TransportMaps for Statistical Computations

D. Bigoni, A. Spantini, R. Morrison, R. Baptista, Y. Marzouk
Massachusetts Institute of Technology, `http://transportmaps.mit.edu`

Abstract: *We develop a **statistical inference** framework based on the transportation of measures, where potentially intractable distributions of interest are mapped to tractable distributions. The framework allows for rich trade-offs of **complexity and accuracy**, and can be used for variational inference and statistical learning.*
*Many sources of low-dimensional structure can be exploited to design **scalable algorithms** within this framework.*

Motivation

Inference, particularly in the Bayesian setting, entails integration with respect to a complex **target** distribution ν_π. Choosing a tractable **reference** distribution ν_ρ, we solve the **transport problem**: find a map T that pushes forward ν_ρ to ν_π, denoted $T_\sharp \nu_\rho = \nu_\pi$. In other words, we seek:

$$T : \mathbb{R}^d \to \mathbb{R}^d \quad \text{s.t.} \quad \nu_\rho(A) = \nu_\pi(T(A)) \, , \quad \forall A \in \mathcal{B}(\mathbb{R}^d) \, .$$

This renders challenging integration problems tractable:

$$I[f] = \int f(\boldsymbol{x}) \nu_\pi(d\boldsymbol{x}) = \int f\left(T(\boldsymbol{x})\right) \nu_\rho(d\boldsymbol{x}) \, .$$

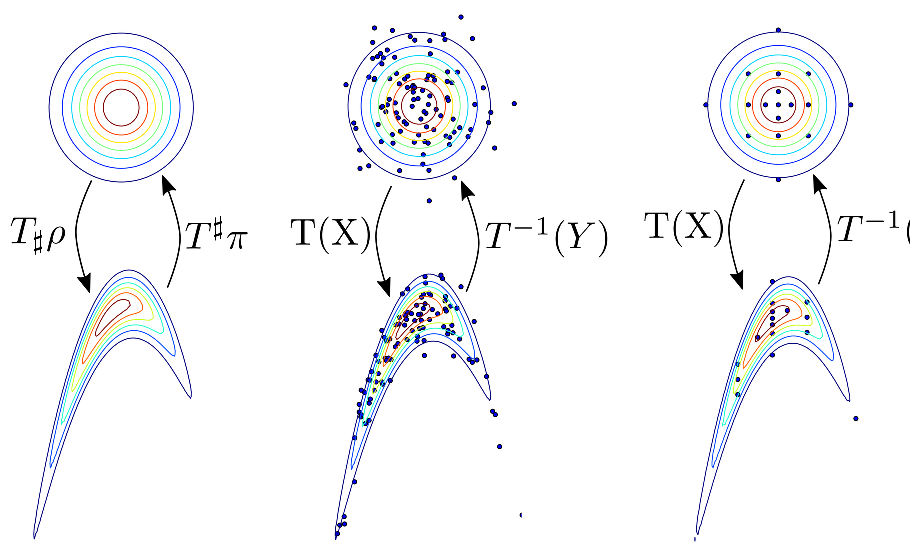
The transport map framework

Let ρ be the density of ν_ρ and π be the density of ν_π.
The map T s.t. $T_\sharp \nu_\rho = \nu_\pi$ defines the following identities:

Pushforward: $T_\sharp \rho(\boldsymbol{x}) = \rho \circ T^{-1}(\boldsymbol{x}) \, |\nabla T^{-1}(\boldsymbol{x})| = \pi \, ,$

Pullback: $T^\sharp \pi(\boldsymbol{x}) = \pi \circ T(\boldsymbol{x}) \, |\nabla T(\boldsymbol{x})| = \rho \, ,$

and for $\mathbf{X} \sim \nu_\rho$, $T(\mathbf{X}) \sim \nu_\pi$.

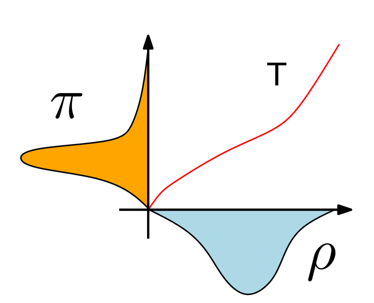


(a) densities (b) Monte Carlo (c) sparse grid

Knothe-Rosenblatt rearrangement

For any ν_ρ, ν_π absolutely continuous, there exists a **triangular monotone** map $T \in \mathcal{T}_>$ s.t. $T(d\nu_\rho) = d\nu_\pi$.

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



$$T^{(i)}(x_{1:i}) = c^{(i)}(x_{1:i-1}) + \int_0^{x_i} \left(h^{(i)}(x_{1:i-1}, t) \right)^2 + \varepsilon \, dt \, .$$

The transport problem is cast as a minimization of the **Kullback-Leibler divergence** [1, 2]:

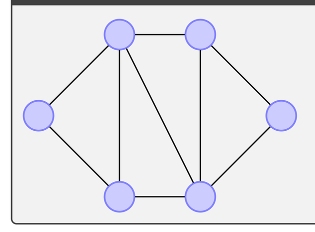
$$T^\star = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathcal{D}_{\mathrm{KL}}(T_\sharp \nu_\rho \| \nu_\pi) = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathbb{E}_\rho \left[-\log T^\sharp \tilde{\pi} \right]$$

The **variance diagnostic** is a global convergence criterion for the approximation of $T^\star$ in the space $\mathcal{T}_>$:

$$\mathcal{D}_{\mathrm{KL}}(T_\sharp \nu_\rho \| \nu_\pi) \approx \frac{1}{2} \mathbb{V}\mathrm{ar} \left[\log \frac{\rho}{T^\sharp \tilde{\pi}} \right] \quad \text{as} \quad T \to T^\star$$

The low-dimensional structure of transports

Conditional independence relationships [3, 4]



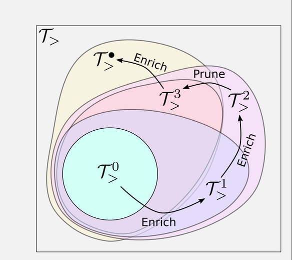
From the Markov structure of π, we prove the existence of transports that are **decomposable** and/or **sparse.**

Smoothness and marginal independence

We seek $\mathcal{T}_>^\bullet \subset \mathcal{T}_>$ $(\dim \mathcal{T}_>^\bullet < \infty)$ s.t.

$$\exists \, T \in \mathcal{T}_>^\bullet \quad \text{s.t.} \quad \mathbb{V}\left[\log \rho / T^\sharp \pi \right] < \varepsilon_\bullet$$

- **First variation** for enrichment
- **Sample-average approximation**

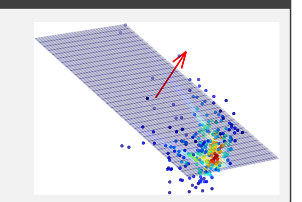


Low rank (e.g., likelihood informed subspaces)

For $\rho = \mathcal{N}(0, \mathbf{I})$ and a **rotation** U s.t.

$$U^\sharp \pi(\boldsymbol{x}) = \mu(\boldsymbol{x}_{1:\kappa}) \eta(\boldsymbol{x}_{\kappa+1:d}) \, , \qquad \eta = \mathcal{N}(0, \mathbf{I}) \, ,$$

$\exists$ a κ-dimensional map T s.t. $(U \circ T)_\sharp \rho = \pi$

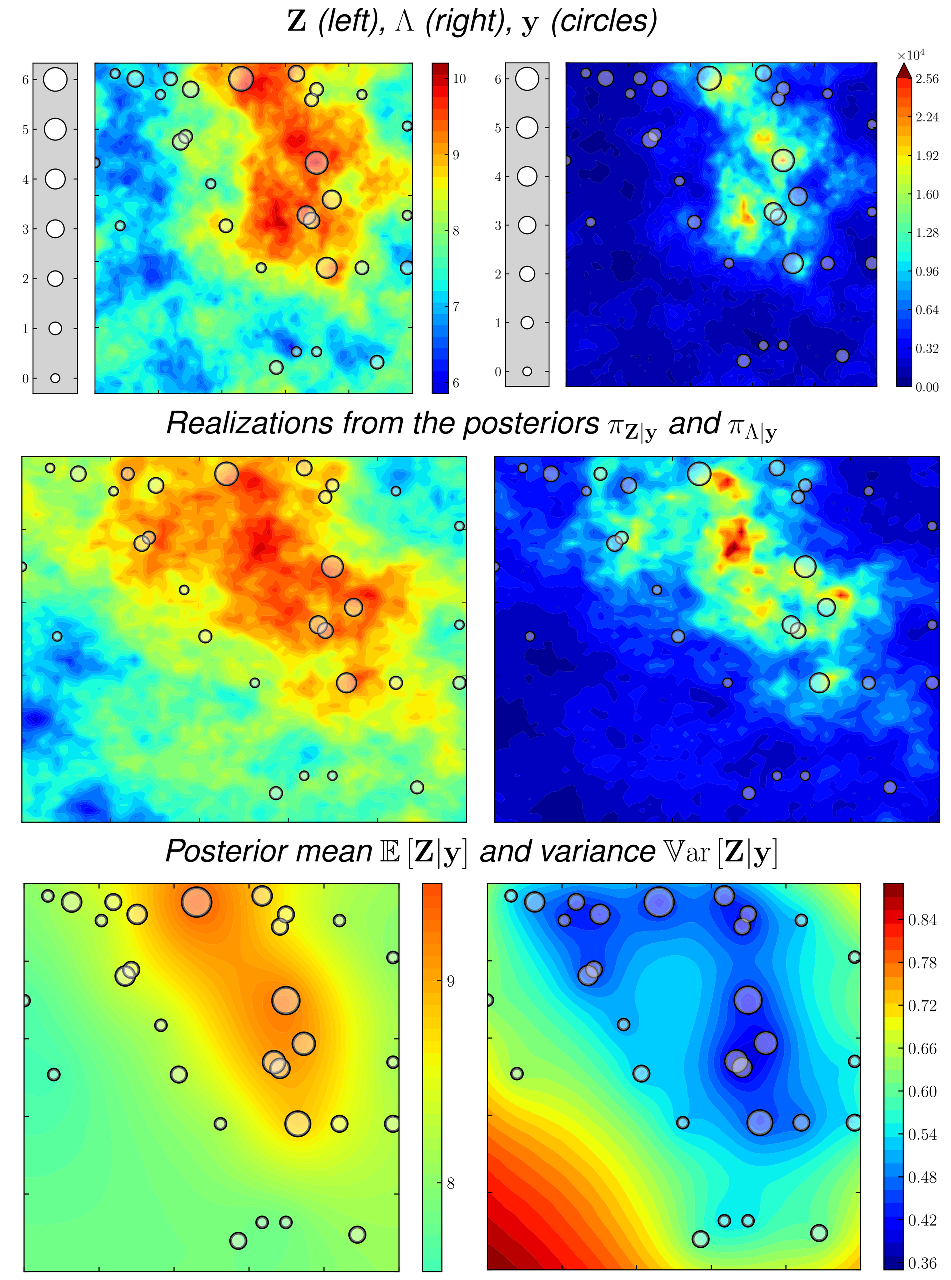


Multilevel, multi-fidelity, multi-scale preconditioning

For a sequence of increasingly accurate and expensive distributions $\pi_1, \ldots, \pi_\ell$, we use the **hierarchy** of problems $(T_i)_\sharp \rho = \mathfrak{T}_{i-1}^\sharp \pi_i$, $\mathfrak{T}_k = T_1 \circ \ldots \circ T_k$, to obtain $(\mathfrak{T}_\ell)_\sharp \rho = \pi_\ell$.

Low-rank distributed parameters

$\mathbf{Z}$ (left), Λ (right), $\mathbf{y}$ (circles)



Realizations from the posteriors $\pi_{\mathbf{Z}|\mathbf{y}}$ and $\pi_{\Lambda|\mathbf{y}}$

Posterior mean $\mathbb{E}[\mathbf{Z}|\mathbf{y}]$ and variance $\mathbb{V}\mathrm{ar}[\mathbf{Z}|\mathbf{y}]$

Consider the log-Gaussian Cox process $(\mathbf{Y}_i)_{i=1}^n$, discretized on an $n = 64 \times 64$ grid with cell centers $\mathbf{s}_i$:

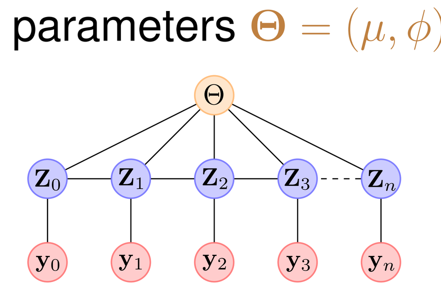
$$\mathbf{Y}_i \sim \text{Poisson}(\boldsymbol{\Lambda}_i) \, , \quad \mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \boldsymbol{\Lambda} \, , \quad \boldsymbol{\Lambda} = \exp \mathbf{Z}$$

$$\mathbf{Z} \sim \mathcal{N}(2 \log 64, \boldsymbol{\Sigma}) \, , \quad \boldsymbol{\Sigma}_{ij} = \exp \left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2}{\ell} \right)$$

- Given observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d)$ we characterize the posterior distribution
$$\pi_{\mathbf{Z}|\boldsymbol{y}}(\boldsymbol{z}) \propto \pi_{\boldsymbol{y}|\mathbf{Z}}(\boldsymbol{z}) \pi_{\mathbf{Z}}(\boldsymbol{z})$$
- The likelihood $\pi_{\mathbf{y}|\mathbf{Z}}(\mathbf{z}) = g(\mathbf{z}_1, \ldots, \mathbf{z}_d)$ is local. Thus a "prior map"
$$T_{\mathrm{pr}}(\boldsymbol{x}) = \begin{Bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{Bmatrix} + \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{12} & \mathbf{L}_{22} \end{bmatrix} \boldsymbol{x} \, , \quad \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top \, ,$$
yields $T_{\mathrm{pr}}^\sharp \pi_{\mathbf{Z}|\boldsymbol{y}}(\boldsymbol{z}) = g(\boldsymbol{z}_{1:d}) \eta(\boldsymbol{z})$, where $\eta = \mathcal{N}(0, \mathbf{I})$.
- In other words, $T_{\mathrm{pr}}^\sharp \pi_{\mathbf{Z}|\mathbf{y}}$ departs from $\mathcal{N}(0, \mathbf{I})$ **in a** $d \ll n$ **dimensional subspace**.
- Hence a d-dimensional map $T = [\tau(\boldsymbol{x}_{1:d}), \mathrm{Id}]$ completes the inference, $(T_{\mathrm{pr}} \circ T)_\sharp \rho = \pi_{\mathbf{Z}|\mathbf{y}}$.

State-space models

- Characterize the **joint state + parameter posterior** distribution of a dynamical model, $\Theta, \mathbf{Z}_{1:N} | \mathbf{y}_{1:N} \sim \pi$
- Illustration: stochastic volatility (SV) model with parameters $\Theta = (\mu, \phi)$:



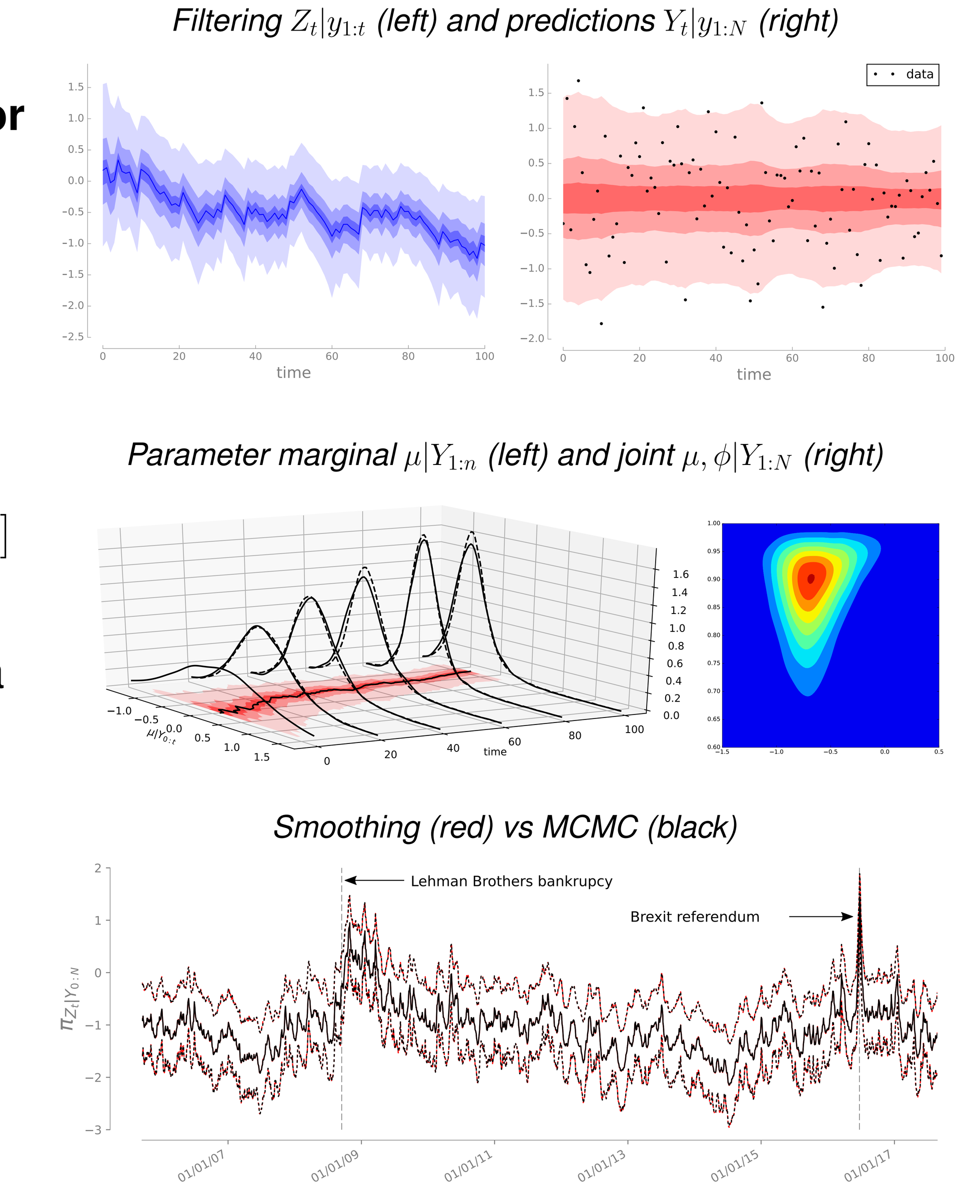
$$Z_{n+1} = \mu + \phi(Z_n - \mu) + \eta_n \, ,$$
$$Y_n = \varepsilon_n \exp(Z_n / 2) \, ,$$
$$\mu, \phi, Z_1 \sim \pi_{\mathrm{prior}}$$
$$\eta_n, \, \varepsilon_n \sim \pi_{\mathrm{noise}} \, , \quad n = [1, \ldots, N]$$

- The **conditional independence** structure of the $(N \cdot d_{\mathbf{Z}} + d_\Theta)$-dimensional posterior π guarantees a decomposable representation:
$$(T_1 \circ T_2 \circ \ldots \circ T_{N-1})_\sharp \, \rho = \pi$$
where each T_i is a map of (effectively) $2 \cdot d_{\mathbf{Z}} + d_\Theta$ dimensions
- For the SV model, we thus solve $(N-1)$ 4-dimensional problems
- Result: a fully Bayesian variational algorithm for nonlinear **smoothing** and **sequential** parameter inference

Filtering $Z_t | y_{1:t}$ (left) and predictions $Y_t | y_{1:N}$ (right)

Parameter marginal $\mu | Y_{1:n}$ (left) and joint $\mu, \phi | Y_{1:N}$ (right)

Smoothing (red) vs MCMC (black)



Learning non-Gaussian graphical models

Generalized precision

For continuous, non-Gaussian multivariate distributions, conditional independence between two variables can be revealed as

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{V \setminus \{ij\}} \iff \partial_i \partial_j \log \pi(\boldsymbol{x}) = 0 \; \forall \boldsymbol{x} \in \mathbb{R}^d$$

Let Ω be the **generalized precision** matrix:

$$\Omega_{ij} = \mathbb{E}_\pi \left\{ |\partial_i \partial_j \log \pi| \right\}$$

The non-zero entries of Ω correspond exactly to the Markov structure of π. In this way, the matrix is a generalization of the precision matrix in the Gaussian setting, which likewise encodes **conditional independence**.

Objective: Given a data set (from some unknown distribution π) $\{\mathbf{x}^{(i)}\}, i = 1, \ldots, n$, find the Markov structure of π

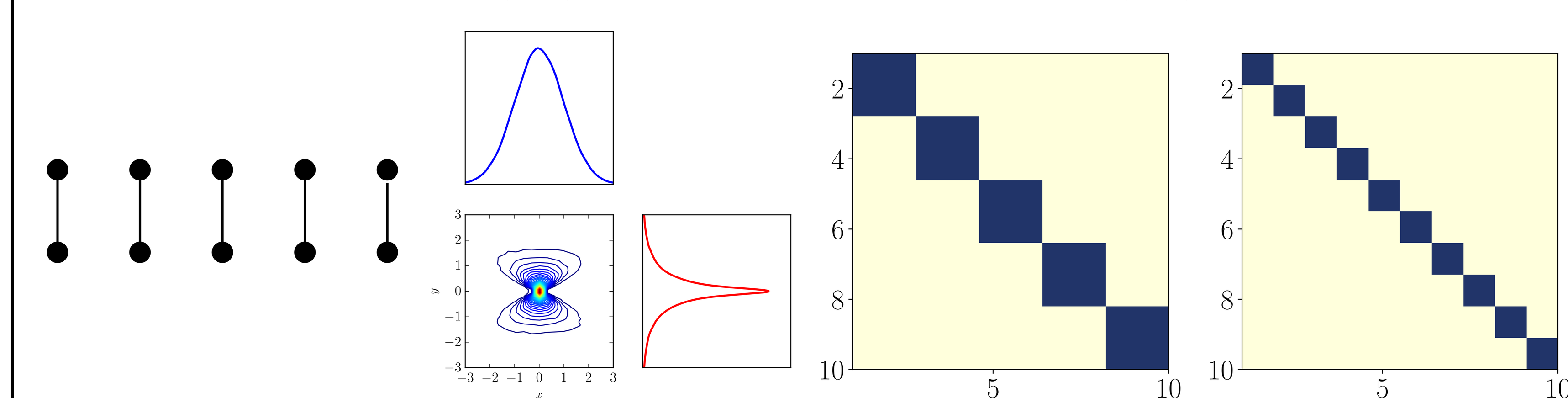
Proposed approach:

- Use transport maps to represent and estimate the density
- Given density, compute score of conditional independence via Ω
- Threshold scores to identify (sparse) graph
- Reorder variables to approximate natural ordering
- Recompute map, given candidate graphical structure. Iterate...

Algorithm: Sparsity Identification in Non-Gaussian distributions (SING)

Numerical example: Take pairs of random variables (X, Y):

$$X \sim \mathcal{N}(0, 1) \, , \quad Y = WX \, , \quad \text{where} \quad W \sim \mathcal{N}(0, 1)$$



From the left: the undirected graphical model; one- and two-dimensional marginal distributions for one pair (X, Y); adjacency matrix of true graph, as found with SING; result assuming data is normal.

Inverse stochastic PDEs

Consider the Electrical Impedance Tomography model $\mathcal{F} : \sigma(\mathbf{x}) \mapsto \mathbf{V}$ that relates the impedance $\sigma(\mathbf{x})$ to Voltage observations $\mathbf{V}$ at $s = 1, \ldots, S$ boundary sensors

$$V_{sk} = \int_{\partial\Omega} \frac{u_k(\mathbf{x})}{C_s} \exp\left(-\frac{(\mathbf{c}_s - \mathbf{x})^2}{2\varepsilon^2} \right) d\mathbf{x} \, ,$$

$$\begin{cases} -\nabla \cdot \sigma(\mathbf{x}) \nabla u_k(\mathbf{x}) = 0 \, , & \mathbf{x} \in \Omega \\ \sigma(\mathbf{x}) \nabla u_k(\mathbf{x}) = I_k(\mathbf{x}) \, , & \mathbf{x} \in \partial\Omega \end{cases}$$

for $k = 1, \ldots, K$ current patterns.
We characterize

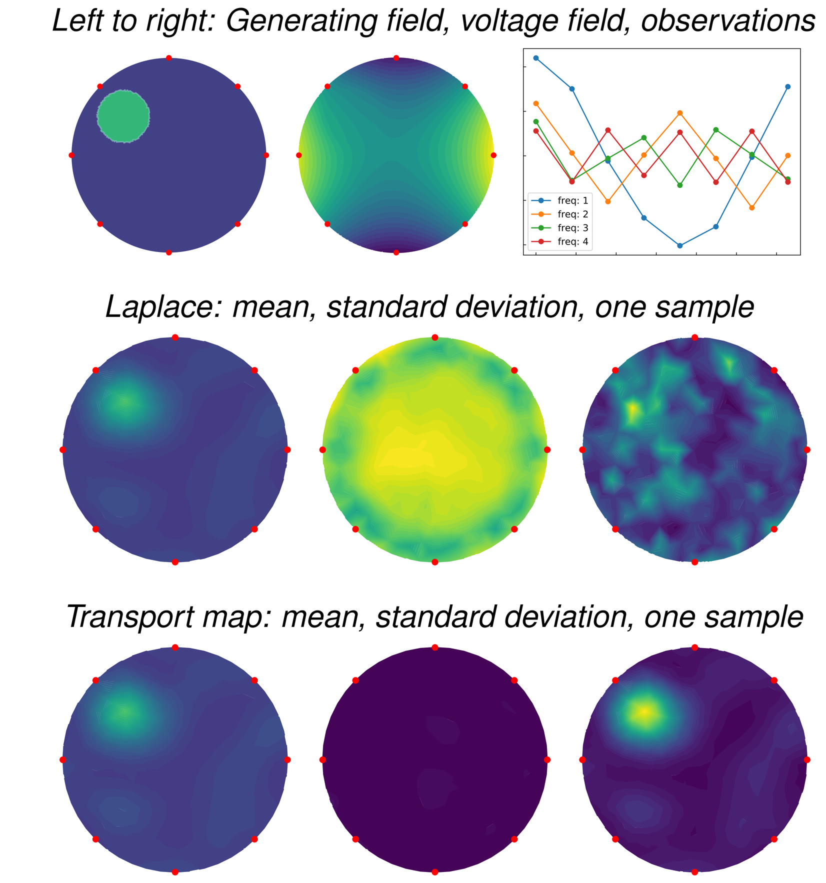
$$\pi(\sigma | \mathbf{d}) \propto \pi_{\mathrm{noise}} \left(\mathcal{F}(\sigma) - \mathbf{d} \right) \pi_{\mathrm{prior}}(\sigma)$$
$$\mathbf{d} = \mathbf{V} + \varepsilon \, , \quad \varepsilon \sim \pi_{\mathrm{noise}} = \mathcal{N}(\mathbf{0}, \delta\mathbf{I})$$
$$\sigma \sim \pi_{\mathrm{prior}} = \log \mathcal{GP}(K_{OU})$$

We **adaptively** construct $T_\sharp \rho(\sigma) \approx \pi(\sigma | \mathbf{d})$.

Left to right: Generating field, voltage field, observations

Laplace: mean, standard deviation, one sample

Transport map: mean, standard deviation, one sample



`TransportMaps` provides algorithms for the characterization of probability measures via **deterministic couplings** and **optimization**, exploiting **low-rank** structure, **conditional independence**, **smoothness** and **marginal independence**

[1] T. Moselhy and Y. Marzouk. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (2012). [2] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification.* Ed. by R. G. Ghanem, D. Higdon, and H. Owhadi. Springer International Publishing, 2016. arXiv: 1602.05023. [3] A. Spantini, D. Bigoni, and Y. Marzouk. "Inference via low-dimensional couplings". In: *ArXiv e-prints* (2017). arXiv: 1703.06131 [stat.ME]. [4] R. Morrison, R. Baptista, and Y. Marzouk. "Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting". In: *Advances in Neural Information Processing Systems (NIPS) 30.* 2017.