Optimisation-based Sampling Approaches for Hierarchical Bayesian Inverse Problems

Tiangang Cui\textsuperscript{1}, John Bardsley\textsuperscript{2}, Youssef Marzouk\textsuperscript{3}, Zheng Wang\textsuperscript{3}

\textsuperscript{1}Monash University
\textsuperscript{2}University of Montana
\textsuperscript{3}Massachusetts Institute of Technology

IMS, Singapore, August, 2018
Inverse Problems

Data

Parameter

\[ y = F(x) + e \]

Forward model (PDE)

Observation/model errors

Data \( y \) are limited in number, noisy, and indirect.

Parameter \( x \) is often a function, and discretised on some mesh.

Continuous, bounded, and 1st order differentiable.

Ill-posedness \( \implies \) non-uniqueness and uncertainty

\( y \in \mathbb{R}^d \)

\( x \in \mathcal{H} \)

\( F : \mathcal{H} \rightarrow \mathbb{R}^d \)
Model and predict fluid end energy transport in subsurface

- Estimating heterogenous rock properties (porosity $\phi$, permeability $k$, relative perm. $k_{r\beta}$) and boundary conditions from well observations

\[
\frac{d}{dt} \int_{\Omega} M(\cdot) \, dV = \int_{\partial\Omega} Q(\cdot) \cdot \hat{n} \, d\Gamma + \int_{\Omega} q(\cdot) \, dV
\]

- $M_m = \phi(\rho_l S_l + \rho_v S_v)$
- $Q_m = \sum_{\beta=1,v} \frac{kk_{r\beta}}{\nu_{\beta}} (\nabla p - \rho_{\beta} \hat{g})$
- $M_e = (1 - \phi)\rho_r c_r T + \phi(\rho_l u_l S_l + \rho_v u_v S_v)$
- $Q_e = \sum_{\beta=1,v} \frac{kk_{r\beta}}{\nu_{\beta}} (\nabla p - \rho_{\beta} \hat{g}) h_{\beta} - K \nabla T$
Model ice sheet movements to predict the amount of ice entering ocean

Estimating **bottom boundary conditions** from velocity measured at top

\[-\nabla \cdot [2\eta(u)\dot{\varepsilon} - I_p] = \rho g \quad \text{in } \Omega \]
\[\nabla \cdot u = 0 \quad \text{in } \Omega \]
\[\sigma_u n = 0 \quad \text{on } \Gamma_t \]
\[u \cdot n = 0 \quad \text{on } \Gamma_b \]
\[T \sigma_u n + \exp(x)Tu = 0 \quad \text{on } \Gamma_b \]

where

\[\sigma_u = -I_p + 2\eta(u)\dot{\varepsilon} \]
\[\dot{\varepsilon} = \frac{1}{2} (\nabla u + \nabla u^\top) \]
\[\eta(u) = \frac{1}{2} A - \frac{1}{n} \left( \frac{1}{2} \text{tr}(\dot{\varepsilon}^2) \right)^{\frac{1-n}{2n}} \]
A systematic new model development approach is needed

**Desiderata**

- Systematically compare all plausible kinetic models
- Penalize model complexity
- Assimilate prior information about model preference and parameter values with available data
- Not just the "best" model: quantify uncertainties in models and parameters

Bayesian statistics provides a natural framework

Bayes' rule:

\[
p(k|D|\{z\}) = \frac{p(D|k)p(k)}{p(D)} \propto \text{likelihood} \cdot \text{prior}
\]

**Prior:** Expert knowledge and smooth assumptions that defines a weighted function space for \(x\). E.g. Gaussian process, \(\pi_0(x|\alpha) = \mathcal{N}(0, \Gamma_{pr}(\alpha))\).

**Likelihood:** knowledge of \(e\), quantifies the probability of \(y\) being true for a given \(x\). E.g., assuming \(e\) follows Gaussian, \(e \sim \mathcal{N}(0, \Gamma_{obs}(\alpha))\)

\[
L(y|F(x), \alpha) \propto \exp\left(-\frac{1}{2} \|y - F(x)\|_{\Gamma^{-1}_{obs}(\alpha)}^2\right)
\]

**Hyper prior:** prior and likelihood are parametrised by \(\alpha\), e.g., correlation length and variance are unknown.
Summarize information over the posterior by calculating the expected value of function of interest

$$E_\pi [g(x)] = \int \int g(x) \pi(x, \alpha | y) \, dx \, d\alpha$$

Example: mean, variance, predicted flow rate ...

- High-dimensional (in principle, infinite dimensional) parameters $x$
- Low-dimensional $\alpha$ determines structures of posterior. For example, changing prior correlation length will result in different function spaces.

For fixed $\alpha$, we convert optimisation to construct posterior approximations

- Scalable importance sampling (computing normalising constant, etc.)
- Scalable MCMC proposals for sampling $x$
- (Possibly scalable) algorithms for explore $\pi(x, \alpha | y)$
For fixed $\alpha$, we want to explore

**Conditional Posterior:**

$$\pi(x|y, \alpha) = \frac{1}{\pi(y|\alpha)} \ L(y|F(x), \alpha) \ \pi_0(x|\alpha)$$

We want to exploit optimisation methods to build a transformation of random variables.

$$T_\# \pi_{\text{ref}} \overset{d}{=} \pi_{\text{tar}}$$

If it is exact:

$$\xi \sim \pi_{\text{ref}} \quad \xrightarrow{T} \quad x \sim \pi_{\text{tar}}$$

$$T(\xi) = x \quad S(x) := T^{-1}(x) = \xi$$
In reality, optimisation-based sampler leads to approximate transport

\[ T_{\#} \pi_{\text{ref}}(x) = |\nabla T| \pi_{\text{ref}}(S(x)) \]

\[ S_{\#} \pi_{\text{tar}}(\xi) = |\nabla S| \pi_{\text{tar}}(T(\xi)) \]
Optimisation Based Samplers

Conditional Posterior: \[ \pi(x|y, \alpha) = \frac{1}{\pi(y|\alpha)} L(y|F(x), \alpha) \pi_0(x|\alpha) \]

- Can evaluate the log of conditional posterior, up to a constant \( \pi(y|\alpha) \)
- Can evaluate actions of linearised model \( J(x) = \nabla_x F(x) \) and its adjoint.

Optimization-based samplers (RTO, RML, Implicit Filtering, ...)

1. Construct objective functions using:
   - the conditional posterior
   - a random sample from a reference distribution
2. Minimize the objective functions to get proposal samples
3. Compute log-densities of the resulting proposal samples
4. Samples (+ densities) are used as:
   - Biasing distribution in importance sampling (e.g., for \( \pi(y|\alpha) \))
   - Independent proposal in Metropolis-Hastings
Randomize-then-Optimize (RTO)

Required form of the target

$$\log \pi_{\text{tar}}(x) = -\frac{1}{2}\|H(x)\|^2 + \text{const}$$

Consider the conditional posterior with Gaussian likelihood and prior

$$\log \pi(x|y, \alpha) = -\frac{1}{2}\|F(x) - y\|_{\Gamma_{\text{obs}}^{-1}}^2 - \frac{1}{2}\|x\|_{\Gamma_{\text{pr}}^{-1}}^2 + c(\alpha)$$

$$= -\frac{1}{2}\left\| \begin{bmatrix} \Gamma_{\text{pr}}^{-\frac{1}{2}} x \\ \Gamma_{\text{obs}}^{-\frac{1}{2}} F(x) \end{bmatrix} - \begin{bmatrix} 0 \\ \Gamma_{\text{obs}}^{-\frac{1}{2}} y \end{bmatrix} \right\|^2 + c(\alpha)$$

$$= -\frac{1}{2}\|H(x)\|^2 + c(\alpha)$$

where $c(\alpha)$ is a function depending on $\alpha$. 
Randomize-then-Optimize (RTO)

**RTO’s mapping (ansatz)**

\[ \xi = S(x) := \Gamma_{\text{pr}}^{\frac{1}{2}} Q^\top H(x) \]

- Reference samples \( \xi \) are drawn from the Gaussian prior \( \pi_0(x|\alpha) \)
- Matrix \( Q \) has the same range as \( \nabla H(x) \), e.g., using QR factorization

**Algorithm**

1. Find the conditional posterior mode \( x_{\text{MAP}} \) and solve a reduced QR

\[
QR = \nabla H(x_{\text{MAP}}) = \begin{bmatrix} \Gamma_{\text{pr}}^{\frac{1}{2}} \\ \Gamma_{\text{obs}}^{\frac{1}{2}} J(x_{\text{MAP}}) \end{bmatrix}
\]

2. Solve nonlinear system \( \xi = Q^\top H(x) \) for \( x \) using non-linear optimisation, e.g., Newton-CG, subspace trust-region, interior points...

3. Compute the probability density of the resulting \( x \).

\[
q_{\text{RTO}}(x) = (2\pi)^{-\frac{n}{2}} \left| \Gamma_{\text{pr}}^{\frac{1}{2}} \right| \left| Q^\top \nabla H(x) \right| \exp \left( -\frac{1}{2} \left\| Q^\top H(x) \right\|^2 \right) \left( \Gamma_{\text{pr}}^{\frac{1}{2}} \xi \right)
\]
A different choice of $Q$

**Algorithmic modification**

Exploit the structure of $\nabla H(x_{\text{MAP}})$ to reduce computation.

Using an SVD $\tilde{J}(x_{\text{MAP}}) := \Gamma_{\text{obs}}^{-\frac{1}{2}} J(x_{\text{MAP}}) \Gamma_{\text{pr}}^{\frac{1}{2}} = \Psi \Lambda \Phi^\top$, the range of $Q$,

$$QR = \nabla H(x_{\text{MAP}}) = \begin{bmatrix} \Gamma_{\text{pr}}^{-\frac{1}{2}} \\ \Gamma_{\text{obs}}^{-\frac{1}{2}} J(x_{\text{MAP}}) \end{bmatrix} = \begin{bmatrix} I_n \\ \Psi \Lambda \Phi^\top \end{bmatrix} \Gamma_{\text{pr}}^{-\frac{1}{2}}$$

can be expressed in a decomposed form.

Use a rank-$r$ SVD of $\tilde{J}(x_{\text{MAP}})$ and a $B$ s.t. $B^\top B = \nabla H(x_{\text{MAP}})^\top \nabla H(x_{\text{MAP}})$

$$\tilde{Q} = \nabla H(x_{\text{MAP}}) B^{-1} = \begin{bmatrix} \Phi(\Lambda^2 + I_r)^{-\frac{1}{2}} \Phi^\top + (I_n - \Phi \Phi^\top) \\ \Psi \Lambda(\Lambda^2 + I_r)^{-\frac{1}{2}} \Phi^\top \end{bmatrix}$$

$\tilde{Q}$: orthogonal, same range, represented by $\Psi, \Lambda, \Phi$.

Adjoint methods can be used to find the SVD without forming $J(x_{\text{MAP}})$. 
Scalable implementation of RTO

Define a rank-\(r\) projector \(P_r = \Gamma_{pr}^{\frac{1}{2}} \Phi \Phi^\top \Gamma_{pr}^{\frac{1}{2}}\), solve \(\xi = \Gamma_{pr}^{\frac{1}{2}} \tilde{Q}^\top H(x)\):

\[
\begin{align*}
(I_n - P_r)\xi &= (I_n - P_r)x, \\
P_r\xi &= \Gamma_{pr}^{\frac{1}{2}} \Phi \left[ (\Lambda^2 + I_r)^{-\frac{1}{2}} \Phi^\top \Gamma_{pr}^{\frac{1}{2}} x + \Lambda (\Lambda^2 + I_r)^{-\frac{1}{2}} \Psi^\top \Gamma_{obs}^{-1} (F(x) - y) \right] \\
&= \Gamma_{pr}^{\frac{1}{2}} \Phi G(P_r x + (I_n - P_r)x)
\end{align*}
\]

System of equations splits in two. Optimize only for \(r\)-dim. part.

Simplify the calculation of the proposal density:

\[
\left| \tilde{Q}^\top \nabla H(x) \right|_{n \times n} = \left| (\Lambda^2 + I_r)^{-\frac{1}{2}} \right| \cdot \left| I_r + \Gamma_{obs}^{-\frac{1}{2}} \Lambda \Psi^\top J(x) \Gamma_{pr}^{\frac{1}{2}} \right|_{r \times r}
\]

Find determinant of an \(r \times r\) matrix. \textit{Reduced from} \(n \times n\) matrix.

Takeaway

- RTO’s mapping keeps most parameter directions fixed.
- Directions that move \textit{depending} on those that are fixed.
- \(\rho_{\text{RTO}}(x)\) is normalised.
Linear model $y = Fx + e$, using rank $r = 0$:
Linear model $y = Fx + e$, using rank $r = 1$: 

![Diagram showing linear model with prior, posterior, and RTO proposal distributions](image)
Truncated SVD for a linear model

Linear model $y = Fx + e$, using rank $r = 2$:
Truncated SVD for a nonlinear model

Nonlinear model $y = F(x) + e$, using rank $r = 0$:
Nonlinear model $y = F(x) + e$, using rank $r = 1$:
Truncated SVD for a nonlinear model

Nonlinear model $y = F(x) + e$, using rank $r = 2$: 

![Graph showing nonlinear model with prior, posterior, and RTO proposal distributions.](image)
Does RTO's mapping yield well-defined density in function space?

\[
\begin{cases}
(I - P) \xi = (I - P) x, \\
P \xi = \Gamma^{\frac{1}{2}} \Pr \Phi G (P x + (I - P) x)
\end{cases}
\]

**Well-posedness**

1. Assume $\mathcal{G}$ is Lipschitz continuous, injective, and its inverse is Lipschitz continuous.

2. The RTO density $\rho_{\text{RTO}}(x)$ is absolute continuous w.r.t. prior

3. Sufficient to show the ratio (or the Radon–Nikodym derivative):

\[
\frac{\pi(x|y, \alpha)}{\rho_{\text{RTO}}(x)}
\]

is positive almost surely at the infinite dimensional limit.

The acceptance probability of M-H and the weights of importance sampling will not decay to zero as parameter dimension increases.
Example: Poisson Equation Inverse Problem

Estimate diffusion coefficient $x(s)$ from measurements of the solution $u(s)$ of

$$-rac{d}{ds} \left( x(s) \frac{du}{ds} \right) = f(s), \quad 0 < s < 1,$$

- We generate data as point observation of the $u(s), s = s_1, s_2, \ldots, s_d$.
- Nonlinear forward model: $y = F(x) + e$. 
Example: Poisson Equation Inverse Problem

Estimate diffusion coefficient $x(s)$ from measurements of the solution $u(s)$ of

$$-rac{d}{ds} \left( x(s) \frac{du}{ds} \right) = f(s), \quad 0 < s < 1,$$

- We generate data as point observation of the $u(s), s = s_1, s_2, \ldots, s_d$.
- Nonlinear forward model: $y = F(x) + e$.

Numerical check:

1. Draw a realization $\xi \in \mathcal{H}$ from the prior (Gaussian process).
2. Discretise this sample $\xi$ on multiple grids.
3. Apply RTO’s prior-to-proposal mapping to each discretised sample.
4. For fixed grid size, apply RTO to sample posteriors with different observation noise level.
RTO, varying parameter dimension, \( n \)

**ESS**

![ESS graph]

**Acceptance Rate**

![Acceptance Rate graph]

**Variance of Log-importance-weight**

![Variance of Log-importance-weight graph]

**Optimization Iterations**

![Optimization Iterations graph]
RTO, varying parameter dimension, $n$

$n = 161$

$n = 641$

$n = 2561$

$n = 10241$
RTO, varying noise level, $\sigma$

- **Numerical ESS**:
  - $10^{-7}$ to $10^1$
  - Values: 4000, 4000, 3000, 3000, 2000, 2000, 1000

- **Acceptance Rate**:
  - $10^{-7}$ to $10^1$
  - Values: 1.0, 1.0, 1.0, 0.8, 0.6, 0.4, 0.2

- **Optimization Iterations per MCMC Step**:
  - Values: 600, 400, 400, 300, 200, 100, 0
Example: Poisson Equation Inverse Problem

Changing parameter dimensions:

1. MCMC acceptance rate and ESS scale with parameter dimension.
2. Variance of log-importance-weights scales with parameter dimension.
3. The number of optimisation iterations stabilises with parameter refinement.

Change observation noise level (posterior concentrates):

1. MCMC acceptance rate and ESS scale with observation noise level.
2. The number of optimisation iterations stabilises with decreasing noise level.

In both cases, the acceptance rate is high.
Hierarchical Bayes

Apply RTO to a simple hierarchical posterior $\alpha = (\delta, \lambda)$

Prior and Hyper Prior

Prior: $\pi_0(x|\delta) \propto \exp \left( -\frac{\delta}{2} \|x\|_2^2 \Gamma_{pr}^{-1} \right)$,

where $\pi_0(x|\delta) = \mathcal{N}(0, (\delta)^{-1} \Gamma_{pr})$.

Gamma hyper priors: $\pi_0(\lambda) \propto \lambda^{a_\lambda-1} \exp(-b_\lambda \lambda)$,
$\pi_0(\delta) \propto \delta^{a_\delta-1} \exp(-b_\delta \delta)$,

where $a_\lambda = a_\delta = 1$ and $b_\lambda = b_\delta = 10^{-4}$.

Posterior

Cond. post: $\pi(x|y, \lambda, \delta) = \frac{1}{\pi(\lambda, \delta|y)} \exp \left( -\frac{\lambda}{2} \|F(x) - y\|_{\Gamma_{obs}}^2 - \frac{\delta}{2} \|x\|_2^2 \Gamma_{pr}^{-1} \right)$

Post: $\pi(x, \lambda, \delta|y) \propto \pi(\lambda, \delta|y) \pi(x|y, \lambda, \delta) \pi_0(\lambda) \pi_0(\delta)$. 
RTO within Hierarchical Gibbs

Algorithm

0. Choose $x_0$, and set $k = 0$;

1. Sample from $\pi(\lambda, \delta|y, x_k)$ via conjugate distributions:
   
   a. $\lambda_{k+1} \sim \Gamma \left( m/2 + \alpha_{\lambda}, \frac{1}{2} \|F(x_k) - y\|_2^2 \Gamma_{\text{obs}}^- + \beta_{\lambda} \right)$;
   
   b. $\delta_{k+1} \sim \Gamma \left( n/2 + \alpha_{\delta}, \frac{1}{2} \|x_k\|_L^2 + \beta_{\delta} \right)$;

2. Sample from $\pi(x|y, \lambda_{k+1}, \delta_{k+1})$ using RTO-MH:
   
   a. compute $x_* \sim q_{\text{RTO}}(x; \lambda_{k+1}, \delta_{k+1})$;
   
   b. set $x_{k+1} = x_*$ with probability
      
      $$ r = \min \left( 1, \frac{\pi(x_*|y, \lambda_{k+1}, \delta_{k+1}) \rho_{\text{RTO}}(x_k)}{\pi(x_k|y, \lambda_{k+1}, \delta_{k+1}) \rho_{\text{RTO}}(x_*)} \right), $$
      
      else set $x_{k+1} = x_k$.

3. Set $k = k + 1$ and return to step 1.
Sample from $\pi(x, \lambda, \delta | y)$ using RTO-within-Gibbs

$x$-samples

Histograms of $\lambda$, $\delta$, and $x_7$
Conclusion

RTO for fixed hyper-parameters:
- Generates explicit maps from prior to a density close to posterior
- Dimension independent
- Rather restrictive assumptions (modified model $\mathcal{G}$ is injective)
  
  Develop trust region methods to safe guard this.
  Develop nonlinear preconditioners to satisfy the assumption.

For sampling hyper-parameters:
- RTO-within-Gibbs is not dimension independent!
  
  Use RTO to obtain marginal posterior and apply pseudo-marginal
  Non-centred RTO-within-Gibbs
Future works:

- Use RTO to map quadrature points and QMC points.
- Multilevel MC for accelerating the optimisation and variance reduction.
- Extension to non-Gaussian likelihood.
- Applications in subsurface and ice sheet flow.
- Applications in ODEs and SDEs.