Bayesian Methods and Uncertainty Quantification for Nonlinear Inverse Problems

John Bardsley, University of Montana

Collaborators:

H. Haario, J. Kaipio, M. Laine, Y. Marzouk, A. Seppänen, A. Solonen, Z. Wang

Technical University Denmark, December 2016

Outline

- Nonlinear Inverse Problems Setup
- Randomize-then-Optimize (RTO)
- Test Cases:
    - small # of parameters examples
    - electrical impedance tomography
    - ℓ_1 priors, i.e., TV and Besov priors

Now Consider a Nonlinear Statistical Model

Now assume the non-linear, Gaussian statistical model

$$\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\epsilon},$$

where

- $\mathbf{y} \in \mathbb{R}^m$ is the vector of observations;
- $\mathbf{x} \in \mathbb{R}^n$ is the vector of unknown parameters;
- $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ is nonlinear;
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}_m)$, i.e., $\boldsymbol{\epsilon}$ is i.i.d. Gaussian with mean 0 and variance λ^{-1}.

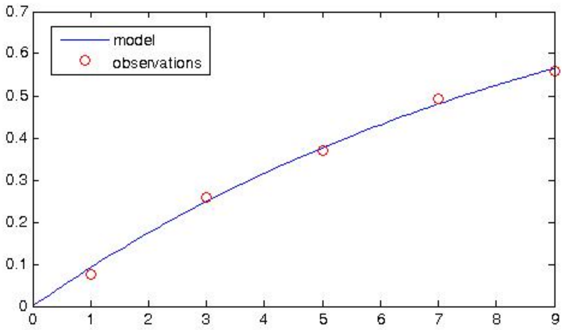
Toy example

Consider the following nonlinear, two-parameter **pre-whitened** model.

$$y_i = x_1(1 - \exp(-x_2 t_i)) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, 3, 4, 5,$$

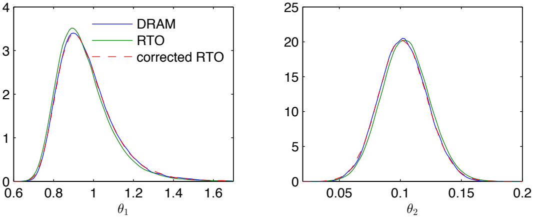
with $t_i = 2i - 1$, $\sigma = 0.0136$, and $\mathbf{y} = [.076, .258, .369, .492, .559]$.

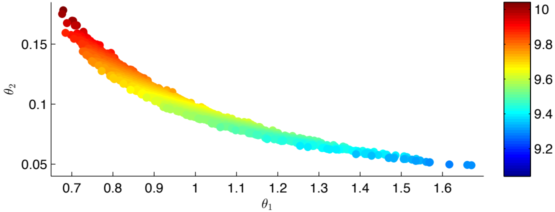
GOAL: estimate a probability distribution for $\mathbf{x} = (x_1, x_2)$.



Toy example continued: the Bayesian posterior $p(x_1, x_2|\mathbf{y})$

$$p(x_1|\mathbf{y}) = \int_{x_2} p(x_1, x_2|\mathbf{y})dx_2 \qquad p(x_2|\mathbf{y}) = \int_{x_1} p(x_1, x_2|\mathbf{y})dx_1$$





$$p(x_1, x_2|\mathbf{y})$$

Compute Samples Using Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a framework for sampling from a (potentially un-normalized) probability distribution.

Some Classical MCMC algorithms

- Gibbs sampling (talk 1: for sampling from $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$)
- Metropolis-Hastings
- Adaptive Metropolis (talk 1: for sampling from $p(\lambda, \delta|\mathbf{y})$)

- Inverse Problems: high-dimensional posterior
- Posterior is harder to explore with classical algorithms
- Chains become more correlated, sampling becomes inefficient

Metropolis-Hastings

Definitions:

$p(\mathbf{x}\|\mathbf{y})$	posterior (target) density
$\mathbf{x}^k$	random variable of the Markov chain at step k
$q(\mathbf{x}^*\|\mathbf{x}^k)$	proposal density given $\mathbf{x}^k$
$\mathbf{x}^*$	random variable from the proposal

A chain of samples $\{\mathbf{x}^0, \mathbf{x}^1, \cdots\}$ is generated by:

1. Start at $\mathbf{x}^0$
2. For $k = 1, 2, \cdots K$
   - 2.1 sample $\mathbf{x}^* \sim q(\mathbf{x}^*|\mathbf{x}^{k-1})$
   - 2.2 calculate $\alpha = \min\left\{\frac{p(\mathbf{x}^*|\mathbf{y})q(\mathbf{x}^{k-1}|\mathbf{x}^*)}{p(\mathbf{x}^{k-1}|\mathbf{y})q(\mathbf{x}^*|\mathbf{x}^{k-1})}, 1\right\}$
   - 2.3 $\mathbf{x}^k = \begin{cases} \mathbf{x}^* & \text{with probability } \alpha \\ \mathbf{x}^{k-1} & \text{with probability } 1-\alpha \end{cases}$

Metropolis-Hastings Demonstration:

http://chifeng.scripts.mit.edu/stuff/mcmc-demo/

chifeng.scripts.mit.edu/stuff/mcmc-demo/

Randomize-then-Optimize (RTO): defines a proposal q

Assumption: RTO requires that the posterior to have least squares form, i.e.,

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}}\|^2\right).$$

Randomize-then-Optimize (RTO): defines a proposal q

Assumption: RTO requires that the posterior to have least squares form, i.e.,

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}}\|^2\right).$$

Given that the likelihood function has the form

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}(\mathbf{x}) - \mathbf{y}\|^2\right),$$

for which priors $p(\mathbf{x})$ will the posterior density function

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

have least squares form?

Test Case 1: Uniform prior

In small parameter cases, it is often true that

$$p(\mathbf{y}|\mathbf{x}) = 0 \quad \text{for} \quad \mathbf{x} \notin \Omega.$$

Then we can choose as a prior $p(\mathbf{x})$ defined by

$$\mathbf{x} \sim U(\Omega),$$

where U denotes the multivariate uniform distribution.

Test Case 1: Uniform prior

In small parameter cases, it is often true that

$$p(\mathbf{y}|\mathbf{x}) = 0 \quad \text{for} \quad \mathbf{x} \notin \Omega.$$

Then we can choose as a prior $p(\mathbf{x})$ defined by

$$\mathbf{x} \sim U(\Omega),$$

where U denotes the multivariate uniform distribution.

Then $p(\mathbf{x})$ = constant on Ω, and we have

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &\propto \exp\left(-\frac{1}{2}\|\mathbf{A}(\mathbf{x}) - \mathbf{y}\|^2\right). \end{aligned}$$

$\star$ Thus can use RTO to sample from $p(\mathbf{x}|\mathbf{y})$.

Test Case 2: Gaussian prior

When a Gaussian prior is chosen,

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right),$$

the posterior can also be written in least squares form:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &\propto \exp\left(-\frac{1}{2}\left\|\mathbf{A}(\mathbf{x})-\mathbf{y}\right\|^2-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right) \end{aligned}$$

Test Case 2: Gaussian prior

When a Gaussian prior is chosen,

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right),$$

the posterior can also be written in least squares form:

$$\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\
&\propto \exp\left(-\frac{1}{2}\left\|\mathbf{A}(\mathbf{x})-\mathbf{y}\right\|^2-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right) \\
&= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix}\mathbf{A}(\mathbf{x})\\ \mathbf{L}^{1/2}\mathbf{x}\end{bmatrix}-\begin{bmatrix}\mathbf{y}\\ \mathbf{L}^{1/2}\mathbf{x}_0\end{bmatrix}\right\|^2\right) \\
&\stackrel{\text{def}}{=} \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{x})-\bar{\mathbf{y}}\|^2\right),
\end{aligned}$$

⋆ Thus we can use RTO to sample from $p(\mathbf{x}|\mathbf{y})$.

Extension of optimization-based approach to nonlinear problems: Randomized maximum likelihood

Recall that when $\bar{\mathbf{A}}$ is linear, we can sample from $p(\mathbf{x}|\mathbf{y})$ via:

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\bar{\mathbf{A}}(\boldsymbol{\psi}) - (\bar{\mathbf{y}} + \boldsymbol{\epsilon})\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m+n}).$$

Comment: For nonlinear models, this is called *randomized maximum likelihood.*

Problem: It is an open question what the probability of $\mathbf{x}$ is.

Extension to nonlinear problems

As in the linear case, we create a nonlinear mapping

$$\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n).$$

Extension to nonlinear problems

As in the linear case, we create a nonlinear mapping

$$\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n).$$

What are $\mathbf{Q}$ and $\mathbf{F}$? First, define

$$\mathbf{x}_{\text{MAP}} = \arg\min_{\mathbf{x}} \|\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}}\|^2,$$

then first-order optimality yields

$$\mathbf{J}(\mathbf{x}_{\text{MAP}})^T (\bar{\mathbf{A}}(\mathbf{x}_{\text{MAP}}) - \bar{\mathbf{y}}) = \mathbf{0}.$$

So $\mathbf{x}_{\mathrm{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{x}_{\mathrm{MAP}})^T \bar{\mathbf{A}}(\mathbf{x}) = \mathbf{J}(\mathbf{x}_{\mathrm{MAP}})^T \bar{\mathbf{y}}.$$

So $\mathbf{x}_{\mathrm{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{x}_{\mathrm{MAP}})^T \bar{\mathbf{A}}(\mathbf{x}) = \mathbf{J}(\mathbf{x}_{\mathrm{MAP}})^T \bar{\mathbf{y}}.$$

QR-rewrite: this equation can be equivalently expressed

$$\mathbf{Q}^T \bar{\mathbf{A}}(\mathbf{x}) = \mathbf{Q}^T \bar{\mathbf{y}},$$

where $\mathbf{J}(\mathbf{x}_{\mathrm{MAP}}) = \mathbf{QR}$ is the 'thin' $\mathbf{QR}$ factorization of $\mathbf{J}(\mathbf{x}_{\mathrm{MAP}})$.

So $\mathbf{x}_{\text{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{x}_{\text{MAP}})^T \bar{\mathbf{A}}(\mathbf{x}) = \mathbf{J}(\mathbf{x}_{\text{MAP}})^T \bar{\mathbf{y}}.$$

QR-rewrite: this equation can be equivalently expressed

$$\mathbf{Q}^T \bar{\mathbf{A}}(\mathbf{x}) = \mathbf{Q}^T \bar{\mathbf{y}},$$

where $\mathbf{J}(\mathbf{x}_{\text{MAP}}) = \mathbf{Q}\mathbf{R}$ is the 'thin' $\mathbf{Q}\mathbf{R}$ factorization of $\mathbf{J}(\mathbf{x}_{\text{MAP}})$.

Nonlinear mapping: define $\mathbf{F} \stackrel{\text{def}}{=} \mathbf{Q}^T \bar{\mathbf{A}}$ and

$$\begin{aligned} \mathbf{x} &= \mathbf{F}^{-1}\left(\mathbf{Q}^T(\bar{\mathbf{y}} + \boldsymbol{\epsilon})\right), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m+n}) \\ &\stackrel{\text{def}}{=} \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n). \end{aligned}$$

RTO: use optimization to compute $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v})$

Compute a sample $\mathbf{x}$ from the RTO proposal $q(\mathbf{x})$:

1. Randomize: compute $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n)$;
2. Optimize: solve

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\mathbf{F}(\boldsymbol{\psi}) - \mathbf{v}\|^2$$

3. Reject $\mathbf{x}$ when $\mathbf{v}$ is not in the range of $\mathbf{F}$.

RTO: use optimization to compute $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v})$

Compute a sample $\mathbf{x}$ from the RTO proposal $q(\mathbf{x})$:

1. Randomize: compute $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n)$;
2. Optimize: solve

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\mathbf{F}(\boldsymbol{\psi}) - \mathbf{v}\|^2$$

3. Reject $\mathbf{x}$ when $\mathbf{v}$ is not in the range of $\mathbf{F}$.

Comment: steps 1 & 2 can be equivalently expressed

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\mathbf{Q}^T(\bar{\mathbf{A}}(\boldsymbol{\psi}) - (\bar{\mathbf{y}} + \boldsymbol{\epsilon}))\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m+n}).$$

PDF for $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}_n)$

First, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}_n)$ implies $p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)$.

PDF for $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n)$

First, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}_n)$ implies $p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T \bar{\mathbf{y}}\|^2\right)$.

Next we need $\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{n \times n}$ to be invertible. Then

$$\begin{aligned} q(\mathbf{x}) &\propto \left|\det\left(\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x})\right)\right| p_{\mathbf{v}}(\mathbf{F}(\mathbf{x})) \\ &= \left|\det\left(\mathbf{Q}^T \mathbf{J}(\mathbf{x})\right)\right| \exp\left(-\frac{1}{2}\|\mathbf{Q}^T(\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}})\|^2\right) \end{aligned}$$

PDF for $\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}_n)$

First, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}_n)$ implies $p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)$.

Next we need $\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ to be invertible. Then

$$
\begin{aligned}
q(\mathbf{x}) &\propto \left|\det\left(\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x})\right)\right| p_{\mathbf{v}}(\mathbf{F}(\mathbf{x})) \\
&= \left|\det\left(\mathbf{Q}^T\mathbf{J}(\mathbf{x})\right)\right| \exp\left(-\frac{1}{2}\|\mathbf{Q}^T(\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}})\|^2\right) \\
&= \left|\det\left(\mathbf{Q}^T\mathbf{J}(\mathbf{x})\right)\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}})\|^2\right) \\
&\qquad \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{x}) - \bar{\mathbf{y}}\|^2\right) \\
&= c(\mathbf{x})p(\mathbf{x}|\mathbf{y}),
\end{aligned}
$$

where the columns of $\bar{\mathbf{Q}}$ are orthonormal and $C(\bar{\mathbf{Q}}) \perp C(\mathbf{Q})$.

Theorem (RTO probability density)

Let $\bar{\mathbf{A}} : \mathbb{R}^n \to \mathbb{R}^{m+n}$, $\bar{\mathbf{y}} \in \mathbb{R}^{m+n}$, *and assume*

- $\bar{\mathbf{A}}$ *is continuously differentiable;*
- $\mathbf{J}(\mathbf{x}) \in \mathbb{R}^{(m+n)\times n}$ *is rank* n *for every* $\mathbf{x}$*;*
- $\mathbf{Q}^T\mathbf{J}(\mathbf{x})$ *is invertible for all relevant* $\mathbf{x}$*.*

Then the random variable

$$\mathbf{x} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}_n),$$

has probability density function

$$q(\mathbf{x}) \propto c(\mathbf{x})p(\mathbf{x}|\mathbf{y}),$$

where

$$c(\mathbf{x}) = \left|\det(\mathbf{Q}^T\mathbf{J}(\mathbf{x}))\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{y}} - \bar{\mathbf{A}}(\mathbf{x}))\|^2\right),$$

where the columns of $\bar{\mathbf{Q}}$ *are orthonormal and* $C(\bar{\mathbf{Q}}) \perp C(\mathbf{Q})$*.*

RTO Metropolis-Hastings

Definitions:

$p(\mathbf{x}\|\mathbf{y})$	posterior (target) density
$\mathbf{x}^k$	random variable of the Markov chain at step k
$q(\mathbf{x}^*)$	RTO (independence) proposal density
$\mathbf{x}^*$	random variable from the proposal

A chain of samples $\{\mathbf{x}^0, \mathbf{x}^1, \cdots\}$ is generated by:

1. Start at $\mathbf{x}^0$
2. For $k = 1, 2, \cdots K$
   2.1 sample $\mathbf{x}^* \sim q(\mathbf{x}^*)$ from the RTO proposal density
   2.2 calculate $\alpha = \min\left\{\frac{p(\mathbf{x}^*|\mathbf{y})q(\mathbf{x}^{k-1})}{p(\mathbf{x}^{k-1}|\mathbf{y})q(\mathbf{x}^*)}, 1\right\}$
   2.3 $\mathbf{x}^k = \begin{cases}\mathbf{x}^* & \text{with probability } \alpha \\ \mathbf{x}^{k-1} & \text{with probability } 1-\alpha\end{cases}$

Metropolis-Hastings using RTO

Given $\mathbf{x}^{k-1}$ and proposal $\mathbf{x}^* \sim q(\mathbf{x})$, accept with probability

$$\begin{aligned} r &= \min\left(1, \frac{p(\mathbf{x}^*|\mathbf{y})q(\mathbf{x}^{k-1})}{p(\mathbf{x}^{k-1}|\mathbf{y})q(\mathbf{x}^*)}\right) \\ &= \min\left(1, \frac{p(\mathbf{x}^*|\mathbf{y})c(\mathbf{x}^{k-1})p(\mathbf{x}^{k-1}|\mathbf{y})}{p(\mathbf{x}^{k-1}|\mathbf{y})c(\mathbf{x}^*)p(\mathbf{x}^*|\mathbf{y})}\right) \\ &= \min\left(1, \frac{c(\mathbf{x}^{k-1})}{c(\mathbf{x}^*)}\right), \end{aligned}$$

where recall that

$$c(\mathbf{x}) = \left|\det(\mathbf{Q}^T\mathbf{J}(\mathbf{x}))\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{y}} - \bar{\mathbf{A}}(\mathbf{x}))\|^2\right).$$

Metropolis-Hastings using RTO, Cont.

T **The RTO Metropolis-Hastings Algorithm**

1. Choose $\mathbf{x}^0 = \mathbf{x}_{\text{MAP}}$ and number of samples N. Set $k = 1$.
2. Compute an RTO sample $\mathbf{x}^* \sim q(\mathbf{x}^*)$.
3. Compute the acceptance probability

$$r = \min\left(1, \frac{c(\mathbf{x}^{k-1})}{c(\mathbf{x}^*)}\right).$$

4. With probability r, set $\mathbf{x}^k = \mathbf{x}^*$, else set $\mathbf{x}^k = \mathbf{x}^{k-1}$.
5. If $k < N$, set $k = k + 1$ and return to Step 2.

Understanding RTO (thanks to Zheng Wang)

Consider the simple, scalar 'inverse problem':

$$\underbrace{y}_{\text{observation}} = \overbrace{f(x)}^{\text{forward model}} + \underbrace{\epsilon}_{\text{noise}}, \quad x \sim N(0,1), \quad \epsilon \sim N(0,1)$$

$$\begin{aligned}
\underbrace{p(x|y)}_{\text{posterior}} &\propto \exp\left(-\frac{1}{2}\left(f(x)-y\right)^2\right)\exp\left(-\frac{1}{2}x^2\right) \\
&\propto \exp\left(-\frac{1}{2}\Bigg\| \underbrace{\begin{bmatrix} x \\ f(x) \end{bmatrix}}_{\bar{\mathbf{A}}(x)} - \underbrace{\begin{bmatrix} 0 \\ y \end{bmatrix}}_{\bar{\mathbf{y}}} \Bigg\|^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left\|\bar{\mathbf{A}}(x) - \bar{\mathbf{y}}\right\|^2\right)
\end{aligned}$$

Understanding RTO

Least-squares form:

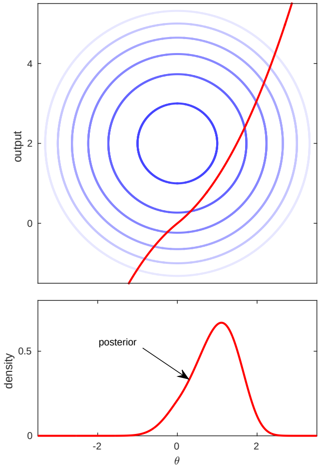
$$p(x|y) \propto \exp\left(-\frac{1}{2} \Big\| \underbrace{\begin{bmatrix} x \\ f(x) \end{bmatrix}}_{\bar{\mathbf{A}}(x)} - \underbrace{\begin{bmatrix} 0 \\ y \end{bmatrix}}_{\bar{\mathbf{y}}} \Big\|^2 \right)$$

$p(x|y)$ is the height of the path

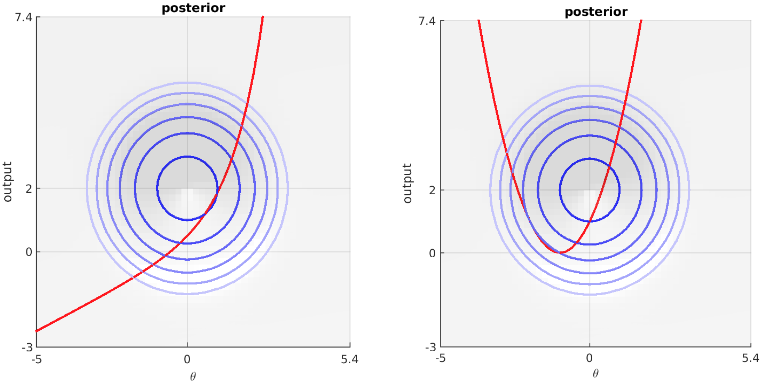
$$\bar{\mathbf{A}}(x) = [x, f(x)]^T$$

on the Gaussian

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ y \end{bmatrix}, \mathbf{I}_2 \right).$$

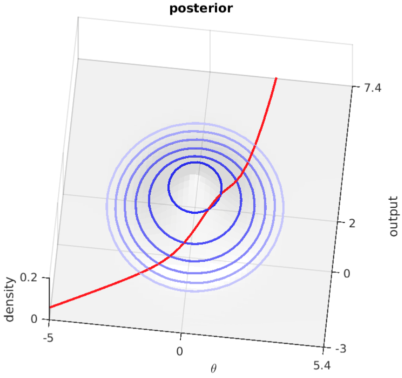


Understanding RTO



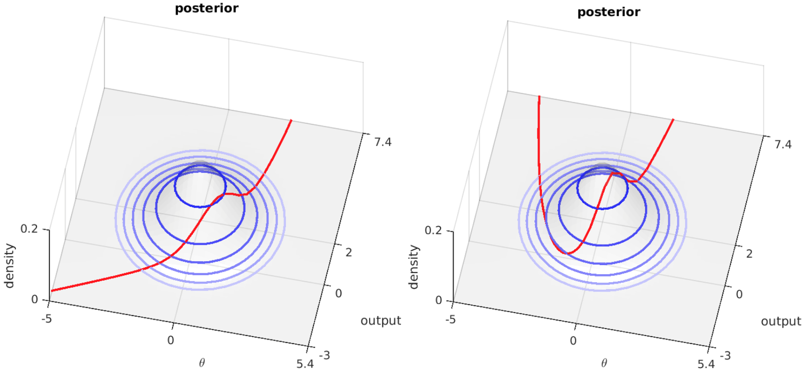
Algorithm's task: sample from the posterior

Understanding RTO



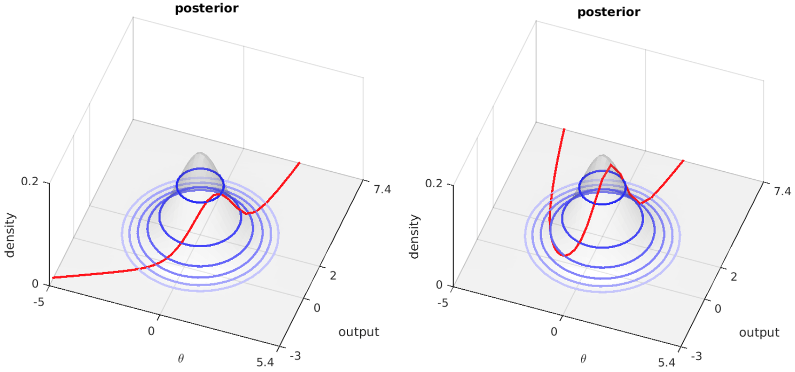
Algorithm's task: sample from the posterior

Understanding RTO



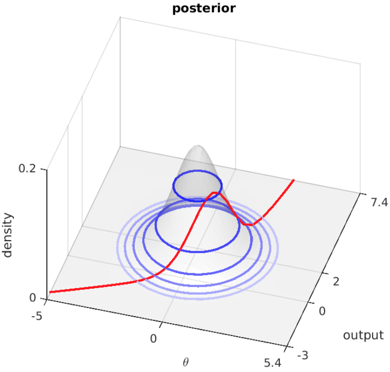
Algorithm's task: sample from the posterior

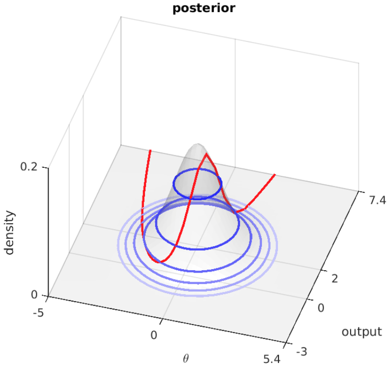
Understanding RTO



Algorithm's task: sample from the posterior

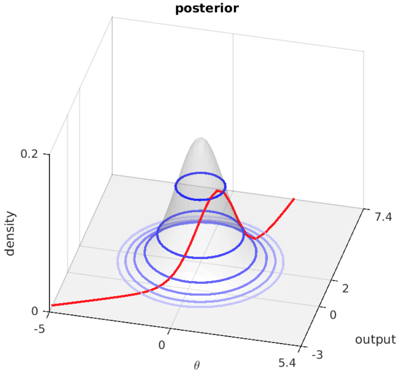
Understanding RTO

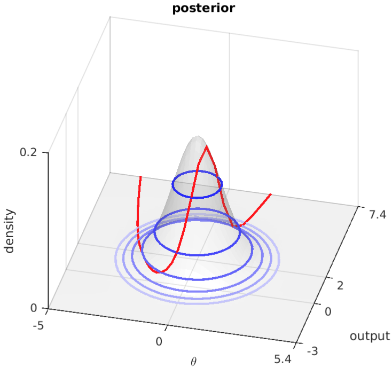




Algorithm's task: sample from the posterior

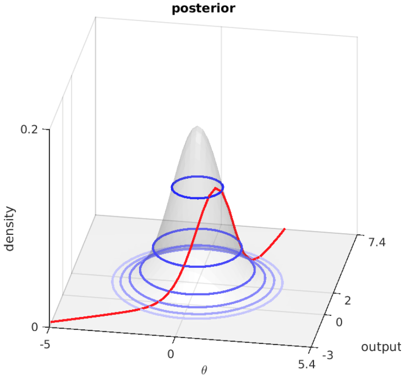
Understanding RTO





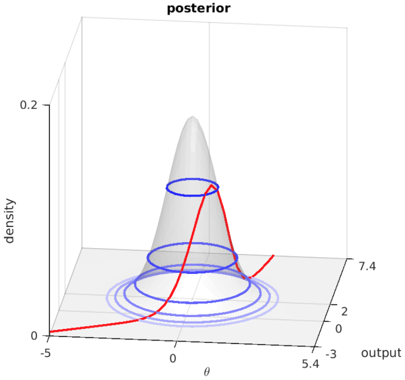
Algorithm's task: sample from the posterior

Understanding RTO



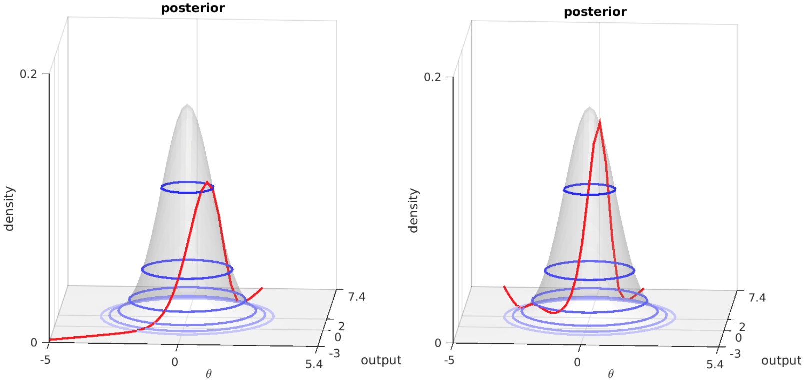
Algorithm's task: sample from the posterior

Understanding RTO



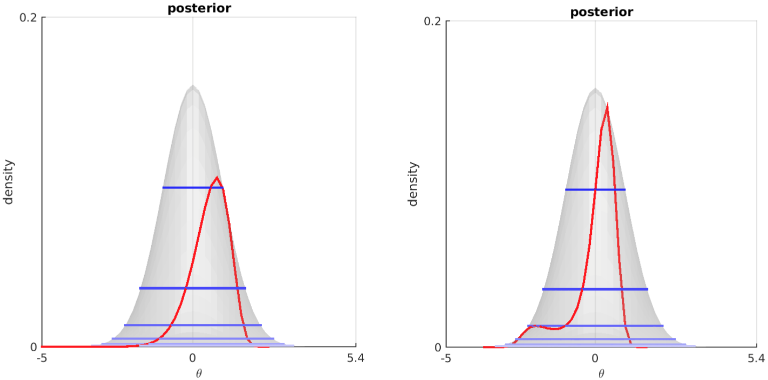
Algorithm's task: sample from the posterior

Understanding RTO



Algorithm's task: sample from the posterior

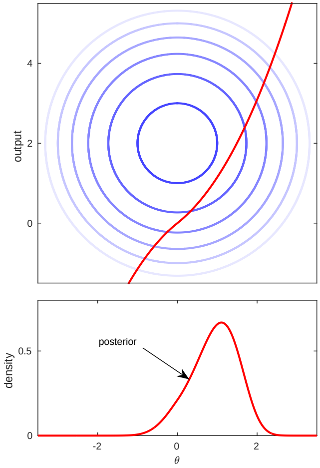
Understanding RTO



Algorithm's task: sample from the posterior

Randomize-then-optimize

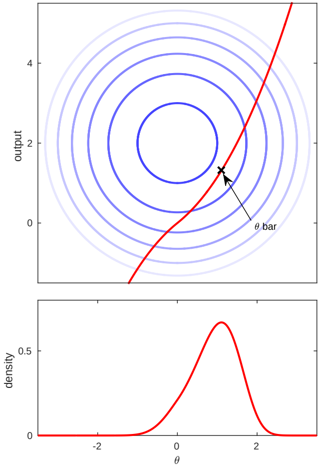
Generate RTO samples $\{x^k\}$:



Randomize-then-optimize

Generate RTO samples $\{x^k\}$:

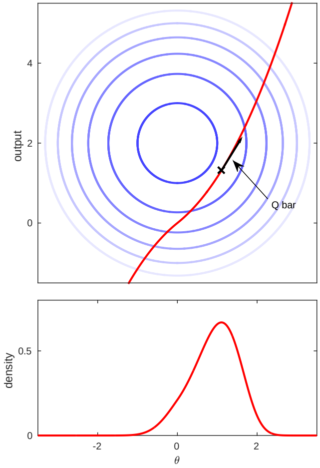
1. Compute x_{MAP}.



Randomize-then-optimize

Generate RTO samples $\{x^k\}$:

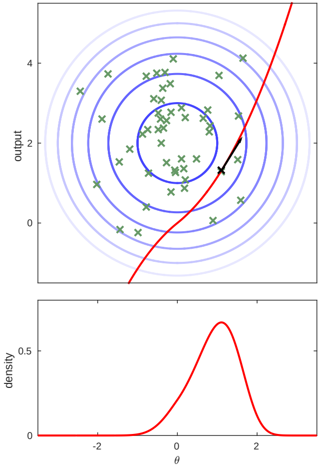
1. Compute x_{MAP}.
2. Compute $\mathbf{Q} = \mathbf{J}(x_{\text{MAP}})/\|\mathbf{J}(x_{\text{MAP}})\|$.



Randomize-then-optimize

Generate RTO samples $\{x^k\}$:

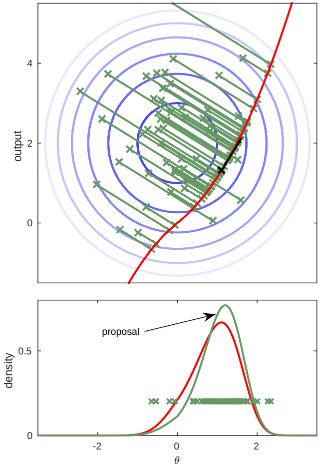
1. Compute x_{MAP}.
2. Compute $\mathbf{Q} = \mathbf{J}(x_{\mathrm{MAP}})/\|\mathbf{J}(x_{\mathrm{MAP}})\|$.
3. For $k = 1, 2, \cdots, K$
   3.1 Sample $\boldsymbol{\xi} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{I}_2)$



Randomize-then-optimize

Generate RTO samples $\{x^k\}$:

1. Compute x_{MAP}.
2. Compute $\mathbf{Q} = \mathbf{J}(x_{\text{MAP}})/\|\mathbf{J}(x_{\text{MAP}})\|$.
3. For $k = 1, 2, \cdots, K$
   3.1 Sample $\boldsymbol{\xi} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{I}_2)$
   3.2 Compute $x^k = \arg\min_x \left\| \mathbf{Q}^T \left(\bar{\mathbf{A}}(x) - \boldsymbol{\xi} \right) \right\|^2$.



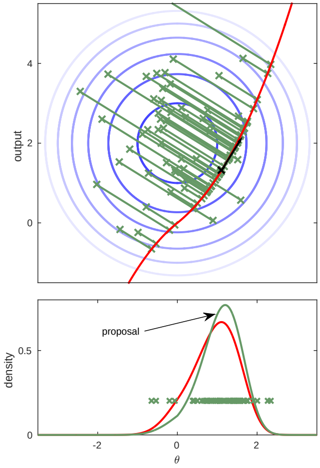
Randomize-then-optimize

Generate RTO samples $\{x^k\}$:

1. Compute x_{MAP}.
2. Compute $\mathbf{Q} = \mathbf{J}(x_{\text{MAP}})/\|\mathbf{J}(x_{\text{MAP}})\|$.
3. For $k = 1, 2, \cdots, K$
   - 3.1 Sample $\boldsymbol{\xi} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{I}_2)$
   - 3.2 Compute $x^k = \arg\min_x \left\|\mathbf{Q}^T\left(\bar{\mathbf{A}}(x) - \boldsymbol{\xi}\right)\right\|^2$.

RTO proposal density:

$$q(x^k) \propto \left|\mathbf{Q}^T\mathbf{J}(x^k)\right| \exp\left(-\frac{1}{2}\left\|\mathbf{Q}^T\left(\bar{\mathbf{A}}(x^k) - \bar{\mathbf{y}}\right)\right\|^2\right)$$



Uniform prior test cases

Choose prior $p(\mathbf{x})$ defined by

$$\mathbf{x} \sim U(\Omega),$$

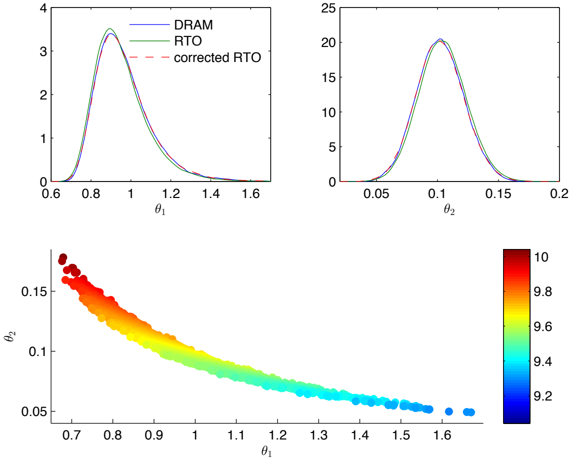
where U is a multivariate uniform distribution on Ω. Then $p(\mathbf{x}) = \text{constant}$ on Ω, and we have

$$p(\mathbf{x}|\mathbf{y}) \quad \propto \quad \exp\left(-\frac{1}{2}\|\mathbf{A}(\mathbf{x}) - \mathbf{y}\|^2\right).$$

$\star$ Thus can use RTO to sample from $p(\mathbf{x}|\mathbf{y})$.

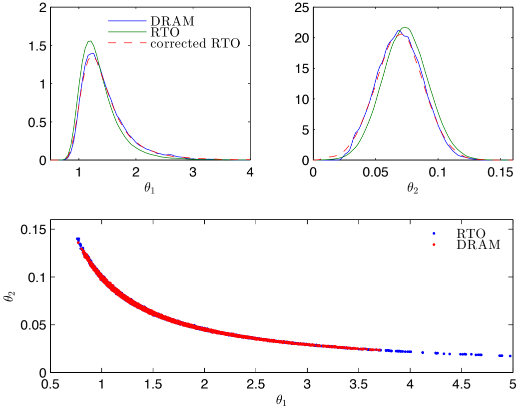
BOD, Good: $\mathbf{A}(x_1, x_2) = x_1(1 - \exp(-x_2\mathbf{t}))$

- $\mathbf{t} = 20$ linearly spaced observations in $1 \le x \le 9$;
- $\mathbf{y} = \mathbf{A}(x_1, x_2) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 0.01$;
- $[x_1, x_2] = [1, 0.1]^T$.



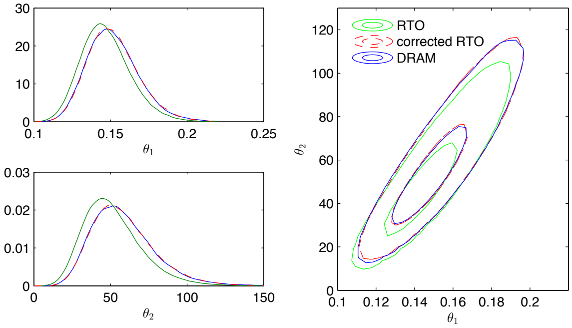
BOD, Bad: $\mathbf{A}(x_1, x_2) = x_1(1 - \exp(-x_2\mathbf{t}))$

- $\mathbf{t}$ = 20 linearly spaced observations in $1 \leq x \leq 5$;
- $\mathbf{y} = \mathbf{A}(x_1, x_2) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 0.01$;
- $[x_1, x_2] = [1, 0.1]^T$.

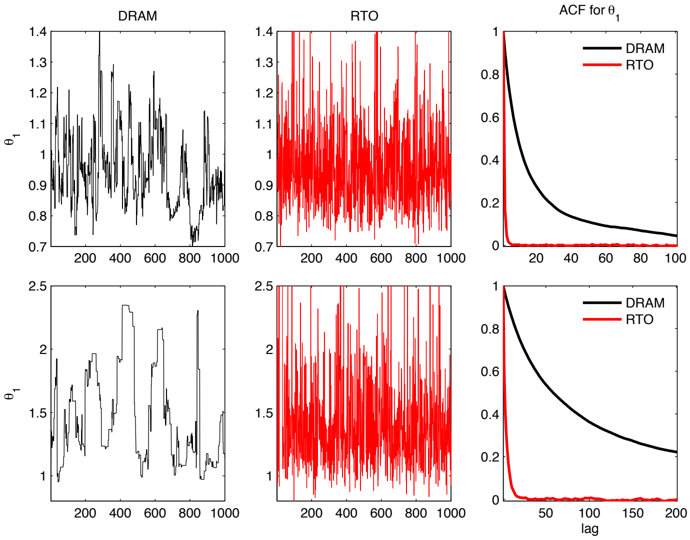


MONOD: $\mathbf{A}(x_1, x_2) = x_1\mathbf{t}/(x_2 + \mathbf{t})$

$$\begin{aligned}\mathbf{t} &= [28, 55, 83, 110, 138, 225, 375]^T \\ \mathbf{y} &= [0.053, 0.060, 0.112, 0.105, 0.099, 0.122, 0.125]^T.\end{aligned}$$



Autocorrelation plots for x_1 for Good and Bad BOD



Gaussian prior test case

When a Gaussian prior is chosen,

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right),$$

the posterior can be written in least squares form:

$$\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}\|\mathbf{A}(\mathbf{x})-\mathbf{y}\|^2 - \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\mathbf{L}(\mathbf{x}-\mathbf{x}_0)\right) \\
&= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix}\mathbf{A}(\mathbf{x}) \\ \mathbf{L}^{1/2}\mathbf{x}\end{bmatrix} - \begin{bmatrix}\mathbf{y} \\ \mathbf{L}^{1/2}\mathbf{x}_0\end{bmatrix}\right\|^2\right) \\
&\stackrel{\text{def}}{=} \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{x})-\bar{\mathbf{y}}\|^2\right).
\end{aligned}$$

$\star$ Thus we can use RTO to sample from $p(\mathbf{x}|\mathbf{y})$.

Electrical Impedance Tomography Seppänen, Solonen, Haario, Kaipio

$$\nabla \cdot (x \nabla \varphi) = 0, \qquad \vec{r} \in \Omega$$

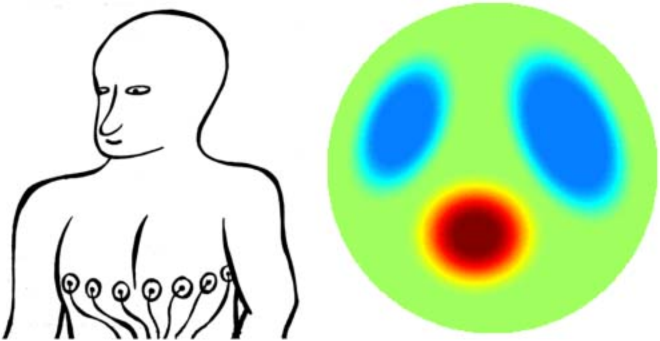
$$\varphi + z_\ell x \frac{\partial \varphi}{\partial \vec{n}} = y_\ell, \quad \vec{r} \in e_\ell,\ \ell = 1, \ldots, L$$

$$\int_{e_\ell} x \frac{\partial \varphi}{\partial \vec{n}} \mathrm{d}S = I_\ell, \quad \ell = 1, \ldots, L$$

$$x \frac{\partial \varphi}{\partial \vec{n}} = 0, \qquad \vec{r} \in \partial\Omega \setminus \cup_{\ell=1}^{L} e_\ell$$

- $x = x(\vec{r})$ & $\varphi = \varphi(\vec{r})$: electrical conductivity & potential.
- $\vec{r} \in \Omega$: spatial coordinate.
- e_ℓ: area under the ℓth electrode.
- z_ℓ: contact impedance between ℓth electrode and object.
- y_ℓ & I_ℓ: amplitudes of the electrode potential and current.
- $\vec{n}$: outward unit normal
- L: number of electrodes.

EIT, Forward/Inverse Problem (image by Siltanen)



Left: current $\mathbf{I}$ and electrode potential $\mathbf{y}$; Right: conductivity $\mathbf{x}$.

Forward Problem: Given the conductivity $\mathbf{x}$, compute

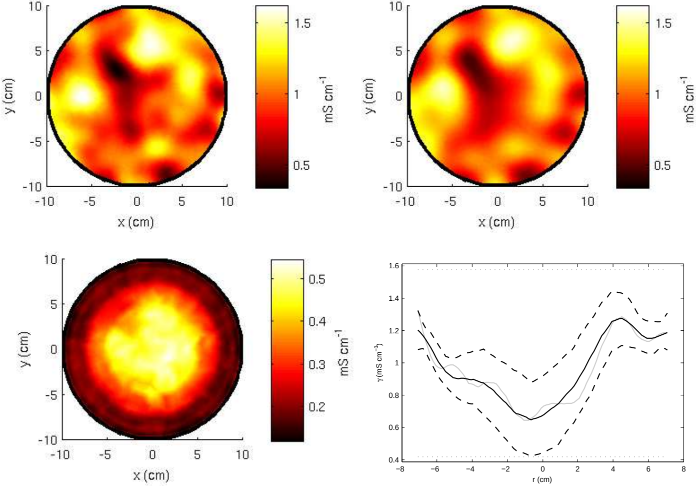
$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon}.$$

Evaluating $\mathbf{f}(\mathbf{x})$ requires solving a complicated PDE.

Inverse Problem: Given $\mathbf{y}$, construct the posterior density $p(\mathbf{x}|\mathbf{y})$.

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Realization from Smoothness Prior

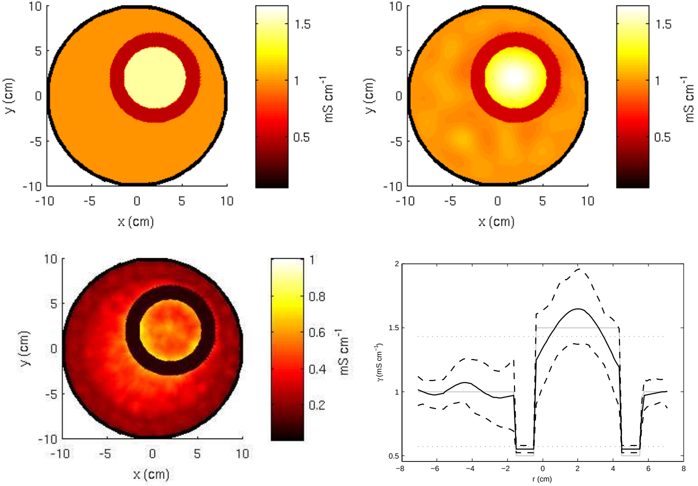


Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Internal Structure #1

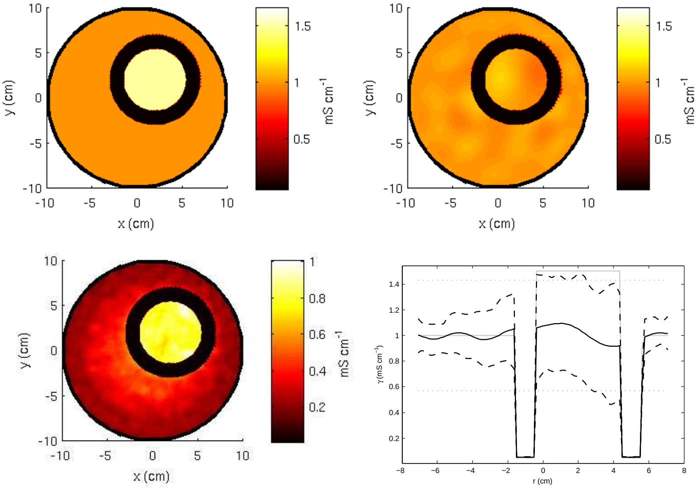


Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Internal Structure #2



Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

Laplace (Total Variation and Besov) Priors

Finally, we consider the ℓ_1 prior case:

$$p(\mathbf{x}) \propto \exp\left(-\delta\|\mathbf{D}\mathbf{x}\|_1\right),$$

where $\mathbf{D}$ is an invertible matrix. Then the posterior then takes the form

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\mathbf{A}(\mathbf{x}) - \mathbf{y}\|^2 - \delta\|\mathbf{D}\mathbf{x}\|_1\right).$$

Note that total variation in one-dimension and the Besov $B^s_{1,1}$-space priors in one- and higher-dimensions have this form.

Laplace (Total Variation and Besov) Priors

Finally, we consider the ℓ_1 prior case:

$$p(\mathbf{x}) \propto \exp\left(-\delta\|\mathbf{D}\mathbf{x}\|_1\right),$$

where $\mathbf{D}$ is an invertible matrix. Then the posterior then takes the form

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\mathbf{A}(\mathbf{x}) - \mathbf{y}\|^2 - \delta\|\mathbf{D}\mathbf{x}\|_1\right).$$

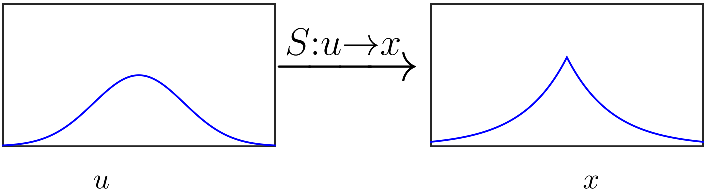
Note that total variation in one-dimension and the Besov $B^s_{1,1}$-space priors in one- and higher-dimensions have this form.

⋆ But $p(\mathbf{x}|\mathbf{y})$ does <u>not</u> have least squares form.

Prior Transformation for ℓ_1 Priors

Main idea: Transform the problem to one that RTO can solve

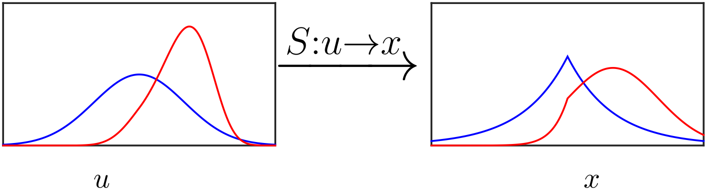
- Define a map between a **reference** parameter u and the **physical** parameter x.
- Choose the mapping so that the prior on u is Gaussian.
- Sample from the transformed posterior, in u, using RTO, then transform the samples back.



Prior Transformation for ℓ_1 Priors

Main idea: Transform the problem to one that RTO can solve

- Define a map between a **reference** parameter u and the **physical** parameter x.
- Choose the mapping so that the prior on u is Gaussian.
- Sample from the transformed posterior, in u, using RTO, then transform the samples back.



The One-Dimensional Transformation

The prior transformation is analytic and is defined

$$x = S(u) \stackrel{\text{def}}{=} \mathrm{F}^{-1}_{p(x)}\left(\varphi(u)\right),$$

where

- $\mathrm{F}^{-1}_{p(x)}$ is the inverse-CDF of the L^1-type prior $p(x)$;
- φ is the CDF of a standard Gaussian.

The One-Dimensional Transformation

The prior transformation is analytic and is defined

$$x = S(u) \stackrel{\text{def}}{=} \mathrm{F}^{-1}_{p(x)}\left(\varphi(u)\right),$$

where

- $\mathrm{F}^{-1}_{p(x)}$ is the inverse-CDF of the L^1-type prior $p(x)$;
- φ is the CDF of a standard Gaussian.

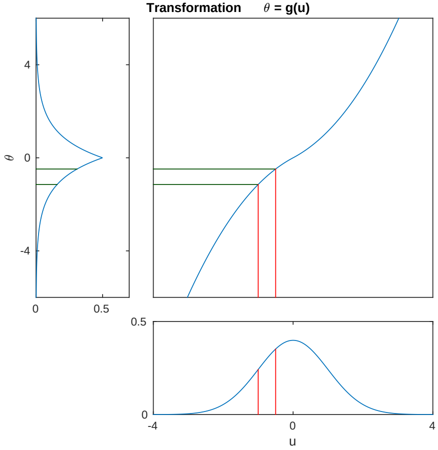
Then the posterior density $p(x|y)$ can be expressed in terms of r:

$$\begin{aligned} p(S(u)|y) &\propto \exp\left(-\frac{1}{2}(f\left(S(u)\right) - y)^2 - \frac{1}{2}u^2\right) \\ &= \exp\left(-\frac{1}{2}\left\| \begin{bmatrix} f\left(S(u)\right) \\ u \end{bmatrix} - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2\right) \end{aligned}$$

Prior Transformation: 1D Laplace Prior

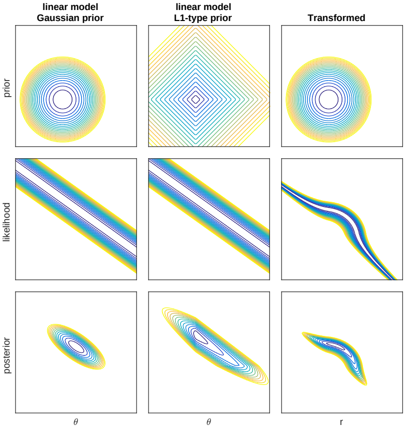
$p(x) \propto \exp\left(-\lambda|x|\right)$

$p(u) \propto \exp\left(-\frac{1}{2}u^2\right)$



For multiple independent x_i, transformation is repeated

2D Laplace Prior



Transformation moves complexity from prior to likelihood

Laplace Priors in Higher-Dimensions

1. Define a change of variables

$$\mathbf{D}\mathbf{x} = S(\mathbf{u})$$

such that the transformed prior is a standard Gaussian, i.e.,

$$p(\mathbf{D}^{-1}S(\mathbf{u})) \propto \exp\left(-\frac{\delta}{2}\|\mathbf{u}\|_2^2\right).$$

Laplace Priors in Higher-Dimensions

1. Define a change of variables

$$\mathbf{D}\mathbf{x} = S(\mathbf{u})$$

such that the transformed prior is a standard Gaussian, i.e.,

$$p(\mathbf{D}^{-1}S(\mathbf{u})) \propto \exp\left(-\frac{\delta}{2}\|\mathbf{u}\|_2^2\right).$$

2. Sample from the transformed posterior, with respect to $\mathbf{u}$,

$$p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{D}^{-1}S(\mathbf{u}))p(\mathbf{D}^{-1}S(\mathbf{u}));$$

Laplace Priors in Higher-Dimensions

1. Define a change of variables

$$\mathbf{D}\mathbf{x} = S(\mathbf{u})$$

   such that the transformed prior is a standard Gaussian, i.e.,

$$p(\mathbf{D}^{-1}S(\mathbf{u})) \propto \exp\left(-\frac{\delta}{2}\|\mathbf{u}\|_2^2\right).$$

2. Sample from the transformed posterior, with respect to $\mathbf{u}$,

$$p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{D}^{-1}S(\mathbf{u}))p(\mathbf{D}^{-1}S(\mathbf{u}));$$

3. Transform the samples back via $\mathbf{x} = \mathbf{D}^{-1}S(\mathbf{u})$.

Test Case 3, L^1-type priors: High-Dimensional Problems

The transformed posterior, with $\mathbf{D}$ an invertible matrix, takes the form

$$\begin{aligned} p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}(f\left(\mathbf{D}^{-1}S(\mathbf{u})\right)-\mathbf{y})^2-\frac{1}{2}\mathbf{u}^2\right) \\ &= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix}\mathbf{A}\left(\mathbf{D}^{-1}S(\mathbf{u})\right) \\ \mathbf{u}\end{bmatrix}-\begin{bmatrix}\mathbf{y} \\ \mathbf{0}\end{bmatrix}\right\|^2\right), \end{aligned}$$

where

$$S(\mathbf{u}) = (S(u_1), \ldots, S(u_n))$$

as defined above.

Test Case 3, L^1-type priors: High-Dimensional Problems

The transformed posterior, with $\mathbf{D}$ an invertible matrix, takes the form

$$\begin{aligned} p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}(f\left(\mathbf{D}^{-1}S(\mathbf{u})\right) - \mathbf{y})^2 - \frac{1}{2}\mathbf{u}^2\right) \\ &= \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} \mathbf{A}\left(\mathbf{D}^{-1}S(\mathbf{u})\right) \\ \mathbf{u} \end{bmatrix} - \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}\right\|^2\right), \end{aligned}$$

where

$$S(\mathbf{u}) = (S(u_1), \ldots, S(u_n))$$

as defined above.

$\star$ $p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y})$ is in least squares form with respect to $\mathbf{u}$ so we can apply RTO!

RTO Metropolis-Hastings to Sample from $p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y})$

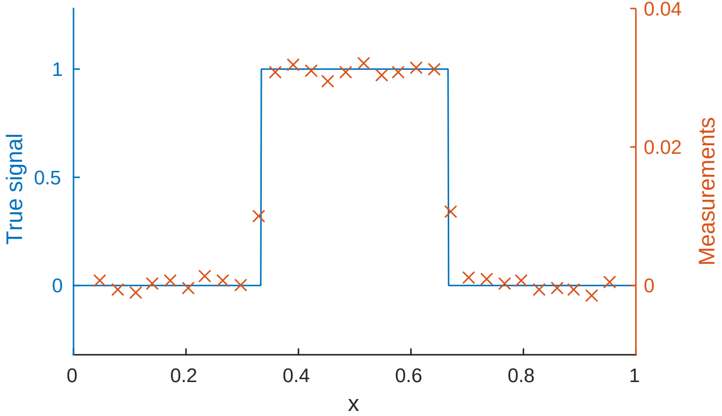
T **The RTO Metropolis-Hastings Algorithm**

1. Choose $\mathbf{u}^0 = \mathbf{u}_{\text{MAP}} = \arg\min_{\mathbf{u}} p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y})$ and number of samples N. Set $k = 1$.
2. Compute an RTO sample $\mathbf{u}^* \sim q(\mathbf{u}^*)$.
3. Compute the acceptance probability

$$r = \min\left(1, \frac{c(\mathbf{u}^{k-1})}{c(\mathbf{u}^*)}\right).$$

4. With probability r, set $\mathbf{u}^k = \mathbf{u}^*$, else set $\mathbf{u}^k = \mathbf{u}^{k-1}$.
5. If $k < N$, set $k = k + 1$ and return to Step 2.

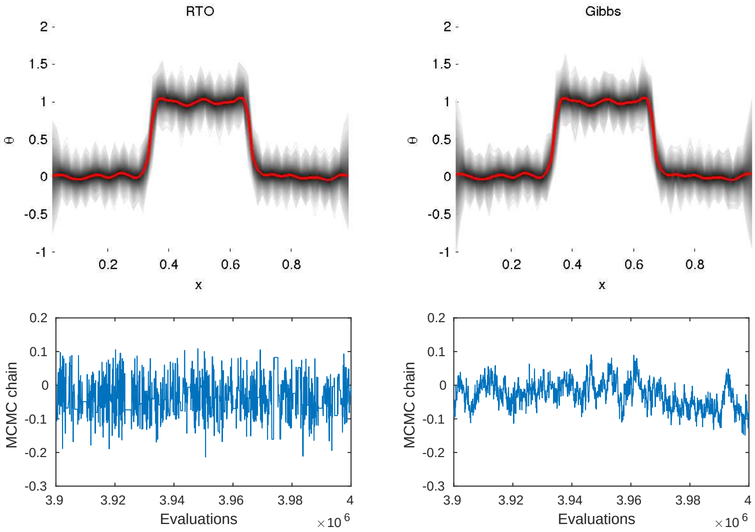
Deconvolution of a Square Pulse w/ TV Prior



$$\mathbf{x} \in \mathbb{R}^{63} \qquad \mathbf{y} \in \mathbb{R}^{32}$$

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \delta\|\mathbf{D}\mathbf{x}\|_1\right)$$

Deconvolution of a Square Pulse w/ TV Prior



2D elliptic PDE inverse problem

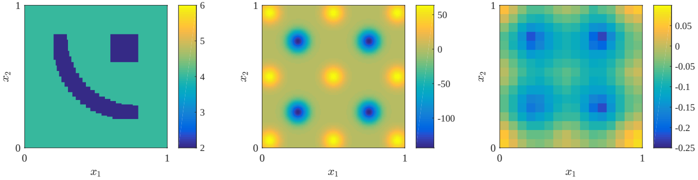
$$-\nabla \cdot (\exp(x(t))\nabla y(t)) = h(t), \quad t \in [0,1]^2,$$

with boundary conditions

$$\exp(x(t))\nabla y(t) \cdot \vec{n}(t) = 0.$$

After discretization, this defines the model

$$\mathbf{y} = \mathbf{A}(\mathbf{x}).$$



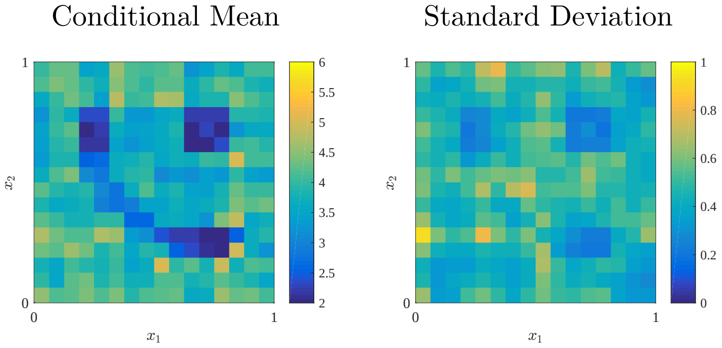
$\mathbf{x}_{\text{true}}$ $\mathbf{h}$ $\mathbf{y}$

2D PDE inverse problem: mean and STD

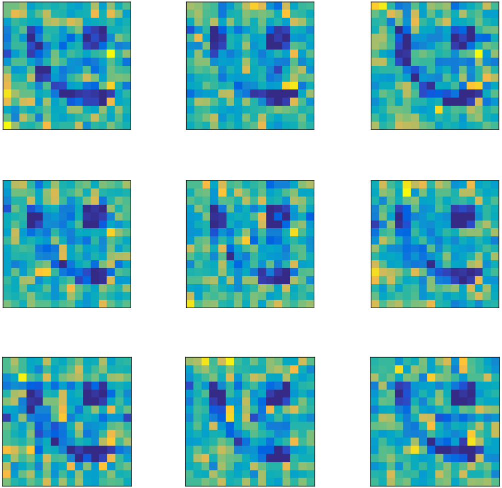
Use RTO-MH to sample from the transformed posterior:

$$p(\mathbf{D}^{-1}S(\mathbf{u})|\mathbf{y}) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}(\mathbf{D}^{-1}S(\mathbf{u})) - \mathbf{y}\|_2^2 - \delta\|\mathbf{u}\|^2\right),$$

where $\mathbf{D}$ is a wavelet transform matrix, then transform the samples back via $\mathbf{x} = \mathbf{D}^{-1}S(\mathbf{u})$.



2D PDE inverse problem: Samples



Conclusions/Takeaways

- The development of computationally efficient MCMC methods for nonlinear inverse problems is challenging.
- RTO was presented as a proposal mechanism within Metropolis-Hastings.
- RTO was described in some detail and then test on several examples, including EIT and ℓ_1 priors such as total variation.