Bayesian inference in inverse problems: forward approximation and dimension reduction

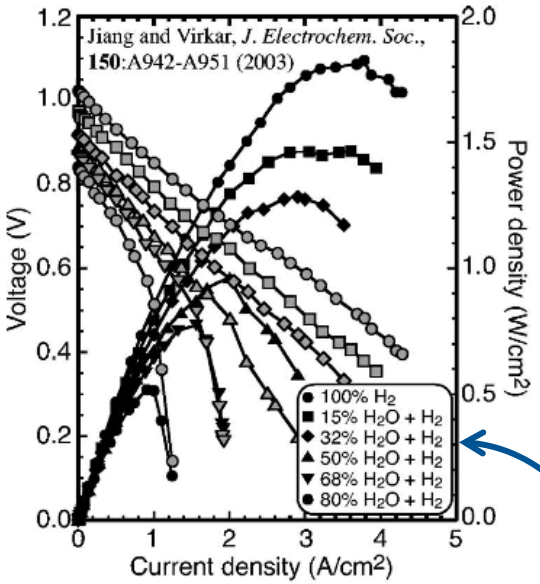
Youssef Marzouk[1]

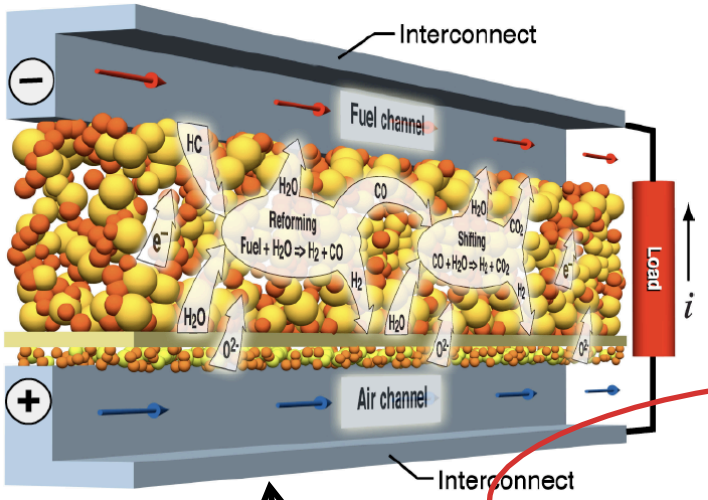
joint work with Patrick Conrad, Tiangang Cui, Jinglai Li, James Martin, Habib Najm

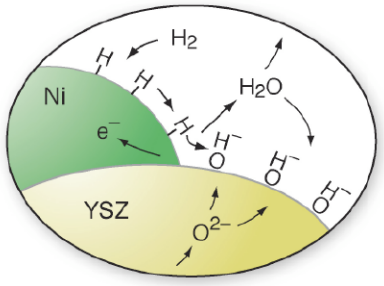
[1]Massachusetts Institute of Technology
Department of Aeronautics & Astronautics
Center for Computational Engineering

UQ for inverse problems

- ***Example:*** *electrochemical energy conversion*







$$d = G(x) + \epsilon$$

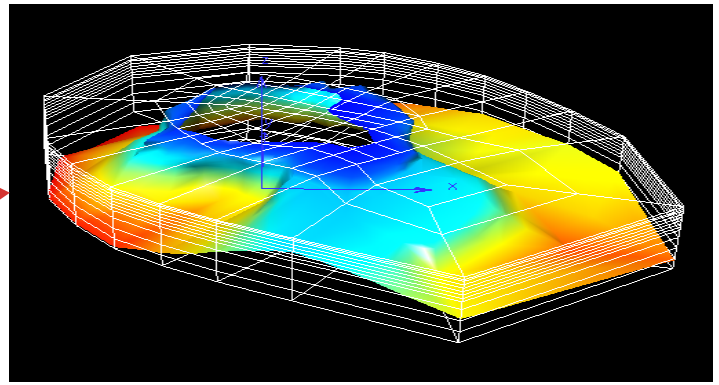
forward model

observation and model errors

- Data are limited in number, noisy, and *indirect*
- Forward model may be computationally intensive

UQ for inverse problems

- ***Example:*** *subsurface flow and transport*



$$d = G(x) + \epsilon$$

forward model (PDE)

observation/model errors

- Parameter x is **high-dimensional**—in principle, infinite-dimensional, i.e., a function $x(s)$

UQ for inverse problems

- A ***statistical perspective*** is essential to uncertainty quantification for inverse problems:
  - To *characterize uncertainty* in the inverse solution, and to understand how this uncertainty depends on observations and other sources of information (e.g., prior distributions)
  - To make probabilistic *predictions*
  - To choose “good” observations or experiments (optimal experimental design)
  - To address questions of model error, model validation, and model selection

Bayesian inference

- We adopt a **Bayesian** approach:

$$p(x \mid d) = \frac{p(d \mid x)\, p(x)}{p(d)}$$

Key idea: model parameters x are treated as random variables

- Ingredients of Bayes' rule:
  - x are model parameters; d are the data (here, assume both to be finite-dimensional)
  - $p(x)$ is the *prior* probability density
  - $L(x) \equiv p(d \mid x)$ is the *likelihood function*
  - $p(d)$ is the evidence or marginal likelihood
  - $p(x \mid d)$ is the posterior probability density: a complete description of uncertainty in the inverse solution

Bayesian inference

- *Likelihood functions* for inverse problems:
  - In general, $p(d \mid x)$ derives from a ***probabilistic model for the data***
  - Examples:

$$d = G(x) + \epsilon$$

deterministic model + measurement noise

$$d_i = G(x, s_i) + \epsilon_i$$

deterministic model + measurement noise (observations indexed by *s*)

$$d_i = G(x, s_i) + \eta(s_i) + \epsilon_i$$

deterministic model + **model discrepancy** + measurement noise

$$d_i = G(x, s_i; \omega)$$

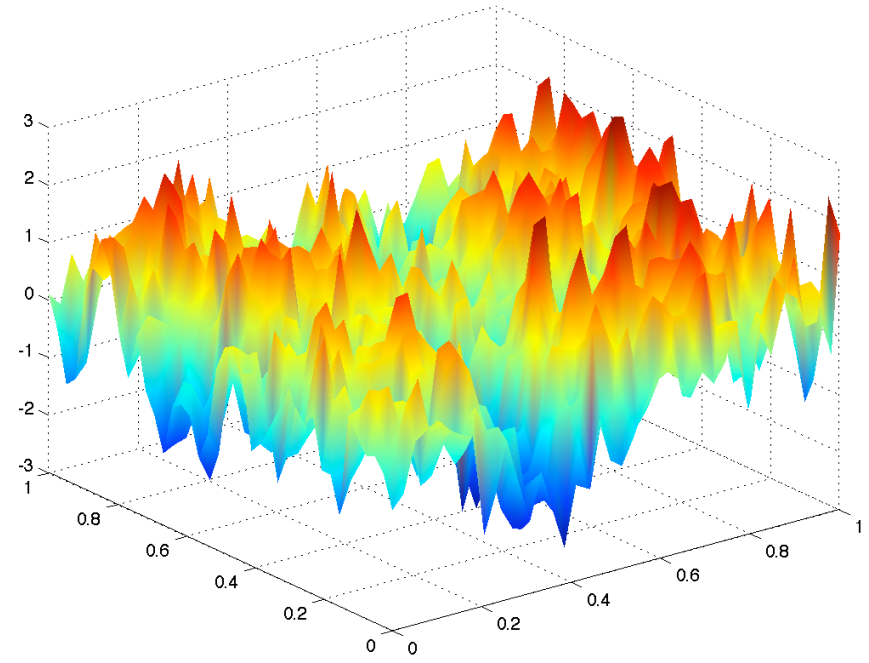
more complicated noise structure OR stochastic forward model

Bayesian inference

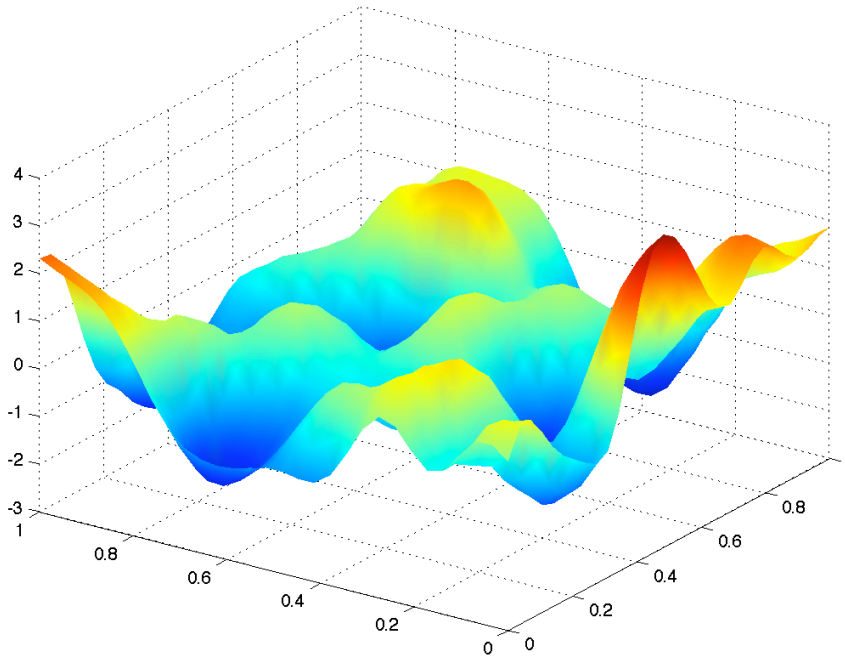
- *Prior distributions* for inverse problems
  - For point parameters: subjective priors and expert judgment; Jeffreys priors; other reference (‘non-informative’) priors
  - For distributed parameters:
    - Gaussian processes with specified covariance kernel
    - Gaussian Markov random fields [Rue/Held 2005]
    - Gaussian priors derived from differential operators [Stuart 2010]
    - Wavelet-based Besov space priors [Lassas 2009]

Bayesian inference

- *Prior distributions* for inverse problems
  - Example: stationary Gaussian random fields



(exponential covariance kernel)



(Gaussian covariance kernel)

Bayesian inference

- *Prior distributions* for inverse problems
  - **Hierarchical priors** can be very useful
  - Example:

$$x(s) \sim \mathcal{GP}\left(\mu(s), C(s, s')\right)$$

$$C(s, s') = \sigma^2 \exp\left(-\frac{1}{p}\left|\frac{s - s'}{L}\right|^p\right)$$

  - ➢ Jointly infer σ^2, L, and some finite-dimensional parameterization of x (for instance, coefficients of its Karhunen-Loève expansion), e.g.

$$p(x, \sigma^2, L \mid d) \propto p(d \mid x)\, p(x \mid \sigma^2, L) p(\sigma^2) p(L)$$

Computational challenges

- How to simulate from or *explore* the posterior distribution?
  - Posterior mode, mean, higher moments; quantiles; credible intervals; realizations…

- How to make Bayesian inference computationally tractable when *statistical* models contain expensive *physical* models (e.g., PDEs)?

- This lecture will focus on two approaches (out of many) for addressing the second question:
  1. **Approximations** of the forward model
  2. **Dimension reduction** and its relationship to posterior sampling schemes

First: Markov chain Monte Carlo

In general, MCMC provides a means of sampling ("simulating") from an arbitrary distribution.

- The density $\pi(x)$ need be known only up to a normalizing constant
- Utility in *inference* and *prediction*: write both as posterior expectations, $\mathbb{E}_\pi f$.

Then

$$\mathbb{E}_\pi f \approx \frac{1}{n} \sum_i^n f\left(x^{(i)}\right)$$

- $x^{(i)}$ will be asymptotically distributed according to π
- $x^{(i)}$ will **not** be i.i.d. In other words, we must pay a price!

Metropolis-Hastings algorithm

A simple recipe!

1. Draw a proposal y from $q(y|x_n)$
2. Calculate acceptance ratio

$$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)}\right\}$$

3. Put

$$x_{n+1} = \begin{cases} y, & \text{with probability } \alpha(x_n, y) \\ x_n, & \text{with probability } 1 - \alpha(x_n, y) \end{cases}$$

With appropriate conditions on the proposal and target, this defines the *transition kernel* of a Markov chain with π as its stationary and limiting distribution

MCMC estimates

What about the **quality** of MCMC estimates?
What is the price one pays for correlated samples?

Compare Monte Carlo (iid) and MCMC estimates of $\mathbb{E}_\pi h$:

Monte Carlo

$$\text{Var}\left[\bar{h}_n\right] = \frac{\text{Var}_\pi\left[h(X)\right]}{n}$$

MCMC

$$\text{Var}\left[\bar{h}_n\right] = \frac{\text{Var}_\pi\left[h(X)\right]}{n}\theta$$

where

$$\theta = 1 + 2\sum_{s>0}^{\infty} \text{corr}\left(h(X_i), h(X_{i+s})\right)$$

is the **integrated autocorrelation**.

Some observations

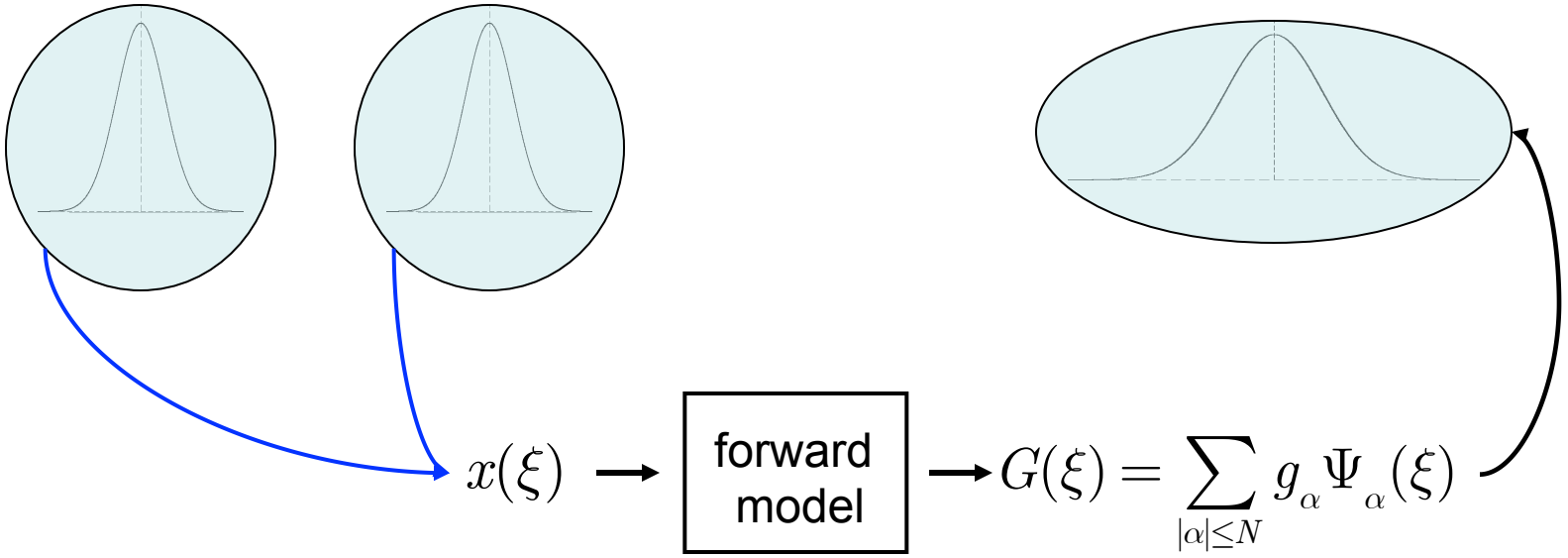
- MCMC requires *many* evaluations of the unnormalized posterior density

- Achieving "good mixing" (getting closer to i.i.d. sampling) is essential. This is an area of **enormous effort and innovation**
  - Langevin MCMC
  - Preconditioned Langevin MCMC, using Hessian information
  - Differential geometric MCMC (Fisher information metric)
  - Hamiltonian MCMC
  - Adaptive Metropolis-Hastings schemes; adaptive Metropolis independence samplers
  - MCMC on function space; discretization invariance [Cotter 2012]
  - Much more…

1. Forward model approximations

- **First idea:** approximate the forward model over the *prior support* of the parameters x
- **How?** Stochastic spectral methods are quite useful in this context. Exploit *regularity* in the parameter dependence of the forward model.

Stochastic spectral methods for BIPs



- *Propagate prior uncertainty through the forward model*
- Equivalently, solve the stochastic ODE/PDE (with uncertain parameters, initial conditions, boundary conditions) determined by the prior
- Use your favorite stochastic spectral approach (intrusive, non-intrusive)
- Spectral expansion *replaces* the forward model in the likelihood function. No further forward model solutions!

Forward model approximation

- Simplest surrogate posterior density (assuming $\xi \equiv x$), with approximation order N

$$\pi^N(\xi) = \prod_{i=1}^{m} p_\epsilon\left(d_i - G_i^N(\xi)\right) p_\xi(\xi)$$

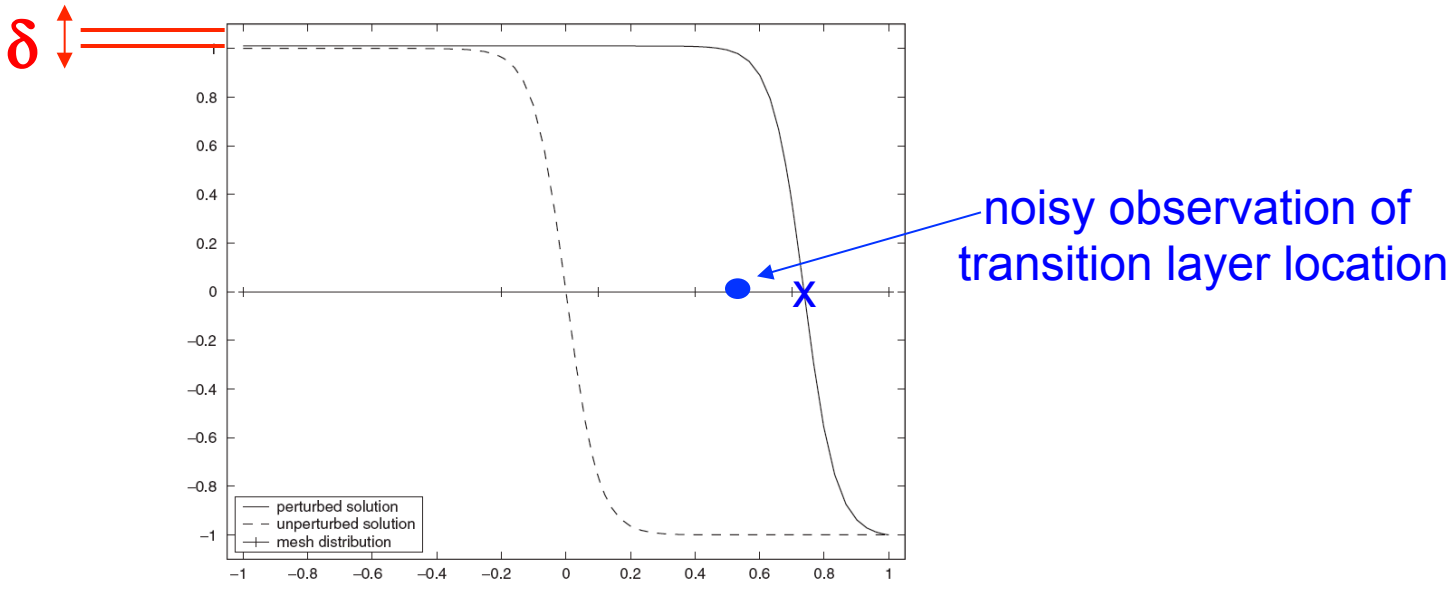
- Convergence of the **forward** approximation implies convergence of the **posterior** distribution:
  - Assume observational error ε is i.i.d. Gaussian
  - If $\left\| G_i(\xi) - G_i^N(\xi) \right\|_{L_p^2} \le CN^{-\alpha}, \quad 1 \le i \le m, \quad \alpha > 0$

    then $D_{KL}\left(\pi^N \middle\| \pi\right) \lesssim N^{-\alpha}$ for sufficiently large *N*.

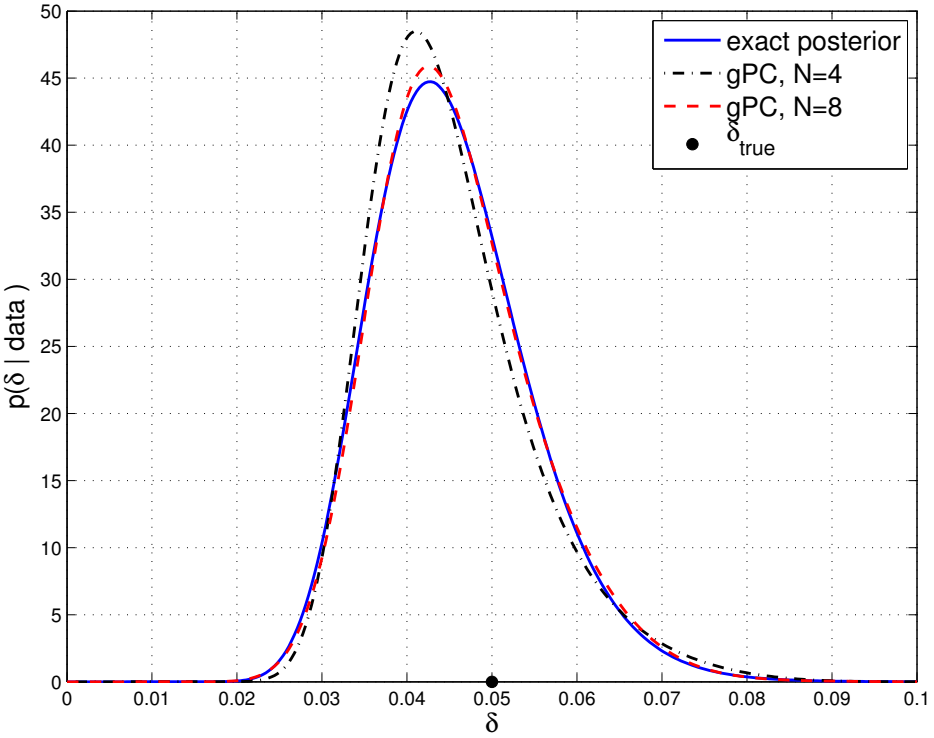
[M & Najm 2007, M & Xiu 2009]

Example: Burgers equation

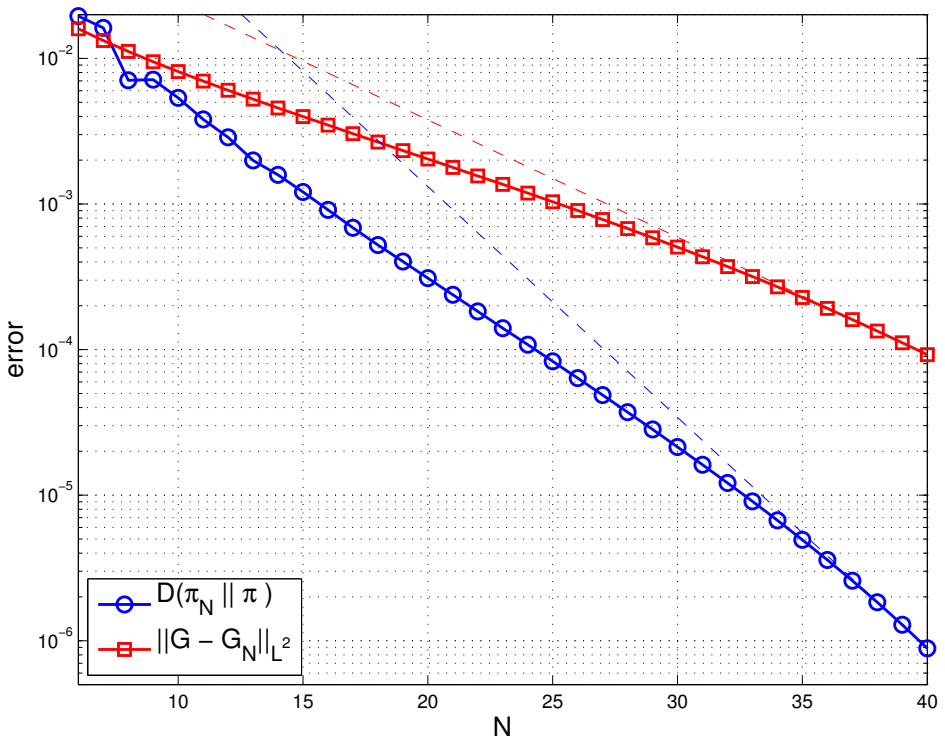
- **Example:** estimate *boundary condition* of viscous **Burgers equation**
  - $u_t + uu_x = \nu u_{xx}, \quad x \in [-1,1], \quad u(-1) = 1 + \delta$
  - Super-sensitivity to perturbation δ
  - Steady-state solution:



Example: Burgers equation



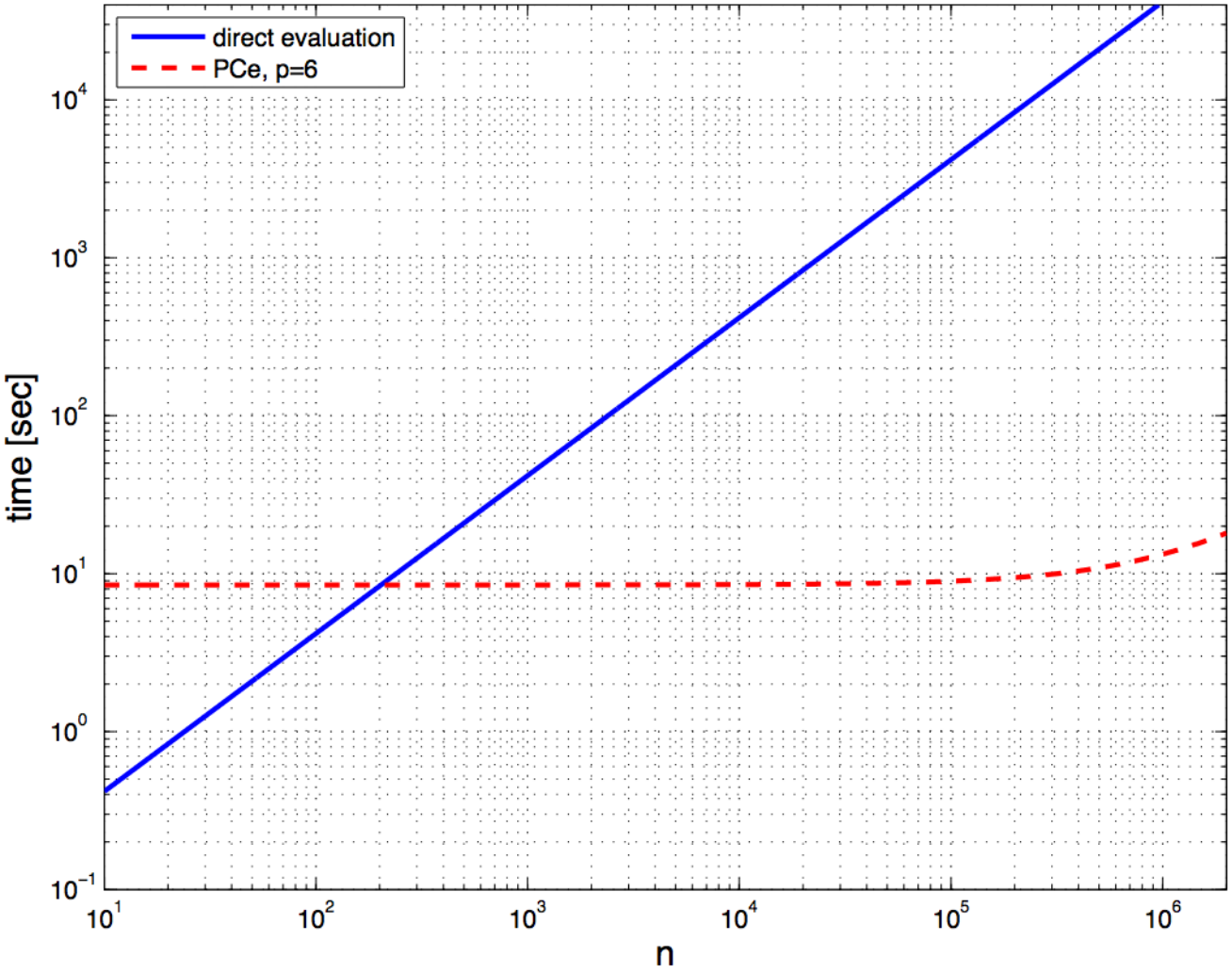
posterior density



convergence of the forward model and the posterior distribution
(factor of 2 increase in rate for uniform priors; Birolleau et al. 2012)

Speedup

- Total computational time *vs* number of posterior samples



- Forward uncertainty propagation (red line offset) occurs *offline*
- Per-sample cost reduced by 3–4 orders of magnitude

Other approximation approaches

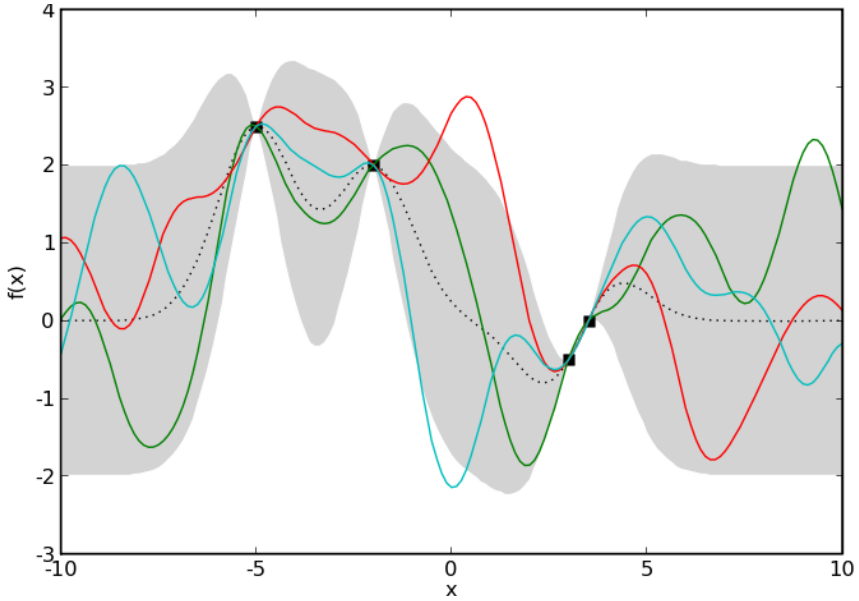
- Stochastic spectral methods are very useful, but not the only option!

- Projection-based reduced order models: POD and reduced-basis methods for parameterized PDEs
  - Again, ensure accuracy over the prior distribution; greedy snapshot selection procedures
  - Nguyen/Patera 2010, Lieberman/Willcox 2010

- *Delayed acceptance* MCMC schemes add a second stage to the Metropolis scheme [Christen & Fox 2005, Cui 2011]
  - Use the full forward model to screen proposals that are accepted using the reduced model
  - Ensure sampling from the exact posterior distribution
  - *At a price:* continual evaluations of the forward model during MCMC

Other approximation approaches

- The statistics community often takes a very different perspective: *Gaussian process regression*
  - Roots in ‘design and analysis of computer experiments’ (Sacks et al. 1989), emulation of computer models (Kennedy & O’Hagan 2001)
  - At any input parameter value, the forward model output is a random variable (Bayesian perspective); contributes to posterior uncertainty…



  - Requires experimental design to choose model evaluation points

Adaptive approx for inference

- Constructing an accurate surrogate over the entire parameter space is still somewhat wasteful
  - Posterior concentrates on a small fraction of the prior support; particularly for high-dimensional problems
  - Localizing a surrogate mitigates the impact of nonlinearity
- Can we construct a surrogate only over the support of the **posterior**? How to do this before characterizing the posterior?

Adaptive approx for inference

- Constructing an accurate surrogate over the entire parameter space is still somewhat wasteful
  - Posterior concentrates on a small fraction of the prior support; particularly for high-dimensional problems
  - Localizing a surrogate mitigates the impact of nonlinearity
- Can we construct a surrogate only over the support of the **posterior**? How to do this before characterizing the posterior?
- **Adaptive** approach, based on the cross-entropy method and importance sampling:
  - Construct a sequence of "cheap" surrogates and biasing distributions that converges to the posterior
  - Surrogates (e.g., polynomial chaos expansions) remain *local* and *low-order*

Adaptive approx for inference

- Overall procedure:
  - Seek a biasing distribution that is close to the posterior $\pi^{\mathrm{d}}(x) \propto L(x)p(x)$
  - Pick biasing distribution $q(x)$ from a simple family of distributions, parameterized by v

$$\min_v D_{\mathrm{KL}}\left(\pi^d(x) \,\|\, q(x;v)\right) \;\leftrightarrow\; \max_v \int L(x)p(x)\log q(x;v)\,dx$$

- Iterative approach:
  - Estimates of $D(v) = \int L(x)\log q(x,v)p(x)dx$ based on naïve sampling from the prior have enormous variance
  - Instead use *sequential importance sampling* to estimate $D(v)$ via a sequence of biasing distributions $q(x;v_m)$

$$v_{m+1} = \arg\max_v \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)})\log q(x^{(i)},v)\frac{p(x^{(i)})}{q(x^{(i)},v_m)}, \quad \text{with } x^{(i)} \sim q(x^{(i)},v_m)$$

  - Maximization problem at each step can be solved easily

Adaptive approx for inference

- Iterative approach (cont.):

$$v_{m+1} = \arg\max_{v} \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}) \log q(x^{(i)}, v) \frac{p(x^{(i)})}{q(x^{(i)}, v_m)}, \quad \text{with } x^{(i)} \sim q(x^{(i)}, v_m)$$

  - At each iteration use a localized ***surrogate*** for the forward model, based on $q(x, v_m)$, to evaluate the likelihood function
  - Can further accelerate convergence via a *tempering* approach, replacing likelihood with $L(x;\lambda) = L^{1/\lambda}(x)$

- Convergence: if forward approximations $G_m^N(x) \to G(x)$ in $L^2_{q(x,v_m)}$ as $N \to \infty$ then

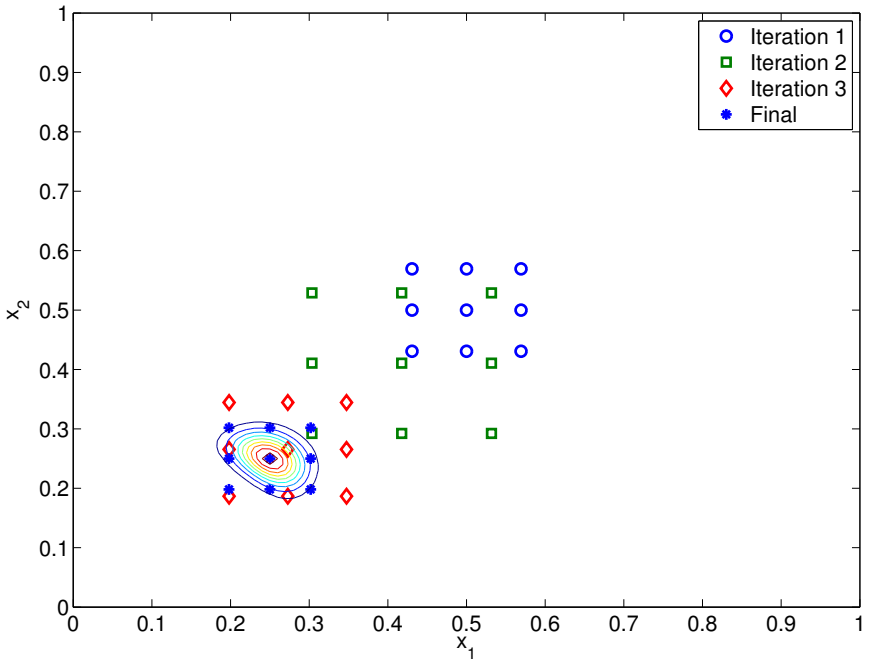
$$v^* \to \arg\min D_{KL}\left(\pi^d(x) \,\|\, q(x,v)\right) \text{ as } \lambda \to 1,\ n \to \infty, N \to \infty$$

- Example: Gaussian biasing distributions and Hermite polynomial chaos surrogates

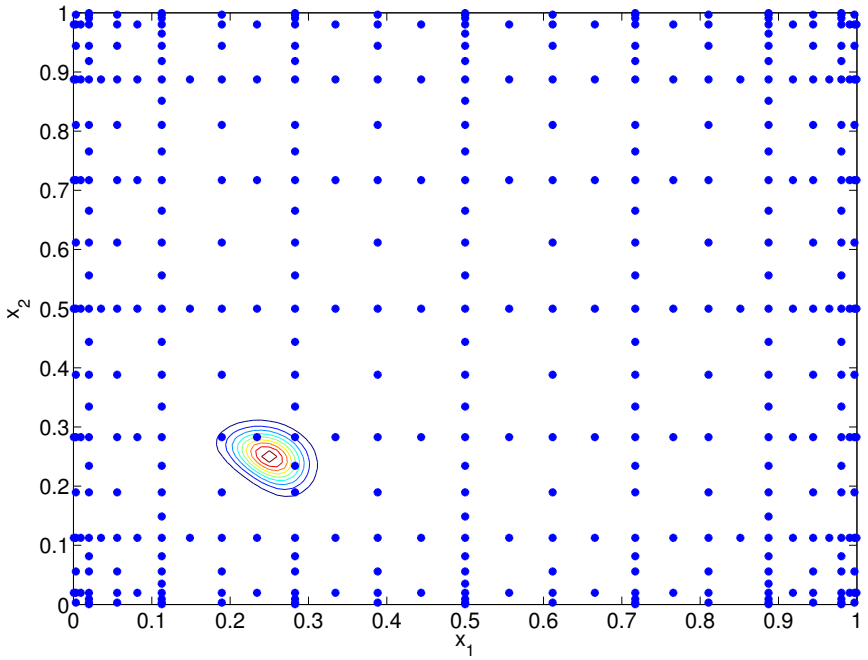
[Li & M 2013]

Adaptive approximations

- Example: 2-D source inversion problem
  (model evaluation points and *posterior density contours)*



adaptive surrogate



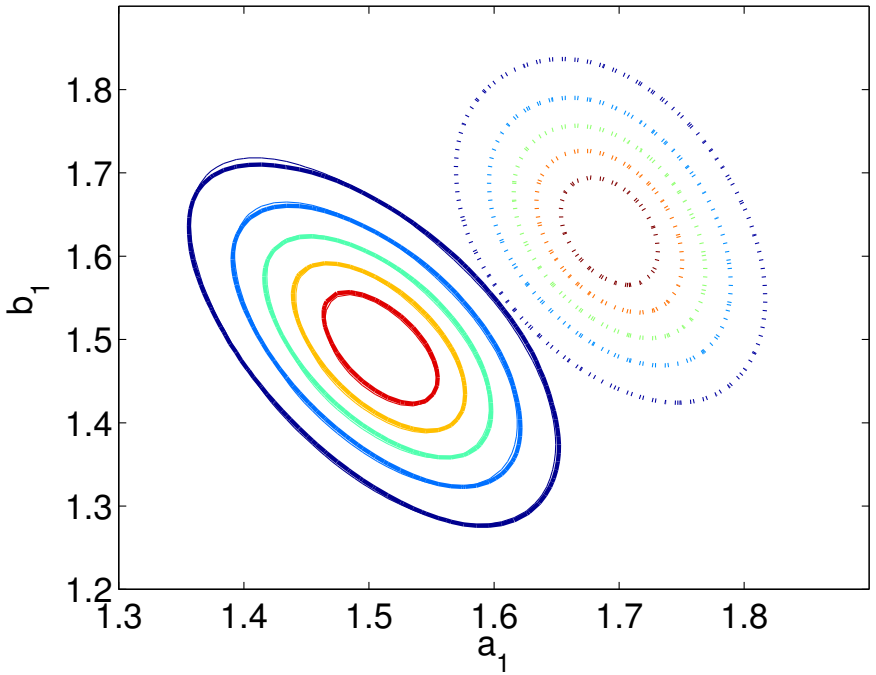
global surrogate

- Sparse grids used to construct polynomial chaos surrogates in both cases
- Number of model evaluations/polynomial order selected to achieve comparable accuracy!

Adaptive approximations

- Example: nonlinear inverse heat conduction problem
  - Infer boundary heat flux from internal temperature measurements; temperature-dependent conductivity $c(u)=1/(1+u^2)$
  - Heat flux parameterized with Fourier modes (11 dimensions)

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(c(u)\frac{\partial u}{\partial x}\right)$$

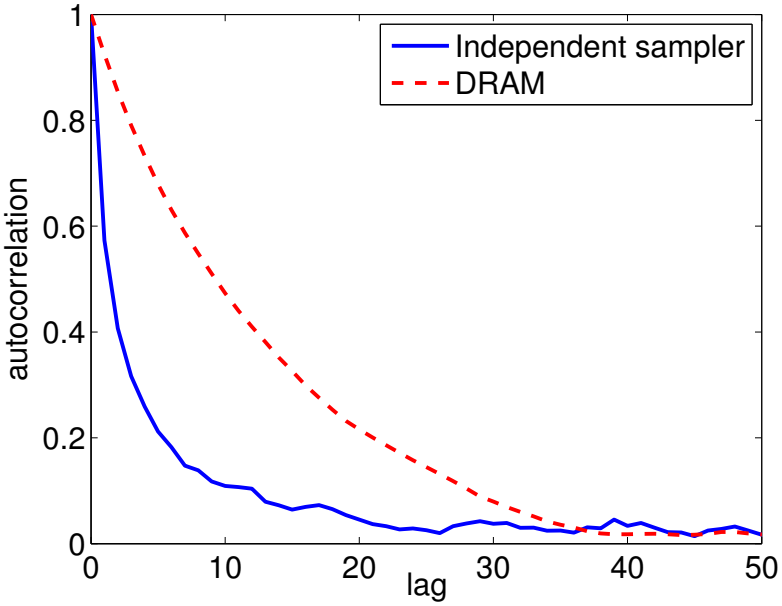


(thick solid line) full model
(thin solid line) **adaptive surrogate**
(dotted line) global surrogate

	# model evals	poly order	$D_{KL}\left(\pi \,\|\, \tilde{\pi}_{surr}\right)$
global surrogate	35929	5	8.37
adaptive surrogate	5763	2	0.0032

Adaptive approximations

- Final biasing distribution also provides a good foundation for efficient MCMC sampling (e.g., use as proposal in an *independence sampler)*



- This approach is not limited to polynomial chaos surrogates or Gaussian biasing distributions:
  - Projection-based reduced order models
  - *Mixtures* of *exponential family* distributions (e.g., for multi-modal posteriors)

Computational challenges

- How to simulate from or *explore* the posterior distribution?
  - Posterior mode, mean, higher moments; quantiles; credible intervals; realizations…

- How to make Bayesian inference computationally tractable when *statistical* models contain expensive *physical* models (e.g., PDEs)?

- This lecture will focus on two approaches (out of many!) for addressing the second question:
  1. **Approximations** of the forward model
  2. **Dimension reduction** and its relationship to posterior sampling schemes

2. Dimension reduction

- Suppose the object of inversion is a function $x(s)$ endowed with a Gaussian process prior $x \sim \mathcal{GP}(\mu, C)$
- What is a convenient finite-dimensional parameterization of x?

- Karhunen-Loève expansion:

$$x(s,\omega) = \mu(s) + \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, c_k(\omega)\phi_k(s)$$

where $\int_D C(s_1, s_2)\phi_k(s_2)ds_2 = \lambda_k \phi_k(s_1), \quad c_k \sim N(0,1)$

- If C is smoothing, one may truncate expansion at $K \ll n$ terms (where n is some grid/discretization size); one thus has an uncorrelated and lower-dimensional parameterization based on the prior

Dimension reduction

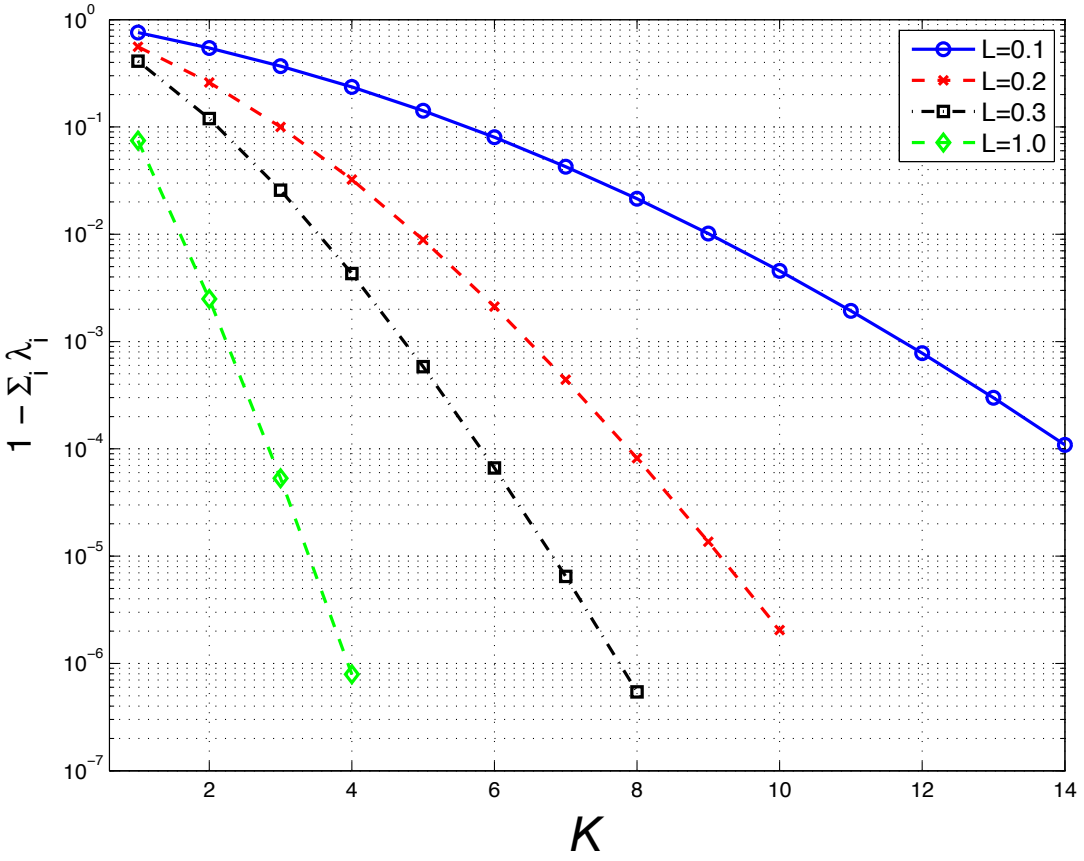


Figure : Decay of K-L eigenvalues with different prior correlation lengths L. Vertical axis shows the missing fraction of the prior variance, $1 - \sum_i^K \lambda_i$, versus K.

Dimension reduction

- Can we go further? What controls the *"intrinsic dimensionality"* of the inference problem?
  1. Smoothing from the prior distribution (e.g., correlation)
  2. How many observations are used
  3. How much the forward model "filters" the parameters (ill-posedness)
     - More formally, #2 and #3 contribute to ***low rank*** of the data misfit Hessian $H(x) = -\nabla_x^2 \log p(d \mid x)$

- *A priori* dimension reduction (via K-L expansion of the prior) will not always work:
  - Non-smooth priors
  - Important information in high-index modes (truth *not* drawn from prior)

- Low dimensionality instead lies in the **change** from prior to posterior…

Dimension reduction

- Why is reducing dimension useful?
  - Achieve complexity that is independent of grid/discretization
  - More efficient posterior sampling *(dimension scaling issues in MCMC)*
  - **Rao-Blackwellization** will reduce the variance in the Monte Carlo estimate of any posterior expectation

- Quick overview:
  - Introduce ideas in the linear-Gaussian case
  - Develop algorithms for **nonlinear** statistical inverse problems
  - Numerical demonstrations (elliptic PDE inverse problem)

Linear-Gaussian problem

- Begin with the simple linear-Gaussian model:

$$d = Gx + \epsilon, \;\; x \sim N(0, \Gamma_{pr}), \;\; \epsilon \sim N(0, \Gamma_{obs}); \;\; H \equiv G^T \Gamma_{obs}^{-1} G$$

- Consider the generalized Rayleigh quotient:

$$\frac{w^T H w}{w^T \Gamma_{pr}^{-1} w}$$

  - When quotient is large, likelihood limits variability in the w direction more strongly than the prior
  - When quotient is small, prior is more constraining (e.g., smoothing prior and/or a rough mode w to which the forward model is insensitive)

Linear-Gaussian problem

- This motivates a ***generalized eigenvalue problem:***

$$Hw = \lambda \Gamma_{pr}^{-1} w$$

  - H is symmetric and Γ is symmetric positive definite
  - Solutions simultaneously diagonalize the log-likelihood Hessian and the prior:

$$W^T H W = \mathrm{diag}(h_1, \ldots, h_n)$$
$$W^T \Gamma_{pr}^{-1} W = \mathrm{diag}(g_1, \ldots, g_n)$$
$$\lambda_i = h_i \,/\, g_i$$

  - Can equivalently solve "prior-preconditioned Hessian" eigenproblem:

$$L^T H L z = \lambda z$$

    where $\Gamma_{pr} = L L^T$ and $W = LZ$ (put $g_i = 1$)

  - Analogy with ***balanced truncation***

Linear-Gaussian problem

- Solution of this generalized eigenproblem yields a **low-rank** expression for the change between prior and posterior covariance

$$\Gamma_{post} = \Gamma_{pr} - W \Sigma W^T \approx \Gamma_{pr} - W_r \Sigma_r W_r^T$$

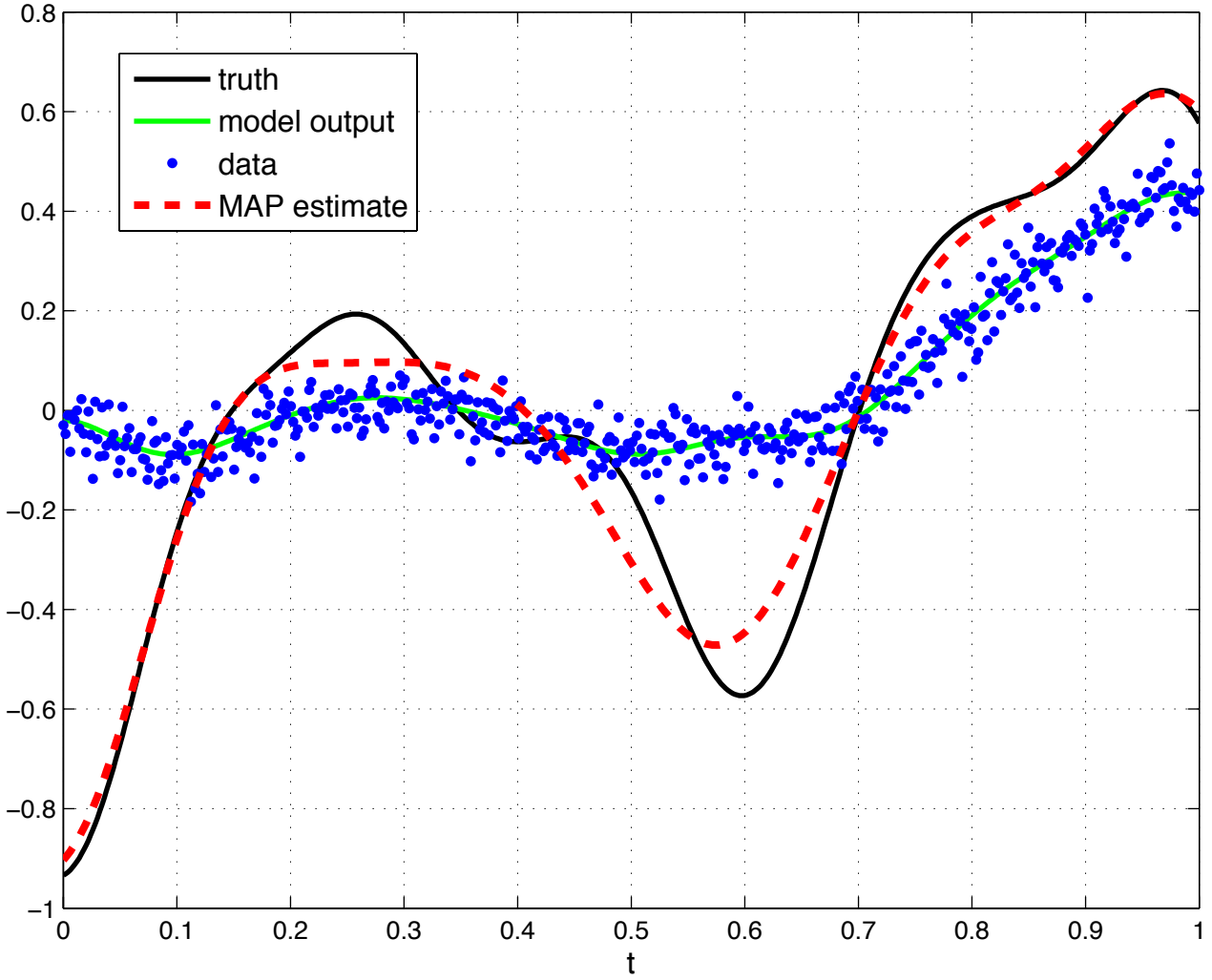
$$\text{where } \Sigma = \text{diag}\left(\ldots, \frac{\lambda_i}{1+\lambda_i}, \ldots \right)$$

- Basis $\mathbf{W}$ appropriately combines information from likelihood and prior (alignment of eigenspaces, low-rank structure of each)
  - Large λ is likelihood-dominated; and vice-versa
  - $\lambda=1$ is roughly balanced (Rayleigh quotient)

- Conjecture: in linear-Gaussian problems, $\mathbf{W}$ yields the best possible rank-*r* approximation of the posterior (e.g., in Hellinger distance)

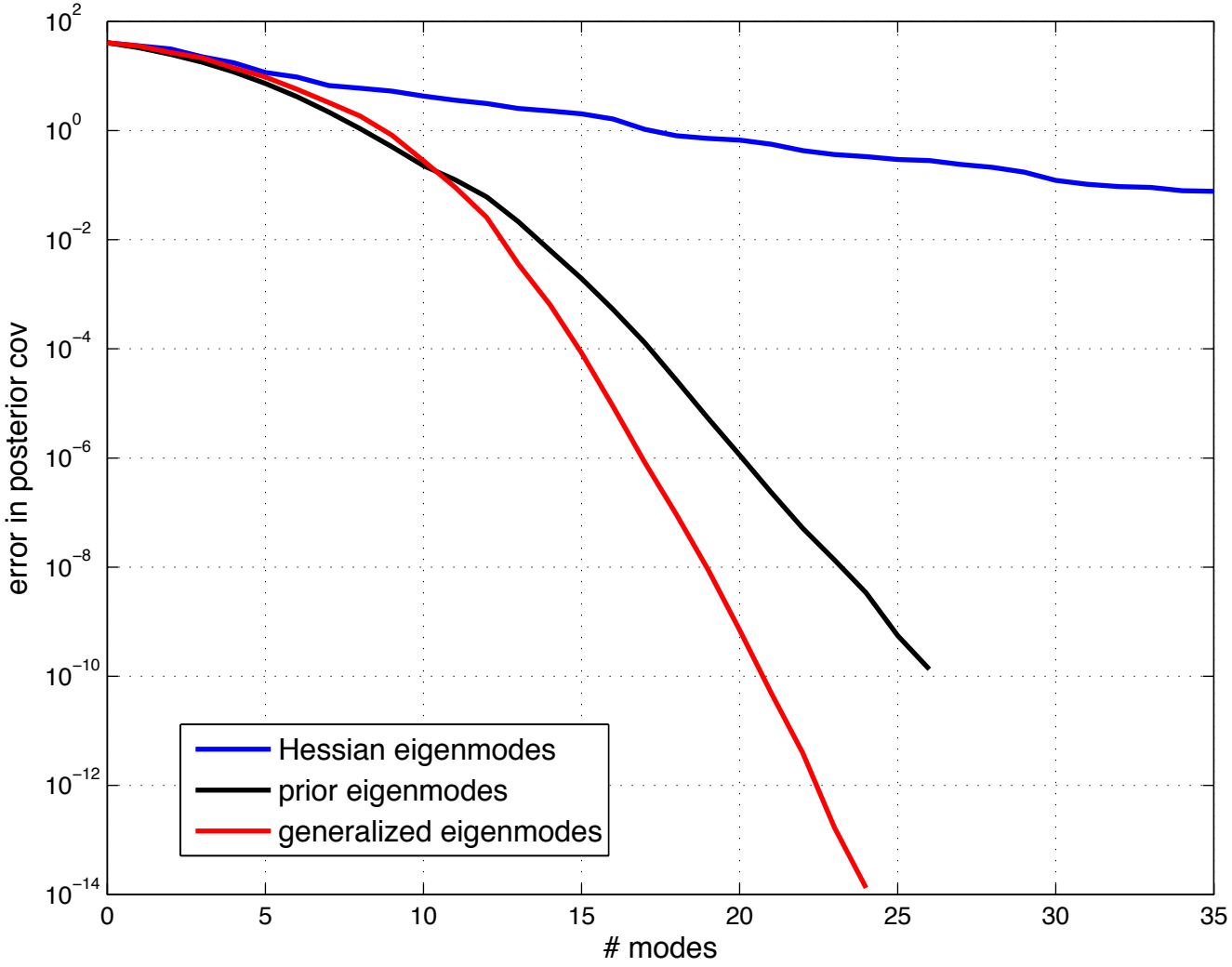
Dimension reduction

- Example: deconvolution problem, *smoothing prior*



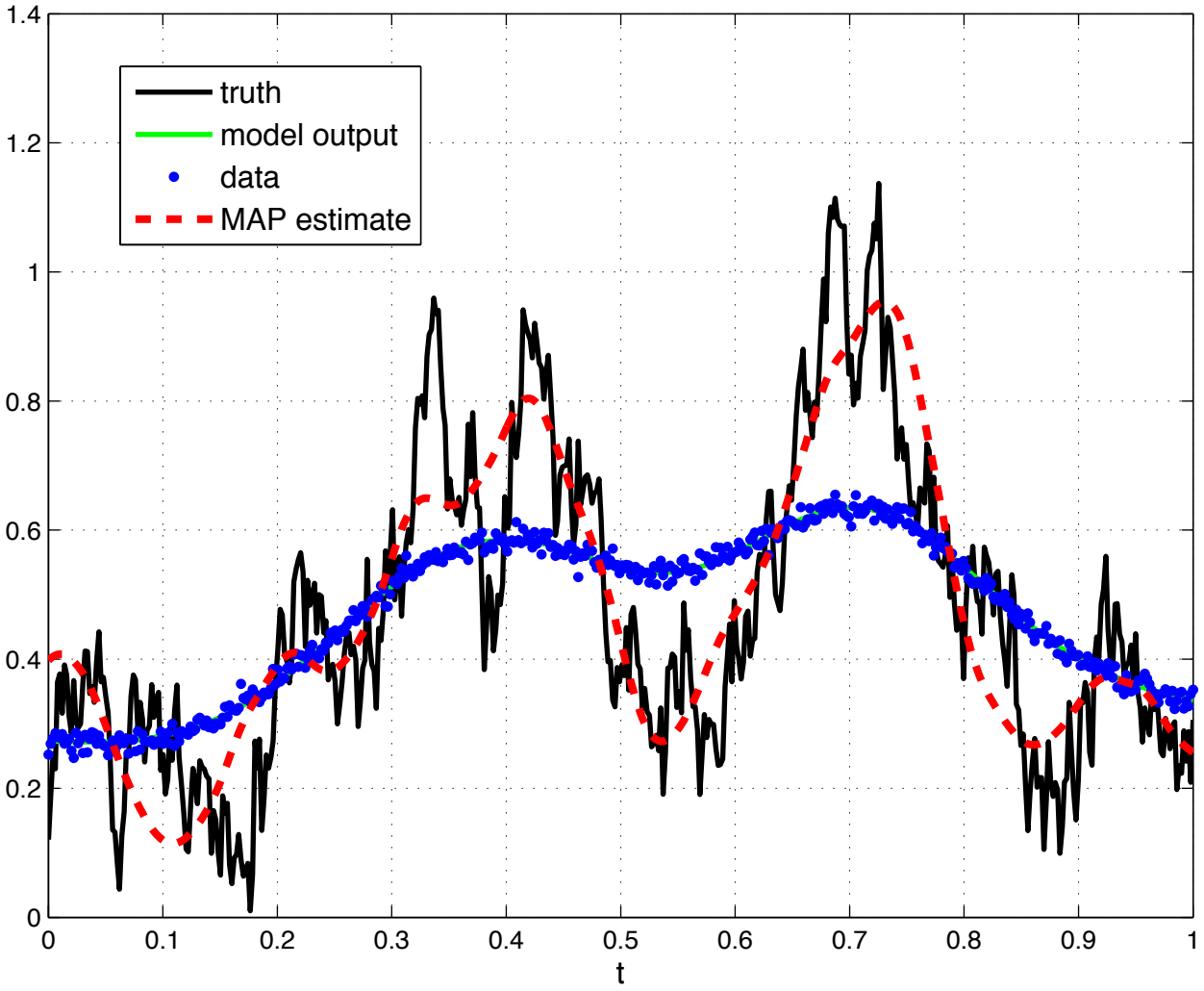
Dimension reduction

- Example: deconvolution problem, *smoothing prior*



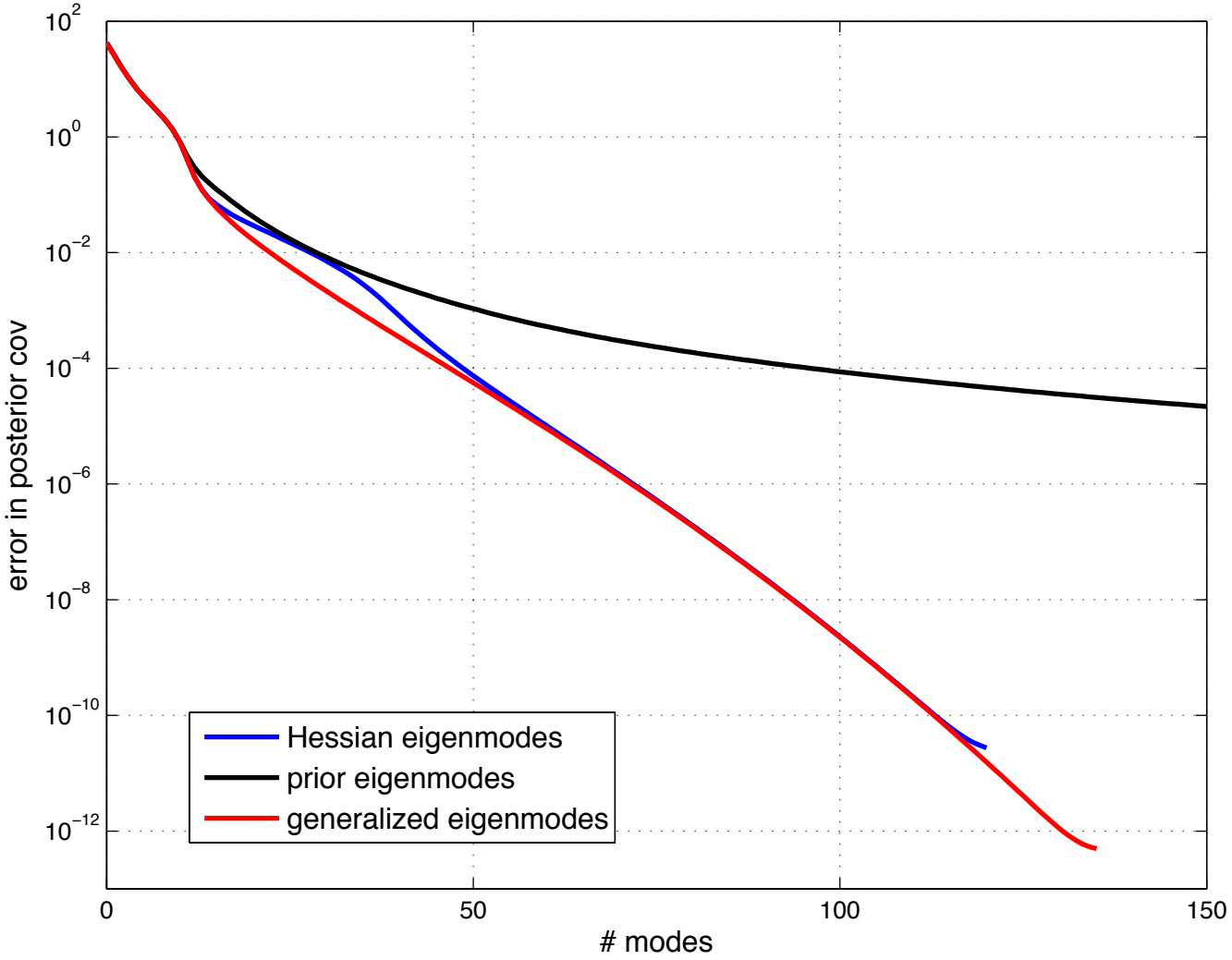
Dimension reduction

- Example: deconvolution problem, *rough prior*



Dimension reduction

- Example: deconvolution problem, *rough prior*



Nonlinear problems

- OK, but can we apply this idea to ***nonlinear*** inverse problems?
  - Key challenge: log-likelihood Hessian $H(x)$ now *varies* over the parameter space
- Simple idea: *combine* locally important directions, over the *support of the posterior*, to yield a *global reduced basis*

Nonlinear problems

- Suppose we have posterior samples $\{x_i\}, \ i = 1 \ldots K$
  - Solve eigenproblem at each sample

$$L^T H(x_i) L z = \lambda z \qquad \Rightarrow Z_{r,i}, \Lambda_{r,i}$$

  - Truncate and collect local dominant directions in matrix M

$$M = \left[\ \frac{1}{\sqrt{K}} Z_{r,1} \Lambda_{r,1}^{1/2} \ \middle| \ \ldots \ \middle| \ \frac{1}{\sqrt{K}} Z_{r,K} \Lambda_{r,K}^{1/2} \ \right]$$

  - Take SVD $M \approx U_r S_r V_r^T$ to get global reduced basis

  - Thresholds: if local truncation is at eigenvalue λ^*, truncate global SVD at $s^* = \mathcal{O}(\sqrt{\lambda^*})$

Nonlinear problems

- How to obtain samples x_i?
- **Algorithm:**
  - Compute posterior mode via deterministic optimization
  - Compute local reduced basis at $x_1 = x_{map}$
  - for $\mathrm{k}=2\ldots\mathrm{K}$ do
    - *subchain:* perform several steps of MCMC in current global reduced basis $\mathrm{span}(\mathrm{U_r})$
    - *perturb:* propose from the prior in complementary dirs; Metropolize
    - *local eigenproblem:* collect x_k at end of MCMC subchain; compute local reduced basis $Z_{r,k}$, $\Lambda_{r,k}$
    - *update:* update the global reduced basis via $\mathrm{svd}(\mathrm{M_{1:k}})$
    - Project sample onto new global reduced basis
- Useful features:
  - Good initial proposal covariance for MCMC subchain is $\mathrm{diag}(\ldots, 1/(1+s_i^2), \ldots)$
  - Global modes $W_r = \Gamma_{pr}^{1/2} U_r$ are approximately uncorrelated

Posterior decomposition

- Posterior after dimension reduction:

$$p(x \mid d) \propto L(x)\, p(x)$$
$$\approx L(\mathcal{P}x)\, p(x)$$

where $\mathcal{P} = \Gamma^{1/2}_{pr} U_{r} U^{T}_{r} \Gamma^{-1/2}_{pr}$

- Alternatively, think of decomposing x:

$$x = \Gamma^{1/2}_{pr} U_{r} c_{r} + \Gamma^{1/2}_{pr} U^{\perp}_{r} c^{\perp}_{r}$$

Posterior decomposition

- Posterior after dimension reduction:

$$p(x \mid d) \propto L(x)\, p(x)$$
$$\approx L(\mathcal{P}x)\, p(x)$$

where $\mathcal{P} = \Gamma^{1/2}_{pr} U_r U^T_r \Gamma^{-1/2}_{pr}$

- Alternatively, think of decomposing x:

$$x = \Gamma^{1/2}_{pr} U_r c_r + \Gamma^{1/2}_{pr} U^{\perp}_r c^{\perp}_r$$

condition on data

independent of data

Posterior decomposition

- Posterior after dimension reduction:

$$p(x \mid d) \propto L(x)\, p(x)$$
$$\approx L(\mathcal{P}x)\, p(x)$$

where $\mathcal{P} = \Gamma^{1/2}_{pr} U_{r} U^{T}_{r} \Gamma^{-1/2}_{pr}$

- Alternatively, think of decomposing x:

$$x = \Gamma^{1/2}_{pr} U_{r} c_{r} + \Gamma^{1/2}_{pr} U^{\perp}_{r} c^{\perp}_{r}$$
$$= \Gamma^{1/2}_{pr} U_{r} c_{r} + \Gamma^{1/2}_{pr} (I - U_{r} U^{T}_{r}) z, \quad z \sim N(0, I)$$

Numerical examples

- Elliptic PDE in two spatial dimensions

$$\nabla \cdot \Big(\kappa(s) \, \nabla u \Big) = -f(s)$$

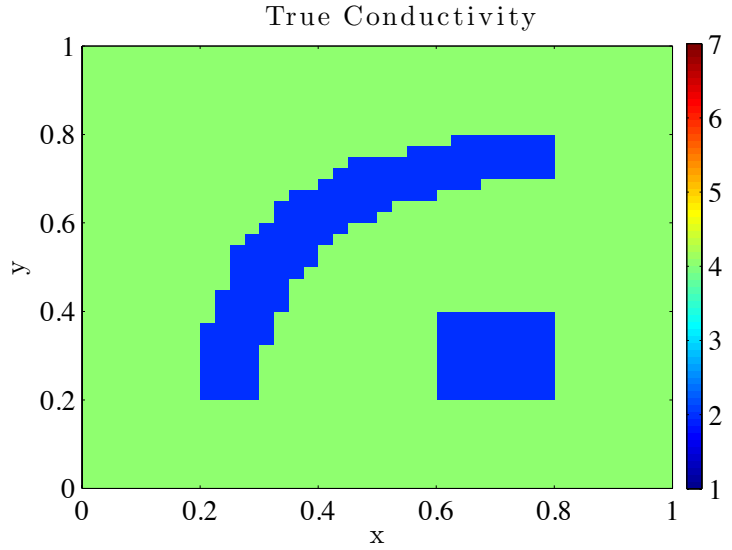
  - Estimate κ from noisy observations of u
  - Log-normal prior on $\kappa(s)$ with an **exponential** covariance kernel

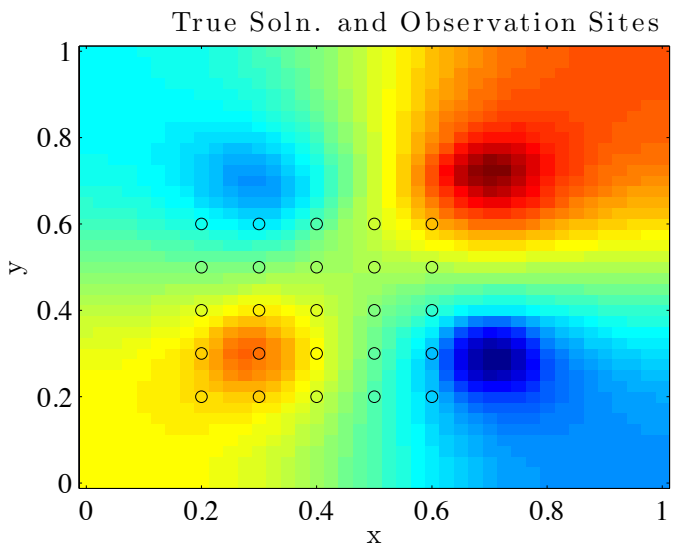
$$\log \kappa \sim GP\Big(0, C\Big)$$

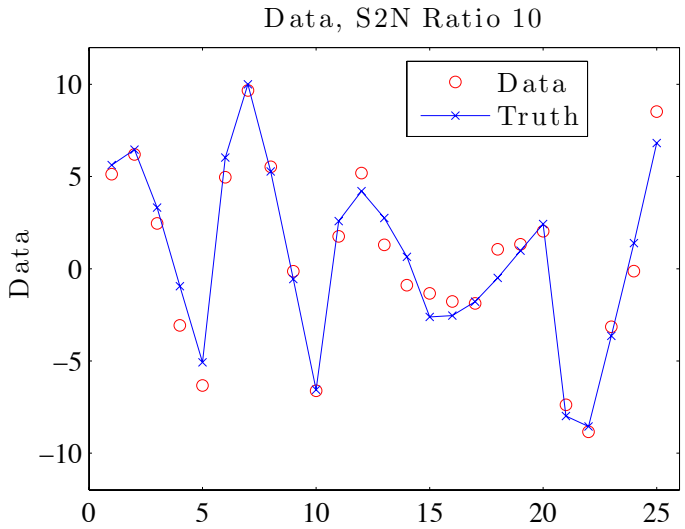
$$C\Big(s_1, s_2\Big) = \sigma^2 \exp\left(- \frac{\big\| s_1 - s_2 \big\|}{L_c} \right)$$

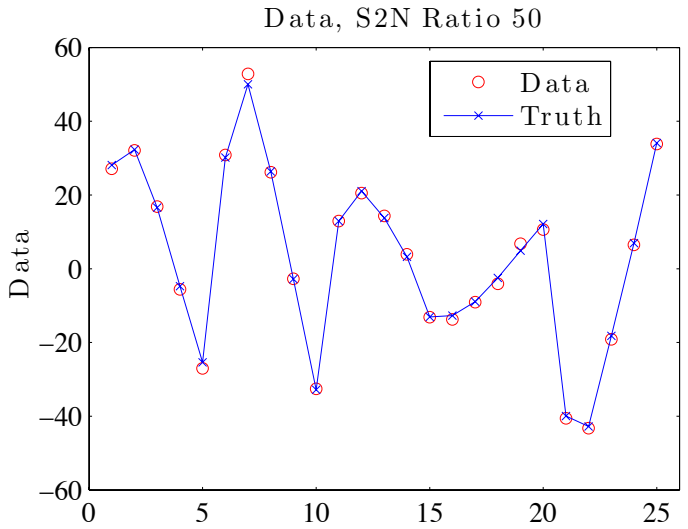
  - Use L_c = 0.25, discretize problem on a 40x40 or 80x80 grid

Problem setup



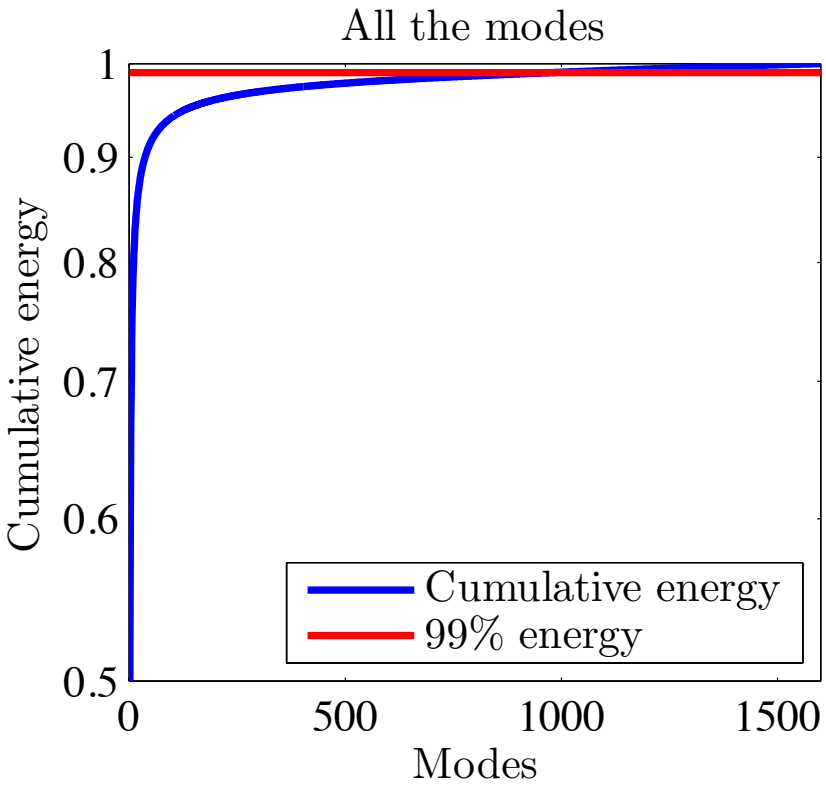


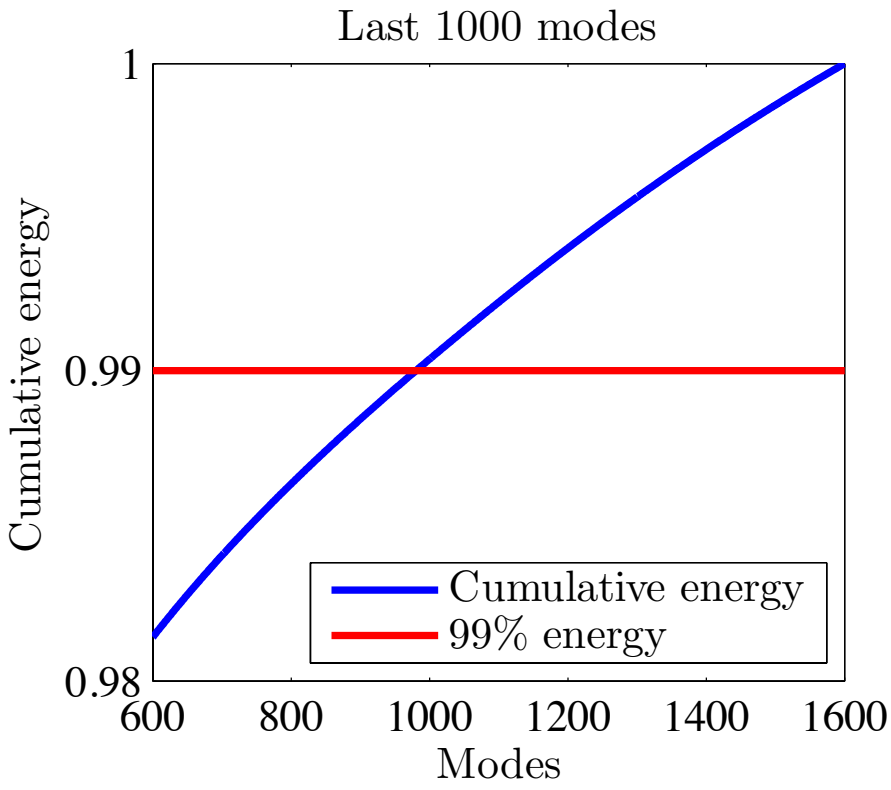




Numerical examples

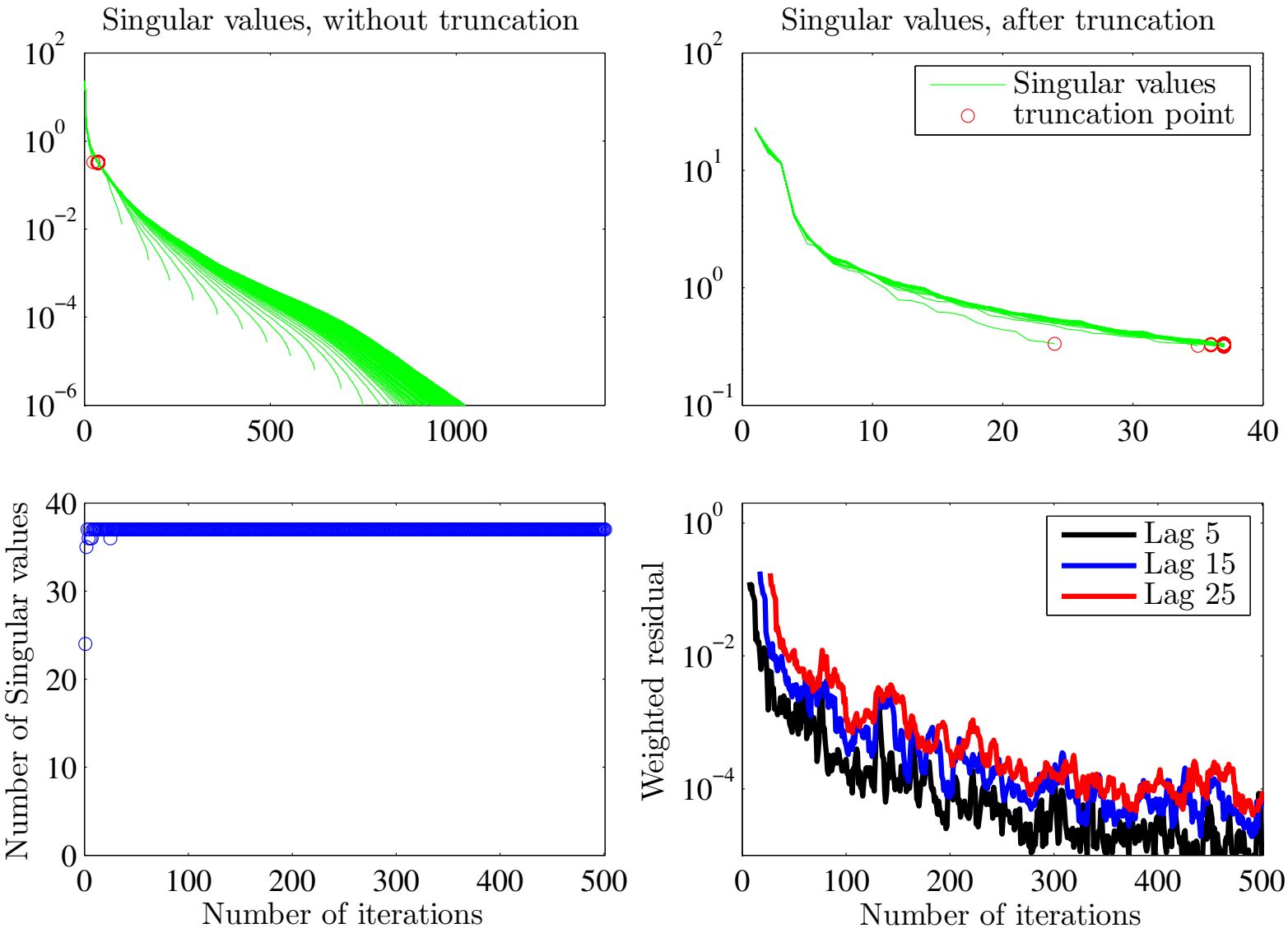
- Truncation based on the prior is insufficient for this problem





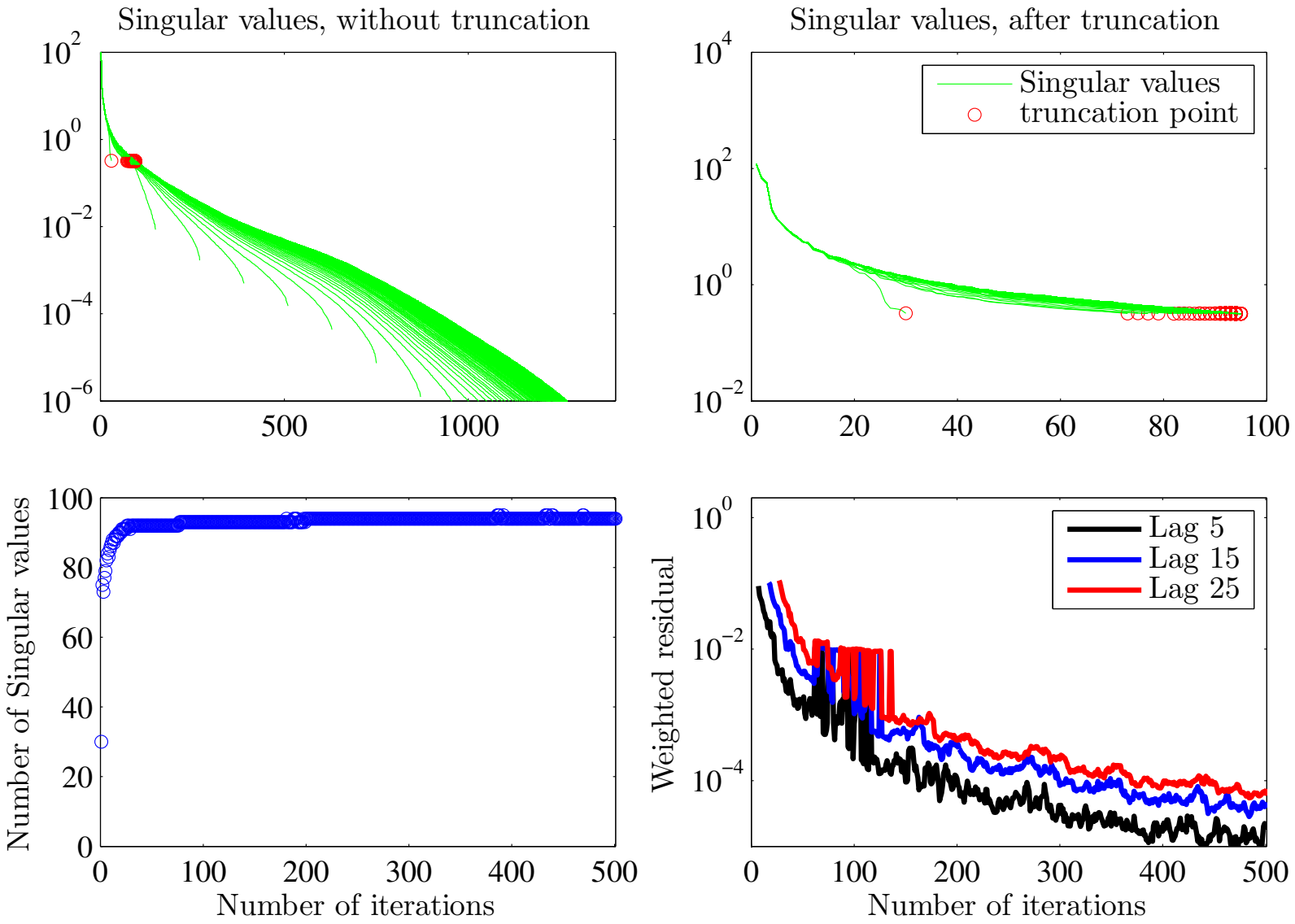
Subspace construction

S2N ratio = 10



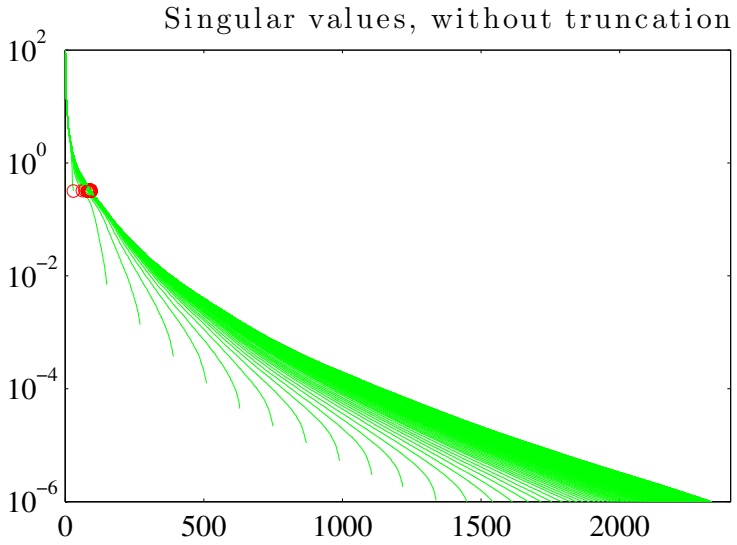
Subspace construction

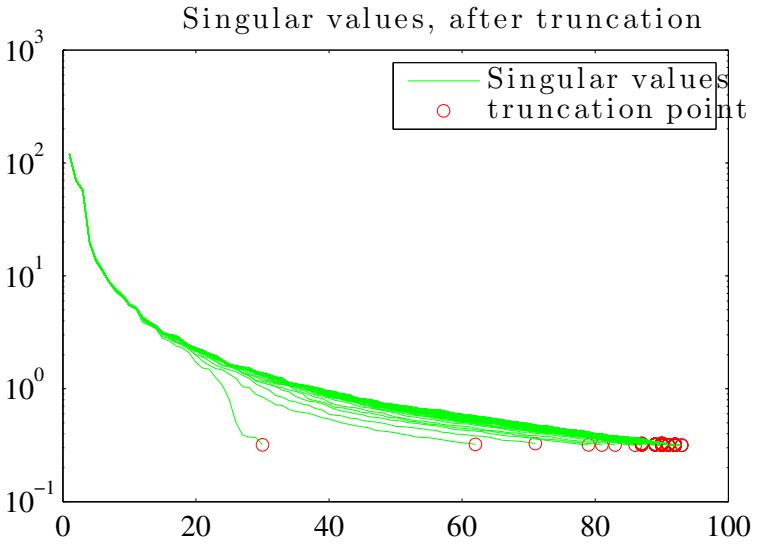
S2N ratio = 50

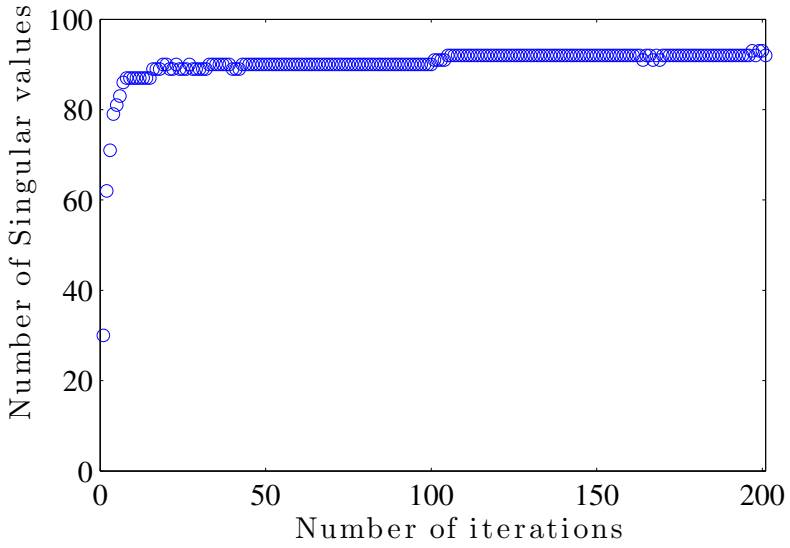


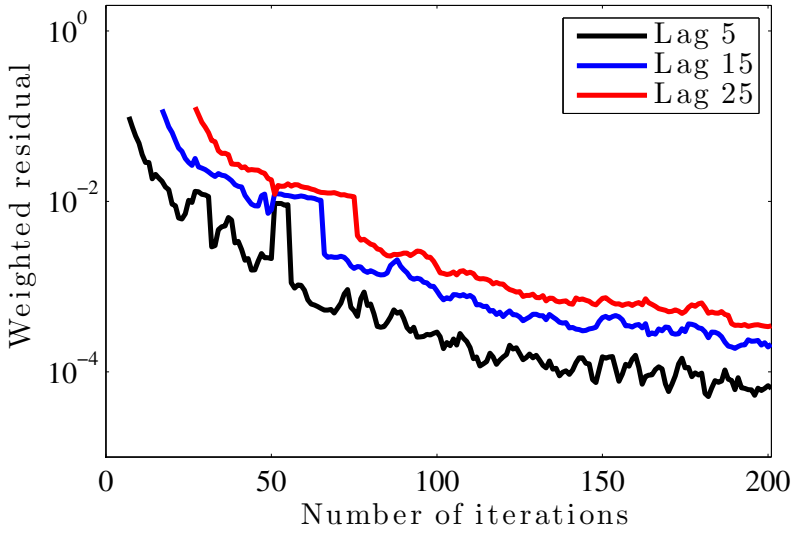
Subspace construction

S2N ratio = 50, finer grid



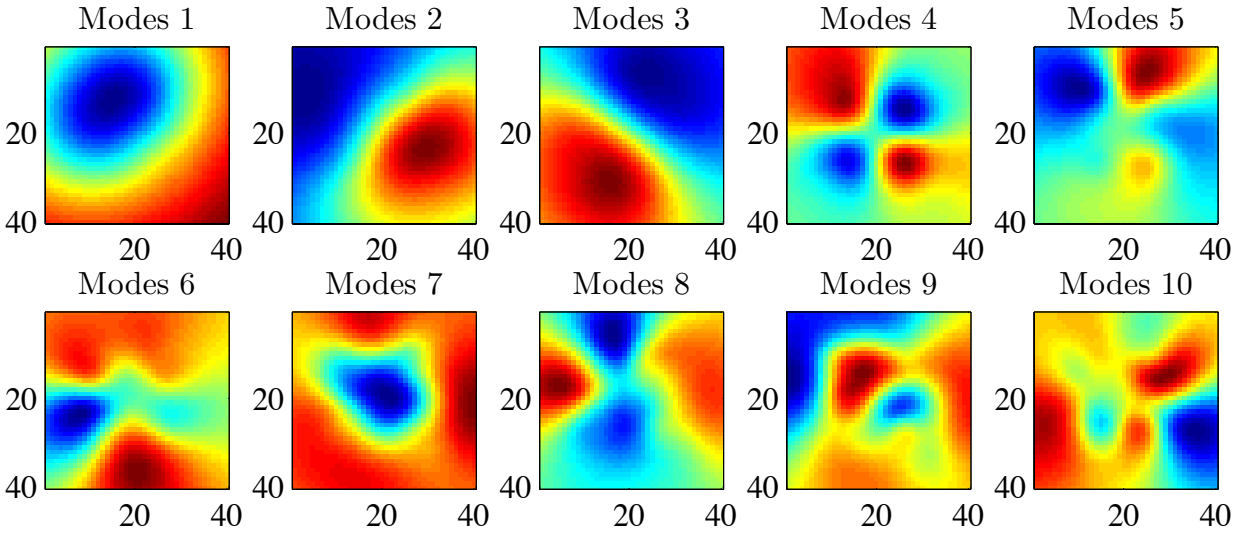




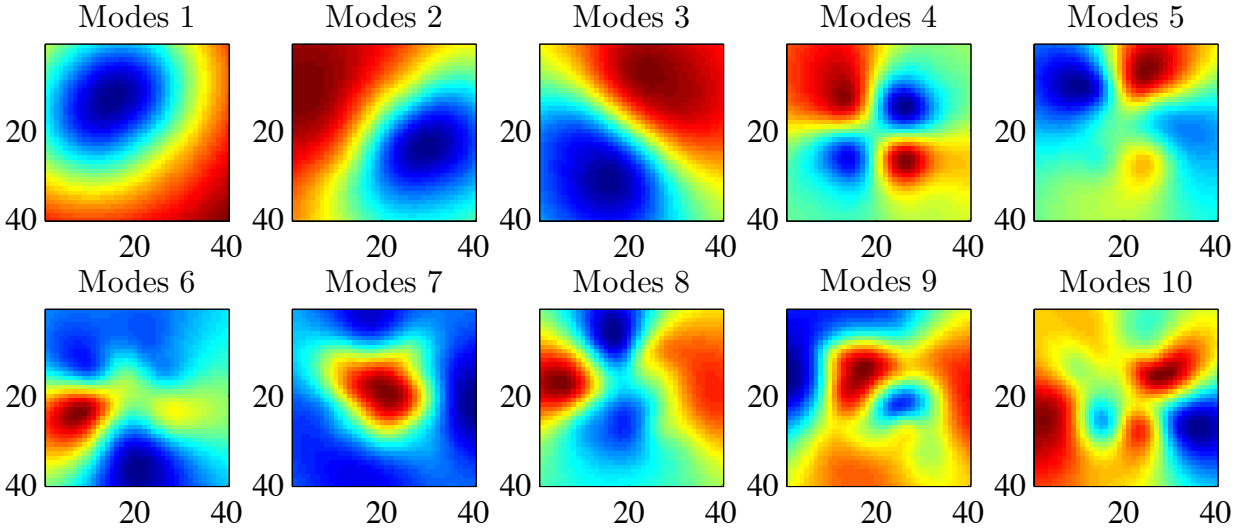


Global reduced modes

after 200 iterations

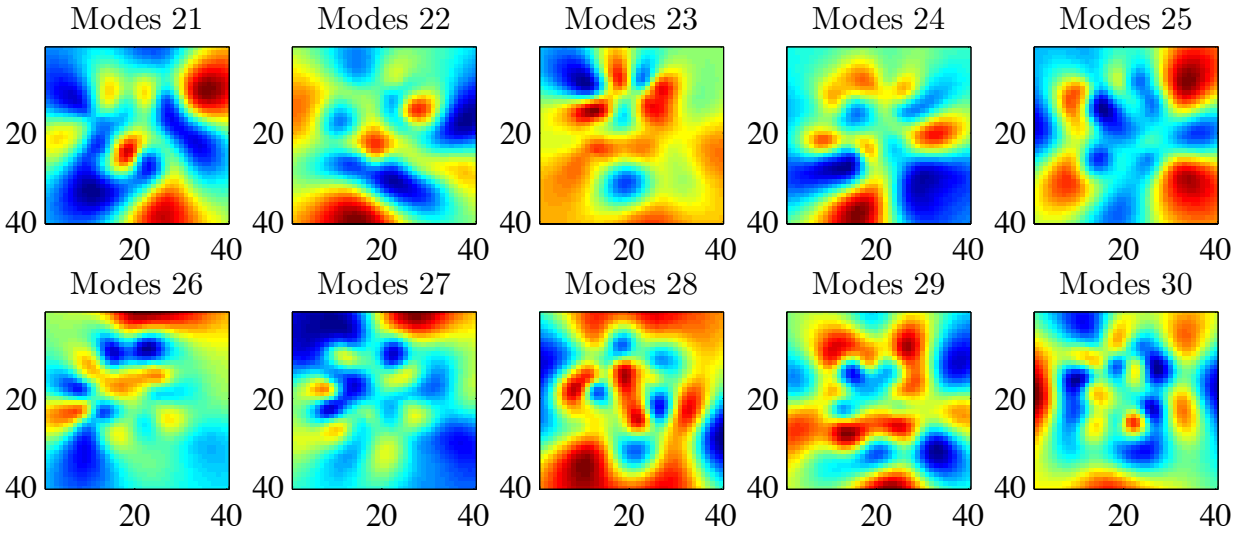


after 400 iterations

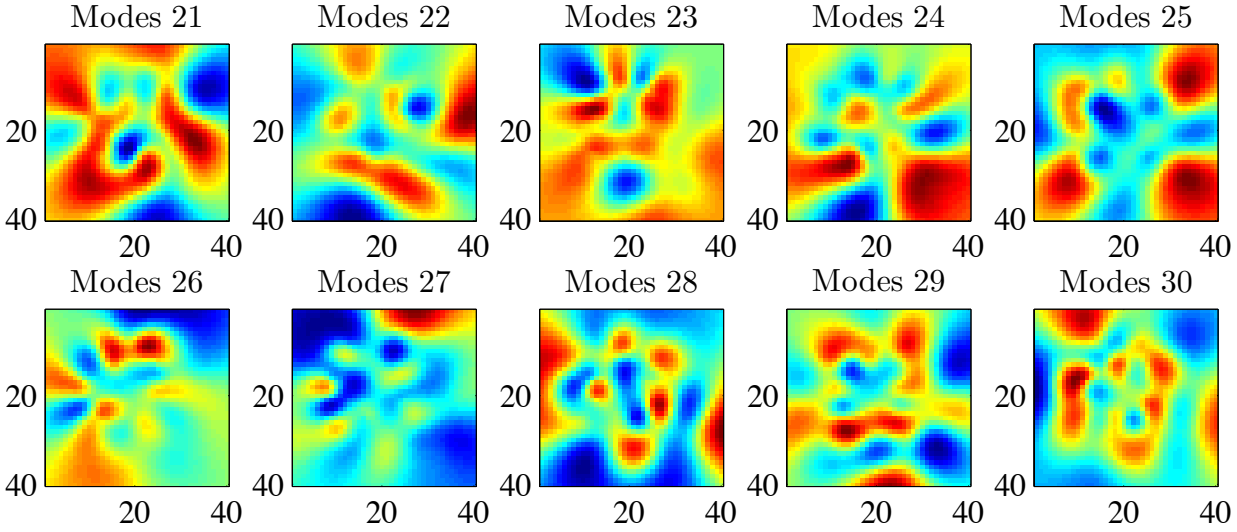


Global reduced modes

after 200 iterations



after 400 iterations

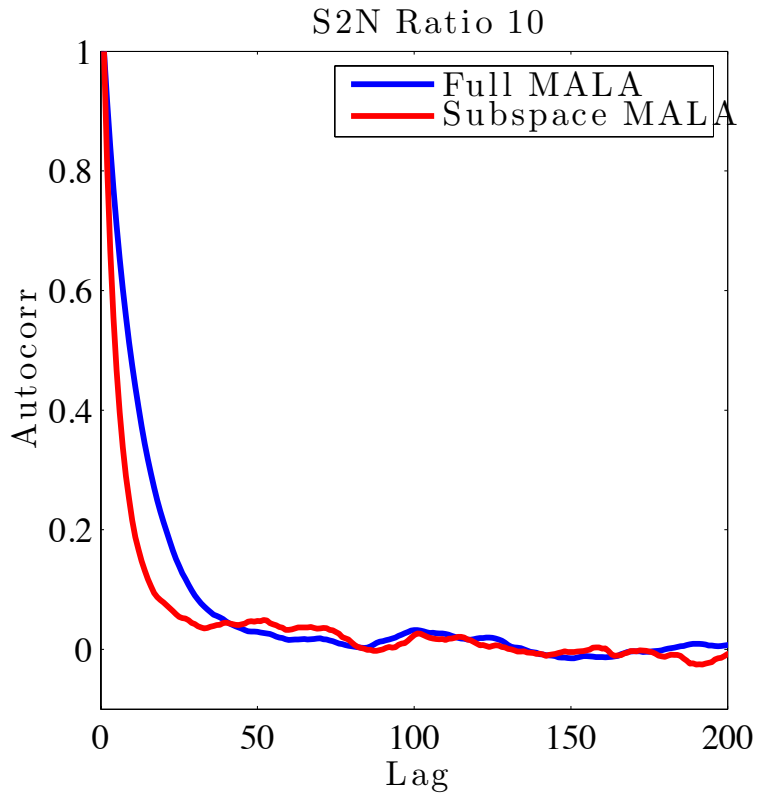


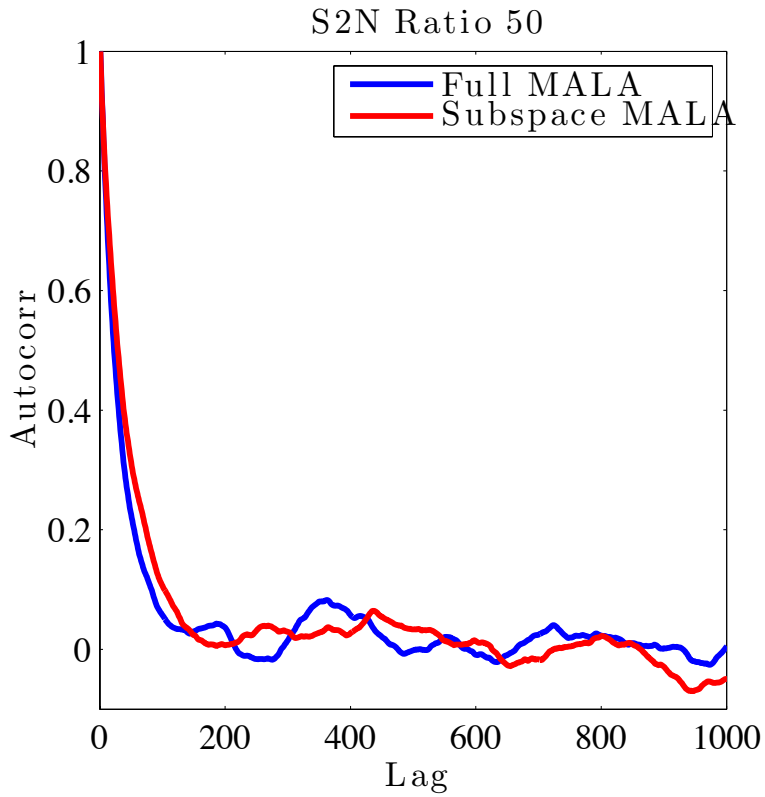
Full-dimensional problem

- Now compare with MCMC on the full (1600-dimensional) problem:
  - Use Metropolis-adjusted Langevin (MALA), preconditioned by Hessian at the MAP

- Two “performance” questions:
  1. How well does the reduced-dimension chain *mix* (versus the full-dimensional chain)?
  2. How accurately do we estimate posterior expectations?

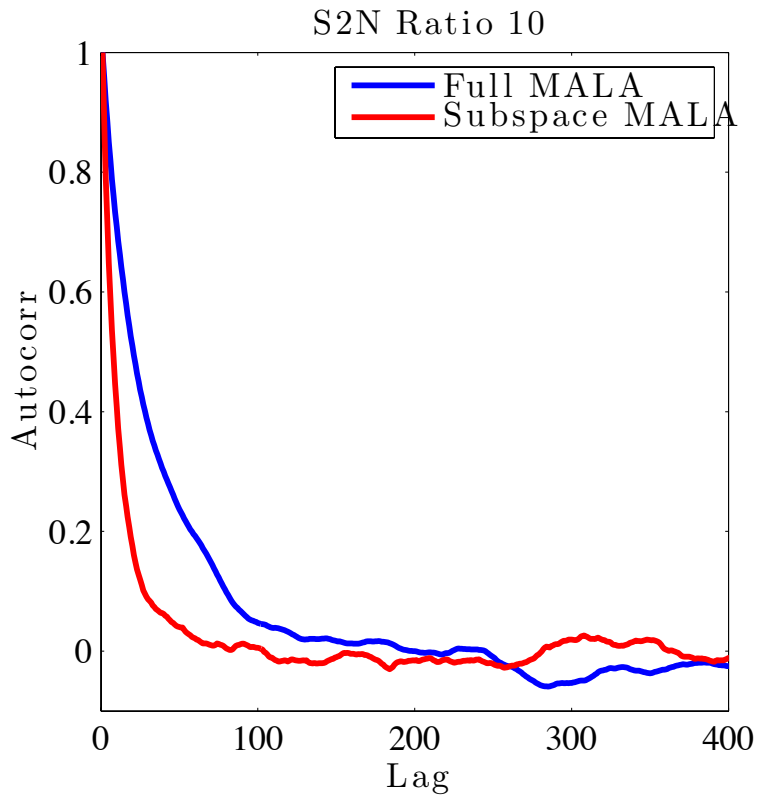
Chain mixing

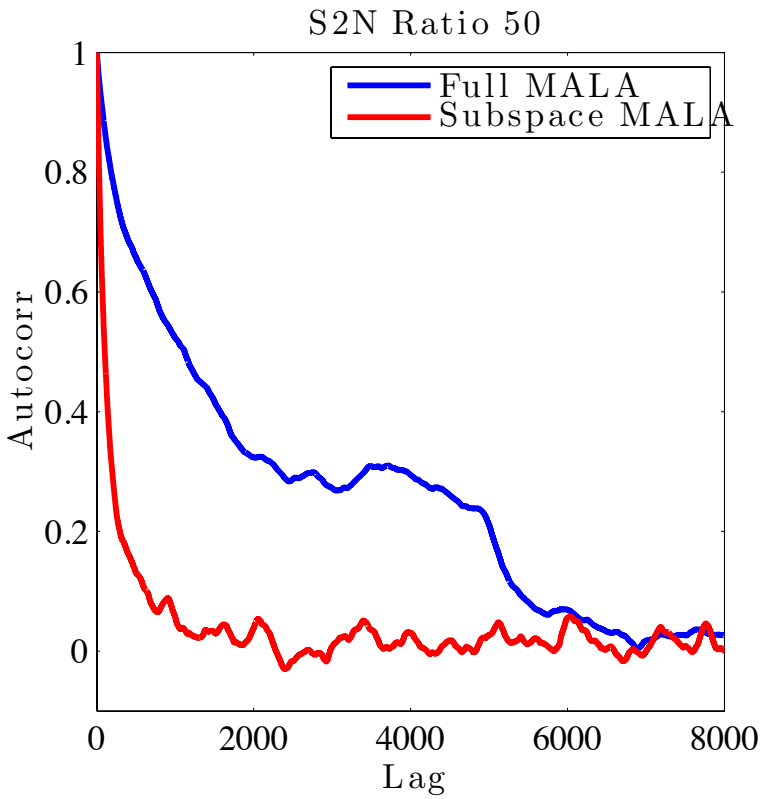




- Autocorrelation of log-posterior values

Chain mixing





- Autocorrelation of *projection onto first mode!*

Posterior estimates

- Estimating posterior expectations:
  - **Variance reduction** due to Rao-Blackwellization is key!
  - Recall law of total variance (for some MCMC estimate $\hat{h}$)

$$\mathrm{Var}[\hat{h}] = \mathrm{Var}_{\tilde{c}_r}\left[\mathbb{E}_{\tilde{c}_r^{\perp}}\left[\hat{h}(\tilde{c}_r, \tilde{c}_r^{\perp}) \mid \tilde{c}_r\right]\right] + \mathbb{E}_{\tilde{c}_r}\left[\mathrm{Var}_{\tilde{c}_r^{\perp}}\left[\hat{h}(\tilde{c}_r, \tilde{c}_r^{\perp}) \mid \tilde{c}_r\right]\right]$$

  - We can entirely eliminate the second term above!

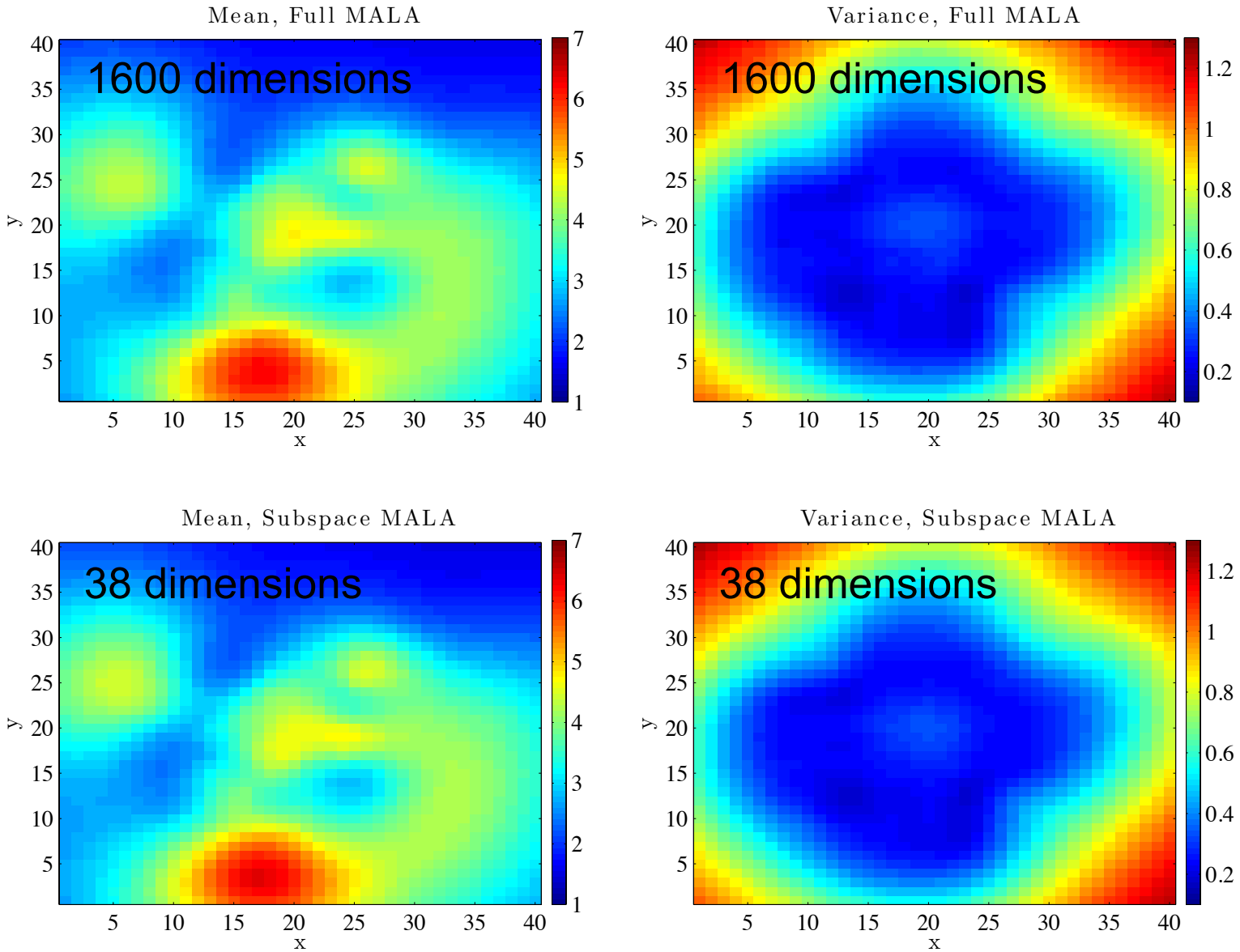
$$x = \Gamma_{pr}^{1/2} U_r c_r + \Gamma_{pr}^{1/2}(I - U_r U_r^T) z\,; \quad z \sim N(0, I)$$

from MCMC samples

everything known analytically!

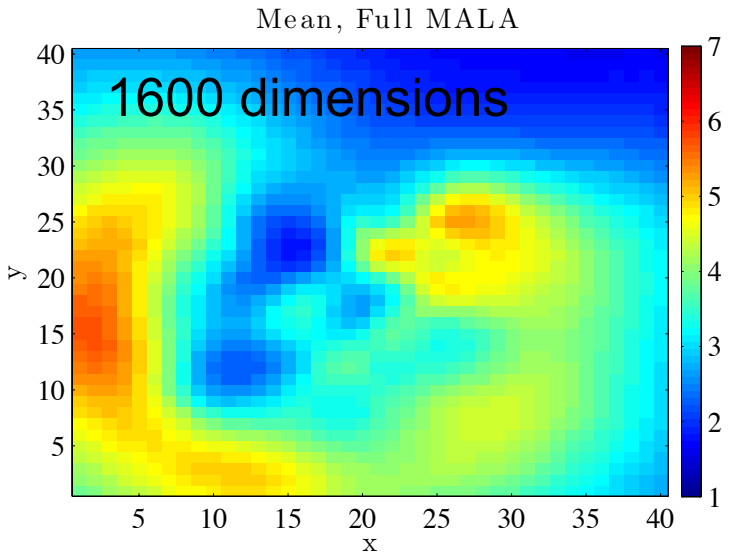
Elliptic PDE inverse problem

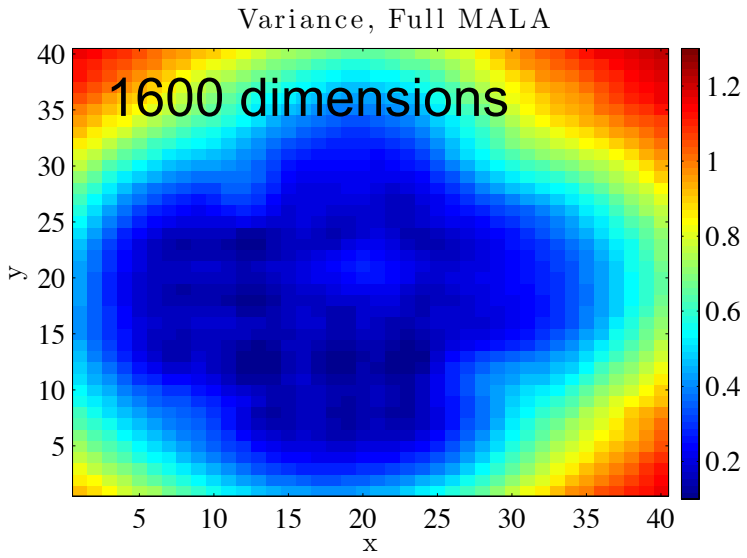
S2N ratio = 10

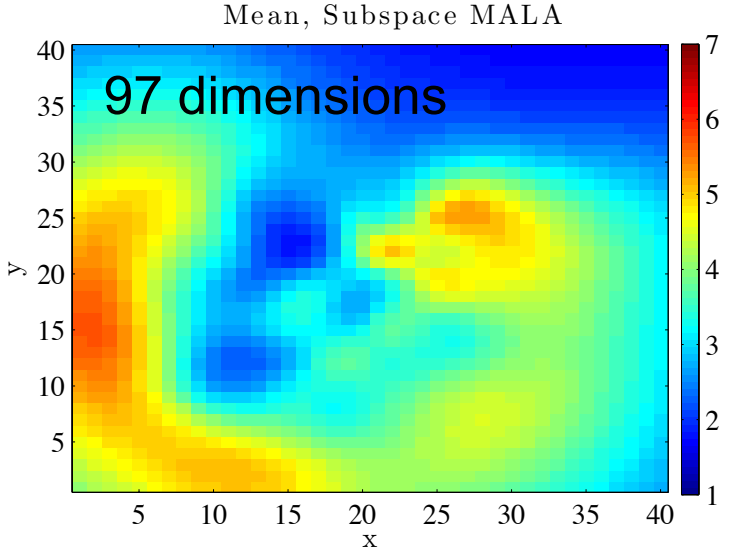


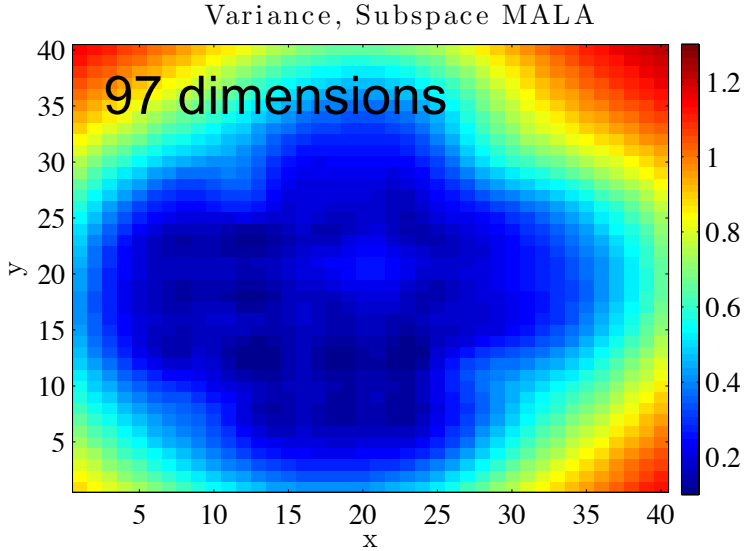
Elliptic PDE inverse problem

S2N ratio = 50









Next steps

- Formal error estimates
- Full-dimensional but *likelihood-informed* and *discretization-invariant* MCMC:
  - Need not eliminate prior-dominated directions entirely; can instead employ different proposals (e.g., Metropolis-within-Gibbs)
  - Integrate global reduced basis with discretization-invariant pCN approach of Stuart *et al.* [joint work with T. Cui and K. Law]
  - Resulting samplers are *exact* yet very efficient

Conclusions

- A broad overview of Bayesian computation for inverse problems
- **Approximations of the forward model:**
  - Stochastic spectral methods and other approaches
  - Construct approximations with respect to the prior; or instead, adaptively construct posterior-focused approximations
- **Dimensionality reduction:**
  - *Change* from prior to posterior confined to a smaller number of directions
  - This is the “intrinsic dimensionality” of the problem; shows grid-independence
  - Capture with a global basis
  - Improved MCMC mixing, plus Rao-Blackwellization in complementary directions

Acknowledgments

- **Support** from *US Department of Energy, Office of Advanced Scientific Computing Research*

Subspace construction

- Note definition of weighted subspace residual:

$$\mathrm{res} = \sum_i s_{1,i} \left\| (I - U_2 U_2^T) u_{1,i} \right\|^2 + \sum_i s_{2,i} \left\| (I - U_1 U_1^T) u_{2,i} \right\|^2$$

  - Evaluated at different lags between subspaces $(\mathrm{U}_1, \mathrm{S}_1)$ and $(\mathrm{U}_2, \mathrm{S}_2)$