Layers of lazy maps for large-scale inference

`https://arxiv.org/abs/1906.00031`

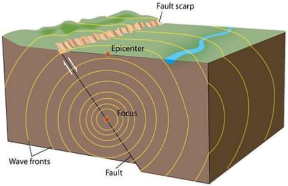
D. Bigoni[†] (**dabi@mit.edu**), O. Zahm[‡], A. Spantini[†], Y.M. Marzouk[†]

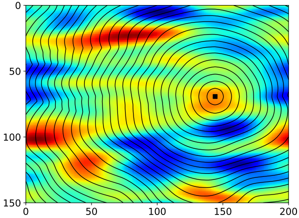
[†] Massachusetts Institute of Technology, USA

[‡] INRIA, France

Applied Inverse Problems Conference
Grenoble, France – July 12, 2019

Bayesian inference – an oversimplified example

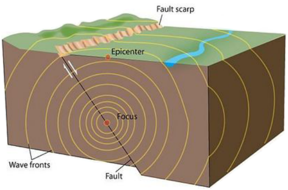


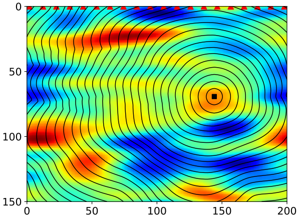


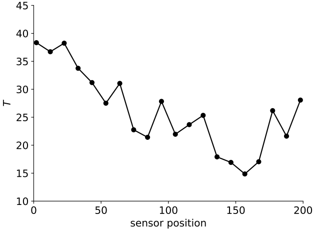
Mathematical model

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

Bayesian inference – an oversimplified example







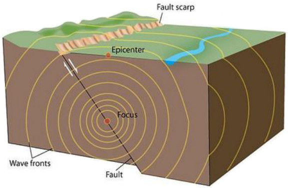
Mathematical model

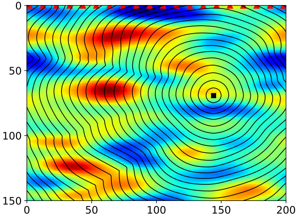
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

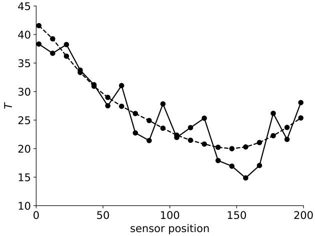
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

velocity field

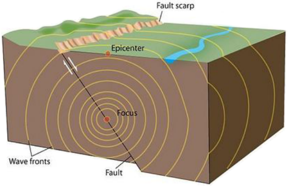
Observational model

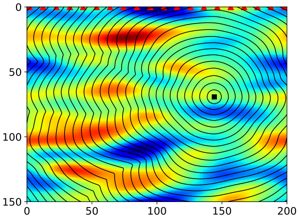
data

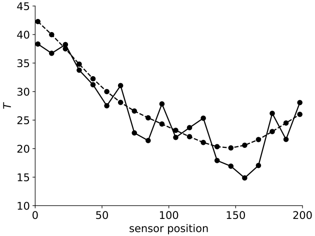
$$\underbrace{\mathbf{d}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

noise

Bayesian inference – an oversimplified example







Mathematical model

travel time

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

velocity field

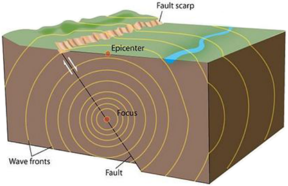
Observational model

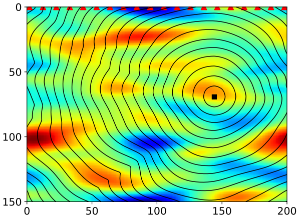
data

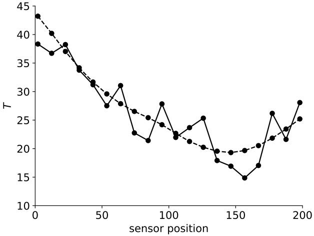
$$\mathbf{d} = \mathbf{G}(\mathbf{v}) + \varepsilon$$

noise

Bayesian inference – an oversimplified example







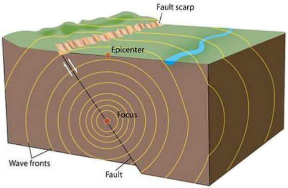
Mathematical model

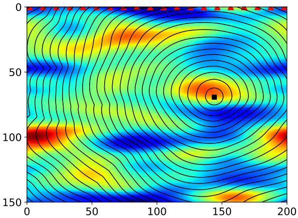
$$|\nabla \overbrace{G(\mathbf{x})}^{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

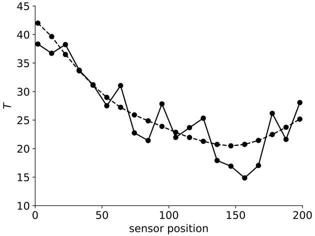
Observational model

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

Bayesian inference – an oversimplified example







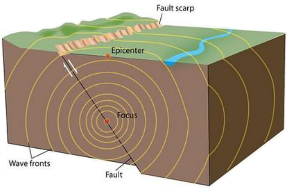
Mathematical model

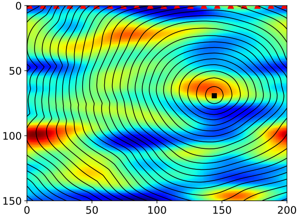
$$|\nabla \underbrace{G(\mathbf{x})}_{\text{travel time}}| = \underbrace{v(\mathbf{x})}_{\text{velocity field}}{}^{-1}$$

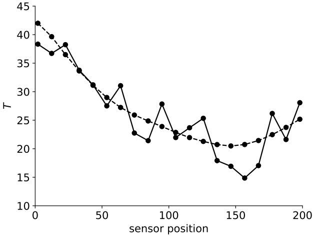
Observational model

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

Bayesian inference – an oversimplified example







Mathematical model

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

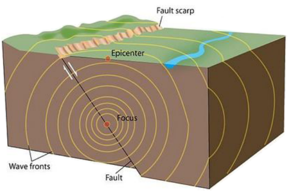
Observational model

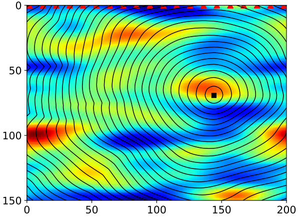
$$\underbrace{\mathbf{d}}_{\text{data}} = \mathbf{G}(\mathbf{v}) + \underbrace{\varepsilon}_{\text{noise}}$$

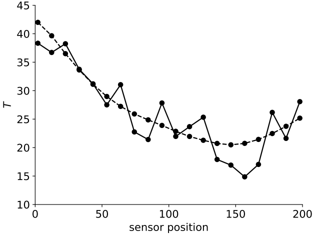
Bayesian inference model

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

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

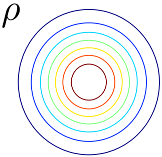
Transport maps

Layers of lazy maps

Results

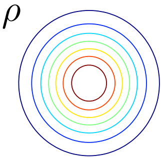
Transport maps – Pullbacks [PB] and Pushforwards [PF]

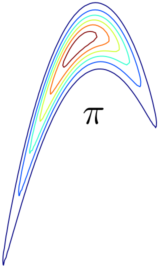
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$



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



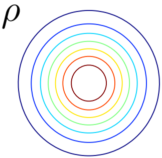


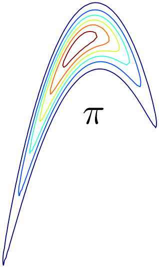
Transport maps – Pullbacks [PB] and Pushforwards [PF]

- 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

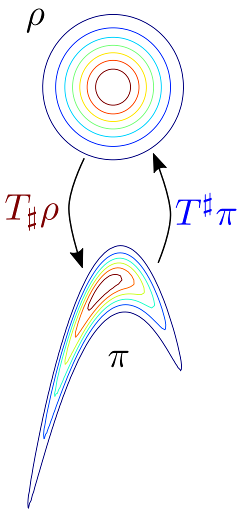
PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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]

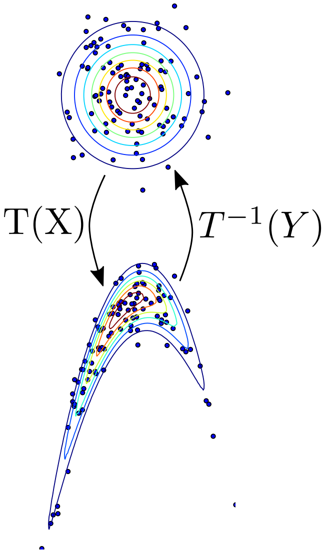
- Distribution $\boldsymbol{\nu}_\rho$ with density $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- Distribution $\boldsymbol{\nu}_\pi$ with density $\pi : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$
- For $T : \mathbb{R}^d \to \mathbb{R}^d$ we define

  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 $\quad$ 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

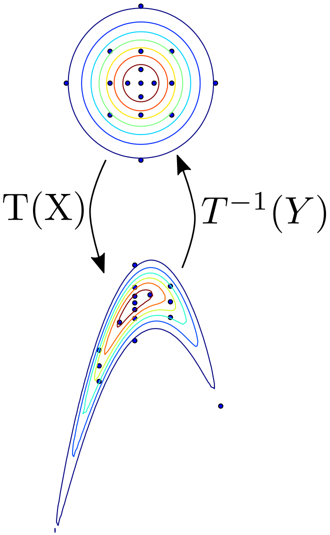
  PF $\quad T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

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



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

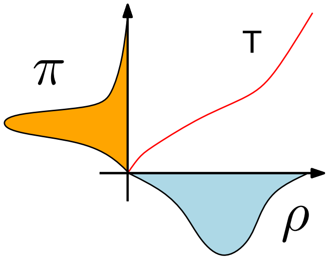
  PF $T_\sharp \rho = \rho \circ T^{-1} |\nabla T^{-1}|$

  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

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

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

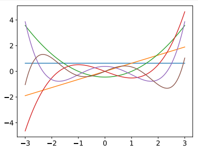
$$\boxed{\mathcal{T}_>^n} = \Big\{ T : \mathbb{R}^d \to \mathbb{R}^d : \overbrace{[T(\mathbf{x})]_k = T^{(k)}(x_1, \ldots, x_k)}^{\text{triangular}} \text{ and } \overbrace{\partial_{x_k} T^{(k)} > 0}^{\text{monotone}} \Big\}$$

Integrated squared representation – $\varepsilon > 0$

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

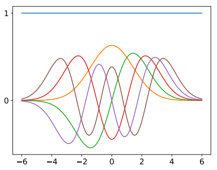
Constant part

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



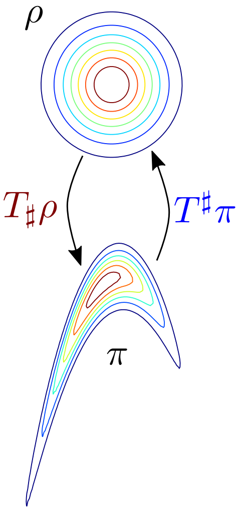
Squared part

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



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

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

Minimize KL-divergence to find optimal map

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**

Minimize KL-divergence to find optimal map

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **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

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

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

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

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

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

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

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

Pros & cons

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**
+ We can **assess convergence**!

Pros & cons

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**
+ We can **assess convergence**!

$\widehat{T}_\sharp \rho$ is a **biased** approximation of π.
How to reduce such bias?

Use $\widehat{T}$ as a preconditioner for ... Importance Sampling

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

Use $\widehat{T}$ as a preconditioner for ... Importance Sampling

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

If $\widehat{T}^\sharp\widetilde{\pi} \underset{\sim}{\propto} \rho$ then $\mathbb{V}\left[\frac{\widehat{T}^\sharp\widetilde{\pi}(\mathbf{x})}{\rho(\mathbf{x})}\right] \approx 0$ and

$$\sum_{i=1}^N f\circ\widehat{T}(\mathbf{x}_i)\,\mathbf{w}_i\ , \qquad \mathbf{w}_i = \frac{\widetilde{\mathbf{w}}_i}{\sum_{i=1}^N \widetilde{\mathbf{w}}_i}\ , \quad \widetilde{\mathbf{w}}_i = \frac{\widehat{T}^\sharp\widetilde{\pi}(\mathbf{x}_i)}{\rho(\mathbf{x}_i)}$$

will be an **accurate estimator** of $\int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$

Use $\widehat{T}$ as a preconditioner for ... Markov Chain Monte Carlo

1. Generate the Markov chain $\{\mathbf{x}_i\}$ with invariant distribution $\widehat{T}^\sharp \pi$
   (use your favorite MCMC method/proposals)
2. The Markov chain $\{T(\mathbf{x}_i)\}$ has invariant distribution π

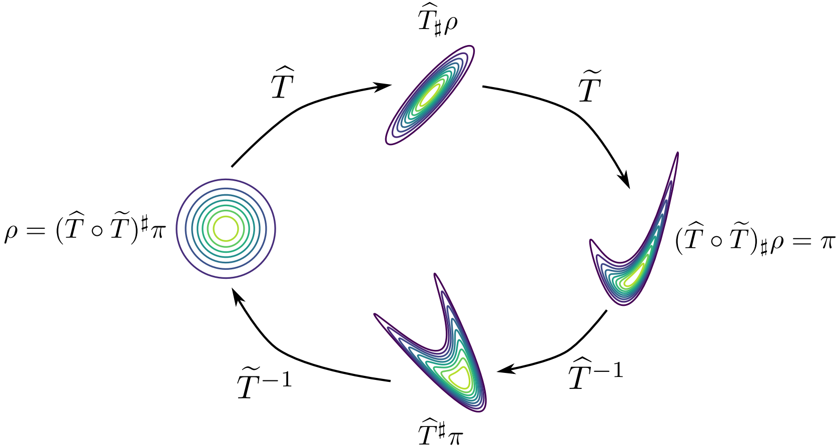
Use $\widehat{T}$ as a preconditioner for ... Markov Chain Monte Carlo

1. Generate the Markov chain $\{\mathbf{x}_i\}$ with invariant distribution $\widehat{T}^\sharp\pi$ (use your favorite MCMC method/proposals)
2. The Markov chain $\{T(\mathbf{x}_i)\}$ has invariant distribution π

If $\widehat{T}^\sharp\widetilde{\pi} \underset{\sim}{\propto} \rho$ then ρ is a **good proposal** for $\widehat{T}^\sharp\widetilde{\pi}$.

Use $\widehat{T}$ as a preconditioner for ... Transport Maps!

1. Solve $\widehat{T} = \arg\min_{T \in \mathcal{T}_>^n} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \pi}\right]$
2. Solve $\widetilde{T} = \arg\min_{T \in \mathcal{T}_>^n} \mathbb{E}_\rho \left[\log \frac{\rho}{T^\sharp \widehat{T}^\sharp \pi}\right]$



Pros & cons

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**
+ We can **assess convergence**!
+ The map can be used as a **preconditioner**

Pros & cons

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

+ **Gradient-based unconstrained optimization** if gradients are available
+ We can **explore π in parallel**
+ We can **generate i.i.d. samples** from $\widehat{T}_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$ **in parallel**
+ We can **assess convergence**!
+ The map can be used as a **preconditioner**
− We need to **approximate d functions of up to d variables!**

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

Exploit source of low-dimensional structure

1. Smoothness and marginal independence
   - Ongoing work...
2. Conditional independence
   - Variational filtering/smoothing and parameter estimation [Spantini, B, Marzouk 2018; Houssineau, Jasra, Singh 2018]
   - Ensamble filtering and smoothing [Spantini, Baptista, Marzouk 2019]
3. Multilevel/multifidelity structure [Parno, Moselhy, Marzouk 2018; Peherstorfer, Marzouk 2019]
4. Low-rank structure [B, Zahm, Spantini, Marzouk 2019]

Layers of lazy maps

Incrementally construct improving maps
by working on residual distributions.

What is a lazy map?

Few ($r \ll d$) complex components and many "lazy" identity components:

$$T(\mathbf{x}) = \begin{bmatrix} T^{(1)}(x_1) \\ T^{(2)}(x_1, x_2) \\ \vdots \\ T^{(r)}(x_1, \ldots, x_r) \\ x_{r+1} \\ \vdots \\ x_d \end{bmatrix} \in \mathcal{T}_r \subset \mathcal{T}_>$$

Maps of this type are effective if ρ and π agree along $d - r$ coordinates.

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

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

Then there exist a lazy map $T \in \mathcal{T}_r$ 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 $\text{rank}(\mathbf{H}_\eta) = r$ and $\boldsymbol{\nu}_\rho = \mathcal{N}(0, \mathbf{I})$, then

there exist a rotation $\mathbf{Q}$ and a lazy map $T \in \mathcal{T}_r$
such that $T_\sharp \rho = \mathbf{Q}^\sharp \pi$

Certified approximation $\pi^\star$ and optimal rotation $\mathbf{Q}$

[Zahm 2018, Bigoni 2019]

Let the columns of $\mathbf{U} \in \mathbb{R}^{d \times r}$ be the eigenvectors corresponding to the largest r eigenvalues $\{\lambda_i\}_{i=1}^r$ 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,

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

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)

Construction of one lazy map

procedure $\textsc{LazyMap}(\pi,\ \rho,\ \varepsilon,\ r_{\mathsf{max}})$
    Compute $H = \int (\nabla \log \frac{\pi}{\rho})(\nabla \log \frac{\pi}{\rho})^\top \mathrm{d}\rho$
    Solve the eigenvalue problem $H u_i = \lambda_i u_i$
    Let $r = r_{\mathsf{max}} \wedge \min\{ r \leq d \ : \frac{1}{2} \sum_{i>r} \lambda_i \leq \varepsilon \}$
    Assemble $\mathbf{Q} = [\mathbf{U} | \mathbf{U}_\perp]$.
    Find T solution to $\min_{T \in \mathcal{T}_r} \mathcal{D}_{\mathrm{KL}}(\rho || T^\sharp \mathbf{Q}^\sharp \pi)$
    return $\mathbf{Q} \circ T$
end procedure

Greedy construction of lazy maps

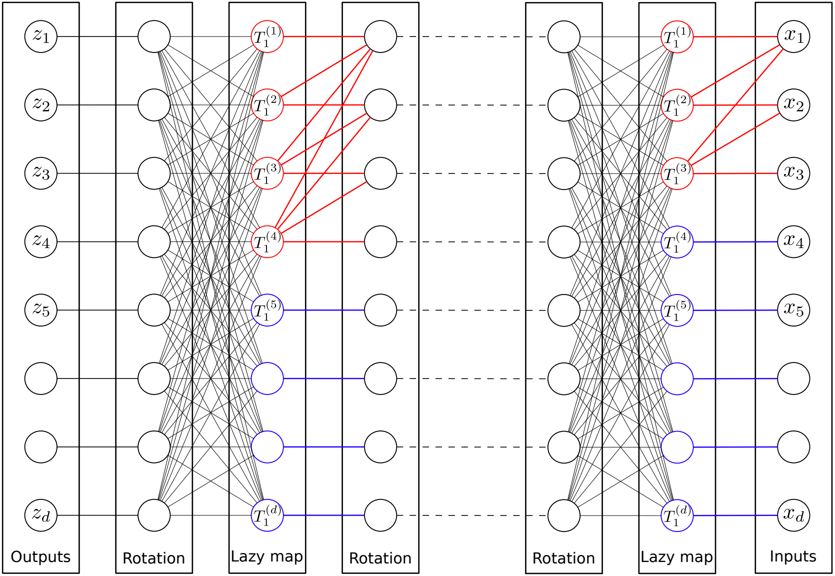
```
procedure LayersOfLazyMaps(π, ρ, ε, r, ℓ_max)
    Set π_0 = π and ℓ = 0
    while ℓ ≤ ℓ_max and ½ Tr(H_ℓ) ≥ ε do
        ℓ ← ℓ + 1
        Compute T_ℓ = LazyMap(π_{ℓ−1}, ρ, 0, r)
        Update 𝔗_ℓ = 𝔗_{ℓ−1} ∘ T_ℓ
        Compute π_ℓ = (𝔗_ℓ)^♯ π
        Compute H_ℓ = ∫(∇ log π_ℓ/ρ)(∇ log π_ℓ/ρ)^⊤ dρ
    end while
    return 𝔗_ℓ = T_1 ∘ ··· ∘ T_ℓ
end procedure
```

Greedy construction of lazy maps

procedure $\textsc{LayersOfLazyMaps}(\pi,\, \rho,\, \varepsilon,\, r,\, \ell_{\mathsf{max}})$
    Set $\pi_0 = \pi$ and $\ell = 0$
    while $\ell \leq \ell_{\mathsf{max}}$ and $\frac{1}{2}\,\mathrm{Tr}(H_\ell) \geq \varepsilon$ **do**
        $\ell \leftarrow \ell + 1$
        Compute $T_\ell = \textsc{LazyMap}(\pi_{\ell-1},\, \rho,\, 0,\, r)$
        Update $\mathfrak{T}_\ell = \mathfrak{T}_{\ell-1} \circ T_\ell$
        Compute $\pi_\ell = (\mathfrak{T}_\ell)^\sharp \pi$
        Compute $H_\ell = \int (\nabla \log \frac{\pi_\ell}{\rho})(\nabla \log \frac{\pi_\ell}{\rho})^\top \mathrm{d}\rho$
    end while
    return $\mathfrak{T}_\ell = T_1 \circ \cdots \circ T_\ell$
end procedure

$\mathfrak{T}$ progressively "Gaussianizes" π.

Composition of layers of lazy transport maps



In practice...

Log-Gaussian Cox process with sparse observations ($d = 64^2$)

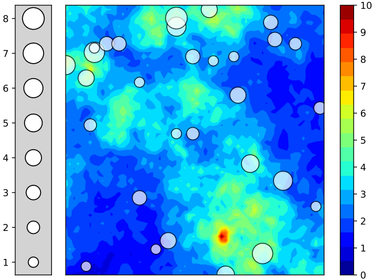
Statistical model

Observables: $\mathbf{Y} := (Y_i)_{i=1}^{30}\,, \quad Y_i \sim \mathrm{Poisson}(\mathbf{\Lambda}_i/d)$

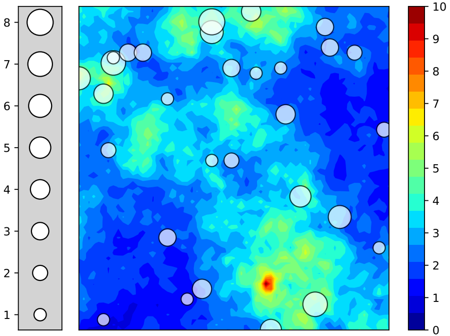
Latent field: $(\mathbf{\Lambda}_i)_{i=1}^{d} \sim \log \mathcal{N}(\mu, \mathrm{cov}(\mathbf{z}, \mathbf{z}'))$

$$\mathrm{cov}(\mathbf{z}, \mathbf{z}') = \sigma^2 \exp\left(-\left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2 / (64\beta)\right)$$

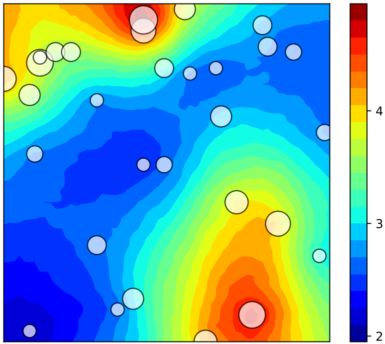
Posterior: $\pi(\mathbf{\Lambda}|\mathbf{Y} = \mathbf{y}^\star) \propto \pi(\mathbf{Y} = \mathbf{y}^\star|\mathbf{\Lambda})\pi(\mathbf{\Lambda})$



Log-Gaussian Cox process with sparse observations ($d = 64^2$)

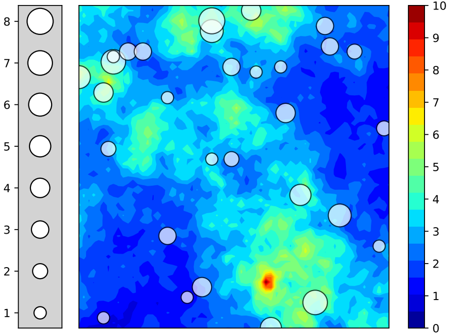


Field $\mathbf{\Lambda}^\star$ and observations $\mathbf{y}^\star$

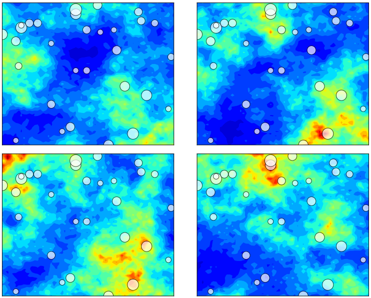


$\mathbb{E}[\mathbf{\Lambda}|\mathbf{y}^\star]$

Log-Gaussian Cox process with sparse observations ($d = 64^2$)

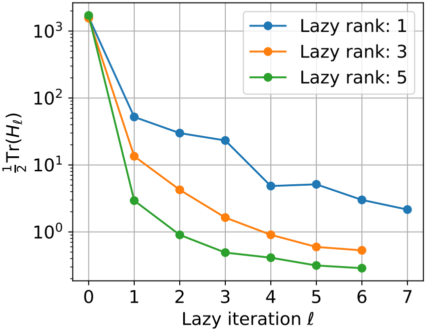


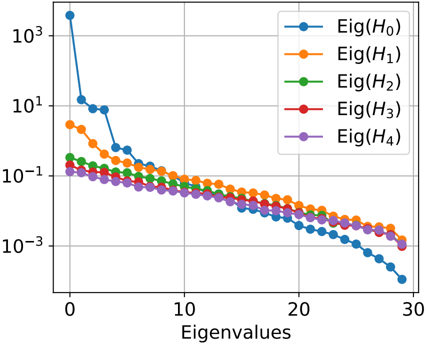
Field $\mathbf{\Lambda}^\star$ and observations $\mathbf{y}^\star$



Realizations of $\mathbf{\Lambda} \sim \pi_{\mathbf{\Lambda}|\mathbf{y}^\star}(\boldsymbol{\lambda})$

Log-Gaussian Cox process with sparse observations ($d = 64^2$)





Metropolis-Hastings with independent proposals of $\mathfrak{T}^\sharp \pi$

A/R	**ESS (worst)**
72.6%	26.6%

Elliptic problem with unknown coefficients ($d = 2601$)

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 ($d = 2601$)

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

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

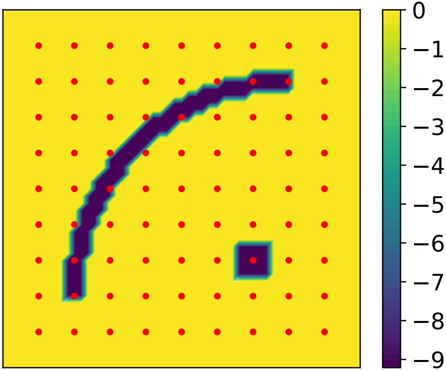
$$\kappa(\mathbf{x},\omega) = \exp\left(g(\mathbf{x},\omega)\right), \quad g(\mathbf{x},\omega) \sim \mathcal{N}\left(\mathbf{0}, C_g(\mathbf{x},\mathbf{x}')\right)$$

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

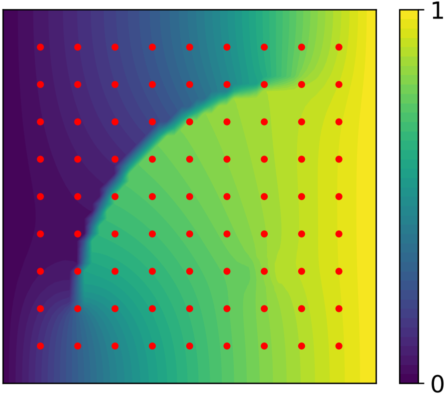
Bayesian inverse problem

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

Elliptic problem with unknown coefficients ($d = 2601$)

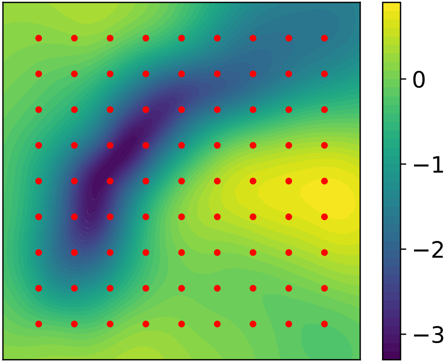


Synthetic field $\log \boldsymbol{\kappa}^\star(\mathbf{x})$

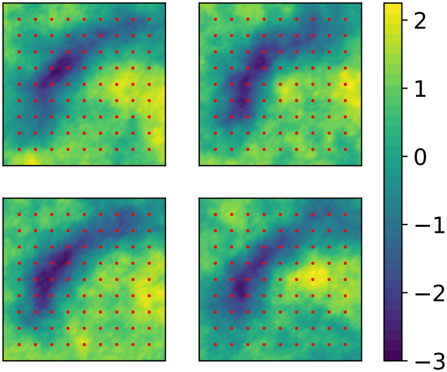


Synthetic solution $\mathbf{G}(\boldsymbol{\kappa}^\star)$

Elliptic problem with unknown coefficients ($d = 2601$)

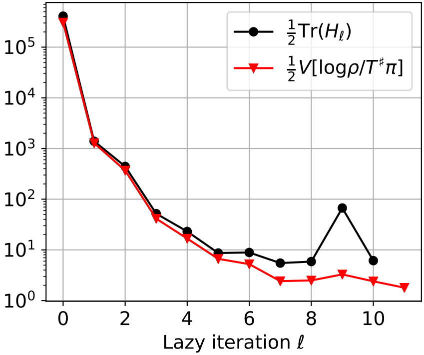


Posterior mean $\mathbb{E}\left[\log \boldsymbol{\kappa}(\mathbf{x}) | \mathbf{d}\right]$



Realizations of $\log \boldsymbol{\kappa}$

Elliptic problem with unknown coefficients ($d = 2601$)



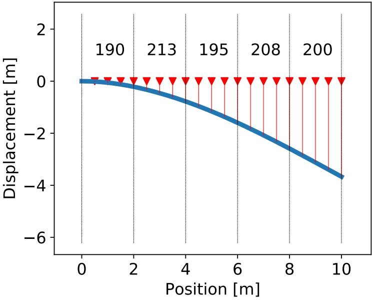
		ESS		
	A/R	**worst**	**best**	**avg.**
π	0.4%	$\sim 0\%$	$\sim 0\%$	$\sim 0\%$
$\mathfrak{T}^\sharp \pi$	28%	0.2%	1.6%	1.5%

Metropolis-Hastings with pCN proposal ($\beta = 0.5$)

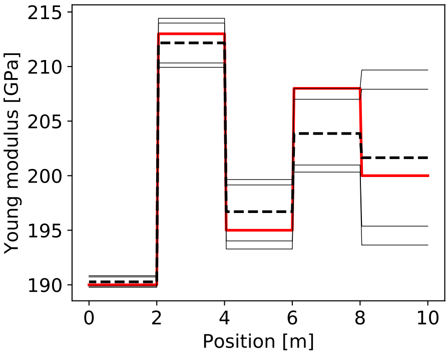
Estimation of the Young's modulus of a cantilever beam ($d = 5$)

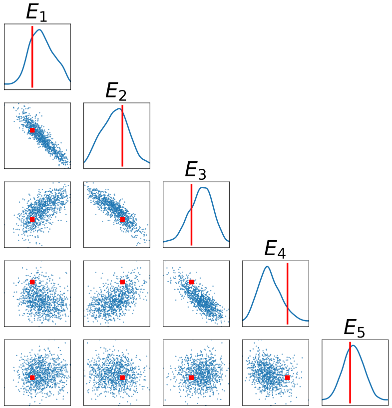
Forward model $\mathbf{G} : \boldsymbol{E} \mapsto \mathbf{u}$

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{E(x)}{2(1+\nu)}\left(\varphi(x) - \frac{\mathrm{d}}{\mathrm{d}x}u(x)\right)\right] = \frac{q(x)}{\kappa A} \,, \\ \frac{\mathrm{d}}{\mathrm{d}x}\left(E(x) I \frac{\mathrm{d}}{\mathrm{d}x}\varphi(x)\right) = \kappa A \frac{E(x)}{2(1+\nu)}\left(\varphi(x) - \frac{\mathrm{d}}{\mathrm{d}x}u(x)\right) \,. \end{cases}$$

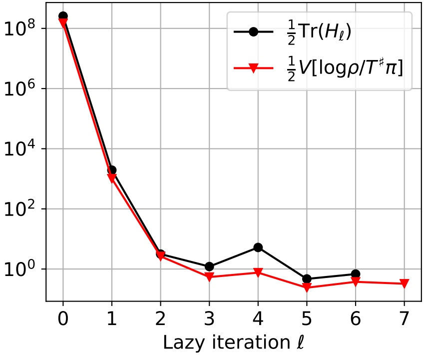


Estimation of the Young's modulus of a cantilever beam ($d = 5$)





Estimation of the Young's modulus of a cantilever beam ($d = 5$)



	ESS		
A/R	**worst**	**best**	**avg.**
68.3%	7.0%	38.7%	20.1%

Metropolis-Hastings with indipendent proposals

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`

Bigoni et al. "Greedy inference with layers of lazy maps" (arXiv)
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