Dimension reduction in nonlinear Bayesian inverse problems

Youssef Marzouk
joint work with Daniele Bigoni, Michael Brennan, Tiangang Cui, Kody Law, Alessio Spantini, Olivier Zahm

Department of Aeronautics and Astronautics
Center for Computational Science and Engineering
Statistics and Data Science Center

Massachusetts Institute of Technology
http://uqgroup.mit.edu

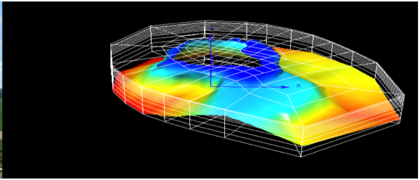
Support from AFOSR, DOE, NSF, ONR

2 November 2021

Motivation: inverse problems in the Bayesian setting

Observations y

Parameters x

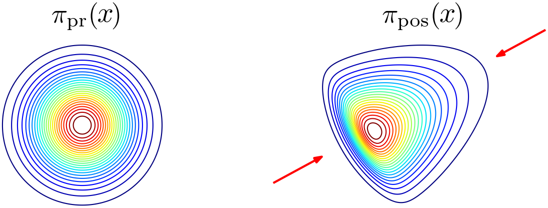


$$\pi_{\text{pos}}(x) := \underbrace{\pi(x|y) \propto \mathcal{L}_y(x)\, \pi_{\text{pr}}(x)}_{\textbf{Bayes' rule}}$$

- Characterize the posterior distribution (density π_{pos})
- This is a challenging task since:
  - $x \in \mathbb{R}^d$ is typically **high-dimensional** (e.g., a discretized function)
  - π_{pos} is **non-Gaussian**
  - evaluations of the likelihood (hence π_{pos}) may be **expensive**
- π_{pos} can be evaluated up to a normalizing constant

A conjecture

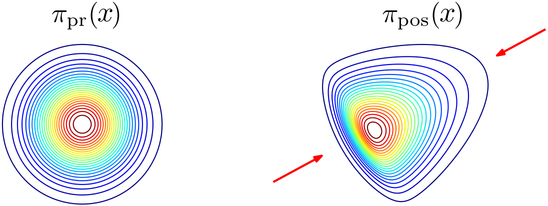
In many situations, the data are “informative” only on a low-dimensional subspace



$$\mathbb{R}^d = \underbrace{X_r}_{\pi_{\mathrm{pos}} \neq \pi_{\mathrm{pr}}} + \underbrace{X_\perp}_{\pi_{\mathrm{pos}} \approx \pi_{\mathrm{pr}}}$$

A conjecture

In many situations, the data are "informative" only on a low-dimensional subspace



This structure is now well understood in the **linear–Gaussian case**,
$x \sim N(0, \Sigma_{\mathrm{pr}})$, $y|x \sim N(Gx, \Sigma_{\mathrm{obs}})$ [Spantini et al. 2015]:

- *Optimal* approximations of the posterior covariance as a low-rank update of the prior, $\widetilde{\Sigma}_{\mathrm{pos}} = \Sigma_{\mathrm{pr}} - K_r K_r^\top$, for any $r \leq d$
- Optimal posterior mean approximations, $\widetilde{\mu}_{\mathrm{pos}} = A_r y$
- Central role of generalized eigenproblems, e.g., $\left(G^T \Sigma_{\mathrm{obs}}^{-1} G, \Sigma_{\mathrm{pr}}^{-1} \right)$

Low effective dimensionality of Bayesian inverse problems

More general idea: the posterior distribution can be well approximated by

$$\widetilde{\pi}_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\, \pi_{\mathrm{pr}}(x)$$

for some **positive function** $\widetilde{\mathcal{L}}$ and rank r **linear projector** $P_r \in \mathbb{R}^{d\times d}$

Low effective dimensionality of Bayesian inverse problems

More general idea: the posterior distribution can be well approximated by

$$\widetilde{\pi}_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\, \pi_{\text{pr}}(x)$$

for some **positive function** $\widetilde{\mathcal{L}}$ and rank r **linear projector** $P_r \in \mathbb{R}^{d\times d}$

P_r induces a decomposition of the space

$$x = x_r + x_\perp \qquad \begin{cases} x_r & \in \ \text{Im}(P_r) \\ x_\perp & \in \ \text{Ker}(P_r) \end{cases}$$

By construction, $x \mapsto \widetilde{\mathcal{L}}(P_r x) = \widetilde{\mathcal{L}}(x_r)$ is only a function of $x_r \in \text{Im}(P_r) \equiv \mathbb{R}^r$.

Low effective dimensionality of Bayesian inverse problems

More general idea: the posterior distribution can be well approximated by

$$\widetilde{\pi}_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\, \pi_{\text{pr}}(x)$$

for some **positive function** $\widetilde{\mathcal{L}}$ and rank r **linear projector** $P_r \in \mathbb{R}^{d \times d}$

P_r induces a decomposition of the space

$$x = x_r + x_\perp \qquad \begin{cases} x_r & \in \ \text{Im}(P_r) \\ x_\perp & \in \ \text{Ker}(P_r) \end{cases}$$

By construction, $x \mapsto \widetilde{\mathcal{L}}(P_r x) = \widetilde{\mathcal{L}}(x_r)$ is only a function of $x_r \in \text{Im}(P_r) \equiv \mathbb{R}^r$.

If $r \ll d$, we can:

- Design **structure-exploiting** MCMC algorithms to sample from π_{pos} (e.g., DILI samplers [Cui, Law, M 2016])
- More easily build surrogates (i.e., **fast approximations**) of $x_r \mapsto \widetilde{\mathcal{L}}(x_r)$
- Develop tractable **variational characterizations** of the posterior (second part of this talk)

Many previous proposals

- P_r can be defined as a projector onto the **dominant eigenspace** of a matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ which contains "relevant information"

Many previous proposals

- P_r can be defined as a projector onto the **dominant eigenspace** of a matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ which contains "relevant information"
  - **Likelihood-informed subspace (LIS)** [Cui et al. 2014]

$$\mathbf{H}_{\mathrm{LIS}} = \int (\nabla G)^T \Sigma_{\mathrm{obs}}^{-1} (\nabla G) \, \mathrm{d}\pi_{\mathrm{pos}}$$

where $\mathcal{L}_y$ follows from $y \sim \mathcal{N}(G(x), \Sigma_{\mathrm{obs}})$

  - **Active subspace (AS)** [Constantine et al. 2015]

$$\mathbf{H}_{\mathrm{AS}} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \, \mathrm{d}\pi_{\mathrm{pr}}$$

Many previous proposals

- P_r can be defined as a projector onto the **dominant eigenspace** of a matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ which contains "relevant information"
  - **Likelihood-informed subspace (LIS)** [Cui et al. 2014]

$$\mathbf{H}_{\text{LIS}} = \int (\nabla G)^T \Sigma_{\text{obs}}^{-1} (\nabla G) \, \mathrm{d}\pi_{\text{pos}}$$

where $\mathcal{L}_y$ follows from $y \sim \mathcal{N}(G(x), \Sigma_{\text{obs}})$

  - **Active subspace (AS)** [Constantine et al. 2015]

$$\mathbf{H}_{\text{AS}} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \, \mathrm{d}\pi_{\text{pr}}$$

- Similarly, various definitions of $\widetilde{\mathcal{L}}$:
  - (LIS) Fix complementary parameters $\widetilde{\mathcal{L}}(P_r x) = \mathcal{L}_y(P_r x + (I - P_r) m_0)$
  - (AS) Take conditional expectation of the log-likelihood

$$\widetilde{\mathcal{L}}(P_r x) = \exp \mathbb{E}_{\pi_{\text{pr}}}(\log \mathcal{L}_y | P_r x)$$

Broad objective

Build an approximation of π_{pos} of the form

$$\widetilde{\pi}_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\text{pr}}(x) \qquad \text{with} \begin{cases} \widetilde{\mathcal{L}} : \mathbb{R}^d \to \mathbb{R}^+ \\ P_r \in \mathbb{R}^{d\times d} \text{ rank-}r \text{ projector} \end{cases}$$

such that

$$\boxed{D_{\text{KL}}(\pi_{\text{pos}} || \widetilde{\pi}_{\text{pos}}) \leq \varepsilon}$$

with $r = r(\varepsilon)$ much smaller than d.

See full details in [ZCLSM 21].

Decomposition of the error

A "Pythagorean" theorem

For any P_r and $\widetilde{\mathcal{L}}$ we have

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}\middle|\middle|\widetilde{\pi}_{\mathrm{pos}}\right) = \underbrace{D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}\middle|\middle|\pi^*_{\mathrm{pos}}\right)}_{=\,\mathrm{function}(P_r)} + \underbrace{D_{\mathrm{KL}}\left(\pi^*_{\mathrm{pos}}\middle|\middle|\widetilde{\pi}_{\mathrm{pos}}\right)}_{=\,\mathrm{function}(P_r,\widetilde{\mathcal{L}})}$$

where

$$\pi^*_{\mathrm{pos}}(x) \propto \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)\pi_{\mathrm{pr}}(x)$$

Decomposition of the error

A "Pythagorean" theorem

For any P_r and $\widetilde{\mathcal{L}}$ we have

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}\middle|\middle|\widetilde{\pi}_{\mathrm{pos}}\right) = \underbrace{D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}\middle|\middle|\pi^*_{\mathrm{pos}}\right)}_{=\,\mathrm{function}(P_r)} + \underbrace{D_{\mathrm{KL}}\left(\pi^*_{\mathrm{pos}}\middle|\middle|\widetilde{\pi}_{\mathrm{pos}}\right)}_{=\,\mathrm{function}(P_r,\widetilde{\mathcal{L}})}$$

where

$$\pi^*_{\mathrm{pos}}(x) \propto \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)\pi_{\mathrm{pr}}(x)$$

This allows decoupling the construction of $\widetilde{\mathcal{L}}$ and P_r.

- Given P_r, the function $\widetilde{\mathcal{L}}$ such that $\widetilde{\mathcal{L}}(P_r x) = \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)$ yields
$$D_{\mathrm{KL}}\left(\pi^*_{\mathrm{pos}}\middle|\middle|\widetilde{\pi}_{\mathrm{pos}}\right) = 0$$
- How to construct P_r such that
$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}\middle|\middle|\pi^*_{\mathrm{pos}}\right) \le \varepsilon$$
with a rank $r \ll d$?

Constructing the projector P_r

Assumption on the prior distribution

There exist functions V and Ψ such that

$$\pi_{\text{pr}}(x) \propto \exp\big(- V(x) - \Psi(x)\big) \quad \text{with} \quad \begin{cases} \nabla^2 V \succeq \Gamma \\ \exp(\sup \Psi - \inf \Psi) \leq \kappa \end{cases}$$

for some SPD matrix $\Gamma \in \mathbb{R}^{d\times d}$ and some $\kappa \geq 1$.

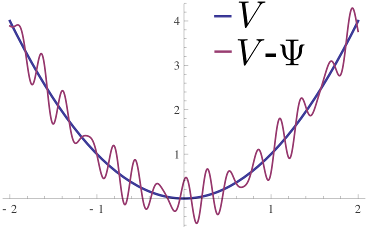
Constructing the projector P_r

Assumption on the prior distribution

There exist functions V and Ψ such that

$$\pi_{\mathrm{pr}}(x) \propto \exp\big(-V(x) - \Psi(x)\big) \quad \text{with} \quad \begin{cases} \nabla^2 V \succeq \Gamma \\ \exp(\sup \Psi - \inf \Psi) \leq \kappa \end{cases}$$

for some SPD matrix $\Gamma \in \mathbb{R}^{d\times d}$ and some $\kappa \geq 1$.



- Gaussian prior satisfies this assumption with $\Gamma = \Sigma_{\mathrm{pr}}^{-1}$ and $\kappa = 1$
- Gaussian mixture $\pi_{\mathrm{pr}} \propto \sum_i \mathcal{N}(\mu_i, \Sigma_i)$ also satisfies this assumption
- Uniform prior on convex bounded domain also allowed [ZCLSM21]

Constructing the projector P_r

Based on this assumption, π_{pr} satisfies the **logarithmic Sobolev inequality**

$$\int h^2 \log \frac{h^2}{\int h^2 \, \mathrm{d}\pi_{\text{pr}}} \, \mathrm{d}\pi_{\text{pr}} \leq 2\kappa \int \|\nabla h\|^2_{\Gamma^{-1}} \mathrm{d}\pi_{\text{pr}}$$

for any function h with sufficient regularity.

- Putting $h^2 = \mathcal{L}_y / \int \mathcal{L}_y \, \mathrm{d}\pi_{\text{pr}}$ bounds the KL divergence from prior to posterior:

$$\mathcal{D}_{\text{KL}}(\pi_{\text{pos}} || \pi_{\text{pr}}) \leq \frac{\kappa}{2} \int \|\nabla \log \mathcal{L}_y\|^2_{\Gamma^{-1}} \, \mathrm{d}\pi_{\text{pos}}$$

Constructing the projector P_r

Proposition: subspace logarithmic Sobolev inequality

π_{pr} also satisfies

$$\int h^2 \log \frac{h^2}{\mathbb{E}(h^2|P_r x)} \, \mathrm{d}\pi_{\text{pr}} \leq 2\kappa \int \|(I_d - P_r^T)\nabla h\|^2_{\Gamma^{-1}} \, \mathrm{d}\pi_{\text{pr}}$$

for any function h with sufficient regularity and any projector P_r.

Constructing the projector P_r

Proposition: subspace logarithmic Sobolev inequality

π_{pr} also satisfies

$$\int h^2 \log \frac{h^2}{\mathbb{E}(h^2|P_r x)} \, \mathrm{d}\pi_{\text{pr}} \leq 2\kappa \int \|(I_d - P_r^T)\nabla h\|_{\Gamma^{-1}}^2 \, \mathrm{d}\pi_{\text{pr}}$$

for any function h with sufficient regularity and any projector P_r.

Corollary

For any projector P_r we have

$$D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi_{\text{pos}}^*\right) \leq \frac{\kappa}{2} \mathcal{R}_{\pi_{\text{pos}}}(P_r)$$

where

$$\mathcal{R}_{\pi_{\text{pos}}}(P_r) = \int \|(I_d - P_r^T)\nabla \log \mathcal{L}_y\|_{\Gamma^{-1}}^2 \, \mathrm{d}\pi_{\text{pos}}$$

Constructing the projector P_r

Corollary

For any projector P_r we have

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \pi^*_{\mathrm{pos}}\right) \leq \frac{\kappa}{2} \mathcal{R}_{\pi_{\mathrm{pos}}}(P_r)$$

where

$$\mathcal{R}_{\pi_{\mathrm{pos}}}(P_r) = \int \|(I_d - P_r^T)\nabla \log \mathcal{L}_y\|^2_{\Gamma^{-1}} \, \mathrm{d}\pi_{\mathrm{pos}}$$

Constructing the projector P_r

Corollary

For any projector P_r we have

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle\| \pi^*_{\mathrm{pos}}\right) \leq \frac{\kappa}{2} \mathcal{R}_{\pi_{\mathrm{pos}}}(P_r)$$

where

$$\mathcal{R}_{\pi_{\mathrm{pos}}}(P_r) = \int \|(I_d - P_r^T)\nabla \log \mathcal{L}_y\|^2_{\Gamma^{-1}} \, \mathrm{d}\pi_{\mathrm{pos}}$$

Finding P_r that **minimizes** this bound corresponds to **PCA** of $\nabla \log \mathcal{L}_y(X)$.

- For a fixed r, the minimizer P_r^* of the **reconstruction error** $\mathcal{R}_{\pi_{\mathrm{pos}}}(P_r)$ is the Γ-orthogonal projector onto the dominant generalized eigenspace of

$$\mathbf{H} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \, \mathrm{d}\pi_{\mathrm{pos}}$$

- Furthermore, we have $\mathcal{R}_{\pi_{\mathrm{pos}}}(P_r^*) = \sum_{i>r} \lambda_i$, where λ_i is the i-th generalized eigenvalue of $(\mathbf{H}, \Gamma)$

An idealized algorithm

1. Compute
$$\mathbf{H} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \, \mathrm{d}\pi_{\mathrm{pos}}$$
2. Define P_r as the projector on the dominant eigenspace of $\mathbf{H}$
3. Compute the conditional expectation
$$\widetilde{\mathcal{L}}(P_r x) = \mathbb{E}_{\mathrm{pr}}(\mathcal{L}_y | P_r x)$$

Then $\pi^*_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\mathrm{pr}}(x)$ satisfies

$$\boxed{D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \pi^*_{\mathrm{pos}}\right) \leq \frac{\kappa}{2} \sum_{i>r} \lambda_i}$$

- At step 2, we can choose the rank $r = r(\varepsilon)$ of P_r such that
$$D_{\mathrm{KL}}(\pi_{\mathrm{pos}} || \pi^*_{\mathrm{pos}}) \leq \varepsilon$$
- A strong decay in λ_i implies $r(\varepsilon) \ll d$

An idealized algorithm

1. Compute

$$\mathbf{H} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \; d\pi_{\text{pos}}$$

2. Define P_r as the projector on the dominant eigenspace of $\mathbf{H}$

3. Compute the conditional expectation

$$\widetilde{\mathcal{L}}(P_r x) = \mathbb{E}_{\text{pr}}(\mathcal{L}_y | P_r x)$$

Practical issues

- Evaluating $\mathbf{H}$ requires computing an integral over the posterior
- Computing the conditional expectation requires some effort

Sample approximations

1. Monte Carlo approximation of $\mathbf{H}$:

$$\mathbf{H} \approx \widehat{\mathbf{H}}_K := \frac{1}{K} \sum_{i=1}^{K} \nabla \log \mathcal{L}_y(X_i) \otimes \nabla \log \mathcal{L}_y(X_i) \quad \text{with} \quad X_i \overset{\text{iid}}{\sim} \pi_{\text{pos}}$$

Proposition

Under some assumptions, **quasi-optimal projectors** are obtained with high probability $1 - \delta$ if

$$K \geq \mathcal{O}\left(\sqrt{\text{rank}(H)} + \sqrt{\log(2\delta^{-1})}\right)^2$$

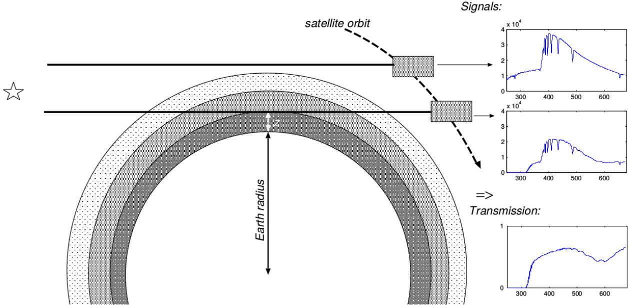
- Key assumption: $\nabla \log \mathcal{L}_y(X)$ is *sub-Gaussian*, for $X \sim \pi_{\text{pos}}$

2. Sample approximations of the conditional expectation $\mathbb{E}_{\text{pr}}(\mathcal{L}_y | P_r x)$
   - Error controlled by same factors; details in [ZCLSM21]

Example: GOMOS atmospheric remote sensing [Tamminen 2004]

- Estimate gas densities $x = \varrho^{\text{gas}}(z)$ from transmission spectra $y_\omega(z)$
- Beer's law:

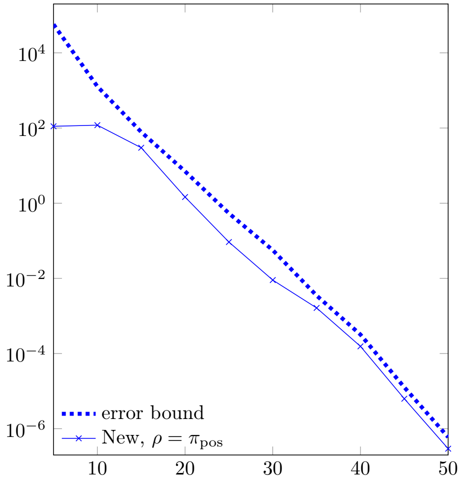
$$y_\omega(z) = \exp\left(- \int_{\text{light path}} \sum_{\text{gas}} \alpha_\omega^{\text{gas}}(z(\zeta))\, \varrho^{\text{gas}}(z(\zeta))\, \mathrm{d}\zeta \right) + \xi$$



- Gaussian prior $\mathcal{N}(\mu_{\text{pr}}, \Sigma_{\text{pr}})$ (hence $\Gamma = \Sigma_{\text{pr}}^{-1}$ and $\kappa = 1$)
- After discretization of the atmosphere, $\dim(x) = 200$

GOMOS: results

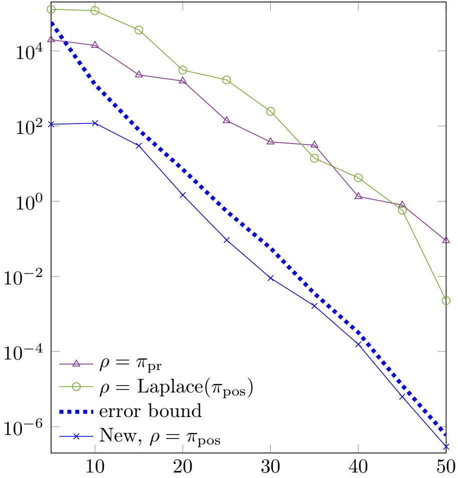
$D_{\mathrm{KL}}(\pi_{\mathrm{pos}} || \widetilde{\pi}_{\mathrm{pos}}) = \mathit{function}(r)$



$$\mathbf{H} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \; \mathrm{d}\pi_{\mathrm{pos}}$$

GOMOS: results

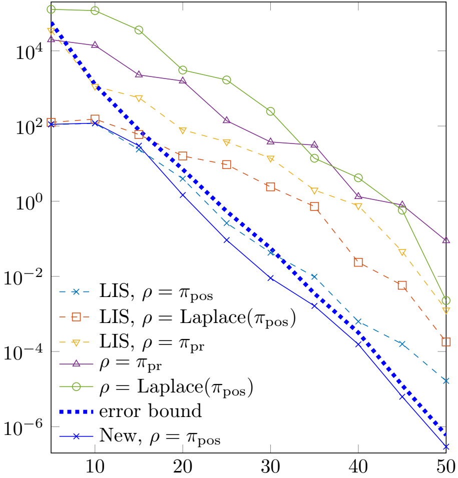
$D_{\mathrm{KL}}(\pi_{\mathrm{pos}} || \widetilde{\pi}_{\mathrm{pos}}) = \mathit{function}(r)$



$$\mathbf{H}^{(\rho)} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \; \mathrm{d}\rho$$

GOMOS: results

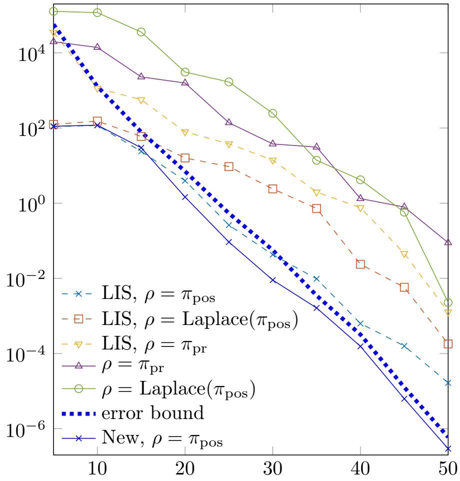
$D_{KL}(\pi_{pos}||\widetilde{\pi}_{pos}) = function(r)$



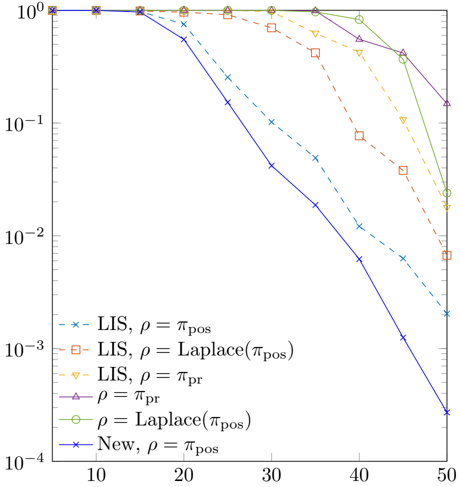
$$\mathbf{H}^{(\rho)} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \; d\rho \qquad \mathbf{H}^{(\rho)}_{LIS} = \int (\nabla G)^T \Gamma_{obs}^{-1} (\nabla G) \; d\rho$$

GOMOS: results

$D_{\mathrm{KL}}(\pi_{\mathrm{pos}} || \widetilde{\pi}_{\mathrm{pos}}) = \mathit{function}(r)$



$d_{\mathrm{Hell}}(\pi_{\mathrm{pos}}, \widetilde{\pi}_{\mathrm{pos}}) = \mathit{function}(r)$



$$\mathbf{H}^{(\rho)} = \int \nabla \log \mathcal{L}_y \otimes \nabla \log \mathcal{L}_y \, \mathrm{d}\rho$$

$$\mathbf{H}^{(\rho)}_{\mathrm{LIS}} = \int (\nabla G)^T \Gamma_{\mathrm{obs}}^{-1} (\nabla G) \, \mathrm{d}\rho$$

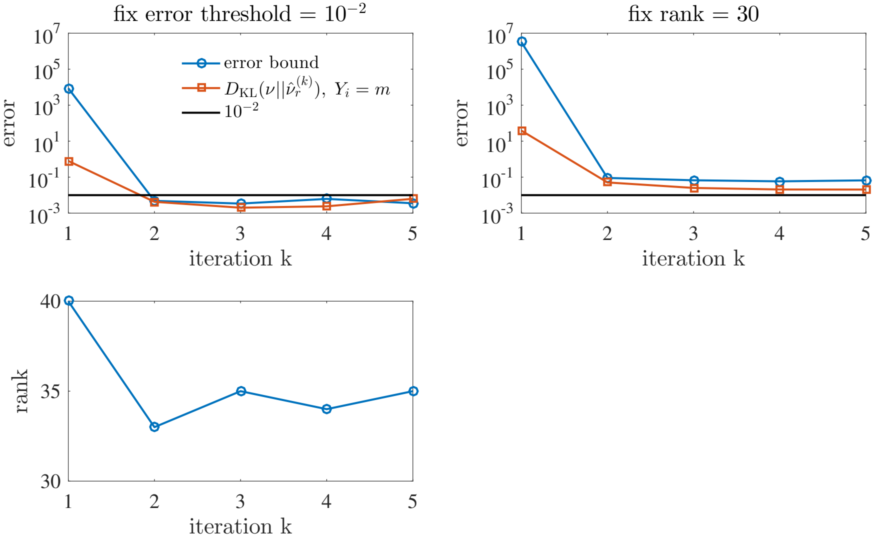
An iterative algorithm

In practice, to avoid drawing samples from π_{pos}, we can iterate *directly* towards a low-dimensional approximation $\widetilde{\pi}_{\text{pos}}$:

Conceptually:

$$\left(\rho^{\ell} \equiv \widetilde{\pi}_{\text{pos}}^{r,\ell}\right) \xrightarrow{\textit{sampling}} H^{(\rho^{\ell+1})} \xrightarrow{\textit{eigenprob}} P_r^{\ell+1} \longrightarrow \left(\rho^{\ell+1} \equiv \widetilde{\pi}_{\text{pos}}^{r,\ell+1}\right) \rightarrow \cdots$$

Iterative algorithm: results



(left) fixed threshold; *(right)* fixed rank

Questions about these low-dimensional approximations

Some open or interesting questions:

- Many MCMC algorithms use the subspace $\mathrm{Im}(P_r)$ to derive proposals and/or splitting (Metropolis-within-Gibbs) schemes (e.g., DILI [Cui et al. 2016])
  - Impact of subspace quality on computational performance of MCMC algorithms? Some inital results in [Cui & Tong 2021]
- Understanding the convergence of iterative algorithms for identifying the projector P_r, and the associated computational tradeoffs
- Extension to the *infinite-dimensional* setting
- Possibility of handling heavier-tailed priors?

Questions about these low-dimensional approximations

Some open or interesting questions:

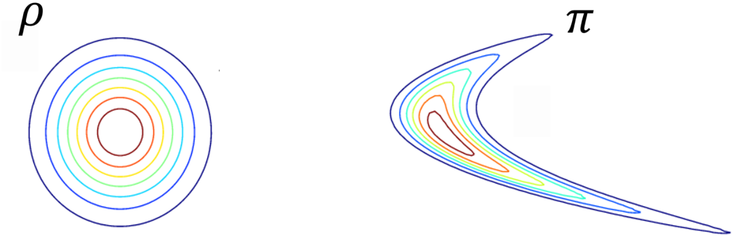
- Many MCMC algorithms use the subspace $\mathrm{Im}(P_r)$ to derive proposals and/or splitting (Metropolis-within-Gibbs) schemes (e.g., DILI [Cui et al. 2016])
  - Impact of subspace quality on computational performance of MCMC algorithms? Some inital results in [Cui & Tong 2021]
- Understanding the convergence of iterative algorithms for identifying the projector P_r, and the associated computational tradeoffs
- Extension to the *infinite-dimensional* setting
- Possibility of handling heavier-tailed priors?

Next: an application of these ideas to transport. . .

Inference through transport

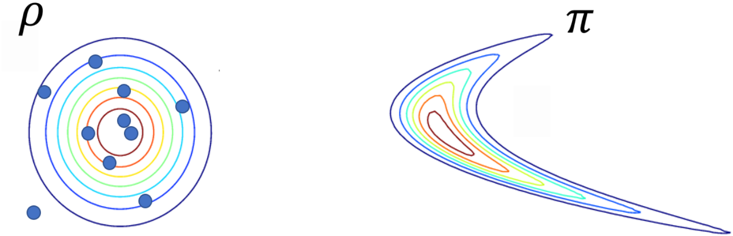
Main idea: Characterize π_{pos} (henceforth π) as a transformation of some simple distribution ρ.

Goal: Find a function T s.t. if $X \sim \rho$, then $T(X) \sim \pi$.



Main idea: Characterize π_{pos} (henceforth π) as a transformation of some simple distribution ρ.

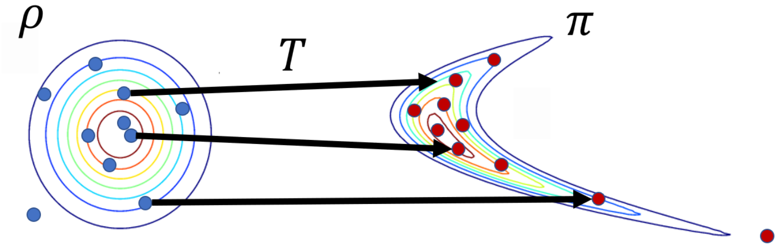
Goal: Find a function T s.t. if $X \sim \rho$, then $T(X) \sim \pi$.



Inference through transport

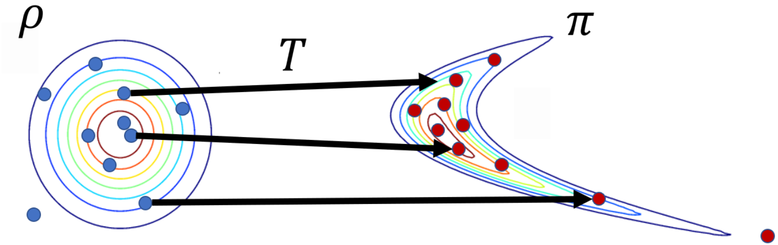
Main idea: Characterize π_{pos} (henceforth π) as a transformation of some simple distribution ρ.

Goal: Find a function T s.t. if $X \sim \rho$, then $T(X) \sim \pi$.



Main idea: Characterize π_{pos} (henceforth π) as a transformation of some simple distribution ρ.

Goal: Find a function T s.t. if $X \sim \rho$, then $T(X) \sim \pi$.



Notation: $\overset{\text{pushforward}}{T_\sharp \rho = \pi} \longleftrightarrow \overset{\text{pullback}}{\rho = T^\sharp \pi}$

How to construct a suitable map?

Maps from unnormalized densities, i.e., *variational characterization* of the map T:

How to construct a suitable map?

Maps from unnormalized densities, i.e., *variational characterization* of the map T:

$$\min_{T \in \mathcal{T}^h} \mathcal{D}_{KL}(\, T_\sharp \, \rho \,||\, \pi \,) = \min_{T \in \mathcal{T}^h} \mathcal{D}_{KL}(\, \rho \,||\, T_\sharp^{-1} \, \pi \,)$$

- π is the "target" density on $\mathbb{R}^d$; ρ is, e.g., $\mathcal{N}(0, \mathbf{I}_d)$
- $\mathcal{T}^h$ is a parameterized class of maps from $\mathbb{R}^d$ to itself
  - For instance, **monotone lower triangular maps** (approximate the Knothe–Rosenblatt rearrangement)
- Expectation is with respect to the *reference* measure ρ
  - Compute via, e.g., Monte Carlo, sparse quadrature
- Use unnormalized evaluations of π and its gradients
- No MCMC or importance sampling
- In general non-convex, unless π is log-concave

How to construct a suitable map?

Maps from unnormalized densities, i.e., *variational characterization* of the map T:

$$\min_{T \in \mathcal{T}^h} \mathcal{D}_{KL}(\, T_\sharp \, \rho \,||\, \pi \,) = \min_{T \in \mathcal{T}^h} \mathcal{D}_{KL}(\, \rho \,||\, T_\sharp^{-1} \, \pi \,)$$

- π is the "target" density on $\mathbb{R}^d$; ρ is, e.g., $\mathcal{N}(0, \mathbf{I}_d)$
- $\mathcal{T}^h$ is a parameterized class of maps from $\mathbb{R}^d$ to itself
  - For instance, **monotone lower triangular maps** (approximate the Knothe–Rosenblatt rearrangement)
- Expectation is with respect to the *reference* measure ρ
  - Compute via, e.g., Monte Carlo, sparse quadrature
- Use unnormalized evaluations of π and its gradients
- No MCMC or importance sampling
- In general non-convex, unless π is log-concave

- **Key steps:** (1) parameterize, (2) optimize

Low-dimensional structure of transport maps

Underlying challenge: maps in high dimensions

- Essential trade-off between expressiveness and computational effort/tractability!

Low rank structure

(See [BBZSM 2020] for details.)

- Let $U = [U_r \;\; U_\perp] \in \mathbb{R}^{d \times d}$ be a unitary matrix, with $U_r \in \mathbb{R}^{d \times r}$. A **lazy map** $T : \mathbb{R}^d \to \mathbb{R}^d$ takes the form:
$$T(z) = U_r \tau(z_1, \ldots, z_r) + U_\perp z_\perp$$
for some diffeomorphism $\tau : \mathbb{R}^r \to \mathbb{R}^r$.
  - Map $T \in \mathcal{T}_r(U)$ departs from the identity only on an r-dimensional subspace
- **Proposition:** For any lazy map $T \in \mathcal{T}_r(U)$, there exists a strictly positive function $f : \mathbb{R}^r \to \mathbb{R}_+$ such that
$$T_\sharp \rho(x) = f(U_r^\top x)\, \rho(x),$$
for all $x \in \mathbb{R}^d$ where $\rho = \mathcal{N}(0, \mathbf{I}_d)$. Conversely, any density of the form $f(U_r^\top x)\, \rho(x)$ for some $f : \mathbb{R}^r \to \mathbb{R}_+$ admits a lazy map representation.

Discovering structure in π before optimization

How to find a good U_r?

- Define
$$H_\pi := \mathbb{E}_\pi\left[\left(\nabla \log \frac{\pi}{\rho}\right)\left(\nabla \log \frac{\pi}{\rho}\right)^\top\right]$$
- Let (λ_i, u_i) be the ith eigenpair of H_π and put $U_r = [u_1\, u_2 \cdots u_r]$.
- **From previous results:** There exists a map $T^\star \in \mathcal{T}_r(U)$ such that
$$\mathcal{D}_{KL}(\pi || T^\star_\sharp \rho) \le \frac{1}{2}(\lambda_{r+1} + \ldots + \lambda_d).$$

Discovering structure in π before optimization

How to find a good U_r?

- Define
$$H_\pi := \mathbb{E}_\pi\left[\left(\nabla \log \frac{\pi}{\rho}\right)\left(\nabla \log \frac{\pi}{\rho}\right)^\top\right]$$
- Let (λ_i, u_i) be the ith eigenpair of H_π and put $U_r = [u_1\, u_2 \cdots u_r]$.
- **From previous results:** There exists a map $T^\star \in \mathcal{T}_r(U)$ such that
$$\mathcal{D}_{KL}(\pi || T^\star_\sharp \rho) \leq \frac{1}{2}(\lambda_{r+1} + \ldots + \lambda_d).$$
- Good approximation when the spectrum of H_π decays quickly
- $T^\star$ uses a *ridge approximation* of the likelihood $\frac{\mathrm{d}\pi}{\mathrm{d}\rho} \approx f^\star(U_r^\top x)$, with optimal profile function $f^\star(z_r) = \mathbb{E}_{X\sim\rho}\left[\frac{\pi(X)}{\rho(X)} \,\middle|\, U_r^\top X = z_r\right]$.

Error bound after optimization ("trace diagnostic")

Consider the matrix

$$H_{T^\sharp\pi} := \mathbb{E}_{T^\sharp\pi}\left[\left(\nabla \log \frac{T^\sharp\pi}{\rho}\right)\left(\nabla \log \frac{T^\sharp\pi}{\rho}\right)^\top\right]$$

Then

$$\mathcal{D}_{KL}(\pi || T_\sharp\rho) \leq \frac{1}{2}\,\mathrm{Tr}\,(H_T).$$

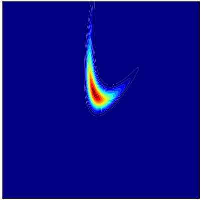
Limiting case: if $T^\sharp\pi = \rho$, then $H_T = \mathbf{0}$ and $\mathcal{D}_{KL}(\pi || T_\sharp\rho) = 0$.

Bound on the forward KL divergence for a given map.

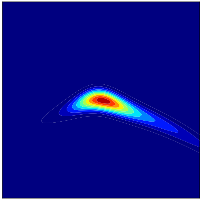
Layers of lazy maps

- What if (λ_i) do not decay quickly? What if we are limited to small r?
- Answer: build a **composition** of lazy maps, via a greedy construction
$$\mathfrak{T}_\ell = T_1 \circ T_2 \circ \cdots \circ T_\ell$$
- Algorithm ("deeply lazy" maps):
  - Given (π, ρ, r_1): compute H_π and construct a first lazy map T_1
  - Pull back π by T_1: $\pi_2 := (T_1^{-1})_\sharp \pi$
  - Given (π_2, ρ, r_2): compute H_{π_2} and construct a next lazy map T_2 ...
  - **Generic iteration**: at stage ℓ, build a lazy map to the pullback $\pi_\ell := (T_1 \circ T_2 \circ \cdots \circ T_{\ell-1})_\sharp^{-1} \pi$
  - **Stop** when $\frac{1}{2}\,\mathrm{Tr}(H_{\pi_\ell}) < \epsilon$

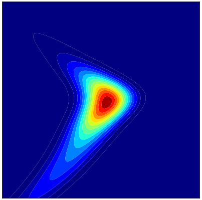
Example: rotated "banana" target distribution, $r = 1$ maps



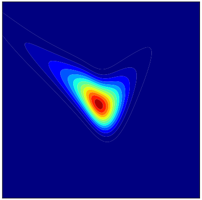
Target π



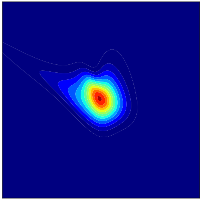
$\mathfrak{T}_1^\sharp \pi$



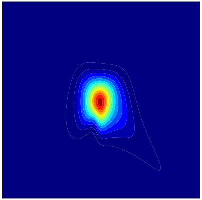
$\mathfrak{T}_2^\sharp \pi$



$\mathfrak{T}_3^\sharp \pi$

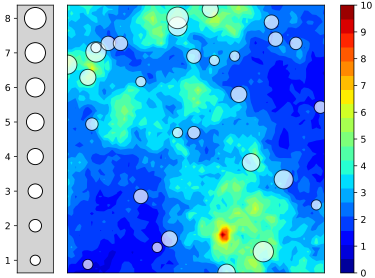


$\mathfrak{T}_5^\sharp \pi$

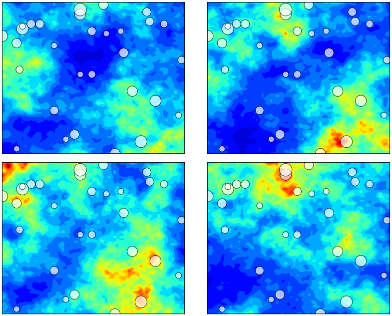


$\mathfrak{T}_8^\sharp \pi$

Example: log-Gaussian Cox process



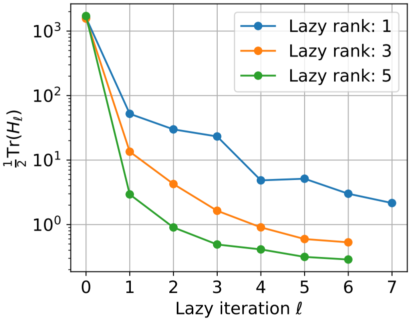
Field $\boldsymbol{\Lambda}^{\star}$ and observations $\mathbf{y}^{\star}$



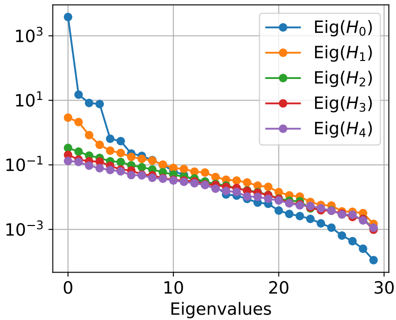
Realizations of $\boldsymbol{\Lambda} \sim \pi_{\boldsymbol{\Lambda}|\mathbf{y}^{\star}}$

Example: log-Gaussian Cox process

- Parameter dimension $n = 4096$, 30 observations; fixed ranks r



Convergence

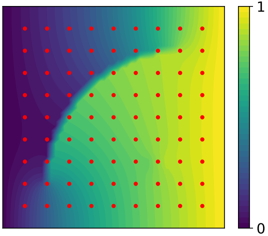


Spectrum of H_{π_ℓ}

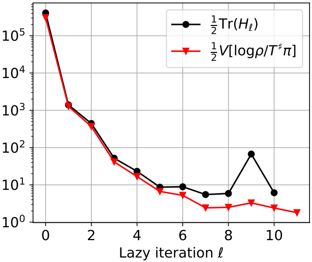
Example: elliptic PDE Bayesian inverse problem

$$\begin{cases} \nabla \cdot (e^{\kappa(\mathbf{x})} \nabla u(\mathbf{x})) = 0, \quad \text{for } \mathbf{x} \in \mathcal{D} := [0,1]^2 , \\ u(\mathbf{x}) = 0 \text{ for } x_1 = 0, \;\; u(\mathbf{x}) = 1 \text{ for } x_1 = 1, \;\; \frac{\partial u(\mathbf{x})}{\partial \mathbf{n}} = 0 \text{ for } x_2 \in \{0,1\} \end{cases}$$

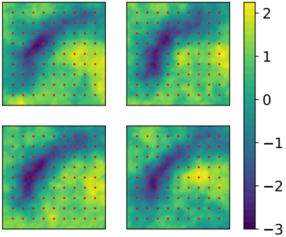
- Infer $\kappa(\mathbf{x})$, discretized with $n = 2601$ parameters; 81 observations; lazy maps of $r \leq 4$ and polynomial degree up to 2



$u(\mathbf{x})$ and observations



Convergence



Posterior realizations of $\kappa(\mathbf{x})$

Summary

- Identify and exploit *low-dimensional structure* in "updates" between distributions (from prior to posterior, from reference to target):
  - Derive an upper bound on the forward KL divergence
  - Minimize this upper bound using PCA on $\nabla \log \mathcal{L}_y$
  - Better performance than heuristic gradient-based methods (e.g., likelihood-informed subspace or active subspaces)

- Transport methods: exploiting the **pullback** distribution
  - Compositions of low-dimensional maps, constructed greedily ("deeply lazy" maps)

Summary

- Identify and exploit *low-dimensional structure* in “updates” between distributions (from prior to posterior, from reference to target):
  - Derive an upper bound on the forward KL divergence
  - Minimize this upper bound using PCA on $\nabla \log \mathcal{L}_y$
  - Better performance than heuristic gradient-based methods (e.g., likelihood-informed subspace or active subspaces)
- Transport methods: exploiting the **pullback** distribution
  - Compositions of low-dimensional maps, constructed greedily (“deeply lazy” maps)

Thanks for your attention!

References

- M. Brennan, D. Bigoni, O. Zahm, A. Spantini, Y. Marzouk. “Greedy inference with structure-exploiting lazy maps.” *NeurIPS 2020*.
- O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk. “Certified dimension reduction in nonlinear Bayesian inverse problems.” arXiv:1807.03712v3, 2021.
- T. Cui, O. Zahm, “Data-free likelihood-informed dimension reduction of Bayesian inverse problems.” *Inverse Problems*, 2021.
- T. Cui, X. Tong, “A unified performance analysis of likelihood-informed subspace methods.” arXiv:2101.02417, 2021.
- O. Zahm, P. Constantine, C. Prieur, Y. Marzouk. “Gradient-based dimension reduction of multivariate vector-valued functions,” *SISC*, 2020.
- A. Spantini, D. Bigoni, Y. Marzouk. “Inference via low-dimensional couplings.” *JMLR* 19(66): 1–71, 2018.
- P. Constantine, C. Kent, T. Bui-Thanh. “Accelerating Markov chain Monte Carlo with active subspaces.” SISC, 2016.
- A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, Y. Marzouk, “Optimal low-rank approximations of Bayesian linear inverse problems,” *SISC*, 2015.
- T. Cui, J. Martin, Y. Marzouk, A. Solonen. A. Spantini, “Likelihood-informed dimension reduction for nonlinear inverse problems,” *Inverse Problems*, 2014.

Approximation of $\pi^*_{\text{pos}}(x) \propto \mathbb{E}_{\text{pr}}(\mathcal{L}_y|P_r x)\pi_{\text{pr}}(x)$

- The conditional expectation $\mathbb{E}_{\text{pr}}(\mathcal{L}_y|P_r x)$ can be expressed as
$$x \mapsto \int \mathcal{L}_y(P_r x + (I_d - P_r)z)\ \pi_{\text{pr}}(z|P_r x)\mathrm{d}z$$
where $\pi_{\text{pr}}(\cdot|P_r x)$ denotes the conditional prior, which depends on x.

Approximation of $\pi^*_{\text{pos}}(x) \propto \mathbb{E}_{\text{pr}}(\mathcal{L}_y | P_r x) \pi_{\text{pr}}(x)$

- The conditional expectation $\mathbb{E}_{\text{pr}}(\mathcal{L}_y | P_r x)$ can be expressed as
$$x \mapsto \int \mathcal{L}_y(P_r x + (I_d - P_r) z) \, \pi_{\text{pr}}(z | P_r x) \mathrm{d}z$$
where $\pi_{\text{pr}}(\cdot | P_r x)$ denotes the conditional prior, which depends on x.

- Consider the following Monte Carlo estimate
$$\widetilde{\mathcal{L}} : x \mapsto \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_y(P_r x + (I_d - P_r) Z_i) \quad , \quad Z_i \overset{\text{iid}}{\sim} \pi_{\text{pr}}$$
In general, $\widetilde{\mathcal{L}}(P_r x)$ is a biased estimator for $\mathbb{E}_{\text{pr}}(\mathcal{L}_y | P_r x)$.

Approximation of $\pi^*_{\text{pos}}(x) \propto \mathbb{E}_{\text{pr}}(\mathcal{L}_y|P_r x)\pi_{\text{pr}}(x)$

- The conditional expectation $\mathbb{E}_{\text{pr}}(\mathcal{L}_y|P_r x)$ can be expressed as
$$x \mapsto \int \mathcal{L}_y(P_r x + (I_d - P_r)z)\ \pi_{\text{pr}}(z|P_r x)\mathrm{d}z$$
where $\pi_{\text{pr}}(\cdot|P_r x)$ denotes the conditional prior, which depends on x.

- Consider the following Monte Carlo estimate
$$\widetilde{\mathcal{L}} : x \mapsto \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_y(P_r x + (I_d - P_r)Z_i) \quad , \quad Z_i \overset{\text{iid}}{\sim} \pi_{\text{pr}}$$
In general, $\widetilde{\mathcal{L}}(P_r x)$ is a biased estimator for $\mathbb{E}_{\text{pr}}(\mathcal{L}_y|P_r x)$.

Proposition

The random distribution $\widetilde{\pi}_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\text{pr}}(x)$ is such that
$$\mathbb{E}\Big(D_{\text{KL}}(\pi^*_{\text{pos}}||\widetilde{\pi}_{\text{pos}})\Big) \lesssim \left(C_1 + \frac{C_2}{M}\right) \mathcal{R}_{\pi_{\text{pos}}}(P_r)$$

Convergence of the greedy construction

Theorem (BBZSM21)

Let $U^1, U^2, \ldots$ be a sequence of unitary matrices. For any $\ell \geq 1$, let $T_\ell \in \mathcal{T}_r(U^\ell)$ be a lazy map that minimizes $\mathcal{D}_{KL}(\pi_{\ell-1}||(T_\ell)_\sharp\rho)$, where $\pi_{\ell-1} = (T_1 \circ \ldots \circ T_{\ell-1})^\sharp\pi$. If there exists $0 < t \leq 1$ such that for any $\ell \geq 1$

$$\mathcal{D}_{KL}((U_r^{\ell\top})_\sharp\pi_{\ell-1}||\rho_r) \geq t \sup_{\substack{U \in \mathbb{R}^{d\times d} \\ s.t.\ UU^\top = I_d}} \mathcal{D}_{KL}((U_r^\top)_\sharp\pi_{\ell-1}||\rho_r),$$

then $(T_1 \circ \ldots \circ T_\ell)_\sharp\rho$ converges weakly to π.

Convergence of the greedy construction

Theorem (BBZSM21)

Let $U^1, U^2, \ldots$ be a sequence of unitary matrices. For any $\ell \geq 1$, let $T_\ell \in \mathcal{T}_r(U^\ell)$ be a lazy map that minimizes $\mathcal{D}_{KL}(\pi_{\ell-1}||(T_\ell)_\sharp\rho)$, where $\pi_{\ell-1} = (T_1 \circ \ldots \circ T_{\ell-1})^\sharp \pi$. If there exists $0 < t \leq 1$ such that for any $\ell \geq 1$

$$\mathcal{D}_{KL}((U_r^{\ell\top})_\sharp \pi_{\ell-1}||\rho_r) \geq t \sup_{\substack{U \in \mathbb{R}^{d\times d} \\ s.t.\ UU^\top = I_d}} \mathcal{D}_{KL}((U_r^\top)_\sharp \pi_{\ell-1}||\rho_r),$$

then $(T_1 \circ \ldots \circ T_\ell)_\sharp \rho$ converges weakly to π.

Comments:

- This is a sufficient, not necessary, condition for convergence
- $t = 1$ corresponds to an "ideal" greedy algorithm, but suboptimal choices for U^ℓ corresponding to $0 < t < 1$ are also sufficient
- Bound should apply simultaneously to *all* layers