Layers of low-rank couplings for large-scale Bayesian inference

D. Bigoni[†] (**dabi@mit.edu**), O. Zahm[‡], A. Spantini[†], Y.M. Marzouk[†]

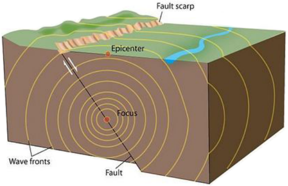
[†] Massachusetts Institute of Technology, USA

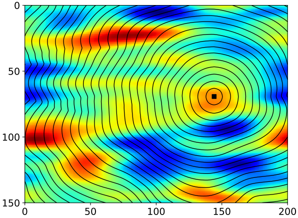
[‡] INRIA, France

SIAM CSE 2019

Spokane, WA – February 27, 2019

Bayesian inference – an oversimplified example

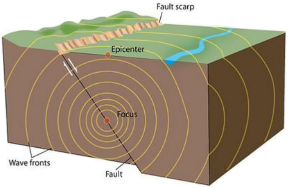


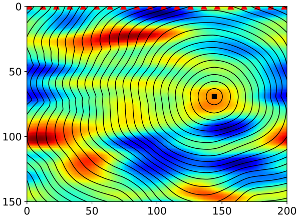


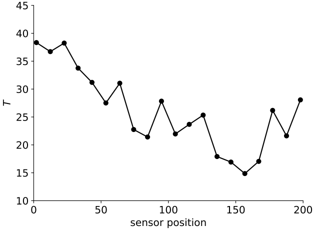
Mathematical model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

velocity field

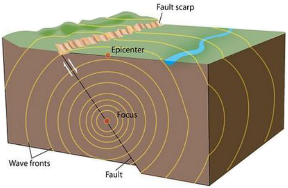
Observational model

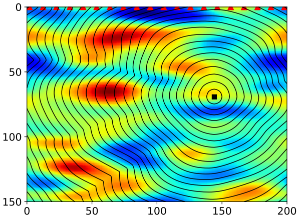
data

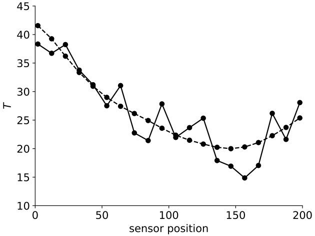
$$\overbrace{\mathbf{d}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}$$

noise

Bayesian inference – an oversimplified example







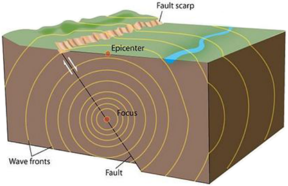
Mathematical model

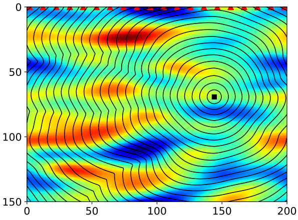
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

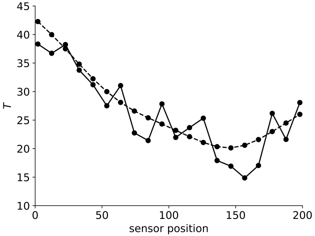
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

velocity field

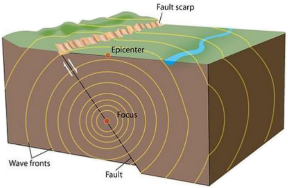
Observational model

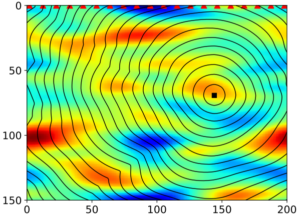
data

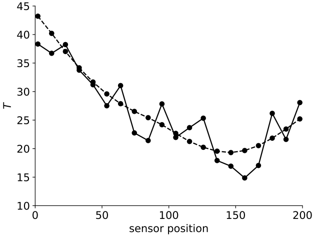
$$\mathbf{d} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}$$

noise

Bayesian inference – an oversimplified example







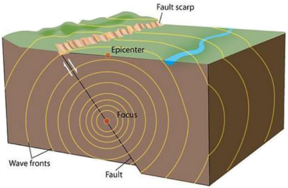
Mathematical model

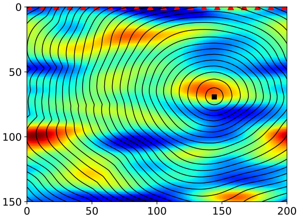
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

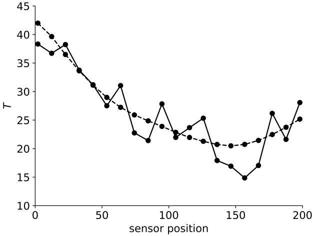
Observational model

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

Bayesian inference – an oversimplified example







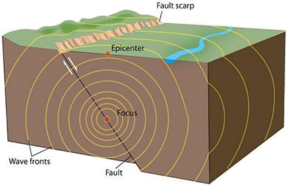
Mathematical model

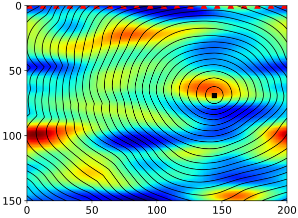
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

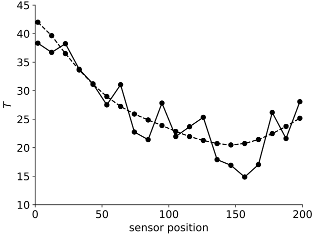
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

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

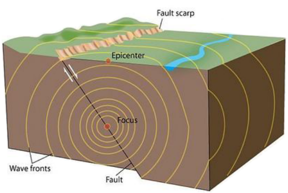
Observational model

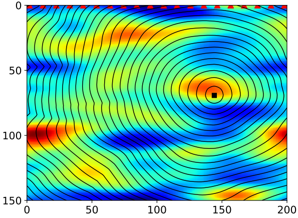
$$\overbrace{\mathbf{d}}^{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

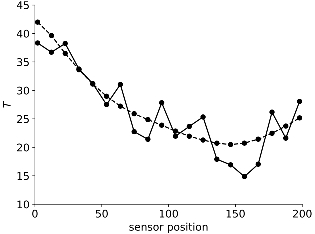
Bayesian inference model

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

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

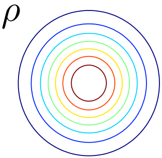
Transport maps

Deep lazy maps

Results

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

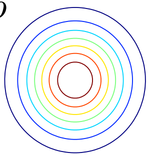
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$

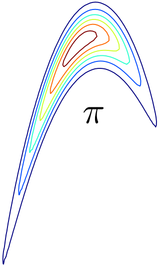


Transport maps – Pullbacks [PB] and Pushforwards [PF]

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

ρ



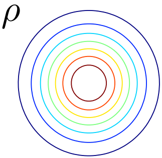


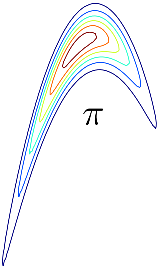
Transport maps – Pullbacks [PB] and Pushforwards [PF]

- 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

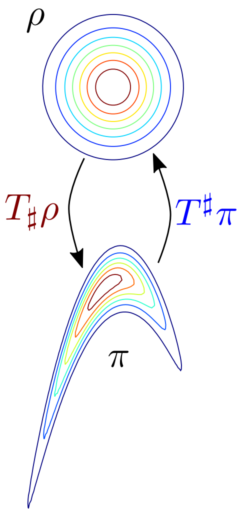
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  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

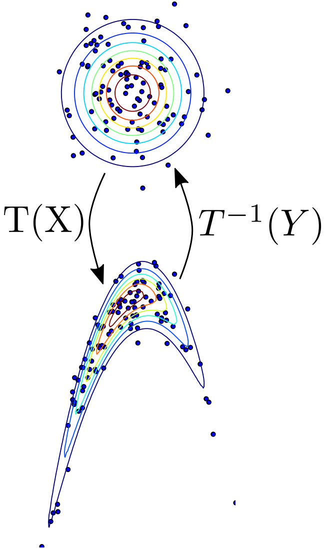
  PF $T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  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

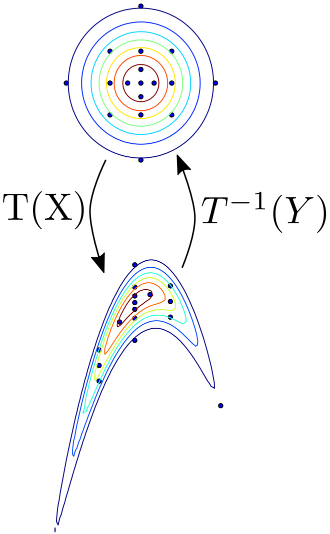
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  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

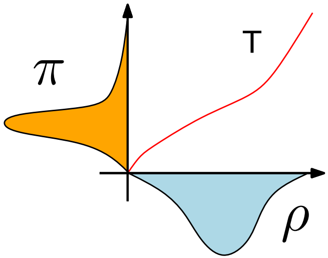
  PF $T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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



Knothe-Rosenblatt rearrangement

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

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

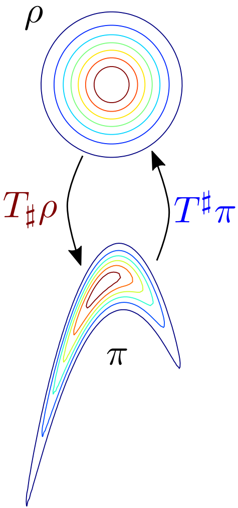
Triangular monotone maps

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

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

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

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

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

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

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!**

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

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 [Chen, MS238]
- Conditional independence [Baptista, MS327]
- Marginal independence
- Low-rank structure

Deep Lazy maps

Incrementally construct improving maps
by working on residuals distributions.

What is a lazy map?

Few ($k \ll d$) complex components and many "lazy" linear components:

$$T(\mathbf{x}) = \begin{bmatrix} T^{(1)}(x_1) \\ T^{(2)}(x_1, x_2) \\ \vdots \\ T^{(k)}(x_1, \dots, x_k) \\ a_{k+1} + b_{k+1} x_{k+1} \\ \vdots \\ a_d + b_d x_d \end{bmatrix}$$

This map is effective if ρ and π agree[1] along $d - k$ coordinates.

[1] but for a linear re-scaling

Assume there exists a rotation matrix $\mathbf{Q}$ such that

$$\int \pi \circ \mathbf{Q}(\boldsymbol{\xi}_{1:k}, \boldsymbol{x}_{k+1:d}) \, \mathrm{d}\boldsymbol{\xi}_{1:k} = \int \rho(\boldsymbol{\xi}_{1:k}, \boldsymbol{x}_{k+1:d}) \, \mathrm{d}\boldsymbol{\xi}_{1:k},$$

Then there exist a "low-rank map"

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

such that

$$T_\sharp \rho = \mathbf{Q}^\sharp \pi$$

Finding a good rotation $\mathbf{Q}$

For any distribution $\boldsymbol{\nu}_\eta$ with finite second moment, let

$$(\mathbf{H}_\eta)_{ij} = \int \partial_i \mathfrak{r}(\boldsymbol{x})\, \partial_j \mathfrak{r}(\boldsymbol{x})\, \eta(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\ , \qquad \mathfrak{r} := \log(\pi/\rho).$$

If $\mathrm{rank}(\mathbf{H}_\eta) = k$ and $\boldsymbol{\nu}_\rho = \mathcal{N}(0, \mathbf{I})$, then

there exist a rotation $\mathbf{Q}$ and a rank-k map T

such that $T_\sharp \rho = \mathbf{Q}^\sharp \pi$

Certified approximation $\pi^\star$ and optimal rotation Q [Zahm2018]

Let the columns of $\mathbf{U} \in \mathbb{R}^{d\times k}$ be the eigenvectors corresponding to the largest k eigenvalues $\{\lambda_i\}_{i=1}^k$ of $\mathbf{H}_\eta$ and let

$$\pi^\star(\boldsymbol{x}) := f(\mathbf{U}^\top \boldsymbol{x})\rho(\boldsymbol{x}) \ ,$$

for f given by the conditional expectation

$$f(\mathbf{z}) := \mathbb{E}\left[\pi(\mathbf{X})/\rho(\mathbf{X}) \middle| \mathbf{U}^\top \mathbf{X} = \boldsymbol{z}\right], \qquad \mathbf{X} \sim \rho \ .$$

Then,

$$\mathcal{D}_{\mathsf{KL}}\left(\pi \| \pi^\star\right) \leq \lambda_{k+1} + \ldots + \lambda_d \qquad \text{and} \qquad \mathbf{Q} = [\mathbf{U}|\mathbf{U}_\perp]$$

In practical problems...

$$(\mathbf{H}_\eta)_{ij} = \int \partial_i \mathfrak{r}(\boldsymbol{x}) \, \partial_j \mathfrak{r}(\boldsymbol{x}) \, \eta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \ , \qquad \mathfrak{r} := \log(\pi/\rho).$$

- $\mathbf{H}_\eta$ will need to be approximated using some quadrature
- $\mathbf{H}_\eta$ will only be approximately low-rank
- The spectrum of $\mathbf{H}_\eta$ will depend on the sampling distribution $\boldsymbol{\nu}_\eta$ (the optimal distribution would be $\boldsymbol{\nu}_\pi$ itself)

In practical problems...

$$(\mathbf{H}_\eta)_{ij} = \int \partial_i \mathfrak{r}(\boldsymbol{x})\, \partial_j \mathfrak{r}(\boldsymbol{x})\, \eta(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\ , \qquad \mathfrak{r} := \log(\pi/\rho).$$

- $\mathbf{H}_\eta$ will need to be approximated using some quadrature
- $\mathbf{H}_\eta$ will only be approximately low-rank
- The spectrum of $\mathbf{H}_\eta$ will depend on the sampling distribution $\boldsymbol{\nu}_\eta$ (the optimal distribution would be $\boldsymbol{\nu}_\pi$ itself)

We will have to resort to lazy maps rather than low-rank maps

procedure DeepLowRankConstruction(π, k_{max}, ε_k, ε_{map}, $\varepsilon_\bullet$)

  $\mathfrak{T} \leftarrow I_n$, where I_n is the identity map

  while $\mathbb{V}\left[\log \frac{T^\sharp \pi}{\eta}\right] > \varepsilon_\bullet$ **do**

    Build quadrature $(\boldsymbol{x}_i, \boldsymbol{w}_i)_{i=1}^{2k_{\text{max}}}$ with respect to $\mathcal{N}(0, \mathbf{I})$

    $\mathbf{U}, k \leftarrow$ ComputeSubspace($(\boldsymbol{x}_i, \boldsymbol{w}_i)_{i=1}^{2\,k_{\text{max}}}$, $\mathfrak{T}^\sharp \pi$, ε_k)

    Characterize the lazy map T such that

$$\mathbb{V}\left[\log \frac{T^\sharp (\mathbf{U}|\mathbf{U}_\perp)^\sharp \pi}{\eta}\right] < \varepsilon_{\text{map}}$$

    $\mathfrak{T} \leftarrow \mathfrak{T} \circ ((\mathbf{U}|\mathbf{U}_\perp) \cdot T)$

  end while

  return T

end procedure

procedure DeepLowRankConstruction(π, k_{max}, ε_k, $\varepsilon_{\mathsf{map}}$, $\varepsilon_\bullet$)

  $\mathfrak{T} \leftarrow I_n$, where I_n is the identity map

  while $\mathbb{V}\left[\log \frac{T^\sharp \pi}{\eta}\right] > \varepsilon_\bullet$ **do**

    Build quadrature $(\boldsymbol{x}_i, \boldsymbol{w}_i)_{i=1}^{2k_{\mathsf{max}}}$ with respect to $\mathcal{N}(0, \mathbf{I})$

    $\mathbf{U}, k \leftarrow$ ComputeSubspace($(\boldsymbol{x}_i, \boldsymbol{w}_i)_{i=1}^{2\,k_{\mathsf{max}}}$, $\mathfrak{T}^\sharp \pi$, ε_k)

    Characterize the lazy map T such that

$$\mathbb{V}\left[\log \frac{T^\sharp (\mathbf{U}|\mathbf{U}_\perp)^\sharp \pi}{\eta}\right] < \varepsilon_{\mathsf{map}}$$

    $\mathfrak{T} \leftarrow \mathfrak{T} \circ ((\mathbf{U}|\mathbf{U}_\perp) \cdot T)$

  end while

  return T

end procedure

$\mathfrak{T}$ progressively “Gaussianizes” π.

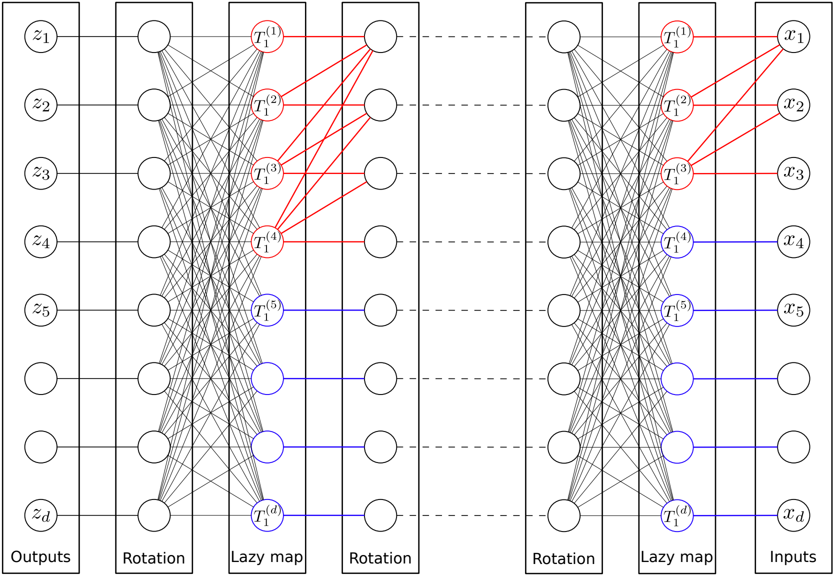
As the $\mathfrak{T}$ improves the subspace approximation improves...

```
procedure ComputeSubspace((x_i, w_i)_{i=1}^m, T^♯π, ε)
    Assemble
```

$$\mathbf{H}_\rho = \sum_{i=1}^{m} \left(\nabla_{\mathbf{x}} \log \frac{\mathfrak{T}^\sharp \pi}{\rho}(\boldsymbol{x}_i) \right) \left(\nabla_{\mathbf{x}} \log \frac{\mathfrak{T}^\sharp \pi}{\rho}(\boldsymbol{x}_i) \right)^T \boldsymbol{w}_i$$

```
    Solve the eigenvalue problem H_ρ X = Λ X
    Define U = [X_{:,1}, ..., X_{:,k}] for k s.t. Σ_{i=k+1}^n λ_i < ε
    return U, k
end procedure
```

Composition of layers (deep) of lazy transport maps



In practice...

Elliptic problem with unknown coefficients

Forward model

$\mathbf{G} : \boldsymbol{\kappa} \mapsto \boldsymbol{u}$

$$
\begin{cases}
-\nabla \cdot (\kappa(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) = 0 & \text{in} \quad \Gamma \times \Omega \\
u(\mathbf{x},\omega) = 0 & \text{on} \quad \mathbf{x}_1 = 0 \\
u(\mathbf{x},\omega) = 1 & \text{on} \quad \mathbf{x}_1 = 1 \\
-\frac{\partial u}{\partial n}(\mathbf{x}) = 0 & \text{on} \quad \mathbf{x}_2 \in \{0,1\}
\end{cases}
$$

$$
\kappa(\mathbf{x},\omega) = \exp\left(g(\mathbf{x},\omega)\right) \ , \quad g(\mathbf{x},\omega) \sim \mathcal{N}\left(\mathbf{0}, C_g(\mathbf{x},\mathbf{x}')\right)
$$

$$
C_g(\mathbf{x},\mathbf{x}') = \exp\left(-|\mathbf{x}-\mathbf{x}'|\right)
$$

Elliptic problem with unknown coefficients

Forward model $\mathbf{G} : \boldsymbol{\kappa} \mapsto \boldsymbol{u}$

$$\begin{cases} -\nabla \cdot (\kappa(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) = 0 & \text{in} \quad \Gamma \times \Omega \\ u(\mathbf{x},\omega) = 0 & \text{on} \quad \mathbf{x}_1 = 0 \\ u(\mathbf{x},\omega) = 1 & \text{on} \quad \mathbf{x}_1 = 1 \\ -\frac{\partial u}{\partial n}(\mathbf{x}) = 0 & \text{on} \quad \mathbf{x}_2 \in \{0,1\} \end{cases}$$

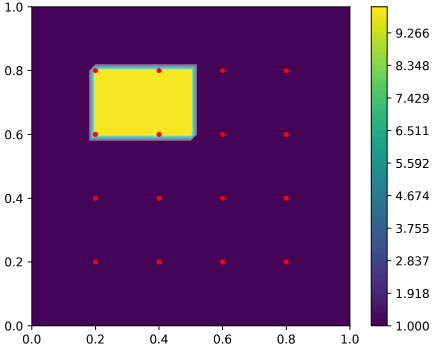
$$\kappa(\mathbf{x},\omega) = \exp\left(g(\mathbf{x},\omega)\right) , \quad g(\mathbf{x},\omega) \sim \mathcal{N}\left(\mathbf{0}, C_g(\mathbf{x},\mathbf{x}')\right)$$

$$C_g(\mathbf{x},\mathbf{x}') = \exp\left(-|\mathbf{x}-\mathbf{x}'|\right)$$

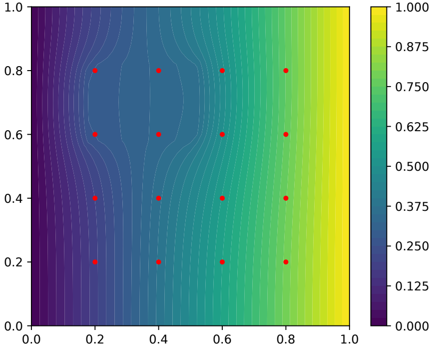
Bayesian inverse problem

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

Elliptic problem with unknown coefficients

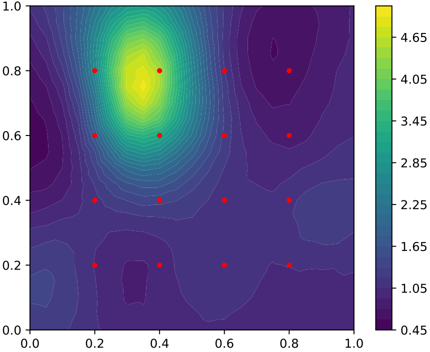


Synthetic field $\boldsymbol{\kappa}^\star(\mathbf{x})$

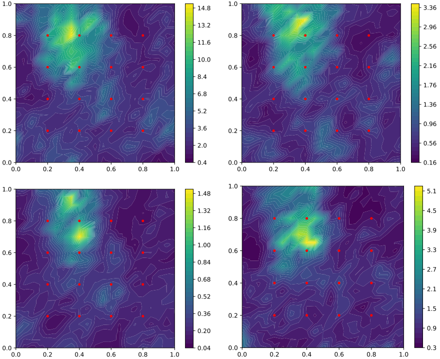


Synthetic solution $\mathbf{G}(\boldsymbol{\kappa}^\star)$

Elliptic problem with unknown coefficients

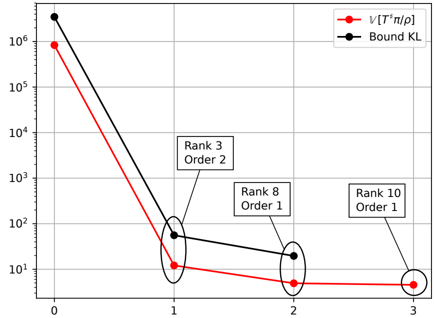


Posterior mean $\mathbb{E}\left[\boldsymbol{\kappa}(\mathbf{x})|\mathbf{d}\right]$

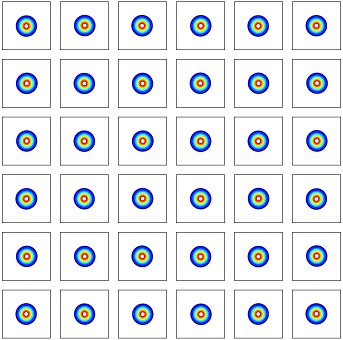


Realizations of $\boldsymbol{\kappa} \sim \pi_{\text{pos}}$

Elliptic problem with unknown coefficients



Variance diagnostic $\mathbb{V}\left[\log \frac{\mathfrak{T}^\sharp \pi}{\rho}\right]$



Random conditionals of $\mathfrak{T}^\sharp \pi \approx \mathcal{N}(0, \mathbf{I})$

Biochemical Oxygen Demand

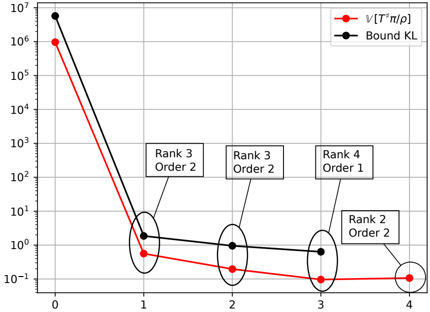
We model the oxygen level at time t by

$$X(t) = A(1 - \exp(-Bt)) + \varepsilon \, , \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \, ,$$
$$A \sim \log \mathcal{N}(0.9, 0.3) \qquad \text{and} \qquad B \sim \log \mathcal{N}(0.16, 0.3) \, ,$$

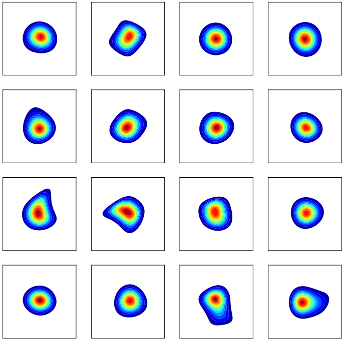
and we want to

Characterize the joint distribution $(X(1), \dots, X(4), A, B) \sim \boldsymbol{\nu}_\pi$.

Elliptic problem with unknown coefficients



Variance diagnostic $\mathbb{V}\left[\log \frac{\mathfrak{T}^\sharp \pi}{\rho}\right]$



Random conditionals of $\mathfrak{T}^\sharp \pi \approx \mathcal{N}(0, \mathbf{I})$

Key contributions

Algorithms for characterizing probability measures
via layers of low-dimensional **deterministic couplings**

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

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

Zahm et al. "Certified dimension reduction in nonlinear Bayesian inverse problems" (arXiv)
Bigoni et al. "On the computation of monotone transports" (preprint)
Spantini et al. "Inference via low-dimensional couplings" (JMLR)
Marzouk et al. "Sampling via measure transport: an introduction" (Springer)
Parno et al. "Transport map accelerated Markov chain Monte Carlo" (JUQ)
El Moselhy et al. "Bayesian inference with optimal maps" (JCP)

Thanks to: U.S. DEPARTMENT OF ENERGY DARPA