Subspace acceleration for large-scale inverse problems

Tiangang Cui

School of Mathematical Sciences, Monash University

Acknowledgments

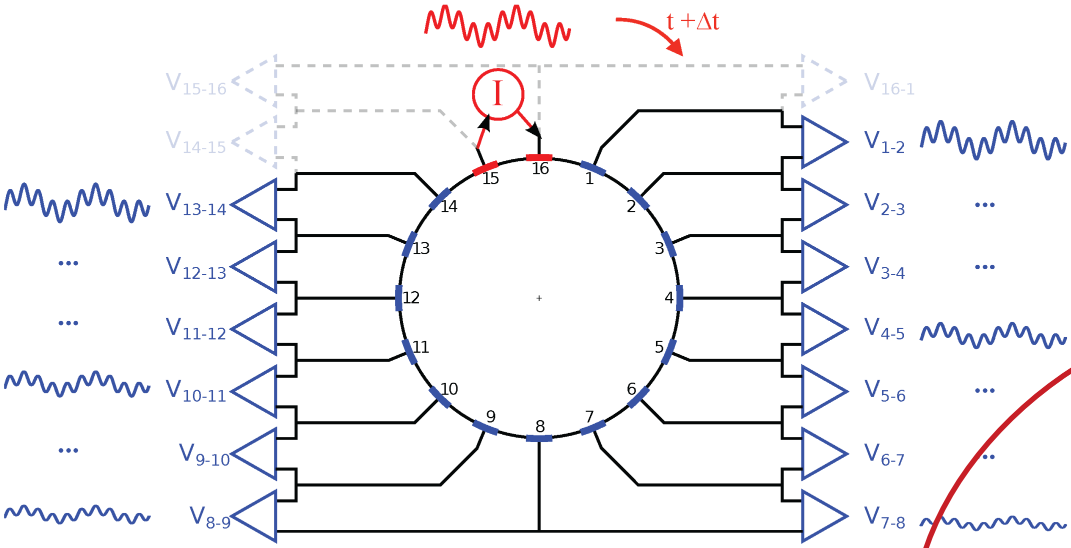
Collaborators:

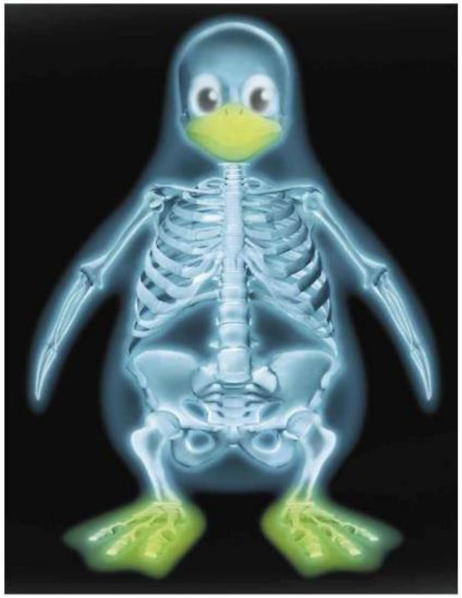
- Kody Law (University of Manchester), Youssef Marzouk (MIT), Alessio Spantini (MIT), Karen Willcox (MIT), Olivier Zahm (INRIA)

- Omar Ghattas (UT-Austin), James Martin (UT-Austin), Benjamin Peherstorfer (MIT), Noémi Petra (UC Merced), Marko Laine (FMI), Zheng Wang (MIT), Luis Tenorio (Mines)

- **Support** from DOE ASCR applied mathematics program, DiaMonD center, Monash startup grant (Math), Robert Bartnik Fellowship (Monash)

Inverse problems





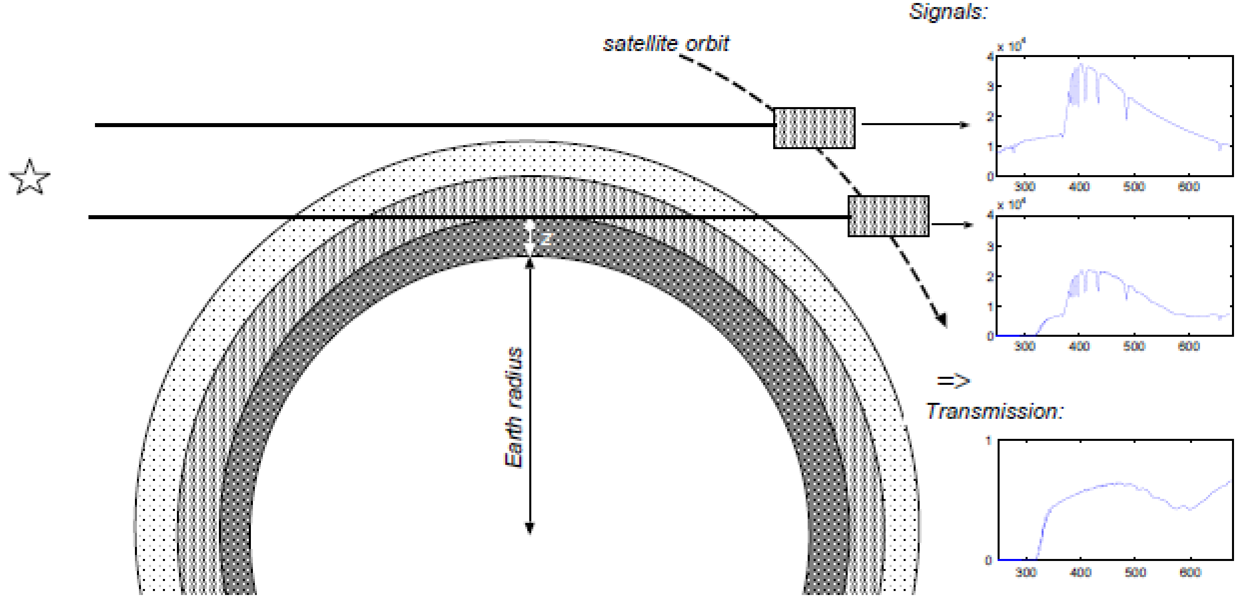
$$y = F(x) + \epsilon$$

forward model (PDE)

observation/model errors

- Data *y* are typically limited, noisy, and indirect (ill-posedness)
- Forward model *F* may be computationally intensive (high- dimensional states)
- Parameter *x* is high-dimensional (in principle, infinite-dimensional)

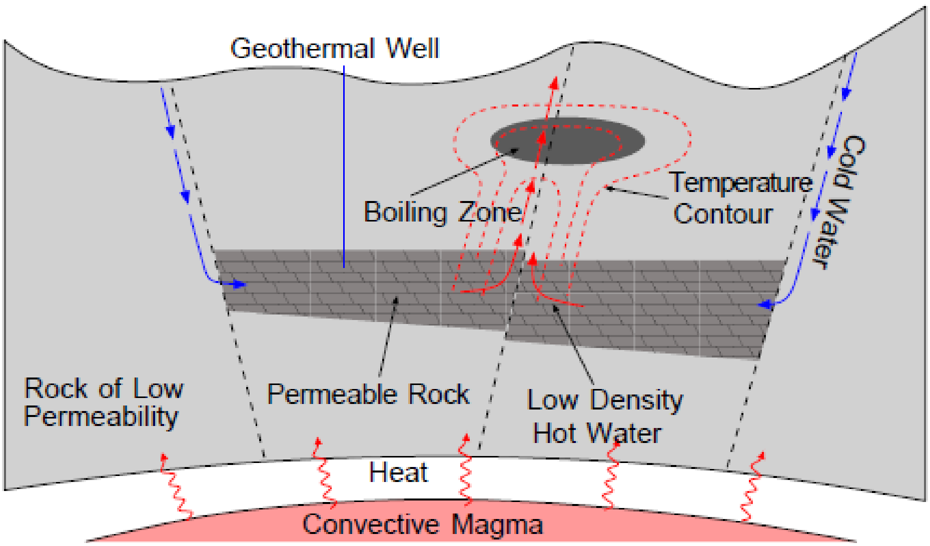
Inverse problems: remote sensing



$$\text{Beer's law} \quad T_{\lambda,l} = \exp\left(- \int_l \sum_{\text{gas}} \alpha_\lambda^{\text{gas}}(h) \rho^{\text{gas}}(h) dh \right)$$

- Monitoring air pollution
- Estimating gas density profiles from satellite data

Inverse problems: subsurface



$$\frac{d}{dt}\int_{\Omega} M_{(\cdot)}\, dV = \int_{\partial\Omega} Q_{(\cdot)}\cdot\hat{n}\, d\Gamma + \int_{\Omega} q_{(\cdot)}\, dV$$

$$M_{\mathrm{m}} = \phi(\rho_{\mathrm{l}} S_{\mathrm{l}} + \rho_{\mathrm{v}} S_{\mathrm{v}})$$

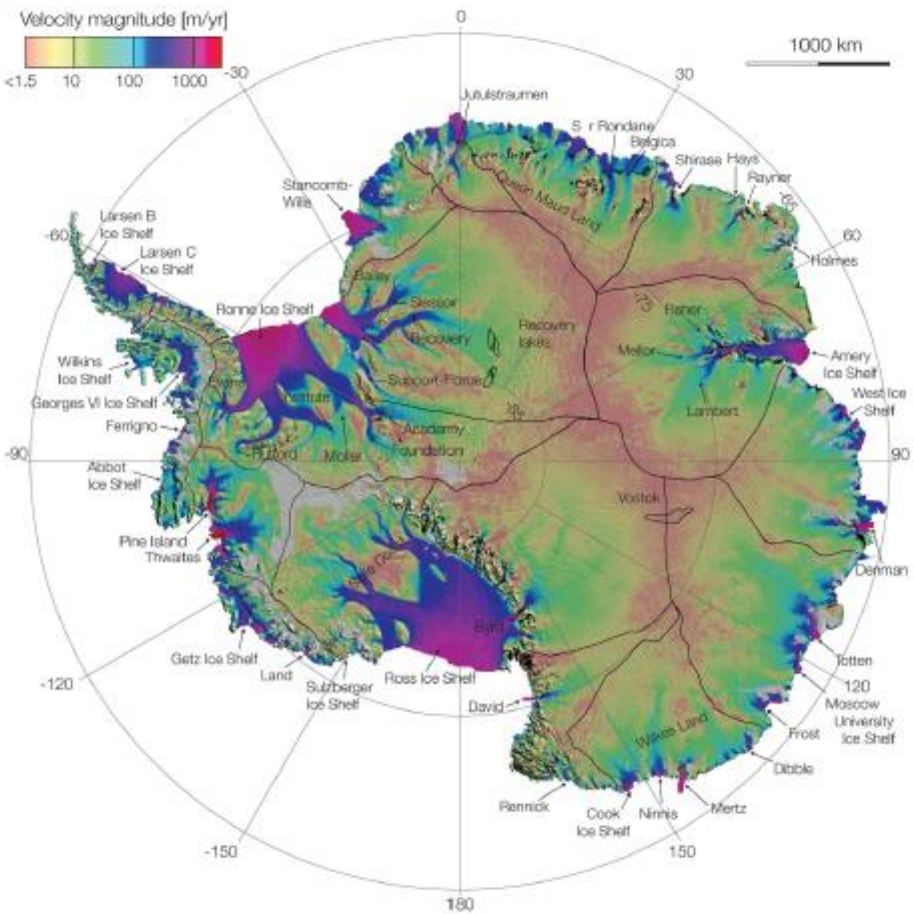
$$Q_{\mathrm{m}} = \sum_{\beta=\mathrm{l},\mathrm{v}} \frac{k k_{\mathrm{r}\beta}}{\nu_\beta}(\nabla p - \rho_\beta \hat{\mathrm{g}})$$

$$M_{\mathrm{e}} = (1-\phi)\rho_{\mathrm{r}} c_{\mathrm{r}} T + \phi(\rho_{\mathrm{l}} u_{\mathrm{l}} S_{\mathrm{l}} + \rho_{\mathrm{v}} u_{\mathrm{v}} S_{\mathrm{v}})$$

$$Q_{\mathrm{e}} = \sum_{\beta=\mathrm{l},\mathrm{v}} \frac{k k_{\mathrm{r}\beta}}{\nu_\beta}(\nabla p - \rho_\beta \hat{\mathrm{g}}) h_\beta - K \nabla T$$

- Model and predict fluid end energy transport in subsurface
- Estimating heterogenous rock properties (porosity, permeability, relative perm.) and boundary conditions from well observations

Inverse problems: polar ice sheet



$$-\nabla \cdot [2\eta(\boldsymbol{u})\,\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} - \boldsymbol{I}p] = \rho \boldsymbol{g} \quad \text{in } \Omega$$

$$\nabla \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega$$

$$\boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n} = \boldsymbol{0} \quad \text{on } \Gamma_t$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_b$$

$$\boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n} + \exp(x)\boldsymbol{T}\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_b$$

where

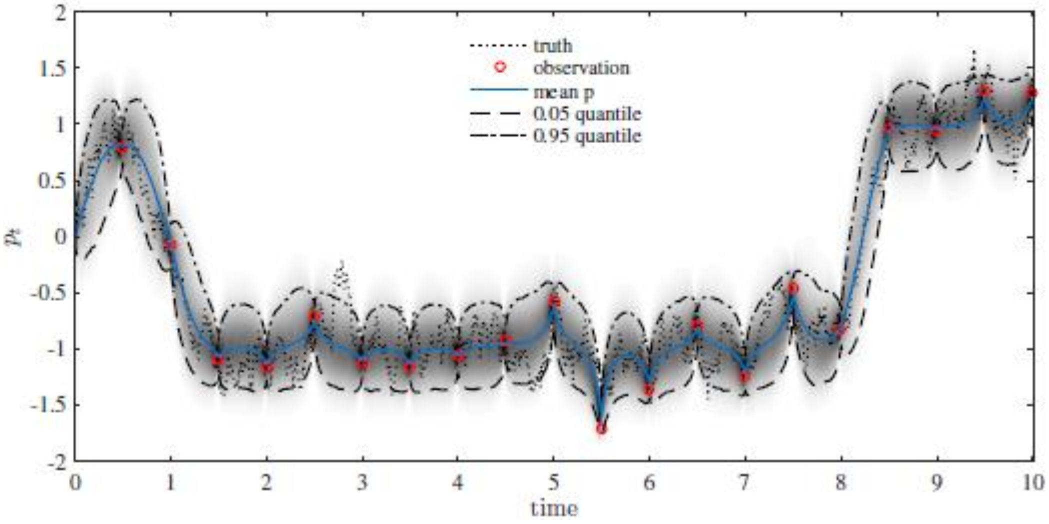
$$\boldsymbol{\sigma}_{\boldsymbol{u}} = -\boldsymbol{I}p + 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}$$

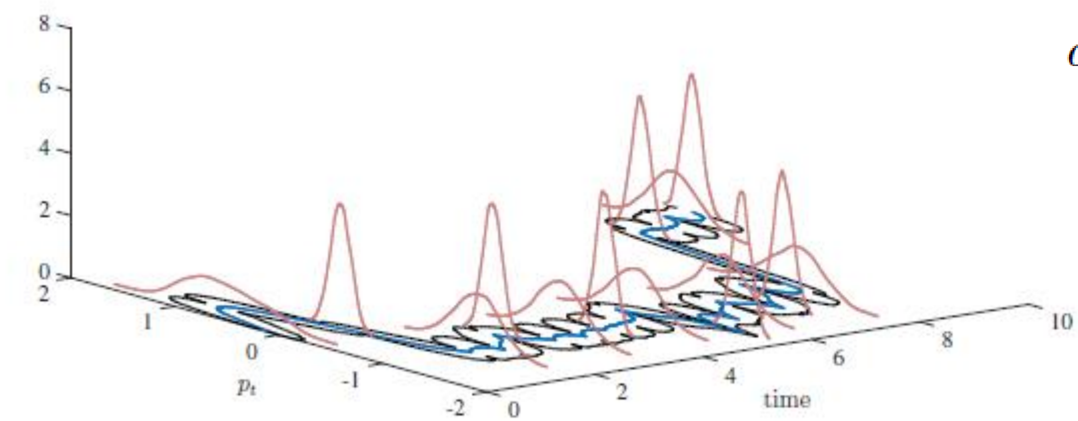
$$\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top)$$

$$\eta(\boldsymbol{u}) = \frac{1}{2} A^{-\frac{1}{n}} \left(\frac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)\right)^{\frac{1-n}{2n}}$$

- Model ice sheet movements for predicting the amount of ice loss
- Estimating bottom boundary conditions from velocity measured at top surface

Inverse problems: SDE



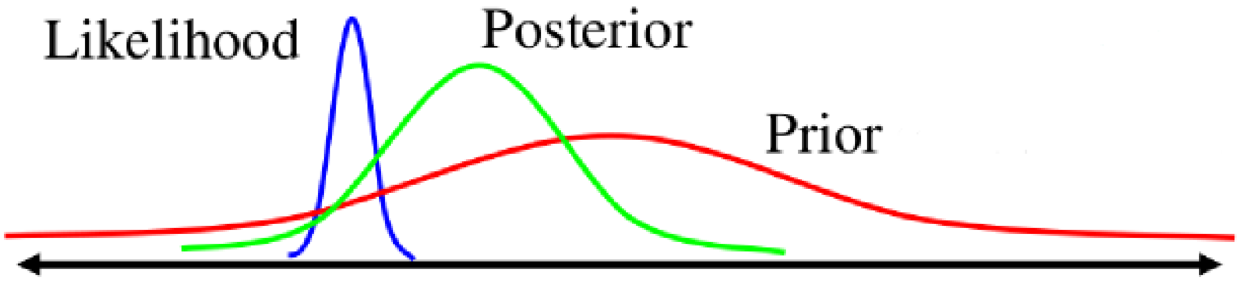


$$dp(t) = f(p(t), \alpha)dt + \sigma(p(t), \beta)du_t$$
$$p(0) = p_0$$
$$y(t) = h(p(t))$$

Estimating parameters and random forcing of SDE to characterize financial and biological systems

UQ for inverse problems

- We adopt a **Bayesian** approach: treat model parameter as random variables



- Posterior distribution: $\pi_{\text{pos}}(x \mid y) \propto \pi_{\text{like}}(y \mid x)\,\pi_0(x)$

- Posterior measure (infinite dimensional):

$$\frac{d\mu_y}{d\mu_0}(x) \propto L(y|x), \qquad \mu_0 = \mathcal{N}(m_0, \Gamma_{\text{pr}})$$

where Γ_{pr} is a trace-class operator.

Computational challenges

- Characterize the Bayesian solution by evaluating **posterior expectations:**

$$\mathbb{E}[h] = \int h(x)\pi_{\text{pos}}(dx \mid y)$$

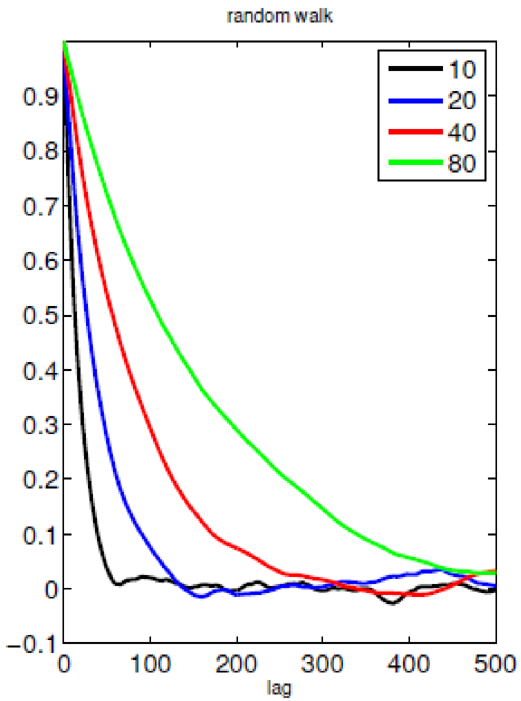
- Explore the posterior distribution:

  Markov chain Monte Carlo (MCMC), quasi Monte Carlo (QMC), sequential Monte Carlo (SMC), importance sampling, sparse quadrature, measure transport

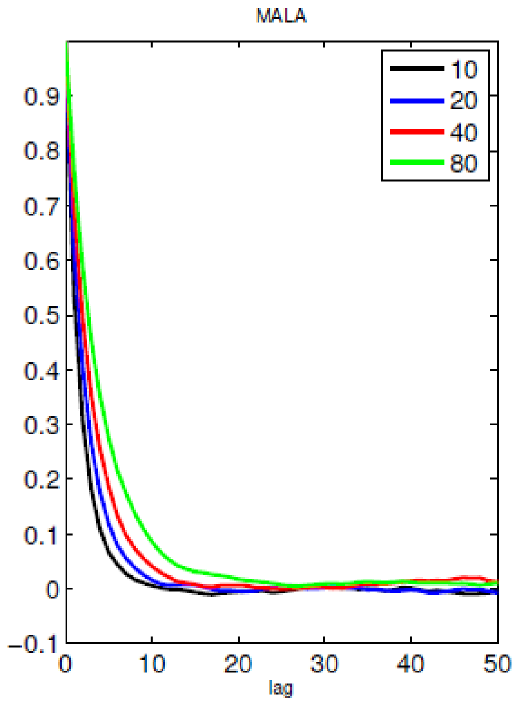
- Parameter dimension scalability is one of the challenges.

Computational challenges

- Parameter dimension scalability (MCMC).



Random walk $O(N_x)$

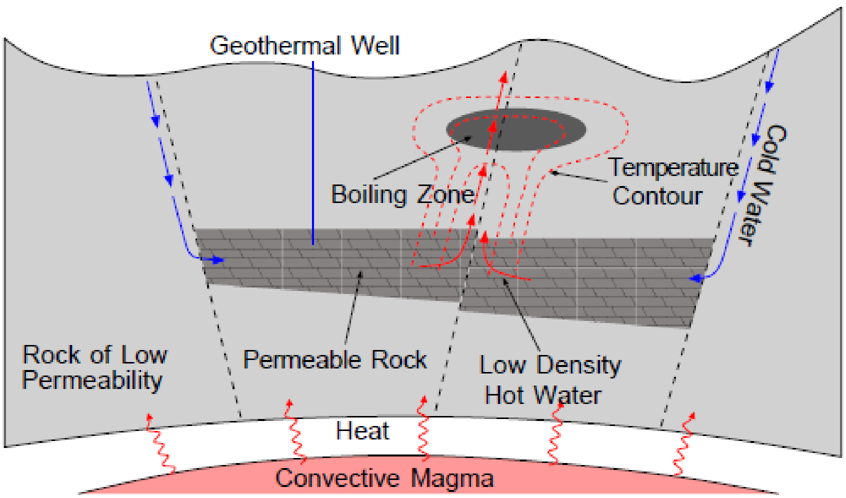


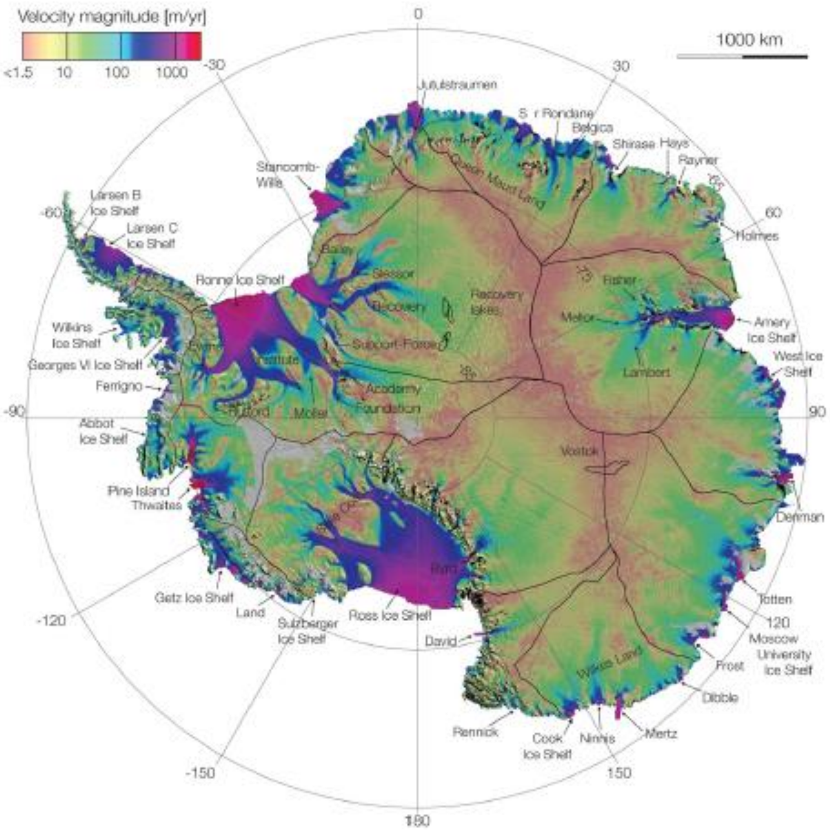
MALA $O(N_X^{\frac{1}{3}})$

- Autocorrelation of Markov chain in various dimensions
- Standard MCMC (or other sampling methods) can scale badly.

Computational challenges

- Simulating the posterior density (involves a forward model) can be very costly





- We typically need thousands or even millions density evaluations.

Computational challenges

Key strategies for making posterior exploration computationally tractable:

- Efficient and structure-exploiting **sampling schemes** to explore the posterior distribution
  - Markov chain Monte Carlo (MCMC), quasi Monte Carlo (QMC), sequential Monte Carlo (SMC), importance sampling, sparse quadrature, measure transport

- ***Approximations*** of the forward model or likelihood:
  - Sparse polynomial expansions, local interpolants, Gaussian process emulators, reduced order models, multi-fidelity approaches

1. Parameter reduction

- Underlying idea: the posterior can be approximated by

$$\pi^a_{\text{pos}}(x \mid y) \propto \pi_{\text{like}}(y \mid \mathcal{P}_r x) \ \ \pi_0(x)$$

for some rank-r projector $\mathcal{P}_r$

- This induces a decomposition of the parameter space

$$x = x_r + x_\perp \begin{cases} x_r \in \mathrm{Im}(\mathcal{P}_r) \\ x_\perp \in \mathrm{Ker}(\mathcal{P}_r) \end{cases}$$

- This decomposition can be used for
  - Design dimension independent MCMC [Cui et al. 2014 & 2016], SMC [Law et al. 2017]
  - Build reduced order models [Cui et al. 2016]
  - Fast solver for linear inverse problems [Spantini et al. 2015]
  - Gaussian approximations of the posterior [Flath *et al.* 2011]
  - Perform QMC, sparse grid quadrature …

1. Parameter reduction

- How to obtain the projector $\mathcal{P}_r$?

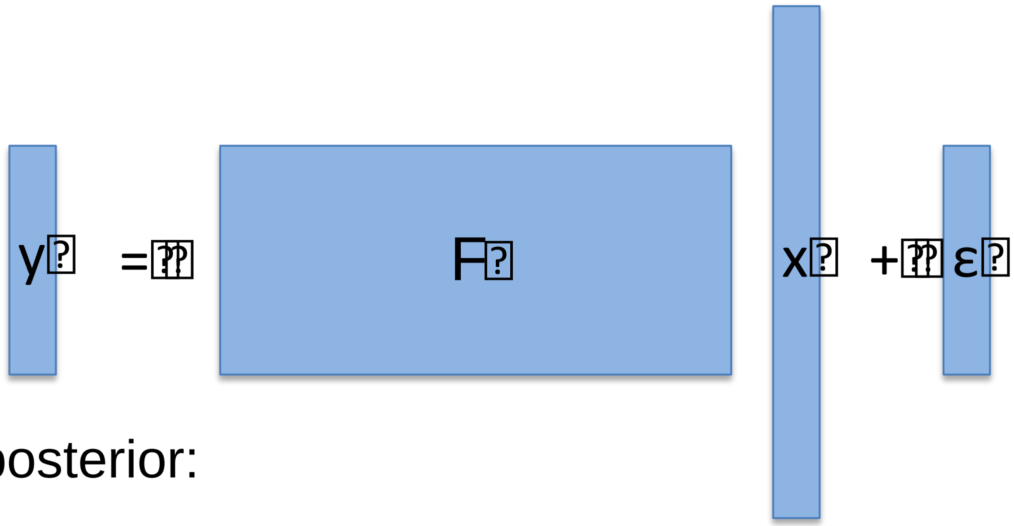
- What controls the intrinsic *parameter dimension* of a Bayesian inverse problem?
  - Structure of the prior (e.g., smoothness and correlation)
  - How many observations are used, and how much the forward model smooths the parameters (i.e., loss of information or ill-posedness of the classical inverse problem)

- **Key idea:** low parameter dimensionality lies in the **change** from prior to posterior…

Linear Bayesian problem

- Simple linear-Gaussian model:

$$y = Fx + \epsilon, \;\; x \sim N(0, \Gamma_{\mathrm{pr}}), \;\; \epsilon \sim N(0, \Gamma_{\mathrm{obs}})$$



- Gaussian posterior:

$$\pi_{\mathrm{pos}}(x \mid y) = N(m_{\mathrm{pos}}, \Gamma_{\mathrm{pos}}) \propto \exp\left(-\frac{1}{2}\left\|\Gamma_{\mathrm{obs}}^{-1/2}(y - Fx)\right\|^2 - \frac{1}{2}\left\|\Gamma_{\mathrm{pr}}^{-1/2}x\right\|^2\right)$$

- Posterior covariance: $\Gamma_{\mathrm{pos}}^{-1} = \Gamma_{\mathrm{pr}}^{-1} + H$
- Data-misfit Hessian: $H \equiv F^{\top}\Gamma_{\mathrm{obs}}^{-1}F$

Linear Bayesian problem

- Consider the generalized Rayleigh quotient:

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

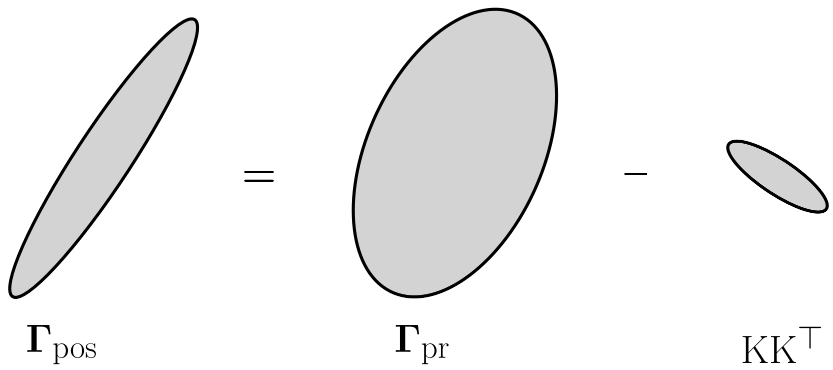
  - When quotient is large, likelihood limits variability in the *w* direction more strongly than the prior
  - When quotient is small, prior is more constraining in the *w* direction

- We seek **data-informed** directions in a *relative* sense…

Linear Bayesian problem

- Consider **approximations** of the form

$$\mathcal{M}_r \equiv \left\{ \hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top \succ 0, \;\; \text{rank}(K) \leq r \right\}$$



- Dominant eigenvectors of $HW_r = \Gamma_{\text{pr}}^{-1} W_r \Lambda_r$ yield **optimal** approximations of the posterior covariance [Spantini *et al.* 2015]
  - Optimality in Riemannian metric on `Sym+`, rather than Frobenius
  - Reinforces importance of **prior-preconditioned Hessian** [Flath *et al.* 2011, Martin *et al.* 2012]
  - Optimality extends to Kullback-Leibler and Hellinger losses

Projection & nonlinear problems

- We can also interpret the optimal low-rank update as the result of an oblique ***projector***

$$\hat{\Gamma}_{\text{pos}}^{-1} = \Gamma_{\text{pr}}^{-1} + \mathcal{P}_r^\top F^\top \Gamma_{\text{obs}}^{-1} F \mathcal{P}_r$$

  – Replace forward model *F* with $F\mathcal{P}_r$, where $\mathcal{P}_r = W_r W_r^\top \Gamma_{\text{pr}}^{-1}$

- Can we extend this idea to ***nonlinear*** inverse problems? — We want to find a matrix H that contains "relative information" of the likelihood.

Projection & nonlinear problems

Some existing approaches:

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

$$H_{\mathrm{LIS}} = \int \nabla F(x)^\top \Gamma_{\mathrm{obs}}^{-1} \nabla F(x) \pi_{\mathrm{pos}}(dx \mid y)$$

- Active subspace (AS) [Constantine et al. 2016]

$$H_{AS} = \int \nabla \log \pi_{\mathrm{like}}(y \mid x) \otimes \nabla \log \pi_{\mathrm{like}}(y \mid x)\, \pi_0(dx)$$

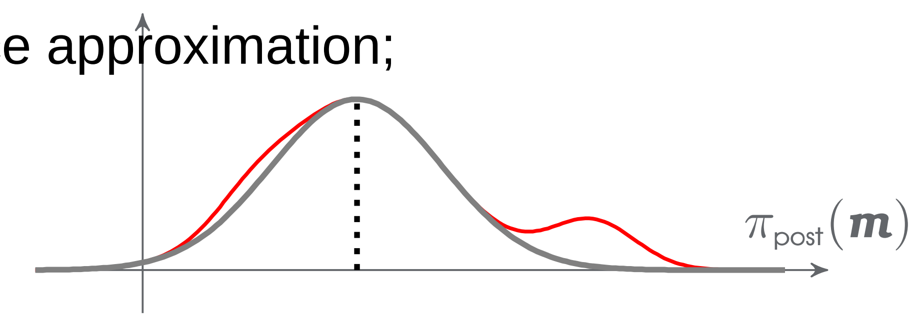
- In general, we are seeking for

$$H = \int G(x)\, \pi_{\mathrm{ref}}(dx)$$

$G(x)$ can be the Fisher information, $\nabla \log \pi_{\mathrm{like}} \otimes \nabla \log \pi_{\mathrm{like}}$, or $x \otimes x$

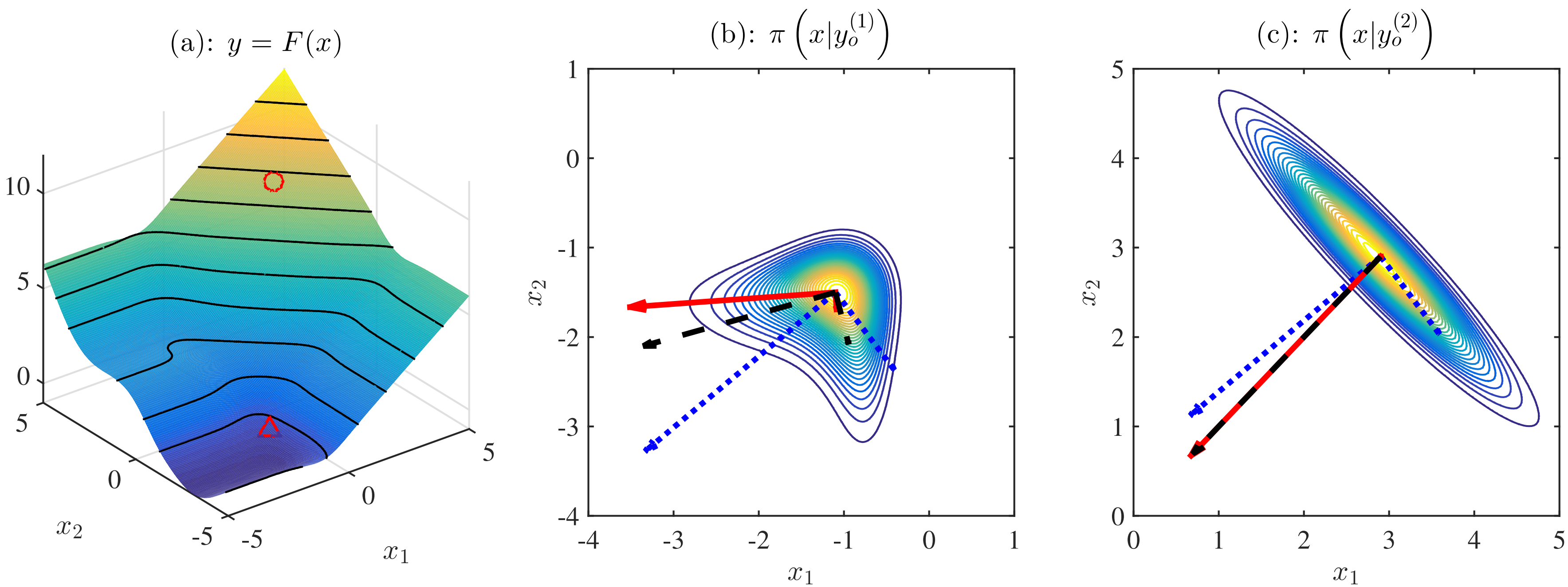
π_{ref} can be prior, posterior, or Laplace approximation;

– Note: $G(x) = x \otimes x$ and $\pi_{\mathrm{ref}} = \pi_0$ recover KL expansion of prior



A 2D example

- Nonlinear forward model $F : \mathbb{R}^2 \to \mathbb{R}$, standard Gaussian prior, Gaussian observation error



- Posteriors from two different observations (triangle = *middle* and circle = *left*)
- Param subspaces with different reference distributions: red = posterior, black = Laplace, blue = prior

Projection & nonlinear problems

- Can we formalize this intuition? — Control the Kullback-Leibler divergence from the approximated posterior to the posterior

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}}(x \mid y) \,\|\, \pi^{\mathrm{a}}_{\mathrm{pos}}(x \mid y)\right) \leq \varepsilon$$

- Given a projector $\mathcal{P}_r$, what is the best $\pi^*_{\mathrm{pos}}(x \mid y) \propto \pi^*_{\mathrm{like}}(y \mid \mathcal{P}_r x)\pi_0(x)$? — Conditional expectation of the likelihood over $\mathrm{Ker}(\mathcal{P}_r)$ w.r.t. the prior

$$\pi^*_{\mathrm{like}}(y \mid \mathcal{P}_r x) = \int \pi_{\mathrm{like}}\left(y \mid \mathcal{P}_r x + (I - \mathcal{P}_r)x\right)\pi_0\left((I - \mathcal{P}_r)dx \mid \mathcal{P}_r x\right)$$

- Then, we apply a Pythagorean theorem on KL divergence

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \,\|\, \pi^a_{\mathrm{pos}}\right) = \underbrace{D_{\mathrm{KL}}\left(\pi_{\mathrm{pos}} \,\|\, \pi^*_{\mathrm{pos}}\right)}_{\text{function of } \mathcal{P}_r} + \underbrace{D_{\mathrm{KL}}\left(\pi^*_{\mathrm{pos}} \,\|\, \pi^a_{\mathrm{pos}}\right)}_{\text{function of } \mathcal{P}_r \text{ and other terms}}$$

Projection & nonlinear problems

- Computing $H = \int \nabla \log \pi_{\text{like}} \otimes \nabla \log \pi_{\text{like}} \; \pi_{\text{pos}}(dx \mid y)$, or H_{LIS} and H_{AS}, in general is subjected to sampling error.

- For an arbitrary $\mathcal{P}_r$, the error bound can be expressed as the PCA recovery error

$$D_{\text{KL}}\left(\pi_{\text{pos}} \,\|\, \pi^*_{\text{pos}}\right) \leq \left\| \Gamma_{\text{pr}}^{-1/2} \left(I - \mathcal{P}_r\right) H \right\|_F^2$$

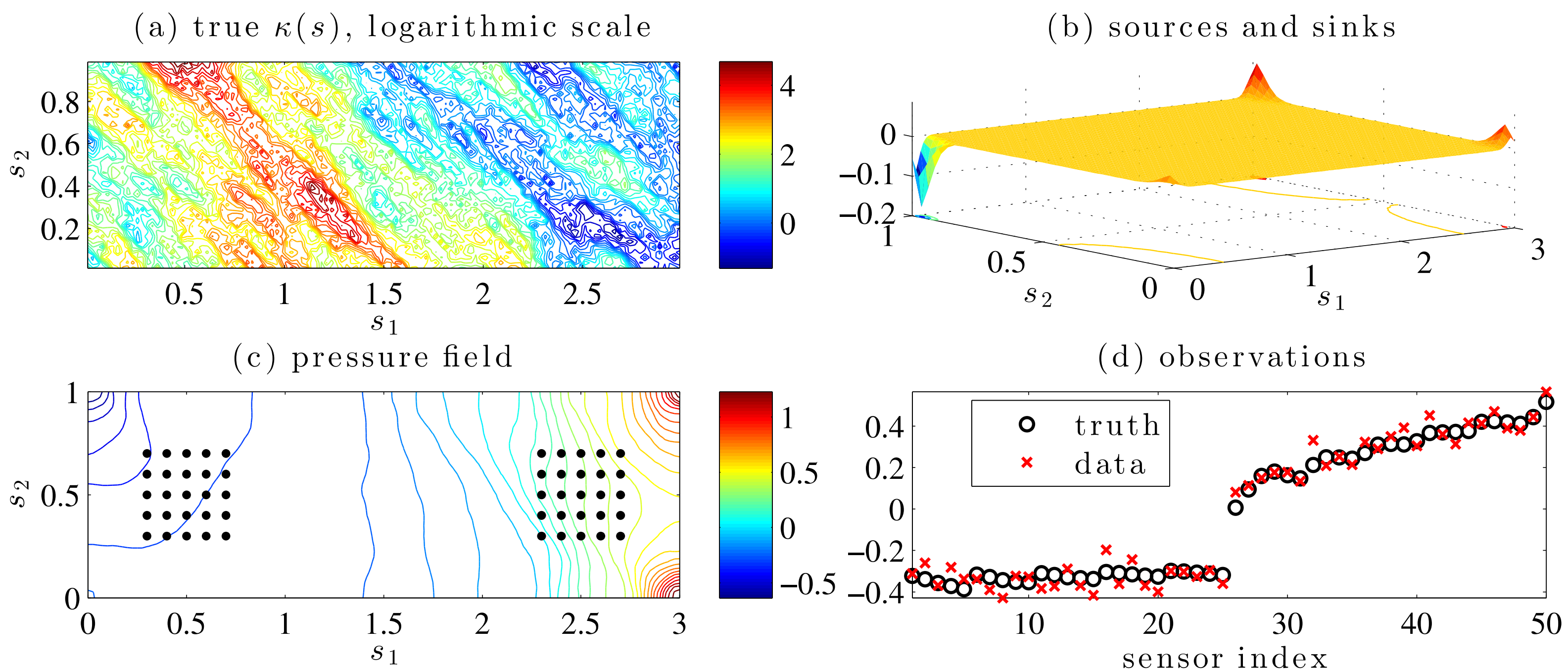
- This provides a way to compare various parameter reduction methods.

Example: elliptic PDE

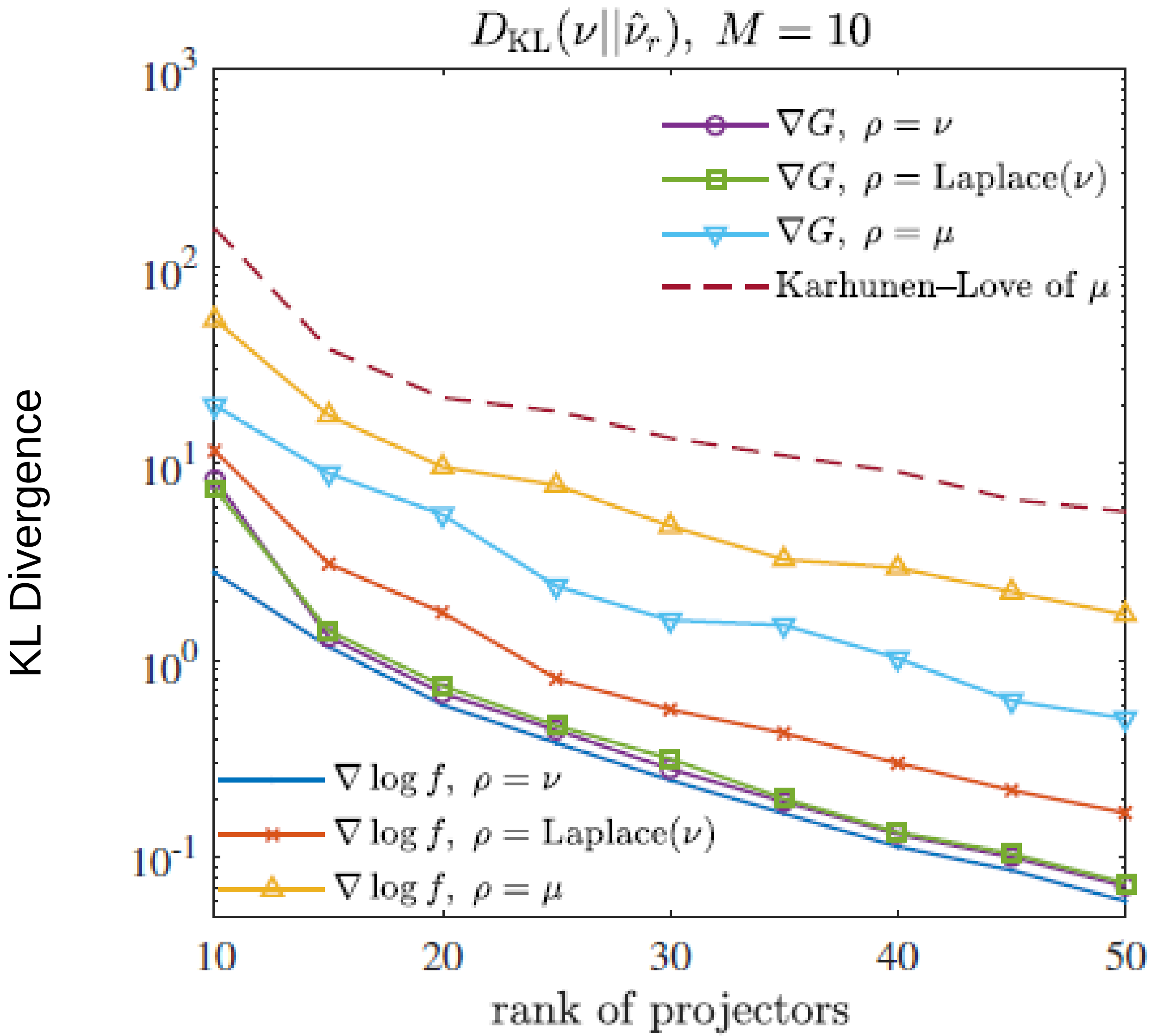
- Elliptic PDE in two spatial dimensions

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

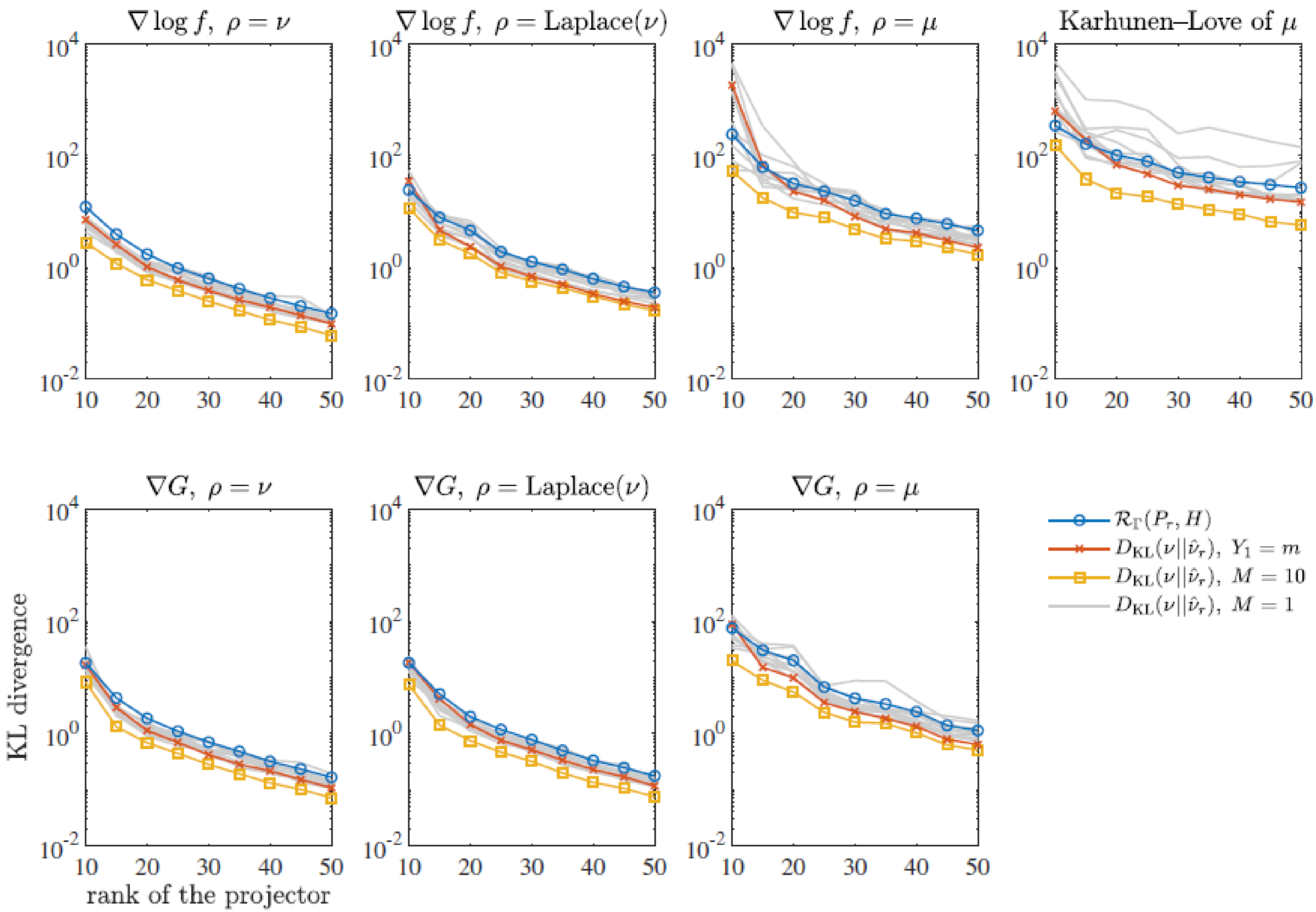
  - Estimate *κ* from noisy observations of *u*
  - Log-normal prior on *κ(s)* with an *anisotropic* **exponential** covariance kernel $\log \kappa \sim N\left(0, C_0\right)$



Example: elliptic PDE



Example: elliptic PDE



Posterior decomposition

- Introducing projector into likelihood yields an *approximate* posterior:

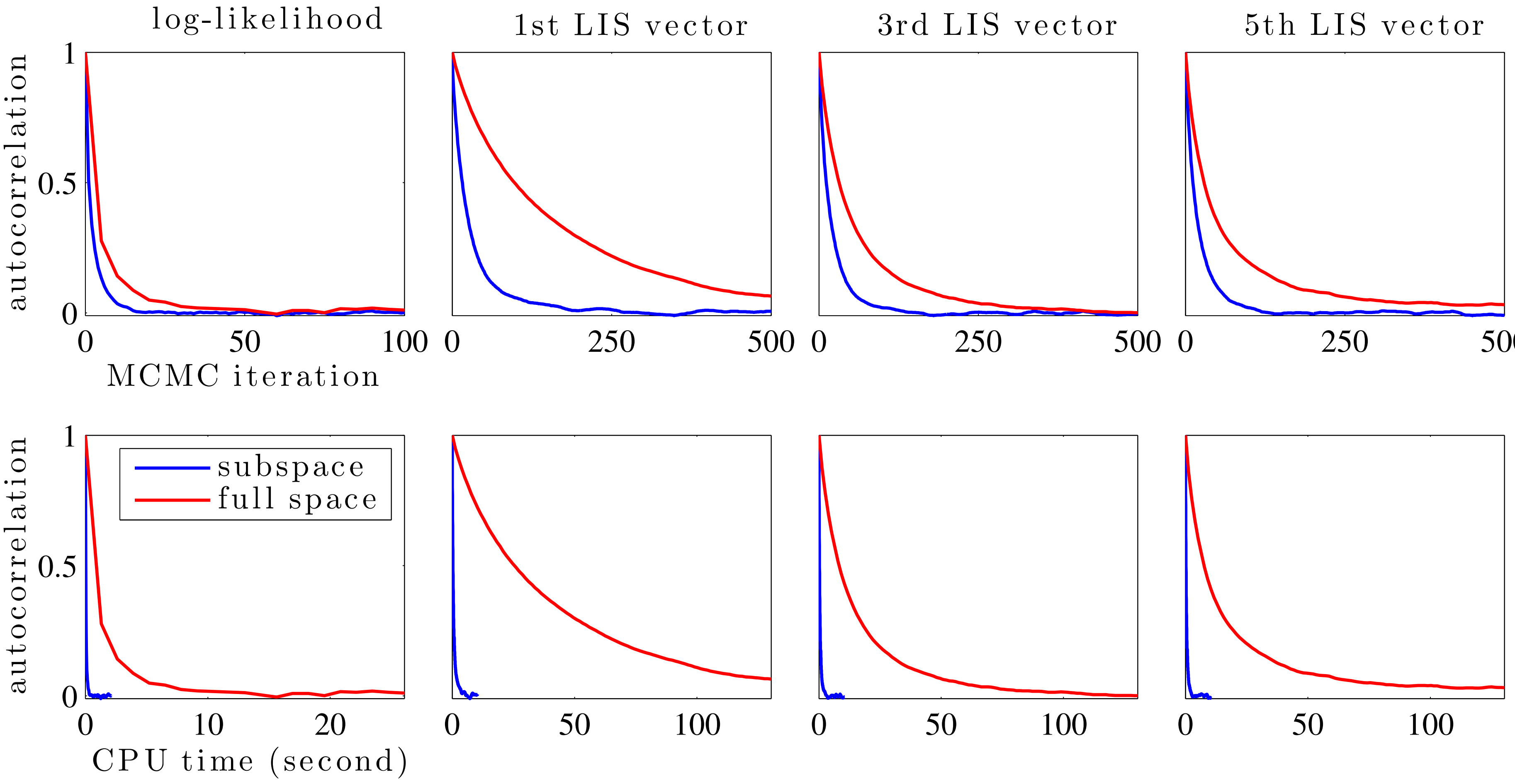
$$\pi_{\text{pos}}(x \mid y) \propto \pi_{\text{like}}\big(y, F(x)\big) \; \pi_0(x)$$
$$\approx \pi_{\text{like}}\big(y, F(\bar{W}_r \bar{W}_r^\top \Gamma_{\text{pr}}^{-1} x)\big) \; \pi_0(x)$$

- By construction, $\bar{W}_r$ diagonalizes the prior precision
- **Decompose** the approximate posterior into a data-conditioned distribution on the LIPS, prior on its complement

$$\pi_{\text{pos}}^a(x_r, x_\perp \mid y) \propto \pi_{\text{like}}(y, F(\bar{W}_r x_r))\pi_0(x_r) \; \pi_0(x_\perp)$$
$$= \pi_{\text{pos}}(x_r \mid y) \; \pi_0(x_\perp)$$

- Explore low-dimensional posterior $\pi_{\text{pos}}(x_r \mid y)$ via sampling
- No need to sample $\pi_0(x_\perp)$; known analytically

MCMC efficiency



- AMALA in subspace (*20 dim*) versus HMALA in full space (*4800 dim*)
- Alternative: use param subspace to design exact samplers (e.g., DILI)

Posterior estimates

- Estimating posterior expectations:
  - Exploit **variance reduction** via de-randomization

$$x = \bar{W}_r x_r + (I - \bar{W}_r \bar{W}_r^\top \Gamma_{\mathrm{pr}}^{-1}) z \;; \quad z \sim N(0, \Gamma_{\mathrm{pr}}), \;\; x_r \sim \pi_{\mathrm{pos}}(x_r \mid y)$$

from MCMC samples

everything known analytically!

  - Example: posterior mean

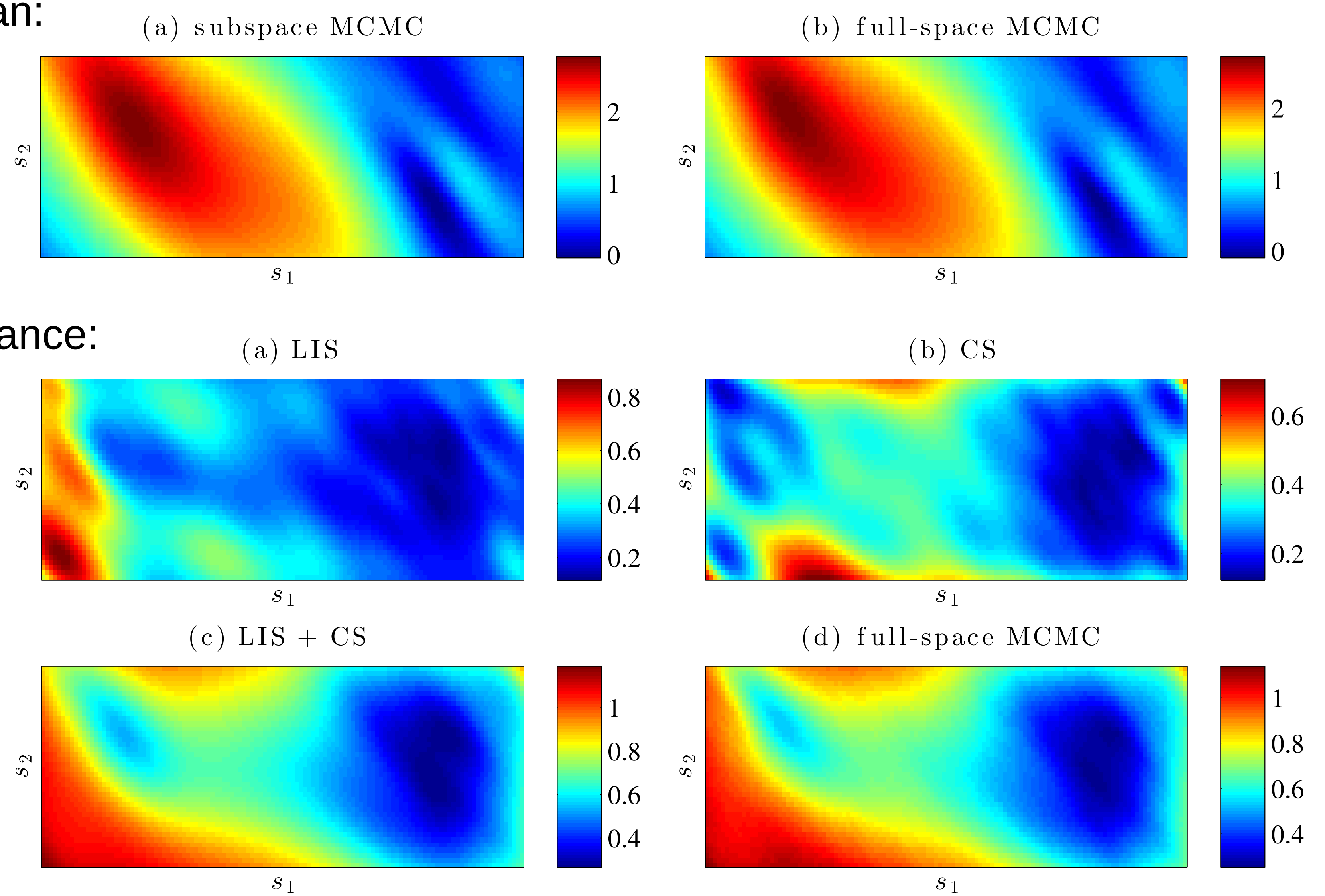
$$\mathbb{E}_\pi[x] = \bar{W}_r \mu_r + (I - \bar{\mathcal{P}}_r) \mu_{pr}$$

  - Example: posterior covariance

$$\mathbb{C}\mathrm{ov}_\pi[x] = \Gamma_{\mathrm{pr}} + \bar{W}_r \left(\tilde{\Gamma}_r - \mathrm{I}_r \right) \bar{W}_r^\top$$

Posterior estimates

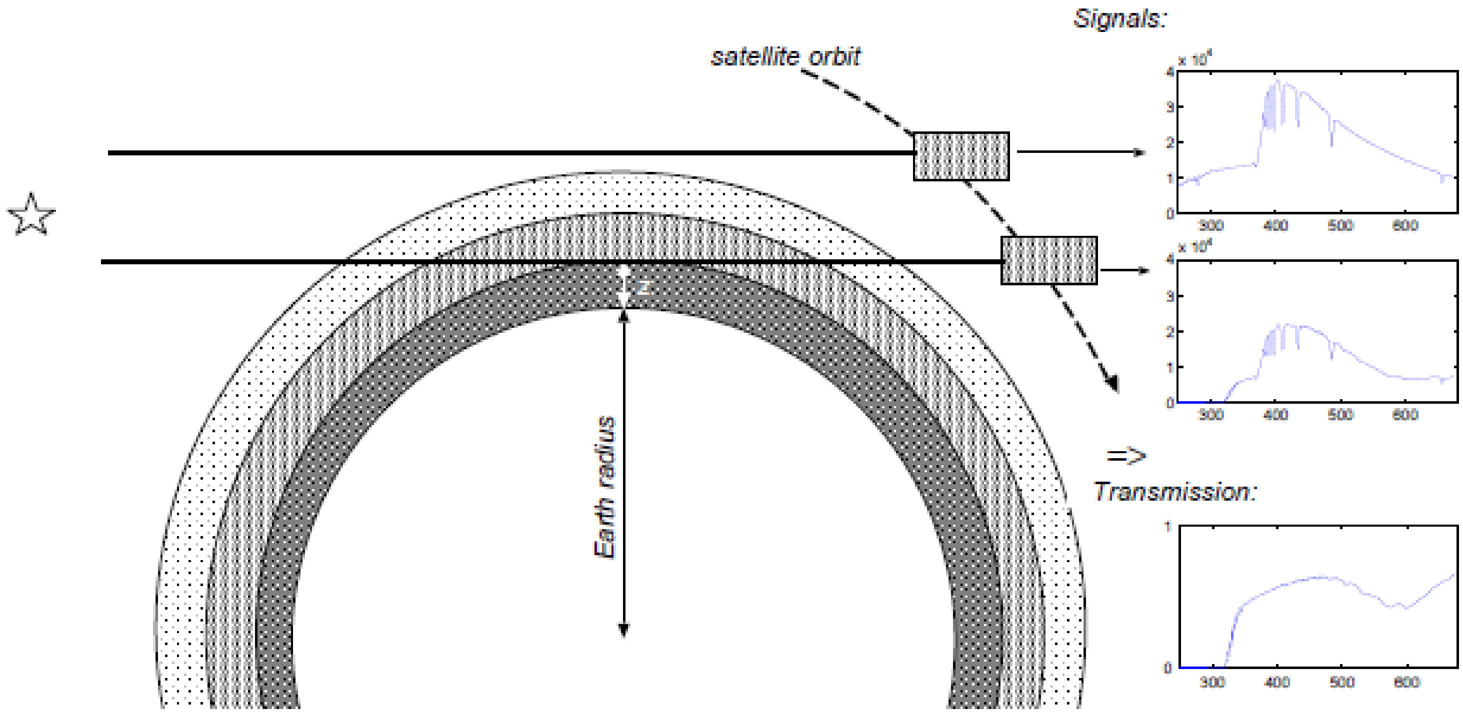
mean:



variance:

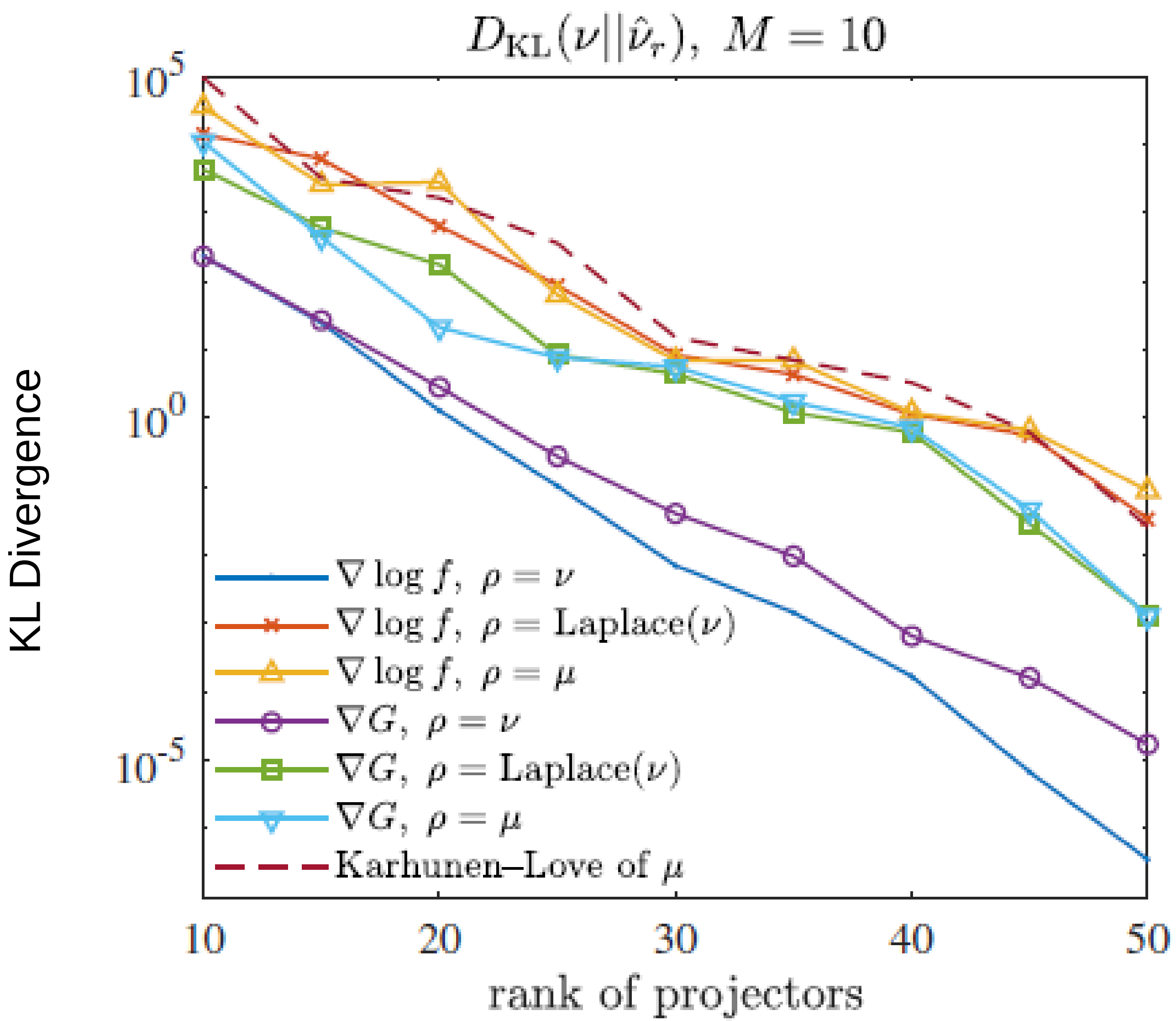
Example: atmospheric sensing

[Haario, Tamminen, et al. 2004]

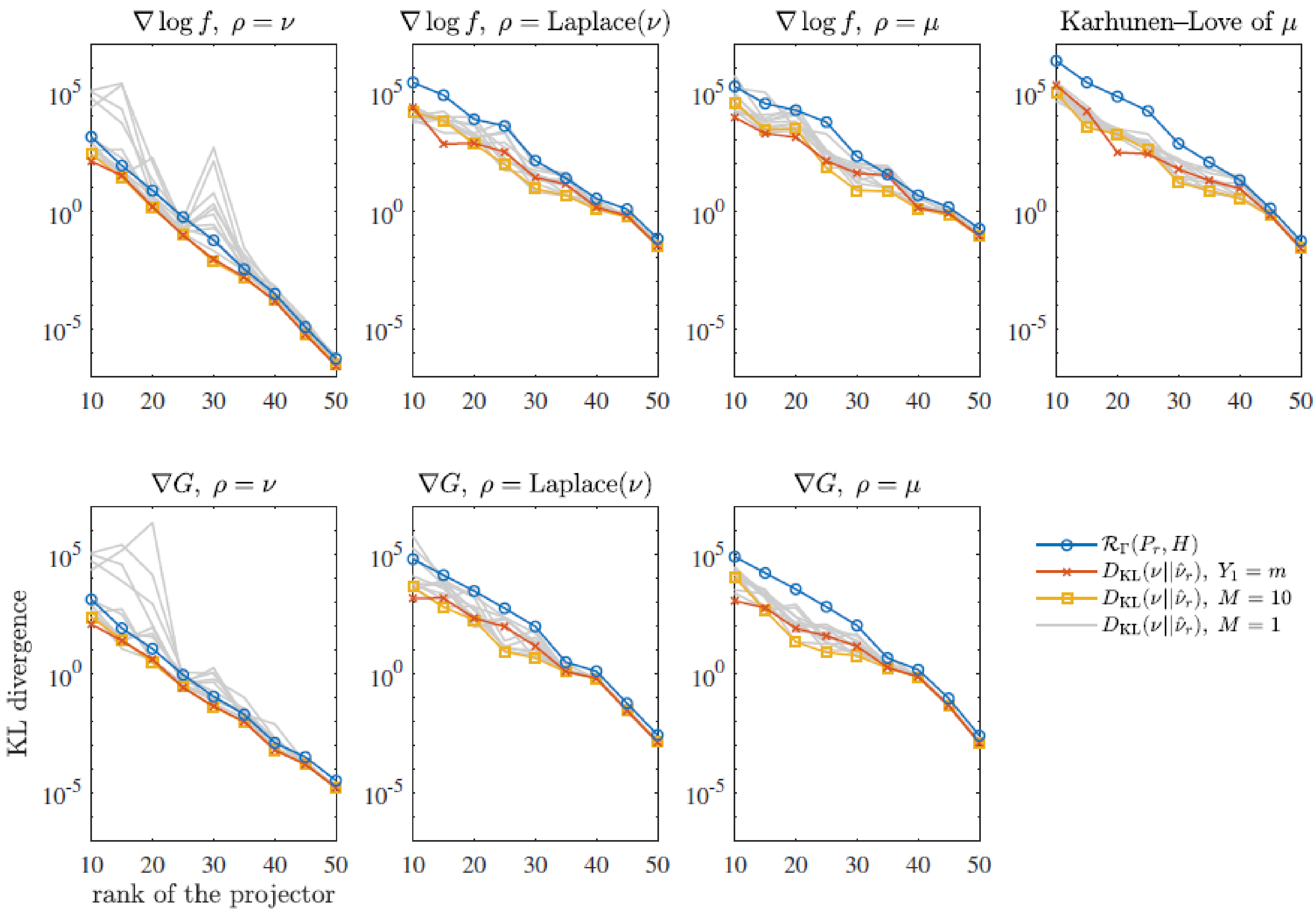


- Estimate gas densities $\rho^{\text{gas}}(z)$ from transmission spectra $T_{\lambda,l}$
- Forward model is nonlinear, $F : \mathbb{R}^{200} \rightarrow \mathbb{R}^{70800}$
- Apply joint reduction

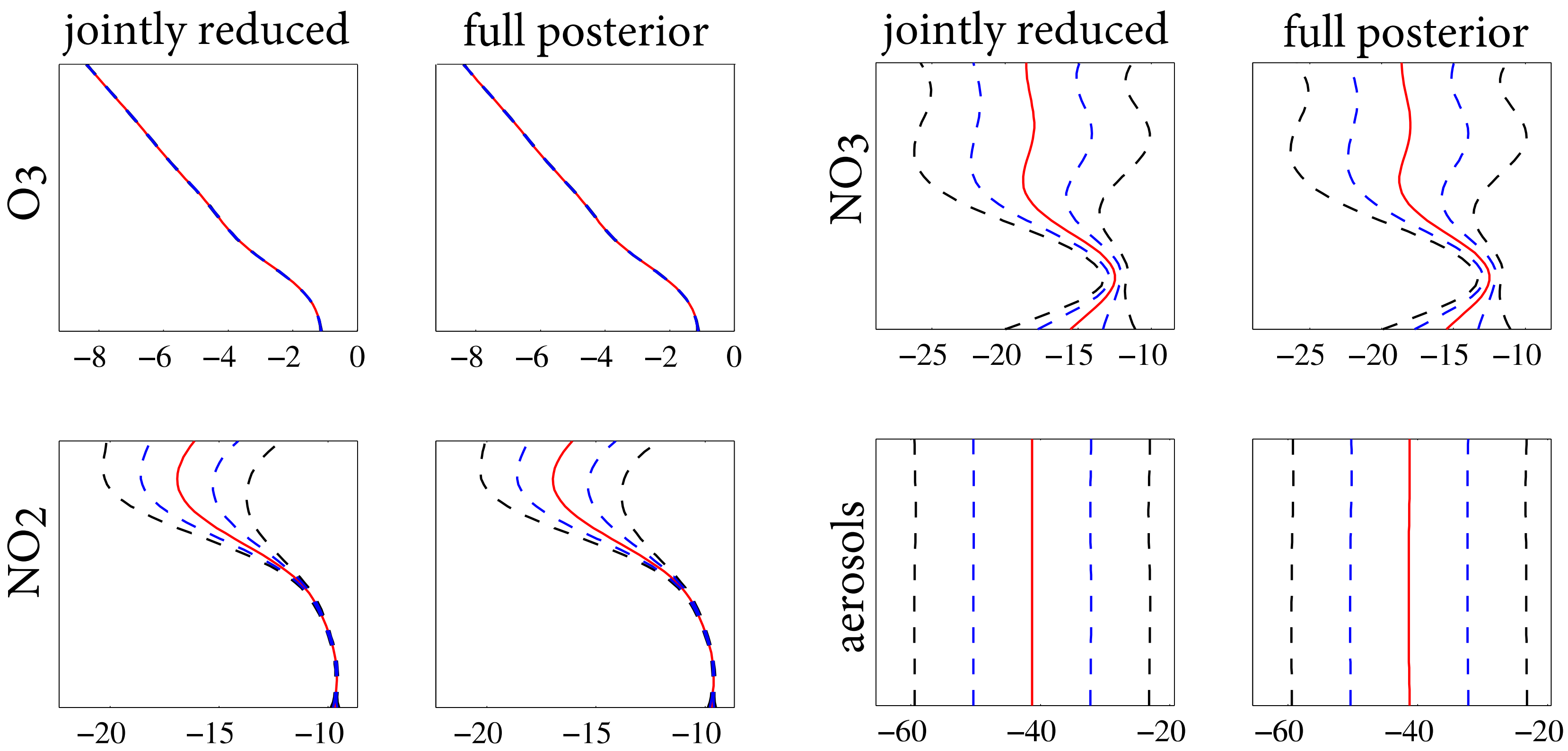
Example: atmospheric sensing



Example: atmospheric sensing



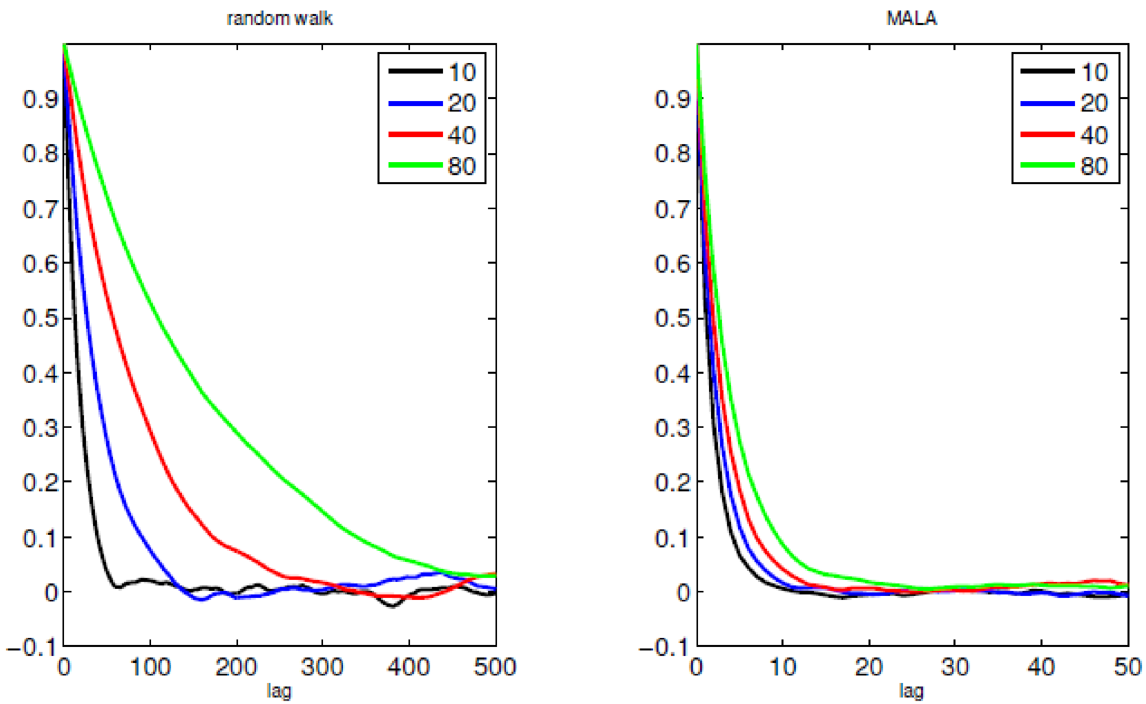
Atmospheric sensing: results



- Estimated gas density profiles
- Full posterior: 200 dim. parameters
- Reduced posterior: 25 dim. parameters

2. Infinite Dimensional MCMC

- Parameter dimension scalability (MCMC).



Random walk $O(N_x)$ MALA $O(N_X^{\frac{1}{3}})$

- Standard MCMC (or other sampling methods) can scale badly.

2. Infinite Dimensional MCMC

Given an MCMC proposal $q(x, dx')$

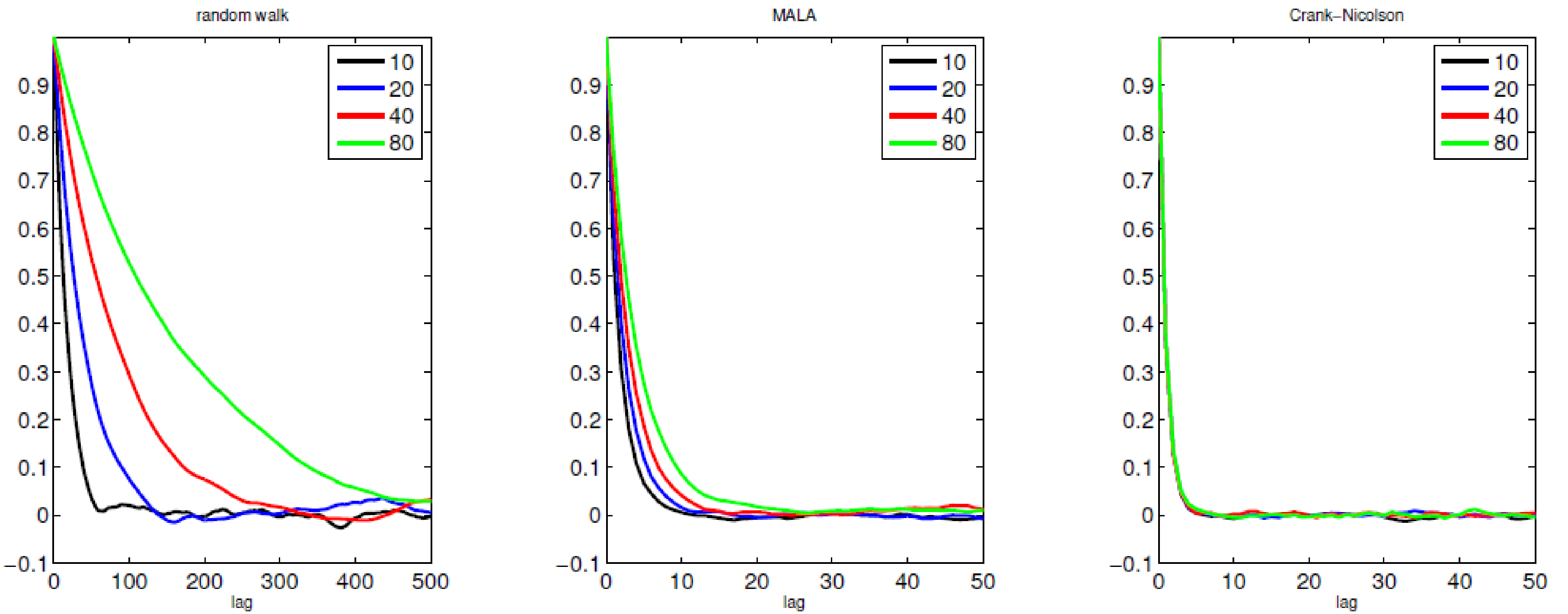
- Transition probability
$$\begin{aligned} \nu(dx, dx') &= \mu_y(dx)\, q(x, dx') \\ \nu^\top(dx, dx') &= \mu_y(dx')q(x', dx) \end{aligned}$$
- Acceptance probability
$$\alpha(x, x') = \frac{d\nu^\top}{d\nu}(x, x')$$
- Requires $\nu \ll \nu^\top$ for a well-defined MCMC for function space
- Many MCMCs defined for the finite dimensional setting $\nu \perp \nu^\top$

(Preconditioned) Crank-Nicholson (PCN) proposal (Beskos et.al. 2008, Cotter et.al. 2013)

$$x' = \sqrt{1 - b^2}\, x + b\,\mathcal{N}(0, \Gamma_{\mathrm{pr}})$$

Satisfies $\nu \ll \nu^\top$

2. Infinite Dimensional MCMC



Random walk $O(N_x)$ MALA $O(N_x^{\frac{1}{3}})$ PCN $O(1)$

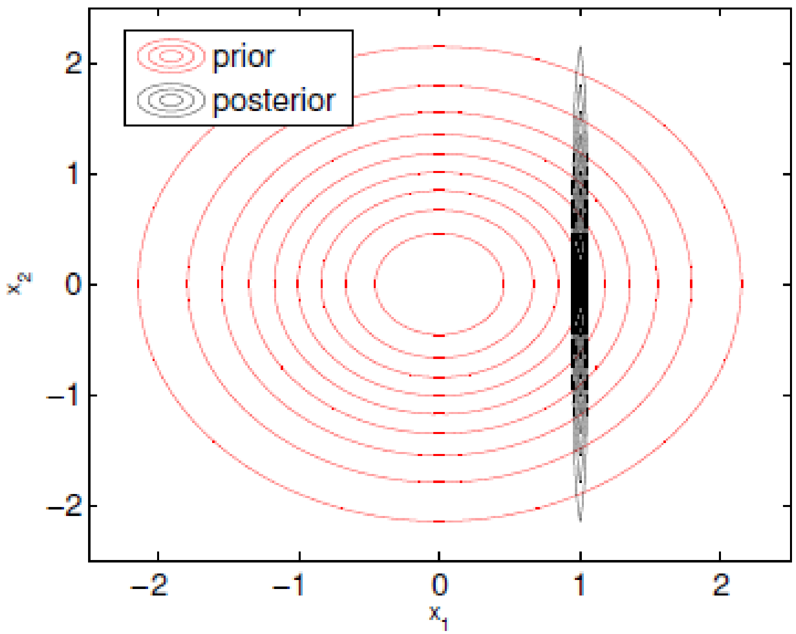
- Autocorrelation of different samplers versus dimensions

2. Infinite Dimensional MCMC

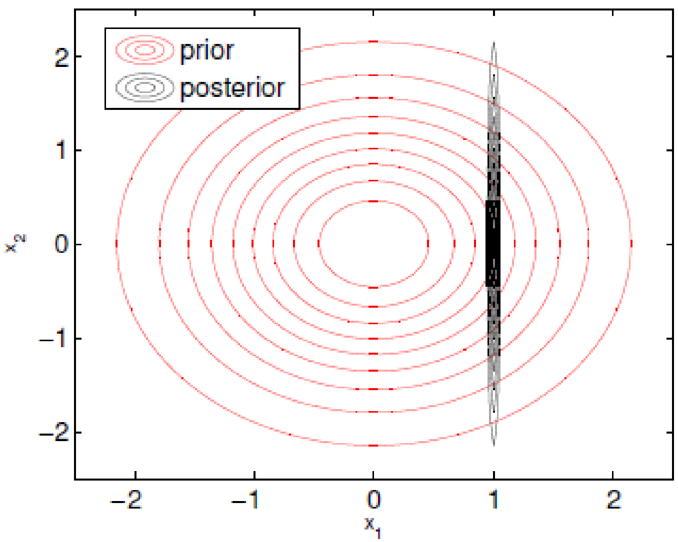
$$x' = \sqrt{1 - b^2}\, x + b\,\mathcal{N}(0, \Gamma_{\mathrm{pr}})$$

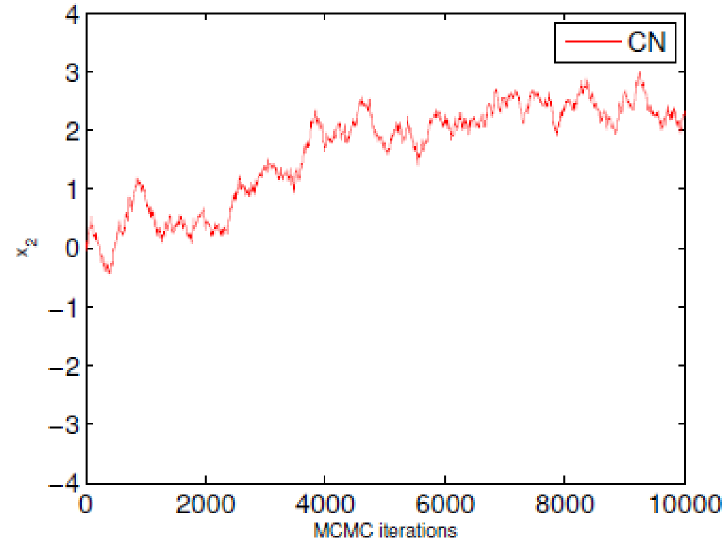
- PCN proposal is isotropic w.r.t. prior
- Likelihood constrains the variability of posterior at some directions.
- What will happen to PCN?
- Consider a linear example:

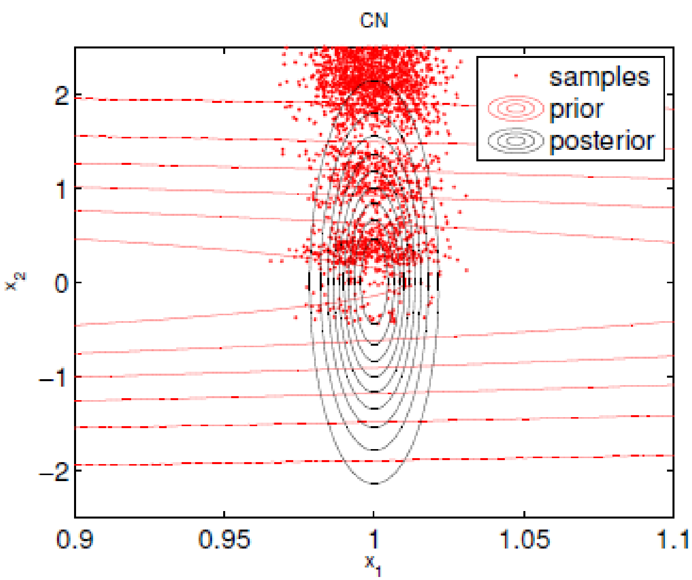
$$y = x_1 + e, \quad e \sim \mathcal{N}(0, \sigma^2), \quad x = (x_1, x_2)^* \sim \mathcal{N}(0, I)$$



2. Infinite Dimensional MCMC





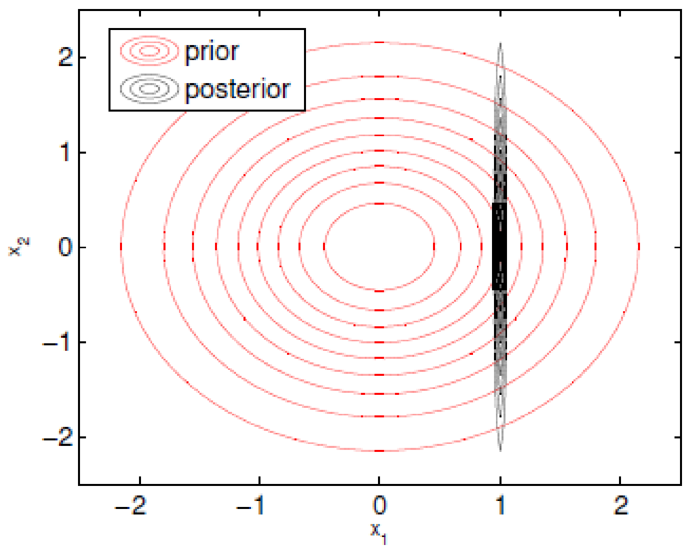


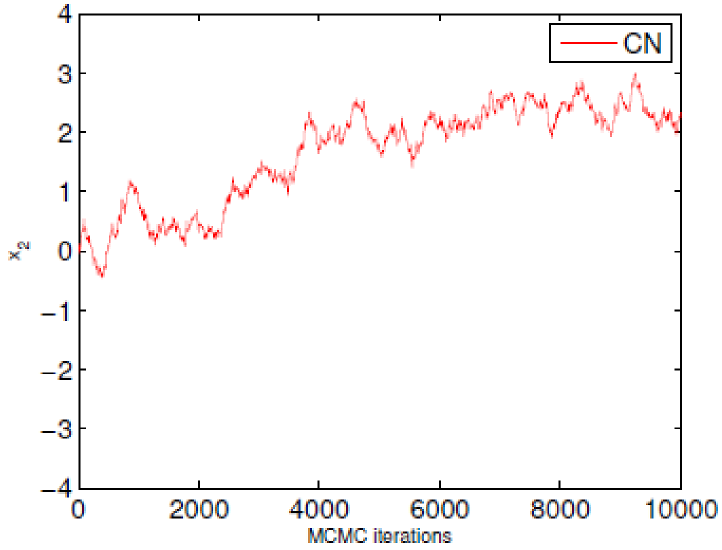
For PCN/CN proposal, the sample correlation

$$\sum_{n=1}^{\infty} \mathrm{corr}(x^{(0)}, x^{(n)}) \approx \frac{\mathrm{const}}{\sigma^2}$$

Problem: μ_y is anisotropic w.r.t. μ_0

2. Infinite Dimensional MCMC

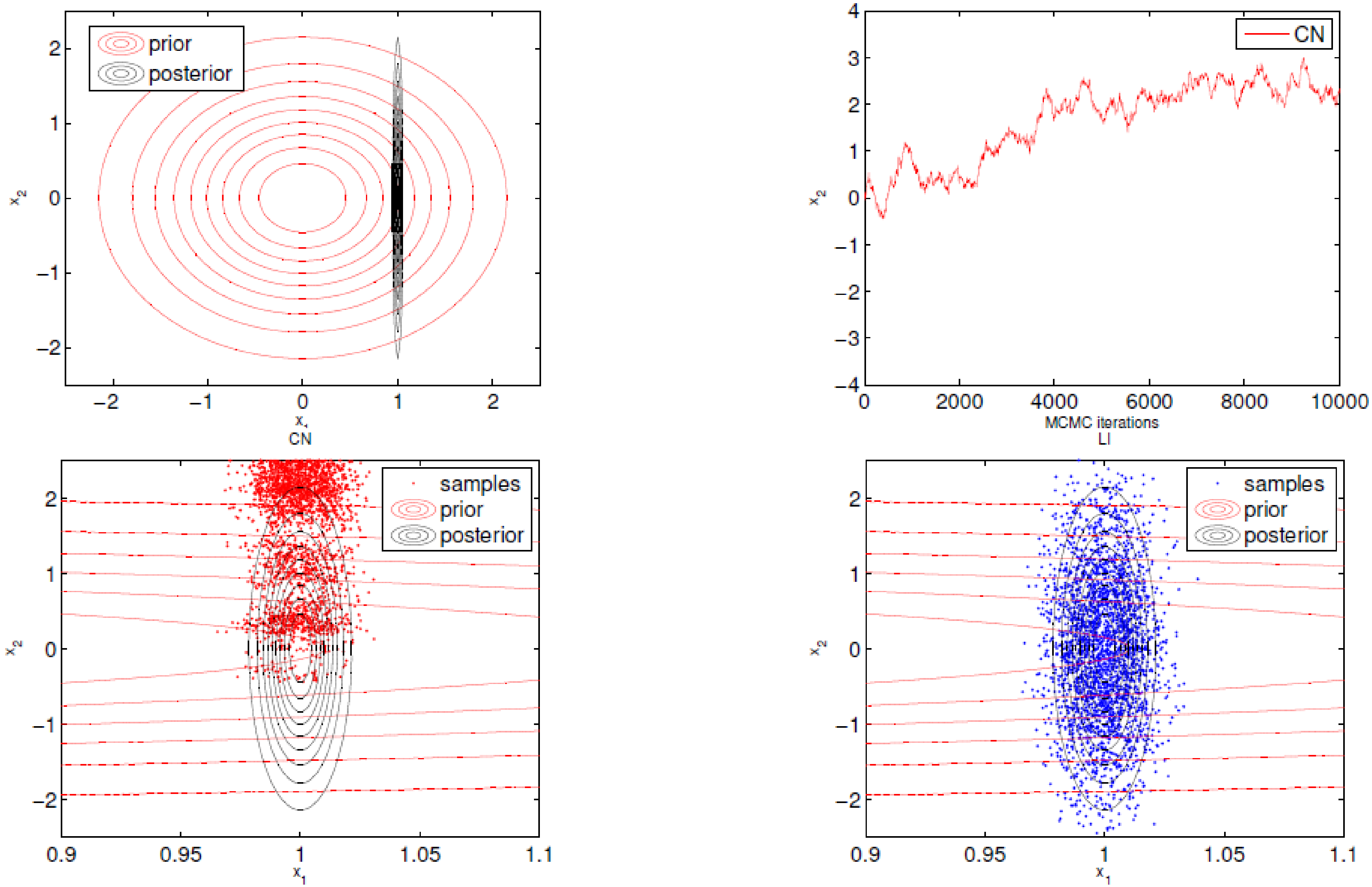




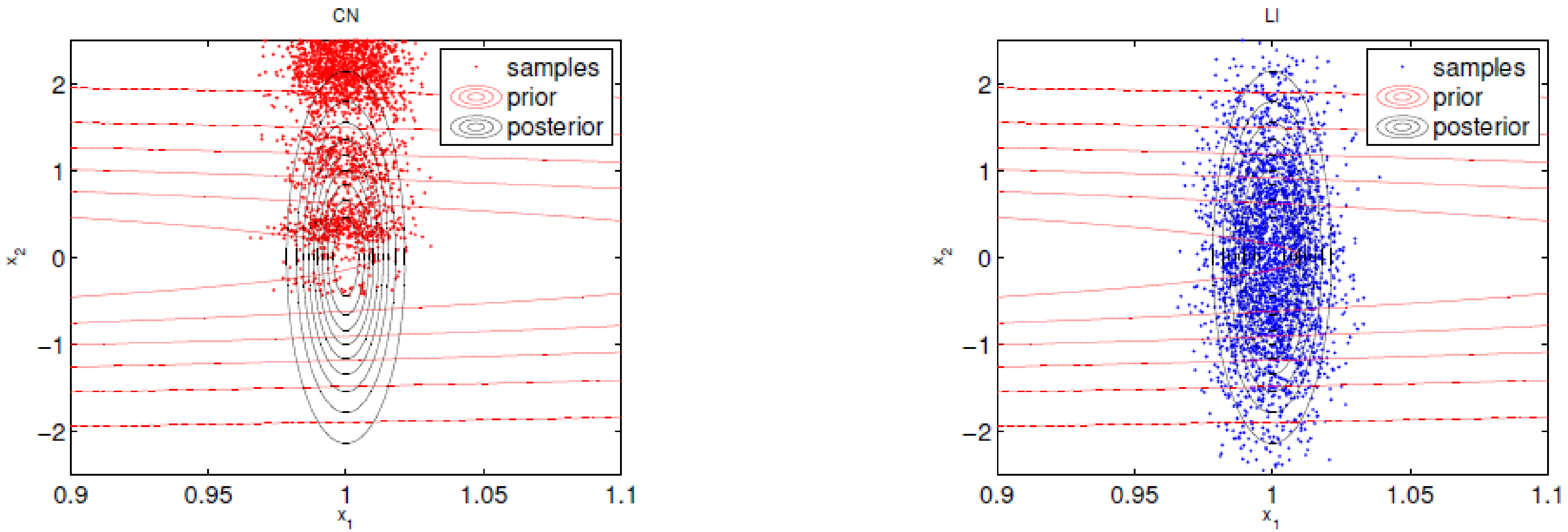
To adapt to this anisotropy, consider an alternative likelihood-informed proposal

$$x' = \begin{bmatrix} \sqrt{1-b^2} & 0 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix} \mathcal{N}(0, I)$$

2. Infinite Dimensional MCMC



2. Infinite Dimensional MCMC



Messages:

- Performance of PCN can be characterized by data dominated directions.
- We want our proposals adapt to the **likelihood information**.
- In function space this leads to operator weighted proposals.

2. Infinite Dimensional MCMC

Operator weighted proposals

$$x' = \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} A \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\right) x + \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} G \Gamma_{\mathrm{pr}}^{\frac{1}{2}}\right) \nabla \log L(y|x) + \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} B\right) \mathcal{N}(0, I)$$

Where A, B and G are are commutative, bounded, self-adjoint operators.

Given $\mathrm{Trace}\left(\left(A^2 + B^2 - I\right)^2\right) < \infty$, and other mild technical conditions, we have $\nu \ll \nu^\top$ (and $\nu^\top \ll \nu$). Thus the operator proposal is well-defined in the function space setting (Cui, Law & Marzouk 2016).

2. Infinite Dimensional MCMC

$$x' = \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} A \Gamma_{\mathrm{pr}}^{-\frac{1}{2}}\right) x + \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} G \Gamma_{\mathrm{pr}}^{\frac{1}{2}}\right) \nabla \log L(y|x) + \left(\Gamma_{\mathrm{pr}}^{\frac{1}{2}} B\right) \mathcal{N}(0, I)$$

Use the likelihood informed subspace (LIS) to build A, B and G.

- The LIS is spanned by the basis $\overline{W} = \Gamma_{\mathrm{pr}}^{\frac{1}{2}} \Psi$
- $[\Gamma_{\mathrm{pr}}^{\frac{1}{2}} \Psi, \Gamma_{\mathrm{pr}}^{\frac{1}{2}} \Psi_\perp]$ form a complete orthogonal system w.r.t. $\langle \cdot, \cdot \rangle_{\Gamma_{\mathrm{pr}}^{-1}}$

$$x = \underbrace{\Gamma_{\mathrm{pr}}^{\frac{1}{2}} \Psi \, v_r}_{\textit{Constrained by data}} + \underbrace{\Gamma_{\mathrm{pr}}^{\frac{1}{2}} \Psi_\perp \, v_\perp}_{\approx \textit{ prior}}$$

- Prescribes different proposal scales to v_r and $v_\perp$
  - v_r Use the best proposal for finite dimensional
  - $v_\perp$ Use Crank-Nicholson

2. Infinite Dimensional MCMC

Split the operators

$$A = A_r + A_\perp \qquad B = B_r + B_\perp \qquad G = G_r + G_\perp$$

LI-Langevin

$$A_r = \Psi_r D_{A_r} \Psi_r{}^* \qquad D_{A_r} = I_r - \triangle t_r D_r \qquad A_\perp = a_\perp (I - \Psi_r \Psi_r{}^*)$$

$$B_r = \Psi_r D_{B_r} \Psi_r{}^* \qquad D_{B_r} = \sqrt{2 \triangle t_r D_r} \qquad B_\perp = b_\perp (I - \Psi_r \Psi_r{}^*)$$

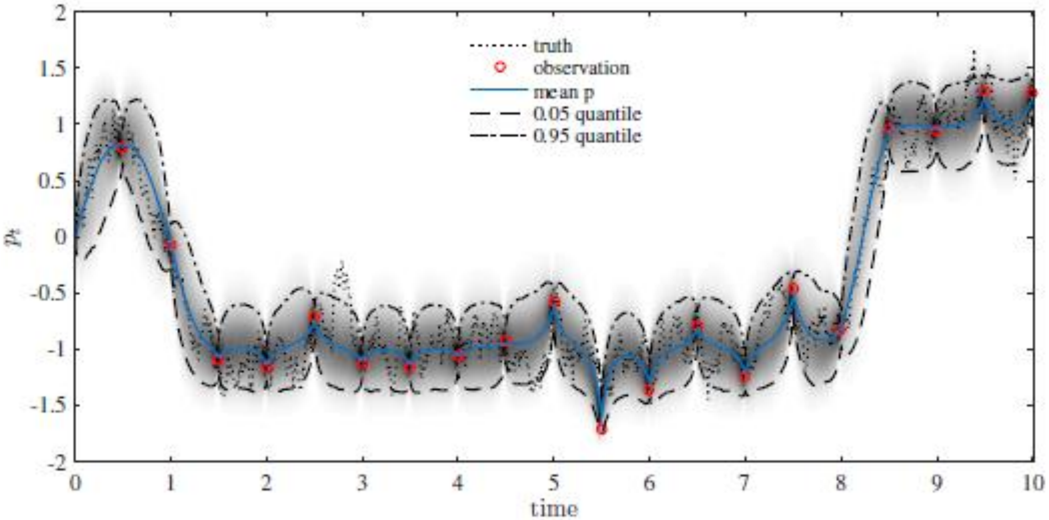
$$G_r = \Psi_r D_{G_r} \Psi_r{}^* \qquad D_{G_r} = \triangle t_r D_r \qquad G_\perp = 0$$

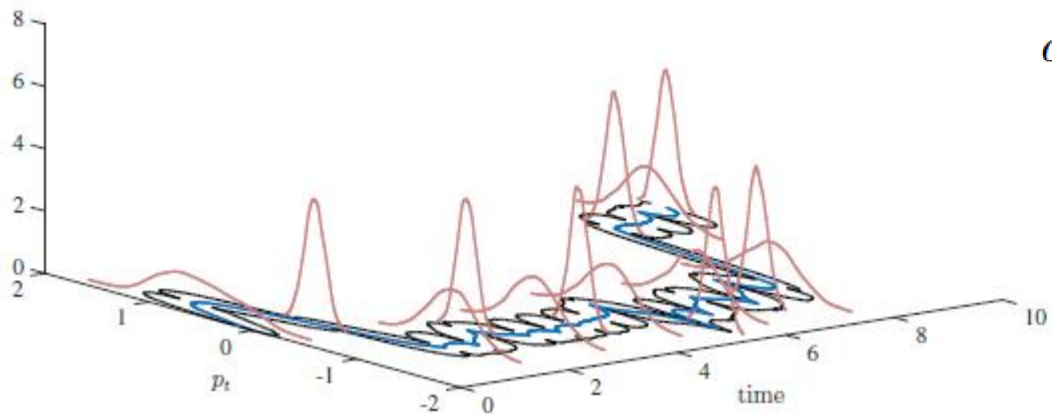
Metropolis-within-Gibbs, alternate on u_r and $u_\perp$

$$A_r = \Psi_r (D_{A_r} - I_r) \Psi_r{}^* + I \qquad B_r = \Psi_r D_{B_r} \Psi_r{}^* \qquad G_r = \Psi_r D_{G_r} \Psi_r{}^*$$

$$A_\perp = \Psi_r \Psi_r{}^* + a_\perp (I - \Psi_r \Psi_r{}^*) \qquad B_\perp = b_\perp (I - \Psi_r \Psi_r{}^*) \qquad G_\perp = 0$$

Example: Conditioned Diffusion

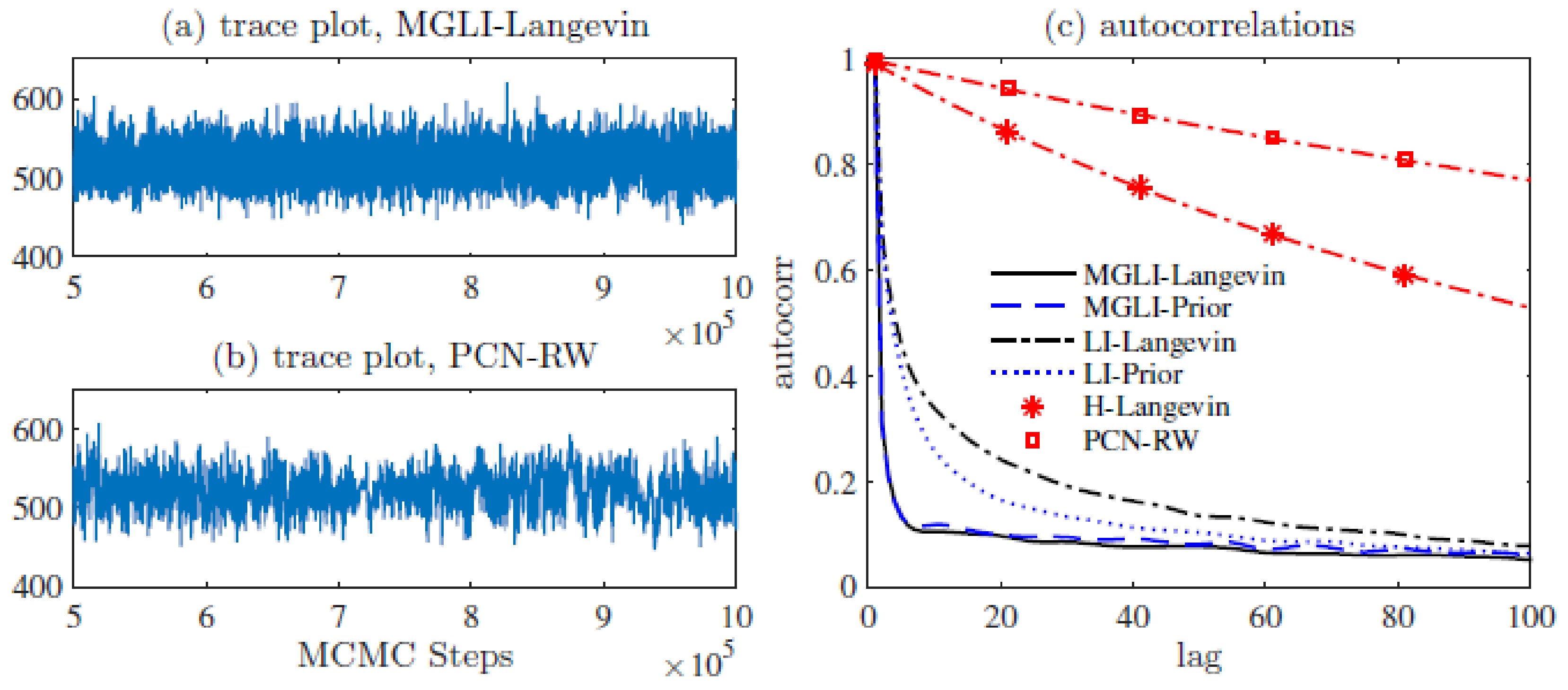




$$dp(t) = f(p(t), \alpha)dt + \sigma(p(t), \beta)du_t$$
$$p(0) = p_0$$
$$y(t) = h(p(t))$$

Estimating parameters and random forcing of SDE to characterize financial and biological systems

Example: Conditioned Diffusion



- Left: MCMC trace. Right: autocorrelation
- H-Langevin: Explicit discretization of Langevin SDE, preconditioned by Hessian at the MAP.

3. Joint model reduction

- Using parameter dimension we can derive parameter scalable solvers, but forward model evaluations are still expensive
- Consider ***joint*** parameter and **model** reduction:

$$\begin{aligned}\pi^a_{\text{pos}}(x_r, x_\perp \mid y) &\propto \pi_{\text{like}}(y, F(\bar{W}_r x_r))\pi_0(x_r) \quad \pi_0(x_\perp)\\ &= \qquad \pi_{\text{pos}}(x_r \mid y) \qquad \pi_0(x_\perp)\\ &\approx \pi_{\text{like}}(y, F^a(\bar{W}_r x_r))\pi_0(x_r)\ \pi_0(x_\perp)\\ &= \qquad \pi^a_{\text{pos}}(x_r \mid y) \qquad \pi_0(x_\perp)\end{aligned}$$

- F^a is a reduced-order model
  - Its input parameters are restricted to lie in the parameter subspace
  - Construct it to ensure accuracy over the support of a chosen reference distribution, marginalized onto the parameter subspace

Model reduction: background

- Consider the PDE model

$$\underbrace{B(x)u}_{\text{linear}} + \underbrace{g(x,u)}_{\text{nonlinear}} = 0, \quad u \in \mathbb{R}^{n_s}, \; n_s \text{ large}$$

- ***State subspace**:* given a distribution on *x*, seek an *r*-dimensional subspace spanned by $\Phi \in \mathbb{R}^{n_s \times r}, \; r \ll n_s$, capturing most of the variation of the solution field *u(x)*

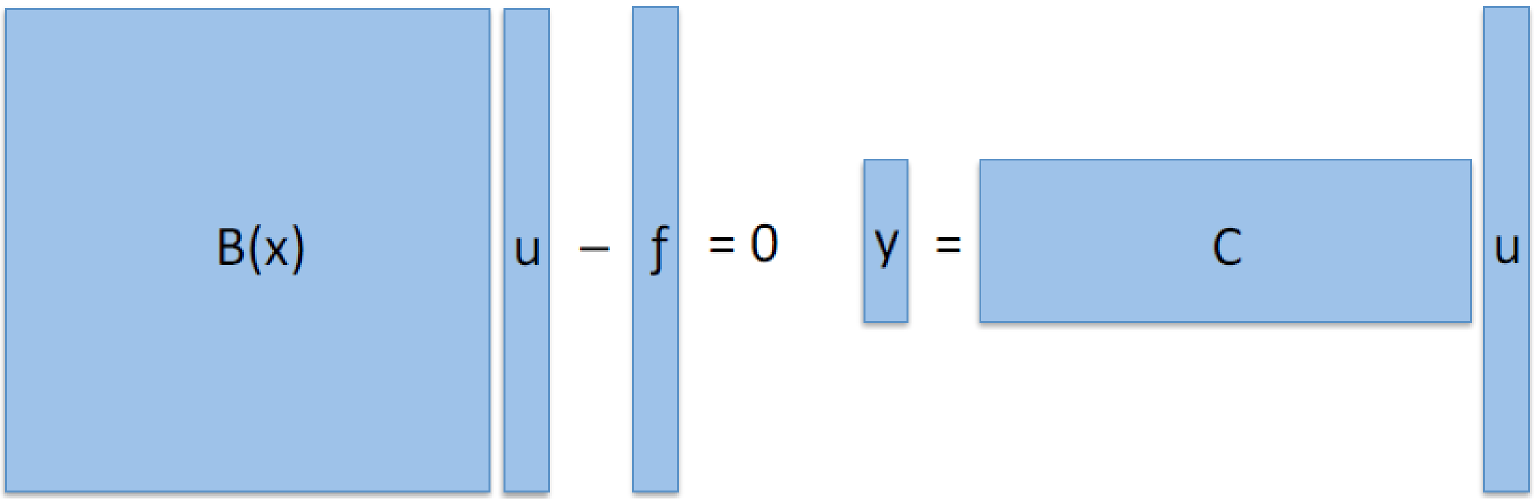
- *Reduced order model* (ROM): obtain $u(x) \approx \Phi u_r(x)$ by solving a smaller system of equations:

$$\text{Galerkin}: \quad \Phi^\top \Big[B(x)\Phi u_r + g(x, \Phi u_r) \Big] = 0$$

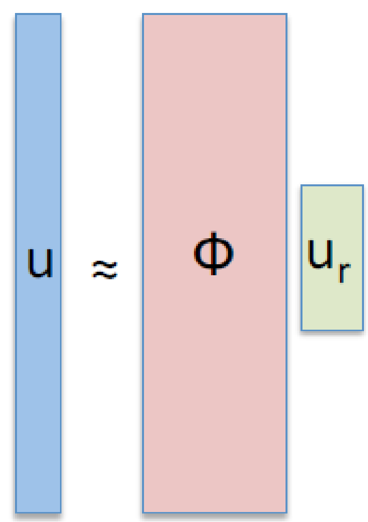
  – EIM/DEIM methods to deal with nonlinear terms
  – Vast literature, for linear and nonlinear PDEs…

Model reduction: example

$-\nabla\cdot(k(x)\,\nabla u) = f$ *and* observation operator $C \Rightarrow y{=}F(x)$



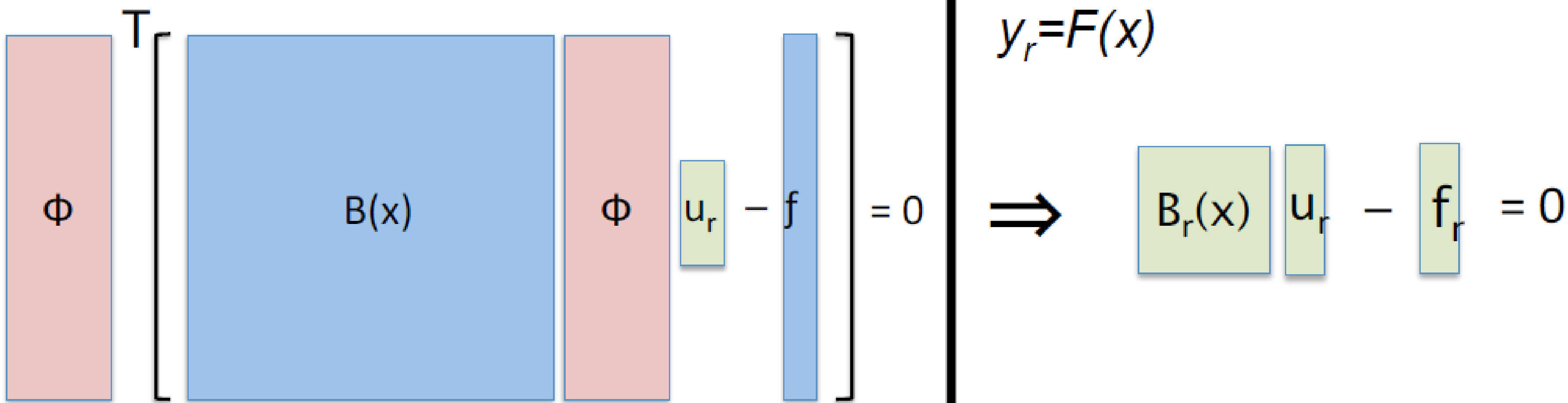
Approximate the state using reduced basis Φ



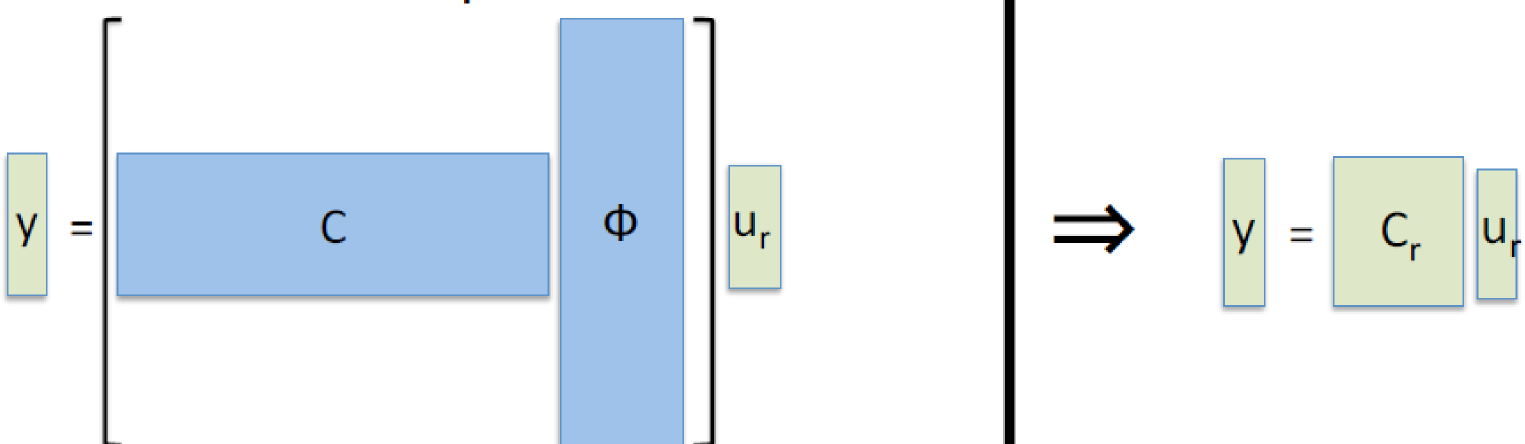
Model reduction: example

Galerkin projection

Reduced forward model
$y_r = F(x)$

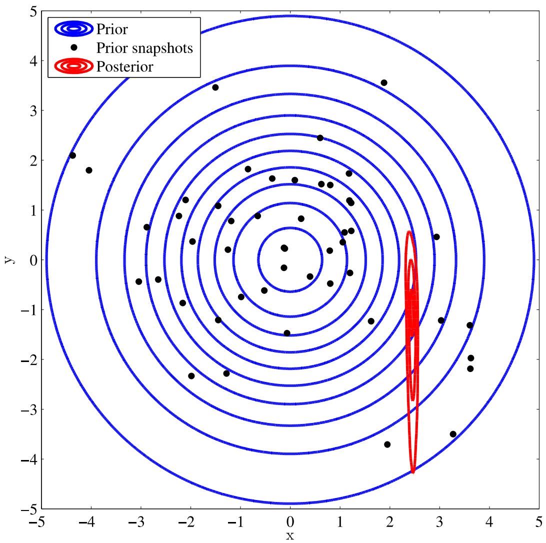


Reduce observation operator



Exploiting locality

- Construct the state basis Φ by selecting *snapshots*
- Take advantage of **locality** in two senses:
  - Posterior can ***concentrate*** significantly with respect to the prior [Cui *et al.* IJNME 2014]



  - Many **directions** of parameter variation induce state variation *unimportant* to inverse problems (i.e., complement of param subspace)
- *Approach:* select a reference measure, marginalize onto the param subspace, use to generate state snapshots

Choice of reference: state subspace

Linear example: spectra of state covariances given different reference distributions

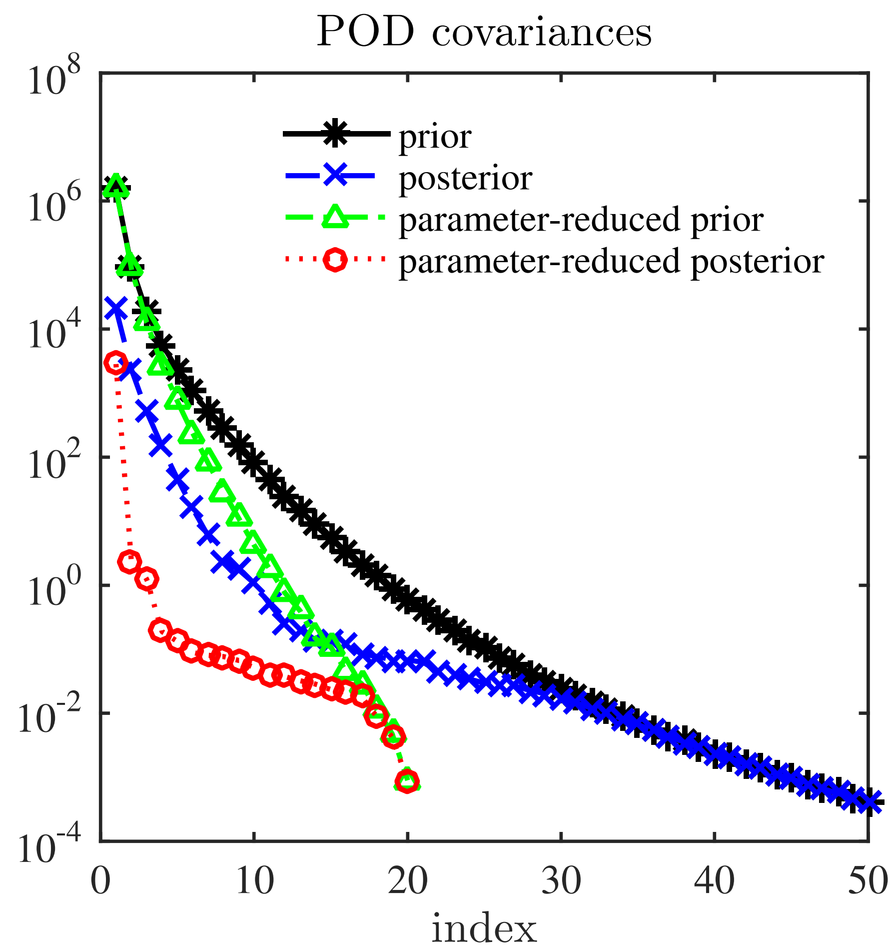
$$\text{POD basis: } K_{(\bullet)}\Phi = \Phi\Lambda$$

$$K_{\text{pr}} = \int u(x)u(x)^\top \pi_0(dx)$$

$$K_{\text{pos}} = \int u(x)u(x)^\top \pi(dx \mid y)$$

$$\tilde{K}_{\text{pr}} = \int u(\bar{W}x_r)u(\bar{W}x_r)^\top \pi_0(dx_r)$$

$$\tilde{K}_{\text{pos}} = \int u(\mathcal{P}_r x)u(\mathcal{P}_r x)^\top \pi(dx \mid y)$$



Joint reduction: algorithms

- How to find parameter subspace and state subspace?

- *Prior* as reference:
  - Evaluate low-rank Hessians at prior samples, solve generalized eigenproblem; obtain LIPS
  - Project prior samples onto LIPS to obtain state snapshots; construct LISS
  - **Offline** (i.e., independent of data) and embarrassingly parallel

- *Laplace* approximation as reference
  - Same as above, but replace prior samples with samples from Laplace approximation of the posterior
  - **Online** (Laplace approx requires data) and embarrassingly parallel

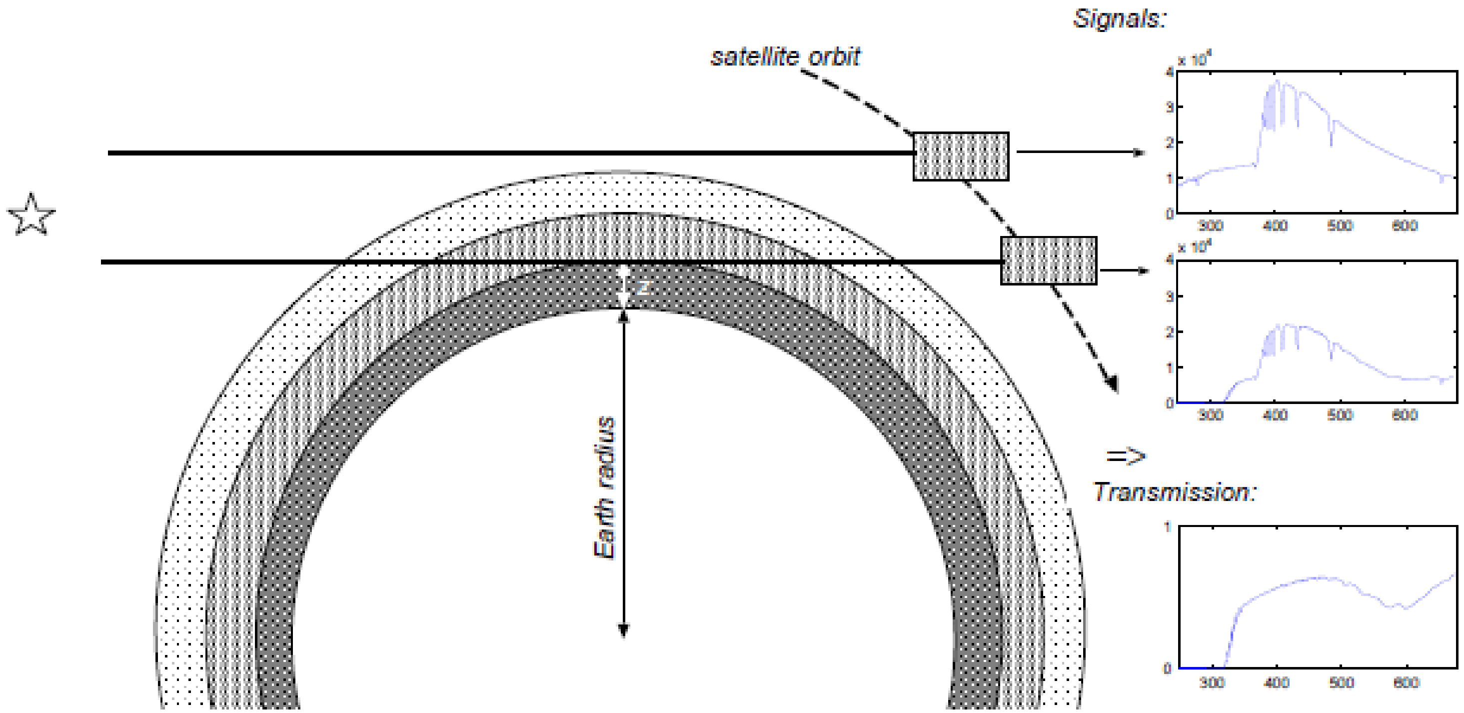
- *Posterior* as reference: need an online and **iterative** approach…

Iterative approach for posterior-ref

- *Given:* initial reduced param basis W_0 and initial reduced state basis Ψ_0, inducing $\tilde{p}_0(x \mid y)$ (projected parameter) and $p_0^*(x \mid y)$ (projected parameter and state)
- At each iteration *n = 0,1,…*
  1. Draw samples from $p_n^*(x \mid y)$ using product-form decomposition: MCMC on $p_n^*(x_r \mid y)$ and direct sampling on $p_n(x_\perp)$
  2. Update param basis W_{n+1} by computing
  $$S_{n+1} = \int \frac{p(x \mid y)}{p_n^*(x \mid y)} H(x) p_n^*(x \mid y)\, dx$$
  via importance sampling
  3. Update state basis Ψ_{n+1} through
  $$K_{n+1} = \int \frac{\tilde{p}_{n+1}(x \mid y)}{p_n^*(x \mid y)} \left[u(\mathcal{P}_{n+1} x)\, u^\top (\mathcal{P}_{n+1} x) \right] p_n^*(x \mid y)\, dx$$
  5. $n \leftarrow n+1$
- Steps #2 and #3 involve the full model, and can easily be parallelized

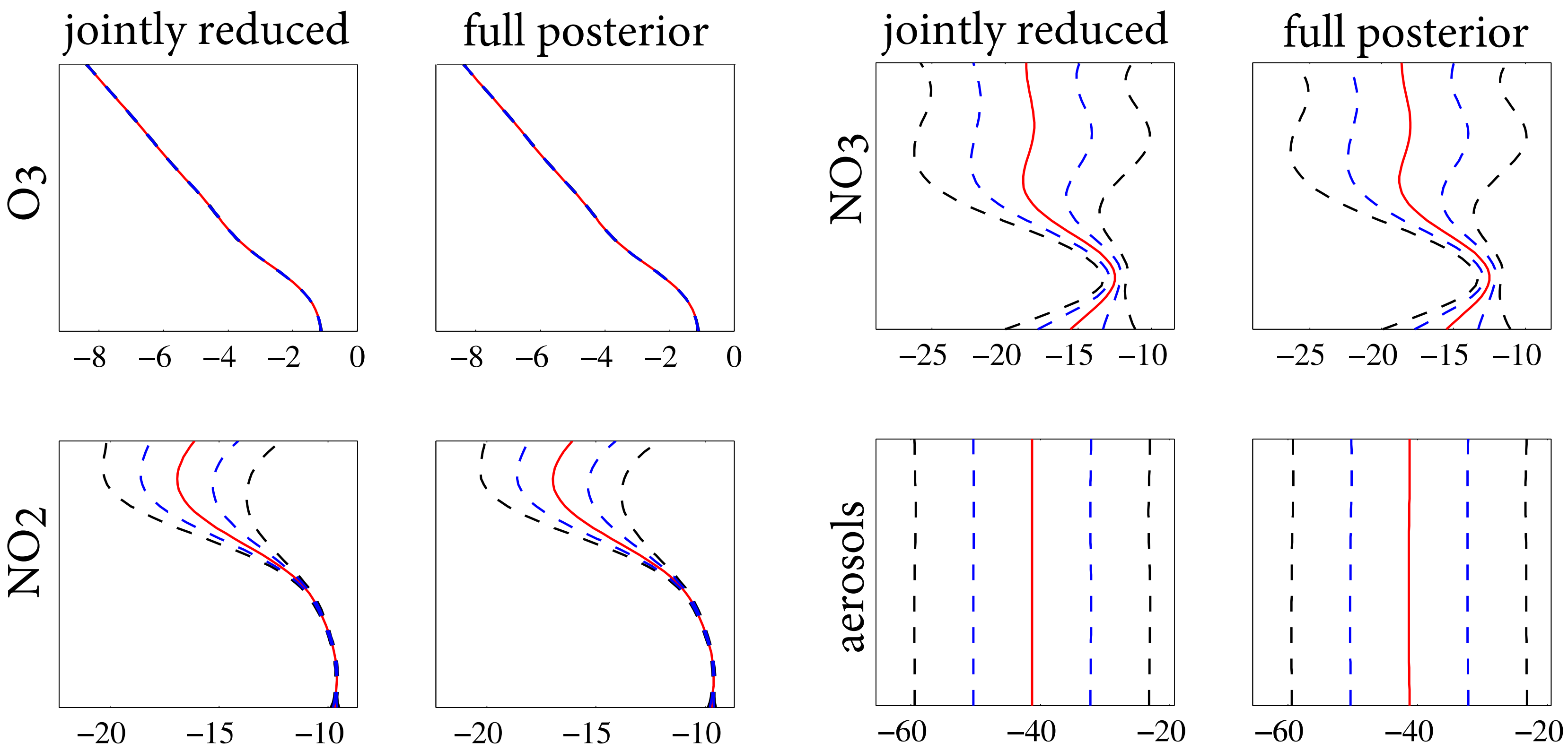
Example: atmospheric sensing

[Haario, Tamminen, et al. 2004]



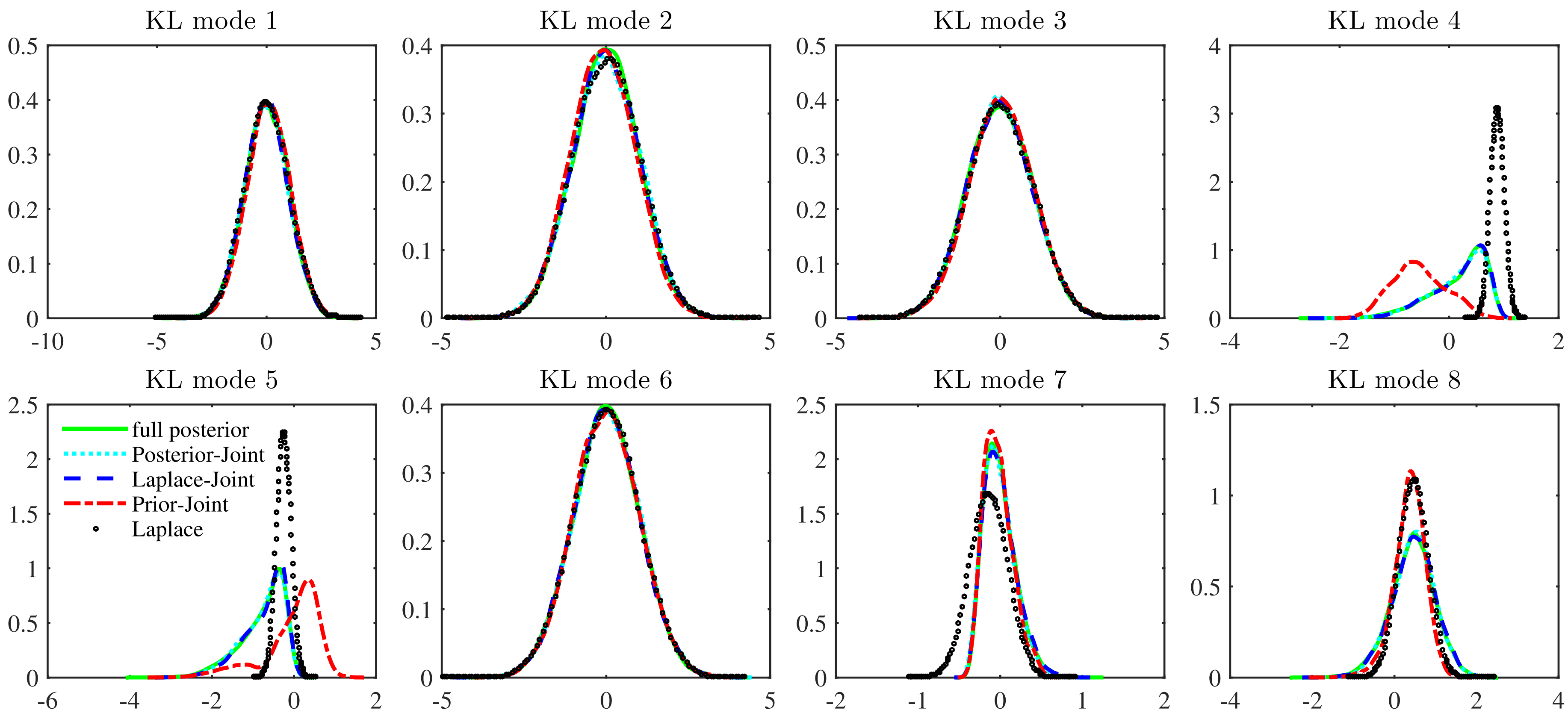
- Estimate gas densities $\rho^{gas}(z)$ from transmission spectra $T_{\lambda,l}$
- Forward model is nonlinear, $F : \mathbb{R}^{200} \rightarrow \mathbb{R}^{70800}$
- Apply joint reduction

Atmospheric sensing: results



- Estimated gas density profiles
- Full posterior: 200 dim. parameters + 70800 dim. states / data
- Jointly reduced posterior: 25 dim. parameters + 45 dim. states / data

Atmospheric sensing: results

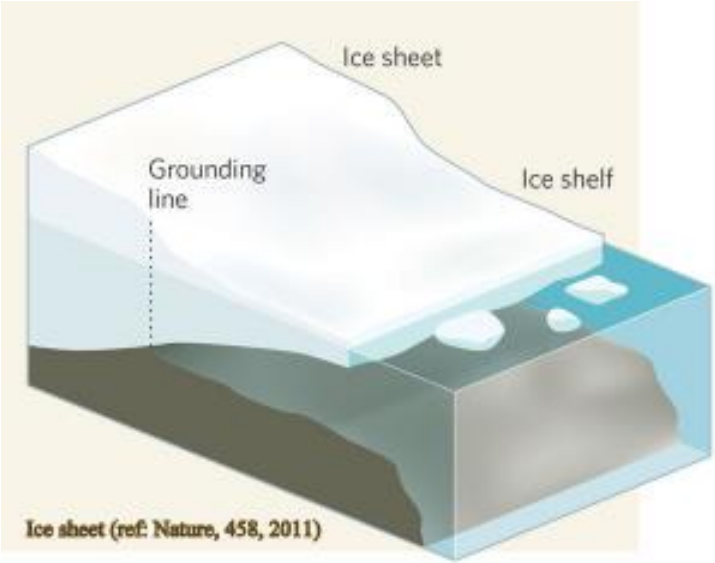


- Posterior distributions marginalized onto the first several prior K-L modes

Example 2: Arolla glacier

Goal: estimating basal sliding coefficients from surface velocity measurements.

$$
\begin{aligned}
-\boldsymbol{\nabla} \cdot [2\eta(\boldsymbol{u})\,\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} - \boldsymbol{I}p] &= \rho \boldsymbol{g} \quad \text{in } \Omega \\
\boldsymbol{\nabla} \cdot \boldsymbol{u} &= 0 \quad \text{in } \Omega \\
\boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n} &= \boldsymbol{0} \quad \text{on } \Gamma_t \\
\boldsymbol{u} \cdot \boldsymbol{n} &= 0 \quad \text{on } \Gamma_b \\
\boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \exp(x)\boldsymbol{T}\boldsymbol{u} &= \boldsymbol{0} \quad \text{on } \Gamma_b
\end{aligned}
$$



- $\boldsymbol{u}$ ice flow velocity, p pressure
- $\boldsymbol{\sigma}_{\boldsymbol{u}} = -\boldsymbol{I}p + 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}$ stress tensor
- $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^\top)$ strain rate tensor
- $\eta(\boldsymbol{u}) = \frac{1}{2} A^{-\frac{1}{n}}\ \dot{\varepsilon}_{\mathrm{II}}^{\frac{1-n}{2n}}$ effective viscosity
- $\dot{\varepsilon}_{\mathrm{II}} = \frac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)$ second invariant of the strain rate tensor

- ρ density, $\boldsymbol{g}$ gravity
- $\boldsymbol{n}$ unit normal vector
- x log basal sliding coefficient
- $\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{n} \otimes \boldsymbol{n}$ tangential operator
- Γ_t and Γ_b top and base boundaries

*Joint work with Ghattas, Petra, Peherstorfer

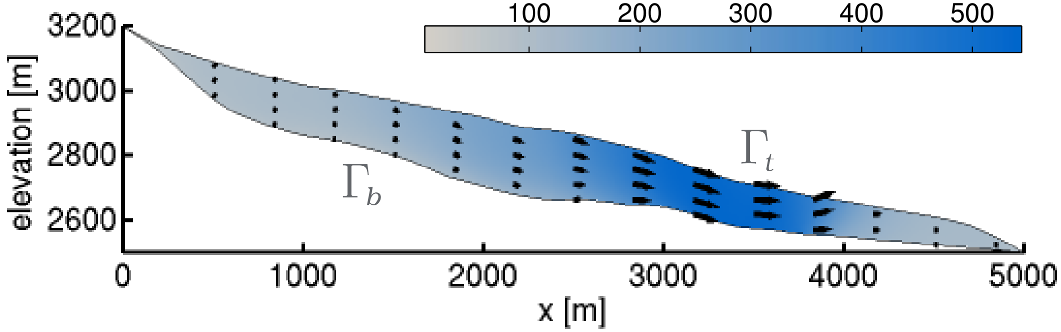
Arolla glacier: setup

- Discretization system:

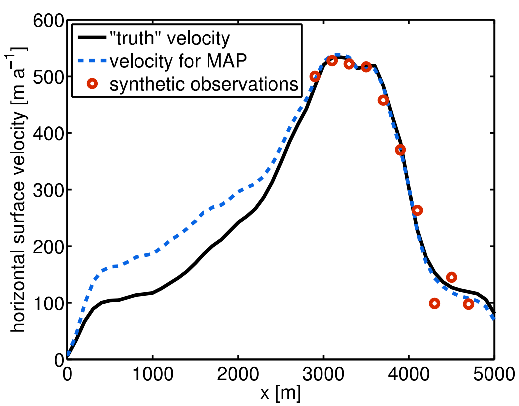
$$\boldsymbol{K}(\underline{\boldsymbol{u}}, \underline{x})\underline{\boldsymbol{u}} + \boldsymbol{B}^\top \underline{p} = -\underline{\vec{r}}(\underline{\boldsymbol{u}}, \underline{p}), \quad \boldsymbol{B}\underline{\boldsymbol{u}} = \boldsymbol{0},$$

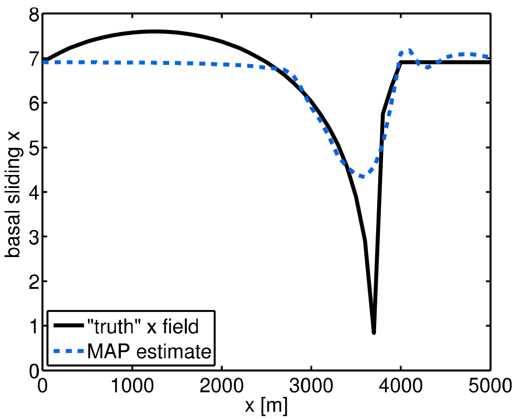
where B is the discretization of the divergence operator.

- One dimensional model to validate our methods

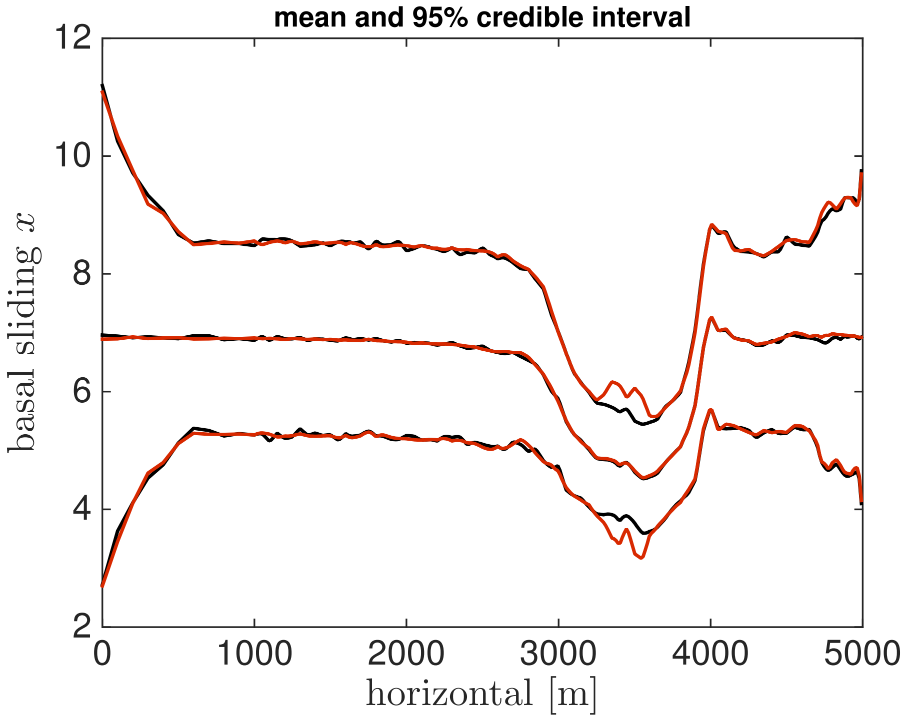


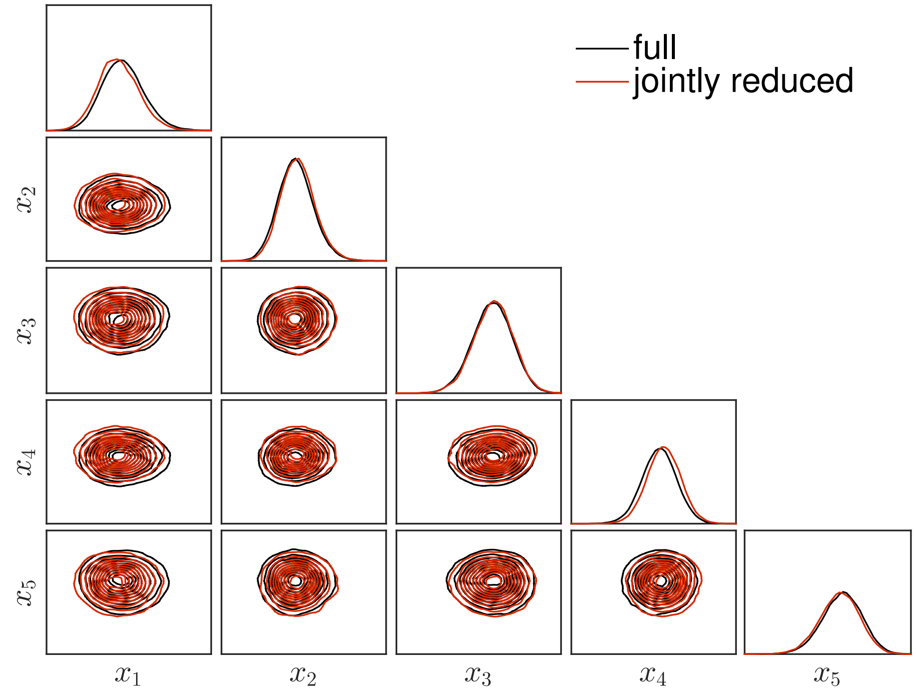
- Synthetic data and MAP estimate (used as the initial guess)





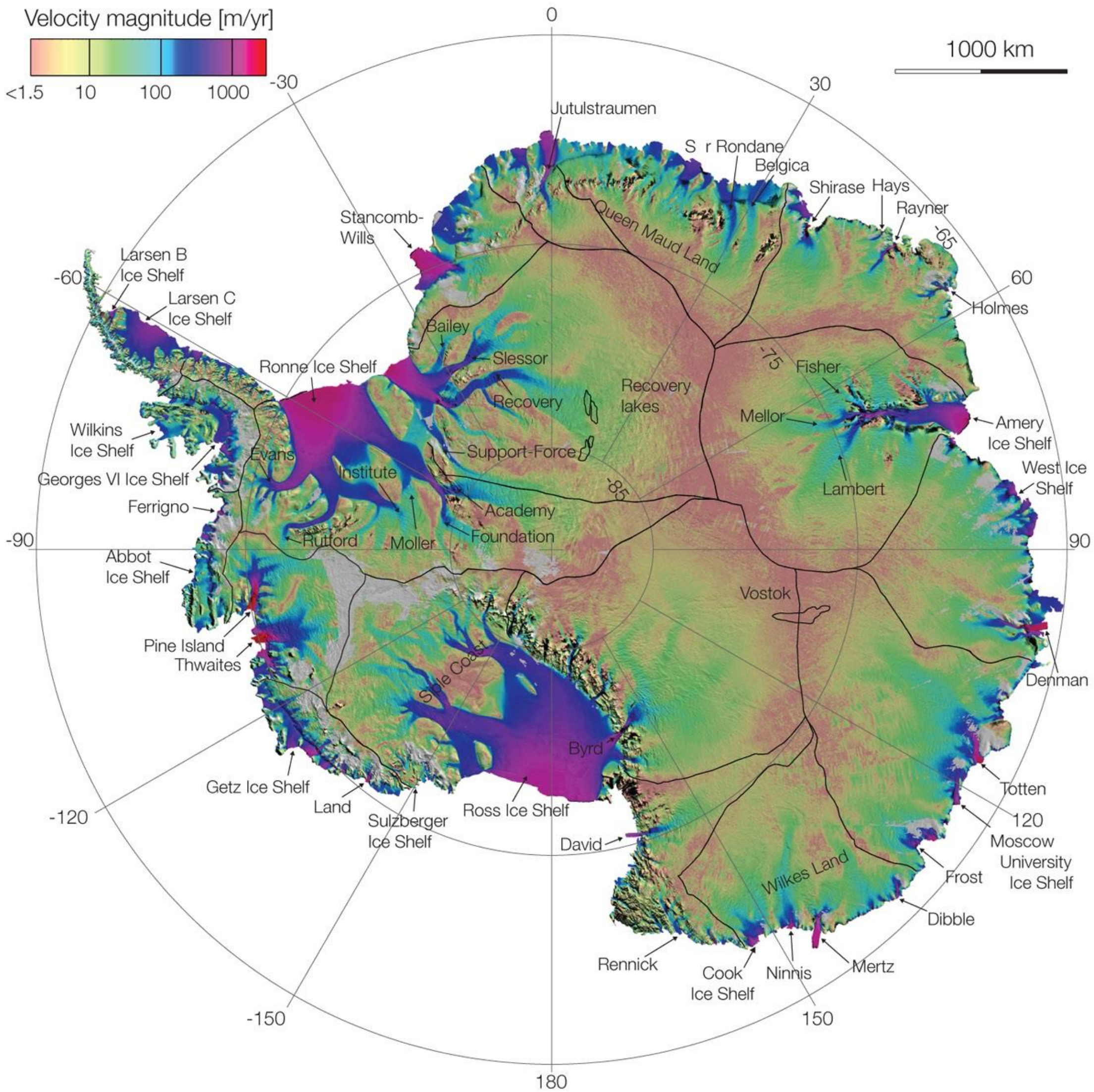
Arolla glacier: results





- Full posterior: 139 dim. parameters + 5373 dim. states
- Jointly reduced posterior: 43 dim. parameters + 50 dim. states
- Left: samples projected onto 5 leading parameter basis vectors
- Right: estimated parameter mean and credible intervals.
- Ongoing: full-scale ice sheet dynamics (on supercomputers)

Next: polar ice sheet



$$-\boldsymbol{\nabla}\cdot[2\eta(\boldsymbol{u})\,\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} - \boldsymbol{I}p] = \rho\boldsymbol{g} \quad \text{in } \Omega$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0 \quad \text{in } \Omega$$

$$\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} = \boldsymbol{0} \quad \text{on } \Gamma_t$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = 0 \quad \text{on } \Gamma_b$$

$$\boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \exp(x)\boldsymbol{T}\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_b$$

where

$$\boldsymbol{\sigma}_{\boldsymbol{u}} = -\boldsymbol{I}p + 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}$$

$$\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^\top)$$

$$\eta(\boldsymbol{u}) = \frac{1}{2}A^{-\frac{1}{n}}\left(\frac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)\right)^{\frac{1-n}{2n}}$$

Conclusions

- Parameter reduction offers a range of possibilities to design scalable solvers for inverse problems
  - We designed a likelihood-informed way to identify reduced parameters
  - With error indicators / error bound
  - Reduced parameter Bayesian estimators (approximation)
  - Discretization invariant likelihood informed MCMC (exact)
  - Randomise-then-optimize (importance sampling, exact, not presented)

- Model reduction
  - Further speed-up by making the forward model dimension independent
  - New model reduction methods, e.g., semi-blackbox, machine learning and balanced truncation?

- Ongoing work:
  - Ice sheet dynamics (energy norm minimization)
  - Time dependent problems (data assimilation)
  - Model-based data reduction

References

- Cui, Martin, Marzouk, Solonen, Spantini. "Likelihood-informed dimension reduction for nonlinear inverse probems." *Inv. Prob.* 30: 114015 (2014).
- Cui, Marzouk, Willcox. "Data-driven model reduction for the Bayesian solution of inverse problems." *Intl. J. for Num. Meth. Eng* 102: 966–990 (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).