Accelerating Bayesian computation with transport maps

Youssef Marzouk, Matthew Parno

Department of Aeronautics and Astronautics
Center for Computational Engineering
Massachusetts Institute of Technology
http://uqgroup.mit.edu

Support from DOE Office of Advanced Scientific Computing Research

13 November 2015

MIT Uncertainty Quantification (UQ) group

Who we are:

- Currently 6 postdoctoral associates, 7 PhD students, 3 SM students (some co-advised), one PI
- Part of the *MIT Center for Computational Engineering*; the *Center for Statistics* within MIT's new *Institute for Data, Systems, and Society* (IDSS); and the *MIT Department of Aeronautics and Astronautics*

Problem domains of interest:

1. *Statistical inference and inverse problems:* large-scale Bayesian computation; model and dimension reduction for Bayesian inference; sequential data assimilation and nonlinear filtering; model selection
    - Applications: subsurface modeling, glaciology and ice-ocean interactions, atmospheric remote sensing, chemical kinetics

Problem domains of interest (continued):

2. *Forward UQ:* uncertainty propagation, solution of random ODEs and PDEs; polynomial chaos, sparse grids, tensor methods; high-dimensional approximation
   - Applications: sensitivity analysis and surrogate modeling in *many* areas, including aerospace systems; stochastic control
3. *Optimal experimental design:* Optimal data collection; Bayesian approaches to model-based batch and sequential experimental design
   - Applications: combustion kinetics, contaminant source detection, UAV navigation and path planning
4. *Optimization under uncertainty:* Derivative-free optimization with risk/robustness measures or constraints; decision-making under uncertainty
   - Applications: chemical process design; energy conversion systems

MIT Uncertainty Quantification (UQ) group

Open-source codes:

- **MUQ**: `http://muq.mit.edu`, MIT Uncertainty Quantification Library
  - A C++/python library for both modelers and algorithm developers; many UQ tools
- **(S)NOWPAC**: `http://bitbucket.org/fmaugust/nowpac`, (Stochastic) Nonlinear Optimization with Path-Augmented Constraints
  - Derivative-free nonlinear constrained optimization with risk and robustness measures

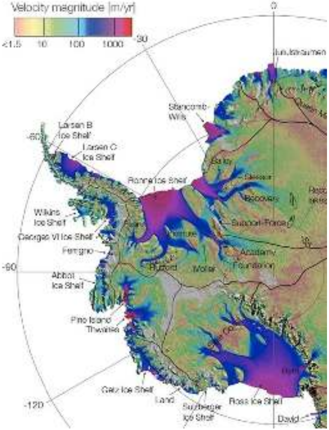
Support from:

- *Government agencies:* DOE, AFOSR, NSF, DARPA
- *Industry and others:* BP, Eni, United Technologies, KAUST

Collaborations with Sandia, Oak Ridge, UT Austin, Harvard, USC, Duke, Montana, Colorado, LIMSI-CNRS, ...

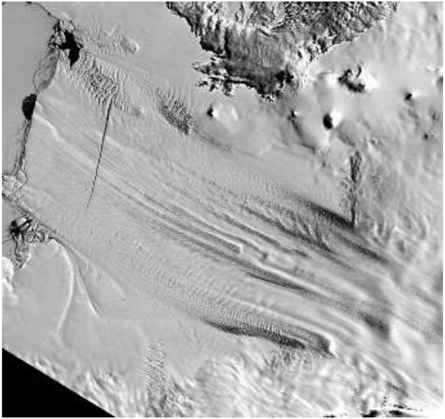
Example: ice sheet dynamics in western Antarctica

Western Antarctic Ice Sheet



[Rignot et al. 2011]

Pine Island Glacier



[NASA]

Bayesian inference setting

Posterior density of the parameters

$$\pi(\theta) := p(\theta|\mathbf{d}) \propto \mathcal{L}(\mathbf{d}, \mathbf{f}(\theta))p(\theta)$$

Ingredients:

- Parameters $\theta \in \mathbb{R}^d$; data $\mathbf{d} \in \mathbb{R}^n$
- Prior density $p(\theta) : \mathbb{R}^d \to \mathbb{R}^+$
- Forward model $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$
  - Often a **black-box** function (the setting for this talk!)
  - Each evaluation is **expensive**
- Likelihood function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$
  - $\mathcal{L}(\mathbf{d}, \mathbf{f}(\theta)) = p(\mathbf{d}|\theta)$; compares model predictions to observed data
  - Each evaluation requires, in principle, an evaluation of $\mathbf{f}$
  - Simple example:

$$\mathbf{d} = \mathbf{f}(\theta) + \epsilon,\ \ \epsilon \sim N(0, \Sigma),\ \ \text{then } \mathbf{d}|\theta \sim N(\mathbf{f}(\theta), \Sigma)$$

Computational challenges

- Extract information from the posterior *(means, covariances, event probabilities, predictions)* by evaluating **posterior expectations:**

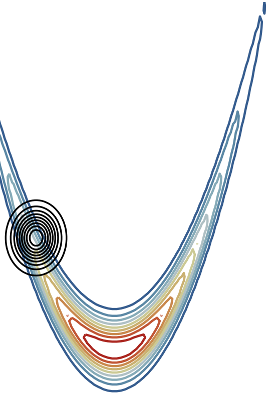
$$\mathbb{E}_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

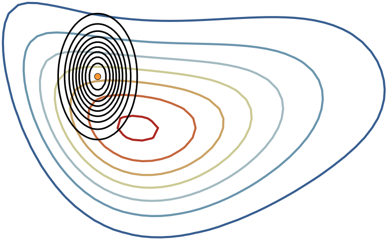
- Key strategies for making this computationally tractable

1. Efficient and structure-exploiting **sampling schemes**
2. **Approximations** of the forward model, e.g., spectral expansions, local interpolants, reduced order models, multi-fidelity approaches

Sampling schemes

- **Markov chain Monte Carlo (MCMC)** algorithms are the workhorse of Bayesian computation

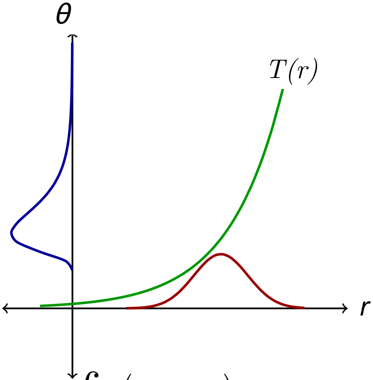




- Effective = ***adapted to the target***
- Can we **transform** proposals *or* targets for better sampling?

Optimal transport

- A different viewpoint: **deterministic coupling** of two random variables $r \sim \mu, \theta \sim \nu$



- Monge problem: $\min_T \int c\Big(r, T(r)\Big)\, \mu(dr)$, where $T_\sharp \mu = \nu$
- A unique and *monotone* solution exists for quadratic (and other) **transport costs** $c(x,y)$ [Brenier 1991, McCann 1995]

Triangular transport

- Useful alternative to the optimal map: *triangular* (Knothe-Rosenblatt) transport

$$T(r) = \begin{bmatrix} T^1(r_1) \\ T^2(r_1, r_2) \\ \vdots \\ T^D(r_1, r_2, \ldots, r_D) \end{bmatrix}$$

  - Exists and is unique (up to ordering) under mild conditions
  - Monotonicity: $\partial_i T^i > 0, \;\; i = 1 \ldots D$
  - Jacobian determinant is easy to evaluate
  - Limit of a *weighted* L^2 - optimal transport [Carlier 2010, Bonnotte 2013]

Transport maps: computation

- Previous work: directly finding a map from prior to posterior [Moselhy & M, JCP 2012]
  - **Reference** = prior or a multivariate standard normal
  - **Target** = posterior

$$\arg\min_{T \in \mathcal{T}_\triangle} D_{\mathrm{KL}}\left(T_\sharp p \,\middle\|\, \pi\right) =$$

$$\arg\max_{T \in \mathcal{T}_\triangle} \mathbb{E}_p\left[\log\left(\pi(T(r))\right) + \log\left|\det \nabla T\right|\right]$$

Combining transport maps with MCMC

- Optimization problem can be costly in high dimensions
- Map must be represented in a finite basis (e.g., polynomials) and is thus in general *approximate*. Can we still achieve *exact* posterior sampling?

- **Key idea:** combining map construction with MCMC
  - Posterior sampling + convex optimization
  - Transport map "preconditions" MCMC sampling; posterior samples enable simpler map construction
  - Can also be understood in the framework of *adaptive* MCMC

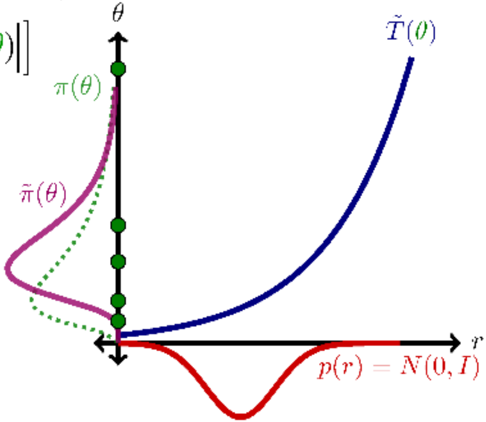
Constructing a map from samples

- Explicitly seek map from target to reference
- Candidate map $\tilde{T}$ yields an approximation $\tilde{\pi} = \tilde{T}_\sharp^{-1} p$ of target dist
- Optimization objective:

$$\min_{\tilde{T} \in \mathcal{T}_\Delta} D_{KL}\left(\pi \middle\| \tilde{T}_\sharp^{-1} p\right) = \min_{\tilde{T} \in \mathcal{T}_\Delta} D_{KL}\left(\tilde{T}_\sharp \pi \middle\| p\right)$$

$$\Rightarrow \max_{\tilde{T} \in \mathcal{T}_\Delta} \mathbb{E}_{\pi}\left[\log p(\tilde{T}(\theta)) + \log\left|\nabla \tilde{T}(\theta)\right|\right]$$

- Samples from π approximate the expectation; p has useful structure



Constructing a map from samples

- Useful structure:
  - Seek a monotone **lower triangular map** (converges to *Knothe-Rosenblatt rearrangement)*
  - Let target $p(r)$ be standard Gaussian
- Yields a **convex** and **separable** optimization problem:

$$\max_{\tilde{T} \in \mathcal{T}_\triangle} \mathbb{E}_\pi \left[\log p \circ \tilde{T} + \log \det \nabla \tilde{T} \right]$$

$$\text{s.t. } \partial_j T^j(\theta) > 0 \quad \pi - \text{a.e.}$$

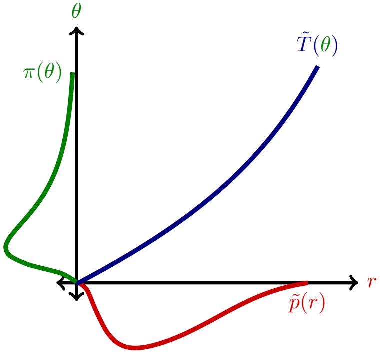
  - Sample-average approximation (SAA) with N samples from π

$$\min_{\tilde{T}^j \in \mathcal{T}_\triangle^j} \sum_{i=1}^{N} \frac{1}{2} \tilde{T}^{j,2}(\theta_i) - \log \left. \frac{\partial \tilde{T}^j}{\partial \theta_j} \right|_{\theta^{(i)}}, \quad \text{s.t.} \quad \left. \frac{\partial \tilde{T}^j}{\partial \theta_j} \right|_{\theta^{(i)}} \geq \lambda_{\min} > 0, \quad \forall i \in \{1, \ldots, N\}$$

  - Linear representation of map $\tilde{T}$ (e.g., polynomial or RBF basis)

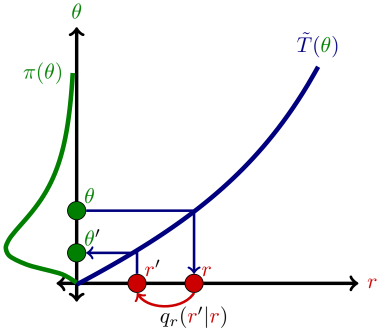
Map-accelerated MCMC

- **Ingredient #1: static map**
  - Idea: perform MCMC in the reference space, on a “preconditioned” density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target



Map-accelerated MCMC

- **Ingredient #1: static map**
  - Idea: perform MCMC in the reference space, on a "preconditioned" density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target

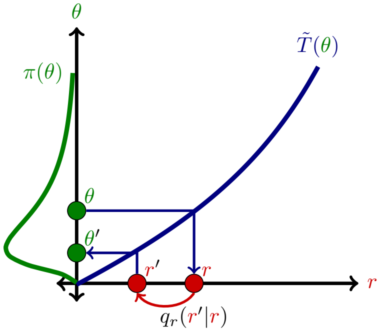


$$\alpha = \frac{\pi(\tilde{T}^{-1}(r')) \,|\, \nabla \tilde{T}^{-1} \,|_{r'} \; q_r(r \,|\, r')}{\pi(\tilde{T}^{-1}(r)) \,|\, \nabla \tilde{T}^{-1} \,|_{r} \; q_r(r' \,|\, r)}$$

simple proposal q_r on pushforward of target through map

Map-accelerated MCMC

- **Ingredient #1: static map**
  - Idea: perform MCMC in the reference space, on a "preconditioned" density
  - Simple proposal in reference space (e.g., random walk) corresponds to a more complex/tailored proposal on target

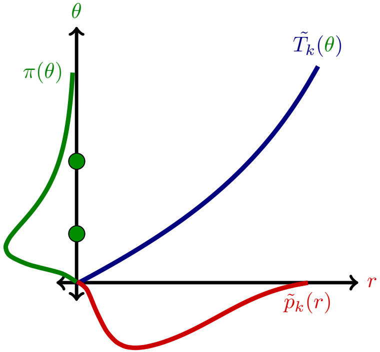


$$\alpha = \frac{\pi(\tilde{T}^{-1}(r')) \,|\, \nabla \tilde{T}^{-1} \,|_{r'}\; q_r(r \,|\, r')}{\pi(\tilde{T}^{-1}(r)) \,|\, \nabla \tilde{T}^{-1} \,|_{r}\; q_r(r' \,|\, r)}$$

more complex proposal, directly on target distribution

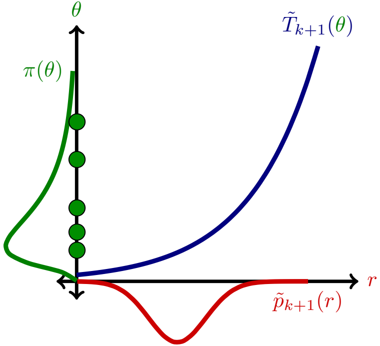
Map-accelerated MCMC

- **Ingredient #2: adaptive map**
  - Update the map with each MCMC iteration: *more samples from π, more accurate $\mathbb{E}_\pi$, better $\tilde{T}$*
  - Analogous to adaptive MCMC [Haario 2001, Andrieu 2006] but with nonlinear transformation to capture non-Gaussian structure



Map-accelerated MCMC

- **Ingredient #2: adaptive map**
  - Update the map with each MCMC iteration:
    more samples from π*, more accurate* $\mathbb{E}_\pi$*, better* $\tilde{T}$
  - Analogous to adaptive MCMC [Haario 2001, Andrieu 2006] but with nonlinear transformation to capture non-Gaussian structure

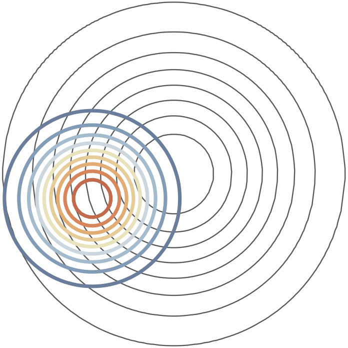


Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior

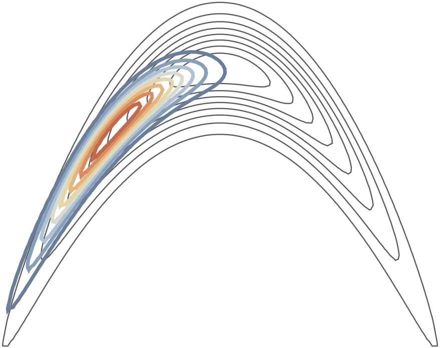
reference random walk proposal

$q_r(r'|r) = N(r, \sigma^2 I)$



mapped random walk proposal

$q_\theta(\theta'|\theta) = q_r\left(\tilde{T}(\theta')|\tilde{T}(\theta)\right)\left|\det D\tilde{T}(\theta')\right|$

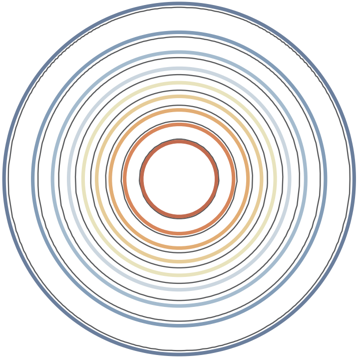


Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior

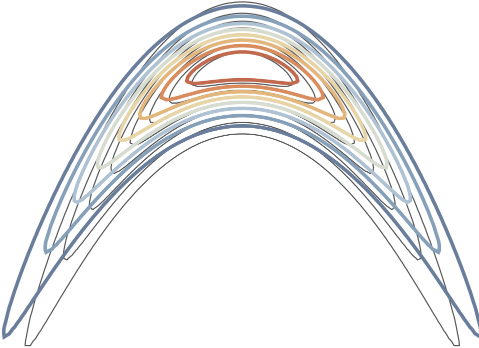
reference independence proposal

$$q_r(r'|r) = N(0, I)$$



mapped independence proposal

$$q_\theta(\theta'|\theta) = q_r\left(\tilde{T}(\theta')|\,\tilde{T}(\theta)\right)\left|\det D\tilde{T}(\theta')\right|$$



Map-accelerated MCMC

- **Ingredient #3: global proposals**
  - If the map becomes sufficiently accurate, would like to avoid random-walk behavior
  - Solution: **delayed rejection** MCMC [Mira 2001]
  - First proposal = independent sample from p (global, more efficient); second proposal = random walk (local, more robust)

- Entire scheme is provably **ergodic** with respect to the exact posterior measure [Parno & M 2015]
  - Requires enforcing a bi-Lipschitz condition on maps, to preserve reasonable tail behavior of target
  - With polynomial maps: revert to *linear* beyond a certain distance from the origin

Example #1: Biological oxygen demand (BOD) model

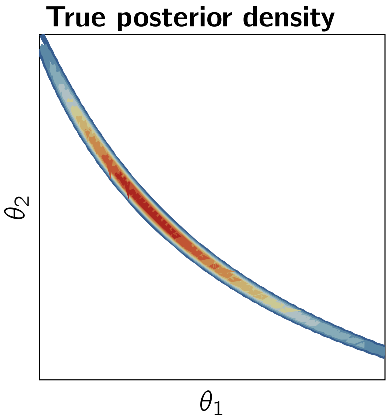
- **Small** inference problem
- Likelihood model:

$$d = \theta_1(1 - \exp(-\theta_2 x)) + \epsilon$$
$$\epsilon \sim N(0, 2 \times 10^{-4})$$

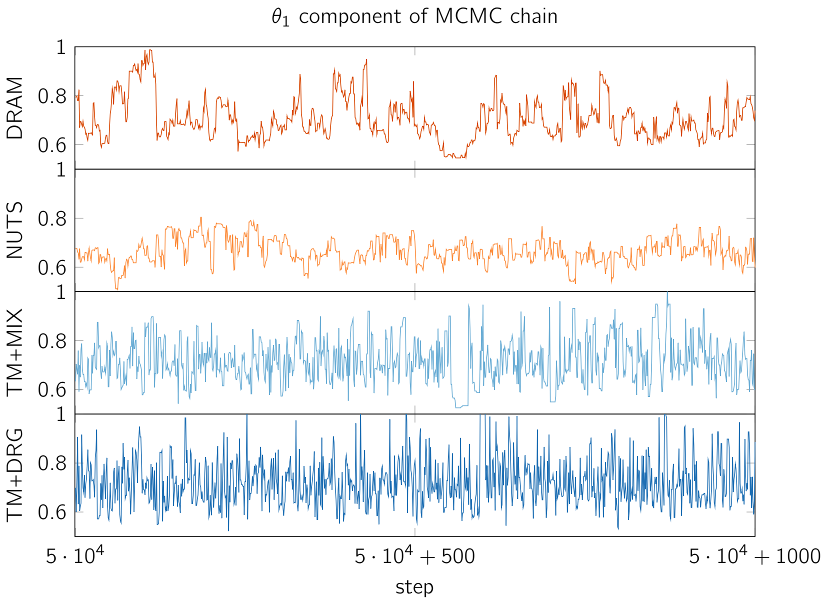
- 20 noisy observations at

$$x = \left\{ \frac{5}{5}, \frac{6}{5}, \ldots, \frac{25}{5} \right\}$$

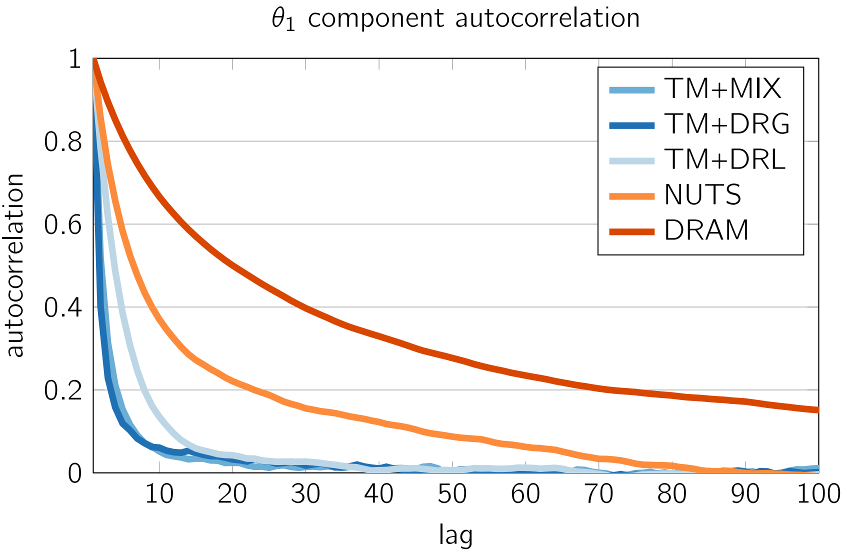
- Third order Hermite polynomial map



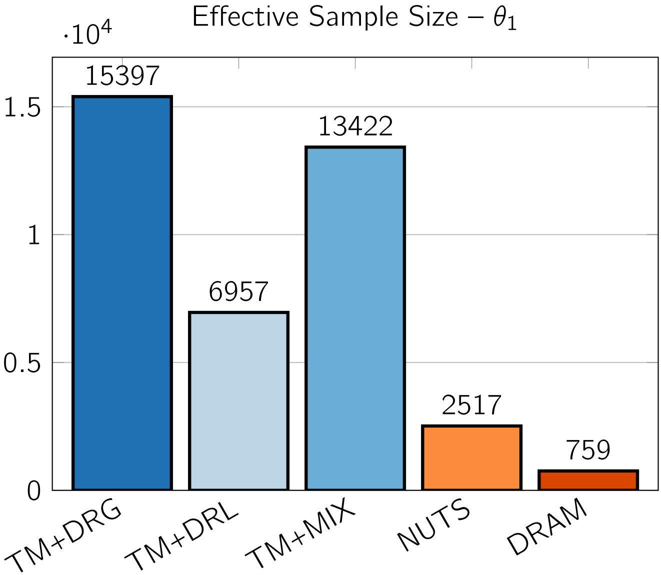
Results: MCMC chain



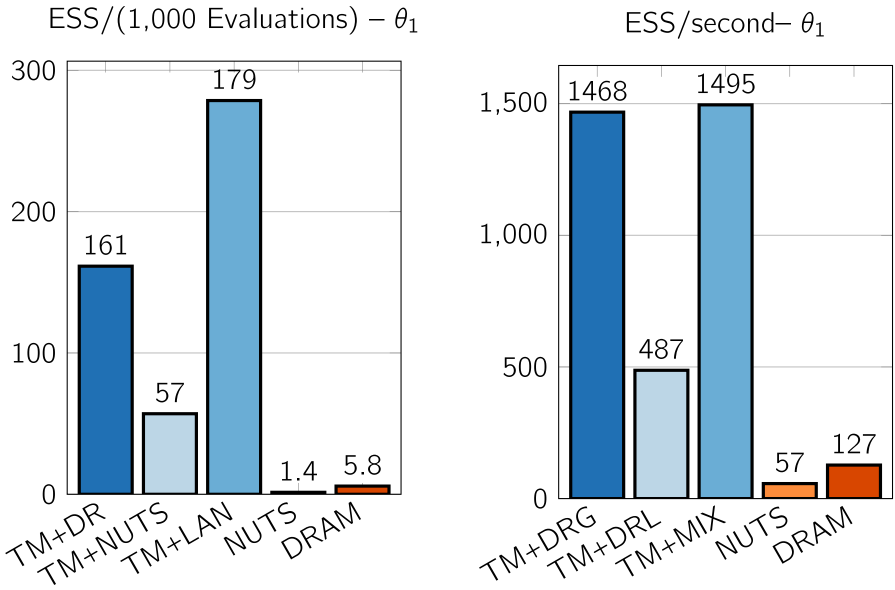
Results: autocorrelation



Results: effective sample size (ESS)



Results: ESS per computational effort



Map-induced distribution

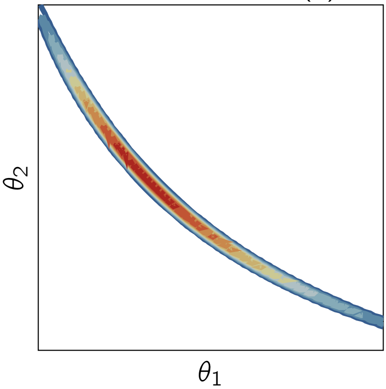
Recall the acceptance ratio:

$$\alpha = \frac{\pi(\tilde{T}^{-1}(r'))|\nabla \tilde{T}^{-1}|_{r'}\, q_r(r|r')}{\pi(\tilde{T}^{-1}(r))|\nabla \tilde{T}^{-1}|_{r}\, q_r(r'|r)}$$

To the standard proposal mechanism, the target looks like:

$$\tilde{p}(r) = \pi(\tilde{T}^{-1}(r))|\nabla \tilde{T}^{-1}|$$

Posterior density $\pi(\theta)$



Map-induced distribution

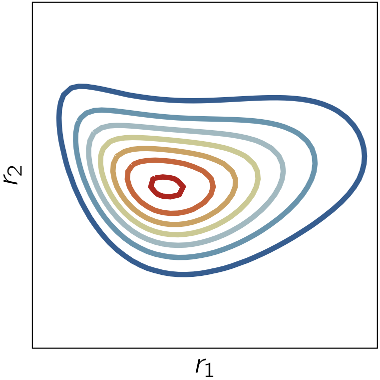
Recall the acceptance ratio:

$$\alpha = \frac{\pi(\tilde{T}^{-1}(r'))|\nabla \tilde{T}^{-1}|_{r'}\, q_r(r|r')}{\pi(\tilde{T}^{-1}(r))|\nabla \tilde{T}^{-1}|_{r}\, q_r(r'|r)}$$

To the standard proposal mechanism, the target looks like:

$$\tilde{p}(r) = \pi(\tilde{T}^{-1}(r))|\nabla \tilde{T}^{-1}|$$

Pushforward of posterior through map $\tilde{p}(r)$



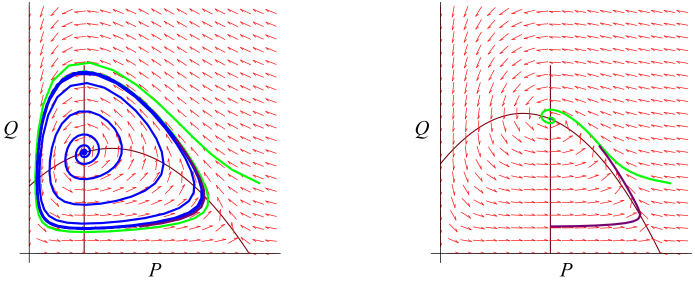
Example #2: predator-prey model

- Six parameter ODE population model

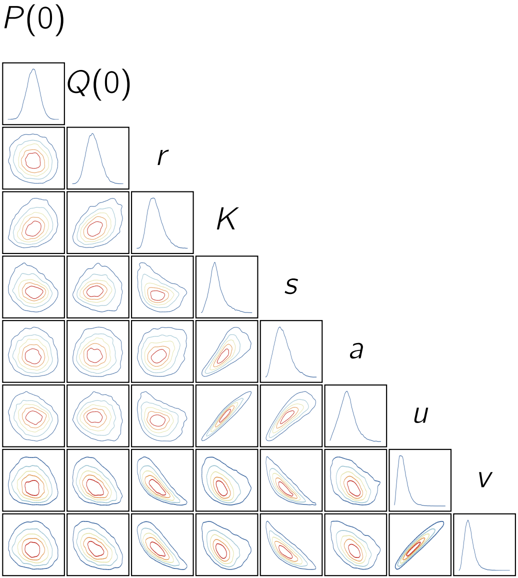
$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) - s\frac{PQ}{a+P}$$

$$\frac{dQ}{dt} = u\frac{PQ}{a+P} - vQ$$

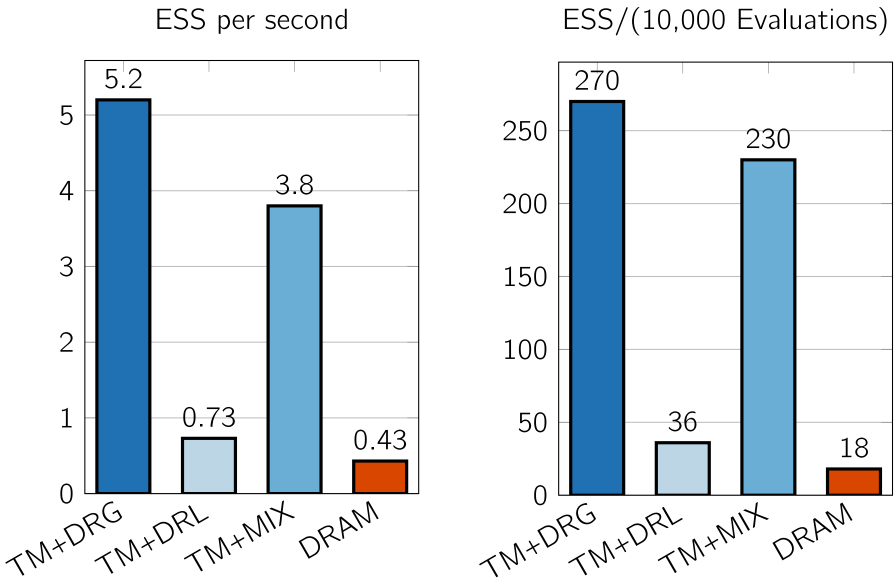
- Ten noisy observations of both populations
- Uniform priors



Predator-prey posterior



Results: ESS per computational effort



Example #3: maple sap dynamics model

- Coupled PDE system for ice, water, and gas locations [Ceseri & Stockie 2013]
- Measure gas pressure in vessel
- Infer 10 physical model parameters
- Very challenging posterior!

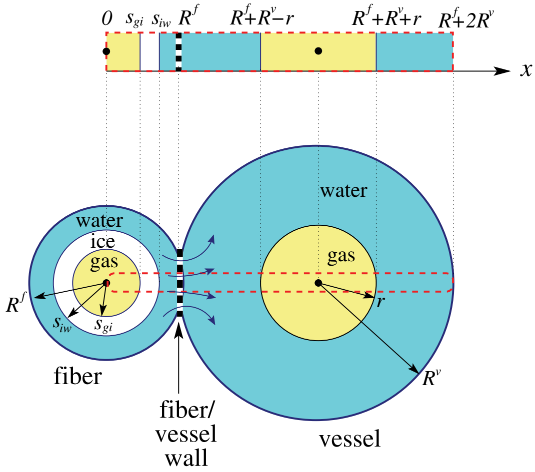
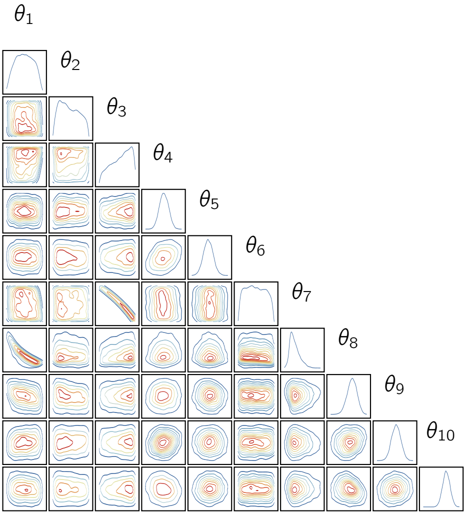
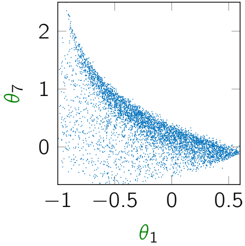
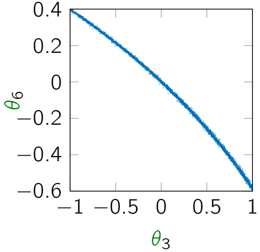


Image from *Ceseri and Stockie, 2013*

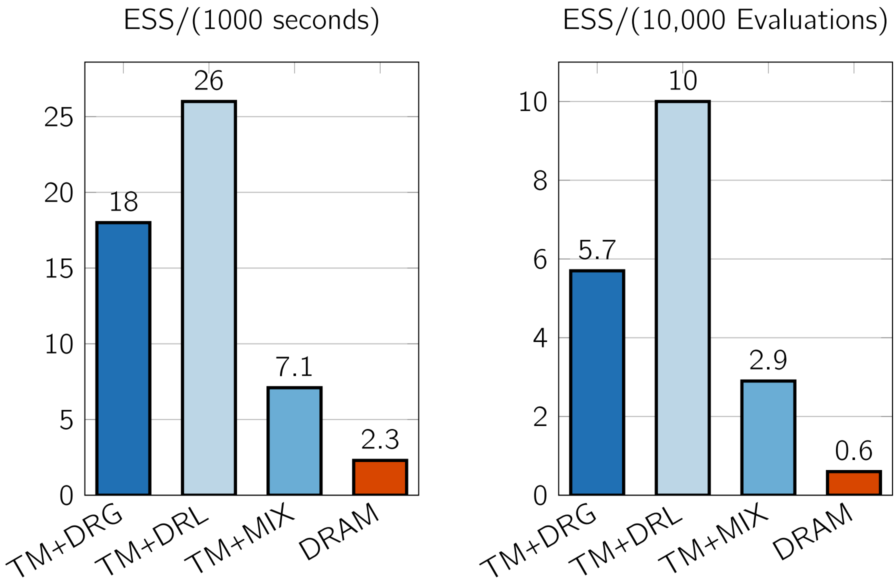
Maple posterior distribution







Results: ESS per computational effort



Comments on MCMC with transport maps

Useful characteristics of the algorithm:

- Map construction is easily parallelizable
- Requires no gradients from posterior density

Generalizes many current MCMC techniques:

- Adaptive Metropolis: map enables **non-Gaussian proposals** *and* a natural mixing between local and global moves
- Manifold MCMC [Girolami & Calderhead 2011]: map defines a Riemannian metric; linear paths in on reference are geodesics on target

Map construction from samples:

- Links with density estimation approaches of [Tabak 2011–14] and iterative Gaussianization/ICA of [Laparra *et al.* 2011]

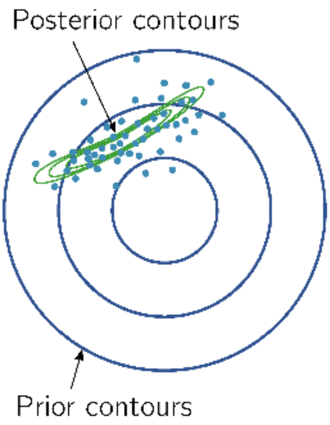
Next steps

- Maps in **high dimensions**
  - Use notion of a *likelihood-informed subspace*, cf. dimension independent likelihood-informed (DILI) MCMC [Cui, Law, & M 2015]; map departs from the identity only in data-informed directions
  - Compose rotations and diagonal maps: basis representation is more scalable than triangular (Rosenblatt) maps

- More fundamentally: relate **structure** of transport maps to essential properties of target distribution
  - Current work: conditional independence (Markov structure) of the target distribution π implies minimal *sparsity* of the inverse map, yields efficient algorithms for *ordering* and *decomposition*

Part 2: Computationally intensive models

- **Surrogates** for $\mathbf{f}$ or $\mathcal{L}$ are very useful for Bayesian inference in this setting...
- *Posterior-focused* surrogates can improve efficiency
  - Posterior-focused polynomial chaos approach [Li & M, SISC 2014]
  - Data-driven model reduction [Cui, M, & Willcox IJNME 2014]
  - RBF approximations [Bliznyuk *et al.* 2012, Joseph 2012]
- In general, samples are then drawn from an **approximate** posterior
- Approximation cost borne *a priori;* must balance with sampling error



Asymptotically exact MCMC via local approximations

Sampling from the **exact** posterior:

- Delayed-acceptance schemes [Christen & Fox 2005]: at least one full model evaluation per accepted sample
- We take a different approach: *asymptotically exact* MCMC, via incremental and infinite refinement of surrogates
  - Posterior exploration and surrogate construction occur *simultaneously*
  - Asymptotic exactness: convergence of surrogate tied to stationarity of the MCMC chain
  - **Joint work** with Patrick Conrad (MIT), Natesh Pillai (Harvard), Aaron Smith (Ottawa)

MCMC with a surrogate and posterior adaptation

Given X_0, initialize a sample set $\mathcal{S}_0$, then simulate chain $\{X_t\}$ with kernel:

MH Kernel $K_t(x, \cdot)$

1. Given X_t, draw $q_t \sim Q(X_t, \cdot)$ from kernel Q with (symmetric) translation invariant density $q(x, \cdot)$
2. Compute acceptance ratio
$$\alpha = \min\left(1, \frac{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t(q_t))p(q_t)}{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t(X_t))p(X_t)}\right)$$
3. As needed, select new samples near q_t or X_t, yielding $\mathcal{S}_t \subseteq \mathcal{S}_{t+1}$. Refine $\tilde{\mathbf{f}}_t \to \tilde{\mathbf{f}}_{t+1}$.
4. Draw $u \sim \mathcal{U}(0, 1)$. If $u < \alpha$, let $X_{t+1} = q_t$, otherwise $X_{t+1} = X_t$.

- Approximation $\tilde{\mathbf{f}}_t$ built from sample set $\mathcal{S}_t = \{\theta_i : \mathbf{f}(\theta_i) \text{ has been run}\}$
- Continue adaptation forever (as $t \to \infty$)

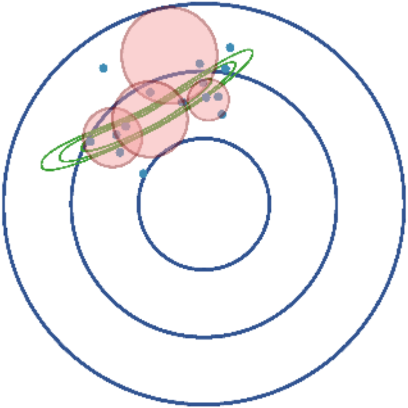
Local approximations

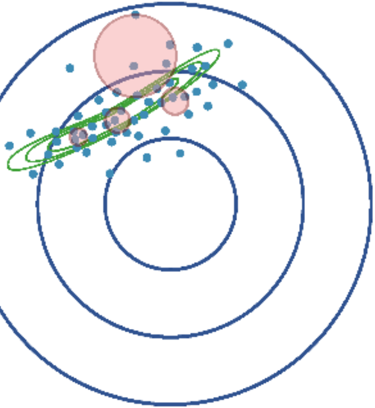
- To compute the approximation $\tilde{\mathbf{f}}(\theta)$, construct a model over the ball $\mathcal{B}_R(\theta)$
- Use samples $\theta_i \in \mathcal{S}$ at distance $r = \|\theta - \theta_i\|$ with weight
$$w(r) = \begin{cases} 0 < w'(r) \leq 1 & r \leq R \\ 0 & \text{else} \end{cases}$$
- Approximations converge locally under loose conditions
  - For example, quadratic approximations over $\mathcal{B}_R(\theta)$ [Conn *et al.*]:
$$\|\mathbf{f} - \mathcal{Q}_R \mathbf{f}\| \leq \kappa(\nu, \lambda, d) R^3$$
- Choose R so that $M(d)$ samples have non-zero weight, e.g., where $M(d)$ ensures that a quadratic is fully determined

Local approximation illustration

earlier times

later times





Experimental design: triggering refinement

1. Random refinement β_t
   - With probability β_t, such that $\sum_t \beta_t = \infty$, refine near X_t or q_t
2. Acceptance probability error indicator γ_t
   - Estimate error in acceptance ratio using cross-validation

$$\alpha_i^+ = \min\left(1, \frac{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t^{\sim i}(q_t))p(q_t)}{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t(X_t))p(X_t)}\right) \quad \alpha_i^- = \min\left(1, \frac{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t(q_t))p(q_t)}{\mathcal{L}(\mathbf{d}, \tilde{\mathbf{f}}_t^{\sim i}(X_t))p(X_t)}\right)$$

   - Compute error indicators

$$\epsilon^+ = \max_i |\alpha - \alpha_i^+| \qquad \epsilon^- = \max_i |\alpha - \alpha_i^-|$$

   - Refine if $\epsilon^+ > \gamma_t$ or $\epsilon^- > \gamma_t$

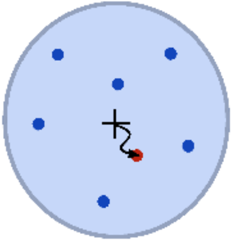
Experimental design: performing refinement

Local space filling refinement

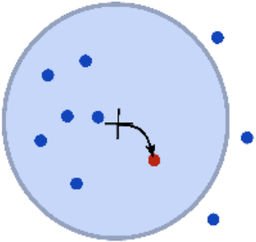
To space fill near $\xi_t = X_t$ or $\xi_t = q_t$, given radius R, locally solve

$$\theta^* = \underset{\|\xi_t - \theta'\|_2 \leq R}{\arg\max} \; \min_{\theta_i \in \mathcal{S}_t} \|\theta' - \theta_i\|_2$$

beginning at ξ_t and add $\theta^* \to \mathcal{S}_{t+1}$



Closer points



Filling in directions

Ergodicity of approximate samplers

Theorem (Conrad, M, Pillai, Smith 2014)

Assume the log-posterior is approximated with local quadratic models and that $\theta \in \mathcal{X} \subseteq \mathbb{R}^d$ for *compact* $\mathcal{X}$, or that $p(\theta|\mathbf{d})$ obeys a *Gaussian envelope* condition

$$\lim_{r\to\infty} \sup_{|\theta|=r} |\log p(\theta|\mathbf{d}) - \log p_\infty(\theta)| = 0$$

for some quadratic form $\log p_\infty$ with negative definite coefficient matrix.

Then under standard regularity assumptions for geometrically ergodic kernel K_∞ and posterior $p(\theta|\mathbf{d})$, the chain X_t is **ergodic** for the **exact posterior**:

$$\lim_{t\to\infty} \|\mathbb{P}(X_t) - p(\theta|\mathbf{d})\|_{TV} = 0$$

A framework for approximate samplers

Many algorithmic variations:

- Target of approximation
  - Forward model: $\mathbf{f}(\boldsymbol{\theta})$
  - Log-likelihood: $\log \mathcal{L}(\mathbf{d}, \mathbf{f}(\boldsymbol{\theta}))$
- Types of local approximations
  - Regression with low-order polynomials
  - Gaussian process regression
  - Quadratic regression given derivatives $\partial \mathbf{f} / \partial \theta$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Gradient-based proposals (e.g., MALA, manifold MALA, stochastic Newton)
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

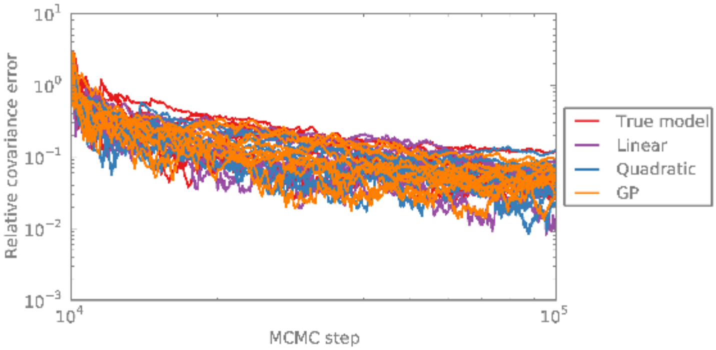
A framework for approximate samplers

Many algorithmic variations:

- Target of approximation
  - Forward model: $\mathbf{f}(\boldsymbol{\theta})$
  - Log-likelihood: $\log \mathcal{L}(\mathbf{d}, \mathbf{f}(\boldsymbol{\theta}))$
- Types of local approximations
  - Regression with low-order polynomials
  - Gaussian process regression
  - Quadratic regression given derivatives $\partial \mathbf{f} / \partial \theta$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Gradient-based proposals (e.g., MALA, manifold MALA, stochastic Newton)
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

Example: elliptic PDE inverse problem

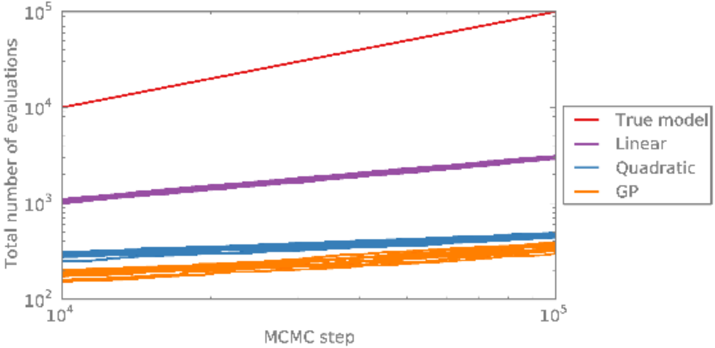
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i, $d = 6$.



Accuracy of chains

Example: elliptic PDE inverse problem

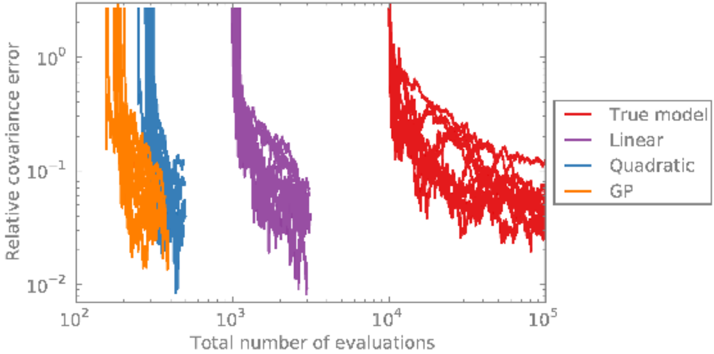
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i, $d = 6$.



Cost of chains

Example: elliptic PDE inverse problem

- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i-1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i, $d = 6$.



Accuracy versus cost

A framework for approximate samplers

Many algorithmic variations:

- Target of approximation
  - Forward model: $\mathbf{f}(\boldsymbol{\theta})$
  - Log-likelihood: $\log \mathcal{L}(\mathbf{d}, \mathbf{f}(\boldsymbol{\theta}))$
- Types of local approximations
  - Regression with low-order polynomials
  - Gaussian process regression
  - Quadratic regression given derivatives $\partial \mathbf{f} / \partial \theta$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Gradient-based proposals (e.g., MALA, manifold MALA, stochastic Newton)
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

Groundwater tracer transport model

- Nonlinear PDE for hydraulic head

$$\nabla \cdot (h\kappa \nabla h) = -f_h$$

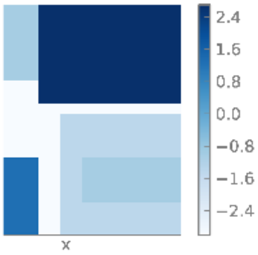
- Darcy velocity $(u, v) = -h\kappa\nabla h$ then enters tracer transport equation:

$$\frac{\partial c}{\partial t} + \nabla \cdot \left(\left(d_m \mathbf{I} + d_l \begin{bmatrix} u^2 & uv \\ uv & v^2 \end{bmatrix} \right) \nabla c \right) - \begin{bmatrix} u \\ v \end{bmatrix} \cdot \nabla c = -f_t,$$

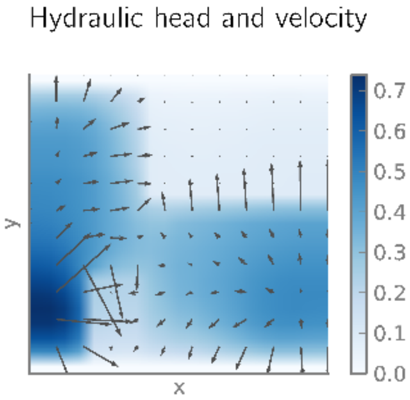
- Tracer advects according to velocity and well forcing

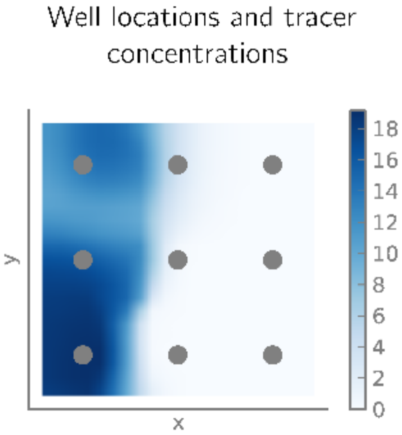
- Observe tracer concentration at well locations, at several times, with Gaussian error
- Infer for piecewise constant conductivities, given log-normal priors
- Forward model takes about 6 seconds to evaluate

Log-conductivity field ($\log \kappa$)

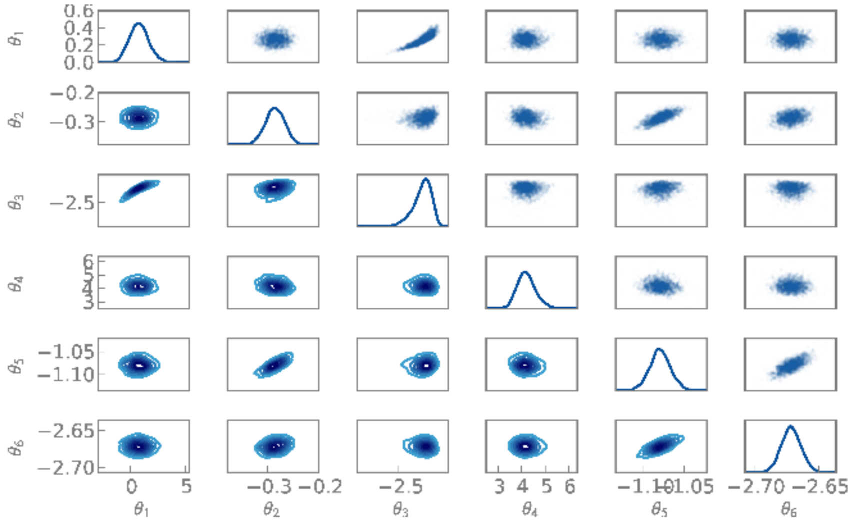


Groundwater tracer transport problem

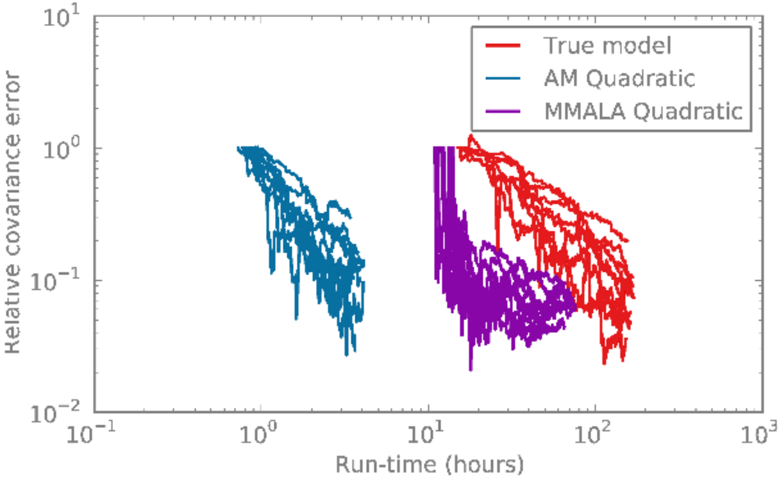




Groundwater hydrology problem: posterior distribution



Single chain performance

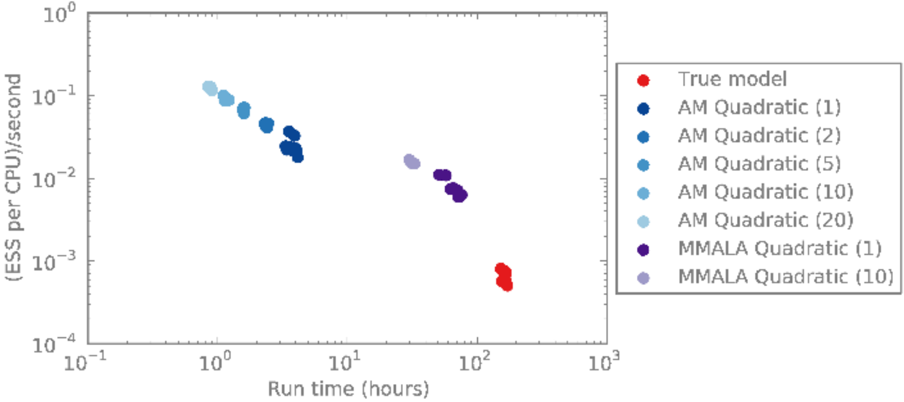


Implementing approximation across parallel chains

- Build a common pool of model runs $\mathcal{S}$ across parallel workers
- Since approximation targets the correct distribution, use *effective sample size (ESS)* to measure efficiency
- ESS per (CPU-second) would be constant with a naïve implementation

- Run N chains of 100,000 steps each
- Discard 10% of each chain as burn-in; evaluate ESS

Parallel efficiency



Conclusions

- Combining **transport maps** with MCMC to accelerate Bayesian computation in non-Gaussian settings
  - *Underlying idea:* Approximate complex distributions via deterministic transformations of a Gaussian distribution
- Introduced a new framework for using **local approximations** within MCMC; proved that the framework produces **asymptotically exact samples**
  - *Underlying idea:* Regularity of the likelihood enables far fewer model evaluations than direct MCMC
- Much ongoing work...
  - Scaling local approximations to high dimensions
  - Building maps in high dimensions
  - Scalable **direct map** (MCMC-free) approaches

References

- Both algorithms implemented in MUQ (MIT Uncertainty Quantification library), `http://muq.mit.edu`

- M. Parno, Y. Marzouk, "Transport map accelerated Markov chain Monte Carlo." Submitted (2015). arXiv:1412.5492.
- P. Conrad, Y. Marzouk, N. Pillai, A. Smith, "Accelerating asymptotically exact MCMC for computationally intensive models via local approximations." *J. Amer. Statist. Assoc.*, in press (2015). arXiv:1402.1694.
- M. Parno, T. Moselhy, Y. Marzouk. "A multiscale strategy for Bayesian inference using transport maps." Submitted (2015). arXiv:1507.07024.
- T. Moselhy, Y. Marzouk, "Bayesian inference with optimal maps." *J. Comp. Phys.*, 231: 7815–7850 (2012).