Bayesian inference for complex models

Lecture 1

Aretha Teckentrup

School of Mathematics, University of Edinburgh

SFB1294 Data assimilation Spring school - March 21, 2022

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

General overview of the course

This course will provide an introduction to the Bayesian approach to inverse problems, including introduction of the Bayesian formulation, computational challenges and related algorithms.

A rough outline is as follows:

- Lecture 1: Introduction
- Lecture 2: Algorithms in finite dimensions
- Lecture 3: Algorithms in infinite dimensions
- Lecture 4: Exercises (Jupyter notebook)

Outline of today's lecture

1. What are inverse problems?
2. Bayesian approach
3. Markov chain Monte Carlo methods

What are inverse problems?

Definition and applications

- An inverse problem is concerned with determining causal factors from observed data.
- In mathematical terms, we want to determine model inputs based on (partial and noisy) observations of model outputs.

What are inverse problems?

Definition and applications

- An inverse problem is concerned with determining causal factors from observed data.
- In mathematical terms, we want to determine model inputs based on (partial and noisy) observations of model outputs.
- Inverse problems appear in many different areas, including:
  - *computational imaging*: recovering the true image from a blurred and noisy observation;
  - *geophysics*: inferring the conductivity of the subsurface from measurements at wells;
  - *machine learning*: building an underlying model from observed data points,
  - *non-destructive testing, astrophysics, medicine, weather prediction, ...*

What are inverse problems?

Mathematical formulation

- Suppose we are given a mathematical model of a process, represented by a (linear or non-linear) map $\mathcal{G}$.
- We are interested in the following inverse problem: given observational data $y \in Y$, determine unknown $u \in U$ such that

$$y = \mathcal{G}(u) + \eta.$$

- Here, η represents observational noise, due to for example measurement error (inaccurate instruments).

What are inverse problems?

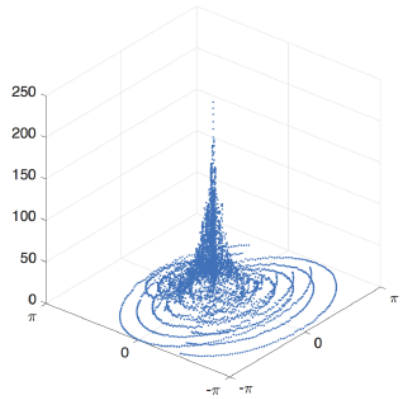
Example 1: Imaging

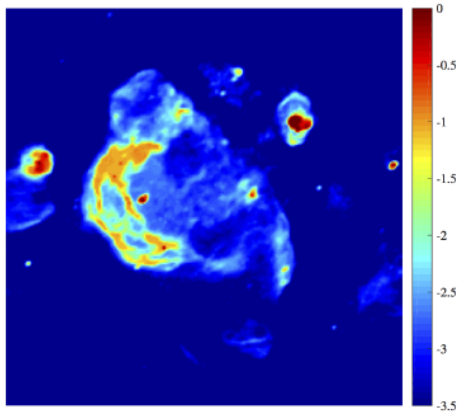
- **Goal**: reconstruct an image u given a noisy, partial observation y

What are inverse problems?

Example 1: Imaging

- **Goal**: reconstruct an image u given a noisy, partial observation y
- Unknown $u \in \mathbb{R}^{d_u}$: the pixel values of the image
- Map $\mathcal{G}$: The map $\mathcal{G} = G \in \mathbb{R}^{d_y \times d_u}$ is linear in many imaging problems, where G incorporates mechanisms such as
  - blurring ($\rightarrow$ averaging over a neighbourhood of pixels)
  - Fourier transform ($\rightarrow$ observations in frequency domain)
  - a mask ($\rightarrow$ partial observations)
- Observations: $y = Gu + \eta \in \mathbb{R}^{d_y}$.





Left: $|y|$ Right: u^*

What are inverse problems?

Example 1: Imaging

- Particular example: Deblurring (without noise).

The map G is defined by its action on each pixel value:

$$y_i = \sum_{j \in N(i)} w_{ij} u_j,$$

where

- y_i is the ith pixel value of the blurred image,
- $N(i)$ is neighbourhood of pixel i,
- w_{ij} are weights indicating the importance of each pixel $j \in N(i)$, and
- u_j is the jth pixel value of the original image.

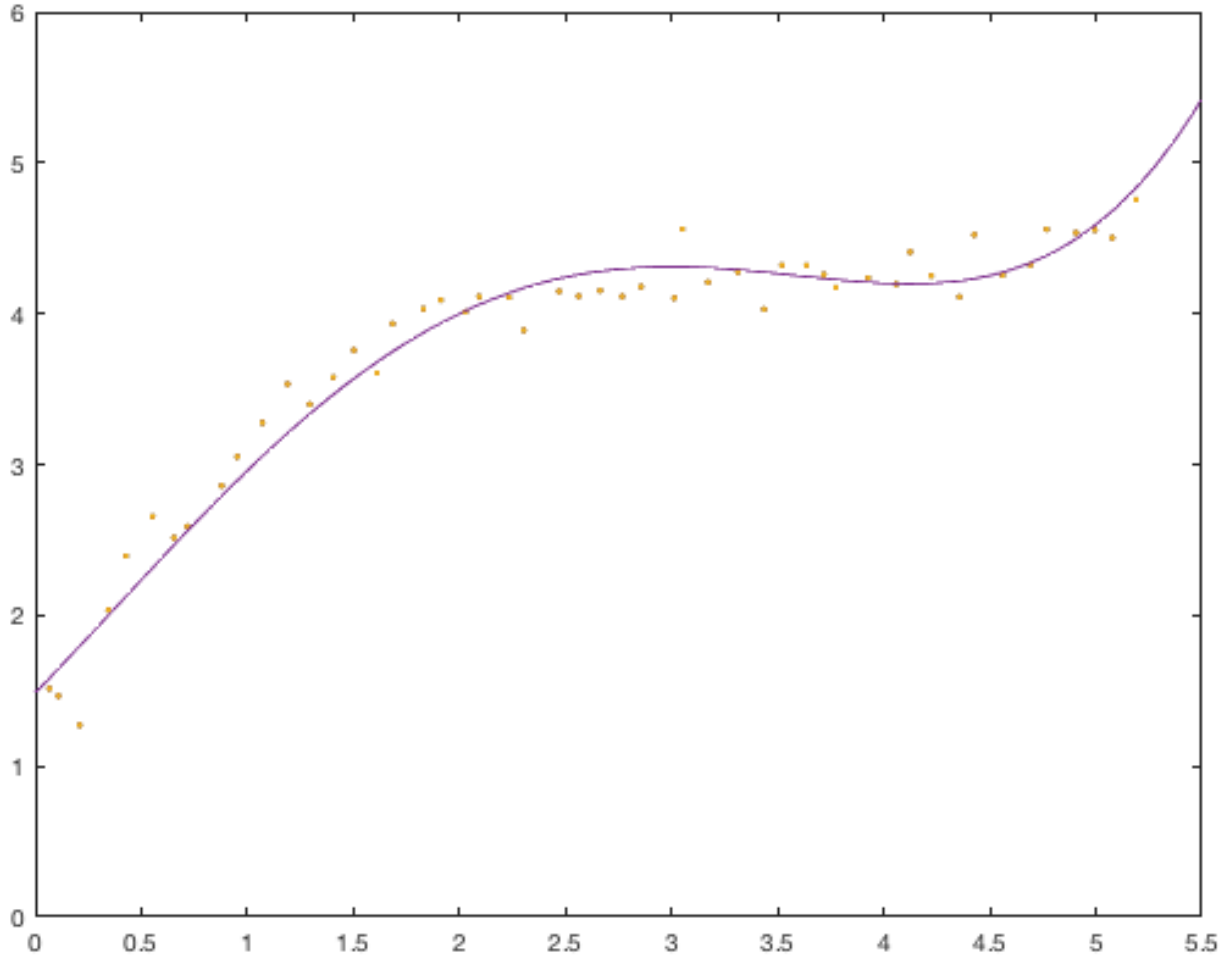
Original

Blurred

What are inverse problems?

Example 2: Regression

- **Goal**: reconstruct a function f given noisy point values $\{f(x_i)\}$.



What are inverse problems?

Example 2: Regression

- **Goal**: reconstruct a function f given noisy point values $\{f(x_i)\}$.
- Unknown $u \in \mathbb{R}^{d_u}$: coefficients in a basis expansion

$$f(x; u) = \sum_{j=1}^{d_u} u_j \phi_j(x),$$

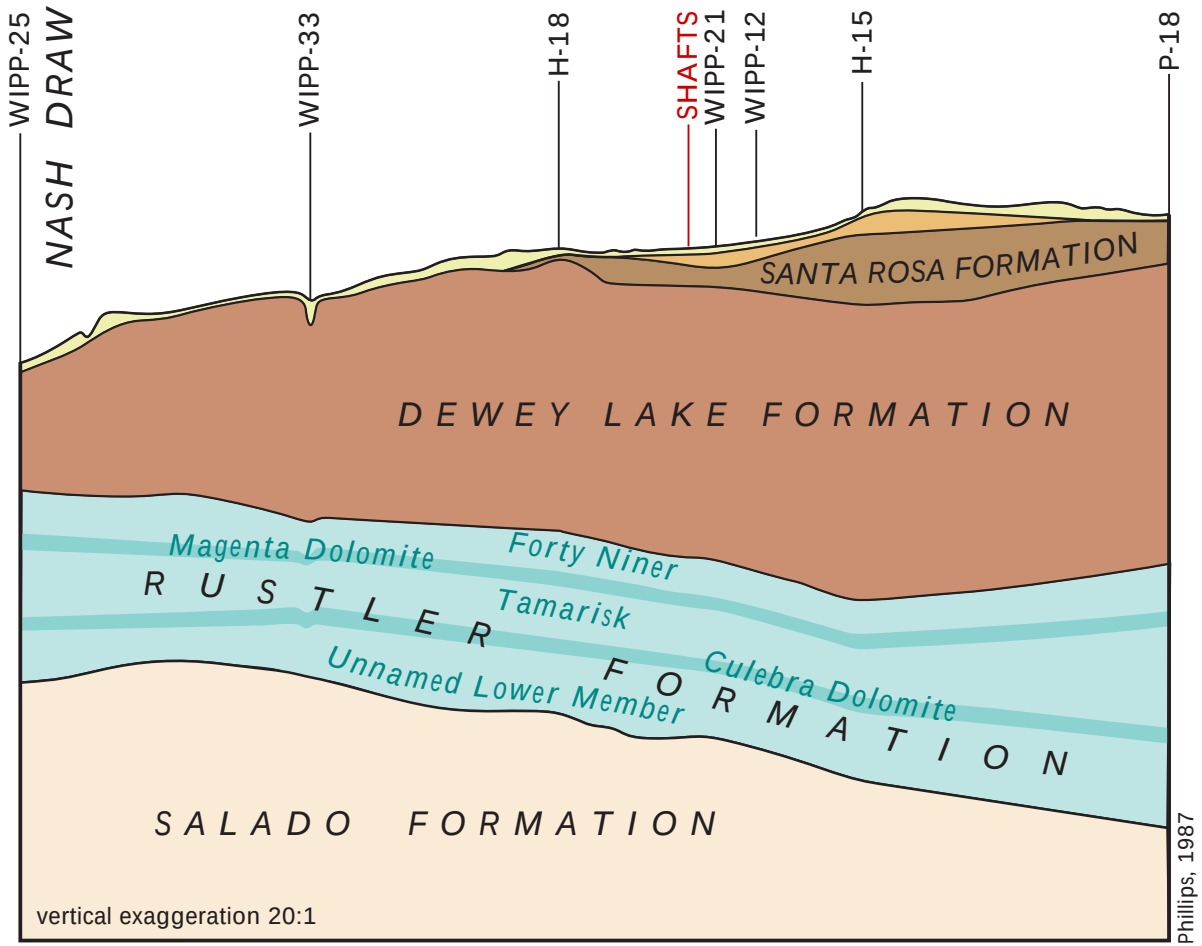
  where $\{\phi_j\}_{j=1}^{d_u}$ are linearly independent, e.g. $\phi_j(x) = x^{j-1}$.

- Map $\mathcal{G}$: implicitly defined by $u \mapsto \{f(x_i; u)\}_{i=1}^{d_y}$. In fact $\mathcal{G} = G$ is linear, with $a_{ij} = \phi_j(x_i)$ (Vandermonde-type matrix).
- Observations: $y = \{f(x_i; u) + \eta_i\}_{i=1}^{d_y} \in \mathbb{R}^{d_y}$.

What are inverse problems?

Example 3: Porous media flow

- **Goal**: reconstruct the hydraulic conductivity k of the subsurface given noisy measurements of the water pressure $\{p(x_i)\}$.



Cross-section at WIPP

What are inverse problems?

Example 3: Porous media flow

- **Goal**: reconstruct the hydraulic conductivity k of the subsurface given noisy measurements of the water pressure $\{p(x_i)\}$.
- Unknown $u \in \mathbb{R}^{d_u}$: coefficients in a basis expansion

$$k(x; u) = \phi_0(x) + \sum_{j=1}^{d_u} u_j \phi_j(x),$$

  where $\{\phi_j\}_{j=1}^{d_u}$ are linearly independent and ϕ_0 is s.t. k is positive.

- Map $\mathcal{G}$: implicitly defined by $u \mapsto \{p(x_i; u)\}_{i=1}^{d_y}$, where p is the solution of

$$-\nabla \cdot (k(x; u) \nabla p(x; u)) = h(x).$$

  (This equation comes from Darcy's law plus conservation of mass: conductivity k, pressure head p, sources/sinks h.)

- Observations: $y = \{p(x_i; u) + \eta_i\}_{i=1}^{d_y} \in \mathbb{R}^{d_y}$.

What are inverse problems?

General inverse problems

- We are interested in the following inverse problem: given observational data $y \in Y$, determine unknown $u \in U$ such that

$$y = \mathcal{G}(u) + \eta.$$

- Simply "inverting $\mathcal{G}$" is not possible, since
  - we do not know the value of η, and
  - typically $\mathcal{G}^{-1}$ does not exist.

    (Think about the case where $\mathcal{G}$ is linear, i.e. $\mathcal{G} = G \in \mathbb{R}^{d_y \times d_u}$. Unless $d_y = d_u$, G^{-1} does not exist.)

What are inverse problems?

Ill-posedness and ill-conditioning of inverse problems

- Intuitively, we want to choose u to minimise the *data misfit functional*

$$J(u; y) := \|y - \mathcal{G}(u)\|_2^2.$$

What are inverse problems?

Ill-posedness and ill-conditioning of inverse problems

- Intuitively, we want to choose u to minimise the *data misfit functional*

$$J(u; y) := \|y - \mathcal{G}(u)\|_2^2.$$

- However, the optimisation problem $u^* := \arg\min_{u \in U} J(u; y)$ is typically ill-posed in the sense of Hadamard:
  - there is no unique solution u^*, or
  - u^* does not depend continuously on y,

  and ill-conditioned:
  - small changes in y can lead to large changes in u^*.

- Note that the forward problem of finding $\mathcal{G}(u)$ given u is typically well-posed.

What are inverse problems?

Ill-posedness: non-unique solutions

- Consider the problem $U = \mathbb{R}$, $Y = \mathbb{R}$ and $\mathcal{G}(u) = u^2$.
- We observe $y = u^2 + \eta$.
- For $y > 0$, the data misfit functional $J(u; y) = (y - u^2)^2$ has two minimisers $u^* = \pm\sqrt{y}$.

What are inverse problems?

Ill-conditioning: Linear example

- Consider the linear case where $\mathcal{G} = G \in \mathbb{R}^{d_y \times d_u}$.

- If $d_y \geq d_u$ (overdetermined system) and G is of full rank, then $\|y - Gu\|_2^2$ has a unique minimiser, given by the solution of the *normal equations*

$$G^{\mathrm{T}} G u^* = G^{\mathrm{T}} y.$$

What are inverse problems?

Ill-conditioning: Linear example

- Consider the linear case where $\mathcal{G} = G \in \mathbb{R}^{d_y \times d_u}$.

- If $d_y \geq d_u$ (overdetermined system) and G is of full rank, then $\|y - Gu\|_2^2$ has a unique minimiser, given by the solution of the *normal equations*

$$G^{\mathrm{T}} G u^* = G^{\mathrm{T}} y.$$

- If the matrix $G^{\mathrm{T}} G$ is ill-conditioned, then small changes in y result in large changes in u^*: $G^{\mathrm{T}} G(u^* + \Delta u^*) = G^{\mathrm{T}}(y + \Delta y)$

$$\Rightarrow \frac{\|\Delta u^*\|}{\|u^*\|} \leq \mathrm{cond}(G^{\mathrm{T}} G) \frac{\|G^{\mathrm{T}} \Delta y\|}{\|G^{\mathrm{T}} y\|},$$

  where $\mathrm{cond}(G^{\mathrm{T}} G) := \|G^{\mathrm{T}} G\| \|(G^{\mathrm{T}} G)^{-1}\|$.

- $G^T G$ is ill-conditioned in many examples, e.g. the Vandermonde-type matrices occuring in polynomial regression.

Bayesian approach

Bayesian inference [Kaipio, Somersalo '06]

- We apply a Bayesian statistical approach to solve inverse problems.

Bayesian approach

Bayesian inference [Kaipio, Somersalo '06]

- We apply a Bayesian statistical approach to solve inverse problems.
- We choose a prior distribution μ_0 on u with pdf $\pi_0(u)$ on U.

Bayesian approach

Bayesian inference [Kaipio, Somersalo '06]

- We apply a Bayesian statistical approach to solve inverse problems.
- We choose a prior distribution μ_0 on u with pdf $\pi_0(u)$ on U.
- Under the measurement model $y = \mathcal{G}(u) + \eta$ with $\eta \sim N(0, \gamma^2 \mathrm{I})$, we have $y|u \sim N(\mathcal{G}(u), \gamma^2 \mathrm{I})$, and the likelihood of the data y is

$$L(y|u) = \frac{1}{\sqrt{(2\pi\gamma^2)^{d_u}}} \exp\Big(-\frac{1}{2\gamma^2} \|y - \mathcal{G}(u)\|_2^2 \Big).$$

In general, $L(y|u) = \exp(-\Phi(y;u))$ for some potential Φ.

Bayesian approach

Bayesian inference [Kaipio, Somersalo '06]

- We apply a Bayesian statistical approach to solve inverse problems.
- We choose a prior distribution μ_0 on u with pdf $\pi_0(u)$ on U.
- Under the measurement model $y = \mathcal{G}(u) + \eta$ with $\eta \sim N(0, \gamma^2 \mathrm{I})$, we have $y|u \sim N(\mathcal{G}(u), \gamma^2 \mathrm{I})$, and the likelihood of the data y is

$$L(y|u) = \frac{1}{\sqrt{(2\pi\gamma^2)^{d_u}}} \exp\Big(-\frac{1}{2\gamma^2} \|y - \mathcal{G}(u)\|_2^2 \Big).$$

  In general, $L(y|u) = \exp(-\Phi(y;u))$ for some potential Φ.

- Using Bayes' Theorem, we obtain the posterior distribution μ^y on $u|y$ with pdf $\pi^y(u)$, given by

$$\pi(u|y) =: \pi^y(u) = \frac{L(y|u)\ \pi_0(u)}{\int_U L(y|u)\pi_0(u)\mathrm{d}u}.$$

- $\pi^y(u)$ is large where $\|y - \mathcal{G}(u)\|_2^2$ is small and $\pi_0(u)$ is large.

Bayesian approach

Advantages of the Bayesian approach

- The solution to the Bayesian inverse problem is the posterior pdf $\pi^y(u)$. This allows for uncertainty quantification in the inferred parameter.

Bayesian approach

Advantages of the Bayesian approach

- The solution to the Bayesian inverse problem is the posterior pdf $\pi^y(u)$. This allows for uncertainty quantification in the inferred parameter.
- In the Bayesian framework, the inverse problem is well-posed:
  - there exists a unique posterior distribution for all $y \in \mathbb{R}^{d_y}$.
    If there are multiple minimisers of $\|y - \mathcal{G}(u)\|_2^2$, then the posterior distribution has multiple modes.
  - the posterior distribution π^y depends continuously on y: if $L(y|u)$ is locally Lipschitz in y, then

$$d_{\mathrm{TV}}(\pi^y, \pi^{y'}) \leq C\|y - y'\|_2.$$

Bayesian approach

Challenges in the Bayesian approach

- The posterior distribution is typically not known in closed form (notable exception: the linear Gaussian case).
- Advanced sampling methods such as Markov chain Monte Carlo methods are required for sampling from the posterior, e.g. for computing the posterior mean $\mathbb{E}[u|y]$.

Bayesian approach

Example: linear, Gaussian, one dimensional

Consider the following example ($d_y = d_u = 1$): $y = gu + \eta$

- Forward model: $g \in \mathbb{R}$,
- Prior: $u \sim N(0, \sigma_0^2)$ with pdf $\pi_0(u) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{\|u\|^2}{2\sigma_0^2})$,
- Noise: $\eta \sim \mathcal{N}(0, \gamma^2)$ with likelihood $L(y|u) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp(-\frac{\|y-gu\|^2}{2\gamma^2})$.

Bayesian approach

Example: linear, Gaussian, one dimensional

Consider the following example ($d_y = d_u = 1$): $y = gu + \eta$

- Forward model: $g \in \mathbb{R}$,
- Prior: $u \sim N(0, \sigma_0^2)$ with pdf $\pi_0(u) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{\|u\|^2}{2\sigma_0^2})$,
- Noise: $\eta \sim \mathcal{N}(0, \gamma^2)$ with likelihood $L(y|u) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp(-\frac{\|y-gu\|^2}{2\gamma^2})$.

Using Bayes' formula gives

$$\pi(u|y) =: \pi^y(u) = \frac{L(y|u)\ \pi_0(u)}{\int_{\mathbb{R}} L(y|u)\pi_0(u)\mathrm{d}u}$$

$$\Rightarrow u|y \sim \mathcal{N}\Big(\frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2} y,\ \sigma_0^2 \frac{\gamma^2}{\gamma^2 + \sigma_0^2 g^2}\Big).$$

The posterior has a shifted mean and smaller variance.

By assumption, we have the following prior density and likelihood:

$$\pi_0(u) \propto \exp\big(-\frac{1}{2\sigma_0^2}u^2\big),$$

$$L(y|u) \propto \exp(-\frac{1}{2\gamma^2}(y-gu)^2)\ .$$

By assumption, we have the following prior density and likelihood:

$$\pi_0(u) \propto \exp\big(-\frac{1}{2\sigma_0^2}u^2\big),$$
$$L(y|u) \propto \exp(-\frac{1}{2\gamma^2}(y-gu)^2)\ .$$

Thus, by Bayes' formula, the posterior is

$$\begin{aligned}\pi^y(u) &\propto \exp\big(-\frac{1}{2\sigma_0^2}u^2 - \frac{1}{2\gamma^2}(y-gu)^2\big)\\ &= \exp\Big(-\frac{1}{2}\Big((\frac{1}{\sigma_0^2}+\frac{g^2}{\gamma^2})u^2 - 2\frac{gy}{\gamma^2}u + \frac{y^2}{\gamma^2}\Big)\Big)\\ &=: \exp\Big(\Big(-\frac{1}{2}\big(au^2-2bu+c\big)\Big),\end{aligned}$$

with

$$a = \frac{1}{\sigma_0^2}+\frac{g^2}{\gamma^2}\ ,\quad b = \frac{gy}{\gamma^2}\ ,\quad c = \frac{y^2}{\gamma^2}\ ,$$

We want to find constants m, K, σ, such that

$$\pi^y(u) \propto \exp(-\frac{1}{2\sigma^2}(u-m)^2 + K) \; .$$

By completing the square, we obtain

$$\begin{aligned} au^2 - 2bu + c &= a(u^2 - 2\frac{b}{a}u + \frac{c}{a}) \\ &= a(u^2 - 2\frac{b}{a}u + (\frac{b}{a})^2 - (\frac{b}{a})^2 + \frac{c}{a}) \\ &= a(u - \frac{b}{a})^2 + c - \frac{b^2}{a}, \end{aligned}$$

We want to find constants m, K, σ, such that

$$\pi^y(u) \propto \exp(-\frac{1}{2\sigma^2}(u-m)^2 + K) \; .$$

By completing the square, we obtain

$$\begin{aligned} au^2 - 2bu + c &= a(u^2 - 2\frac{b}{a}u + \frac{c}{a}) \\ &= a(u^2 - 2\frac{b}{a}u + (\frac{b}{a})^2 - (\frac{b}{a})^2 + \frac{c}{a}) \\ &= a(u - \frac{b}{a})^2 + c - \frac{b^2}{a}, \end{aligned}$$

and thus,

$$\sigma^2 = \frac{1}{a} = \frac{\sigma_0^2\gamma^2}{\gamma^2 + g^2\sigma_0^2} \; , \quad m = \frac{b}{a} = \frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2} y$$

and for the constant K, which does not depend on u, we obtain

$$K = c - \frac{b^2}{a} = \frac{y^2}{\gamma^2} - \frac{\sigma_0^2 g^2 y^2}{y^4 + y^2 g^2 \sigma^2} \; .$$

Bayesian approach

Example: linear, Gaussian, one dimensional

Consider the following example ($d_y = d_u = 1$):

- Forward model: $g \in \mathbb{R}$,
- Prior: $u \sim N(0, \sigma_0^2)$ with pdf $\pi_0(u) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{\|u\|^2}{2\sigma_0^2})$,
- Noise: $\eta \sim \mathcal{N}(0, \gamma^2)$ giving likelihood $L(y|u) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp(-\frac{\|\eta\|^2}{2\gamma^2})$.

Using Bayes' formula gives

$$\pi(u|y) =: \pi^y(u) = \frac{L(y|u)\ \pi_0(u)}{\int_{\mathbb{R}} L(y|u)\pi_0(u)\mathrm{d}u}$$

$$\Rightarrow u|y \sim \mathcal{N}\Big(\frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2} y,\ \sigma_0^2 \frac{\gamma^2}{\gamma^2 + \sigma_0^2 g^2}\Big).$$

Bayesian approach

Example: linear, Gaussian

Generalising the previous example to multiple dimensions:

- Forward model: $G \in \mathbb{R}^{d_u \times d_y}$,
- Prior: $u \sim \mathcal{N}(m_0, C_0)$,
- Noise: $\eta \sim \mathcal{N}(0, \Gamma)$.

Using Bayes' formula gives $u|y \sim \mathcal{N}(m, C)$, with

$$m = m_0 + C_0 G^T (G C_0 G^T + \Gamma)^{-1} (y - G m_0),$$
$$C = C_0 - C_0 G^T (G C_0 G^T + \Gamma)^{-1} G C_0.$$

Markov chain Monte Carlo methods

Motivation

- In many applications, one is interested in computing the expected value of a quantity of interest ϕ under the target distribution π:

$$\mathbb{E}[\phi] = \int_{\mathbb{R}^{d_u}} \phi(u)\,\pi(u)\,du.$$

- The integral is typically intractable, and can be approximated using a sampling method:

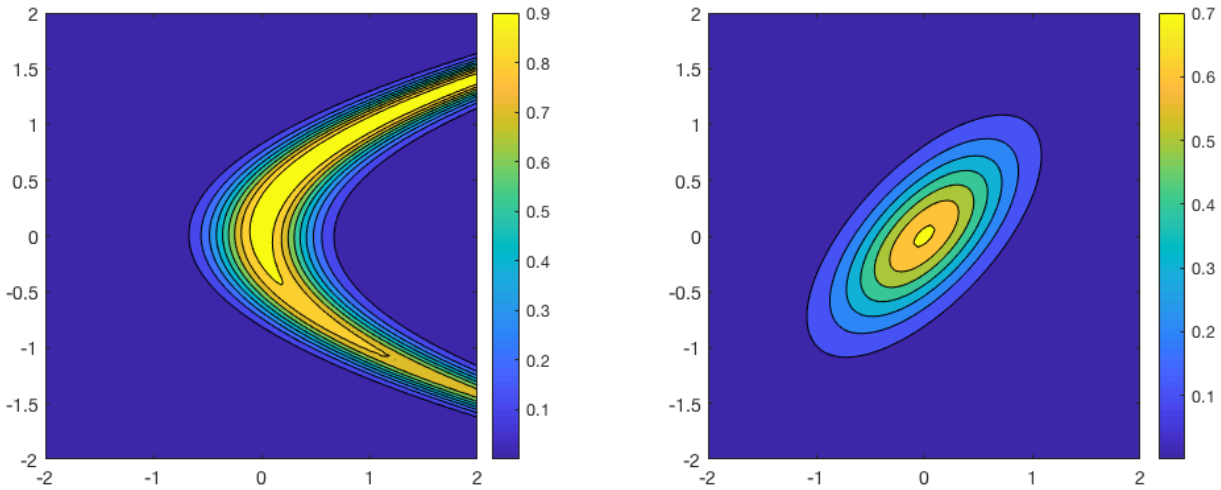
$$\mathbb{E}[\phi] \approx \frac{1}{N}\sum_{i=1}^{N} \phi(u^{(i)}),$$

where $u^{(i)} \sim \pi$, for all $1 \leq i \leq N$.

Markov chain Monte Carlo methods

Sampling methods

- Generating samples $u^{(i)} \sim \pi$ is often difficult since:
  - π is not known in closed form, eg only up to a normalisation constant.
    $\Rightarrow$ not possible to generate independent (i.i.d.) samples from π
  - the state variable $u \in \mathbb{R}^{d_u}$ is high dimensional.
  - π can concentrate on low-dimensional manifolds.



Markov chain Monte Carlo methods

Markov chain[Robert, Casella '99]

A Markov chain Monte Carlo (MCMC) estimator of $\mathbb{E}[\phi]$ is of the form

$$\widehat{E}_N^{\mathrm{MCMC}} := \frac{1}{N}\sum_{i=1}^{N}\phi(u^{(i)}),$$

where $\{u^{(i)}\}_{i=1}^{\infty}$ is a Markov chain.

Definition (Markov chain)

The family of random variables $\{u^{(i)}\}_{i=1}^{\infty}$ is called a *Markov chain* if

$$\Pr\left[\, u^{(i)} = x_i \,\middle|\, u^{(1)} = x_1\,, \ldots,\, u^{(i-1)} = x_{i-1} \,\right] = \Pr\left[\, u^{(i)} = x_i \,\middle|\, u^{(i-1)} = x_{i-1} \,\right],$$

for all $i \geq 2$ and $x_1, \ldots, x_i \in \mathbb{R}^{d_u}$.

Markov chain Monte Carlo methods

Markov chain[Robert, Casella '99]

A Markov chain Monte Carlo (MCMC) estimator of $\mathbb{E}[\phi]$ is of the form

$$\widehat{E}_N^{\mathrm{MCMC}} := \frac{1}{N}\sum_{i=1}^{N}\phi(u^{(i)}),$$

where $\{u^{(i)}\}_{i=1}^{\infty}$ is a Markov chain.

Definition (Markov chain)

The family of random variables $\{u^{(i)}\}_{i=1}^{\infty}$ is called a *Markov chain* if

$$\Pr\left[\, u^{(i)} = x_i \,\middle|\, u^{(1)} = x_1\,, \ldots,\, u^{(i-1)} = x_{i-1} \,\right] = \Pr\left[\, u^{(i)} = x_i \,\middle|\, u^{(i-1)} = x_{i-1} \,\right],$$

for all $i \geq 2$ and $x_1, \ldots, x_i \in \mathbb{R}^{d_u}$.

We want the distribution of each $u^{(i)}$ to be (close to) π.

Markov chain Monte Carlo methods

Why Markov chains?

- Practical advantage:
  - Allows for sequential construction: construct $u^{(i+1)}$ from $u^{(i)}$.
- Theoretical advantages:
  - Stationary distributions: we can construct $\{u^{(i)}\}_{i=1}^{\infty}$ s.t. $u^{(i)} \sim \pi$ as $i \to \infty$ for any $u^{(1)} \in \mathbb{R}^{d_u}$
  - Ergodic average: we can construct $\{u^{(i)}\}_{i=1}^{\infty}$ s.t. $\frac{1}{N}\sum_{i=1}^{N} \phi(u^{(i)}) \to \mathbb{E}[\phi]$ as $N \to \infty$.

Markov chain Monte Carlo methods

Metropolis Hastings Algorithm [Robert, Casella '99]

A particular example is the Metropolis Hastings (MH-MCMC) estimator, which uses the following algorithm to construct $\{u^{(i)}\}_{i=1}^{\infty}$:

ALGORITHM 1. (Metropolis Hastings)

- Choose $u^{(1)}$ with $\pi(u^{(1)}) > 0$.
- At state $u^{(i)}$, sample a proposal u' from density $q(u' \,|\, u^{(i)})$.
- Accept sample u' with probability

$$\alpha(u' \,|\, u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \,|\, u')}{\pi(u^{(i)})\, q(u' \,|\, u^{(i)})}\right),$$

i.e. $u^{(i+1)} = u'$ with probability $\alpha(u' \,|\, u^{(i)})$; otherwise stay at $u^{(i+1)} = u^{(i)}$.

Markov chain Monte Carlo methods

Metropolis Hastings Algorithm [Robert, Casella '99]

A particular example is the Metropolis Hastings (MH-MCMC) estimator, which uses the following algorithm to construct $\{u^{(i)}\}_{i=1}^{\infty}$:

ALGORITHM 1. (Metropolis Hastings)

- Choose $u^{(1)}$ with $\pi(u^{(1)}) > 0$.
- At state $u^{(i)}$, sample a proposal u' from density $q(u' \,|\, u^{(i)})$.
- Accept sample u' with probability

$$\alpha(u' \,|\, u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \,|\, u')}{\pi(u^{(i)})\, q(u' \,|\, u^{(i)})}\right),$$

  i.e. $u^{(i+1)} = u'$ with probability $\alpha(u' \,|\, u^{(i)})$; otherwise stay at $u^{(i+1)} = u^{(i)}$.

- The proposal density q is chosen to be easy to sample from.

Markov chain Monte Carlo methods

Metropolis Hastings Algorithm [Robert, Casella '99]

A particular example is the Metropolis Hastings (MH-MCMC) estimator, which uses the following algorithm to construct $\{u^{(i)}\}_{i=1}^{\infty}$:

ALGORITHM 1. (Metropolis Hastings)

- Choose $u^{(1)}$ with $\pi(u^{(1)}) > 0$.
- At state $u^{(i)}$, sample a proposal u' from density $q(u' \,|\, u^{(i)})$.
- Accept sample u' with probability

$$\alpha(u' \,|\, u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \,|\, u')}{\pi(u^{(i)})\, q(u' \,|\, u^{(i)})}\right),$$

  i.e. $u^{(i+1)} = u'$ with probability $\alpha(u' \,|\, u^{(i)})$; otherwise stay at $u^{(i+1)} = u^{(i)}$.

- The proposal density q is chosen to be easy to sample from.
- The accept/reject step is added in order to obtain samples from π.

Markov chain Monte Carlo methods

Metropolis Hastings Algorithm [Robert, Casella '99]

A particular example is the Metropolis Hastings (MH-MCMC) estimator, which uses the following algorithm to construct $\{u^{(i)}\}_{i=1}^{\infty}$:

ALGORITHM 1. (Metropolis Hastings)

- Choose $u^{(1)}$ with $\pi(u^{(1)}) > 0$.
- At state $u^{(i)}$, sample a proposal u' from density $q(u' \,|\, u^{(i)})$.
- Accept sample u' with probability

$$\alpha(u' \,|\, u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \,|\, u')}{\pi(u^{(i)})\, q(u' \,|\, u^{(i)})}\right),$$

  i.e. $u^{(i+1)} = u'$ with probability $\alpha(u' \,|\, u^{(i)})$; otherwise stay at $u^{(i+1)} = u^{(i)}$.

- The proposal density q is chosen to be easy to sample from.
- The accept/reject step is added in order to obtain samples from π.
- Knowledge of the normalising constant Z of π is not required.

Markov chain Monte Carlo methods

Properties of acceptance probability α

Given current state $u^{(i)}$, we accept proposed state u' with probability

$$\alpha(u' \mid u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \mid u')}{\pi(u^{(i)})\, q(u' \mid u^{(i)})}\right).$$

Markov chain Monte Carlo methods

Properties of acceptance probability α

Given current state $u^{(i)}$, we accept proposed state u' with probability

$$\alpha(u' \mid u^{(i)}) = \min\left(1, \frac{\pi(u')\, q(u^{(i)} \mid u')}{\pi(u^{(i)})\, q(u' \mid u^{(i)})}\right).$$

The proposal u' has a high probability of acceptance when

- $\frac{\pi(u')}{\pi(u^{(i)})}$ is large $\Rightarrow$ u' has a high target probability $\Rightarrow$ we choose samples from regions of high target probability,
- $\frac{q(u^{(i)} \mid u')}{q(u' \mid u^{(i)})}$ is large $\Rightarrow$ there is a high probability of moving back to $u^{(i)}$ from u'.

Markov chain Monte Carlo methods

Central Limit Theorem [Robert, Casella '99]

Define an auxiliary chain $\{\tilde{u}^{(i)}\}_{i=1}^{\infty}$, generated by Algorithm 1 with $\tilde{u}^{(1)} \sim \pi$. Define the *asymptotic variance*

$$\sigma_\phi^2 := \mathbb{V}[\phi(\tilde{u}^{(1)})] + 2\sum_{i=2}^{\infty} \mathbb{C}\mathrm{ov}[\phi(\tilde{u}^{(1)}), \phi(\tilde{u}^{(i)})].$$

Markov chain Monte Carlo methods

Central Limit Theorem [Robert, Casella '99]

Define an auxiliary chain $\{\tilde{u}^{(i)}\}_{i=1}^{\infty}$, generated by Algorithm 1 with $\tilde{u}^{(1)} \sim \pi$. Define the *asymptotic variance*

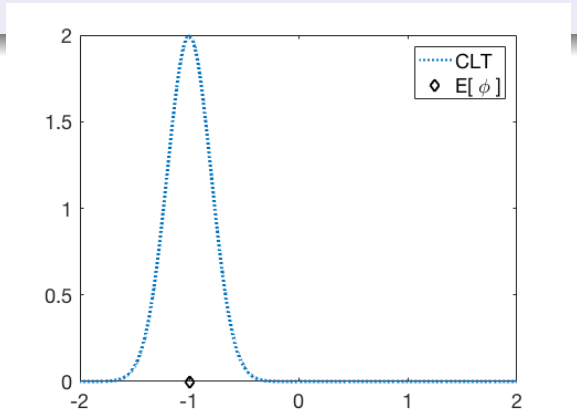
$$\sigma_\phi^2 := \mathbb{V}[\phi(\tilde{u}^{(1)})] + 2\sum_{i=2}^{\infty} \mathbb{C}\mathrm{ov}[\phi(\tilde{u}^{(1)}), \phi(\tilde{u}^{(i)})].$$

Theorem (Central Limit Theorem)

Suppose $\sigma_\phi^2 < \infty$, $\Pr[\alpha = 1] < 1$ *and* $q(u \,|\, u^*) > 0$ *for all* u, u^* *s.t.* $\pi(u), \pi(u^*) > 0$. *Then, as* $N \to \infty$, *we have*

$$\sqrt{N\sigma_\phi^{-2}} \left(\frac{1}{N} \sum_{i=1}^{N} \phi(u^{(i)}) - \mathbb{E}[\phi] \right) \xrightarrow{D} \mathcal{N}(0, 1).$$

The estimator $\frac{1}{N}\sum_{i=1}^{N} \phi(u^{(i)})$ is asymptotically normally distributed, with mean $\mathbb{E}[\phi]$ and variance $\frac{\sigma_\phi^2}{N}$.



Markov chain Monte Carlo methods

Choice of proposal density

The proposal density is chosen to balance

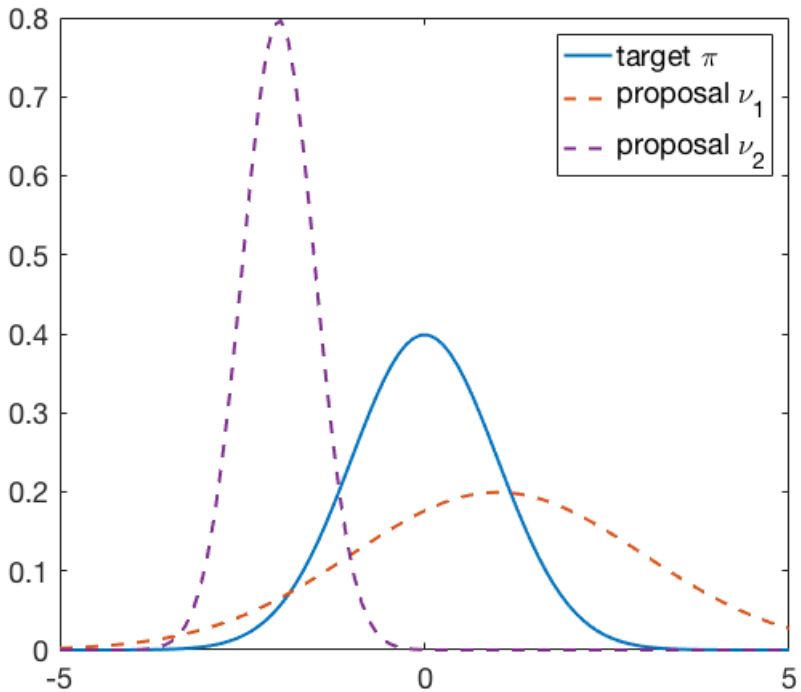
- accuracy:
  - We want to make the asymptotic variance σ_ϕ^2 small.
  - This requires reducing the correlation between samples.
- cost:
  - Sampling from the proposal $q(\cdot \mid u^{(i)})$ has varying cost depending the particular choice.
  - This could involve computing the gradient $\nabla \log \pi(u^{(i)})$ and/or higher order derivatives.

Markov chain Monte Carlo methods

Independent proposal

The independence sampler chooses a proposal distribution independent of the current state $u^{(i)}$: $q(\cdot \mid u^{(i)}) = \nu(\cdot)$.

- This either works very well or very poorly...
- It can work well e.g. in the Bayesian inference problem, with $\nu = \pi_0$, if the prior π_0 and the posterior π are sufficiently close.
- Note that we do not get independent samples $\{u^{(i)}\}_{i=1}^{\infty}$ due to accept/reject step.

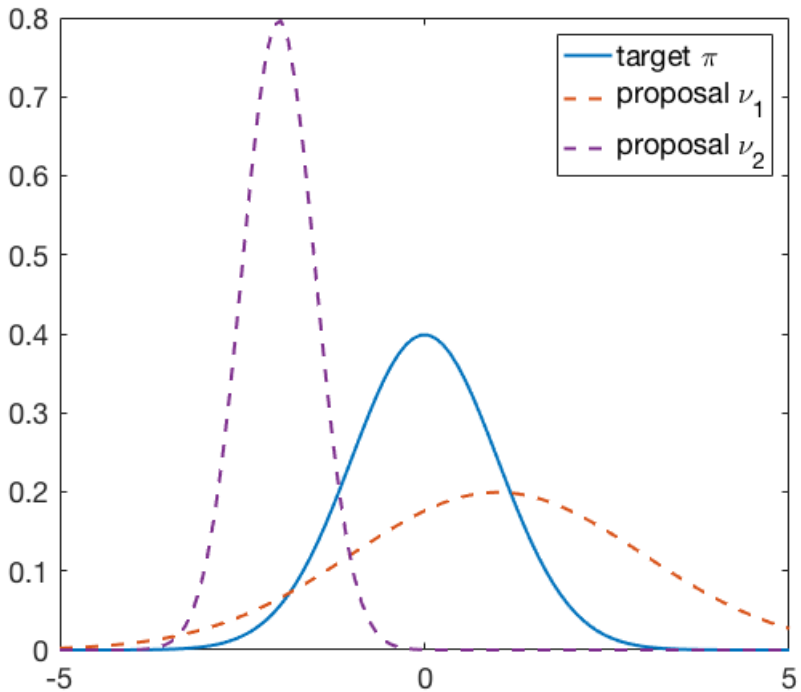


Markov chain Monte Carlo methods

Independent proposal

The independence sampler chooses a proposal distribution independent of the current state $u^{(i)}$: $q(\cdot \mid u^{(i)}) = \nu(\cdot)$.

- This either works very well or very poorly...
- It can work well e.g. in the Bayesian inference problem, with $\nu = \pi_0$, if the prior π_0 and the posterior π are sufficiently close.
- Note that we do not get independent samples $\{u^{(i)}\}_{i=1}^{\infty}$ due to accept/reject step.



The independence sampler does not use the current state $u^{(i)}$...

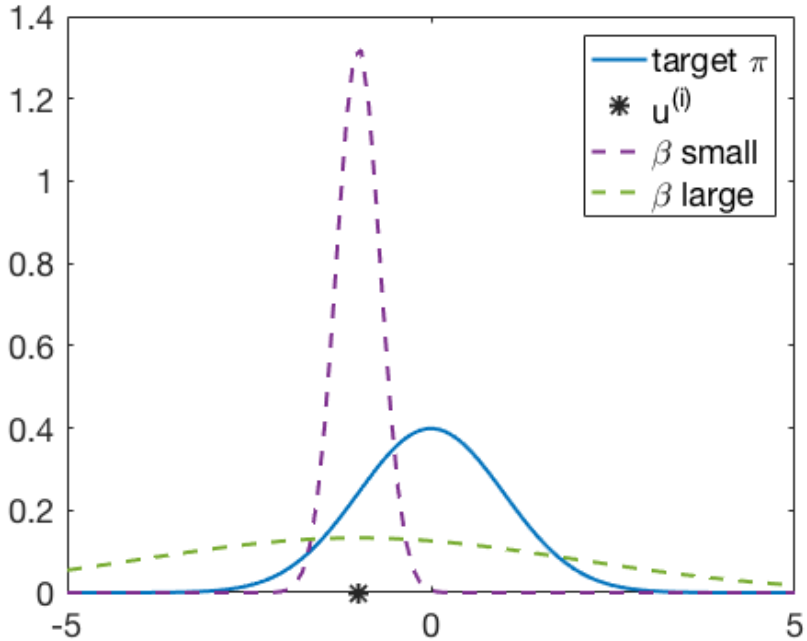
Markov chain Monte Carlo methods

Random walk proposal [Roberts et al, '97]

The random walk proposal is given by $q(u' \mid u^{(i)}) = \mathcal{N}(u^{(i)}, \beta^2 \mathrm{I})$, for some $\beta > 0$, i.e.

$$u' = u^{(i)} + \beta \xi_i, \qquad \text{where} \qquad \xi_i \sim \mathcal{N}(0, \mathrm{I}), \quad \beta > 0,$$

- Here, β is a step size that needs to be tuned:
  - if β is too small, you don't explore the state space.
  - if β is too large, you reject too often.
  - both scenarios lead to large asymptotic variance σ_ϕ^2.



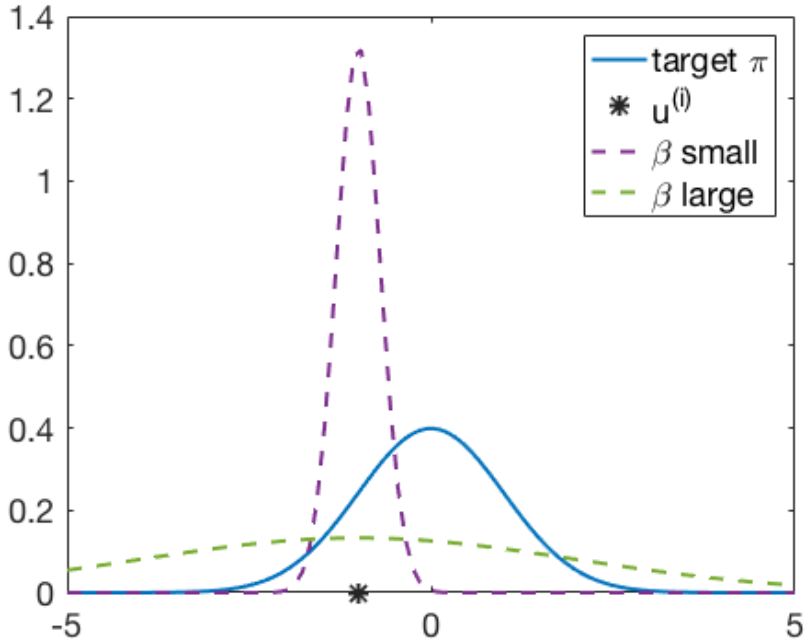
Markov chain Monte Carlo methods

Random walk proposal [Roberts et al, '97]

The random walk proposal is given by $q(u' \mid u^{(i)}) = \mathcal{N}(u^{(i)}, \beta^2 \mathrm{I})$, for some $\beta > 0$, i.e.

$$u' = u^{(i)} + \beta \xi_i, \qquad \text{where} \qquad \xi_i \sim \mathcal{N}(0, \mathrm{I}), \quad \beta > 0,$$

- Here, β is a step size that needs to be tuned:
  - if β is too small, you don't explore the state space.
  - if β is too large, you reject too often.
  - both scenarios lead to large asymptotic variance σ_ϕ^2.



- A general rule of thumb is to tune β such that the average acceptance probability is $\alpha \approx 0.234$, to achieve small asymptotic variance.

Markov chain Monte Carlo methods

Computational challenges

Computational challenges in complex models:

- Challenge in high dimensions u: for any β, the average acceptance rate $\alpha \to 0$ as $d_u \to \infty$. $\Rightarrow \sigma_\phi^2 \to \infty$ as $d_u \to \infty$
- Challenge for computationally expensive likelihoods $e^{-\frac{1}{2\gamma^2}\|y-\mathcal{G}(u)\|_2^2}$: this needs to be evaluated at each iteration of the Metropolis Hastings algorithm. (e.g. PDE-based or big data applications)

These will be addressed in the next lecture!

References I


J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, Springer, 2004.


C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 1999.


A. Stuart, *Inverse Problems: A Bayesian Perspective*, Acta Numerica, 19 (2010), pp. 451–559.