Low-rank Structure in Optimization-based Sampling

Zheng Wang

Massachusetts Institute of Technology
Computational Science and Engineering

February 26, 2019

SIAM Conference on Computational Science and Engineering

Outline

Motivation

Sampling algorithms

- Compute expectations (e.g. mean, variance) using Monte Carlo sum
- Used in Bayesian inference, filtering, experimental design

Scope

- Can evaluate the log-density of the target, up to a constant
- Can evaluate Jacobian actions, perhaps higher derivatives

Types

- Markov chain Monte Carlo (MCMC)
- Self-normalized importance sampling (IS)

Efficiency depends on transition (in MCMC) or biasing distribution (in IS)

Sampling algorithms that involve optimization

Optimization-based samplers

1. Construct objective functions using:
    - parts of target distribution
    - sample from a reference random variable
2. Minimize the objective functions to get proposal samples
3. Compute log-densities of the resulting proposal samples
4. Samples (+ densities) are used as:
    - independent proposal in Metropolis-Hastings
    - biasing distribution in self-normalized IS

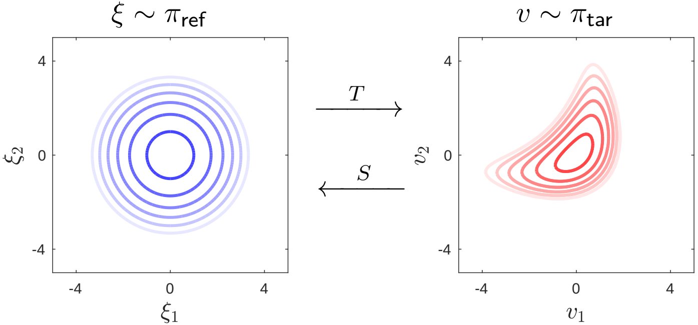
Examples

- Implicit sampling/filtering Chorin et al. (2010); Morzfeld et al. (2012)
- Randomize-then-optimize (RTO) Bardsley et al. (2014)
- Metropolized randomized maximum likelihood Oliver (2017)

Exact transport maps

Seek a transformation of random variables such that:

$$T_\sharp \pi_{\text{ref}} \stackrel{d}{=} \pi_{\text{tar}}$$

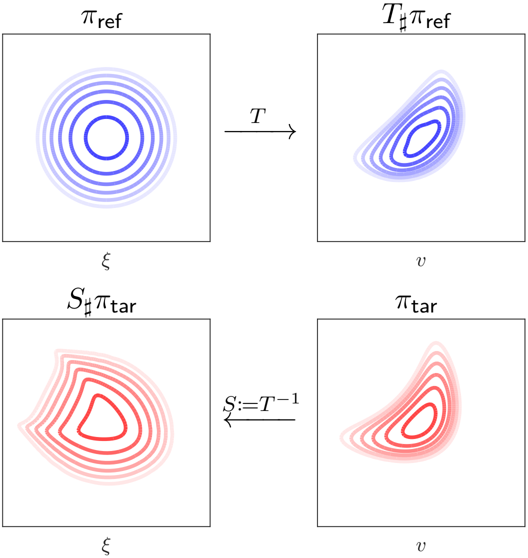


$T(\xi) = v$ $\qquad$ $S(v) := T^{-1}(v) = \xi$

Approximate transport maps

$$T_\sharp \pi_{\text{ref}}\,(v) \\ = \left|\nabla S\right| \pi_{\text{ref}}(S(v))$$

$$S_\sharp \pi_{\text{tar}}\,(\xi) \\ = \left|\nabla T\right| \pi_{\text{tar}}(T(\xi))$$



Randomize-then-Optimize (RTO)

Required form of the target

$$\log\ \pi_{\mathsf{tar}}(v) = -\frac{1}{2}\left\|H(v)\right\|^2$$

RTO's mapping (ansatz)

$$\xi = Q^\top H(v)$$

- Output dimension of H is larger than that of v
- Reference samples ξ are drawn from a standard normal
- Matrix Q is found from thin-QR factorization of $\nabla H\left(v_{\mathrm{MAP}}\right)$

Randomize-then-Optimize (RTO)

RTO's mapping (ansatz)

$$\xi = Q^\top H(v)$$

Guidelines for a useful mapping from Chorin et al. (2010)

1. One-to-one
2. Maps the neighborhood of zero to a set that contains the mode
3. Smooth near $\xi = 0$
4. Easily compute the Jacobian (actions and determinant)

Randomize-then-Optimize (RTO)

RTO's mapping (ansatz)

$$\xi = Q^\top H(v)$$

Guidelines for a useful mapping from Chorin et al. (2010)

1. One-to-one?
2. Maps the neighborhood of zero to a set that contains the mode
3. Smooth near $\xi = 0$
4. Easily compute the Jacobian (actions and determinant)

Simple Bayesian inference example

Consider the following inverse problem

$$\underbrace{d}_{\text{observation}} = \overbrace{G(v)}^{\text{forward model}} + \underbrace{e}_{\text{noise}} \qquad v \sim N(0, I_n) \qquad e \sim N(0, I_m)$$

$$\begin{aligned}
\underbrace{p(v|d)}_{\text{posterior}} &\propto \exp\left(-\frac{1}{2}\left(G(v)-d\right)^2\right)\exp\left(-\frac{1}{2}v^2\right) \\
&= \exp\left(-\frac{1}{2}\Big\| \underbrace{\begin{bmatrix} v \\ G(v) \end{bmatrix} - \begin{bmatrix} 0 \\ d \end{bmatrix}}_{H(v)} \Big\|^2\right) \\
&= \exp\left(-\frac{1}{2}\|H(v)\|^2\right)
\end{aligned}$$

Simple Bayesian inference example

Given problem structure:

$$H(v) = \begin{bmatrix} v \\ G(v) - d \end{bmatrix}$$

1. Find the posterior mode v_{MAP}.
2. Solve a thin-QR factorization to find the matrix Q.

$$QR = \nabla H\left(v_{\text{MAP}}\right) = \begin{bmatrix} I_n \\ \nabla G\left(v_{\text{MAP}}\right) \end{bmatrix}$$

3. Draw a reference sample ξ and solve nonlinear system for v.

$$\xi = Q^\top H(v)$$

4. Compute the proposal density.

$$q(v) = (2\pi)^{-\frac{n}{2}} \left|Q^\top \nabla H(v)\right| \exp\left(-\frac{1}{2}\left\|Q^\top H(v)\right\|^2\right)$$

5. Repeat steps 3 and 4.

A different choice of Q

Algorithmic modification

Exploit the structure of $\nabla H(v_{\mathsf{MAP}})$ to reduce computation.

$$QR = \nabla H(v_{\mathsf{MAP}}) = \begin{bmatrix} I_n \\ \nabla G(v_{\mathsf{MAP}}) \end{bmatrix}$$

A different choice of Q

Algorithmic modification

Exploit the structure of $\nabla H(v_{\text{MAP}})$ to reduce computation.

$$QR = \nabla H(v_{\text{MAP}}) = \begin{bmatrix} I_n \\ \nabla G(v_{\text{MAP}}) \end{bmatrix} = \begin{bmatrix} I_n \\ \Psi\Lambda\Phi^\top \end{bmatrix}$$

Represent range of Q using an SVD. Let $r =$ rank of $\nabla G(v_{\text{MAP}})$.

Choose a symmetric $\widetilde{R} = \left(\nabla H(v_{\text{MAP}})^\top \nabla H(v_{\text{MAP}})\right)^{\frac{1}{2}}$. Then,

$$\widetilde{Q} = \nabla H(v_{\text{MAP}})\widetilde{R}^{-1} = \begin{bmatrix} \Phi(\Lambda^2 + I_r)^{-\frac{1}{2}}\Phi^\top + (I_n - \Phi\Phi^\top) \\ \Psi\Lambda(\Lambda^2 + I_r)^{-\frac{1}{2}}\Phi^\top \end{bmatrix}$$

$\widetilde{Q}$: orthogonal, same range, represented by Ψ, Λ, Φ.

Scalable implementation of RTO

To propose a point, solve $\xi = \widetilde{Q}^\top H(v)$:

$$\begin{cases} (I_n - \Phi\Phi^\top)\xi = (I_n - \Phi\Phi^\top)v, \\ \qquad\qquad \Phi\Phi^\top\xi = \Phi\left[(\Lambda^2 + I_r)^{-\frac{1}{2}}\Phi^\top v + \Lambda(\Lambda^2 + I_r)^{-\frac{1}{2}}\Psi^\top G(v)\right] \end{cases}$$

System of equations splits in two. Optimize only for r-dim. part.

Scalable implementation of RTO

To propose a point, solve $\xi = \widetilde{Q}^\top H(v)$:

$$\begin{cases} (I_n - \Phi\Phi^\top)\xi = (I_n - \Phi\Phi^\top)v, \\ \qquad \Phi\Phi^\top\xi = \Phi\left[(\Lambda^2 + I_r)^{-\frac{1}{2}}\Phi^\top v + \Lambda(\Lambda^2 + I_r)^{-\frac{1}{2}}\Psi^\top G(v)\right] \end{cases}$$

System of equations splits in two. Optimize only for r-dim. part.

To calculate the proposal density:

$$\left|\underbrace{\widetilde{Q}^\top \nabla H(v)}_{n\times n}\right| = \left|(\Lambda^2 + I_r)^{-\frac{1}{2}}\right| \cdot \left|\underbrace{I_r + \Lambda\Psi^\top \nabla G(v)\Phi}_{r\times r}\right|$$

Find determinant of an $r \times r$ matrix. *Reduced from $n \times n$ matrix.*

Unwhitened system of equations

$$
\begin{tikzcd}[column sep=huge]
u \arrow[r, leftrightarrow, "v = S_{\mathrm{pr}}^{-1}(u - m_{\mathrm{pr}})"] & v \arrow[r, leftrightarrow, "\widetilde{Q}^{\top} H(v) = \xi"] & \xi \arrow[r, leftrightarrow, "\zeta = S_{\mathrm{pr}}\xi + m_{\mathrm{pr}}"] & \zeta
\end{tikzcd}
$$

$\sim \pi_{\mathrm{RTO}}(u)$	$\sim q(v)$	$\sim \mathrm{N}(0, \mathrm{I})$	$\sim \mathrm{N}(m_{\mathrm{pr}}, \Gamma_{\mathrm{pr}})$
unwhitened proposal	whitened proposal	whitened reference	prior

RTO's unwhitened mapping:

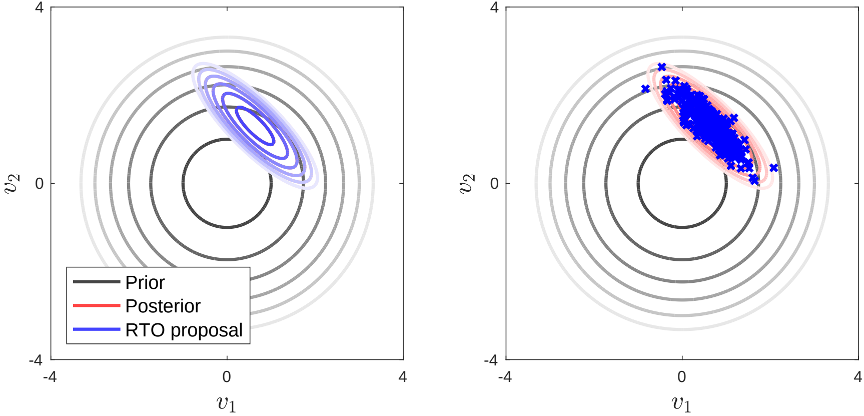
$$
\begin{cases}
(I - P)\,\zeta = (I - P)\,u, \\
P\,\zeta = \mathscr{F}(u)
\end{cases}
$$

Takeaway

- RTO's mapping keeps most parameter directions fixed.
- Directions that move *depend* on those that are fixed.

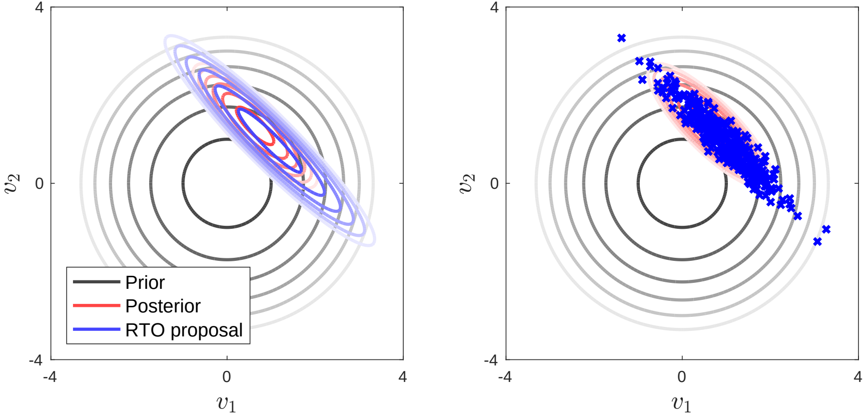
Truncated SVD for a linear model

Linear model G, using rank $r = 2$:



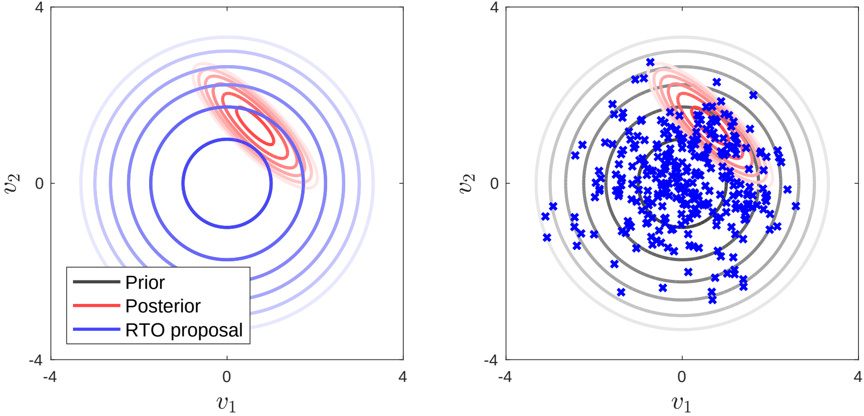
Truncated SVD for a linear model

Linear model G, using rank $r = 1$:



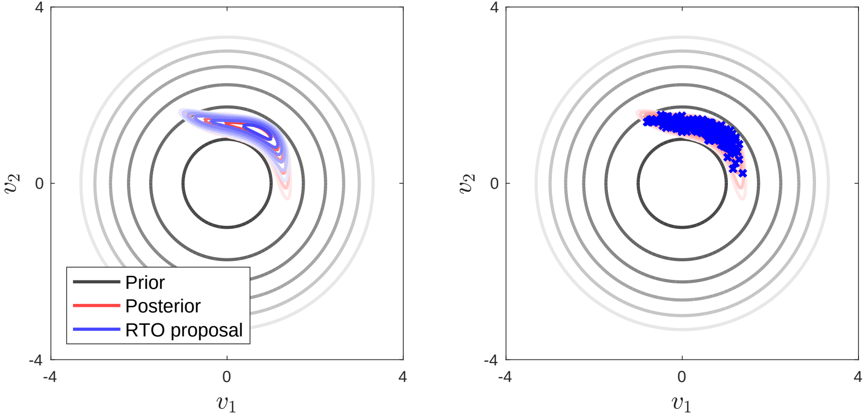
Truncated SVD for a linear model

Linear model G, using rank $r = 0$:



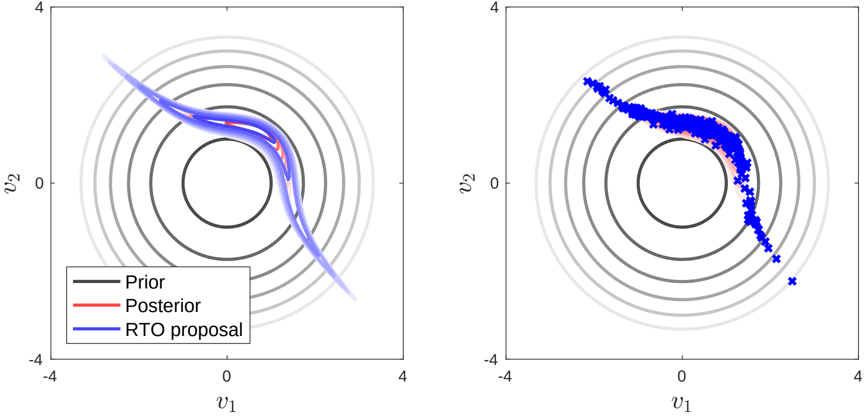
Truncated SVD for a nonlinear model

Nonlinear model G, using rank $r = 2$:



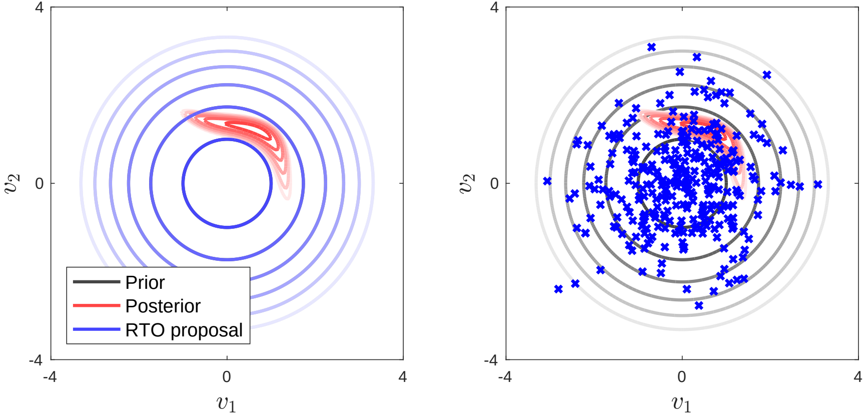
Truncated SVD for a nonlinear model

Nonlinear model G, using rank $r = 1$:



Truncated SVD for a nonlinear model

Nonlinear model G, using rank $r = 0$:



Dimension independence

Empirically, our scalable implementation of RTO appears to have dimension-independent performance.

Table: Numerical ESS for the chain, average acceptance rate, and average optimization iterations for RTO, varying parameter dimension.

Parameter Dim.	321	641	1281	2561	5121	10241
Numerical ESS	4343.5	4544.8	4464.5	4523.3	4484.9	4532.2
Acceptance Rate	0.936	0.948	0.950	0.954	0.950	0.953
Opt. Iterations	324.04	357.76	307.50	198.81	165.06	142.25

RTO in function space

RTO's mapping make sense in function space.

$$\begin{cases} (I - P)\,\zeta = (I - P)\,u, \\ P\,\zeta = \mathscr{F}(u) \end{cases}$$

RTO in function space

RTO's mapping make sense in function space.

$$\begin{cases} (I - P)\,\zeta = (I - P)\,u, \\ P\,\zeta = \mathscr{F}(u) \end{cases}$$

Illustration:

1. Draw a realization $\zeta \in \mathcal{H}$ from the prior (Gaussian process).
2. Discretize this sample ζ on grids of different resolution.
3. Apply RTO's prior-to-proposal mapping to each discretized sample.

Illustration: different discretizations

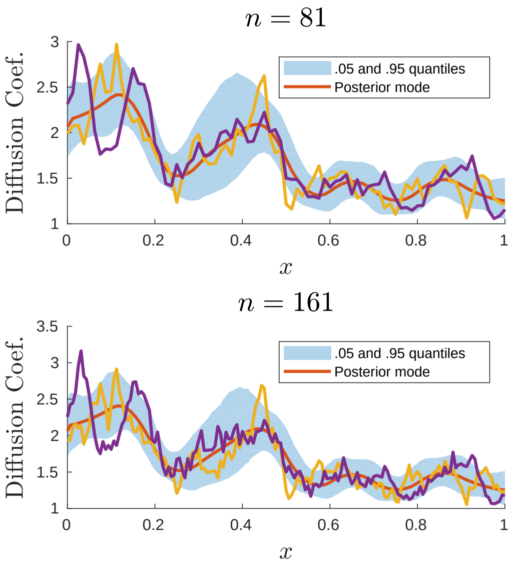
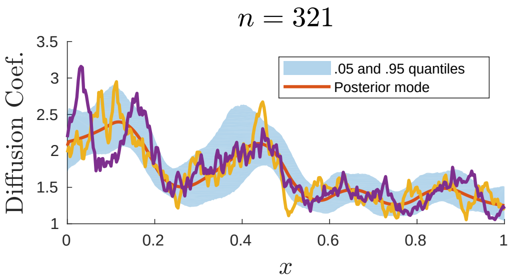
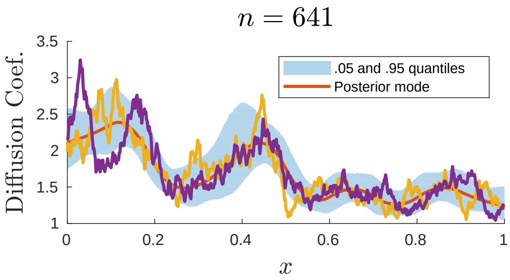


Illustration: different discretizations





Theoretical result

Theorem

(Absolute continuity of RTO's proposal with respect to the prior.) Suppose that the prior μ_{pr} is a non-degenerate Gaussian measure on $\mathcal{H}$, the random variable ζ is distributed according to μ_{pr}, the random variable u is defined through the mapping above, and for every $b \in W^{\perp}$, the mapping

$$a \mapsto a + \Lambda\Psi^T S_{obs}^{-1}(F\left(X(a) + b\right) - y)$$

is Lipshitz continuous, injective, and its inverse is Lipschitz continuous, where μ_{RTO} is the measure induced by u. Then μ_{RTO} is absolutely continuous with respect to μ_{pr}.

Theoretical result

Steps of the proof

1. Formulate finite dimensional mapping in unwhitened coordinates.
2. Recast mapping in function space by defining the projector P through an inner product with sufficiently smooth basis functions χ_i.
3. Determine the proposal measure as the push-forward of the prior Gaussian process through the mapping. Find its Radon-Nikodym derivative with respect to the prior.

Review

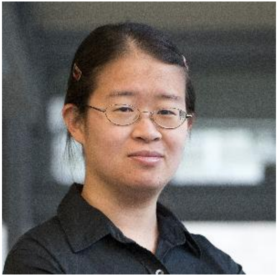
RTO

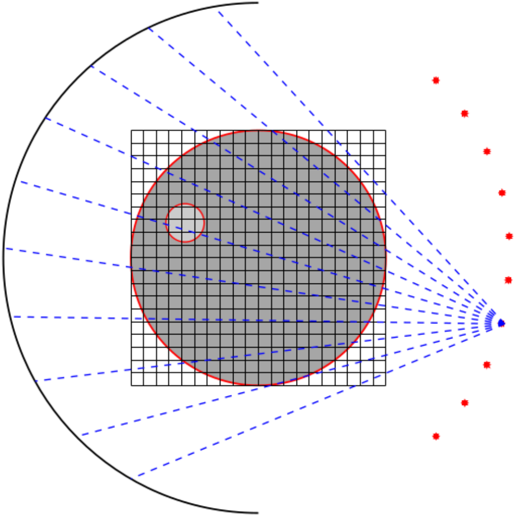
- Specfies an approximate transport map.
- Yields a *non-Gaussian* proposal distribution.
- Scalable implementation makes inference tractable for high parameter dimension.
- Works well for weakly nonlinear problems.
- Provably dimension independent.

However, sampling can be poor if the problem is sufficiently nonlinear.

X-ray tomography

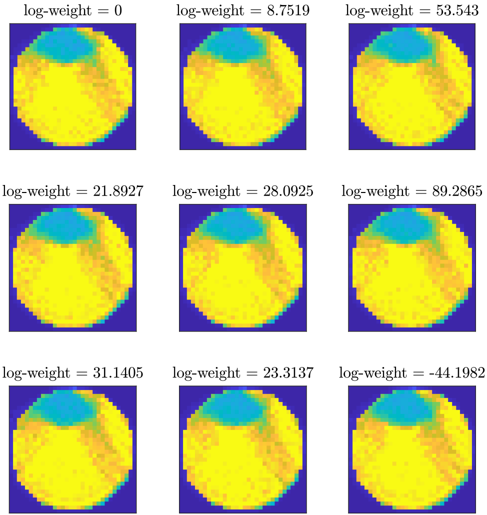
Harriet Li:



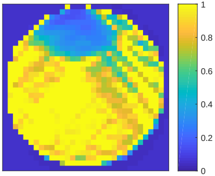


X-ray tomography

Stochastic solution:

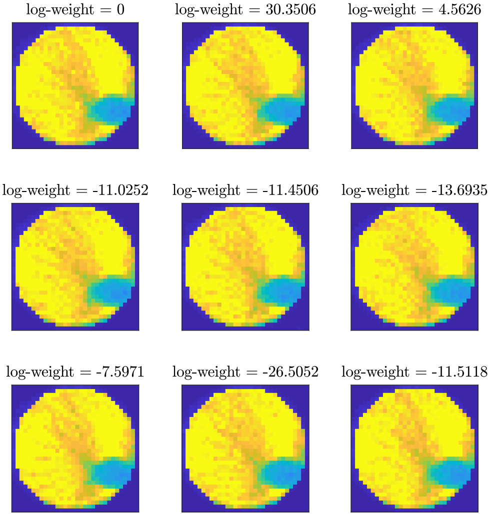


Deterministic solution:

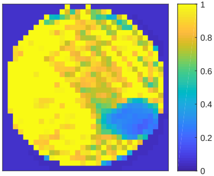


X-ray tomography

Stochastic solution:

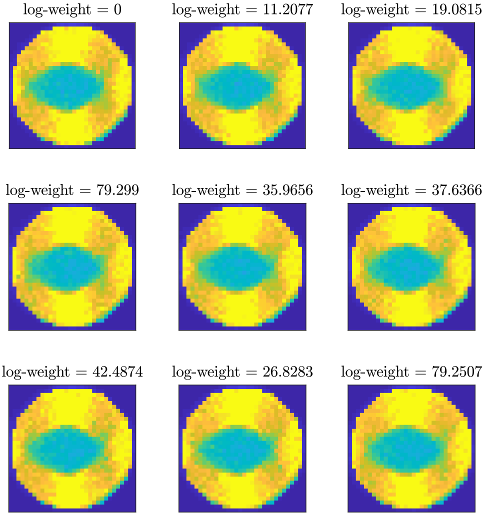


Deterministic solution:

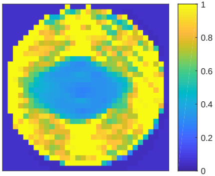


X-ray tomography

Stochastic solution:



Deterministic solution:



Discussion

Recap

- Samplers based on optimization often describe a mapping.
- RTO is an ansatz.
- Parameter dimensions are split into two groups.
- Theoretical result: RTO is dimension independent.
- Correlated samples between discretizations.
- Sampling can be poor for sufficiently nonlinear inverse problems.

Future work

- Improve sampling by using local proposals.
- Use a mixture of the prior to combat invertibility requirements.

References I

Bardsley, J. M., Solonen, A., Haario, H., and Laine, M. (2014). Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1895–A1910.

Chorin, A., Morzfeld, M., and Tu, X. (2010). Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2):221–240.

Morzfeld, M., Tu, X., Atkins, E., and Chorin, A. J. (2012). A random map implementation of implicit filters. *Journal of Computational Physics*, 231(4):2049–2066.

Oliver, D. S. (2017). Metropolized Randomized Maximum Likelihood for sampling from multimodal distributions. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):259–277.

Robust to observational noise

Empirically RTO is robust to observational noise.

Table: Numerical ESS for the chain, average acceptance rate, and average optimization iterations for RTO, varying observational noise.

Observational Noise	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
Numerical ESS	4504.8	4427.4	4349.9	4423.0	4415.1	4187.2
Acceptance Rate	0.946	0.944	0.941	0.945	0.935	0.924
Opt. Iterations	567.64	495.41	363.71	296.55	89.07	8.32