Local approximation MCMC for computationally intensive models

[1]Youssef Marzouk
joint work with [1]Patrick Conrad, [1]Andrew Davis,
[2]Natesh Pillai, [3]Aaron Smith

[1]Department of Aeronautics and Astronautics
Center for Computational Engineering & Center for Statistics
Massachusetts Institute of Technology
http://uqgroup.mit.edu

[2]Department of Statistics
Harvard University

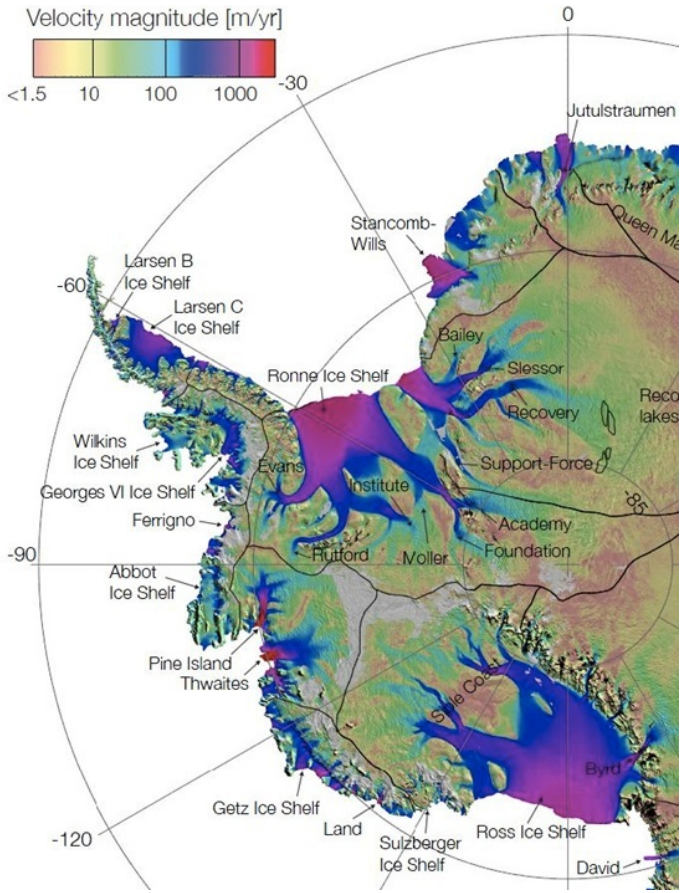
[3]Department of Mathematics and Statistics
University of Ottawa

8 June 2016

Inference with computationally intensive models

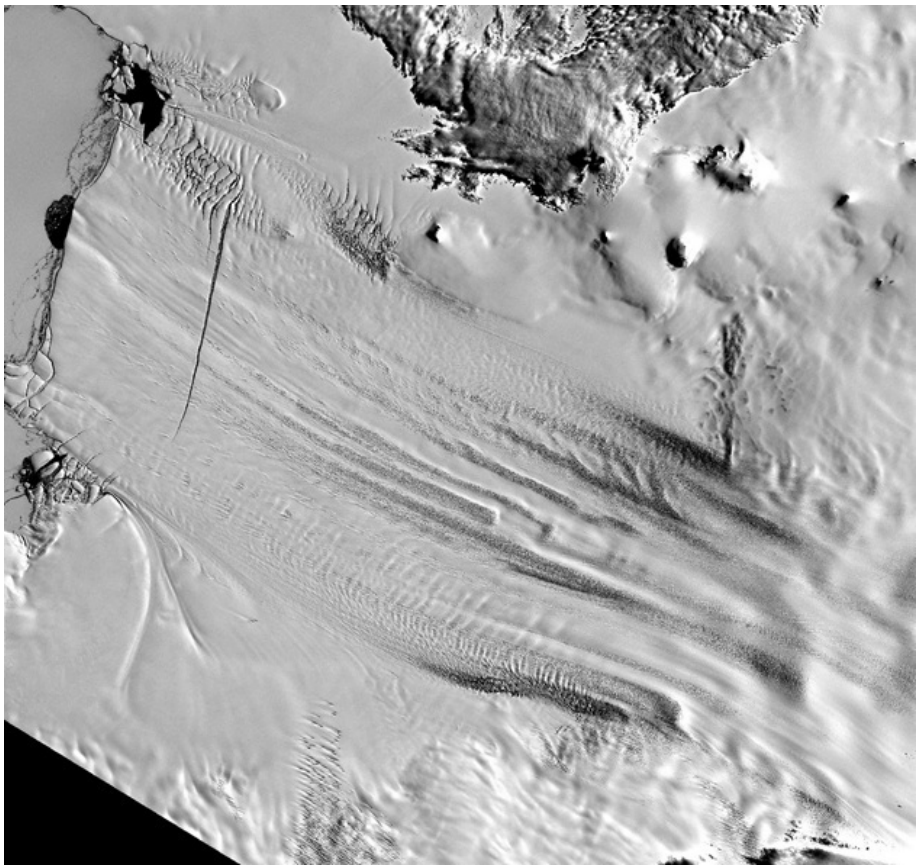
A terrestrial example: ice sheet dynamics in western Antarctica

Western Antarctic Ice Sheet



[Rignot et al. 2011]

Pine Island Glacier



[NASA]

Bayesian inference setting

Posterior density of the parameters

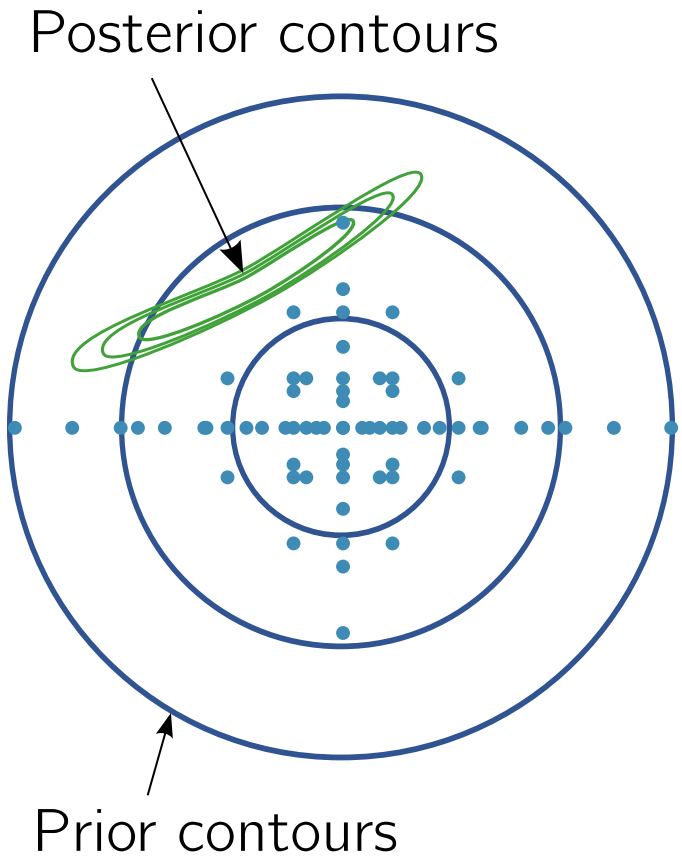
$$\pi(\theta) := p(\theta|\mathbf{d}) \propto \mathcal{L}(\mathbf{d}, \mathbf{f}(\theta))p(\theta)$$

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 $\mathbf{f}$

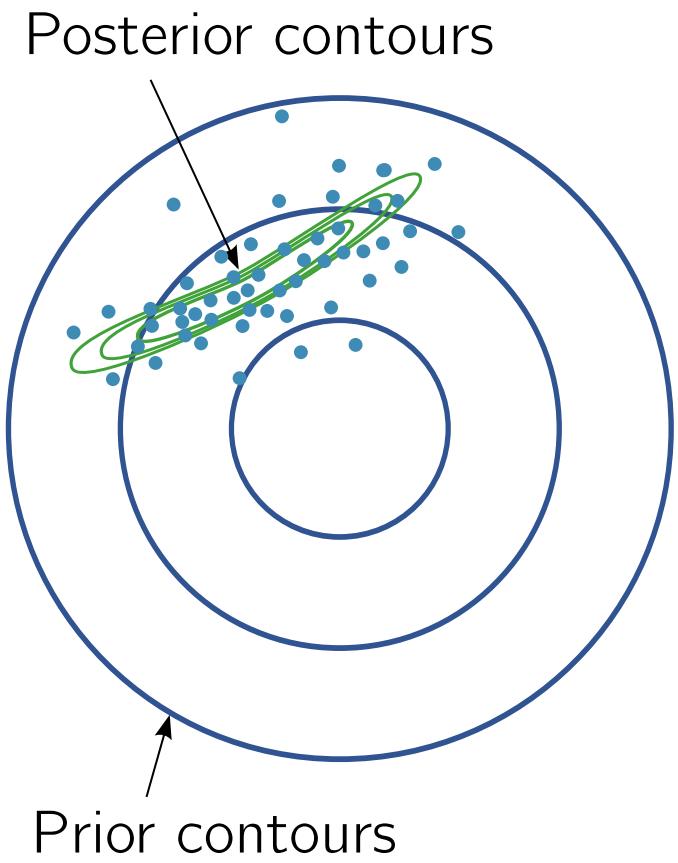
Accelerating MCMC with computationally intensive models

- **Surrogates** for $\mathbf{f}$ or $\mathcal{L}$ are very useful for Bayesian inference in this setting...
- Simple approach: construct an approximation of $\mathbf{f}$ or $\mathcal{L}$ over the *prior* distribution
  - Convergence of this approximation (e.g., in L^2_p) yields convergence to the true posterior



Accelerating MCMC with computationally intensive models

- *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

Metropolis-Hastings MCMC

Given X_0, simulate chain $\{X_t\}_{t \leq N}$ according to transition kernel:

MH Kernel $K_\infty(x, \cdot)$

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
$$\alpha = \min\left(1, \frac{\mathcal{L}(\mathbf{d}, \mathbf{f}(q_t))p(q_t)}{\mathcal{L}(\mathbf{d}, \mathbf{f}(X_t))p(X_t)}\right)$$
3. Draw $u \sim \mathcal{U}(0, 1)$. If $u < \alpha$, let $X_{t+1} = q_t$, otherwise $X_{t+1} = X_t$.

- Evaluates forward model N times
- Run-time can be dominated by cost of $\mathbf{f}$

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

$$w(r) = \begin{cases} 0 < w'(r) \leq 1 & r \leq R \\ 0 & \text{else} \end{cases}$$

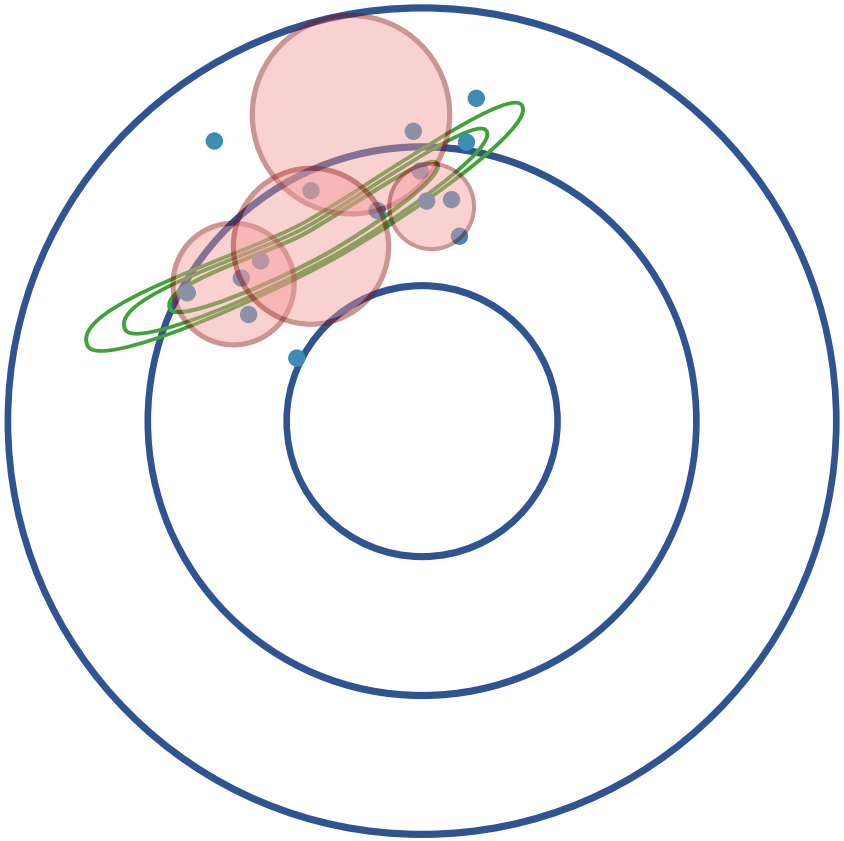
- Choose R so that $M(d)$ samples have non-zero weight, e.g., where $M(d)$ ensures that a quadratic is fully determined

- Approximations converge locally under loose conditions (e.g., $\mathbf{f}$ continuously differentiable with Lipschitz gradients)
  - For example, quadratic approximations over $\mathcal{B}_R(\theta)$ [Conn *et al.*]:

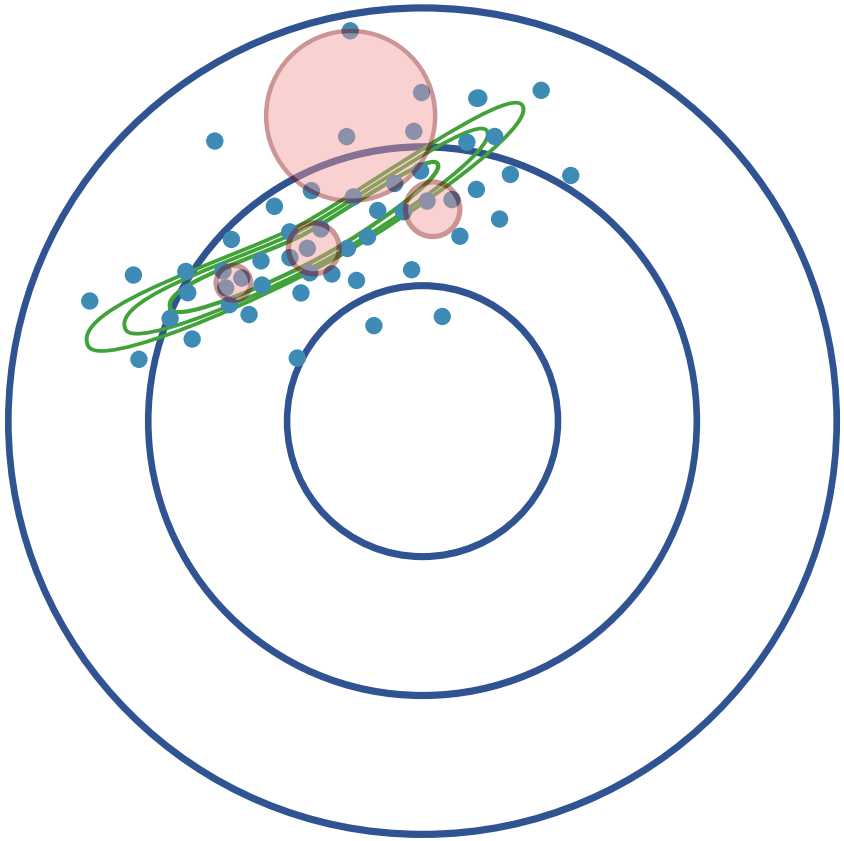
$$\|\mathbf{f} - \mathcal{Q}_R\mathbf{f}\| \leq \kappa(\nu, \lambda, d)R^3$$

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 $\epsilon^- > \gamma_t$

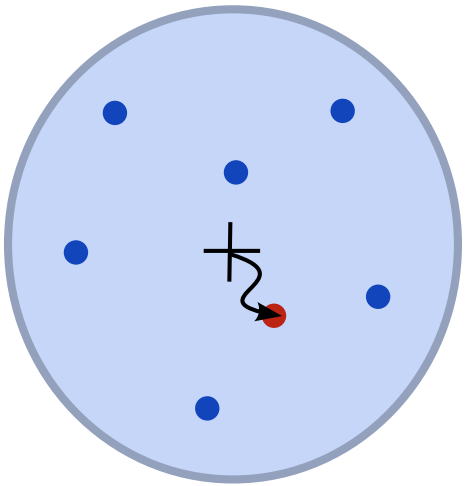
Experimental design: performing refinement

Local space filling refinement

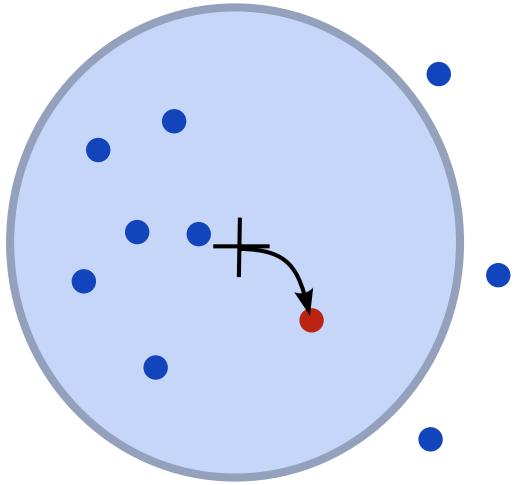
To space fill near $\xi_t = X_t$ or $\xi_t = q_t$, given radius R, locally solve

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

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



Closer points



Filling in directions

- Alternative approach: use [Moré & Sorensen 1983] to add a new point while explicitly controlling poisedness

Ergodicity of approximate samplers

Theorem (Conrad, M, Pillai, Smith 2015)

Let the log-posterior be approximated with local quadratic models; assume that $\theta \in \mathcal{X} \subseteq \mathbb{R}^d$ for *compact* $\mathcal{X}$, or that $\pi(\theta) := p(\theta|\mathbf{d})$ obeys a *Gaussian envelope* condition,

$$\lim_{r\to\infty} \sup_{|\theta|=r} |\log \pi(\theta) - \log p_\infty(\theta)| = 0,$$

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

Then, under standard regularity assumptions for geometrically ergodic kernel K_∞ and posterior π, the chain X_t is **ergodic** for the **exact posterior**:

$$\lim_{t\to\infty} \|\mathbb{P}(X_t) - \pi\|_{TV} = 0$$

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 $\partial_\theta \mathbf{f}$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Proposals (e.g., MALA, manifold MALA, HMC) that extract *derivative information* from the approximation
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

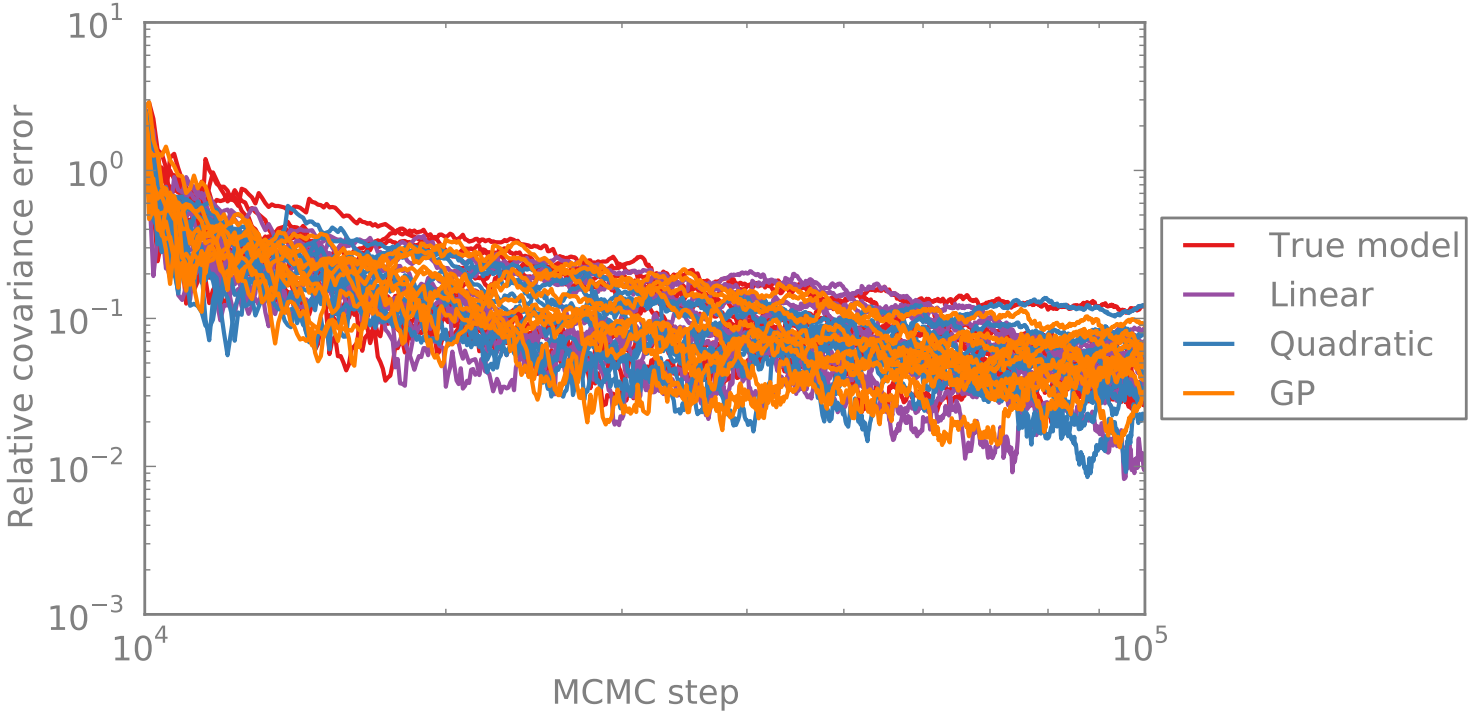
A framework for approximate samplers

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 $\partial_\theta \mathbf{f}$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Proposals (e.g., MALA, manifold MALA, HMC) that extract *derivative information* from the approximation
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

Example: elliptic PDE inverse problem

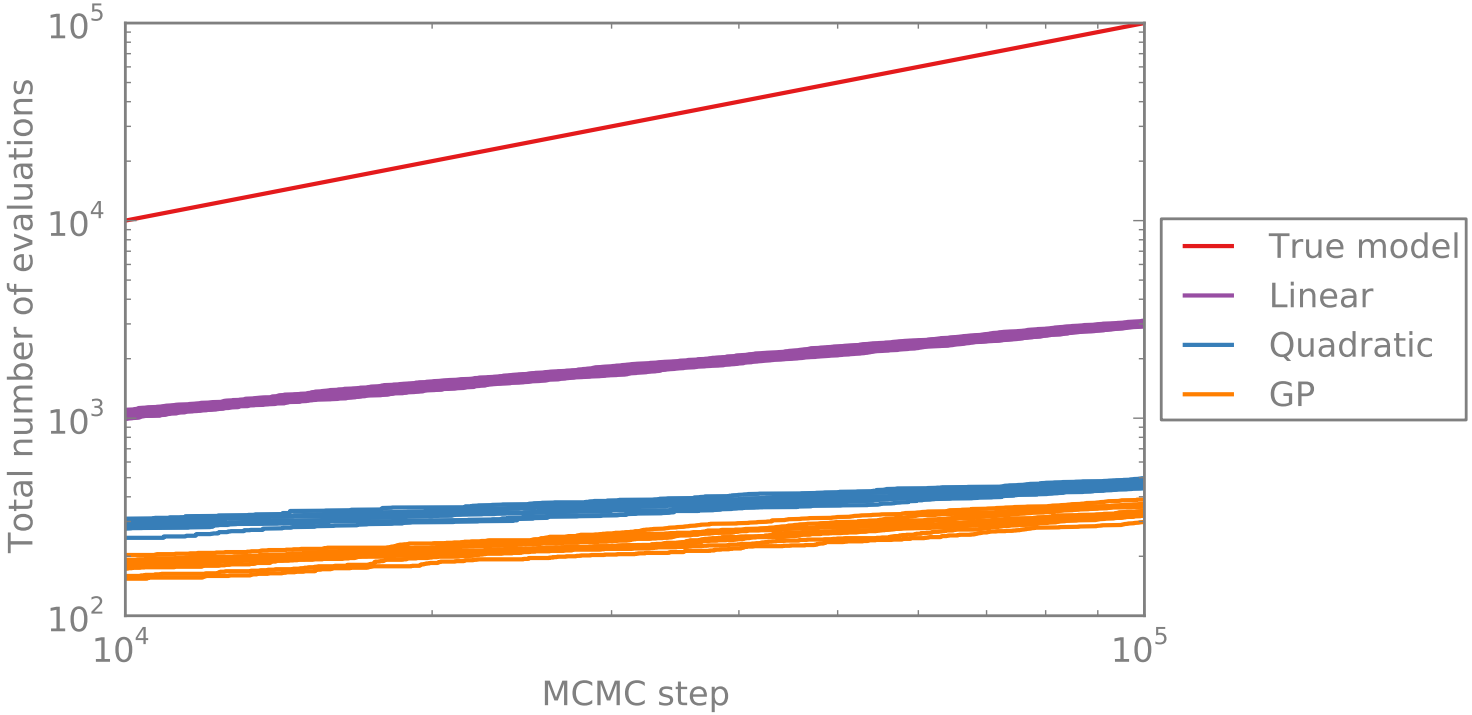
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i.



Accuracy of chains

Example: elliptic PDE inverse problem

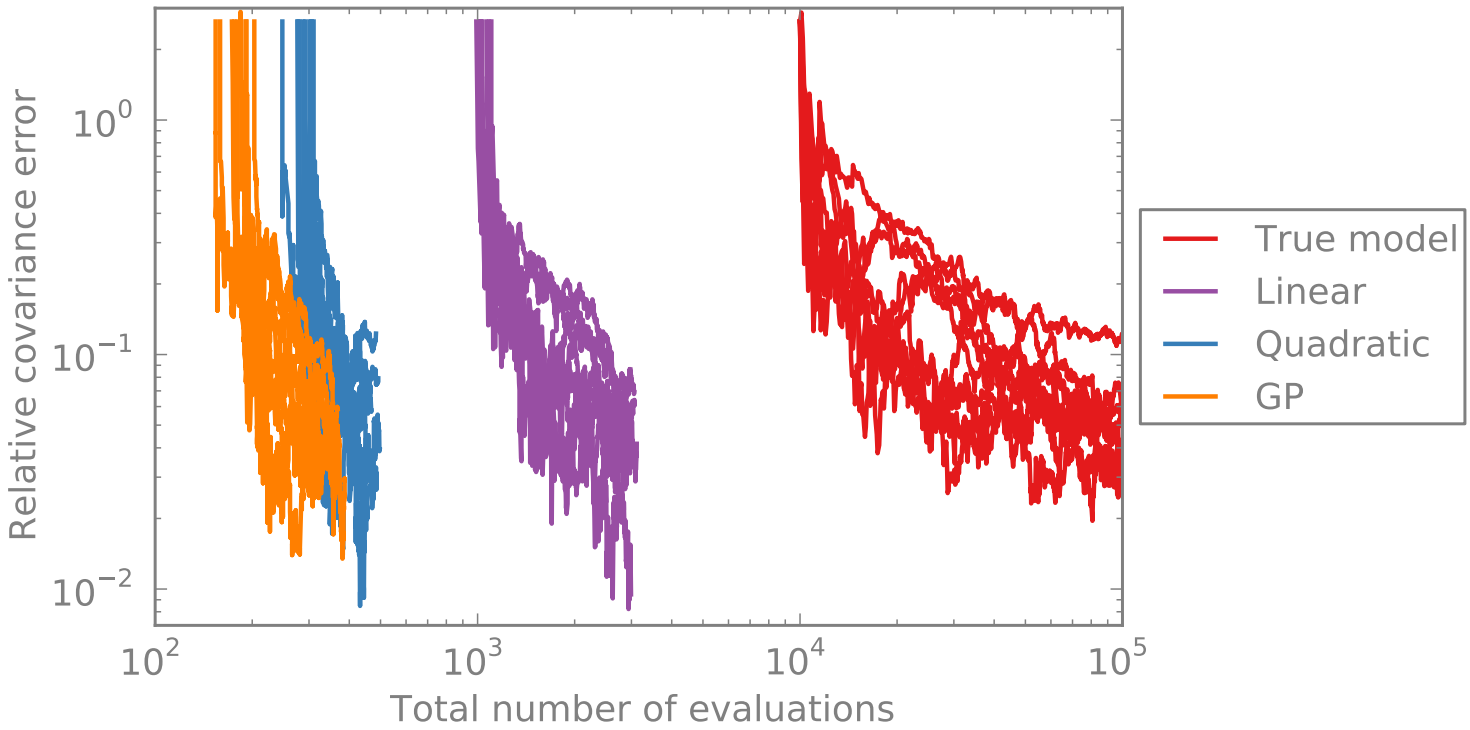
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i.



Cost of chains

Example: elliptic PDE inverse problem

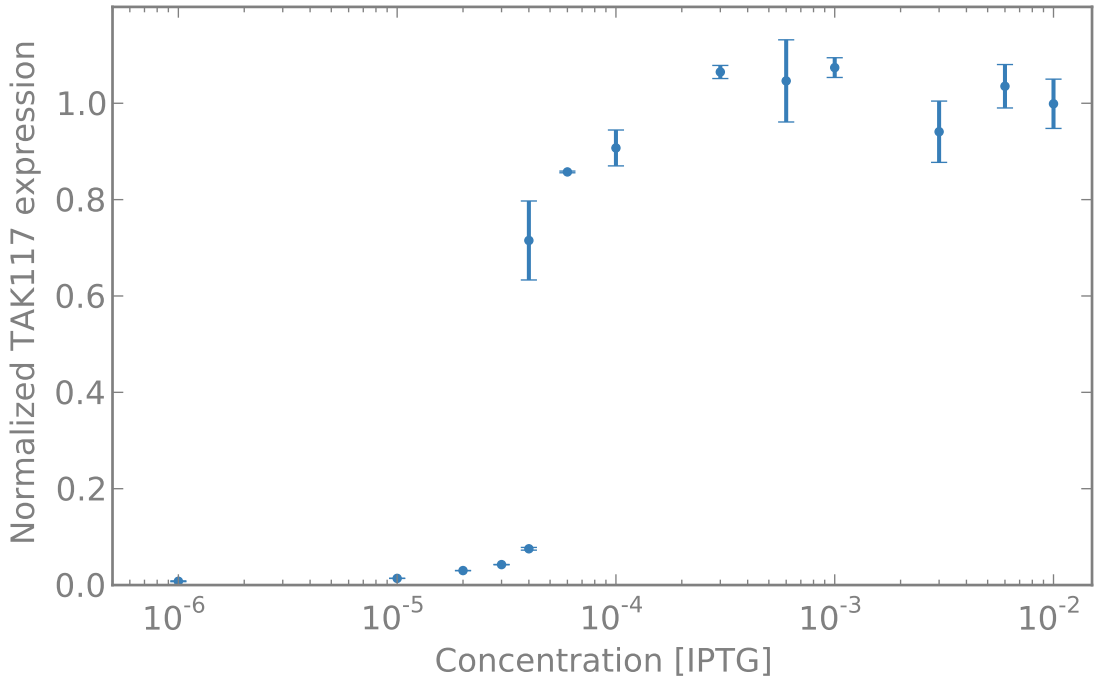
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i.



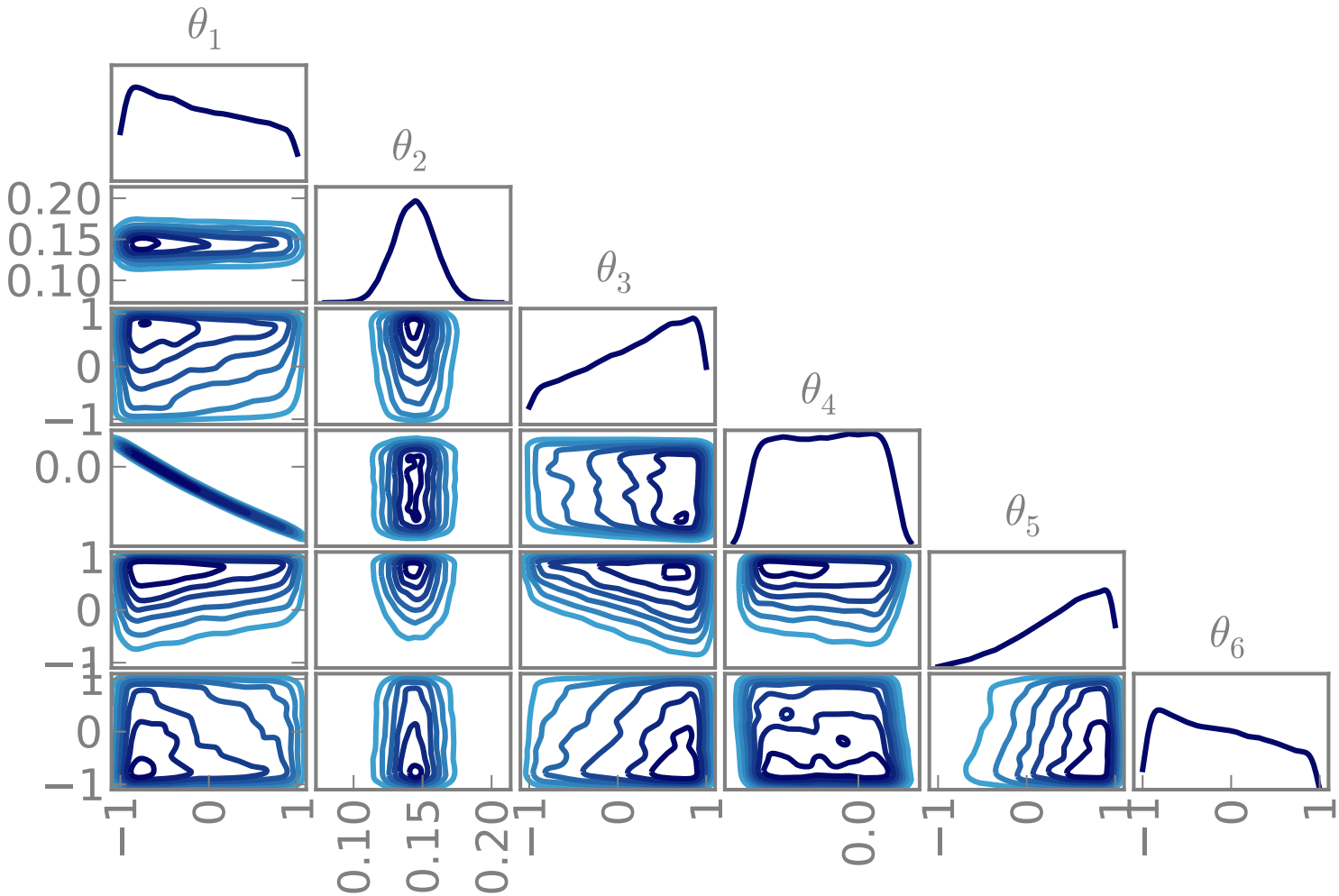
Accuracy versus cost

Example: genetic toggle switch

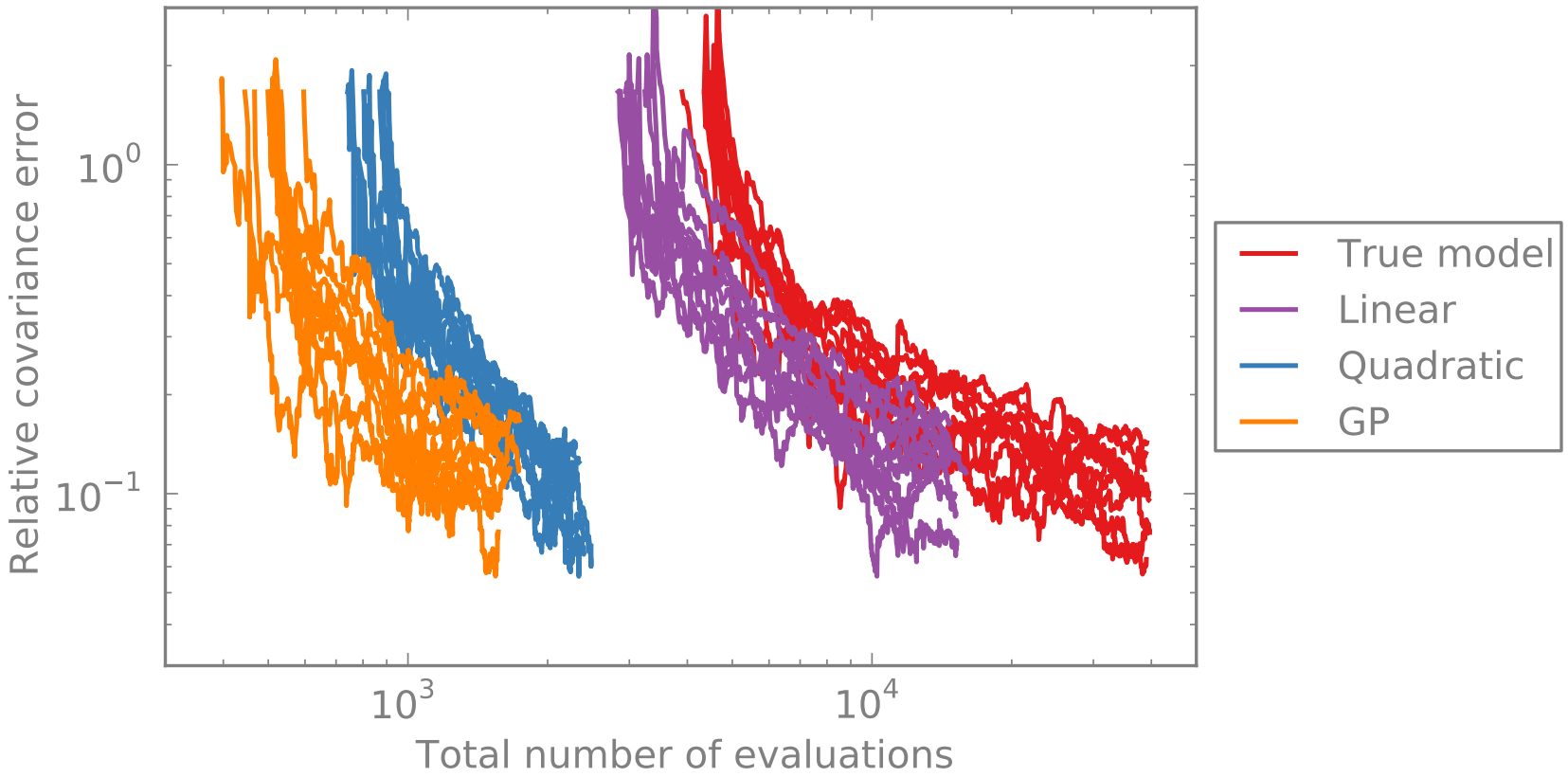
- Model for genetic "toggle switch" synthesized in *E. coli*
- ODE system, six parameters to infer
- Uniform priors, Gaussian observational errors
- Real experimental data



Genetic toggle switch posterior

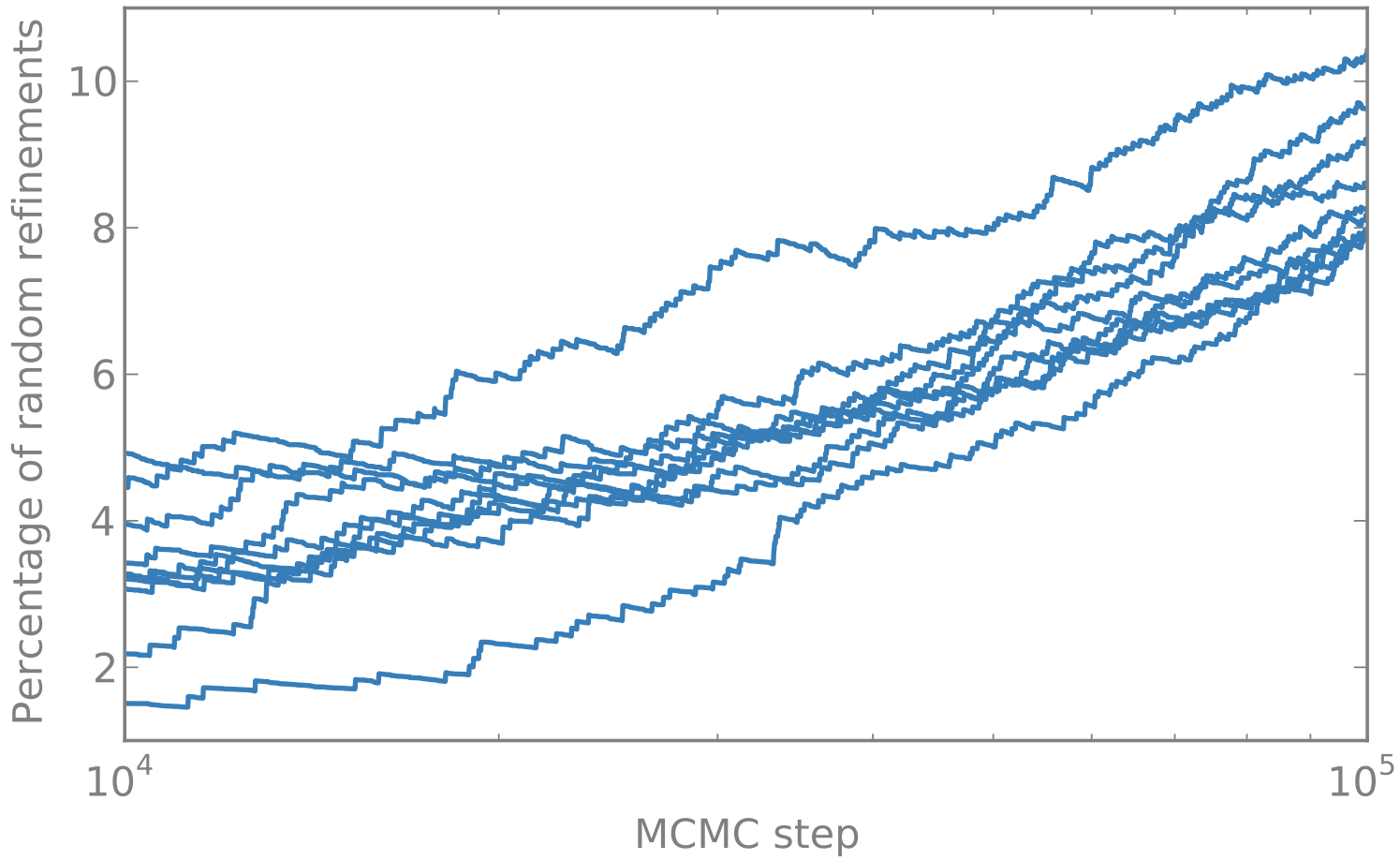


Genetic toggle switch: accuracy versus cost



Error indicator versus random refinements

Percentage of refinements triggered randomly (β_t), rather than by error indicator (γ_t):



$$\gamma_t = 0.1t^{-0.1},\ \beta_t = 0.01t^{-0.2}$$

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 $\partial_\theta \mathbf{f}$
- MCMC kernels
  - Random-walk Metropolis, adaptive Metropolis
  - Proposals (e.g., MALA, manifold MALA, HMC) that extract *derivative information* from the approximation
- Parallel chains, sharing a common pool of model evaluations $\mathcal{S}$

Groundwater tracer transport model

- Nonlinear PDE for hydraulic head

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

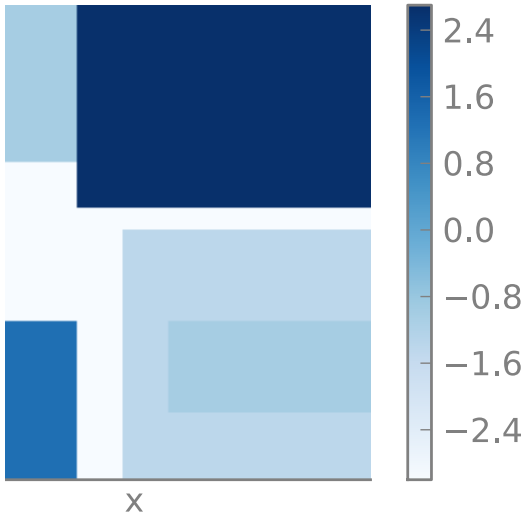
- Darcy velocity $(u, v) = -h\kappa\nabla h$ then enters tracer transport equation:

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

- Tracer advects according to velocity and well forcing

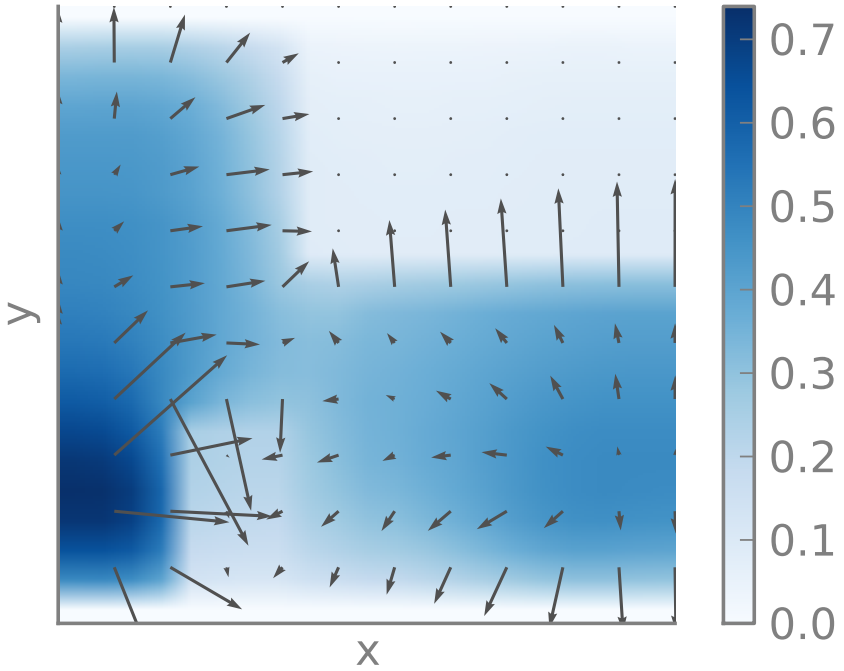
- Observe tracer concentration at well locations, at several times, with Gaussian error
- Infer for piecewise constant conductivities, given log-normal priors
- Forward model takes about 6 seconds to evaluate

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

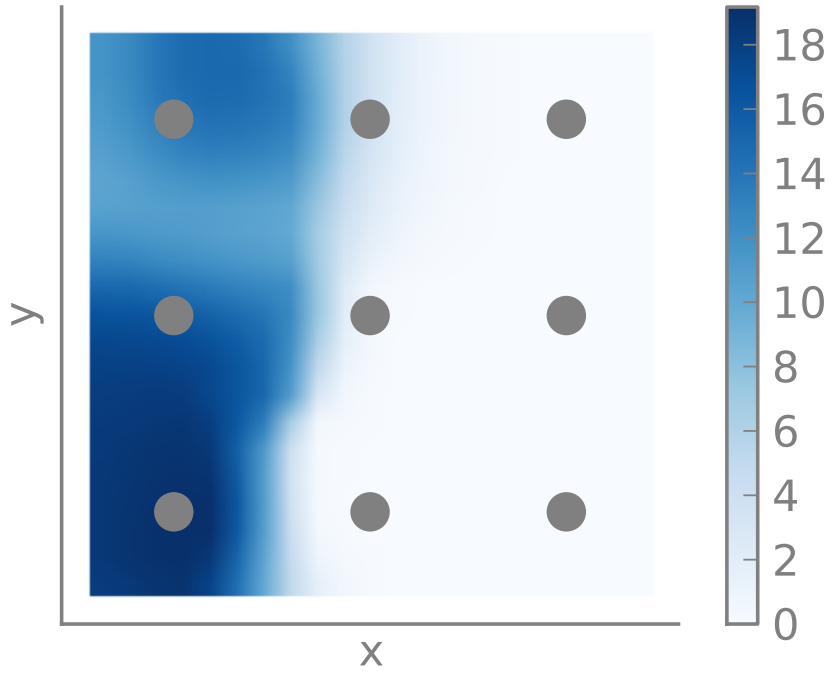


Groundwater tracer transport problem

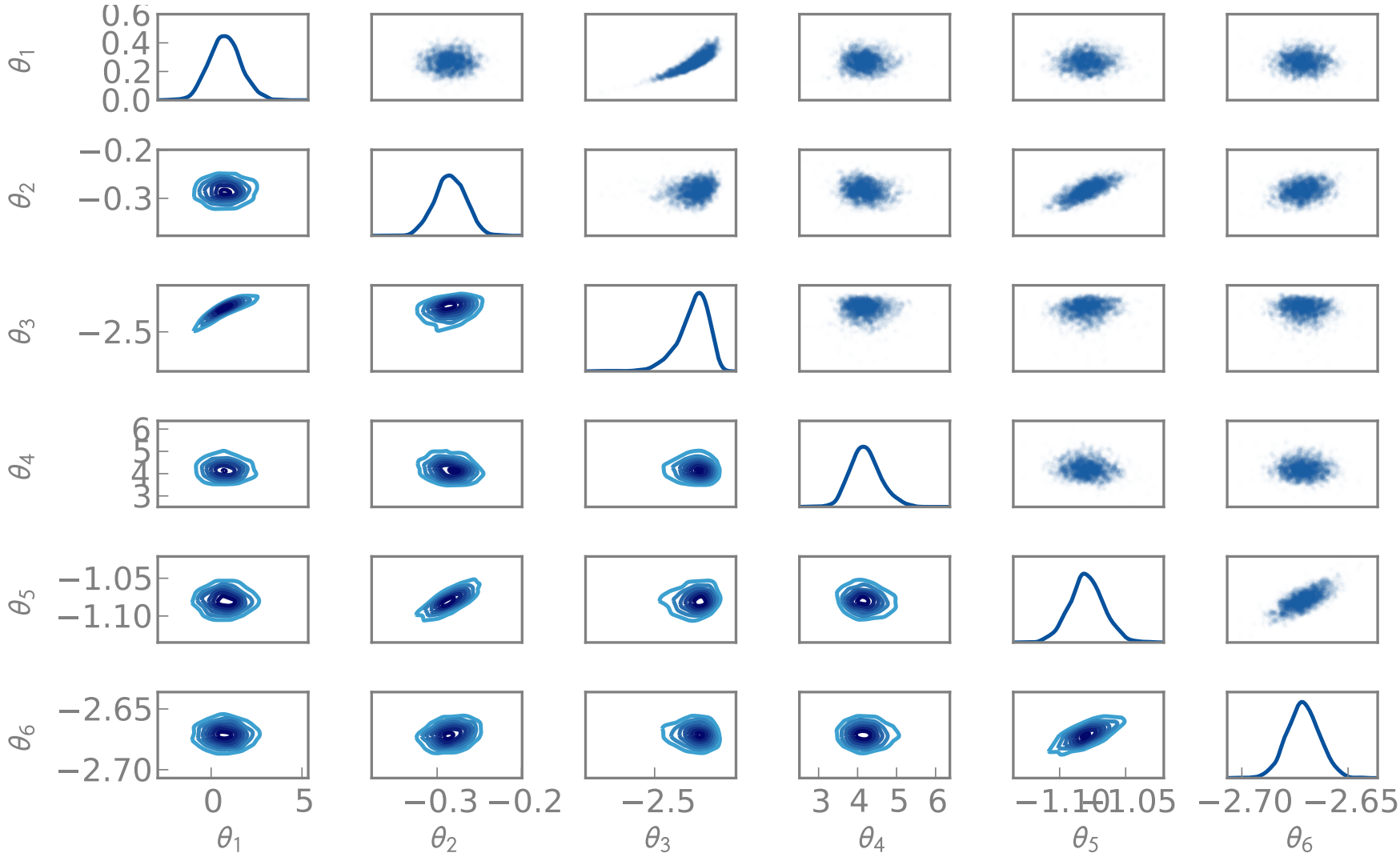
Hydraulic head and velocity



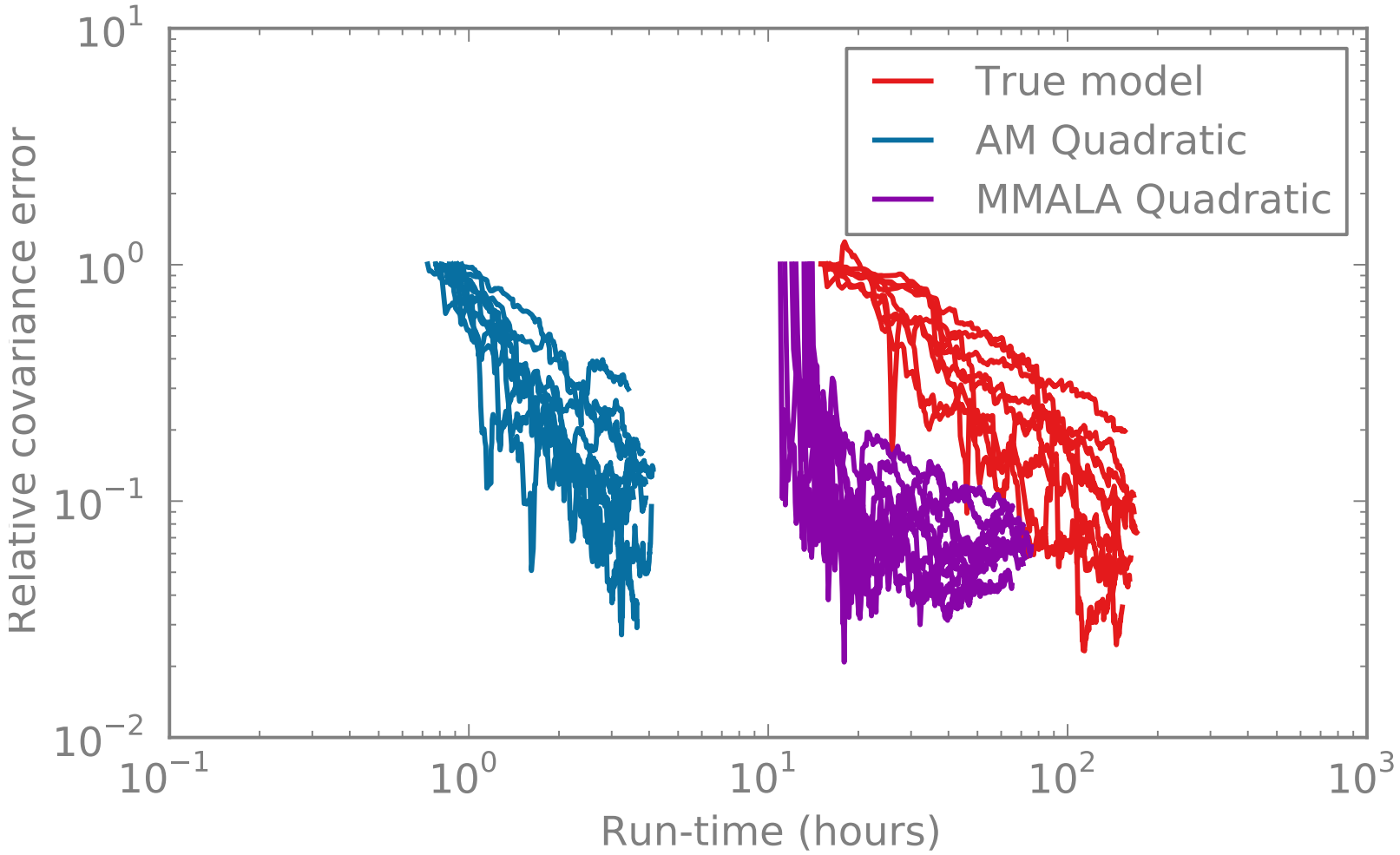
Well locations and tracer concentrations



Groundwater hydrology problem: posterior distribution



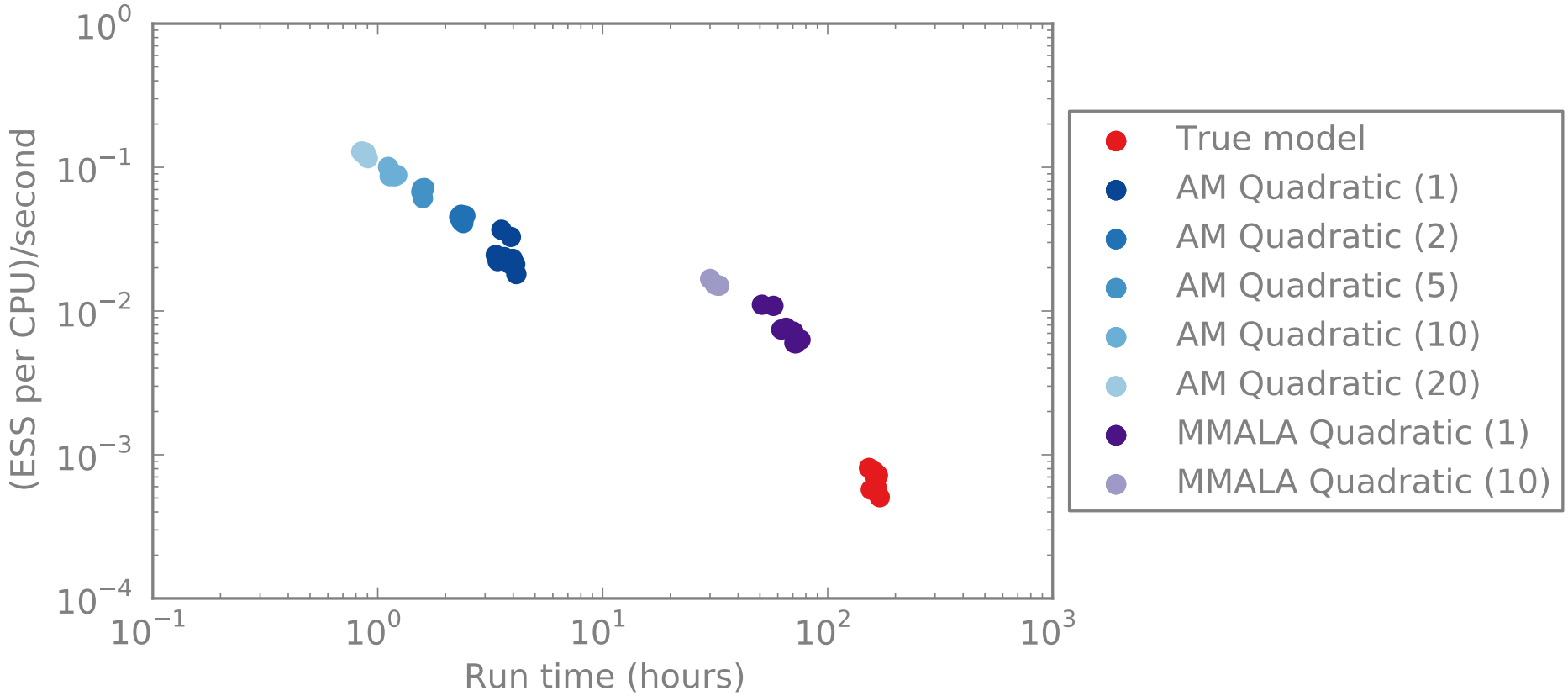
Single chain performance



Implementing approximation across parallel chains

- **Now:** build a common pool of model runs $\mathcal{S}$ across parallel workers
- Run N chains of 100,000 steps each
- Discard 10% of each chain as burn-in; use *effective sample size (ESS)* to measure efficiency
- ESS per CPU-second would be constant with a naïve implementation

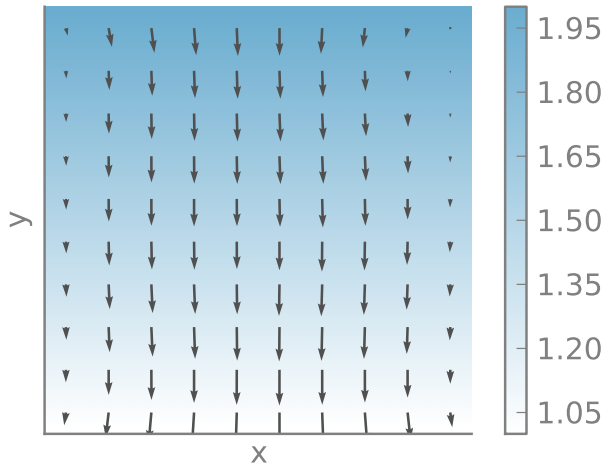
Parallel efficiency



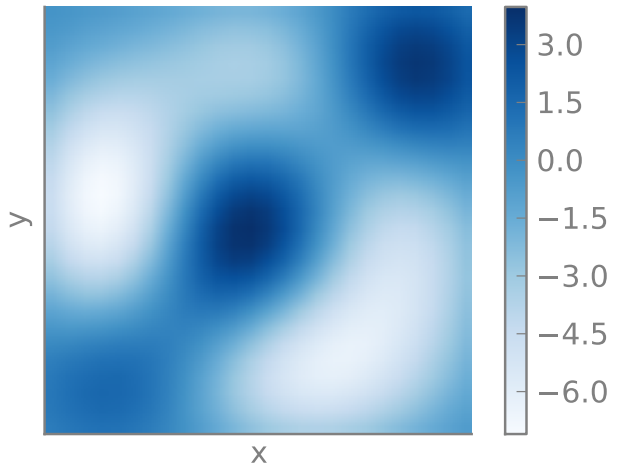
MacAyeal ice stream model

- Model ice as highly viscous non-Newtonian fluid
- Two-dimensional model, thin ice assumption
- Ice flows north to south
- East and west edges have fixed velocity
- Satellites provide high–accuracy velocity observations

Velocity and ice thickness



Log-basal friction



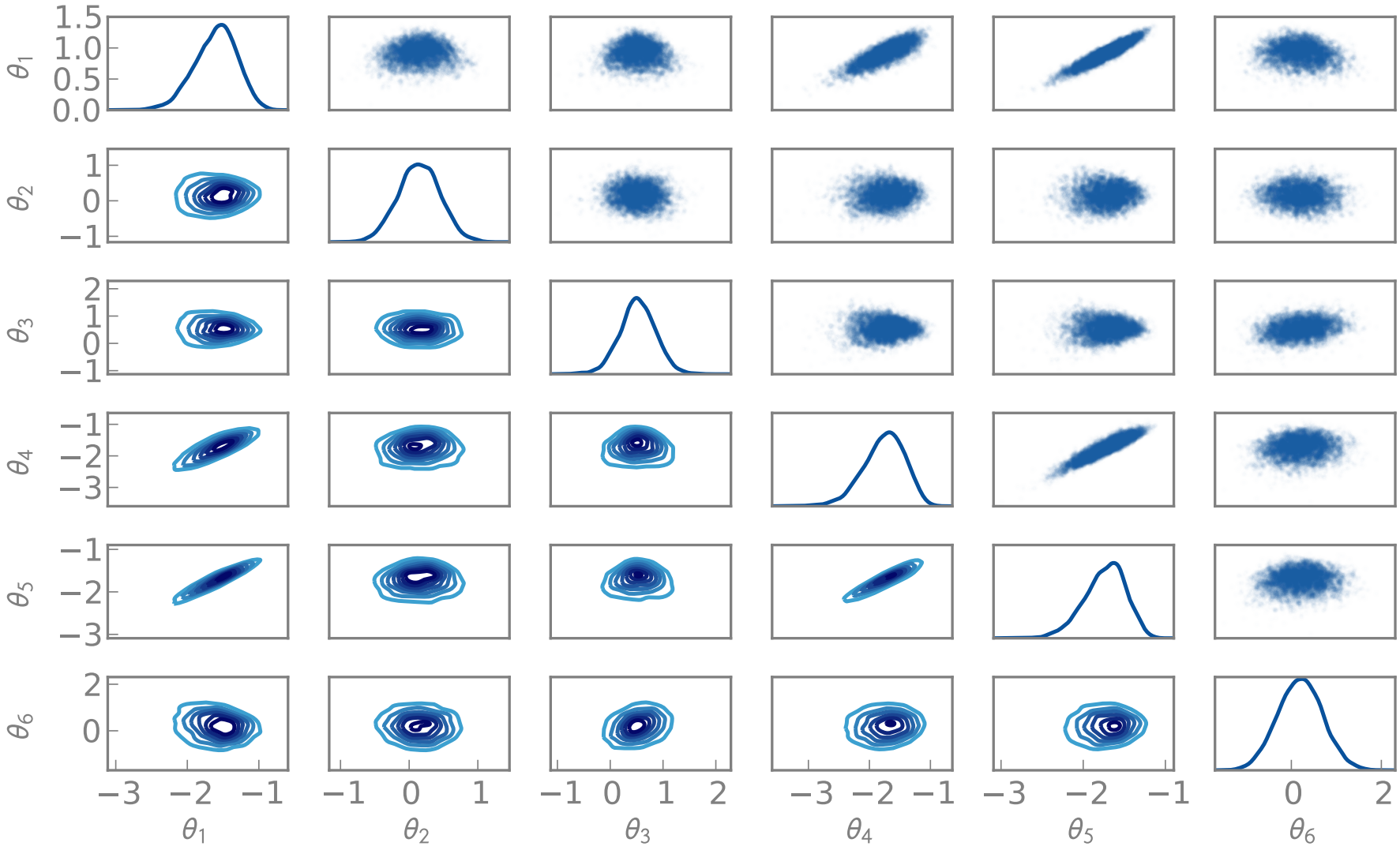
MacAyeal ice stream model

- Coupled system of nonlinear PDEs relates ice stream velocities to basal friction, ice thickness, and surface elevation
- Observe velocity on a 10×10 grid, with small Gaussian error
- Infer for log-friction field as 12 basis coefficients in KL expansion, given log-normal priors

$$\begin{aligned} \log \beta(x, y) &= \sum_{i=1}^{12} \theta_i \sqrt{\lambda_i} \phi_i(x, y) \\ \theta_i &\sim N(0, 1) \end{aligned}$$

MacAyeal ice stream model: posterior distribution

Showing 6 of 12 parameters:



MacAyeal ice stream model: computational performance

- Forward model takes about 25 seconds to evaluate
- Perform MCMC with 10 chains, each of 200 000 steps, discarding first 10% as burn-in
- Standard MCMC would take about **two months**
- Feasible with adaptive Metropolis, quadratic approximation, and parallel chains
- Approximate MCMC runs in about **one day**, roughly 50× improvement
- Used about 35 000 runs of the forward model

Conclusions

- 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
- Introduced a **parallel MCMC** scheme, building local approximations from model evaluations shared across chains
- Much ongoing work:
  - Other experimental design approaches
  - Extensions to higher dimensions
  - Hybrid global/local approximation
  - Regression approximations for pseudomarginal MCMC
  - Other "intractable" likelihoods

References

- Open-source implementation in MUQ, `http://muq.mit.edu`
- 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 (2016). Also arXiv:1402.1694.

Support from US Department of Energy, Office of Advanced Scientific Computing Research (ASCR)