Lecture #4: Posterior approximations for Bayesian inverse problems

Youssef Marzouk

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

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

19–22 March 2018

Agenda

Plan for the lectures:

1. Lectures 1–2: Bayesian inference and MCMC foundations
   - Bayesian modeling
   - MCMC algorithms and demos

2. Lectures 3–4: Bayesian approach to inverse problems
   - Elements of a Bayesian inverse problem formulation
   - Linear–Gaussian problems in detail
   - Surrogate modeling and likelihood approximations
   - Dimension reduction

Approximations in MCMC

Efficient sampling is great, but what if each posterior evaluation is very expensive?

Obvious answer: *approximate* the expensive part, e.g., the forward model.

This raises many interesting issues:

- What kind of approximation scheme to use? What properties of the forward model/likelihood are being exploited?
- When to construct the approximation (offline versus online) and what kind of accuracy to demand from it?
- What is the accuracy of the resulting posterior? Bias in posterior estimates? Can/should we correct for these?

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-dependent) surrogates
- Constructed offline or *online* during posterior sampling

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-dependent) surrogates
- Constructed offline or *online* during posterior sampling

Errors and corrections:

1. Convergence rate of the forward model approximation transfers to the posterior it induces [M & Xiu 2009; Cotter, Dashti, Stuart 2010]
2. Can always correct using a *delayed-acceptance* scheme [Christen & Fox 2005], but at a price
3. Recent work in *asymptotically exact* online approximations

Recap of the Bayesian setting

Posterior density of the parameters

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

Ingredients:

- Parameters $\theta \in \mathbb{R}^d$; data $y \in \mathbb{R}^n$
- Prior density $p(\theta) : \mathbb{R}^d \to \mathbb{R}^+$
- Forward model $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$
  - Sometimes a black-box function
  - Each evaluation is **expensive**
- Likelihood function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$
  - $\mathcal{L}(y, \mathbf{f}(\theta)) = p(y|\theta)$
  - Each evaluation requires, in principle, an evaluation of $\mathbf{f}$

Forward model approximations

- Simple approach: construct an approximation of $\mathbf{f}$ over the *prior* distribution
- Convergence of this approximation (e.g., in L^2_p) yields convergence to the true posterior
- An initial result [M & Xiu 2009]:
  - Let $\mathcal{L}(y, \mathbf{f}(\theta)) = N(y; \mathbf{f}(\theta), I)$ (additive Gaussian noise)
  - Define approximations $(\mathbf{f}^M)$ s.t. $\|\mathbf{f} - \mathbf{f}^M\|_{L^2_p} \leq CM^{-\alpha}$, $\alpha > 0$ (e.g., *polynomial approximations* for smooth and square integrable $\mathbf{f}$)
  - Define corresponding *approximate posteriors* $\pi^M(\theta) \propto \mathcal{L}(y, \mathbf{f}^M(\theta))\, p(\theta)$
  - Then, for sufficiently large M,

$$D_{KL}(\pi_M \| \pi) \lesssim M^{-\alpha}$$

Forward model approximations

- Simple approach: construct an approximation of $\mathbf{f}$ over the *prior* distribution
- Convergence of this approximation (e.g., in L^2_p) yields convergence to the true posterior
- An initial result [M & Xiu 2009]:
  - Let $\mathcal{L}(y, \mathbf{f}(\theta)) = N(y; \mathbf{f}(\theta), I)$ (additive Gaussian noise)
  - Define approximations $(\mathbf{f}^M)$ s.t. $\|\mathbf{f} - \mathbf{f}^M\|_{L^2_p} \leq CM^{-\alpha}$, $\alpha > 0$ (e.g., *polynomial approximations* for smooth and square integrable $\mathbf{f}$)
  - Define corresponding *approximate posteriors* $\pi^M(\theta) \propto \mathcal{L}(y, \mathbf{f}^M(\theta))\, p(\theta)$
  - Then, for sufficiently large M,

$$D_{KL}(\pi_M \| \pi) \lesssim M^{-\alpha}$$

- More general results in [Cotter, Dashti, Stuart 2010]

Forward model approximations

- **Example:** estimate initial perturbation to viscous Burgers' equation

$$\begin{aligned} u_t + uu_x &= \nu u_{xx}, \qquad x \in [-1,1] \\ u(-1) &= 1+\theta, \qquad u(1) = -1. \end{aligned}$$

- Uniform prior on θ, noisy observations of the transition layer location.

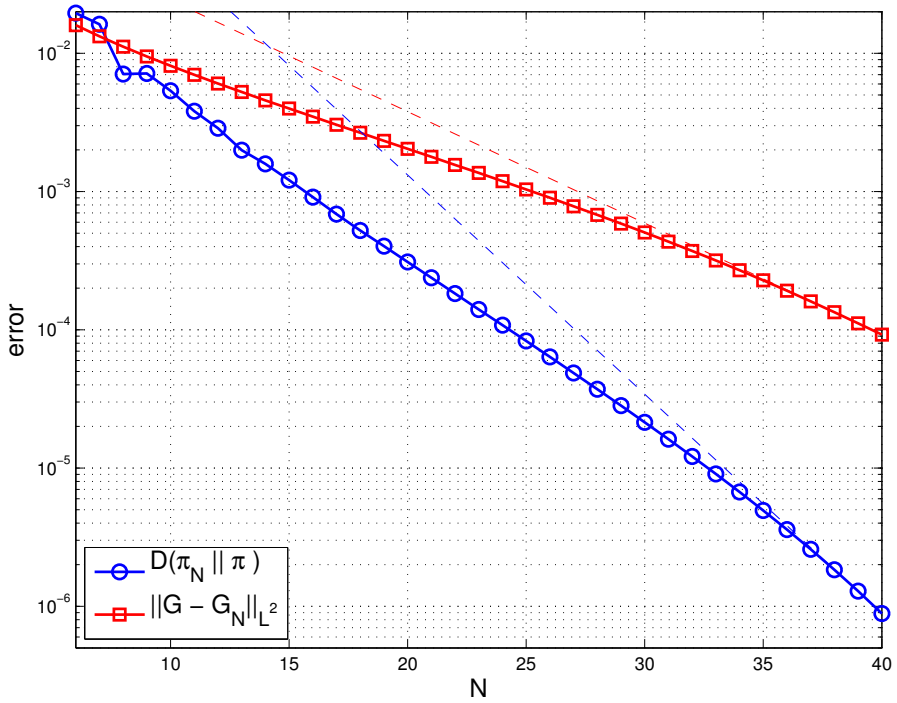
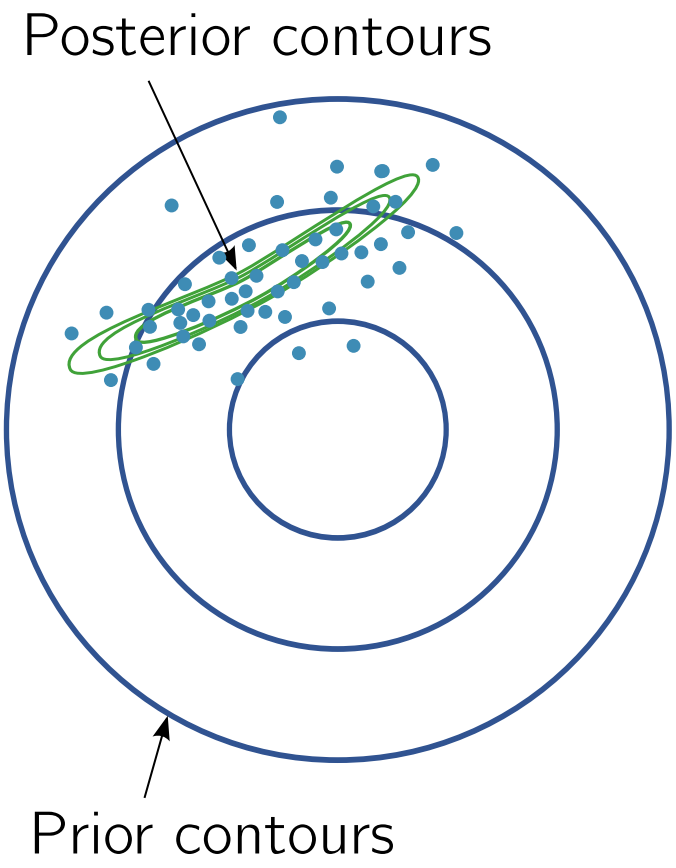


Figure: Convergence of the forward model and the posterior distribution

Forward model and likelihood approximations

- *Posterior-focused* surrogates can improve efficiency
  - Posterior-focused polynomial approximations [Li & M, SISC 2014]
  - Data-driven model reduction [Cui, M, & Willcox IJNME 2014]
  - RBF approximations for $\theta \mapsto \mathcal{L}(y, \mathbf{f}(\theta))$ [Bliznyuk *et al.* 2012, Joseph 2012]
- In general, samples are then drawn from an *approximate* posterior
- Approximation cost borne *a priori;* should balance approx error with sampling error, but difficult to quantify



Delayed-acceptance MCMC

How to use *approximate* likelihoods to accelerate sampling from the **exact posterior?**

- Delayed-acceptance MCMC schemes [Christen & Fox 2005]
- Suppose we have a true target density π and a (cheaper) approximation $\widetilde{\pi}$

Delayed-acceptance MCMC

How to use *approximate* likelihoods to accelerate sampling from the **exact posterior?**

1. Draw a proposal y from $q(y|x_n)$
2. Calculate *first* acceptance ratio
$$\alpha_1(x_n, y) = \min\left\{1, \frac{\widetilde{\pi}(y)q(x_n|y)}{\widetilde{\pi}(x_n)q(y|x_n)}\right\}$$
3. Put
$$z = \begin{cases} y, & \text{with probability } \alpha_1(x_n, y) \\ x_n, & \text{with probability } 1-\alpha_1(x_n, y) \end{cases}$$
and thus define a *second* proposal q^*
4. Calculate *second* acceptance ratio
$$\alpha_2(x_n, z) = \min\left\{1, \frac{\pi(z)q^*(x_n|z)}{\pi(x_n)q^*(z|x_n)}\right\}$$
5. Put
$$x_{n+1} = \begin{cases} z, & \text{with probability } \alpha_2(x_n, z) \\ x_n, & \text{with probability } 1-\alpha_2(x_n, z) \end{cases}$$

Delayed-acceptance MCMC

How to use *approximate* likelihoods to accelerate sampling from the **exact posterior?**

- Delayed-acceptance MCMC schemes [Christen & Fox 2005]
- Suppose we have a true target density π and a (cheaper) approximation $\widetilde{\pi}$

- "Screens" proposals using the cheaper model $\widetilde{\pi}$
- Computed second-stage probability can be close to one for good approximations
- Still calls π at least once per accepted sample

Asymptotically exact MCMC using approximations

A different approach:

- Can we construct an *asymptotically exact* MCMC, via incremental and infinite refinement of approximations?
  - Use **local** approximations of the forward model or log-likelihood
  - Posterior exploration and surrogate construction occur *simultaneously*
  - Asymptotic exactness: convergence of surrogate tied to stationarity of the MCMC chain

(Conrad, M, Pillai, Smith JASA 2016; Conrad, Davis, M, Pillai, Smith JUQ 2018)

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 density $q(x, \cdot)$
2. Compute acceptance ratio
$$\alpha = \min\left(1, \frac{\mathcal{L}(y, \mathbf{f}(q_t))p(q_t)}{\mathcal{L}(y, \mathbf{f}(X_t))p(X_t)}\right)$$
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 density $q(x, \cdot)$
2. Compute acceptance ratio
$$\alpha = \min\left(1, \frac{\mathcal{L}(y, \tilde{\mathbf{f}}_t(q_t))p(q_t)}{\mathcal{L}(y, \tilde{\mathbf{f}}_t(X_t))p(X_t)}\right)$$
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

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

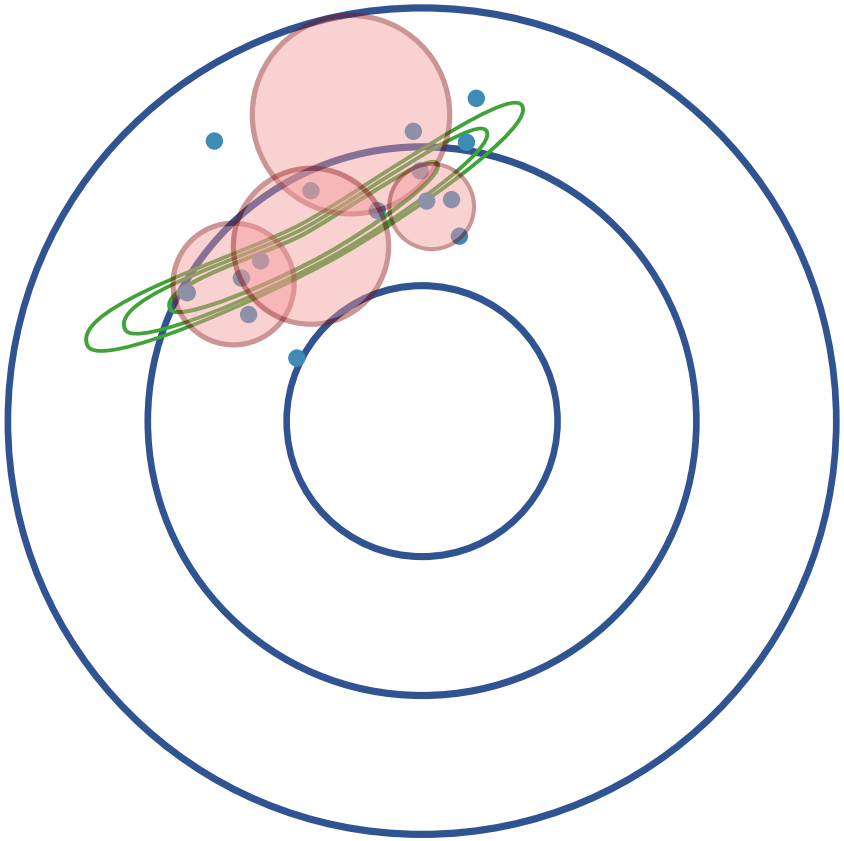
- Choose R so that $M(d)$ samples have non-zero weight, e.g., where $M(d)$ ensures that a quadratic is fully determined

- 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.*]:

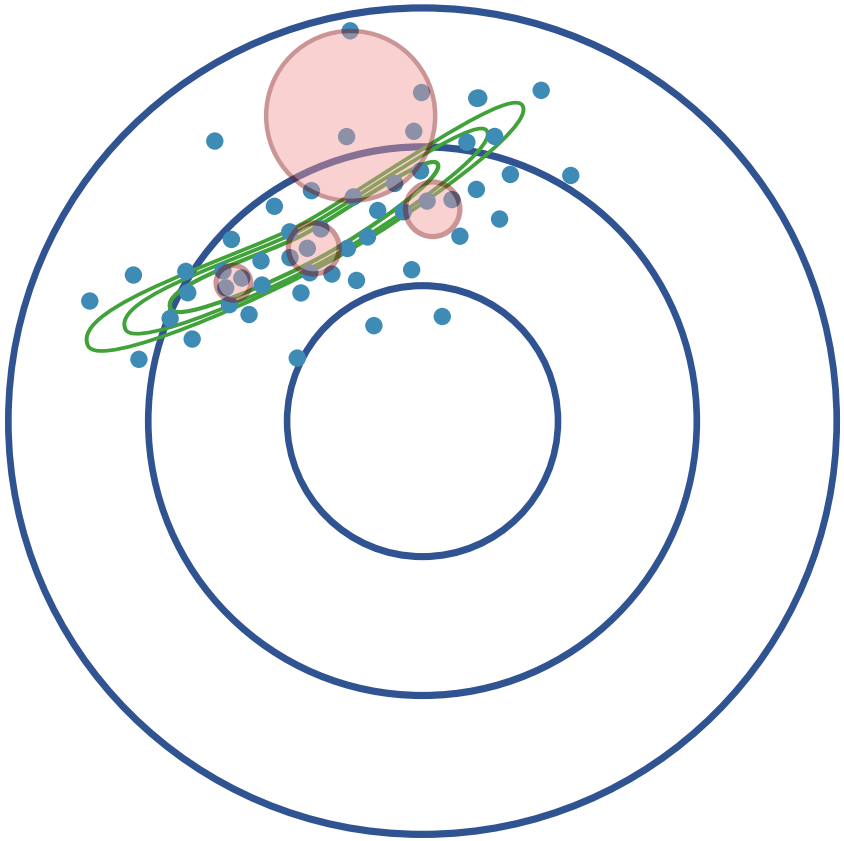
$$\|\mathbf{f} - \mathcal{Q}_R\mathbf{f}\| \leq \kappa(\nu, \lambda, d)R^3$$

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

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

   - Compute error indicators

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

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

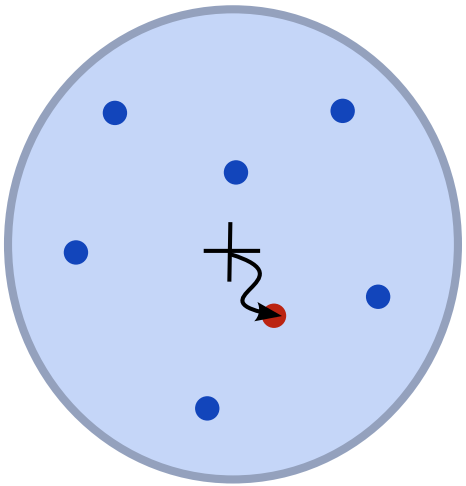
Experimental design: performing refinement

Local space filling refinement

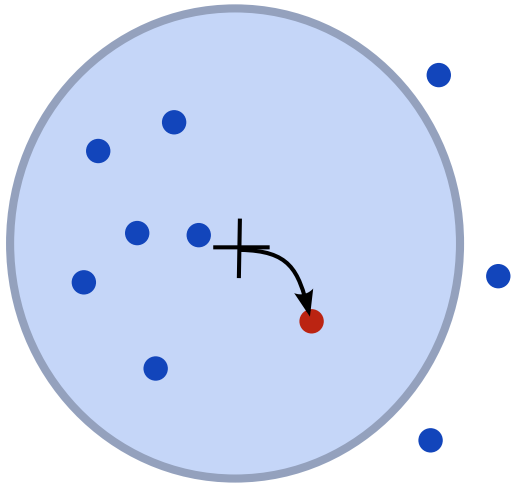
To space fill near $\xi_t = X_t$ or $\xi_t = q_t$, given radius R, locally solve

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

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 2016)

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 $\pi(\theta) := p(\theta|y)$ obeys a *Gaussian envelope* condition.

Then, under standard regularity assumptions for geometrically ergodic kernel K_∞ and posterior π, the chain X_t converges to 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: $\log \mathcal{L}(y, \mathbf{f}(\boldsymbol{\theta}))$
- 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 $\mathcal{S}$

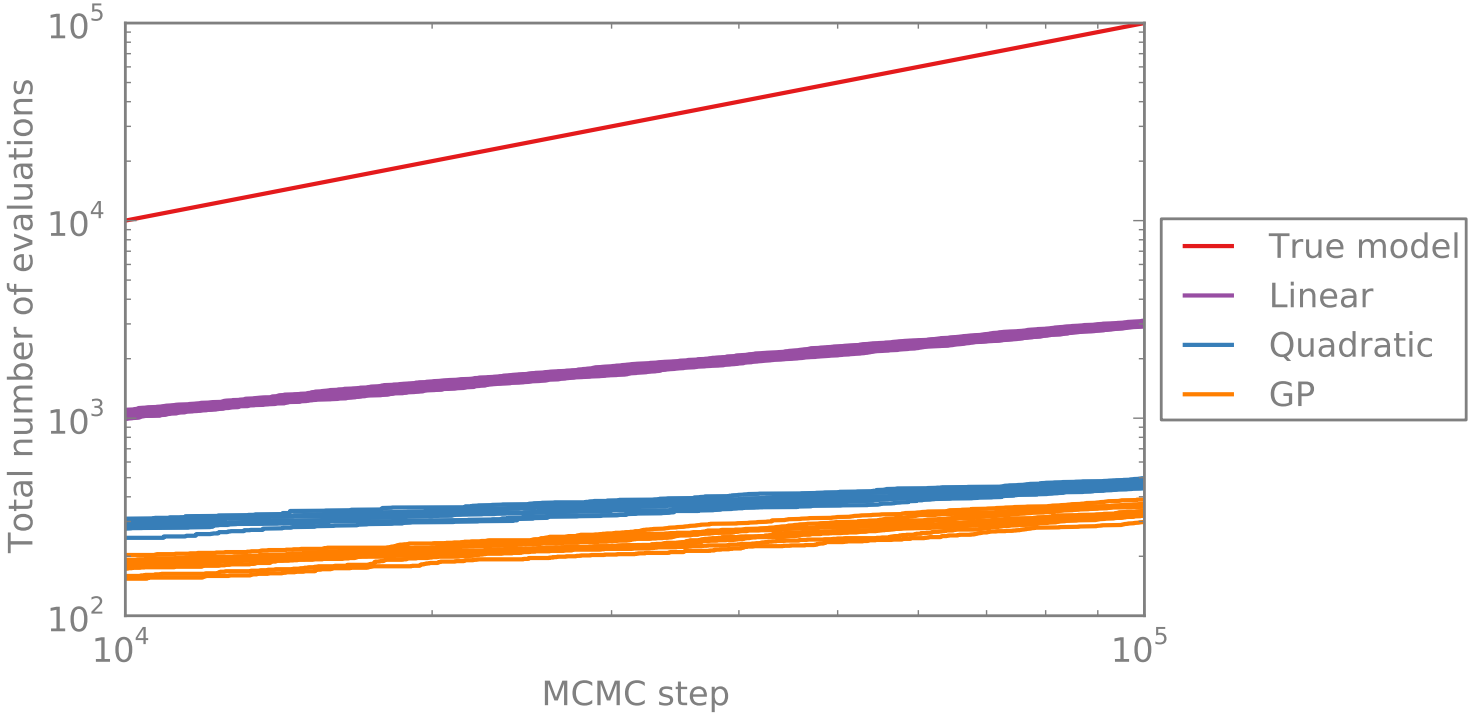
A framework for approximate samplers

Many algorithmic variations:

- Target of approximation
  - Forward model: $\mathbf{f}(\boldsymbol{\theta})$
  - Log-likelihood: $\log \mathcal{L}(y, \mathbf{f}(\boldsymbol{\theta}))$
- 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

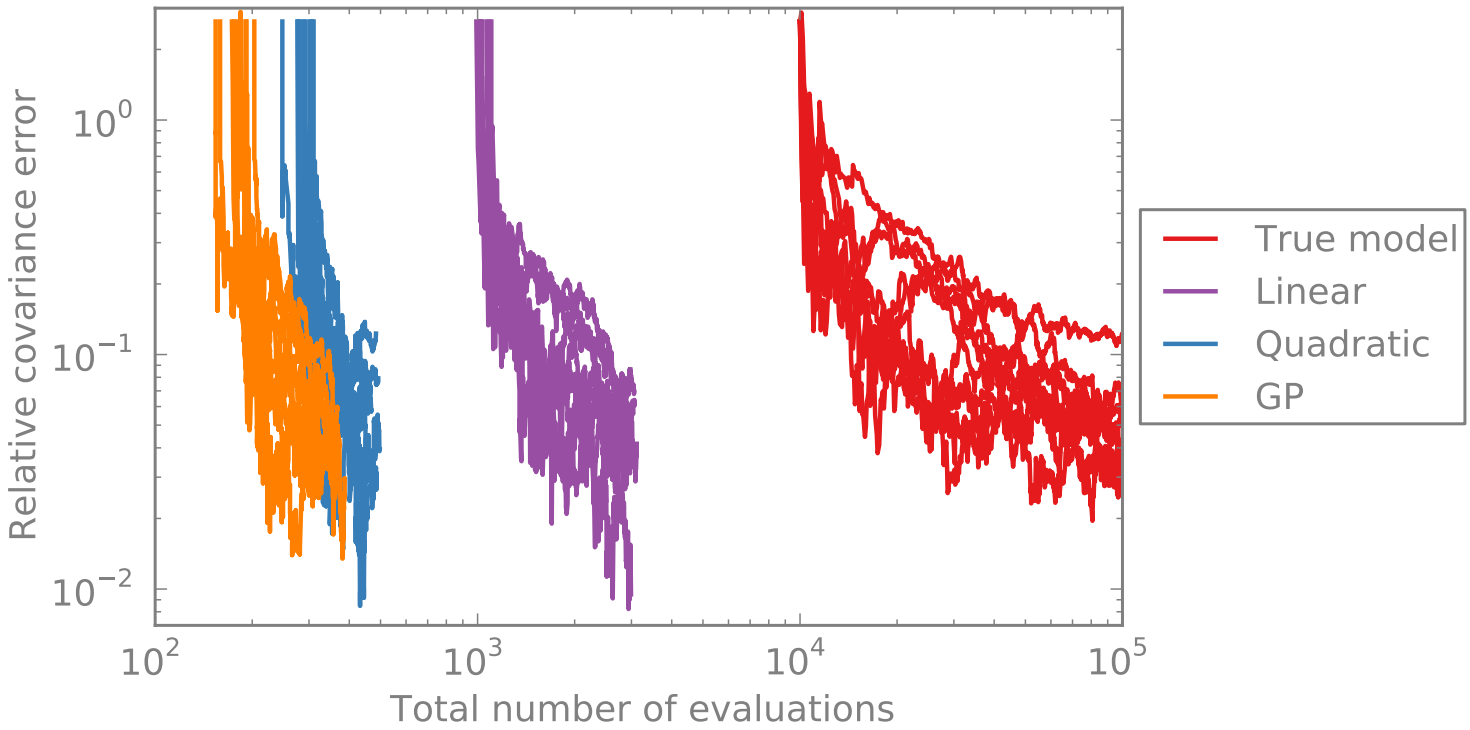
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i.



Cost of chains

Example: elliptic PDE inverse problem

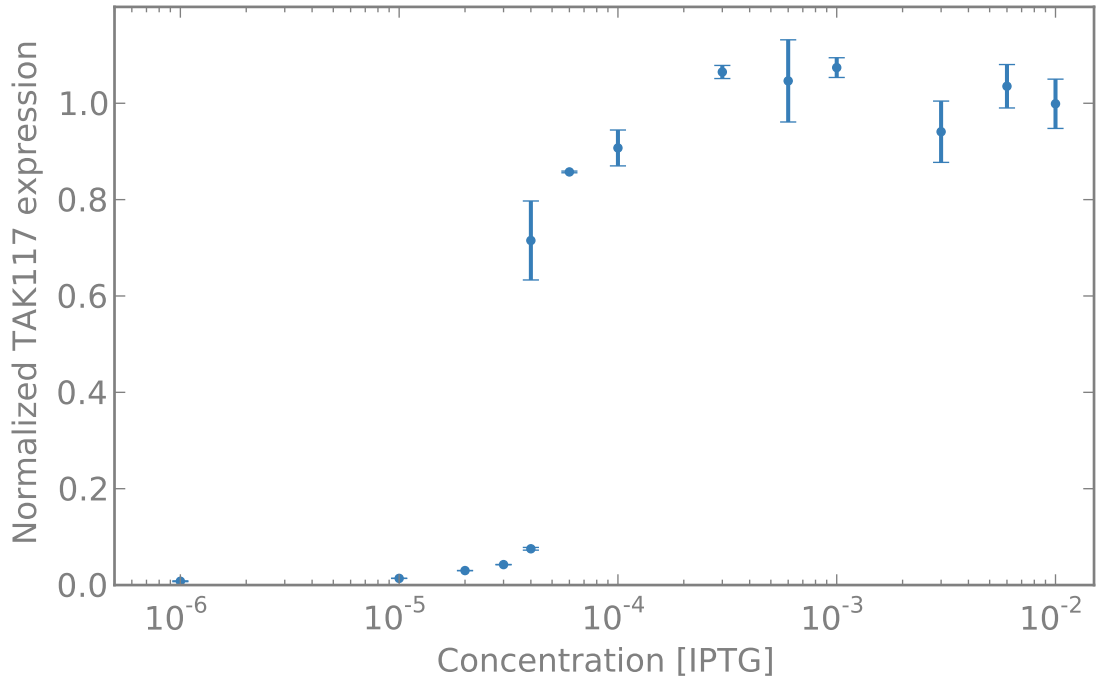
- Elliptic PDE inverse problem: $\nabla \cdot (\kappa(x) \nabla u(x)) = -f$
- Infer permeability field $\kappa(x)$ from limited/noisy observations of pressure u
- Karhunen-Loève expansion: $\log \kappa(x) = \sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} \phi_i(x)$. Standard Gaussian priors on θ_i.



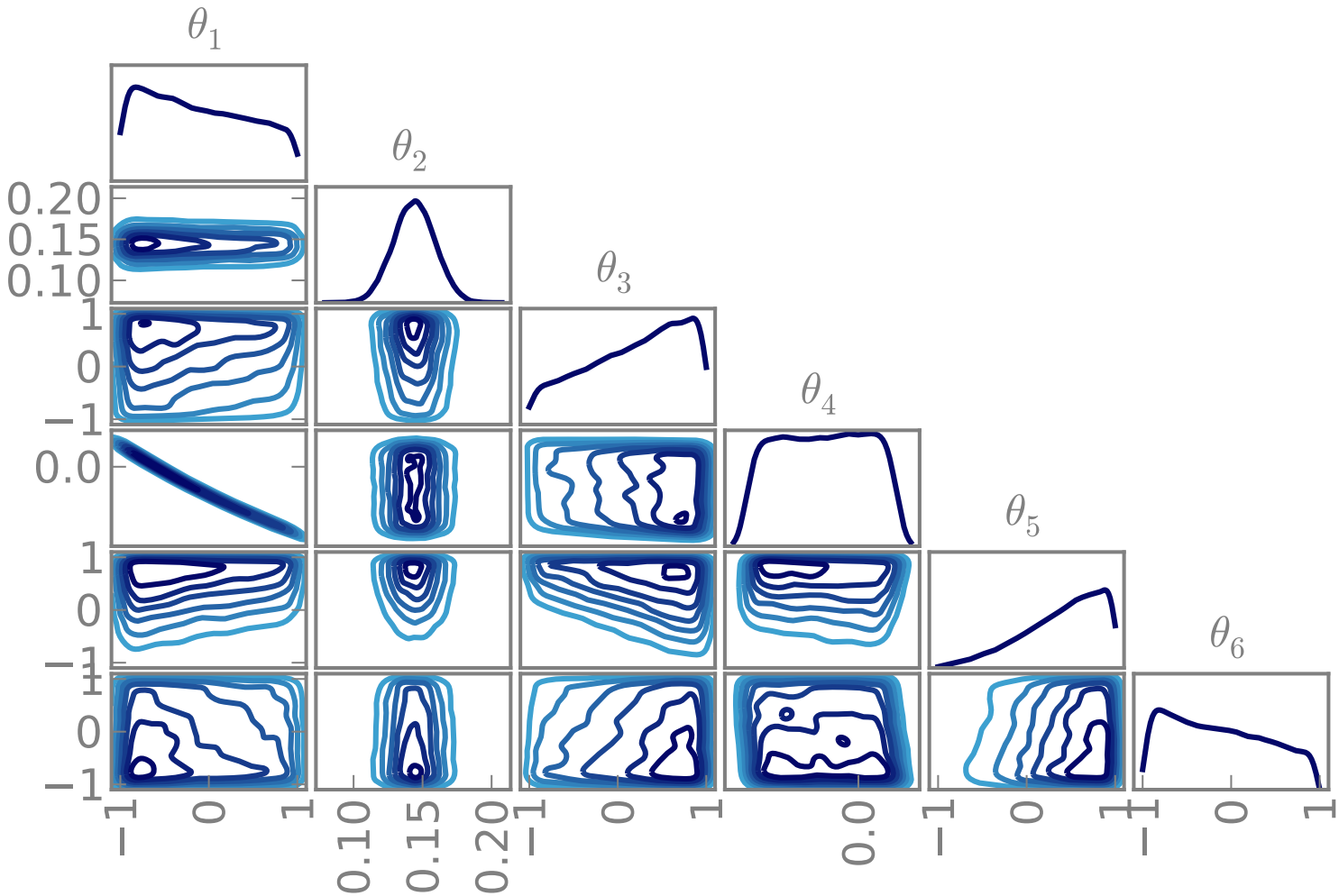
Accuracy versus cost

Example: genetic toggle switch

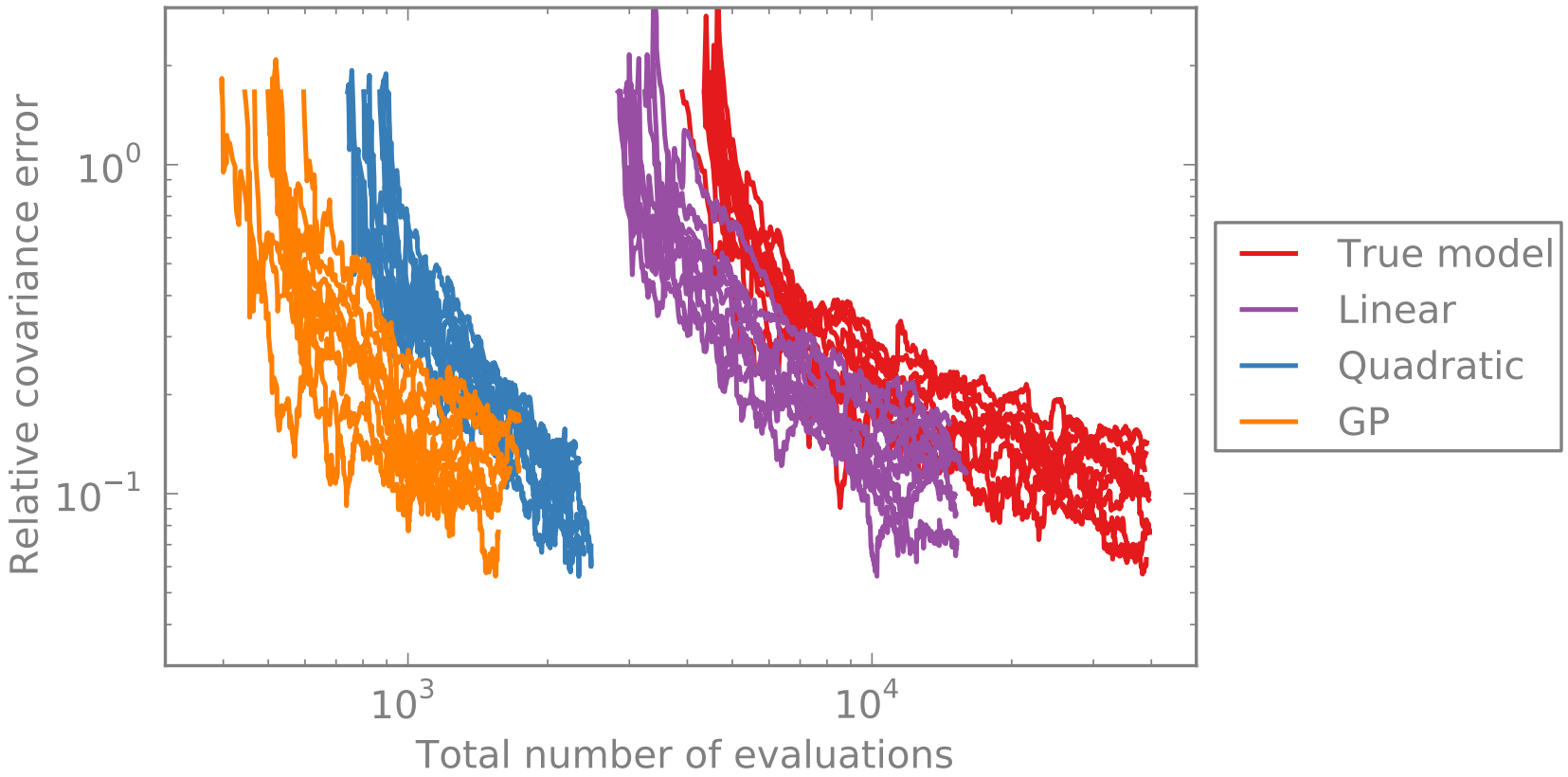
- Model for genetic "toggle switch" synthesized in *E. coli*
- ODE system, six parameters to infer
- Uniform priors, Gaussian observational errors
- Real experimental data



Genetic toggle switch posterior



Genetic toggle switch: accuracy versus cost



A framework for approximate samplers

Many algorithmic variations:

- Target of approximation
  - Forward model: $\mathbf{f}(\boldsymbol{\theta})$
  - Log-likelihood: $\log \mathcal{L}(y, \mathbf{f}(\boldsymbol{\theta}))$
- 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

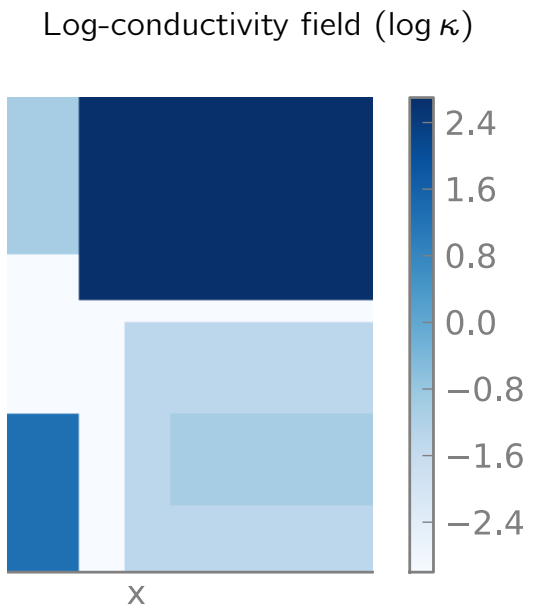
$$\nabla \cdot (h\kappa \nabla h) = -f_h$$

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

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

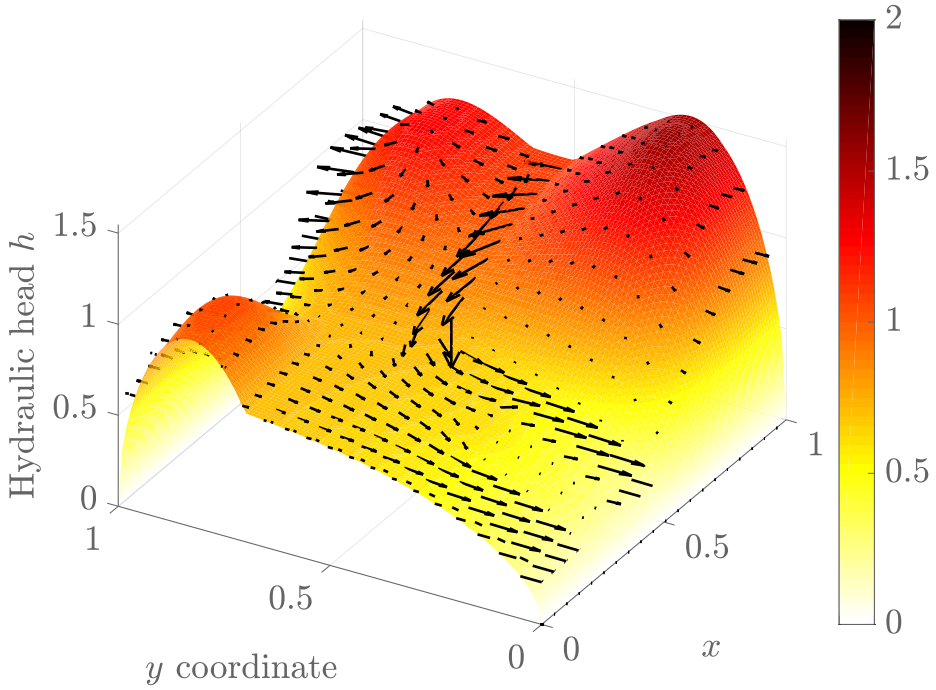
- Tracer advects according to velocity and well forcing

- Observe tracer concentration at well locations, at several times, with Gaussian error
- Infer for piecewise constant conductivities; log-normal priors
- Forward model takes about 13 seconds to evaluate

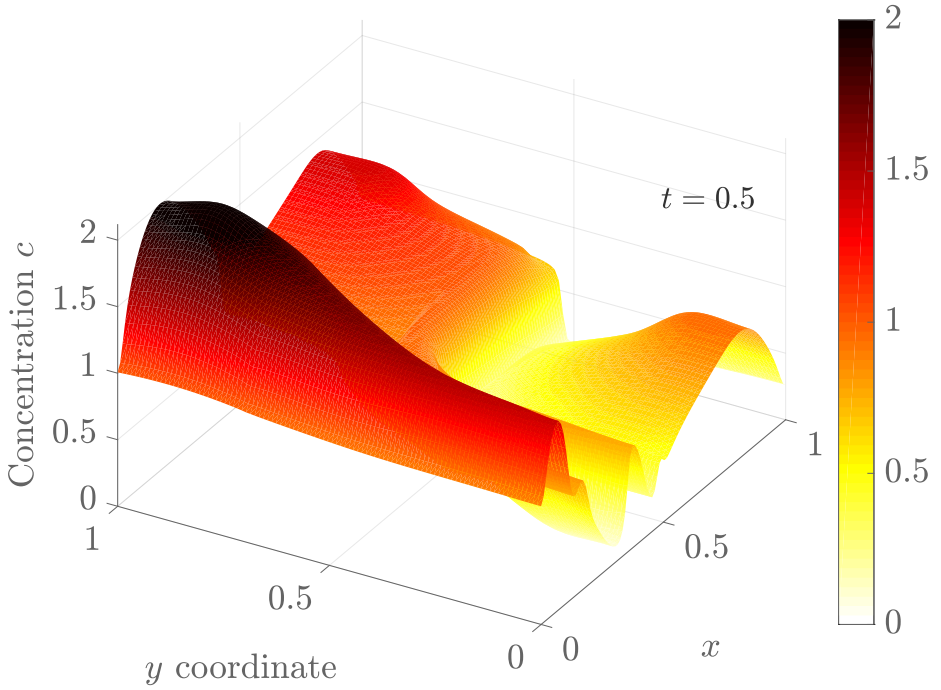


Groundwater tracer transport problem

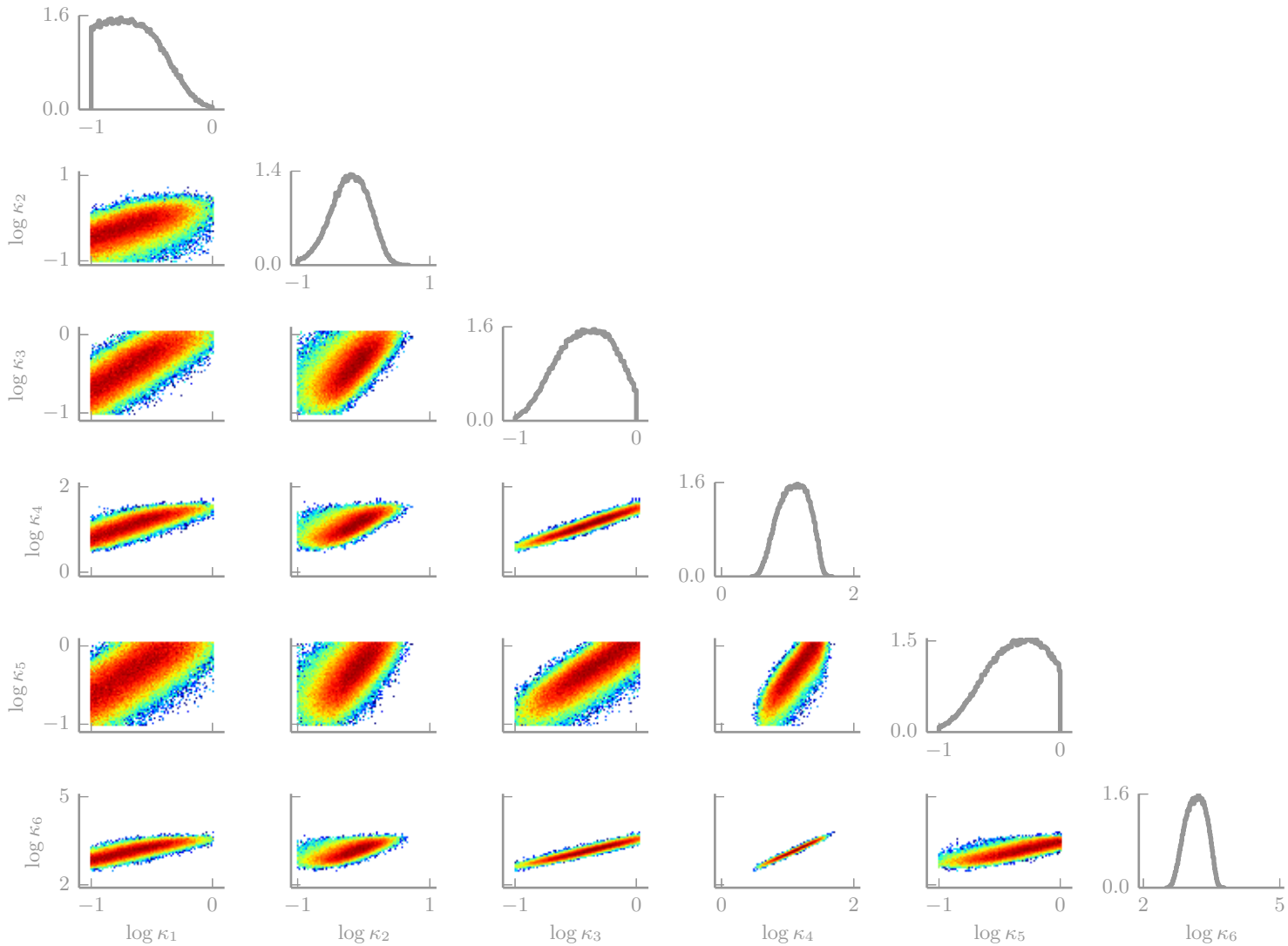
hydraulic head and Darcy velocity



tracer concentrations



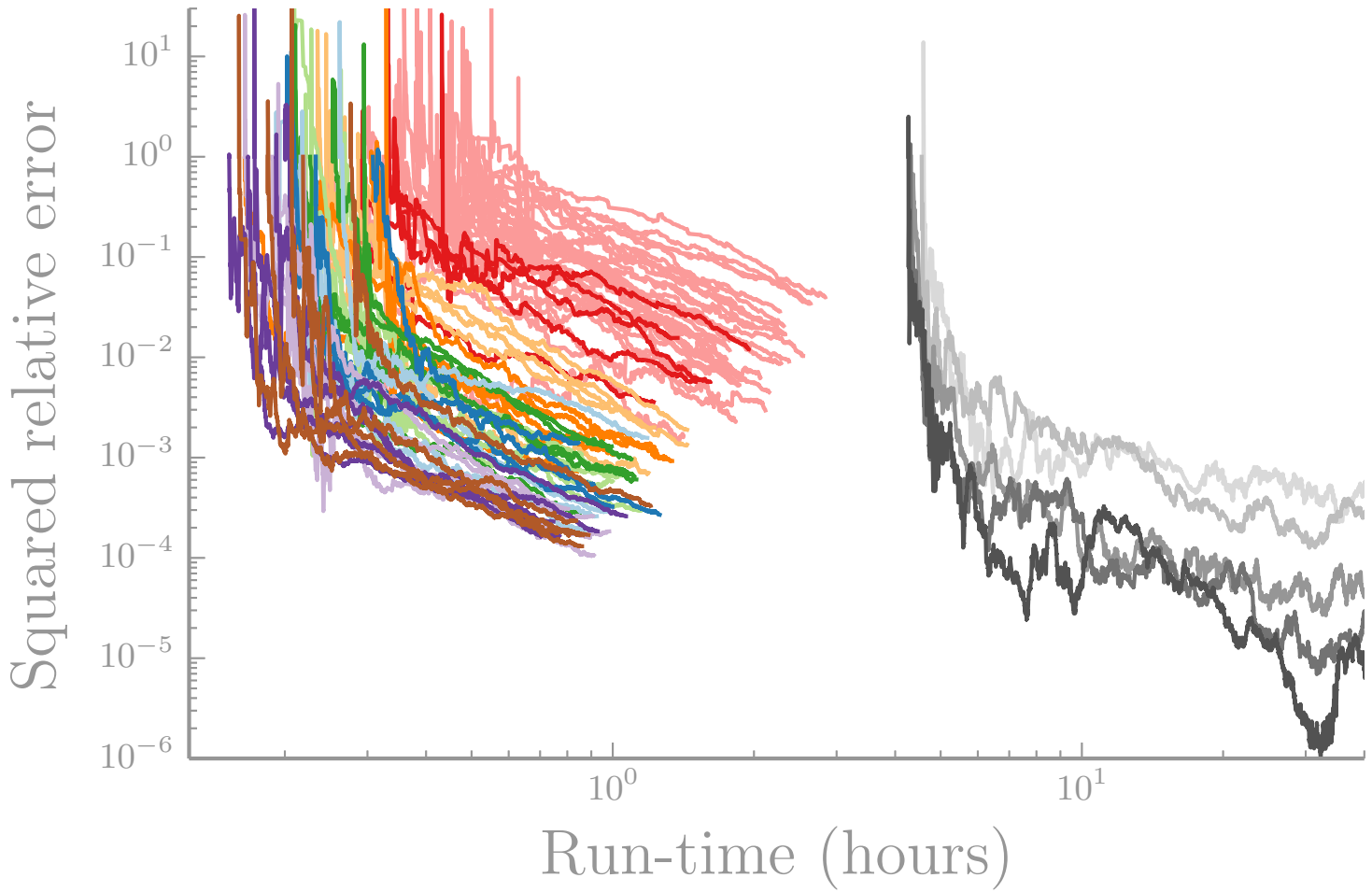
Tracer transport problem: posterior distribution



Parallel local approximation

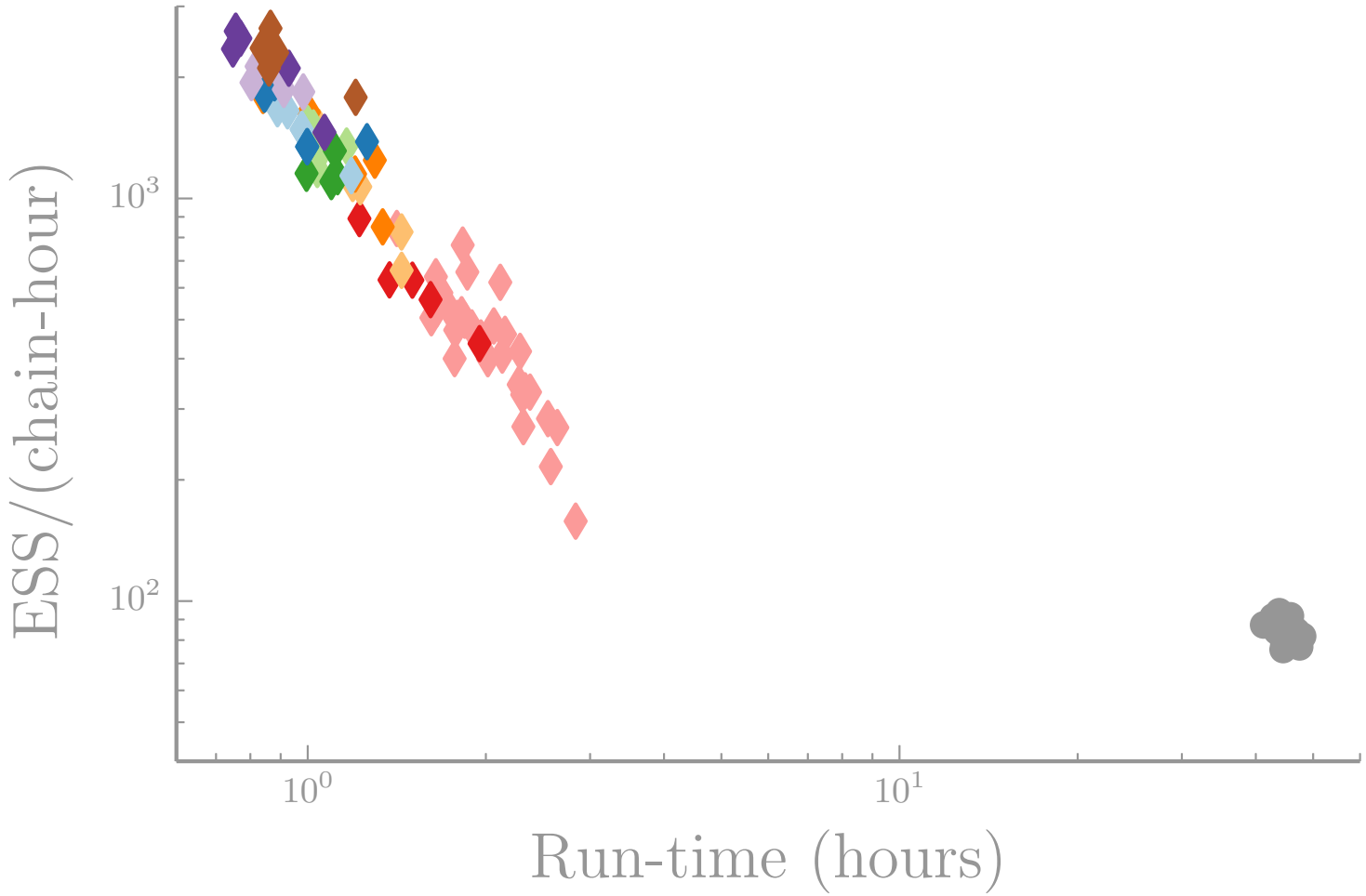
- **Now:** build a common pool of model runs $\mathcal{S}$ across parallel workers
- Run k chains of 10^5 steps each
- Discard 10% of each chain as burn-in; use *effective sample size (ESS)* to measure efficiency
- ESS per chain–hour would be constant with a naïve implementation

Error versus run time



Darker shades = more parallel chains, $k \in \{1, \ldots, 30\}$.

Parallel efficiency

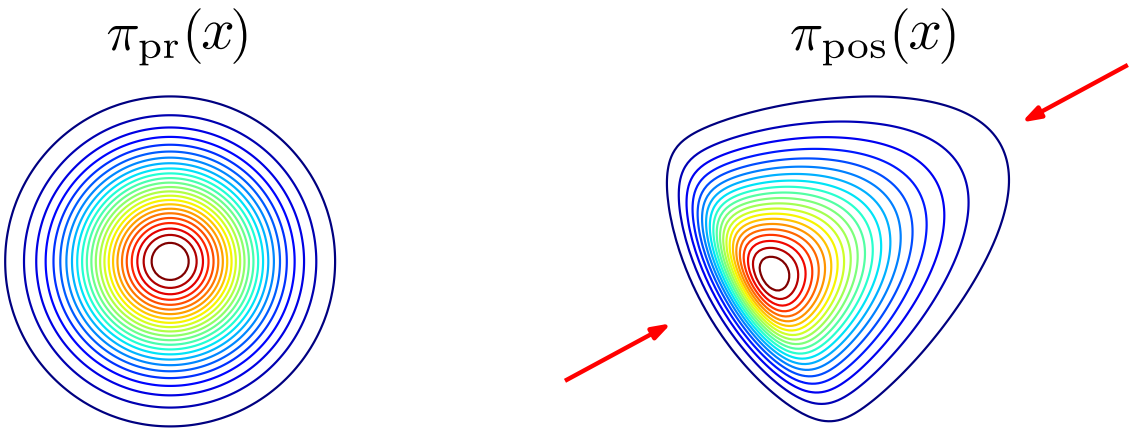


Darker shades = more parallel chains, $k \in \{1, \ldots, 30\}$.

Further directions

- **Refinement rates**: balance *bias* (due to structural error in surrogate) with *variance* (due to finite # of MCMC samples)
  - Both under-refining or over-refining are undesirable
  - Constants are generally unknown; let bias^2 and variance $\to 0$ at the same rate
- Hybrid global + local approximations
- Noisy density evaluations, pseudomarginal MCMC
  - Marginalize out unimportant variables in high-dimensional problems

Conjecture: in many situations, the data are informative only on a low-dimensional subspace



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

Low effective dimensionality of Bayesian inverse problems

Underlying 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 some **linear projector** $P_r \in \mathbb{R}^{d\times d}$ with rank r.

Low effective dimensionality of Bayesian inverse problems

Underlying 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 some **linear projector** $P_r \in \mathbb{R}^{d\times d}$ with rank r.

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

Underlying 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 some **linear projector** $P_r \in \mathbb{R}^{d\times d}$ with rank r.

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 $r \ll d$:

- Build surrogates for the **low-dimensional** function $x_r \mapsto \widetilde{\mathcal{L}}(x_r)$ with a reasonable complexity,
- Design **structure-exploiting** MCMC algorithms to sample from π_{pos} (e.g., [Cui, Law, M 2016; Beskos, Girolami, Lan, Farrell, Stuart 2017])

A zoo of methods for constructing P_r and $\widetilde{\mathcal{L}}$

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

A zoo of methods for constructing P_r and $\widetilde{\mathcal{L}}$

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

$$\mathbf{H} = \Gamma_{\text{pr}}$$

  - **Likelihood informed subspace (LIS)** [Cui et al 2014]

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

  - **Active subspace (AS)** [Constantine, Kent, Bui-Thanh 2015]

$$\mathbf{H}_{\text{AS}}(y) = \int \left(\nabla \log \mathcal{L}_y\right) \left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pr}}$$

A zoo of methods for constructing P_r and $\widetilde{\mathcal{L}}$

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

$$\mathbf{H} = \Gamma_{\text{pr}}$$

  - **Likelihood informed subspace (LIS)** [Cui et al 2014]

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

  - **Active subspace (AS)** [Constantine, Kent, Bui-Thanh 2015]

$$\mathbf{H}_{\text{AS}}(y) = \int \left(\nabla \log \mathcal{L}_y\right)\left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pr}}$$

- Definition of $\widetilde{\mathcal{L}}$:
  - A **common choice** (LIS)

$$\widetilde{\mathcal{L}}(P_r x) = \mathcal{L}_y(P_r x)$$

  - Or via the **conditional expectation** of the log-likelihood (AS)

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

Objective

- Within the approximation class

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

  what is the "best" approximation of π_{pos}?

- In practice, can we build such an approximation?

(Joint work with O. Zahm, T. Cui, K. Law, A. Spantini)

Best approximation problem

For a given rank r, consider the minimization problem

$$\min_{P_r, \widetilde{\mathcal{L}}} \; D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \widetilde{\pi}_{\mathrm{pos}}\right) \qquad \text{with} \quad \widetilde{\pi}_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\mathrm{pr}}(x)$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ denotes the **Kullback-Leibler** divergence.

Best approximation problem

For a given rank r, consider the minimization problem

$$\min_{P_r, \widetilde{\mathcal{L}}} \; D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \widetilde{\pi}_{\mathrm{pos}}\right) \qquad \text{with} \quad \widetilde{\pi}_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\mathrm{pr}}(x)$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ denotes the **Kullback-Leibler** divergence.

Optimal function $\widetilde{\mathcal{L}}$ for a given projector P_r

For any projector $P_r \in \mathbb{R}^{d\times d}$, a minimizer of $\widetilde{\mathcal{L}} \mapsto D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \widetilde{\pi}_{\mathrm{pos}}\right)$ satisfies

$$\widetilde{\mathcal{L}}(P_r x) = \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)$$

Best approximation problem

For a given rank r, consider the minimization problem

$$\min_{P_r, \widetilde{\mathcal{L}}} \; D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \widetilde{\pi}_{\mathrm{pos}}\right) \qquad \text{with} \quad \widetilde{\pi}_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}(P_r x)\pi_{\mathrm{pr}}(x)$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ denotes the **Kullback-Leibler** divergence.

Optimal function $\widetilde{\mathcal{L}}$ for a given projector P_r

For any projector $P_r \in \mathbb{R}^{d\times d}$, a minimizer of $\widetilde{\mathcal{L}} \mapsto D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \widetilde{\pi}_{\mathrm{pos}}\right)$ satisfies

$$\widetilde{\mathcal{L}}(P_r x) = \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)$$

The best approximation problem becomes

$$\min_{P_r} \; D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \middle|\middle| \pi^*_{\mathrm{pos}}\right) \qquad \text{where} \quad \pi^*_{\mathrm{pos}}(x) \propto \mathbb{E}_{\pi_{\mathrm{pr}}}\left(\mathcal{L}_y \middle| P_r x\right)\pi_{\mathrm{pr}}(x)$$

But solving this problem appears intractable in general...

An upper bound

Assumption on the prior: there exists a SPD matrix $\Sigma \in \mathbb{R}^{d \times d}$ such that

$$-\nabla^2 \log \pi_{\text{pr}}(x) \succeq \Sigma \qquad \forall x \in \mathbb{R}^d$$

- In other words, we assume that the prior is **strongly log-concave**
- **Any Gaussian** $\pi_{\text{pr}} = \mathcal{N}(\mu_{\text{pr}}, \Gamma_{\text{pr}})$ satisfies this assumption with $\Sigma = \Gamma_{\text{pr}}^{-1}$

An upper bound

Assumption on the prior: there exists a SPD matrix $\Sigma \in \mathbb{R}^{d\times d}$ such that

$$-\nabla^2 \log \pi_{\text{pr}}(x) \succeq \Sigma \qquad \forall x \in \mathbb{R}^d$$

- In other words, we assume that the prior is **strongly log-concave**
- **Any Gaussian** $\pi_{\text{pr}} = \mathcal{N}(\mu_{\text{pr}}, \Gamma_{\text{pr}})$ satisfies this assumption with $\Sigma = \Gamma_{\text{pr}}^{-1}$

Upper bound for the KL-divergence

For any projector P_r we have

$$D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi^*_{\text{pos}}\right) \leq \frac{1}{2}\text{trace}\Big(\Sigma^{-1}\big(I_d - P_r\big)^T \mathbf{H}(y)\big(I_d - P_r\big)\Big)$$

where $\pi^*_{\text{pos}}(x) \sim \mathbb{E}_{\pi_{\text{pr}}}\big(\mathcal{L}_y \big| P_r x\big)\pi_{\text{pr}}(x)$ and

$$\mathbf{H}(y) = \int \big(\nabla \log \mathcal{L}_y\big)\big(\nabla \log \mathcal{L}_y\big)^T \, \mathrm{d}\pi_{\text{pos}}$$

- The proof relies on logarithmic Sobolev inequalities [Ledoux 1997]
- This upper bound is quadratic w.r.t. P_r

Minimizer of the upper bound

- Let (λ_i, v_i) be the i-th eigenpair of the **generalized eigenvalue problem**:

$$\mathbf{H}(y) v_i = \lambda_i \Sigma v_i$$

- A minimizer of the upper bound is given by

$$P_r^* = \left(\sum_{i=1}^{r} v_i v_i^T \right) \Sigma$$

- With the choice $\pi_{\text{pos}}^*(x) \propto \mathbb{E}(\mathcal{L}_y | P_r^* x) \pi_{\text{pr}}(x)$ we have

$$D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi_{\text{pos}}^*\right) \leq \frac{1}{2} \sum_{i=r+1}^{d} \lambda_i$$

Minimizer of the upper bound

- Let (λ_i, v_i) be the i-th eigenpair of the **generalized eigenvalue problem**:

$$\mathbf{H}(y)v_i = \lambda_i \Sigma v_i$$

- A minimizer of the upper bound is given by

$$P_r^* = \left(\sum_{i=1}^{r} v_i v_i^T \right) \Sigma$$

- With the choice $\pi_{\text{pos}}^*(x) \propto \mathbb{E}(\mathcal{L}_y | P_r^* x) \pi_{\text{pr}}(x)$ we have

$$D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi_{\text{pos}}^*\right) \leq \frac{1}{2} \sum_{i=r+1}^{d} \lambda_i$$

This may not be a solution to the best approximation problem!

However:

- we can choose the rank $r = r(\varepsilon)$ such that $D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi_{\text{pos}}^*\right) \leq \varepsilon$
- a strong decay in λ_i ensures $r(\varepsilon) \ll d$

Approach

1. Compute
$$\mathbf{H}(y) = \int \left(\nabla \log \mathcal{L}_y\right)\left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pos}}.$$
2. Solve the generalized eigenvalue problem
$$\mathbf{H}(y) v_i = \lambda_i \Sigma v_i,$$
and assemble P_r^*.
3. Compute the conditional expectation
$$\widetilde{\mathcal{L}}(P_r^* x) = \mathbb{E}(\mathcal{L}_y | P_r^* x).$$

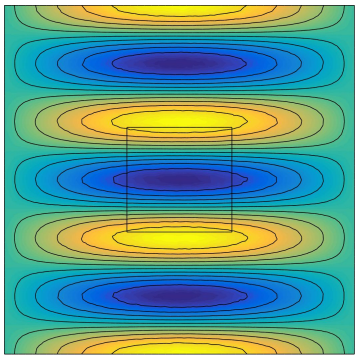
Then $\pi_{\text{pos}}^*(x) \propto \widetilde{\mathcal{L}}(P_r^* x)\pi_{\text{pr}}(x)$ satisfies

$$\boxed{D_{\text{KL}}\left(\pi_{\text{pos}} \middle|\middle| \pi_{\text{pos}}^*\right) \leq \frac{1}{2} \sum_{i=r+1}^{d} \lambda_i}$$

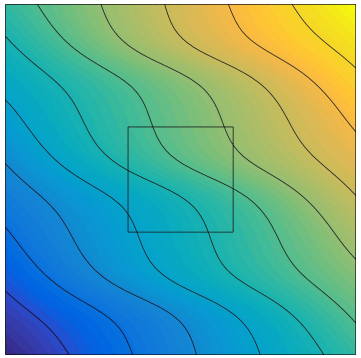
Numerical illustration: heat equation on $\Omega = [0,1]^2$

Parameter	Model	Observation
$x = \log \kappa \sim \mathcal{N}(0, \Gamma_{\text{pr}})$	$-\nabla(\kappa \nabla u) = 0 \quad \text{in } \Omega$	$y = u_{\Omega_{\text{obs}}} + \mathcal{N}(0, \Gamma_{\text{obs}})$
Γ_{pr}: exponential kernel	$u = x_1 + x_2 \quad \text{on } \partial\Omega$	where $\Omega_{\text{obs}} = [.35, .65]^2$

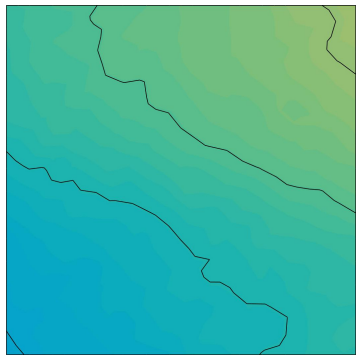
$\log \kappa_{\text{true}}$



u_{true}



y_{obs}



After discretization, the dimension of the problem is $d = 2730$

Numerical illustration: heat equation on $\Omega = [0,1]^2$

Parameter

$x = \log \kappa \sim \mathcal{N}(0, \Gamma_{\text{pr}})$

Γ_{pr}: exponential kernel

Model

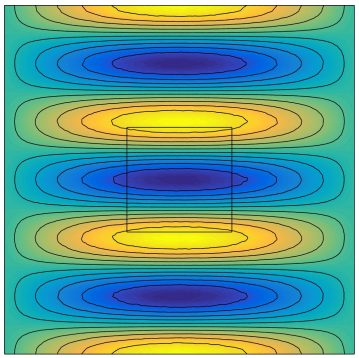
$$-\nabla\big(\kappa \nabla u\big) = 0 \quad \text{in } \Omega$$
$$u = x_1 + x_2 \quad \text{on } \partial\Omega$$

Observation

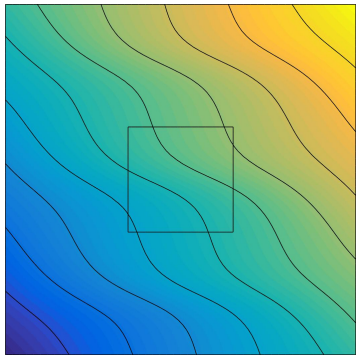
$y = u_{\Omega_{\text{obs}}} + \mathcal{N}(0, \Gamma_{\text{obs}})$

where $\Omega_{\text{obs}} = [.35, .65]^2$

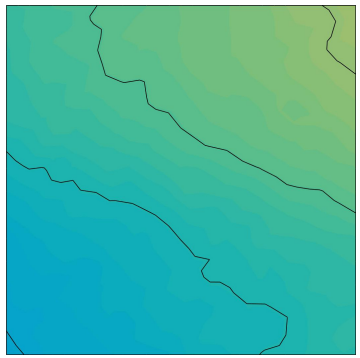
$\log \kappa_{\text{true}}$



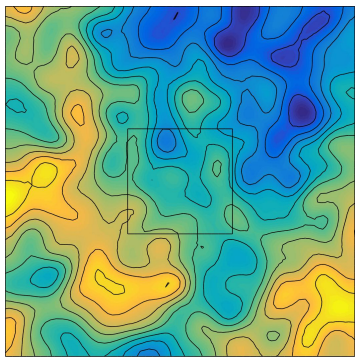
u_{true}



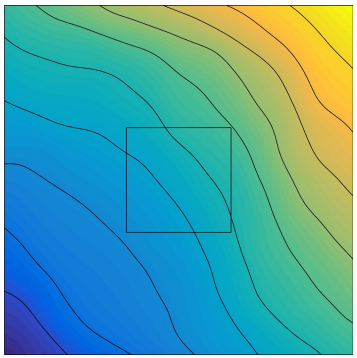
y_{obs}



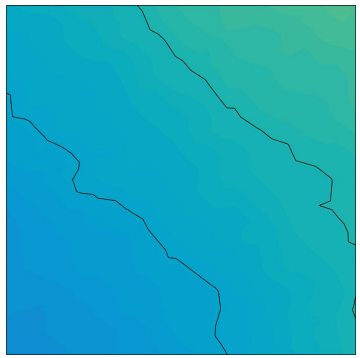
$\log \kappa$



u



y



Numerical illustration: heat equation on $\Omega = [0,1]^2$

Parameter

$$x = \log \kappa \sim \mathcal{N}(0, \Gamma_{\text{pr}})$$

Γ_{pr}: exponential kernel

Model

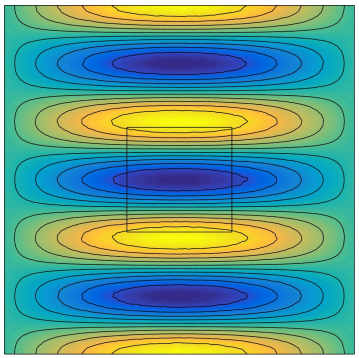
$$-\nabla\big(\kappa \nabla u\big) = 0 \quad \text{in } \Omega$$
$$u = x_1 + x_2 \quad \text{on } \partial\Omega$$

Observation

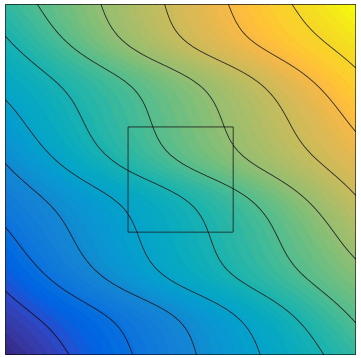
$$y = u_{\Omega_{\text{obs}}} + \mathcal{N}(0, \Gamma_{\text{obs}})$$

where $\Omega_{\text{obs}} = [.35, .65]^2$

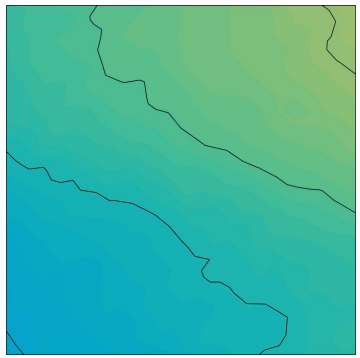
$\log \kappa_{\text{true}}$



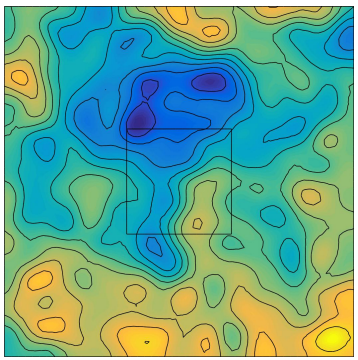
u_{true}



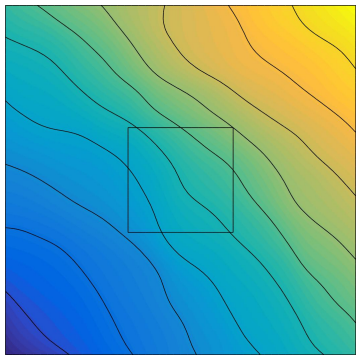
y_{obs}



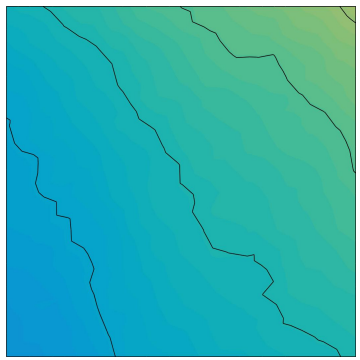
$\log \kappa$



u



y



Numerical illustration: heat equation on $\Omega = [0,1]^2$

Parameter

$$x = \log \kappa \sim \mathcal{N}(0, \Gamma_{\text{pr}})$$

Γ_{pr}: exponential kernel

Model

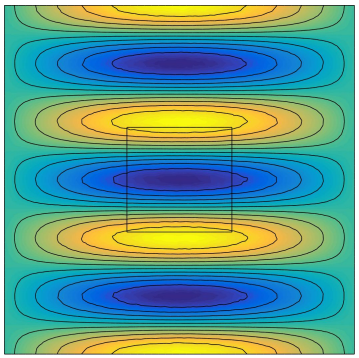
$$-\nabla\big(\kappa \nabla u\big) = 0 \quad \text{in } \Omega$$
$$u = x_1 + x_2 \quad \text{on } \partial\Omega$$

Observation

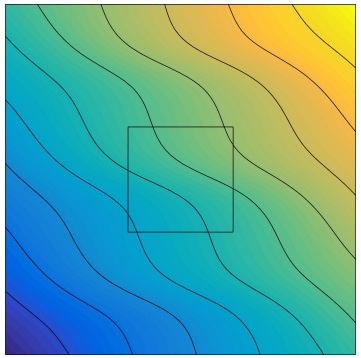
$$y = u_{\Omega_{\text{obs}}} + \mathcal{N}(0, \Gamma_{\text{obs}})$$

where $\Omega_{\text{obs}} = [.35, .65]^2$

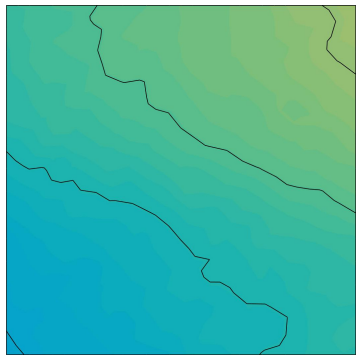
$\log \kappa_{\text{true}}$



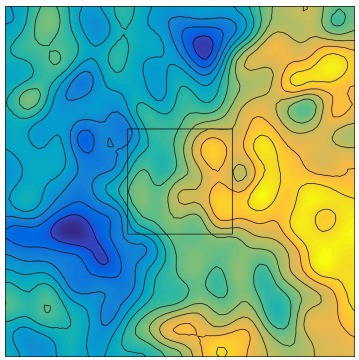
u_{true}



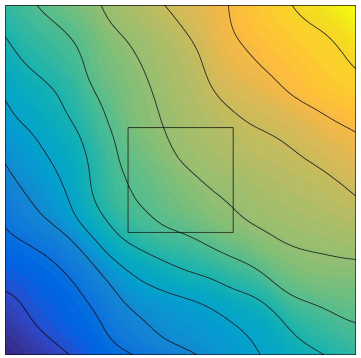
y_{obs}



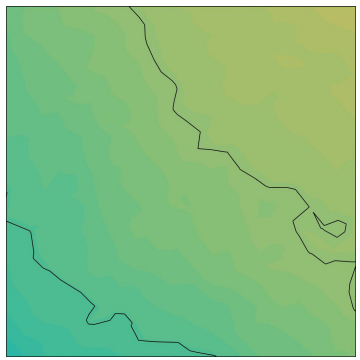
$\log \kappa$



u



y



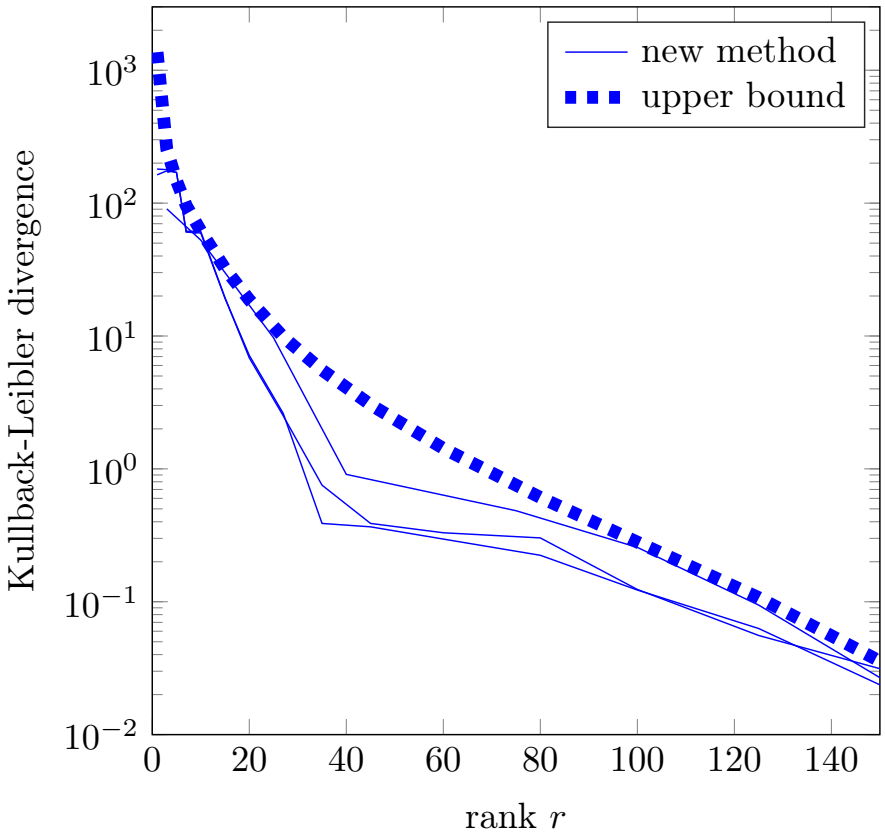
Approximation of the conditional expectation

Assume we can exactly compute

$$\mathbf{H}(y) = \int \left(\nabla \log \mathcal{L}_y\right)\left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pos}}$$

Instead of computing the **expensive** conditional expectation, we use

$$\widetilde{\mathcal{L}}(P_r^* x) = \mathcal{L}_y(P_r^* x + \xi_\perp) \qquad \text{with} \quad \xi \sim \pi_{\text{pr}}(x)$$



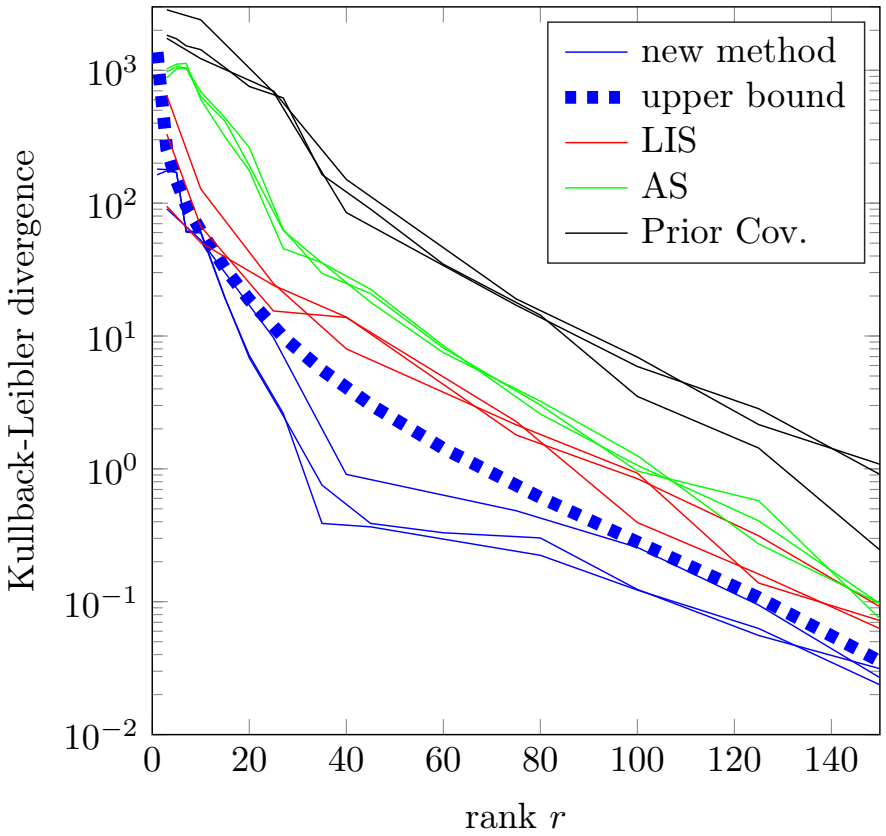
Approximation of the conditional expectation

Assume we can exactly compute

$$\mathbf{H}(y) = \int \left(\nabla \log \mathcal{L}_y\right)\left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pos}}$$

Instead of computing the **expensive** conditional expectation, we use

$$\widetilde{\mathcal{L}}(P_r^* x) = \mathcal{L}_y(P_r^* x + \xi_\perp) \qquad \text{with} \quad \xi \sim \pi_{\text{pr}}(x)$$



Comparison with other methods:

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

$$\mathbf{H}_{\text{AS}}(y) = \int \left(\nabla \log \mathcal{L}_y\right)\left(\nabla \log \mathcal{L}_y\right)^T \, \mathrm{d}\pi_{\text{pr}}$$

$$\mathbf{H} = \Gamma_{\text{pr}}$$

Monte Carlo approximation for H

$$\mathbf{H} = \int f(x)\ \rho(x)\mathrm{d}x \quad \overset{\text{Monte Carlo}}{\approx} \quad \mathbf{H}^{(K)} = \frac{1}{K}\sum_{k=1}^{K} f(x^{(k)}) \quad \text{with} \quad x^{(k)} \overset{\text{iid}}{\sim} \rho(x)$$

$f(x)$	$\rho = \pi_{\text{pr}}$	$\rho = \pi_{\text{pos}}$
$(\nabla \log \mathcal{L})(\nabla \log \mathcal{L})^T$	AS	New method
$(\nabla G)^T \Gamma_{\text{obs}}^{-1} (\nabla G)$	LIS-PR	LIS

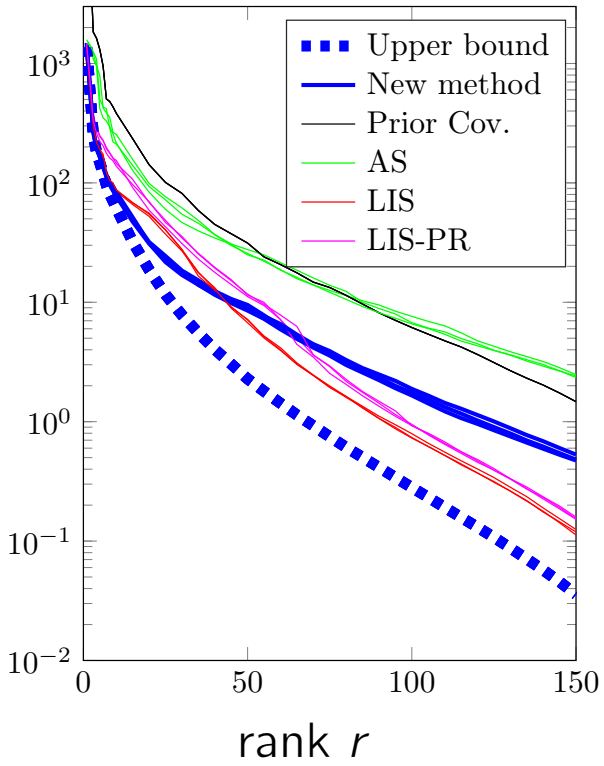
- **Draw the samples** $x^{(1)}, x^{(2)}, \ldots$
  - $\rho = \pi_{\text{pr}}$: readily available
  - $\rho = \pi_{\text{pos}}$: MCMC/importance sampling
- **"Information" per sample**
  - AS/new method : $\text{rank}\big(f(x^{(k)})\big) = 1$
  - LIS-PR/LIS : $\text{rank}\big(f(x^{(k)})\big) \geq 1$

Approximation of the projector

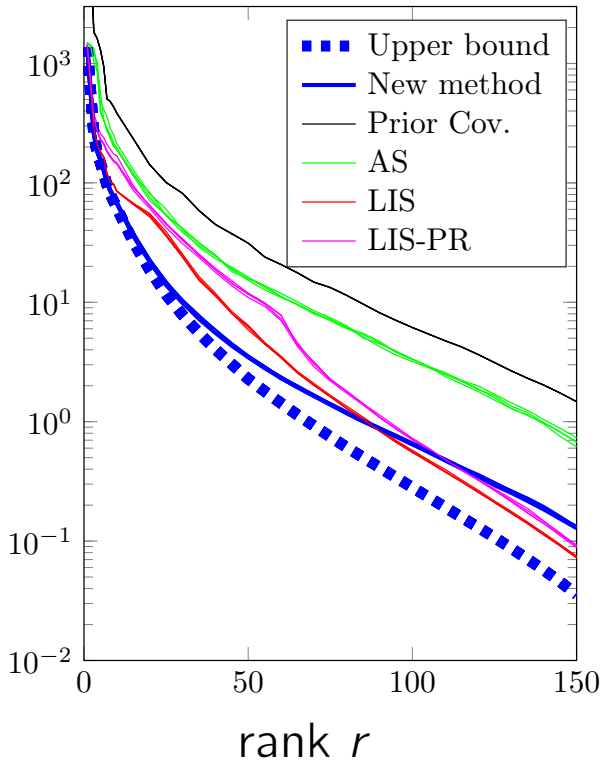
Compute P_r from $\mathbf{H}^{(K)}$ and plot the upper bound

$$\frac{1}{2}\text{trace}\Big(\Sigma^{-1}(I_d - P_r)\mathbf{H}(y)(I_d - P_r)\Big) = \text{function}(r).$$

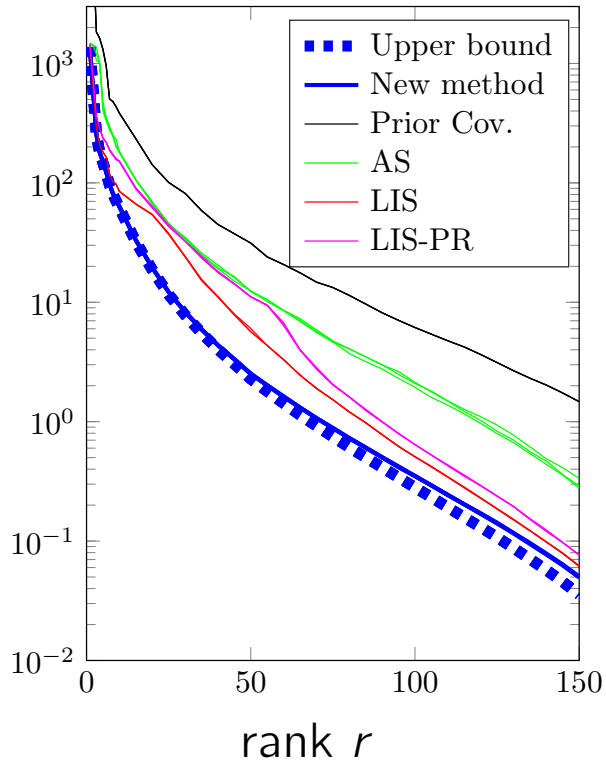
$K = 50$



$K = 200$



$K = 1000$



Implementations and links

Dimension reduction in practice:

- Since posterior samples are required to identify $\mathbf{H}$, one typically uses iterative schemes: MCMC [Cui et al. 2014; Cui, Law, M 2016], or importance sampling [Cui, Willcox, M 2016]
- Other issues: sample size bounds for error due to $\mathbf{H}^{(K)}$ and approximation of conditional expectation

Implementations and links

Linking dimension reduction to model reduction:

- Parameter dimension reduction makes forward model/likelihood approximation **easier**
- Should only care about response of the forward model along parameter dimensions informed by the data
- Model reduction can exploit "locality" in two senses: parameter dimension reduction *and* posterior concentration relative to the prior [Cui, Willcox, M 2016]

References

- Open-source implementations in MUQ, `http://muq.mit.edu`
- Cui, Martin, Marzouk, Solonen, Spantini. "Likelihood-informed dimension reduction for nonlinear inverse probems." *Inv. Prob.* 30: 114015 (2014).
- Spantini, Solonen, Cui, Martin, Tenorio, Marzouk. "Optimal low-rank approximation of linear Bayesian inverse problems." *SIAM J. Sci. Comp.* 37: A2451–A2487 (2015).
- Cui, Law, Marzouk. "Dimension-independent likelihood-informed MCMC." *J. Comp. Phys.* 304: 109–137 (2016).
- Cui, Marzouk, Willcox "Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction." *J. Comp. Phys.* 315: 363–387 (2016)
- Conrad, Marzouk, Pillai, Smith, "Accelerating asymptotically exact MCMC for computationally intensive models via local approximations." *J. Amer. Statist. Assoc.*, 111: 1591–1607 (2016).
- Conrad, Davis, Marzouk, Pillai, Smith "Parallel local approximation MCMC for expensive models." *SIAM JUQ*, to appear (2018). arXiv: 1607.02788
- Zahm, Cui, Law, Spantini, Marzouk. "Certified dimension reduction for nonlinear Bayesian inverse problems." Forthcoming (2018).