\mathcal{T}}^k \supset \mathcal{T}_>^{k+1} \supset \mathcal{T}^k$), using

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

as a guidance criterion.

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 $\mathcal{H} \supset \mathcal{T}_>$
- We find the **"enriched"** space $\mathcal{T}_>^{k+1}$ ($\widetilde{\mathcal{T}}^k \supset \mathcal{T}_>^{k+1} \supset \mathcal{T}^k$), using

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

as a guidance criterion.

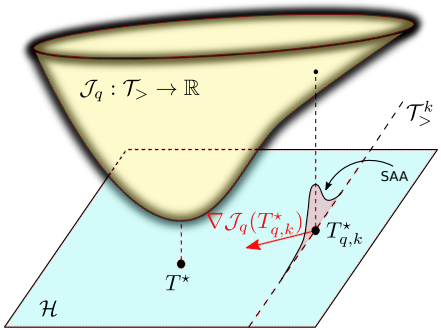
No new evaluation of $\log \pi$ or $\nabla_{\mathbf{x}} \log \pi$ are required

Controlling the sample average accuracy

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

Controlling the sample average accuracy

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

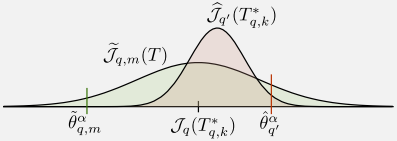


Controlling the sample average accuracy

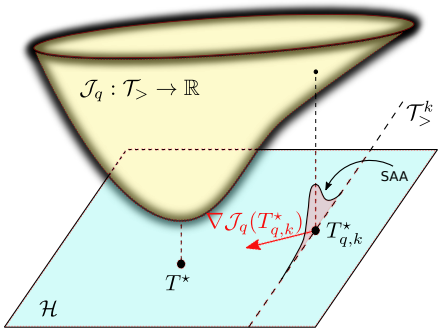
$$T_k^\star = \arg\min_{T \in \mathcal{T}_>^k} -\mathbb{E}_\rho \left[\log T^\sharp \pi \right] \approx \arg\min_{T \in \mathcal{T}_>^k} \overbrace{-\mathcal{Q}_q \left[\log T^\sharp \pi(\mathbf{x}_i) \right]}^{\mathcal{J}_q(T)} =: T_{q,k}^\star$$

Sample average approximation

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

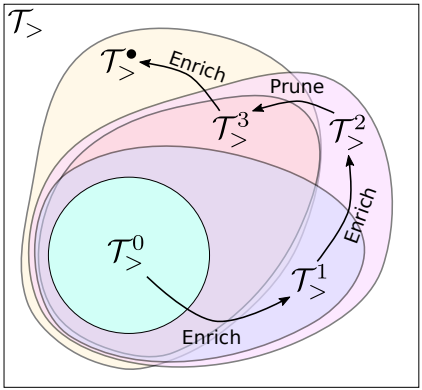


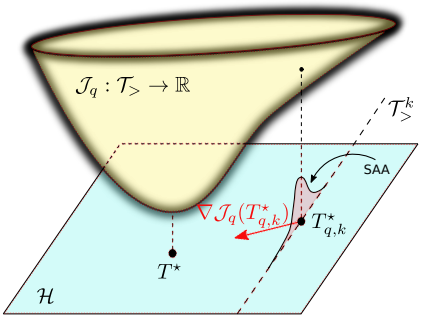
$$\widetilde{J}_{q,m}(T) = \frac{1}{m} \sum_{i=1}^{m} \min_{T \in \mathcal{T}_>^k} \mathcal{J}_q(T)$$



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





Off-line learning / On-line inference

$$\text{Joint distribution } \boldsymbol{\nu}_{\pi'} \; : \qquad \pi'(\mathbf{X}, \boldsymbol{\Theta}) := \pi\left(\mathbf{X}|\boldsymbol{\Theta}\right) \pi(\boldsymbol{\Theta})$$

$$\text{Posterior distribution } \boldsymbol{\nu}_{\pi} \; : \qquad \widetilde{\pi}(\boldsymbol{\Theta}|\mathbf{X} = \hat{\mathbf{x}}) \propto \mathcal{L}_{\hat{\mathbf{x}}}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$$

Off-line learning / On-line inference

$$\text{Joint distribution } \boldsymbol{\nu}_{\pi'} : \quad \pi'(\mathbf{X}, \boldsymbol{\Theta}) := \pi(\mathbf{X}|\boldsymbol{\Theta})\,\pi(\boldsymbol{\Theta})$$

$$\text{Posterior distribution } \boldsymbol{\nu}_{\pi} : \quad \widetilde{\pi}(\boldsymbol{\Theta}|\mathbf{X} = \hat{\mathbf{x}}) \propto \mathcal{L}_{\hat{\mathbf{x}}}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$$

Let $T(\mathbf{y}, \boldsymbol{\theta}) = \begin{bmatrix} T^{(\mathbf{y})}(\mathbf{y}) \\ T^{(\boldsymbol{\theta})}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix}$ be such that $T_\sharp \mathcal{N}(\mathbf{0}, \mathbf{I}) \approx \boldsymbol{\nu}_{\pi'}$,

then given data $\hat{\mathbf{x}}$

$$C_{\hat{\mathbf{y}}}(\boldsymbol{\theta}) := T^{(\boldsymbol{\theta})}(\hat{\mathbf{y}}, \boldsymbol{\theta}) \text{ is such that } \left(C_{\hat{\mathbf{y}}}\right)_\sharp \boldsymbol{\nu}_\rho \approx \boldsymbol{\nu}_\pi$$

$$\text{where } \hat{\mathbf{y}} = \left(T^{(\mathbf{y})}\right)^{-1}(\hat{\mathbf{x}})$$

Biochemical Oxygen Demand

Biochemical Oxygen Demand

We model the oxygen level at time t by

$$X(t) = A(1 - \exp(-Bt)) + \varepsilon \,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \,,$$
$$A \sim \mathcal{U}(0.4, 1.2) \qquad \text{and} \qquad B \sim \mathcal{U}(0.01, 0.31) \,,$$

and we want to

Biochemical Oxygen Demand

We model the oxygen level at time t by

$$X(t) = A(1 - \exp(-Bt)) + \varepsilon\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,,$$
$$A \sim \mathcal{U}(0.4, 1.2) \qquad \text{and} \qquad B \sim \mathcal{U}(0.01, 0.31)\,,$$

and we want to

1. Off-line: characterize the joint distribution $(X(1), \dots, X(4), A, B) \sim \boldsymbol{\nu}_{\pi'}$.

Biochemical Oxygen Demand

We model the oxygen level at time t by

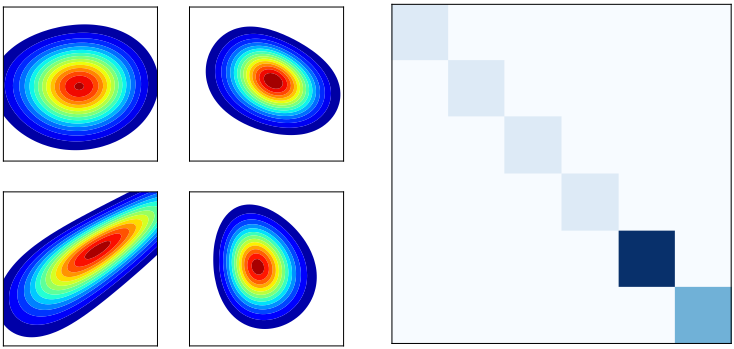
$$X(t) = A(1 - \exp(-Bt)) + \varepsilon\,, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,,$$
$$A \sim \mathcal{U}(0.4, 1.2) \qquad \text{and} \qquad B \sim \mathcal{U}(0.01, 0.31)\,,$$

and we want to

1. Off-line: characterize the joint distribution $(X(1), \dots, X(4), A, B) \sim \boldsymbol{\nu}_{\pi'}$.
2. On-line: for data $\hat{\mathbf{x}}$, generate samples from the posterior $(A, B|\mathbf{X} = \hat{\mathbf{x}}) \sim \boldsymbol{\nu}_{\pi}$.

Biochemical Oxygen Demand – off-line phase

Iteration 1

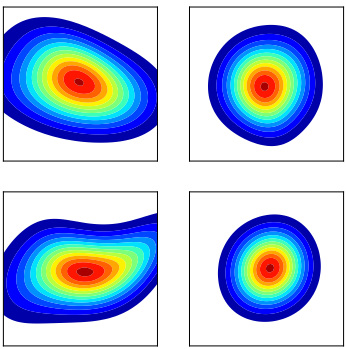


$T^\sharp\pi$ $\nabla_{\mathbf{x}} T$

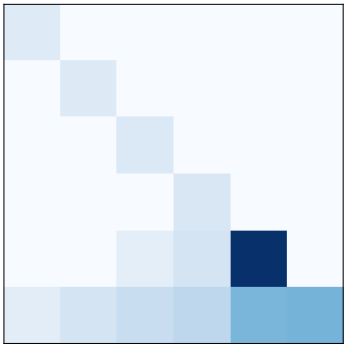
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 2



$T^\sharp\pi$

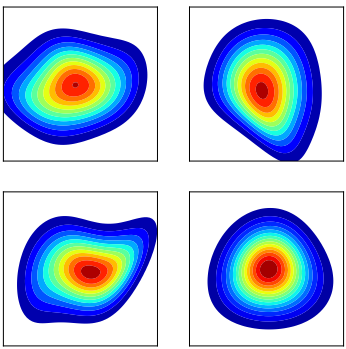


$\nabla_{\mathbf{x}} T$

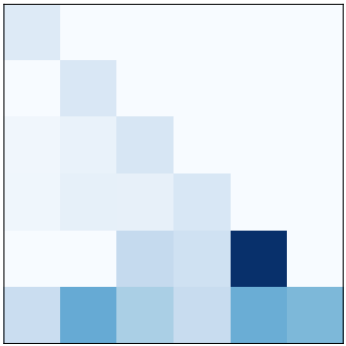
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 3



$T^\sharp\pi$

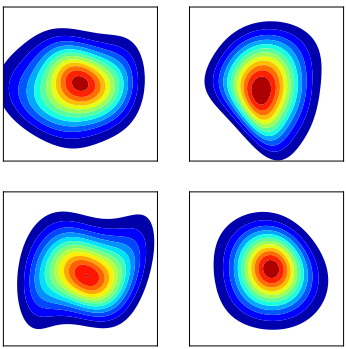


$\nabla_{\mathbf{x}} T$

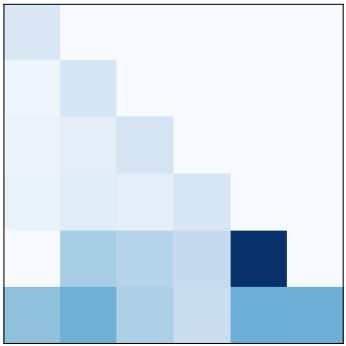
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 4



$T^\sharp\pi$

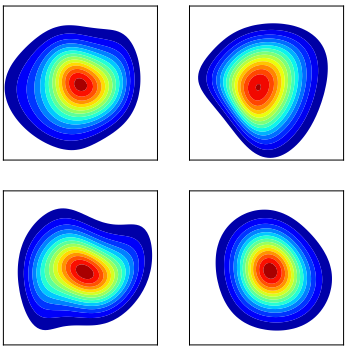


$\nabla_{\mathbf{x}} T$

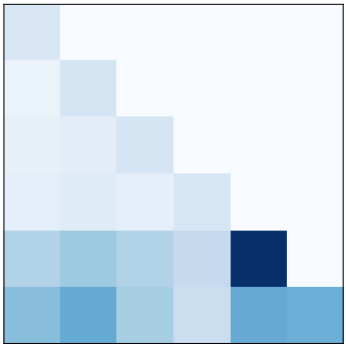
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 5



$T^\sharp\pi$

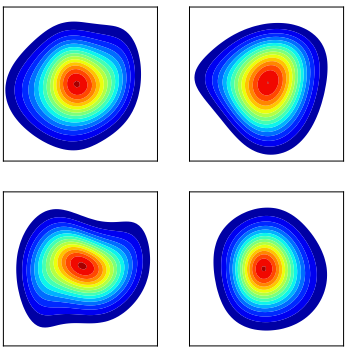


$\nabla_{\mathbf{x}} T$

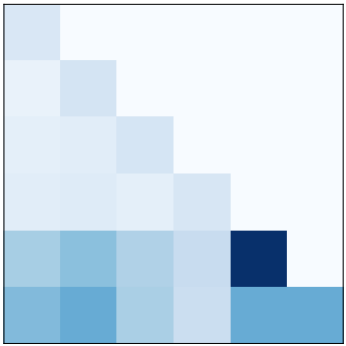
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 6



$T^\sharp\pi$

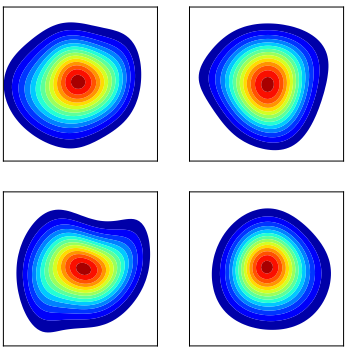


$\nabla_{\mathbf{x}} T$

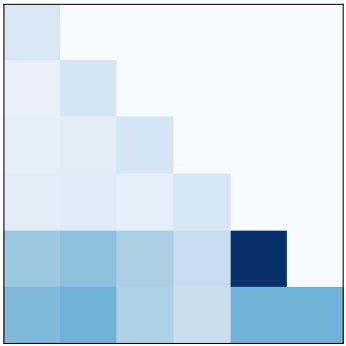
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 7



$T^\sharp\pi$

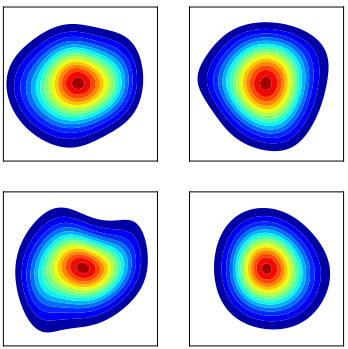


$\nabla_{\mathbf{x}} T$

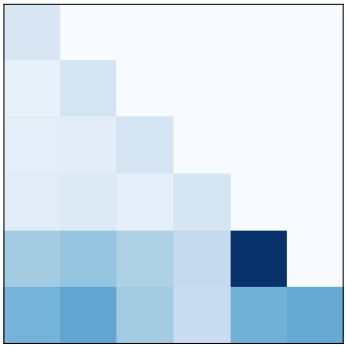
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 8



$T^\sharp\pi$

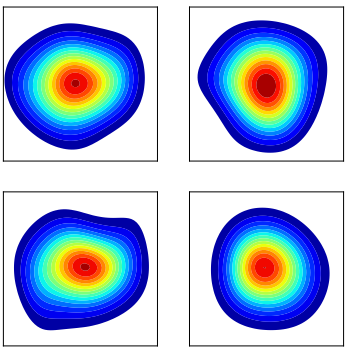


$\nabla_{\mathbf{x}} T$

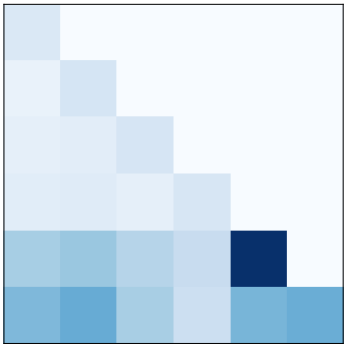
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 9



$T^\sharp\pi$

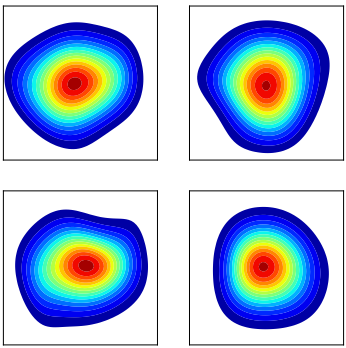


$\nabla_{\mathbf{x}} T$

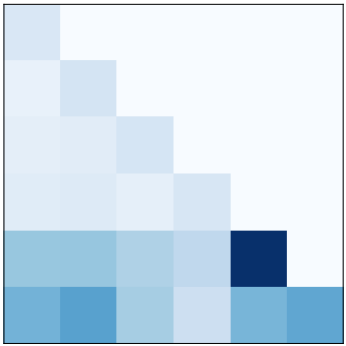
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 10



$T^\sharp\pi$

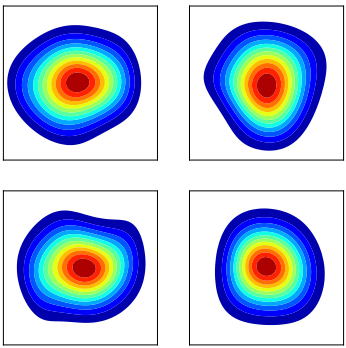


$\nabla_{\mathbf{x}} T$

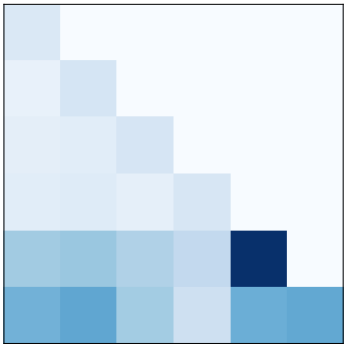
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 11



$T^\sharp\pi$

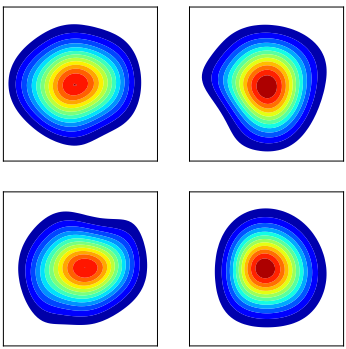


$\nabla_{\mathbf{x}} T$

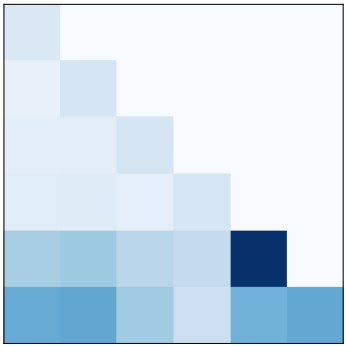
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 12



$T^\sharp\pi$

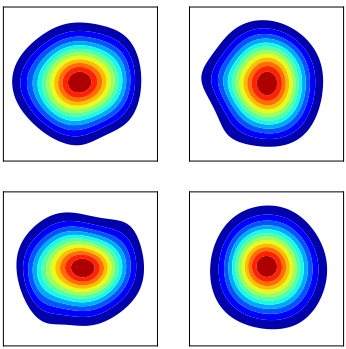


$\nabla_{\mathbf{x}} T$

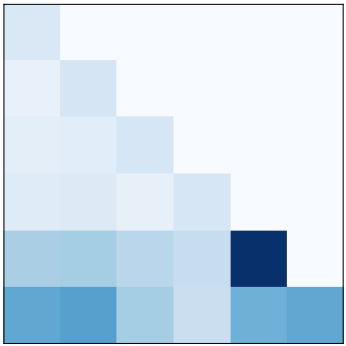
Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

Iteration 13



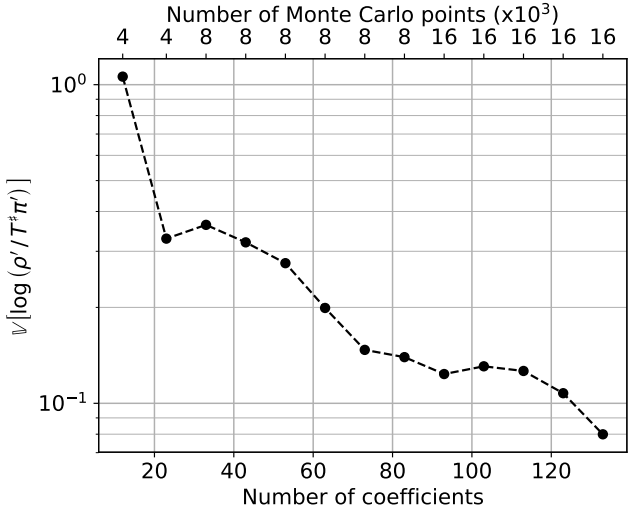
$T^\sharp\pi$



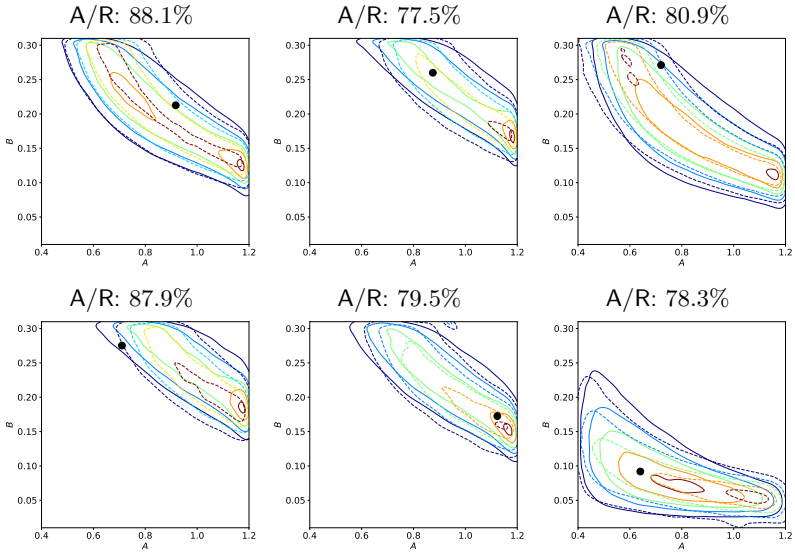
$\nabla_{\mathbf{x}} T$

Reminder: $T^\sharp\pi' \approx \rho'$, where ρ' is the density of $\mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand – off-line phase

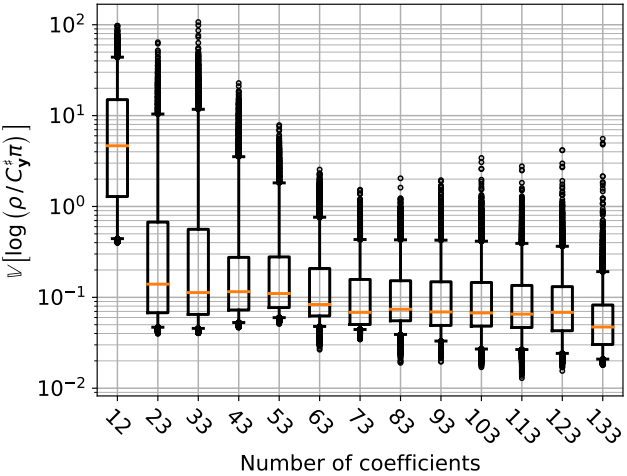


Biochemical Oxygen Demand – on-line phase



Dashed lines are contours of Markov Chains long 10^6
generated with Metropolis-Hastings with independent proposals.

Biochemical Oxygen Demand – on-line phase



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