A map-based approach to Bayesian inference in inverse problems

Youssef Marzouk & Tarek El Moselhy
Massachusetts Institute of Technology
Department of Aeronautics & Astronautics
ymarz@mit.edu
Map-based (and multiscale) approaches to Bayesian inference in inverse problems

Youssef Marzouk, Tarek El Moselhy, & Matthew Parno
Massachusetts Institute of Technology
Department of Aeronautics & Asztronautics
ymarz@mit.edu
Inverse problems

- Infer model parameters from **indirect, noisy, and limited** observations
- Problems are often ill-posed and high-dimensional

**Example:** estimate subsurface properties from observations of pressure and transport
Inverse problems

• Infer model parameters from indirect, noisy, and limited observations
• Problems are often ill-posed and high-dimensional

⇒ An essential step in predictive simulation: endow parameters and subsequent predictions with quantified uncertainties
Statistical inference

- Inversion as statistical inference: a Bayesian approach
  - Model parameters represented by random variable \( x \). Data \( d \).
  - Apply Bayes’ rule:

\[
p(x|d) = \frac{p(d|x)p(x)}{\int p(d|x)p(x)dx}
\]

- The **posterior density** \( \pi^d(x) \equiv p(x|d) \) is the full Bayesian solution to the inference problem
  - Not just a single value for \( x \), but a probability density
  - A complete description of uncertainty
  - An input to future simulations — prediction and experimental design
Bayesian inference

- **Likelihood function**: \( L(x; d) \equiv p(d|x) \)

  - *Example*: deterministic **forward model** \( G(x) \)
    
    additive measurement + model error \( \epsilon \sim p_{\epsilon} \)

    \[
    d = G(x) + \epsilon \quad \Rightarrow \quad L(x) = p_{\epsilon} (G(x) - d)
    \]
Bayesian inference

\[
p(x|d) = \frac{p(d|x)p(x)}{\int p(d|x)p(x)dx}
\]

- **Prior density** \( p(x) \):
  - Incorporates additional information: physical constraints, smoothness, structure, expert judgment; *results of previous experiments*
  - No regularization parameter *per se*
  - Flexible *hierarchical* modeling
Topics for today

1. Bayesian inference without Markov chains
2. Multiscale methods for statistical inverse problems
Bayesian inference for inverse problems

• Computational issues:
  – Forward model (PDE) appears inside the likelihood
  – Need to extract information from the posterior
  – Posterior evaluations may be expensive
  – Parameter $x$ may be high-dimensional

• Markov chain Monte Carlo (MCMC) sampling. Very useful, but…
  – Generates a stream of correlated samples
  – Proposal design is difficult; potential for poor mixing
  – No clear convergence criteria
  – Requires a large number of forward model evaluations—though surrogates and reduced models can help
  – Somewhat serial and not recursive
An alternative viewpoint

forward model $y(x)$
+ error model
+ data $d$
[i.e., likelihood]

prior knowledge

posterior knowledge

$f_X \sim p$

$Z = f(X)$

“prior random variable”

“posterior random variable”

$Z \equiv X \mid d \sim \pi^d$
Random variable transformation

Can we compute an appropriate map $f$?
Properties of the map

• Map “pushes forward” the prior measure to the posterior measure

• Potential advantages and desiderata
  – Generate *arbitrary* numbers of *independent* posterior samples, without additional forward solves
  – Analytical expressions for posterior moments
  – Clear convergence criterion?
  – Computationally less expensive? Opportunities for parallelization?
  – Propagate posterior through subsequent models? Apply recursively?
  – Compute posterior normalizing constant (*evidence* or *marginal likelihood*) for use in model selection?
Formulation

- Some notation:
  - Map is \( Z = f(X) \) \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \)
    
    *(we will discuss the functional form of \( f \) later)*

  - Forward model is \( y : \mathbb{R}^n \rightarrow \mathbb{R}^{nobs} \)

  - Start with the posterior density as

  \[
  \pi^d(z) = \frac{L(z;d)p(z)}{\beta} \propto L(z;d)p(z)
  \]

- What if we *knew* an invertible \( f \) such that \( Z = f(X) \)

- Perform a transformation from the **posterior** to the **prior** to get a probability density for \( X \)
Formulation

\[ \pi(z) = \frac{L(z; d) p(z)}{\beta} \]

\[ Z = f(X) \]

\[ q(x; f) = \frac{L(f(x); d) p(f(x))}{\beta} \left| \det \frac{\partial f}{\partial x} \right| \]

\( q(x) \) is a probability density for the prior random variable \( X \), parameterized by \( f \)
Formulation

• But we already know the density of $X$, namely the prior

$$p(x)$$

• The transformed distribution should then satisfy

$$q(x; f) = p(x)$$

• Find a map $f$ such that $q$ is close to $p$
  – For instance, minimize Kullback-Leibler (KL) divergence, Hellinger distance, etc
  – Some analogy with variational Bayes (though different)
  – Closer relation to implicit filtering [Chorin 2009–2010]
Formulation schematic

prior rv $X$

posterior rv $Z$

$\Pi^d$

Find transformation such that $p = q$

$Z = f(X)$

transformation

prior rv $X$

$q$

$q$

$p$
Formulation

• Putting q close to p…
  – Kullback-Leibler divergence

\[ D_{KL}(p\|q) = \int \log \frac{p(x)}{q(x)} p(x) dx = 0 \iff \mathbb{E}[\exp(T(X; f))] = \exp(\mathbb{E}[T(X; f)]) \]

where \( T(x; f) \equiv \log L(f(x); d) + \log p(f(x)) + \log \left| \det \frac{\partial f}{\partial x} \right| - \log p(x) \)

– Thus \( T \) must be constant in \( x \)
– Same result holds true for Hellinger and other “distances”
– \( T = constant \) also obtained by pointwise equality \( p(x) = q(x) \), but reveals the posterior normalizing constant \( \beta \)

\[ T(x; f) = \log L(f(x); d) + \log p(f(x)) + \log \left| \det \frac{\partial f}{\partial x} \right| - \log p(x) = \log \beta \]

– As a byproduct of inference, we will calculate the evidence!
Example of $T$

- In the case of a Gaussian prior (identity covariance) and additive Gaussian noise, the expression for $T(x; f)$ is

$$
L(f(x); d) = \exp \left( -\frac{1}{2} (y(f(x)) - d)^T \Sigma_n^{-1} (y(f(x)) - d) \right) \quad p(x) = \exp \left( -\frac{1}{2} x^T x \right)
$$

$$
T(x; f) = \log L(f(x); d) + \log p(f(x)) + \log \left| \det \frac{\partial f}{\partial x} \right| - \log p(x)
$$

$$
T(x; f) = -\frac{1}{2} (y(f(x)) - d)^T \Sigma_n^{-1} (y(f(x)) - d) - \frac{1}{2} f(x)^T f(x)
$$

$$
+ \log \left| \det \frac{\partial f}{\partial x} \right| + \frac{1}{2} x^T x
$$
Optimization problems

1. Minimize variance
   \[
   \min_T \text{Var} \left[ T \left( X; f \right) \right] 
   \]

2. Pointwise equality \((T = \text{constant in the } L^2 \text{ sense})\)
   \[
   T(x_i; f) = \mathbb{E}[T(X; f)], \quad i = 1 \ldots N
   \]

3. Explicitly minimize KL-divergence
   \[
   \min_f D_{KL} \left( p(x) \| q(x; f) \right)
   \]

- Note: expectation and variance are all with respect to the prior distribution \(p\)
Existence and uniqueness

- In general, map exists but is **not unique**
  - Example: linear forward model, additive Gaussian noise, zero-mean Gaussian prior:
    \[
    y(x) = Ax, \quad d = y + \epsilon, \quad X \sim N(0, I), \quad \epsilon \sim N(0, \Sigma_n)
    \]
  - Posterior distribution is Gaussian and known in closed form:
    \[
    Z \sim N\left(\mu_z, \Sigma_z\right) \quad \Sigma_z = \left(I + A^T \Sigma_n^{-1} A\right)^{-1}, \quad \mu_z = \Sigma_z A^T \Sigma_n^{-1} d
    \]
  - Any affine transformation \( Z = f(X) = \mu_z + LX \) such that \( LL^T = \Sigma_z \) represents a valid map
  - \( L \) is **not uniquely defined**: Cholesky, matrix square root, etc
Existence and uniqueness

• Connections with optimal transport theory [Caffarelli, McCann, numerous others]:
  – Use distance minimization to guarantee **existence** and **uniqueness** of an **invertible** map

\[
\min_f \mathbb{E} \left[ \left( f(X) - X \right)^T \left( f(X) - X \right) \right]
\]

  – Map is the gradient of a convex scalar function

• Full formulation becomes, e.g.,

\[
\min_f \text{Var}[T] + \lambda \mathbb{E} \left[ \left( f(X) - X \right)^T \left( f(X) - X \right) \right]
\]
Implementation

• Optimization problems
  – For \( \min_f \text{Var}[T(X; f)] \) use Newton’s method and variants
  – Take full advantage of forward adjoints and Hessians to compute derivatives with respect to degrees of freedom in \( f \)
  – Expectation/variance currently computed using samples from the prior

  – For \( T(x_i) = \text{constant} \) again use prior samples and write as nonlinear least squares
  – Use adjoint information (e.g., Gauss-Newton)
Implementation

• Represent $f$ using an orthogonal polynomial expansion (e.g., Hermite chaos)
  – Number of coefficients is equal to the dimension of the parameter space times the number of coefficients

$$f(x) = G^T \Psi(x)$$

  Vector of orthogonal polynomials

  Matrix of unknown coefficients

• Evaluation of $T$ can use efficient \textit{surrogate} forward models $y(z)$
  – Compute surrogate, for example, using polynomial chaos expansion based on propagation of prior uncertainty
Implementation

• Alternate ways of enforcing uniqueness:
  – “Triangular” dependence:
    \[ z_i = f(x_1, \ldots, x_i), \quad i \leq n \]

• Start with identity map \((posterior = prior)\)

• Continuation over total polynomial degree

• Important open question: can we represent monotone functions (or convex functions) efficiently?
Simple linear example

- 100 dimensional problem
- \( A \) is randomly generated
- Gaussian posterior:
  \[
  Z \sim N(\mu_z, \Sigma_z)
  \]
- Start iterations from identity map \( f(X) = X \)
- Convergence to exact solution in 12 iterations

\[
\begin{align*}
y(x) &= Ax, \quad d = y + \epsilon \\
X &\sim N(0, I), \quad \epsilon \sim N(0, \Sigma_n)
\end{align*}
\]

\[
A \in \mathbb{R}^{8 \times 100} \quad y, d \in \mathbb{R}^8
\]
Reaction kinetics

- Five late-time observations of $A$; truth is $k_1 = 1$, $k_2 = 2$
- Gaussian prior
- Infer $k_1$ and $k_2$

\[
\frac{dA}{dt} = -k_1 A + k_2 B
\]
\[
\frac{dB}{dt} = k_1 A - k_2 B
\]

\[
\frac{k_2}{k_1} \approx 2
\]

convergence: $D_{KL} < 10^{-5}$
Reaction kinetics: map

- 7th order polynomial map

First component of posterior

Second component of posterior
Reaction kinetics: map

- $7^{th}$ order polynomial map
- Transformation Jacobian is positive definite except at a few points (red) at the tail of the prior

First component of posterior

Second component of posterior
Cascaded maps

• Can also introduce a **sequence** of maps:
  – Gradually adjust noise variance, number of observations, or forward model fidelity
  – Final map $f$ is a **composition** of cascaded maps

\[
\sum_{n,1} > \sum_{n,2} > \ldots > \sum_{n,k} = \sum_n
\]

\[f = f_k \circ f_{k-1} \circ \ldots \circ f_2 \circ f_1; \quad z = f(x)\]

• Jacobian of transformation follows directly from the chain rule
• Density of intermediate random variable is available after each stage
Cascaded maps

• Low-order maps at each stage result in a high-order final map $f$
• Only $f_i$ is parameterized; other functions are fixed from previous stages

• Two different strategies:
  – Solve each stage exactly
    • Map captures transformation from $x_i$ to $x_{i+1}$ at each stage
  – Solve each stage approximately
    • Use solution as a guide to the next stage
    • Solve only final stage $x_k \rightarrow x_{k+1} \equiv z$ exactly
    • Potential problem which we did not observe: inaccurate solution might be a “bad guide” and complicate subsequent maps
Cascaded maps

- **Example:** estimate kinetic parameters in a *genetic toggle switch*
  - DAE model from [Gardner *et al* 2000]
  - **Real experimental data:** steady-state expression levels of one gene ($v$)

\[
\frac{du}{dt} = \frac{\alpha_1}{1 + v^\beta} - u
\]

\[
\frac{dv}{dt} = \frac{\alpha_2}{1 + w^\gamma} - v
\]

\[
w = \frac{u}{\left(1 + [IPTG]/K\right)^\eta}
\]
Cascaded maps

- **Example:** focus on two parameters \((\alpha_1, \gamma)\) with a particularly complicated distribution

![Scatter plot](image.png)
Cascaded maps

- **Example:** focus on two parameters \((\alpha_1, \gamma)\) with a particularly complicated distribution after stage 2
Cascaded maps

- **Example:** focus on two parameters \((\alpha_1, \gamma)\) with a particularly complicated distribution after stage 3
Cascaded maps

- **Example:** focus on two parameters $(\alpha_1, \gamma)$ with a particularly complicated distribution after stage 4
Cascaded maps

- **Example**: focus on two parameters $(\alpha_1, \gamma)$ with a particularly complicated distribution

![Graph showing exact posterior density](image-url)
Nonlinear PDE

- Reaction-diffusion equation
- Gaussian prior with Gaussian covariance kernel for log-diffusivity ($L_c = 0.25, \sigma = 1.25$)

$$\nabla \cdot (\kappa \nabla p) = 50p(1 - p)$$

$$\log(\kappa - \kappa_0) \sim GP(0, C)$$

$$C(r_1, r_2) = \sigma^2 \exp \left(-\frac{\|r_1 - r_2\|^2}{2L_c^2}\right)$$

- 3481 spatial elements, prior parameterized with 16 Karhunen-Loève modes, 20 observations at random locations
- Third order map
- At solution: KL-divergence = 0.0075, Var[T] = 0.0136
Results

\[ \det\left(\frac{df}{dx}\right) \]

Marginal posterior of K-L mode weights
Results

Map captures posterior dependencies among K-L mode weights
Elliptic PDE with high-dim inputs

- Elliptic PDE in two dimensions, $61^2$ spatial grid
- Log-normal prior on $\kappa(x)$ with an **exponential** covariance kernel and large variance: **58 K-L modes**
  \[ \sigma = 1.75, \quad L_c = 2 \]
- 200 observation points
- Map $f$ up to order $p = 3$

\[
\nabla \cdot \left( \kappa \nabla p \right) = 0
\]

\[
\log(\kappa - \kappa_0) \sim GP\left(0, C\right)
\]

\[
C(r_1, r_2) = \sigma^2 \exp\left(-\frac{\|r_1 - r_2\|}{L_c}\right)
\]

Posterior median and 0.1/0.9 quantiles
Elliptic PDE in high dimensions

true log-permeability

“gold-standard” MCMC (10^7 samples), 4.6 hours yet effective sample size < 2500

map/optimization, 28 minutes
Elliptic PDE in high dimensions

Effective sample size; dominant components of $10^7$-sample MCMC chain
Elliptic PDE in high dimensions

- Compare map-inference to MCMC (DRAM) at equivalent wall-clock time
Elliptic PDE in high dimensions

- Compare map-inference to MCMC (DRAM) at equivalent wall-clock time

map/optimization = asterisk; MCMC without burn-in = triangle; long chain MCMC = line
1 Bayesian inference without Markov chains

2 Multiscale methods for statistical inverse problems
Elliptic/parabolic PDEs

- Familiar example: *inhomogeneous coefficient* $k(x)$

$$p_t = \nabla \cdot (k(x) \nabla p) + s(x,t)$$

$$\nabla \cdot (k(x) \nabla p) = -f(x)$$

- $M(x) = \log k(x)$ is an unknown log-permeability; infer from a few observations of $p$ in the spatial domain $D$

- What is the “true” **dimension** of the inference problem? Can we exploit **ill-posedness** and/or **multiscale** structure?
Inference and scale

• Consider draws from **prior** (e.g., stationary Gaussian random field)

![Exponential covariance kernel](image)

![Gaussian covariance kernel](image)

• Given limited observations of pressure, how much fine scale information can we actually learn?
Multiscale inference

- Inspiration from *multiscale finite element* methods (MsFEM) [Hou, Efendiev, ...]
  - MsFEM builds coarse basis functions $\tilde{\phi}_i(x)$ that encode local variability of coefficient (e.g., permeability) at the fine scale

- Form a smaller system at the coarse scale to obtain PDE solution field (e.g., pressure)
  \[
p(x) = \sum_i p_i \tilde{\phi}_i(x)
\]

- Entries of coarse scale stiffness matrix:
  \[
a_{ij} = \sum_k \int_K k \nabla \tilde{\phi}_i \nabla \tilde{\phi}_j \, dx
\]

- Appropriately chosen coarse-scale quantities are sufficient to predict the solution field
Multiscale inference

- Multiscale methods thus provide a means of identifying **conditional independence**
  - Permeability and data (at coarse nodes) are conditionally independent given A

\[
\pi(k, A | d) \propto \pi(d | k, A) \pi(k, A) = \pi(d | A) \pi(k | A) \pi(A)
\]

- \( k \) = material property at fine scale
- \( A \) = ‘coarse’ quantities, e.g., integrals used to assemble global stiffness matrix in MsFEM
- \( d \) = data

**Redesign MCMC:** need to sample from joint posterior distribution of \((k, A)\)
- Reduced dimensionality? Better mixing? Opportunities for parallelism?
Multiscale inference

- MCMC approach
  - Propose \((k^*, A^*)\) from \(q\) and accept with probability \(\alpha = \min\{\gamma, 1\}\) where

\[
\gamma = \frac{\pi(d|A^*) \pi(k^*|A^*) \pi(A^*)}{\pi(d|A) \pi(k|A) \pi(A)} \frac{q(k|A) q(A|A^*)}{q(k^*|A^*) q(A^*|A)}
\]

- Put \(q(k|A) = \pi(k|A)\), yielding

\[
\gamma = \frac{\pi(d|A^*) \pi(A^*)}{\pi(d|A) \pi(A)} \frac{q(A|A^*)}{q(A^*|A)}
\]

Contains coarse-scale quantities only; MCMC can proceed solely at the coarse scale.

- Overall scheme thus requires
  1. Evaluating \(\pi(A)\), the prior distribution of the coarse quantity \(A\) (i.e., upscaling the prior)
  2. Sampling from \(\pi(k|A)\), where \(A\) is a nonlinear function of \(k\). Need to sample from a multivariate normal distribution conditioned on a nonlinear constraint.
Prior on the coarse scale

• Focus on *elemental integrals* used to assemble coarse stiffness matrix $A$
  
  – Construct their joint distribution using an empirical approach (for now)

\[
e_{ij} = \int_{\Omega} k \nabla \varphi_i \nabla \varphi_j \, dx
= \frac{1}{\int_{x_j}^{x_{j+1}} \frac{1}{k(x)} \, dx}
\]
Prior on the coarse scale

- Sample from prior on $k(x)$; $\{\log e_{ij}\}$ appear jointly Gaussian
Prior on the coarse scale

- Parameterize the joint distribution of the elemental integrals using an exponential-form covariance kernel
Prior on the coarse scale

- Parameterize the joint distribution of the elemental integrals using an exponential-form covariance kernel
Conditional simulation of fine scale

• Need to sample from $\pi\left(\log k \mid f(\log k) = A^{(i)}\right)$
  - $\pi(\log k)$ is Gaussian
  - Specified value of the elemental integrals (at iteration $i$ of the coarse-scale MCMC chain) provides a **nonlinear constraint**
  - Taking a “slice” of the prior

• Procedure:
  - Begin with a prior sample $k^*$
  - Use successive linearizations of the constraint to reach desired value of $f$
  - Perform MCMC along the constraint
  - “Blend” these two steps for greater efficiency
Conditional simulation of fine scale

- Raw iterative conditioning
- MCMC smoothing applied
- Analytic solution
Problem setup: *elliptic equation, Gaussian covariance kernel* on log-permeability, \( l = 0.07 \), 5 coarse elements (50 fine elements)
Multiscale inference: results

• Results of MCMC on coarse scale
Multiscale inference: results

- Now generate posterior realizations on fine scale
Comparisons with \{single-scale inference + FEM\}, \{single-scale inference + MsFEM\}
Multiscale inference: results

- Comparisons with \{single-scale inference + FEM\}, \{single-scale inference + MsFEM\}
Comparisons with \{single-scale inference + FEM\}, \{single-scale inference + MsFEM\}

posterior medians, 0.25/0.75 quantiles, and true permeability (black)
Multiscale inference

- Significant efficiency gains over single-scale sampling \textit{(so far)}
  - ESS is number of \textit{effectively independent} samples in each dimension of the chain, after 160000 MCMC steps

<table>
<thead>
<tr>
<th>Method</th>
<th>min ESS</th>
<th>med ESS</th>
<th>max ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-scale</td>
<td>33</td>
<td>45</td>
<td>60</td>
</tr>
<tr>
<td>Multiscale</td>
<td>5383</td>
<td>31676</td>
<td>38866</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>min ESS/sec</th>
<th>med ESS/sec</th>
<th>max ESS/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-scale</td>
<td>0.24</td>
<td>0.33</td>
<td>0.43</td>
</tr>
<tr>
<td>Multiscale</td>
<td>1.7</td>
<td>9.9</td>
<td>12.1</td>
</tr>
</tbody>
</table>

- These results are all obtained in \textit{serial}; yet conditional simulation of the fine scale is \textit{embarrassingly parallel}...
• “Truth” permeability from layer 5 of SPE10 data set (black)
  – exponential prior covariance, 220 grid points, 11 pressure observations
  – posterior realizations (dashed) and posterior mean (green)
Conclusions

• A new **map-based** approach to Bayesian inference
  – Find a function $f$ that pushes forward the prior measure to the posterior
  – Connections with optimal transport theory
  – Entire posterior now obtained by solving an **optimization** problem
  – Clear convergence criterion; evidence computed “for free”
  – Recursive and easily parallelizable
  – Favorable performance comparison with MCMC on ill-posed PDE-based problems

• **Many** open issues and ideas…
  – Importance sampling or stochastic expansions for $T(x; f)$
  – Better parameterizations of the map
  – *Limiting* behavior of the map in ill-posed and high-D problems
  – More efficient optimization approaches
  – *Sequential* data assimilation
Conclusions

• Multiscale framework for Bayesian inversion
  – Exploits conditional independence of observations and fine scale, given intermediate (coarse) quantities
  – Implemented with MsFEM but quite general
  – More efficient posterior sampling
  – Huge opportunities for parallelization

• Special thanks to Tarek El Moselhy, Matthew Parno (MIT)
• Support from: US Department of Energy, Office of Advanced Scientific Computing Research (ASCR), Sandia National Laboratories, KAUST Global Research Partnership