Markov chain Monte Carlo methods for Uncertainty Quantification
General concept
Lecture 21

March 31, 2016; April 5, 2016; April 7, 2016

Motivation: Bayesian inference & prediction

Data: $D = (D_1, ..., D_N)$
Parameters: $X = (X_1, ..., X_d)$
Likelihood: $\mathcal{L}(D|X)$
Prior model: $\pi(X)$

Posterior model: The basis for inference about X
... via Bayes theorem

$$\pi(X|D) = \frac{\mathcal{L}(D|X)\pi(X)}{\int \mathcal{L}(D|Y)\pi(Y)\mathrm{d}Y}$$

Predictive model: Predictive distribution of a future obs. D_{future}.

$$\mathcal{L}(D_{\text{future}}|D) = \int \mathcal{L}(D_{\text{future}}|X)\pi(X|D)\mathrm{d}X$$

... expected likelihood where uncertainty of X is constrained w.r.t $\pi(X|D)$

The problem: Challenges in Bayesian Inference

For some function $g(\cdot)$,

- the derivation of any posterior quantity requires the computation of integrals of the form:

$$\mathrm{E}_{\pi(X|D)}(g(X)) = \int g(X)\pi(X|D)\mathrm{d}X$$

- the posterior distribution density $\pi(X|D)$ or $\pi(g(X)|D)$ is intractable because of

$$\int \mathcal{L}(D|X)\pi(X)\mathrm{d}X = ??$$

CODE

CODE

Application: Bayesian hierarchical model

Likelihood

$$D_i \sim \text{Bernoulli}(p(t_i|\alpha,\beta)),\ i = 1, ..., 23$$

$$p(t_i|\alpha,\beta) = \frac{\exp(\alpha + \beta t_i)}{1 + \exp(\alpha + \beta t_i)}$$

Prior

$$\alpha \sim \text{N}(\mu = 0, \sigma^2 = 10^2)$$
$$\beta \sim \text{N}(\mu = 0, \sigma^2 = 10^2)$$

Posterior:

$$\pi(X = (\alpha,\beta)|D) = \prod_{i=1}^{23} \left(\frac{\exp(\alpha + \beta t_i)}{1 + \exp(\alpha + \beta t_i)}\right)^{D_i} \left(\frac{1}{1 + \exp(\alpha + \beta t_i)}\right)^{1-D_i}$$
$$\times \exp(-\frac{1}{2}\alpha^2/10^2) \times \exp(-\frac{1}{2}\beta^2/10^2) \times \frac{1}{\text{CONST.}}$$

The cure: Monte Carlo methods

Due to the 'essential correspondence' between

density $\pi(X|D)$ & samples $\{X^{(n)} \sim \pi(X|D)\}$:

- Posterior density could be re-created via
  - histograms estimators,
  - kernel density estimators,
  - Normal mixture models, etc...

- Expectations could be approx. via

$$\mathrm{E}_{\pi(X|D)}(g(X)) \approx \frac{1}{N}\sum_{n=1}^{N} g(x^{(n)})$$

Monte Carlo methods (main idea)

- Generate an i.i.d. sample $X^{(n)} \sim \pi(\mathrm{d}X)$, for $n = 1, ..., N$
  - Inverse probability integral transform
  - Rejection sampling
  - Importance sampling

- Approx. integral $\mathrm{E}_\pi(g(X)) = \int g(X)\pi(X)\mathrm{d}X$

  with $\bar{g}^{(N)} \approx \frac{1}{N}\sum_{n=1}^{N} g(X^{(n)})$

  and standard error $\mathrm{s.e.}(\bar{g}^{(N)}) = \sqrt{\frac{1}{N}\mathrm{Var}_\pi(g(X))}$

... according to the $\sqrt{N}$-CLT

Ripley (2001). Stochastic simulation.

Markov chain Monte Carlo methods (main idea)

- Generate a Markov chain $X^{(n)} \sim P(\mathrm{d}\cdot|X^{(n-1)})$, for $n = 1, ..., N$
- Approx. integral $\mathrm{E}_\pi(g(X)) = \int g(X)\pi(X)\mathrm{d}X$

  with $\bar{g}^{(N)} \approx \frac{1}{N}\sum_{n=1}^{N} g(X^{(n)})$

  and standard error $\text{s.e.}(\bar{g}^{(N)}) = \sqrt{\frac{1}{N}\tau_g \mathrm{Var}_\pi(g(X))}$

  where $\tau_g \in (0, \infty)$ is the integrated autocorrelation time with $\tau_g = 1 + 2\sum_{k=1}^{\infty} \mathrm{Cor}(x_n, x_{n+k})$

... according to the (Markov chain) $\sqrt{N}$-CLT

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

MCMC theory

Main conditions for $P(\mathrm{d} \cdot |\cdot)$

1. Stationarity w.r.t. $\pi(\mathrm{d}\cdot)$
   - Reversibility w.r.t. $\pi(\mathrm{d}\cdot)$
2. ϕ-irredusibility
   - Hurris recurrent
3. Aperiodicity

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Stationarity

Definition: The Markov chain $\{x^{(n)}; n = 1, ..., N\}$ simulated by the transition probability $P(\mathrm{d}\cdot|\cdot)$ has stationary (or invariant) distribution $\pi(\cdot)$ iff

$$\int_{x \in \mathcal{X}} \pi(\mathrm{d}x) P(\mathrm{d}y|x) = \pi(\mathrm{d}y)$$

where $x, y \in \mathcal{X}$.

Explanation: If $x_n \sim \pi(\mathrm{d}\cdot)$ & $x_{n+1} \sim P(\mathrm{d}\cdot|x_n)$, then $x_{n+1} \sim \pi(\mathrm{d}\cdot)$

Hopefully, if we run the Markov chain (started from anywhere) for a long time, then for a long N the distribution of X_N will be approx. stationary: $X_N \overset{\text{appox.}}{\sim} \pi(\mathrm{d}\cdot)$.

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Reversibility

Definition: The Markov chain $\{x^{(n)}; n = 1, ...\}$ simulated by the transition probability $P(\mathrm{d}\cdot|\cdot)$ is reversible w.r.t distribution $\pi(\cdot)$ iff

$$\pi(\mathrm{d}x)P(\mathrm{d}y|x) = \pi(\mathrm{d}y)P(\mathrm{d}x|y)$$

where $x, y \in \mathcal{X}$.

Explanation: It expresses an equilibrium in the flow of the Markov chain: The probability of being in x and moving to y is the same as the probability of being in y and moving to x.

Property: Reversibility implies stationarity

Rational: It is more conservative assumption, but it is easier to be checked, since no integral is involved

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Reversibility implies stationarity

Proposition: If Markov chain $\{x^{(n)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is reversible w.r.t distribution $\pi(\mathrm{d}\cdot)$, then $\pi(\mathrm{d}\cdot)$ is the stationarity distr.

Prof: We compute that

$$\begin{aligned}\int_{x\in\mathcal{X}} \pi(\mathrm{d}x)P(\mathrm{d}y|x) &= \int_{x\in\mathcal{X}} \pi(\mathrm{d}y)P(\mathrm{d}x|y)\\ &= \pi(\mathrm{d}y)\int_{x\in\mathcal{X}} P(\mathrm{d}x|y)\\ &= \pi(\mathrm{d}y)\end{aligned}$$

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

ϕ-irreduciblility

Definition: The Markov chain $\{x^{(n)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is ϕ-irreducible if for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, there exists a positive integer n s.t. $P^n(A|x) > 0$, for all $x \in \mathcal{X}$, where $P^n(A|x)$ is the n-step transition probability of the Markov chain.

Explain: The Markov chain has positive probability of eventually reaching any state from any other state, in a finite number of iterations.

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Harris recurrent

Definition: The Markov chain $\{x^{(n)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is Harris recurrent if for all $A \subseteq \mathcal{X}$ with $\pi(A) > 0$ and for all $x \in \mathcal{X}$, there exists a positive integer n s.t. $P^n(A|x) = 1$, for all $x \in \mathcal{X}$, where $P^n(A|x)$ is the n-step transition probability of the Markov chain.

Explain: For all $A \subseteq \mathcal{X}$ with $\pi(A) > 0$ and for all $x \in \mathcal{X}$, the probability that the Markov chain will eventually reach B from x is 1.

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Aperiodicity

The Markov chain $\{x^{(n)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is aperiodic if there does not exist partition $\{\mathcal{X}_i; i = 1, ..., \kappa\}$ where $\pi(\mathcal{X}_i) > 0$ s.t.

- $P(\mathcal{X}_{i+1}|x) = 1$ for $x \in \mathcal{X}_i$ and
- $P(\mathcal{X}_1|x) = 1$ for $x \in \mathcal{X}_\kappa$

Christian P. Robert and George Casella (2004). Monte Carlo Statistical Methods.

Ergodicity

Theorem: If Markov chain $\{x^{(t)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is Harris recurrent, aperiodic, and has a stationary distribution $\pi(\mathrm{d}\cdot)$ then for every initial distribution $\tilde{\pi}$

$$\lim_{n\to\infty} || \int P^n(\cdot|x)\tilde{\pi}(\mathrm{d}x) - \pi(\cdot)|| = 0$$

Explain: Standard Markov chain theory tells that for any initial seed $x^{(0)}$, the realisation of the chain $\{x^{(1)}, x^{(2)}, x^{(3)}, ...\}$, provides via the ergodic theorem, a realisation of the stationary distribution since

$$x^{(n)} \to \pi(\cdot), \text{ as } n \to \infty$$

Markov chain $\sqrt{N}$−CLT

Theorem: If Markov chain $\{x^{(n)}\}$ with transition probability $P(\mathrm{d}\cdot|\cdot)$ is irreducible, aperiodic, and reversible with stationary distribution $\pi(\mathrm{d}\cdot)$ then the CLT applies: For some function $g(\cdot)$, and $\bar{g}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} g(x^{(n)})$

$$N^{-1/2}(\bar{g}^{(N)} - \mathrm{E}_\pi(g(x))) \Longrightarrow \mathrm{N}(0, \tau_g \mathrm{Var}_\pi(g(x)))$$

where $\tau_g = 1 + 2\sum_{k=1}^{\infty} \mathrm{Cor}(x_n, x_{n+k})$, if $\tau_g < \infty$.

Explain: Standard Markov chain theory tells that for any initial seed $x^{(0)}$, the realisation of the chain $\{x^{(1)}, x^{(2)}, x^{(3)}, ...\}$, provides, an approx. of the required expectations as

$$\bar{g}^{(N)} \to \mathrm{E}_\pi(g(x)), \text{ as } N \to \infty$$

where $\bar{g}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} g(x^{(n)})$.

CODE

CODE

Metropolis-Hastings: the algorithm

Simulate from a Metropolis-Hastings transition probability $P(\mathrm{d}\cdot|\cdot)$, with target distribution $\pi(\mathrm{d}\cdot)$, and proposal distribution $q(\mathrm{d}\cdot|\cdot)$.

i.e. $X^{(n+1)} \sim P(\mathrm{d}\cdot|X^{(n)})$

Given that the current state of the Markov chain is at $X^{(n)} = x$:

1. Generate a 'proposed value' x' from $q(\mathrm{d}\cdot|x)$
2. Calculate
$$a(x, x') = \min(1, \frac{\pi(x')}{\pi(x)} \frac{q(x|x')}{q(x'|x)})$$
3. With probability $a(x, x')$ accept the proposed value and set $X^{(n+1)} = x'$; otherwise reject and set $X^{(n+1)} = x$.

   Generate $u \sim \mathrm{U}(0, 1)$, and set $X^{(n+1)} = \begin{cases} x' & \text{, if } a(x, x') \geqslant u \\ x & \text{, if } a(x, x') < u \end{cases}$

Metropolis-Hastings: the transition probability

The Metropolis-Hastings transition probability is:

$$P(y|x) = q(y|x)a(x,y) + (1 - r(x))\delta_x(y)$$

where $r(x) = \int q(y|x)a(x,y)\mathrm{d}y$,

and $\delta_x(y)$ is the Dirac mass in x.

Metropolis-Hastings: Reversibility

The Metropolis-Hastings (as described above) produces a Markov chain $\{x^{(t)}\}$ which is reversible w.r.t $\pi(\mathrm{d}\cdot)$.

Prof: We need to show that

$$\pi(\mathrm{d}x)P(\mathrm{d}y|x) = \pi(\mathrm{d}y)P(\mathrm{d}x|y)$$

If $x \neq y$,

$$\begin{aligned}
\pi(\mathrm{d}x)P(\mathrm{d}y|x) &= [\pi(x)\mathrm{d}x][q(y|x)a(x,y)\mathrm{d}y] \\
&= \pi(x)q(y|x)\min(1, \frac{\pi(y)}{\pi(x)}\frac{q(x|y)}{q(y|x)})\mathrm{d}x\mathrm{d}y \\
&= \min(\pi(x)q(y|x), \pi(y)q(x|y))\mathrm{d}x\mathrm{d}y \\
&= [\pi(y)\mathrm{d}y][q(x|y)a(y,x)\mathrm{d}x] \\
&= \pi(\mathrm{d}y)P(\mathrm{d}x|y)
\end{aligned}$$

If $x = y$, then the equation is trivial.

Metropolis-Hastings: Main properties

The Metropolis-Hastings (as described above) :

- is reversible and hence admits stationary distribution $\pi(\mathrm{d}\cdot)$.
- is irreducible if

$$q(y|x) > 0, \text{ for every } x \in \mathcal{X},\ y \in \mathcal{X}$$

  since every set of $\mathcal{X}$ can be reached in a single step
- is aperiodic, if it allows events $\{X^{(t+1)} = X^{(t)}\}$,
  i.e. the probability of such an event is not zero

Metropolis-Hastings: Advantages/Challenges

Advantages:

We only need to know density $\pi(\cdot)$ up to a normalisation constant

$$a(x, x') = \min(1, \frac{\pi(x')}{\pi(x)} \frac{q(x|x')}{q(x'|x)})$$

Challenges:

If the proposal distribution $q(\mathrm{d} \cdot |\cdot)$ is poorly chosen:

- the exploration of the sampling space will be slow
- the standard error of the MC estimates will be large; because of high autocorrelations; $\tau_g = 1 + 2\sum_{k=1}^{\infty} \mathrm{Cor}(x_h, x_{h+k})$
- E.g. the number of rejections can be high

Metropolis-Hastings: Special cases

Popular special cases of the Metropolis-Hastings algorithm are:

IMH: The independence Metropolis-Hastings sampler

RWM: The Random Walk Metropolis algorithm*

MALA: The Langevin adjusted Hastings algorithm

... just different proposal distributions

Independence Metropolis-Hastings algorithm (IMH)

The proposal distribution $q(\cdot, \mathrm{d}\cdot)$ is independent on the current state i.e. $q(x'|x) = q(x')$

$X^{(n+1)} \sim P^{(\mathrm{IMH})}(\mathrm{d} \cdot |X^{(n)})$

Given that the current state of the Markov chain is at $X^{(n)} = x$:

1. Generate x' from $q(\mathrm{d}\cdot)$
2. Calculate

$$a(x, x') = \min(1, \frac{\pi(x')}{\pi(x)} \frac{q(x|x')}{q(x'|x)}$$
$$= \min(1, \frac{\pi(x')}{\pi(x)} \frac{q(x)}{q(x')})$$

3. With probability $a(x, x')$ accept and set $X^{(n+1)} = x'$; otherwise reject and set $X^{(n+1)} = x$.

Independence Metropolis-Hastings: notes

1. The proposal distribution $q(\mathrm{d}\cdot|\cdot)$ is independent on the current state; i.e. $q(x'|x) = q(x')$

2. The proposal distribution should be as close as possible to the target (stationary) distribution; i.e. $q(\cdot) \approx \pi(\cdot)$

3. Ideally, if $q(\cdot) = \pi(\cdot)$, then $a(x, x') = 1$, and the algorithm reduces to .... i.i.d. sampling from $\pi(\mathrm{d}\cdot)$

4. It is a little bit .... difficult to find 'good' $q(\cdot)$ s.t. $q(\cdot) \approx \pi(\cdot)$,

   however possible if you try hard...

   E.q.: If $\pi(\cdot)$ is uni-modal, $q(\cdot)$ can be a multivariate Normal distribution $\mathrm{N}(\mu_\pi, \Sigma_\pi)$ where μ_π, Σ_π are properly chosen.

Random walk Metropolis algorithm (RWM)

The proposal distribution $q(\mathrm{d}\cdot|\cdot)$ is s.t.

$$q(x'|x) = \mathrm{N}(x'|x, \sigma^2\mathbb{I}), \text{ for } \sigma^2 > 0$$

$$X^{(n+1)} \sim P^{(\mathrm{RWM})}(\mathrm{d}\cdot|X^{(n)})$$

Given that the current state of the Markov chain is at $X^{(n)} = x$:

1. Generate $x' \sim \mathrm{N}(x, \sigma^2\mathbb{I})$
2. Calculate

$$a(x, x') = \min(1, \frac{\pi(x')}{\pi(x)}\frac{q(x|x')}{q(x'|x)} = \min(1, \frac{\pi(x')}{\pi(x)}\cancel{\frac{\mathrm{N}(x|x', \sigma^2\mathbb{I})}{\mathrm{N}(x'|x, \sigma^2\mathbb{I})}})$$
$$= \min(1, \frac{\pi(x')}{\pi(x)})$$

3. With probability $a(x, x')$ accept and set $X^{(n+1)} = x'$; otherwise reject and set $X^{(n+1)} = x$.

Random walk Metropolis: notes 1

- Rational: "Local" exploration of the sampling space, around the neighbourhood of $X_n = x$.

$$q(x'|x) \ : \quad x' = x + \sigma z \quad ; \quad z \sim \mathrm{N}(0,1)$$

- Move towards modes of $\pi(\cdot)$ more often that moving away from them
  - "Uphill moves" are all accepted w.p. $a(x,x') = 1$
  - "Downhill moves" may be accepted w.p. $a(x,x') < 1$, or rejected w.p. $1 - a(x,x')$

- Advantages:
  - RWM is flexible: the choice of $q(\mathrm{d}\cdot|\cdot)$ is simple
  - RWM uses the previously simulated value x at stage $X^{(n)}$ to generate the proposed value x' for stage $X^{(n+1)}$.

Random walk Metropolis: notes 2

- RWM achieves optimal performance, if the proposal scale σ^2 leads to expected acceptance prob. $\bar{a}_{\text{opt}} \approx 0.234$

  ...if the components of $x := (x_1, ..., x_d)$ are independent.

  ...however this rule leads to satisfactory performance in general cases

- Variations: $q(\mathrm{d}\cdot|\cdot)$ can be any symmetric dist. s.t.

$$q(x'|x) = q(|x - x'|)$$

  E.g. $q(\mathrm{d}\cdot|\cdot)$:

  propose $\quad x' \sim \mathrm{U}(x - \sigma, x + \sigma)$.

  propose $\quad x' = x + \sigma z \quad ; \quad z \sim \mathrm{U}(-1, 1)$

Langevin adjusted Hastings algorithm (MALA)

The proposal distribution $q(\mathrm{d}\cdot|\cdot)$ is s.t.

$$q(x'|x) = \mathrm{N}(x'|x + \frac{\sigma^2}{2}\nabla\log(\pi(x)), \sigma^2\mathbb{I}), \text{ for } \sigma^2 > 0$$

$$X^{(n+1)} \sim P^{(\mathrm{MALA})}(\mathrm{d}\cdot|X^{(n)})$$

Given that the current state of the Markov chain is at $X^{(n)} = x$:

1. Generate $x' \sim \mathrm{N}(x + \frac{\sigma^2}{2}\nabla\log(\pi(x)), \sigma^2\mathbb{I})$
2. Calculate

$$\begin{aligned} a(x, x') &= \min(1, \frac{\pi(x')}{\pi(x)}\frac{q(x|x')}{q(x'|x)}) \\ &= \min(1, \frac{\pi(x')}{\pi(x)}\frac{\mathrm{N}(x|x' + \frac{\sigma^2}{2}\nabla\log(\pi(x')), \sigma^2\mathbb{I})}{\mathrm{N}(x'|x + \frac{\sigma^2}{2}\nabla\log(\pi(x)), \sigma^2\mathbb{I})}) \end{aligned}$$

3. With probability $a(x, x')$ accept and set $X^{(n+1)} = x'$; otherwise reject and set $X^{(n+1)} = x$.

Langevin adjusted Hastings: notes

1. Goal: Direct the proposed values toward areas where density $\pi(\cdot)$ is likely to be larger by using information from $\pi(\cdot)$.

2. Rational: the inclusion of $\nabla \log(\pi(\cdot))$ in the proposal centre encourages moves towards the modes of $\pi(\cdot)$

   $$q(x'|x) \; : \quad x' = x + \frac{\sigma^2}{2}\nabla \log(\pi(x)) + \sigma z \quad ; \quad z \sim \mathrm{N}(0,1)$$

3. In difficult settings, exact gradients $\nabla \log(\pi(\cdot))$ can be replaced by numerical derivatives

4. MALA achieves optimal performance, if the proposal scale σ^2 leads to expected acceptance prob. $\bar{a}_{\text{opt}} \approx 0.57$

   ...if the components of $x := (x_1, ..., x_d)$ are independent.

   ...however this rule leads to satisfactory performance in general cases

CODE

CODE

Tuning Metropolis-Hastings algorithms

The issue:

- How do we select a satisfactory proposal distr. $q(\cdot, \mathrm{d}\cdot)$?
- Well, ... this is not easy in general Metropolis-Hastings algorithm
- But, ... for some special cases, it is possible by adjusting the proposals

Recall that :

- About RWM, the σ^2 is unknown
  - RWM can achieve satisfactory performance, if the proposal scale σ^2 leads to acc. prob. $a_{\text{opt}} \approx 0.234$
- About MALA, the σ^2 is unknown
  - MALA can achieve satisfactory performance, if the proposal scale σ^2 leads to acc. prob. $a_{\text{opt}} \approx 0.57$

An adaptive scheme for RWM, MALA

Goal: Adjust the proposal scaling σ^2 in RWM or MALA algorithms

For $n = 0, 1, 2, \ldots$, iterate:

1. Simulate $X^{(n+1)}$ from $P^{(\mathrm{RWM})}_{\sigma_n^2}(\mathrm{d}\cdot|X^{(n)})$
2. Adjust σ^2 s.t. $\log(\sigma^2_{n+1}) = \log(\sigma^2_n) + \gamma_{n+1}(a^{\mathrm{RWM}}_{n+1} - \bar{a}_{\mathrm{opt}})$

Acceptance prob. of RWM at n-th iteration: a^{RWM}_n

Optimal acc. prob. value: $$a_{\mathrm{opt}} = \begin{cases} 0.234 & \text{, for RWM} \\ 0.57 & \text{, for MALA} \end{cases}$$

Gain sequence: $\gamma_n : \mathbb{N} \to \mathbb{R}_+$, a decreasing function $\gamma_n \searrow 0$

E.g. $\gamma_n = C/n^{\varsigma}$, $C > 0$, $\varsigma \in (0.5, 1)$

Christophe Andrieu and Johannes Thoms (2008). A tutorial on adaptive MCMC

Adaptive RWM and MALA algorithms: notes 1

Gain sequence γ_n

- γ_n must present a smooth slow decay
- As $n \uparrow$, $\gamma_n \downarrow$, and the influence of adaptation vanishes
- A reasonable choice is

$$\gamma_n = C/n^{\varsigma}, \quad C > 0, \quad \varsigma \in (0.5, 1)$$

  - ς controls the speed that γ_n decays to 0

Christophe Andrieu and Johannes Thoms (2008). A tutorial on adaptive MCMC

An adaptive scheme for RWM, MALA: notes 2

Rational: At state n,

- if $a_{n+1}^{\text{RWM}} < \bar{a}_{\text{opt}}$,

  $\implies \log(\sigma_{n+1}^2) < \log(\sigma_n^2)$

  $\implies \sigma_n^2$ decreases

  $\implies$ RWM/MALA will perform smaller steps at stage $n+1$

- if $a_{n+1}^{\text{RWM}} > \bar{a}_{\text{opt}}$,

  $\implies \log(\sigma_{n+1}^2) > \log(\sigma_n^2)$

  $\implies \sigma_n^2$ increases

  $\implies$ RWM/MALA will perform larger steps at stage $n+1$

Christophe Andrieu and Johannes Thoms (2008). A tutorial on adaptive MCMC

CODE

CODE

Blockwise MCMC samplers

Consider r.v. $X := (X_1, ..., X_d)$ that follows $X \sim \pi(\mathrm{d}X)$.

Challenge: In many cases, it is difficult to select appropriate proposals to construct an efficient Metropolis-Hastings algorithm that 'updates' simultaneously the whole $X := (X_1, ..., X_d)$

Reasons: X can be high dimensional.

Different X_i may have different ranges, types, etc.

etc...

Cure: 'Break' sampling of X by combining M-H algorithms targeting the conditional distributions of $\pi(\mathrm{d}\cdot)$

Blockwise MCMC sampler

Consider r.v. $X := (X_1, ..., X_d)$ that follows $X \sim \pi(\mathrm{d}X)$.

How to generate $X^{(n)} \sim P^{(\text{blockwise})}(\mathrm{d} \cdot | X^{(n-1)})$ targeting $\pi(\mathrm{d}X)$??

At iteration n:

- Simulate $X_1^{(n)} \sim P_1^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_1^{(n)} | X_2^{(n-1)}, ..., X_d^{(n-1)})$

  $\vdots$

- Simulate $X_i^{(n)} \sim P_i^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_i^{(n)} | X_1^{(n)}, ..., X_{i-1}^{(n)}, X_{i+1}^{(n-1)}, ..., X_d^{(n-1)})$

  $\vdots$

- Simulate $X_d^{(n)} \sim P_i^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_d^{(n)} | X_1^{(n)}, X_3^{(n)}, ... X_{d-1}^{(n)})$

Set $X^{(n)} = (X_1^{(n)}, X_3^{(n)}, ... X_d^{(n)})$

For example, for the i-th block

Simulate $X_i^{(n)} \sim P_i^{(\mathrm{RWM})}(\mathrm{d}\cdot|\cdot)$ targeting $\pi(\mathrm{d}X_i^{(n)}|X_1^{(n)},\ldots,X_{i-1}^{(n)},X_{i+1}^{(n-1)},\ldots,X_d^{(n-1)})$ (via RWM)

1. Generate $x' \sim \mathrm{N}(x_i^{(n-1)}, \sigma^2\mathbb{I})$
2. Calculate

$$a(x,x') = \min\left(1, \frac{\pi(x'|x_1^{(n)},\ldots,x_{i-1}^{(n)},x_{i+1}^{(n-1)},\ldots,x_d^{(n-1)})}{\pi(x_i^{(n-1)}|x_1^{(n)},\ldots,x_{i-1}^{(n)},x_{i+1}^{(n-1)},\ldots,x_d^{(n-1)})}\right)$$

$$= \min\left(1, \frac{\pi(x_1^{(n)},\ldots,x_{i-1}^{(n)},x',x_{i+1}^{(n-1)},\ldots,x_d^{(n-1)})}{\pi(x_1^{(n)},\ldots,x_{i-1}^{(n)},x_i^{(n-1)},x_{i+1}^{(n-1)},\ldots,x_d^{(n-1)})}\right.$$

3. With probability $a(x_i^{(n-1)}, x')$ accept and set $X_i^{(n)} = x'$; otherwise reject and set $X_i^{(n)} = x_i^{(n-1)}$.

Gibbs sampler (special case of Blockwise MCMC sampler)

Consider r.v. $X := (X_1, ..., X_d)$ that follows $X \sim \pi(\mathrm{d}X)$.

If all the full conditional dist of $\pi(\mathrm{d}X)$ can be sampled directly

How to $X^{(n)} \sim P^{(\text{Gibbs})}(\mathrm{d} \cdot |X^{(n-1)})$ targeting $\pi(\mathrm{d}X)$??

At iteration n:

- Simulate $X_1^{(n)} \sim \pi(\mathrm{d}X_1^{(n)}|X_2^{(n-1)}, ..., X_d^{(n-1)})$

  $\vdots$

- Simulate $X_i^{(n)} \sim \pi(\mathrm{d}X_i^{(n)}|X_1^{(n)}, ..., X_{i-1}^{(n)}, X_{i+1}^{(n-1)}, ..., X_d^{(n-1)})$

  $\vdots$

- Simulate $X_d^{(n)} \sim \pi(\mathrm{d}X_d^{(n)}|X_1^{(n)}, X_3^{(n)}, ...X_{d-1}^{(n)})$

Set $X^{(n)} = (X_1^{(n)}, X_3^{(n)}, ...X_d^{(n)})$

CODE

CODE

Blockwise MCMC sampler: notes

The blockwise MCMC sampler admits $\pi(\mathrm{d}X)$ as stationary distribution

Systematic sweep: (described above)

- The blocks are updated in a fix order: $P^{(\text{blockwise})} = P_1^{(\text{M-H})} P_2^{(\text{M-H})} ... P_d^{(\text{M-H})}$
- The Markov chain is NOT reversible

Permutation sweep:

- The blocks are updated in a random permutation order p $P^{(\text{blockwise})} = P_{p(1)}^{(\text{M-H})} P_{p(2)}^{(\text{M-H})} ... P_{p(d)}^{(\text{M-H})}$
- The Markov chain is reversible

Random sweep:

- At each iteration, randomly select and update ONLY one block $P^{(\text{blockwise})} = \frac{1}{d} \sum_{i=1}^{d} P_i^{(\text{M-H})}$
- The Markov chain is reversible

Blockwise MCMC sampler (Permutation sweep)

Consider r.v. $X := (X_1, ..., X_d)$ that follows $X \sim \pi(\mathrm{d}X)$.

How to generate $X^{(n)} \sim P^{(\text{blockwise})}(\mathrm{d} \cdot | X^{(n-1)})$ targeting $\pi(\mathrm{d}X)$??

At iteration n:

- Generate a random permutation $p = (p(1), ..., p(d))$
  - Simulate $X_{p(1)}^{(n)} \sim P_{p(1)}^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_{p(1)}^{(n)} | \text{all the rest})$

    $\vdots$
  - Simulate $X_{p(i)}^{(n)} \sim P_{p(i)}^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_{p(i)}^{(n)} | \text{all the rest})$

    $\vdots$
  - Simulate $X_{p(d)}^{(n)} \sim P_{p(d)}^{(\text{M-H})}(\mathrm{d} \cdot | \cdot)$ targeting $\pi(\mathrm{d}X_{p(d)}^{(n)} | \text{all the rest})$

Set $X^{(n)} = (X_1^{(n)}, X_3^{(n)}, ... X_d^{(n)})$

Blockwise MCMC sampler (Random sweep)

Consider r.v. $X := (X_1, ..., X_d)$ that follows $X \sim \pi(\mathrm{d}X)$.

Generate $X^{(n)} \sim P^{(\text{blockwise})}(\mathrm{d}\cdot|X^{(n-1)})$ targeting $\pi(\mathrm{d}X)$

At iteration n:

- Select block $i \sim \mathrm{U}\{1, ..., d\}$, at random
- Simulate $X_i^{(n)} \sim P_i^{(\text{M-H})}(\mathrm{d}\cdot|\cdot)$ targeting $\pi(\mathrm{d}X_i^{(n)}|\text{all the rest})$

Set $X^{(n)} = (X_1^{(n-1)}, ..., X_{i-1}^{(n-1)}, X_i^{(n)}, X_{i+1}^{(n-1)}, ..., X_d^{(n-1)})$

Improving the quality of the MCMC sample

After we generate the MCMC sample $\{X_1, X_2, X_3, ...\}$

Burn-in: Use only the generated Markov chain in the stationarity

- Discard the first few iterations (e.g. first b steps) of the Markov chain as a burn-in period
- Keep only the last tail of the Markov chain

  E.g. keep $\tilde{X} = \{X_b, X_{b+1}, X_{b+2}, X_{b+3}, ...\}$

Thinning: Try to reduce the autocorrelation by sub-sampling

- Use only every k-th element of the generated Markov chain

  E.g. use $\tilde{\tilde{X}} = \{\tilde{X}_1, \tilde{X}_{1+k}, \tilde{X}_{1+2k}, ...\} = \{X_{b+1}, X_{b+k}, X_{b+2k}, ...\}$
  ...k-step thinning

Application: Inference on what?

Compute densities:

- $\pi(\alpha, \beta|D)$
- $\pi(\alpha|D) = \int \pi(\alpha, \beta|D) \mathrm{d}\alpha,$
- $\pi(\beta|D) = \int \pi(\alpha, \beta|D) \mathrm{d}\beta$

Compute expectations:

- $\mathrm{E}(\alpha|D) = \int \alpha\pi(\alpha|D) \mathrm{d}\alpha$
- $\mathrm{E}(\beta|D) = \int \beta\pi(\beta|D) \mathrm{d}\beta$
- $$\begin{aligned} \Pr(t = 66.0|D) &= \mathrm{E}(p(t = 66.0|(\alpha, \beta))|D) \\ &= \int p(t = 66.0|(\alpha, \beta))\pi(\alpha, \beta|D) \mathrm{d}\alpha \mathrm{d}\beta \end{aligned}$$

Application: How to facilitate inference?

Generate sample: $\{(\alpha_n, \beta_n) \sim P(\mathrm{d}\cdot|(\alpha_{n-1}, \beta_{n-1})); n = 1, ..., N\}$
Estimate densities:

- $\hat{\pi}(\alpha, \beta|D)$: with the histogram of $\{(\alpha_n, \beta_n); n = 1, 2, ..., N\}$
- $\hat{\pi}(\alpha|D)$: with the histogram of $\{\alpha_n; n = 1, 2, ..., N\}$
- $\hat{\pi}(\beta|D)$: with the histogram of $\{\beta_n; n = 1, 2, ..., N\}$

Estimate expectations:

- $\widehat{\mathrm{E}(\alpha|D)} = \bar{\alpha}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} \alpha_n$, with $\mathrm{s.e.}(\widehat{\mathrm{E}(\alpha|D)}) = \sqrt{\frac{1}{N} s_\alpha^2 \tau_\alpha}$
- $\widehat{\mathrm{E}(\beta|D)} = \bar{\beta}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} \beta_n$, with $\mathrm{s.e.}(\widehat{\mathrm{E}(\beta|D)}) = \sqrt{\frac{1}{N} s_\beta^2 \tau_\beta}$
- $\widehat{\mathrm{Pr}}(t = 66.0|D) = \bar{p}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} \overbrace{p(t = 66.0|(\alpha_n, \beta_n))}^{=p_n}$

  with $\mathrm{s.e.}(\widehat{\mathrm{Pr}}(t = 66.0|D)) = \sqrt{\frac{1}{N} s_p^2 \tau_p}$