Markov Chain Monte Carlo Methods

Christian P. Robert

Université Paris-Dauphine, University of Warwick, & CREStt
CIRM Master Class in Bayesian Statistics

October 22, 2018

Outline

The Metropolis-Hastings Algorithm

Running Monte Carlo via Markov Chains

Far from necessary to use a sample from the distribution f to approximate the integral

$$\mathfrak{I} = \int h(x) f(x) dx \ ,$$

as sample from different target(s) g can be exploited in importance sampling version:

$$\hat{\mathfrak{I}} = {}^{1}\!/_{n} \sum_{i=1}^{n} h(x_i) {}^{f(x_i)}\!/_{g(x_i)}$$

Running Monte Carlo via Markov Chains

Far from necessary to use a sample from the distribution f to approximate the integral

$$\mathfrak{I} = \int h(x) f(x) dx \ ,$$

as non-i.i.d. sample $X_1, \ldots, X_n \sim f$ produced by using an ergodic Markov chain with stationary distribution f

Running Monte Carlo via Markov Chains (2)

Idea

For an arbitrary starting value $x^{(0)}$, an ergodic chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution f

Running Monte Carlo via Markov Chains (2)

Idea

For an arbitrary starting value $x^{(0)}$, an ergodic chain $(X^{(t)})$ is generated using a transition kernel with stationary distribution f

- Insures the convergence in distribution of $(X^{(t)})$ to a random variable from f.
- For a "large enough" T_0, $X^{(T_0)}$ can be considered as distributed from limiting distribution f
- Produce a *dependent* sample $X^{(T_0)}, X^{(T_0+1)}, \ldots$, which is generated from f, sufficient for most approximation purposes

Running Monte Carlo via Markov Chains (2)

- Insures the convergence in distribution of $(X^{(t)})$ to a random variable from f.
- For a "large enough" T_0, $X^{(T_0)}$ can be considered as distributed from limiting distribution f
- Produce a *dependent* sample $X^{(T_0)}, X^{(T_0+1)}, \ldots$, which is generated from f, sufficient for most approximation purposes

Problem:

How can one build a Markov chain with a given stationary distribution?

The Metropolis–Hastings algorithm

Basics

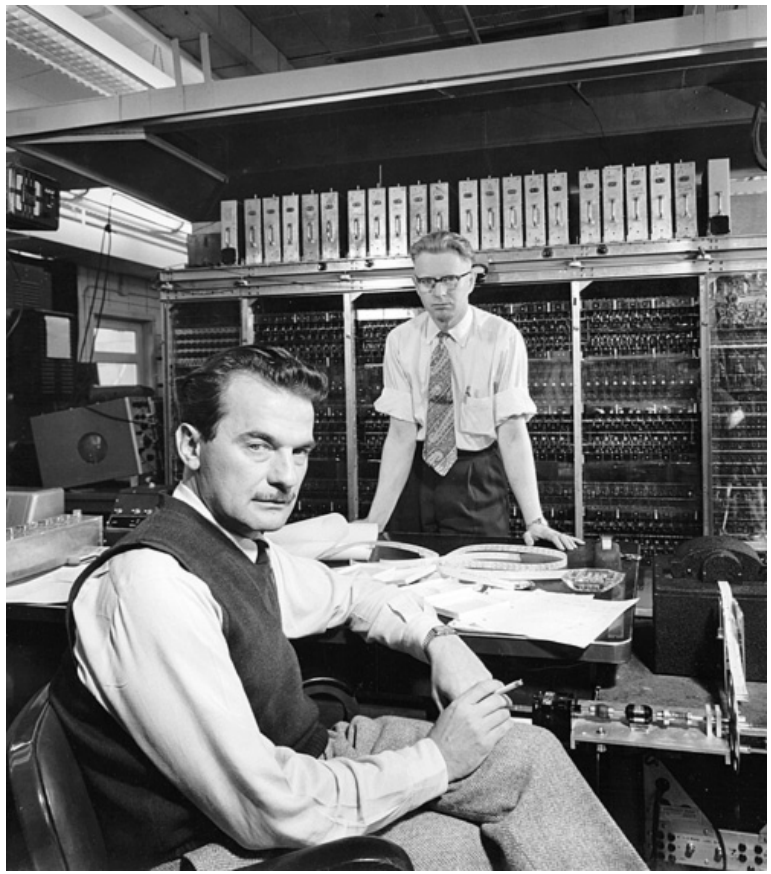
The algorithm uses the **objective (target) density**

$$f$$

and a conditional density

$$q(y|x)$$

called the **instrumental (or proposal) distribution**



[Metropolis & al., 1953]

The MH algorithm

Algorithm (Metropolis–Hastings)

Given $x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with prob. } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

$$\rho(x, y) = \min \left\{ \frac{f(y)}{f(x)} \, \frac{q(x|y)}{q(y|x)} \, , 1 \right\} .$$

Features

- Independent of normalizing constants for both f and $q(\cdot|x)$ (ie, those constants independent of x)
- Never move to values with $f(y) = 0$
- The chain $(x^{(t)})_t$ may take the same value several times in a row, even though f is a density wrt Lebesgue measure
- The sequence $(y_t)_t$ is usually **not** a Markov chain

Convergence properties

1. The M-H Markov chain is **reversible**, with invariant/stationary density f since it satisfies the **detailed balance condition**

$$\boxed{f(y)\,K(y,x) = f(x)\,K(x,y)}$$

Convergence properties

1. The M-H Markov chain is **reversible**, with invariant/stationary density f since it satisfies the **detailed balance condition**

$$\boxed{f(y)\,K(y,x) = f(x)\,K(x,y)}$$

2. As f is a probability measure, the chain is **positive recurrent**

Convergence properties

1. The M-H Markov chain is **reversible**, with invariant/stationary density f since it satisfies the **detailed balance condition**

$$\boxed{f(y)\,K(y,x) = f(x)\,K(x,y)}$$

2. As f is a probability measure, the chain is **positive recurrent**
3. If

$$\Pr\left[\frac{f(Y_t)\;q(X^{(t)}|Y_t)}{f(X^{(t)})\;q(Y_t|X^{(t)})} \geq 1\right] < 1. \qquad (1)$$

that is, the event $\{X^{(t+1)} = X^{(t)}\}$ is possible, then the chain is **aperiodic**

Convergence properties (2)

4. If

$$q(y|x) > 0 \text{ for every } (x, y), \tag{2}$$

the chain is **irreducible**

Convergence properties (2)

4. If

$$q(y|x) > 0 \text{ for every } (x, y), \tag{2}$$

   the chain is **irreducible**

5. For M-H, f-irreducibility implies **Harris recurrence**

Convergence properties (2)

4. If
$$q(y|x) > 0 \text{ for every } (x, y), \tag{2}$$
the chain is **irreducible**
5. For M-H, f-irreducibility implies **Harris recurrence**
6. Thus, for M-H satisfying (1) and (2)
   (i) For h, with $\mathbb{I}_f|h(X)| < \infty$,
$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)df(x) \qquad \text{a.e. } f.$$
   (ii) and
$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\mu(dx) - f \right\|_{TV} = 0$$
for every initial distribution μ, where $K^n(x,\cdot)$ denotes the kernel for n transitions.

The Independent Case

The instrumental distribution q is independent of $X^{(t)}$, and is denoted g by analogy with Accept-Reject.

The Independent Case

The instrumental distribution q is independent of $X^{(t)}$, and is denoted g by analogy with Accept-Reject.

Algorithm (Independent Metropolis-Hastings)

Given $x^{(t)}$,

a Generate $Y_t \sim g(y)$

b Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \min\left\{ \dfrac{f(Y_t)\, g(x^{(t)})}{f(x^{(t)})\, g(Y_t)}\, , 1 \right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Properties

The resulting sample is **not** iid

Properties

The resulting sample is **not** iid but there exist strong convergence properties:

Theorem (Ergodicity)

The algorithm produces a uniformly ergodic chain if there exists a constant M such that

$$f(x) \leq Mg(x)\,, \quad x \in \text{supp } f.$$

In this case,

$$\|K^n(x,\cdot) - f\|_{TV} \leq \left(1 - \frac{1}{M}\right)^n \,.$$

[Mengersen & Tweedie, 1996]

Illustration

Example (Noisy AR(1))

Hidden Markov chain from a regular AR(1) model,

$$x_{t+1} = \varphi x_t + \epsilon_{t+1} \qquad \epsilon_t \sim \mathcal{N}(0, \tau^2)$$

and observables

$$y_t | x_t \sim \mathcal{N}(x_t^2, \sigma^2)$$

Illustration

Example (Noisy AR(1))

Hidden Markov chain from a regular AR(1) model,

$$x_{t+1} = \varphi x_t + \epsilon_{t+1} \qquad \epsilon_t \sim \mathcal{N}(0, \tau^2)$$

and observables

$$y_t | x_t \sim \mathcal{N}(x_t^2, \sigma^2)$$

The distribution of x_t given x_{t-1}, x_{t+1} and y_t is

$$\exp \frac{-1}{2\tau^2} \left\{ (x_t - \varphi x_{t-1})^2 + (x_{t+1} - \varphi x_t)^2 + \frac{\tau^2}{\sigma^2} (y_t - x_t^2)^2 \right\} .$$

Illustration

Example (Noisy AR(1))

Use for proposal the $\mathcal{N}(\mu_t, \omega_t^2)$ distribution, with

$$\mu_t = \varphi \frac{x_{t-1} + x_{t+1}}{1 + \varphi^2} \quad \text{and} \quad \omega_t^2 = \frac{\tau^2}{1 + \varphi^2}\,.$$

Illustration

Example (Noisy AR(1))

Use for proposal the $\mathcal{N}(\mu_t, \omega_t^2)$ distribution, with

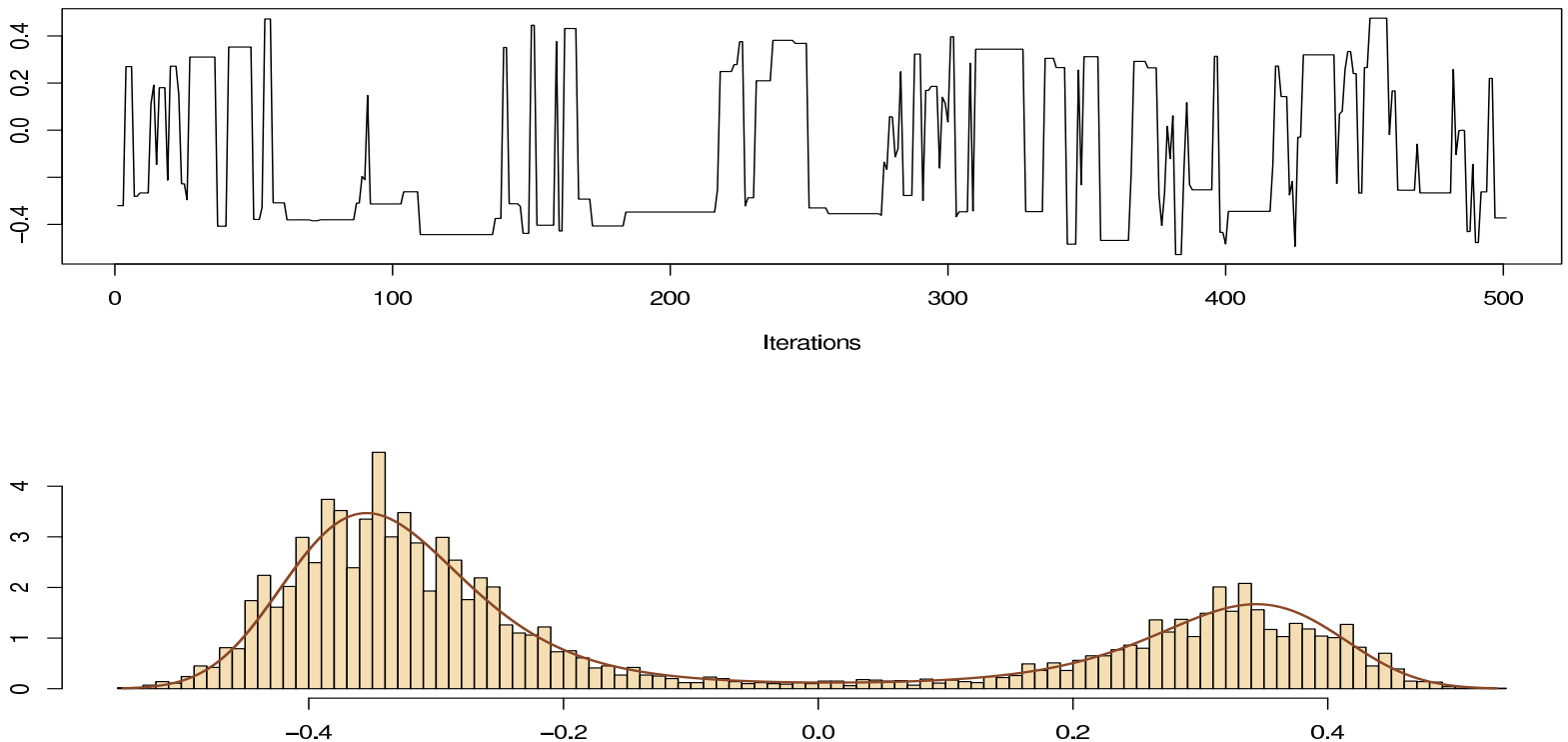
$$\mu_t = \varphi \frac{x_{t-1} + x_{t+1}}{1+\varphi^2} \quad \text{and} \quad \omega_t^2 = \frac{\tau^2}{1+\varphi^2}\,.$$

Ratio

$$\pi(x)/q_{\text{ind}}(x) = \exp -(y_t - x_t^2)^2/2\sigma^2$$

is bounded

Illustration



(top) Last 500 realisations of the chain $\{X_k\}_k$ out of $10,000$ iterations; (bottom) histogram of the chain, compared with the target distribution.

Random walk Metropolis–Hastings

Use of a local perturbation as proposal

$$Y_t = X^{(t)} + \varepsilon_t,$$

where $\varepsilon_t \sim g$, independent of $X^{(t)}$.
The instrumental density is now of the form $g(y - x)$ and the Markov chain is a random walk if we take g to be *symmetric* $g(x) = g(-x)$

Random walk Metropolis–Hastings

Algorithm (Random walk Metropolis)

Given $x^{(t)}$

1. Generate $Y_t \sim g(y - x^{(t)})$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \min\left\{1, \dfrac{f(Y_t)}{f(x^{(t)})}\right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Random walk Metropolis–Hastings

Example (Random walk and normal target)

Generate $\mathcal{N}(0,1)$ based on the uniform proposal $[-\delta, \delta]$

[Hastings (1970)]

The probability of acceptance is then

$$\rho(x^{(t)}, y_t) = \exp\{(x^{(t)^2} - y_t^2)/2\} \wedge 1.$$

Random walk Metropolis–Hastings

Example (Mixture models)

$$\pi(\theta|x) \propto \prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} p_\ell f(x_j|\mu_\ell, \sigma_\ell) \right) \pi(\theta)$$

Random walk Metropolis–Hastings

Example (Mixture models)

$$\pi(\theta|x) \propto \prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} p_\ell f(x_j|\mu_\ell, \sigma_\ell) \right) \pi(\theta)$$

Metropolis-Hastings proposal:

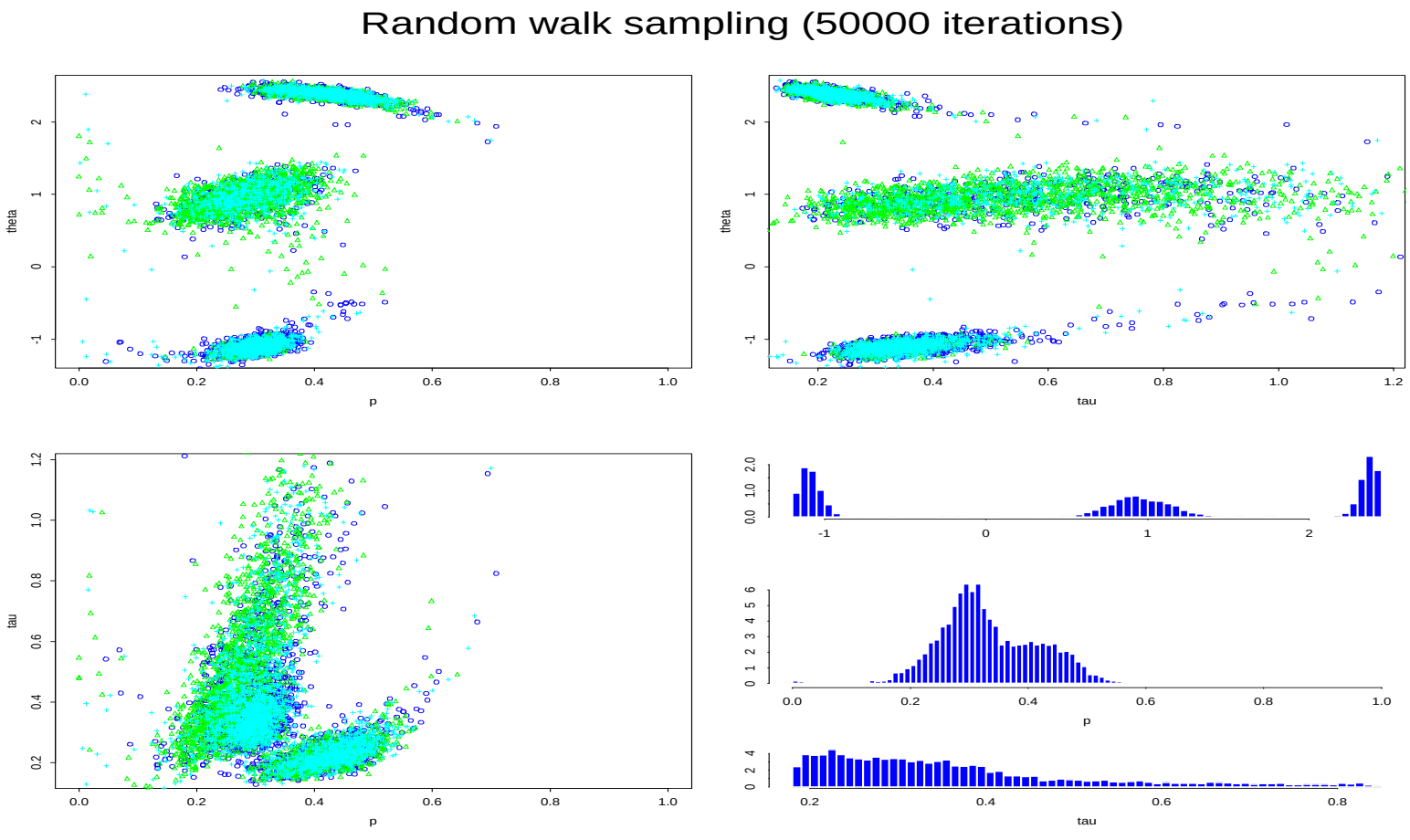
$$\theta^{(t+1)} = \begin{cases} \theta^{(t)} + \omega\varepsilon^{(t)} & \text{if } u^{(t)} < \rho^{(t)} \\ \theta^{(t)} & \text{otherwise} \end{cases}$$

where

$$\rho^{(t)} = \frac{\pi(\theta^{(t)} + \omega\varepsilon^{(t)}|x)}{\pi(\theta^{(t)}|x)} \wedge 1$$

and ω scaled for good acceptance rate

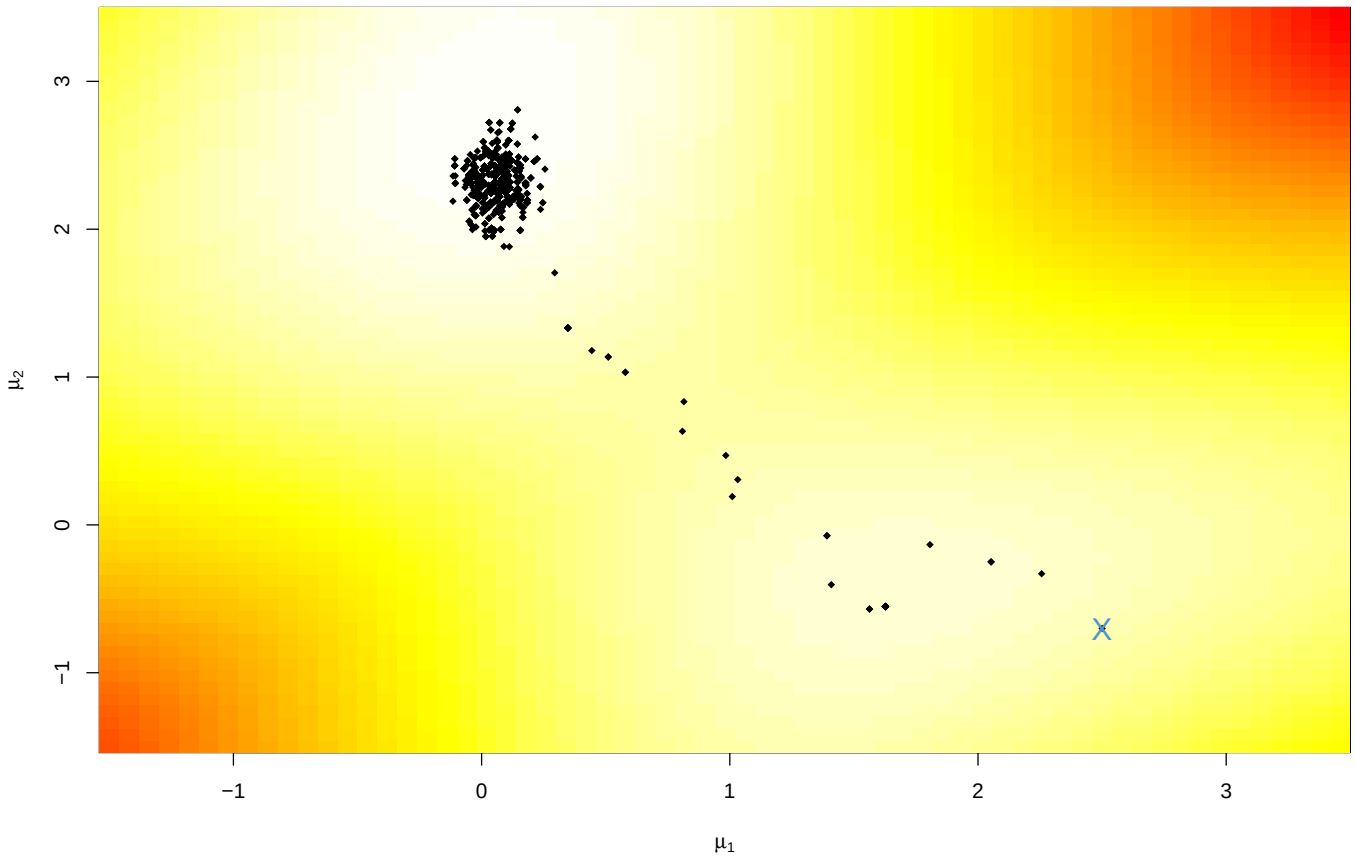
Random walk Metropolis–Hastings



General case of a 3 component normal mixture

[Celeux & al., 2000]

Random walk Metropolis–Hastings



Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$

Random walk Metropolis–Hastings

Example (probit model)

Likelihood of the **probit model**

$$\prod_{i=1}^{n} \Phi(y_i^{\mathsf{T}}\beta)^{x_i}\, \Phi(-y_i^{\mathsf{T}}\beta)^{1-x_i}$$

Random walk Metropolis–Hastings

Example (probit model)

Likelihood of the **probit model**

$$\prod_{i=1}^{n} \Phi(y_i^{\mathsf{T}}\beta)^{x_i}\, \Phi(-y_i^{\mathsf{T}}\beta)^{1-x_i}$$

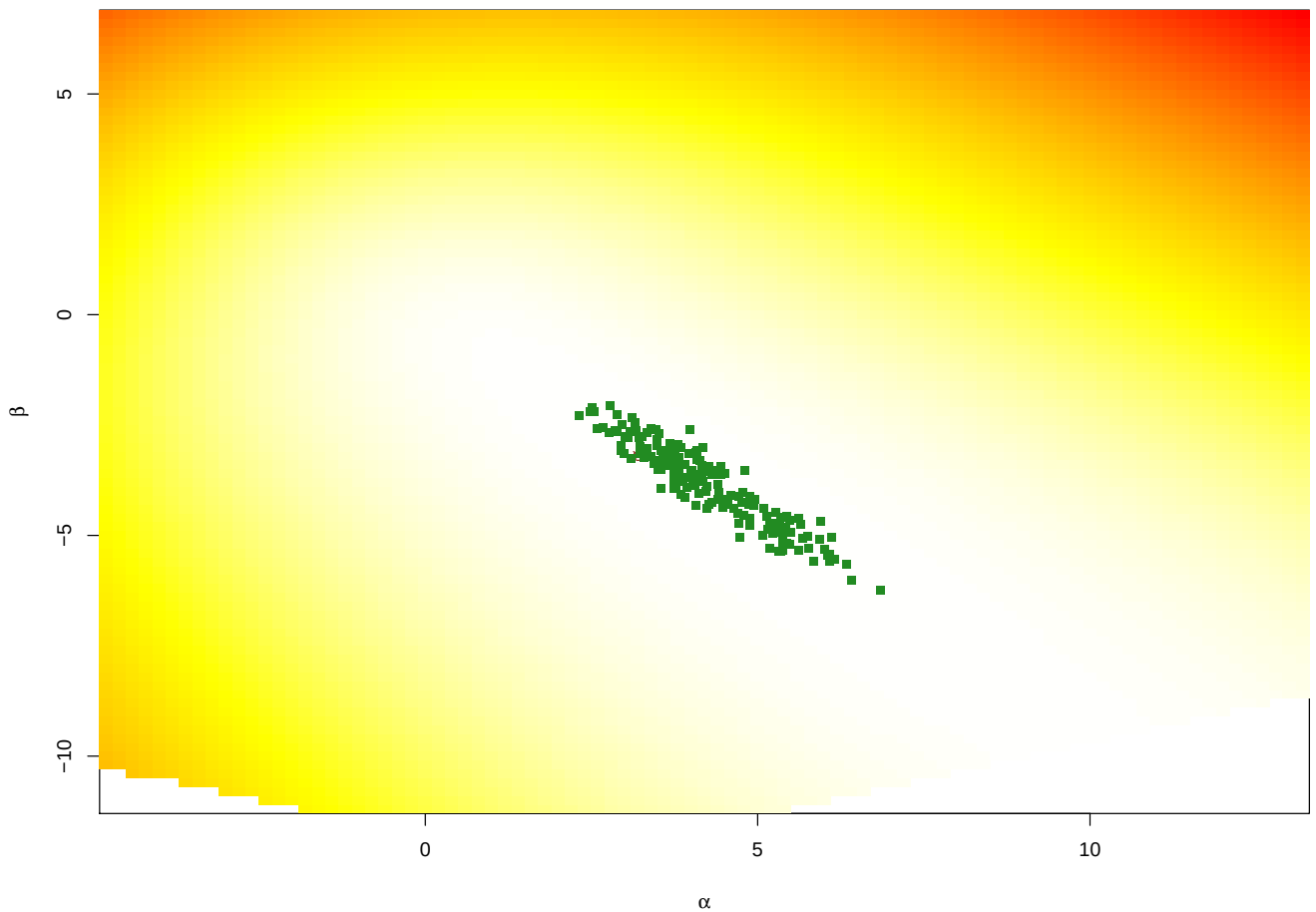
Random walk proposal

$$\beta^{(t+1)} = \beta^{(t)} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}_p(0, \Sigma)$$

where, for instance,

$$\Sigma = \alpha(YY^{\mathsf{T}})^{-1}$$

Random walk Metropolis–Hastings



Likelihood surface and random walk Metropolis-Hastings steps

Convergence properties

Uniform ergodicity prohibited by random walk structure

Convergence properties

Uniform ergodicity prohibited by random walk structure
At best, geometric ergodicity:

Theorem (Sufficient ergodicity)

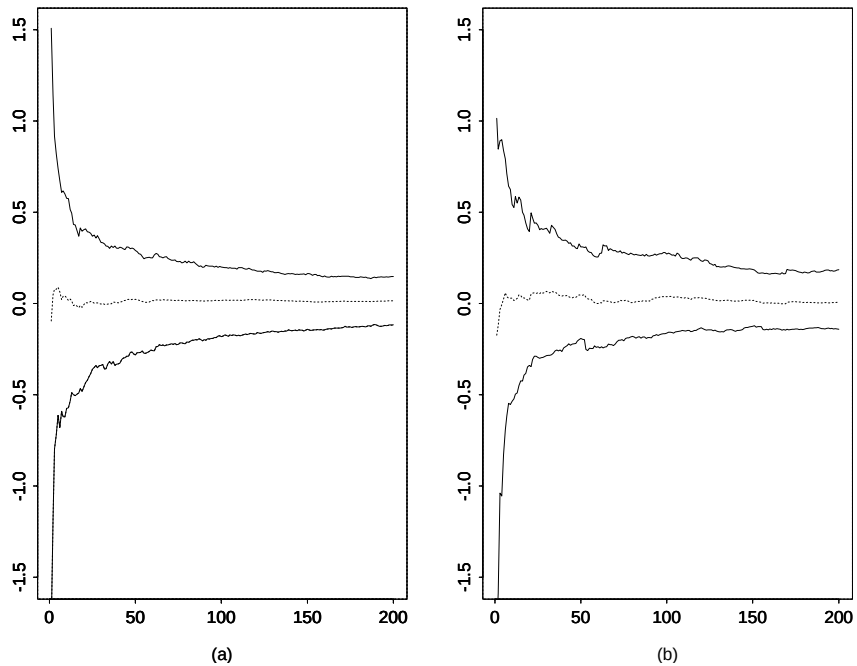
For a symmetric density f, log-concave in the tails, and a positive and symmetric density g, the chain $(X^{(t)})$ is geometrically ergodic.
[Mengersen & Tweedie, 1996]

▸ no tail effect

Convergence properties

Example (Comparison of tail effects)

Random-walk Metropolis–Hastings algorithms based on a $\mathcal{N}(0,1)$ instrumental for the generation of (a) a $\mathcal{N}(0,1)$ distribution and (b) a distribution with density $\psi(x) \propto (1+|x|)^{-3}$



90% confidence envelopes of the means, derived from 500 parallel independent chains

Convergence properties

Example (Cauchy by normal)

Cauchy $\mathscr{C}(0,1)$ target and Gaussian random walk proposal, $\xi' \sim \mathscr{N}(\xi, \sigma^2)$, with acceptance probability

$$\frac{1+\xi^2}{1+(\xi')^2} \wedge 1\,,$$

Overall fit of the Cauchy density by the histogram satisfactory, but poor exploration of the tails: 99% quantile of $\mathscr{C}(0,1)$ equal to 3, but no simulation exceeds 14 out of $10,000$!

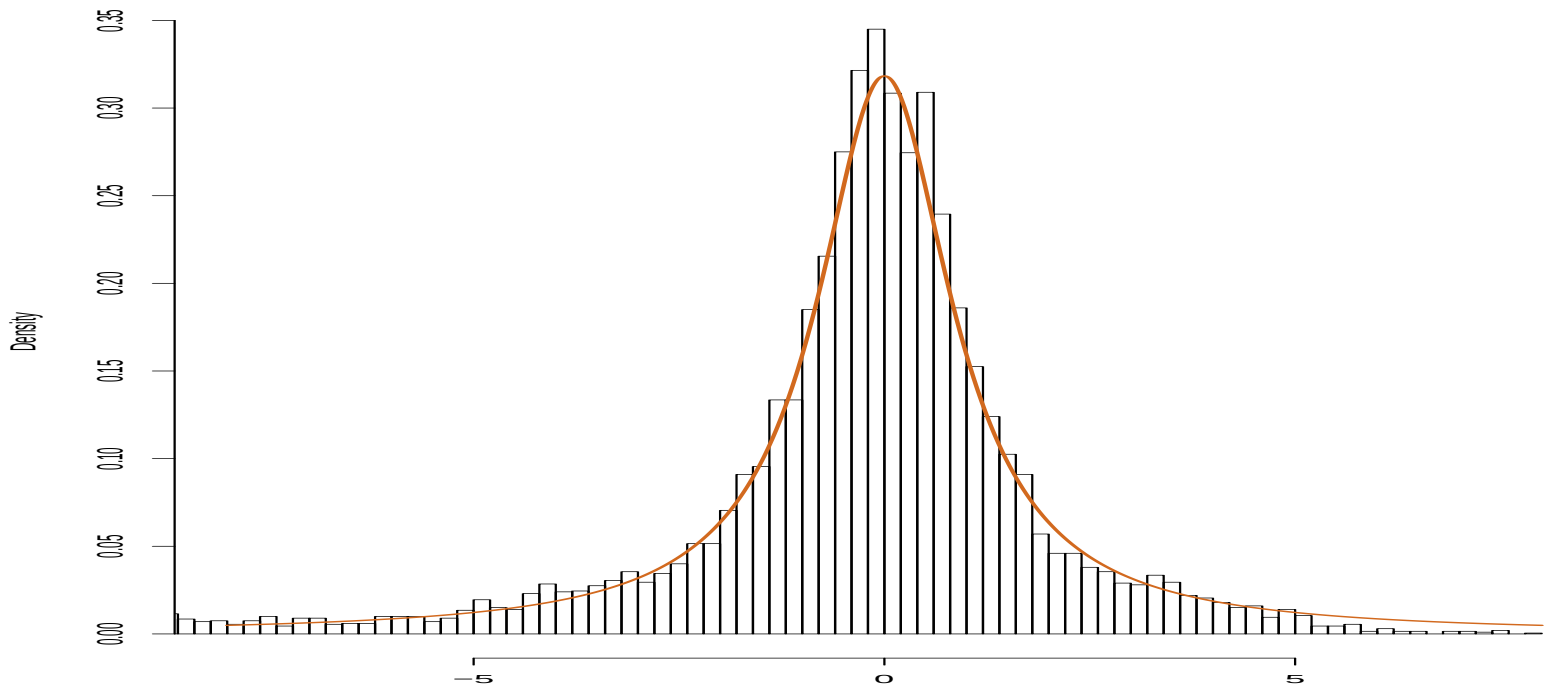
[Roberts & Tweedie, 2004]

Convergence properties

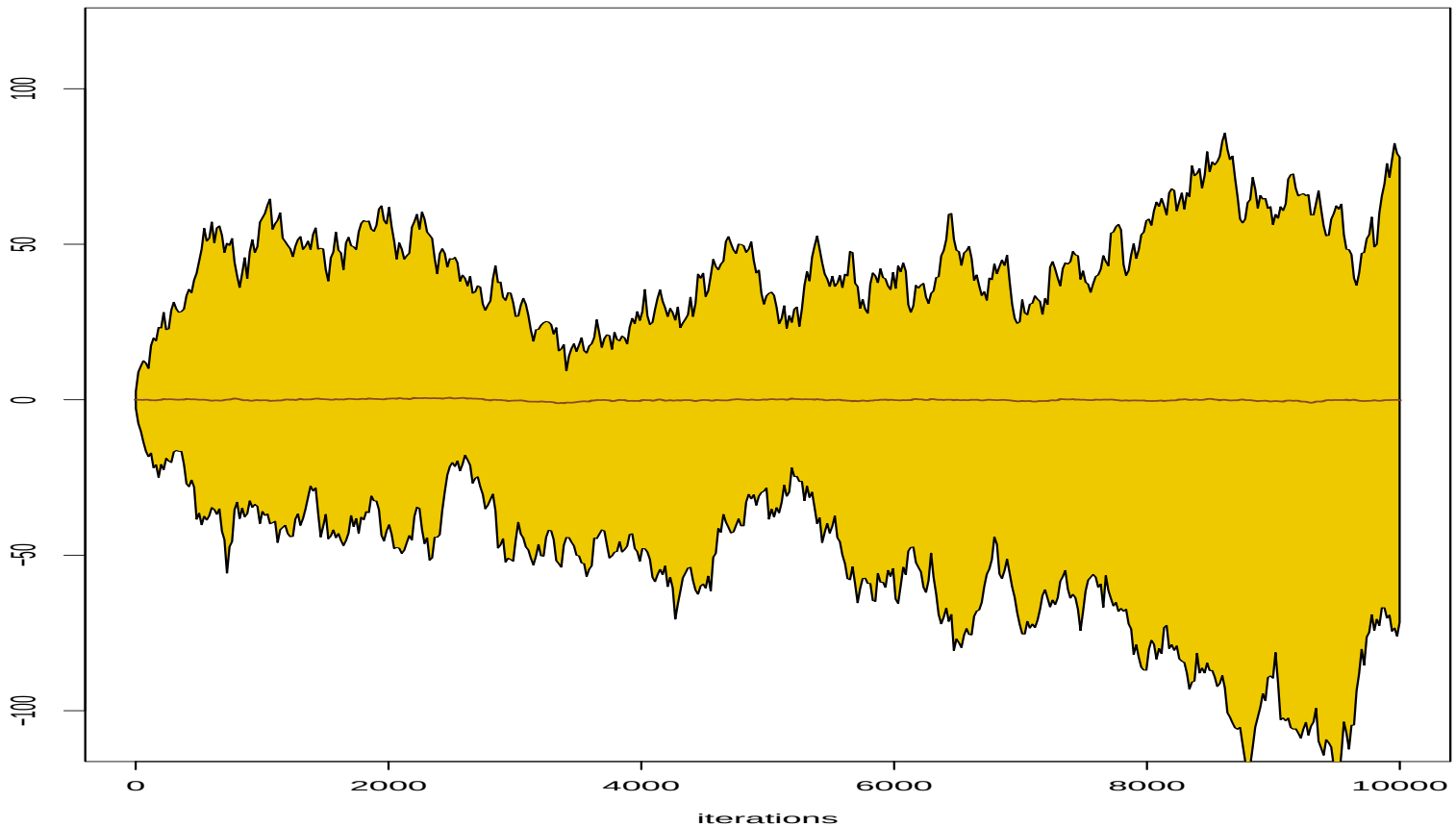
Again, lack of geometric ergodicity!

[Mengersen & Tweedie, 1996]

Slow convergence shown by the non-stable range after $10,000$ iterations.



Convergence properties



Range of 500 parallel runs for the same setup

Comments

- **[CLT, Rosenthal's inequality...]** h-ergodicity implies CLT for additive (possibly unbounded) functionals of the chain, Rosenthal's inequality and so on...
- **[Control of the moments of the return-time]** The condition implies (because $h \geq 1$) that

$$\sup_{x \in C} \mathbb{E}_x[r_0(\tau_C)] \leq \sup_{x \in C} \mathbb{E}_x \left\{ \sum_{k=0}^{\tau_C - 1} r(k) h(X_k) \right\} < \infty,$$

where $r_0(n) = \sum_{l=0}^{n} r(l)$ Can be used to derive bounds for the coupling time, an essential step to determine computable bounds, using coupling inequalities

[Roberts & Tweedie, 1998; Fort & Moulines, 2000]

Extensions

There are many other families of HM algorithms

- *Adaptive Rejection Metropolis Sampling*
- *Reversible Jump (later!)*
- *Langevin algorithms*
- *not HMC*

to name just a few...

Langevin Algorithms

Proposal based on the *Langevin diffusion* L_t is defined by the stochastic differential equation

$$dL_t = dB_t + \frac{1}{2}\nabla \log f(L_t)dt,$$

where B_t is the standard *Brownian motion*

Theorem

The Langevin diffusion is the only non-explosive diffusion which is reversible with respect to f.

Discretization

Instead, consider the sequence

$$x^{(t+1)} = x^{(t)} + \frac{\sigma^2}{2}\nabla \log f(x^{(t)}) + \sigma\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}_p(0, I_p)$$

where σ^2 corresponds to the discretization step

Discretization

Instead, consider the sequence

$$x^{(t+1)} = x^{(t)} + \frac{\sigma^2}{2}\nabla \log f(x^{(t)}) + \sigma\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}_p(0, I_p)$$

where σ^2 corresponds to the discretization step
Unfortunately, the discretized chain may be be transient, for instance when

$$\lim_{x\to\pm\infty} \left|\sigma^2 \nabla \log f(x)|x|^{-1}\right| > 1$$

MH correction (MALA)

Accept the new value Y_t with probability

$$\frac{f(Y_t)}{f(x^{(t)})} \cdot \frac{\exp\left\{-\left\|Y_t - x^{(t)} - \frac{\sigma^2}{2}\nabla \log f(x^{(t)})\right\|^2 \Big/ 2\sigma^2\right\}}{\exp\left\{-\left\|x^{(t)} - Y_t - \frac{\sigma^2}{2}\nabla \log f(Y_t)\right\|^2 \Big/ 2\sigma^2\right\}} \wedge 1 \,.$$

Choice of the scaling factor σ

Should lead to an acceptance rate of 0.574 to achieve optimal convergence rates (when the components of x are uncorrelated)

[Roberts & Rosenthal, 1998]

Optimizing the Acceptance Rate

Problem of choice of the transition kernel from a practical point of view

Most common alternatives:

(a) a fully automated algorithm like ARMS;

(b) an instrumental density g which approximates f, such that f/g is bounded for uniform ergodicity to apply;

(c) a random walk

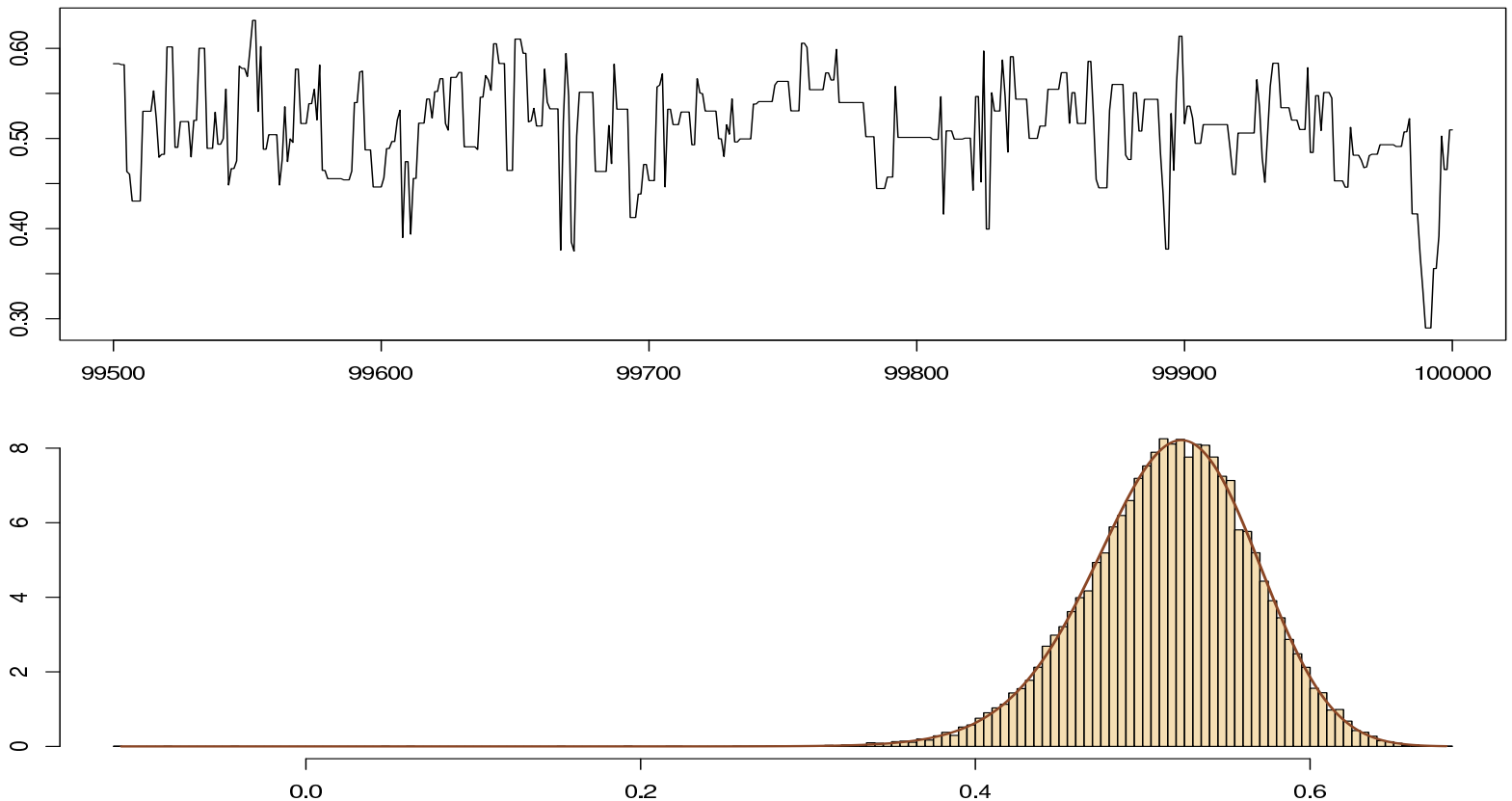
In both cases (b) and (c), the choice of g is critical,

Optimizing the Acceptance Rate

Example (Noisy AR(1))

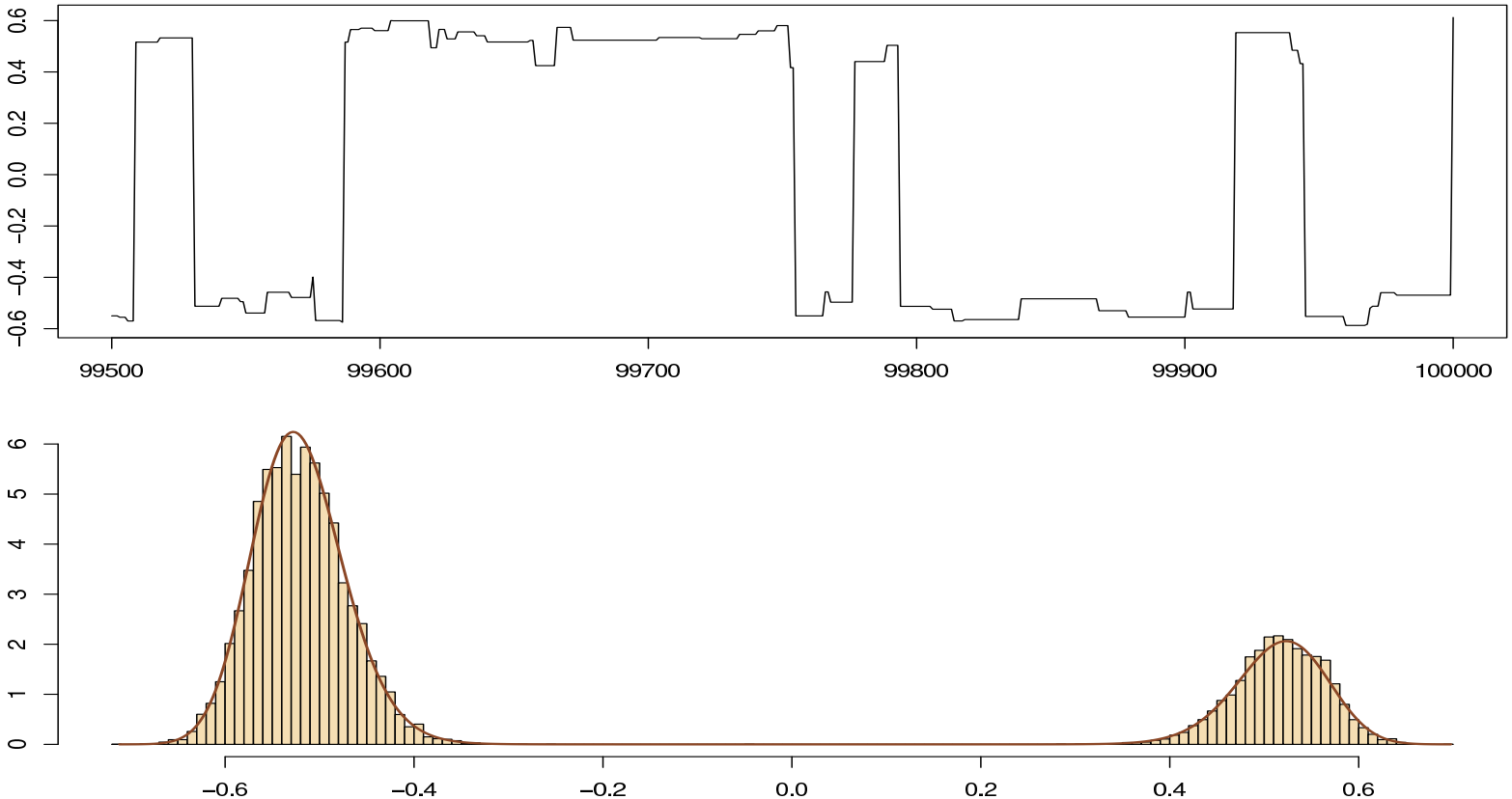
For a Gaussian random walk with scale ω small enough, the random walk never jumps to the other mode. But if the scale ω is sufficiently large, the Markov chain explores both modes and give a satisfactory approximation of the target distribution.

Optimizing the Acceptance Rate



Markov chain based on a random walk with scale $\omega = .1$.

Optimizing the Acceptance Rate



Markov chain based on a random walk with scale $\omega = .5$.

Some properties of the Metropolis–Hastings algorithm

Alternative representation of Metropolis–Hastings estimator δ as

$$\delta = \frac{1}{n}\sum_{t=1}^{n} h(x^{(t)}) = \frac{1}{n}\sum_{i=1}^{M_n} \mathfrak{n}_i h(\mathfrak{z}_i)\,,$$

where

- $\mathfrak{z}_i$'s are the accepted y_j's,
- M_n is the number of accepted y_j's till time n,
- $\mathfrak{n}_i$ is the number of times $\mathfrak{z}_i$ appears in the sequence $(x^{(t)})_t$.

The "accepted candidates"

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
To simulate from $\tilde{q}(\cdot|\mathfrak{z}_i)$

1. Propose a candidate $y \sim q(\cdot|\mathfrak{z}_i)$
2. Accept with probability

$$\tilde{q}(y|\mathfrak{z}_i) \Big/ \left(\frac{q(y|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}\right) = \alpha(\mathfrak{z}_i, y)$$

Otherwise, reject it and starts again.

The "accepted candidates"

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \underbrace{\frac{\pi(x)p(x)}{\int \pi(u)p(u)\mathrm{d}u}}_{\tilde{\pi}(x)} \underbrace{\frac{\alpha(x,y)q(y|x)}{p(x)}}_{\tilde{q}(y|x)}$$

The "accepted candidates"

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \frac{\pi(x)\alpha(x,y)q(y|x)}{\int \pi(u)p(u)\mathrm{d}u}$$

The "accepted candidates"

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i, \cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \frac{\pi(y)\alpha(y,x)q(x|y)}{\int \pi(u)p(u)\mathrm{d}u}$$

The "accepted candidates"

Define

$$\tilde{q}(\cdot|\mathfrak{z}_i) = \frac{\alpha(\mathfrak{z}_i,\cdot)\, q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \leq \frac{q(\cdot|\mathfrak{z}_i)}{p(\mathfrak{z}_i)}$$

where $p(\mathfrak{z}_i) = \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i) dy$
The transition kernel $\tilde{q}$ admits $\tilde{\pi}$ as a stationary distribution:

$$\tilde{\pi}(x)\tilde{q}(y|x) = \tilde{\pi}(y)\tilde{q}(x|y)\,,$$

The "accepted chain"

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*

The "accepted chain"

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*

The "accepted chain"

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,;$$

The "accepted chain"

Lemma (Douc & X., *AoS*, 2011)

The sequence $(\mathfrak{z}_i, \mathfrak{n}_i)$ *satisfies*

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ *is a Markov chain;*
2. $\mathfrak{z}_{i+1}$ *and* $\mathfrak{n}_i$ *are* independent *given* $\mathfrak{z}_i$*;*
3. $\mathfrak{n}_i$ *is distributed as a geometric random variable with probability parameter*

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,;$$

4. $(\mathfrak{z}_i)_i$ *is a Markov chain with transition kernel* $\tilde{Q}(\mathfrak{z}, dy) = \tilde{q}(y|\mathfrak{z})dy$ *and stationary distribution* $\tilde{\pi}$ *such that*

$$\tilde{q}(\cdot|\mathfrak{z}) \propto \alpha(\mathfrak{z}, \cdot)\, q(\cdot|\mathfrak{z}) \quad \text{and} \quad \tilde{\pi}(\cdot) \propto \pi(\cdot)p(\cdot)\,.$$

Importance sampling perspective

1. A natural idea:

$$\delta^* = \frac{1}{n} \sum_{i=1}^{M_n} \frac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)} \,,$$

Importance sampling perspective

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \frac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \frac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} .$$

Importance sampling perspective

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \frac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \frac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} .$$

2. But p not available in closed form.

Importance sampling perspective

1. A natural idea:

$$\delta^* \simeq \frac{\sum_{i=1}^{M_n} \frac{h(\mathfrak{z}_i)}{p(\mathfrak{z}_i)}}{\sum_{i=1}^{M_n} \frac{1}{p(\mathfrak{z}_i)}} = \frac{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)} h(\mathfrak{z}_i)}{\sum_{i=1}^{M_n} \frac{\pi(\mathfrak{z}_i)}{\tilde{\pi}(\mathfrak{z}_i)}} .$$

2. But p not available in closed form.
3. The geometric $\mathfrak{n}_i$ is the replacement, an obvious solution that is used in the original Metropolis–Hastings estimate since $\mathbb{E}[\mathfrak{n}_i] = 1/p(\mathfrak{z}_i)$.

The Bernoulli factory

The crude estimate of $1/p(\mathfrak{z}_i)$,

$$\mathfrak{n}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \mathbb{I}\{u_\ell \ge \alpha(\mathfrak{z}_i, y_\ell)\} \ ,$$

can be improved:

Lemma (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$, the quantity

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \le j} \{1 - \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ which variance, conditional on $\mathfrak{z}_i$, is lower than the conditional variance of $\mathfrak{n}_i$, $\{1 - p(\mathfrak{z}_i)\}/p^2(\mathfrak{z}_i)$.

Rao-Blackwellised, for sure?

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \leq j} \{1 - \alpha(\mathfrak{z}_i, y_\ell)\}$$

1. Infinite sum but finite with at least positive probability:

$$\alpha(x^{(t)}, y_t) = \min \left\{ 1, \frac{\pi(y_t)}{\pi(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})} \right\}$$

   For example: take a symmetric random walk as a proposal.
2. What if we wish to be sure that the sum is finite?

Rao-Blackwellised, for sure?

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \leq j} \{1 - \alpha(\mathfrak{z}_i, y_\ell)\}$$

1. Infinite sum but finite with at least positive probability:

$$\alpha(x^{(t)}, y_t) = \min \left\{ 1, \frac{\pi(y_t)}{\pi(x^{(t)})} \frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})} \right\}$$

   For example: take a symmetric random walk as a proposal.

2. What if we wish to be sure that the sum is finite?

Finite horizon k version:

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

Variance improvement

Theorem (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ with an almost sure finite number of terms.

Variance improvement

Theorem (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|z_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(z_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(z_i, y_\ell)\}$$

is an unbiased estimator of $1/p(z_i)$ with an almost sure finite number of terms. Moreover, for $k \geq 1$,

$$\mathbb{V}\hat{\xi}_i^k z_i = \frac{1 - p(z_i)}{p^2(z_i)} - \frac{1 - (1 - 2p(z_i) + r(z_i))^k}{2p(z_i) - r(z_i)} \left(\frac{2 - p(z_i)}{p^2(z_i)} \right) (p(z_i) - r(z_i)) ,$$

where $p(z_i) := \int \alpha(z_i, y)\, q(y|z_i)\, dy$. and $r(z_i) := \int \alpha^2(z_i, y)\, q(y|z_i)\, dy$.

Variance improvement

Theorem (Douc & X., *AoS*, 2011)

If $(y_j)_j$ is an iid sequence with distribution $q(y|\mathfrak{z}_i)$ and $(u_j)_j$ is an iid uniform sequence, for any $k \geq 0$, the quantity

$$\hat{\xi}_i^k = 1 + \sum_{j=1}^{\infty} \prod_{1 \leq \ell \leq k \wedge j} \{1 - \alpha(\mathfrak{z}_i, y_j)\} \prod_{k+1 \leq \ell \leq j} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\}$$

is an unbiased estimator of $1/p(\mathfrak{z}_i)$ with an almost sure finite number of terms. Therefore, we have

$$\mathbb{V}\hat{\xi}_i \mathfrak{z}_i \leq \mathbb{V}\hat{\xi}_i^k \mathfrak{z}_i \leq \mathbb{V}\hat{\xi}_i^0 \mathfrak{z}_i = \mathbb{V}\mathfrak{n}_i \mathfrak{z}_i \,.$$

The "Big Data" plague

Simulation from posterior with **large** sample size n

- Computing time at least of order O(n)
- solutions using likelihood decomposition

$$\prod_{i=1}^{n} \ell(\theta|x_i)$$

and handling subsets on different processors (CPU), graphical units (GPU), or computers

[Scott et al., 2013, Korattikara et al., 2013]

- no consensus on method of choice, with instabilities from removing most prior input and uncalibrated approximations

[Neiswanger et al., 2013]

Proposed solution

"There is no problem an absence of decision cannot solve."
Anonymous

Given $\alpha(x, y) := 1 \wedge r(x, y)$, factorise

$$r(x, y) = \prod_{k=1}^{d} \rho_k(x, y)$$

under constraint $\rho_k(x, y) = \rho_k(y, x)^{-1}$
Delayed Acceptance Markov kernel given by

$$\tilde{P}(x, A) := \int_A q(x, y)\tilde{\alpha}(x, y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x, y)\tilde{\alpha}(x, y)\mathrm{d}y\right) \mathbf{1}_A(x)$$

where

$$\tilde{\alpha}(x, y) := \prod_{k=1}^{d} \{1 \wedge \rho_k(x, y)\}.$$

Proposed solution

Algorithm 1 Delayed Acceptance

To sample from $\tilde{P}(x, \cdot)$:

1. Sample $y \sim Q(x, \cdot)$.
2. For $k = 1, \ldots, d$:
   - with probability $1 \wedge \rho_k(x, y)$ continue
   - otherwise stop and output x
3. Output y

Arrange terms in product so that most computationally intensive ones calculated 'at the end' hence least often

Proposed solution

Algorithm 1 Delayed Acceptance

To sample from $\tilde{P}(x, \cdot)$:

1. Sample $y \sim Q(x, \cdot)$.
2. For $k = 1, \ldots, d$:
   - with probability $1 \wedge \rho_k(x, y)$ continue
   - otherwise stop and output x
3. Output y

Generalization of Fox & Nicholls (1997) and Christen & Fox (2005), where testing for acceptance with approximation before computing exact likelihood first suggested

The "Big Data" plague

Delayed Acceptance intended for likelihoods or priors, but not a clear solution for "Big Data" problems

1. all product terms must be computed
2. all terms previously computed either stored for future comparison or recomputed
3. sequential approach limits parallel gains...
4. ...unless prefetching scheme added to delays

[Strid (2010)]

Adaptive MCMC may be hazardous to your ergodicity!

- **Algorithms trained on-line usually invalid:**

Adaptive MCMC may be hazardous to your ergodicity!

↯ **Algorithms trained on-line usually invalid:**
using the whole past of the "chain" implies that this is no longer a Markov chain! !

Illustration

Example (Poly t distribution)

Consider a t-distribution $\mathcal{T}(3, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with a flat prior $\pi(\theta) = 1$
If we try fit a normal proposal from empirical mean and variance of the chain so far,

$$\mu_t = \frac{1}{t} \sum_{i=1}^{t} \theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t} \sum_{i=1}^{t} (\theta^{(i)} - \mu_t)^2 \,,$$

Illustration

Example (Poly t distribution)

Consider a t-distribution $\mathcal{T}(3, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with a flat prior $\pi(\theta) = 1$
If we try fit a normal proposal from empirical mean and variance of the chain so far,

$$\mu_t = \frac{1}{t}\sum_{i=1}^{t}\theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t}\sum_{i=1}^{t}(\theta^{(i)} - \mu_t)^2\,,$$

Metropolis–Hastings algorithm with acceptance probability

$$\prod_{j=2}^{n}\left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2}\right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2/2\sigma_t^2}{\exp -(\mu_t - \xi)^2/2\sigma_t^2}\,,$$

where $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

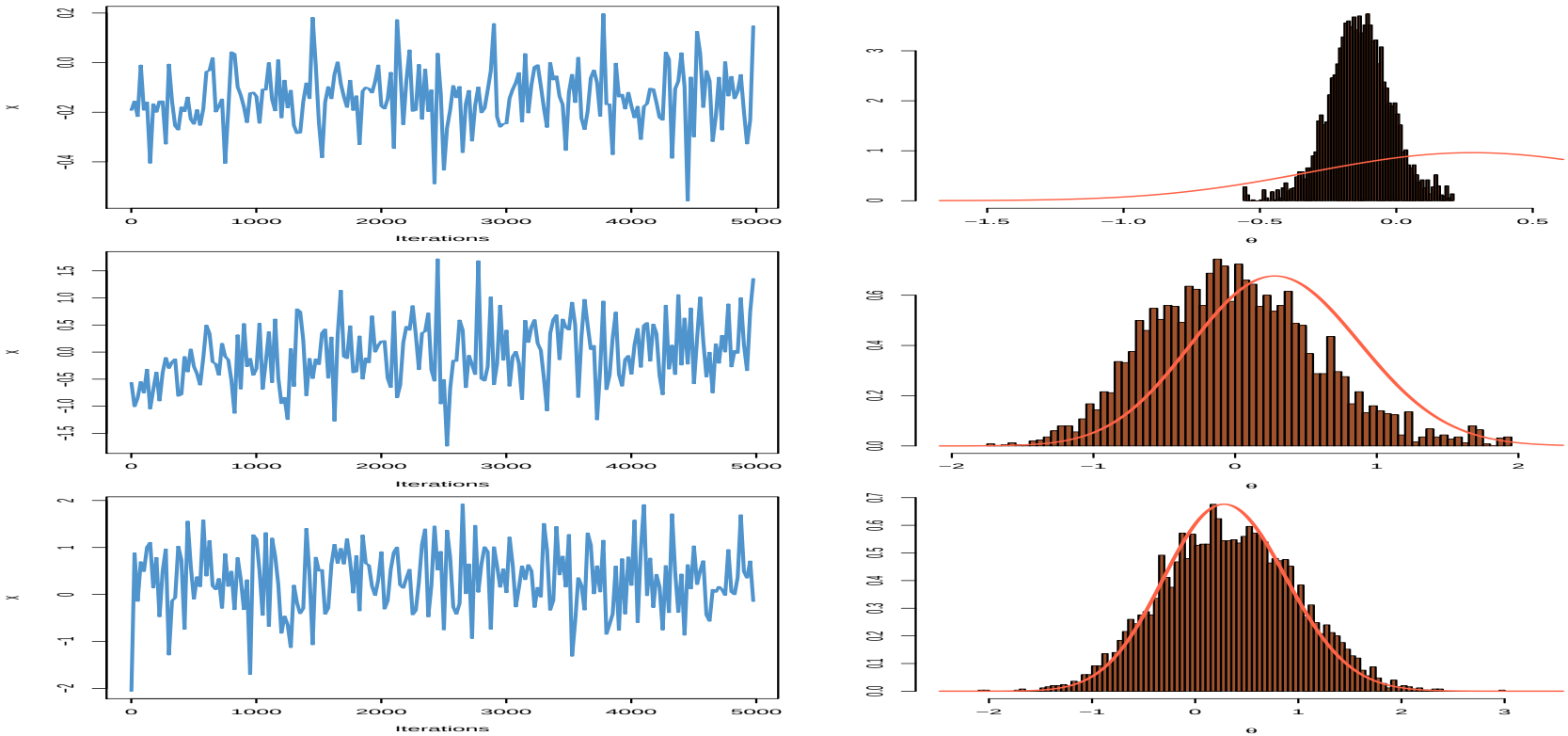
Illustration

Example (Poly t distribution (2))

Invalid scheme:

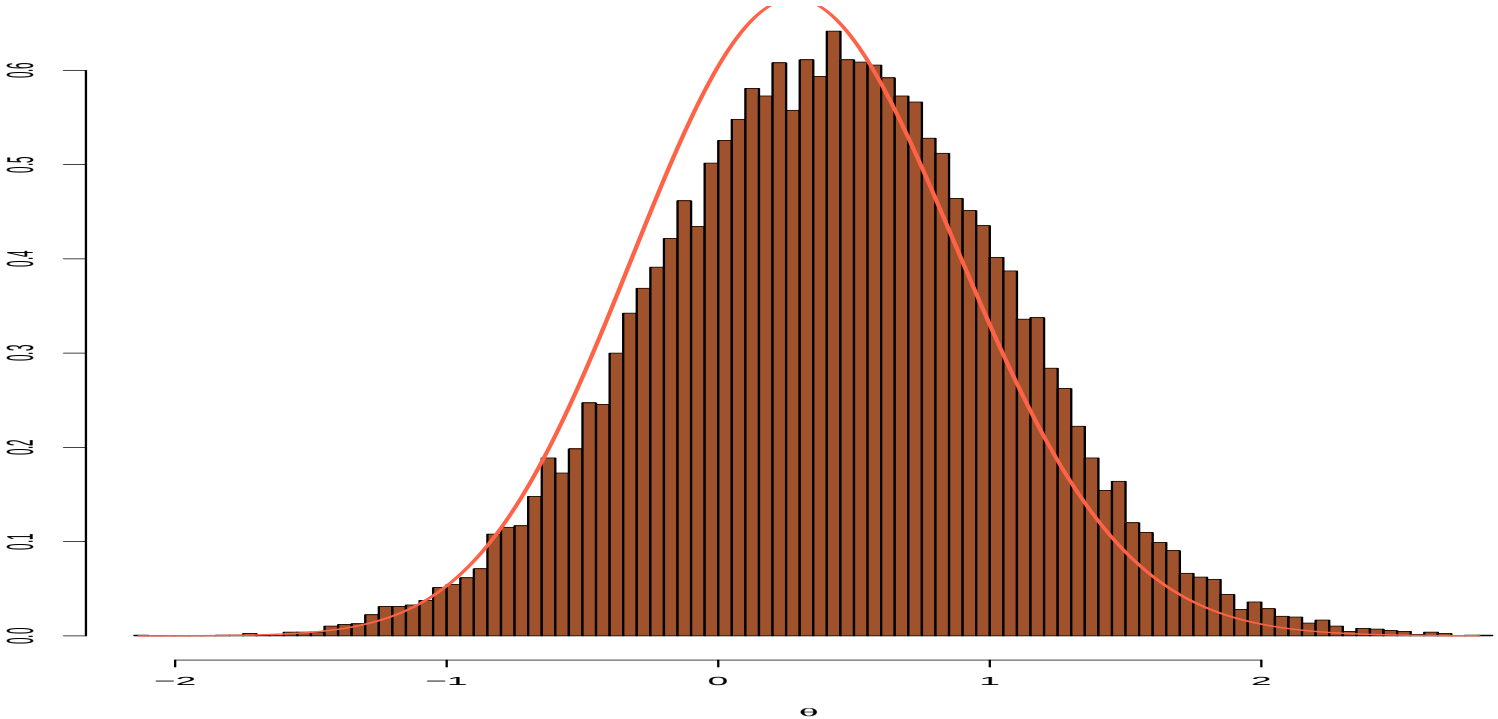
- when range of initial values too small, the $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support.
- long-range dependence on past values modifies the distribution of the sequence.
- using past simulations to create a non-parametric approximation to the target distribution does not work either

Illustration



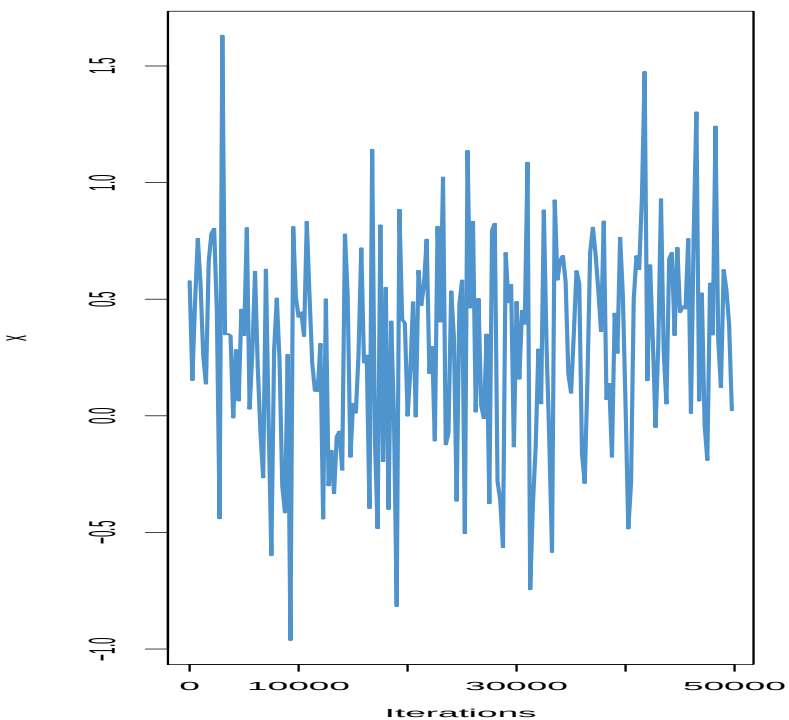
Adaptive scheme for a sample of 10 $x_j \sim \mathcal{T}_\ni$ and initial variances of (top) 0.1, (middle) 0.5, and (bottom) 2.5.

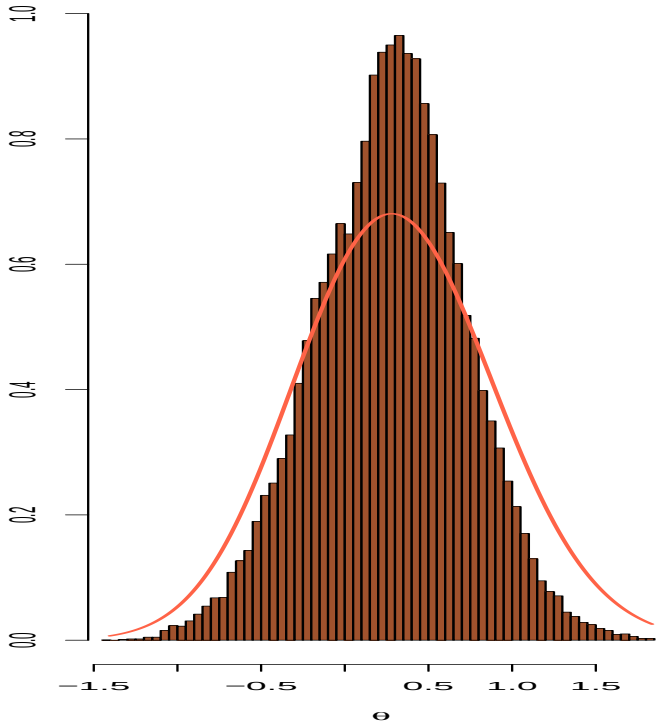
Illustration



Comparison of the distribution of an adaptive scheme sample of $25,000$ points with initial variance of 2.5 and of the target distribution.

Illustration





Sample produced by $50,000$ iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

The Gibbs Sampler

General Principles

A very **specific** simulation algorithm based on the target distribution f:

1. Uses the conditional densities $f_1, \ldots, f_p$ from f

General Principles

A very **specific** simulation algorithm based on the target distribution f:

1. Uses the conditional densities $f_1, \ldots, f_p$ from f
2. Start with the random variable $\mathbf{X} = (X_1, \ldots, X_p)$

General Principles

A very **specific** simulation algorithm based on the target distribution f:

1. Uses the conditional densities $f_1, \ldots, f_p$ from f
2. Start with the random variable $\mathbf{X} = (X_1, \ldots, X_p)$
3. Simulate from the conditional densities,

$$\begin{aligned} X_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \\ \sim f_i(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p) \end{aligned}$$

for $i = 1, 2, \ldots, p$.

General Principles

Algorithm (Gibbs sampler)

Given $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, generate

1. $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)})$;
2. $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$,

$\ldots$

p. $X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$

$$\mathbf{X}^{(t+1)} \to \mathbf{X} \sim f$$

Properties

The full conditionals densities $f_1, \ldots, f_p$ are the only densities used for simulation. Thus, even in a high dimensional problem, all of the simulations may be univariate

Properties

The full conditionals densities $f_1, \ldots, f_p$ are the only densities used for simulation. Thus, even in a high dimensional problem, all of the simulations may be univariate

The Gibbs sampler **is not reversible** with respect to f. However, each of its p components is. Besides, it can be turned into a reversible sampler, either using the *Random Scan Gibbs sampler* or running instead the (double) sequence

$$f_1 \cdots f_{p-1} f_p f_{p-1} \cdots f_1$$

A Very Simple Example: Independent $\mathcal{N}(\mu, \sigma^2)$ Observations

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard familly

A Very Simple Example: Independent $\mathcal{N}(\mu, \sigma^2)$ Observations

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard familly

But...

$$\mu|\boldsymbol{Y}_{0:n}, \sigma^2 \sim \mathcal{N}\left(\mu \left| \frac{1}{n}\sum_{i=1}^n Y_i, \frac{\sigma^2}{n}\right.\right)$$

$$\sigma^2|\boldsymbol{Y}_{1:n}, \mu \sim \mathcal{IG}\left(\sigma^2 \left| \frac{n}{2} - 1, \frac{1}{2}\sum_{i=1}^n (Y_i - \mu)^2\right.\right)$$

assuming constant (improper) priors on both μ and σ^2

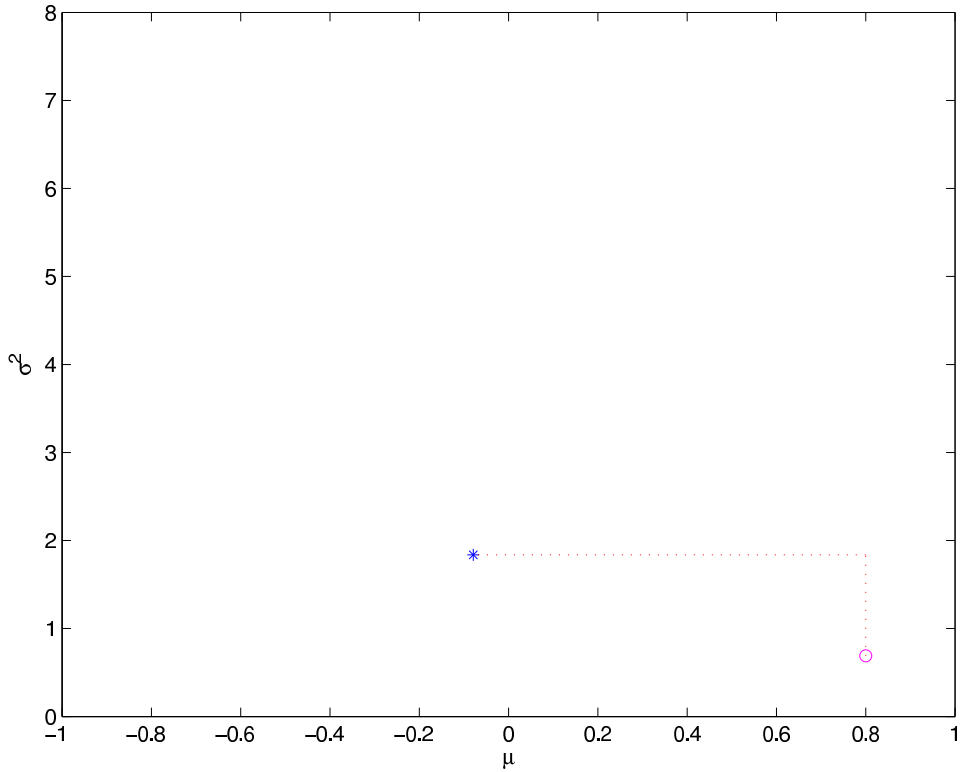
- Hence we may use the Gibbs sampler for simulating from the posterior of (μ, σ^2)

A Very Simple Example: Independent $\mathcal{N}(\mu, \sigma^2)$ Observations

R Gibbs Sampler for Gaussian posterior

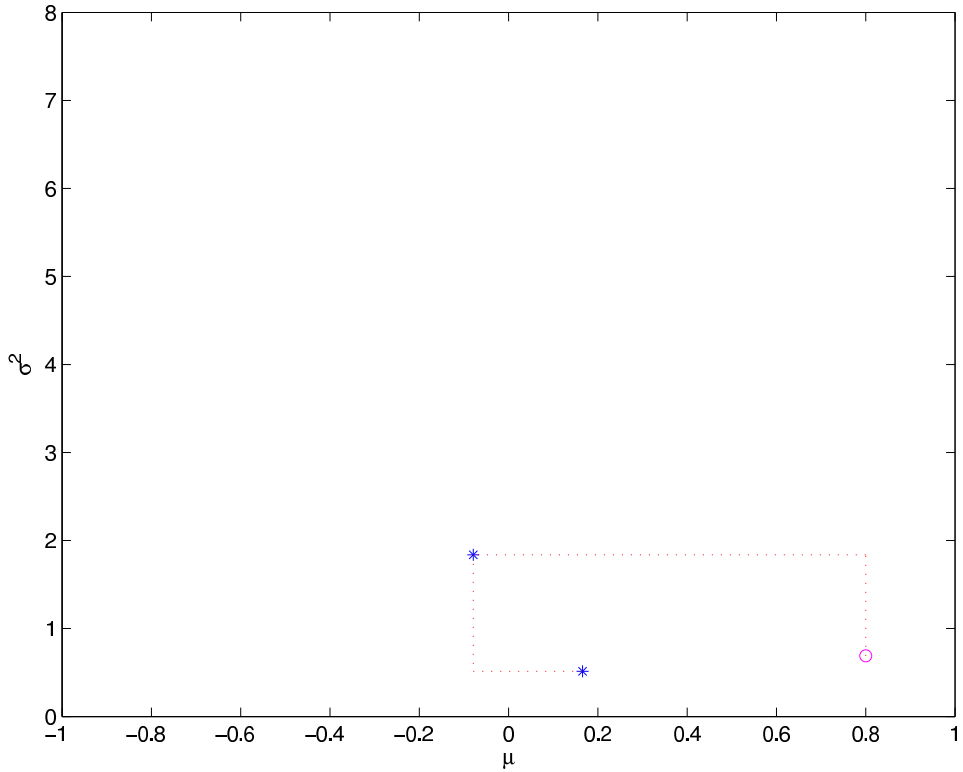
```
n = length(Y);
S = sum(Y);
mu = S/n;
for (i in 1:500)
    S2 = sum((Y-mu)^2);
    sigma2 = 1/rgamma(1,n/2-1,S2/2);
    mu = S/n + sqrt(sigma2/n)*rnorm(1);
```

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



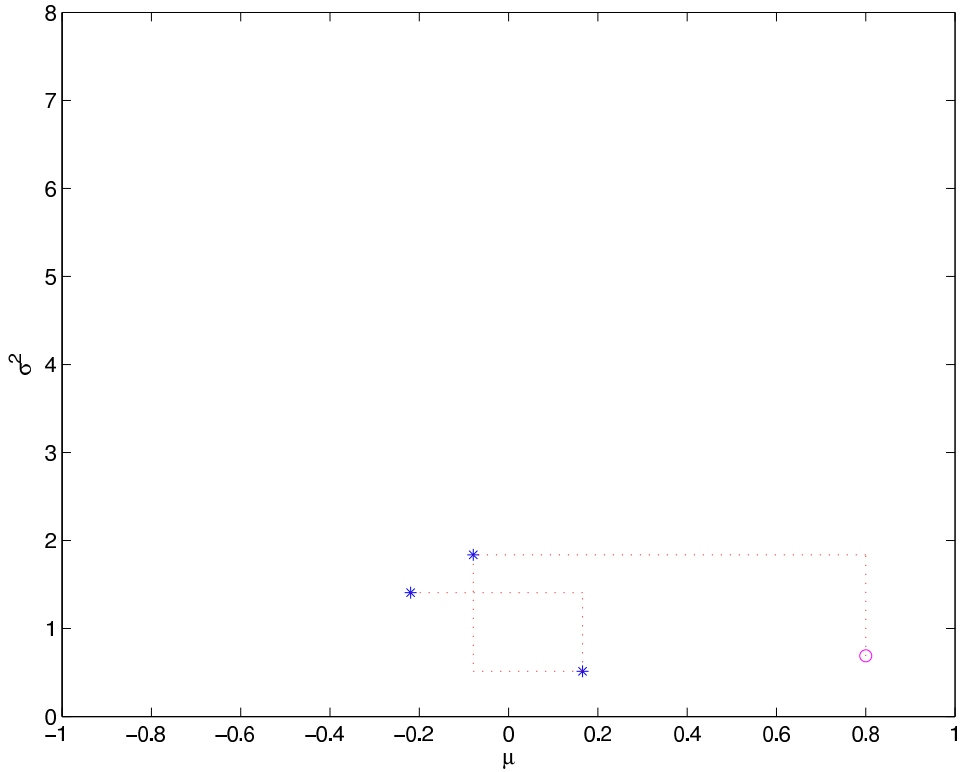
Number of Iterations 1

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



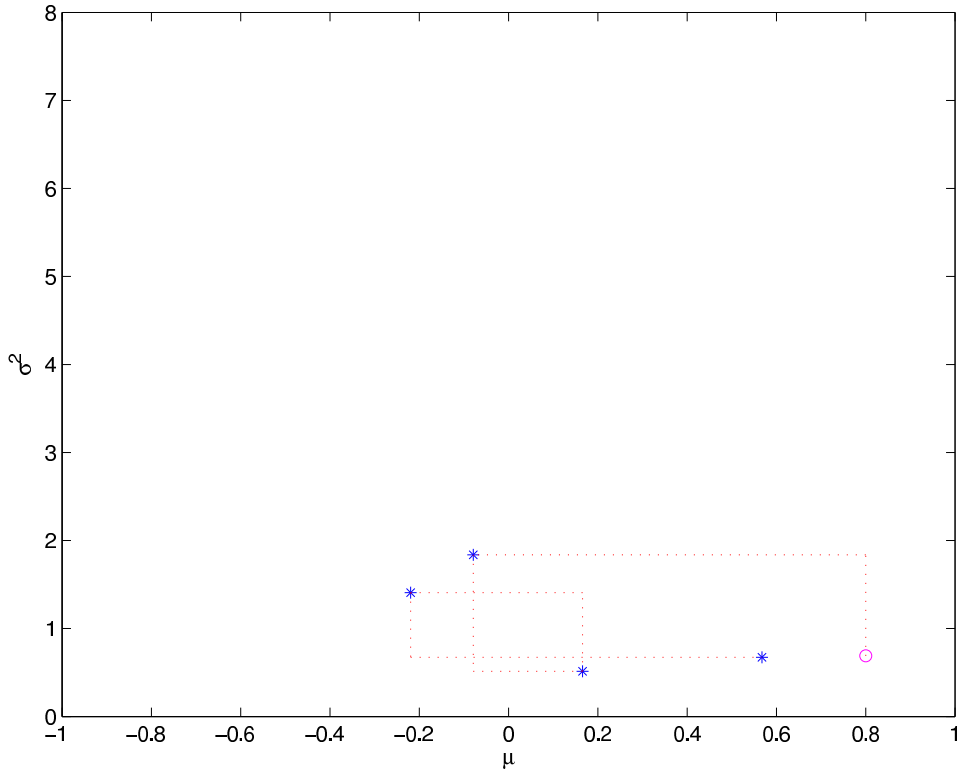
Number of Iterations 1, 2

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



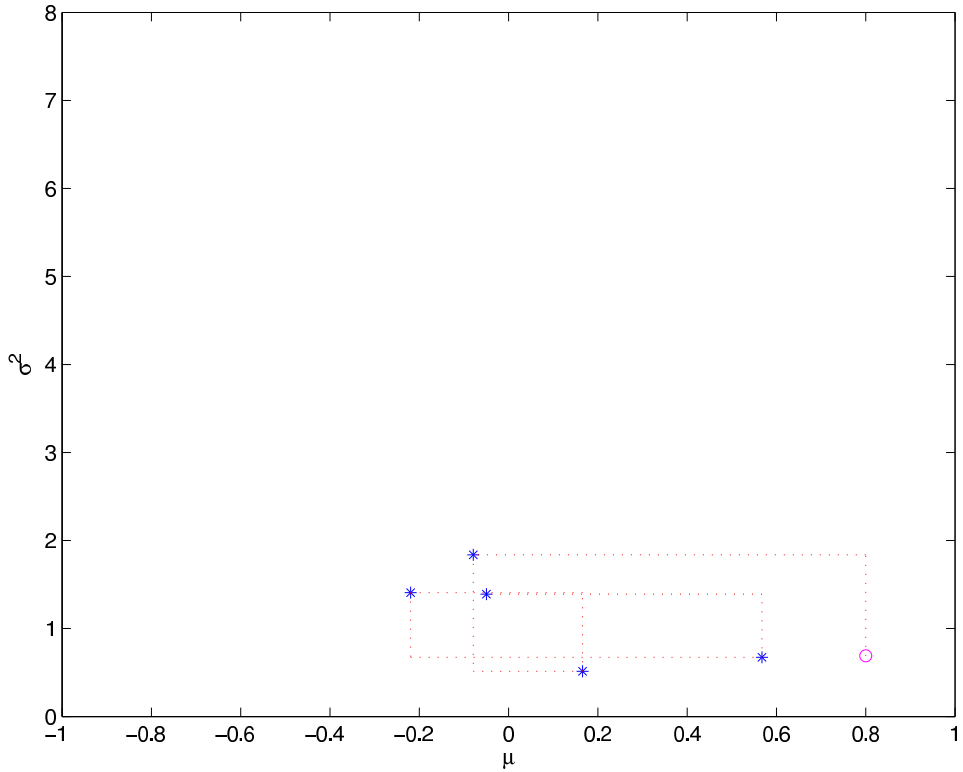
Number of Iterations 1, 2, 3

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



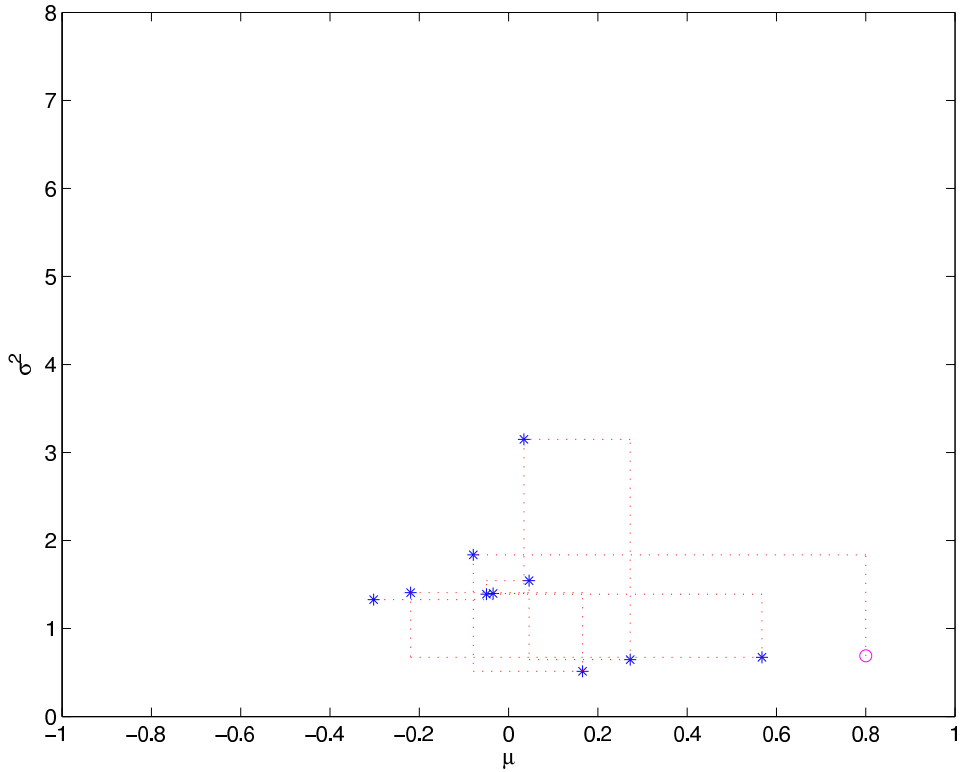
Number of Iterations 1, 2, 3, 4

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



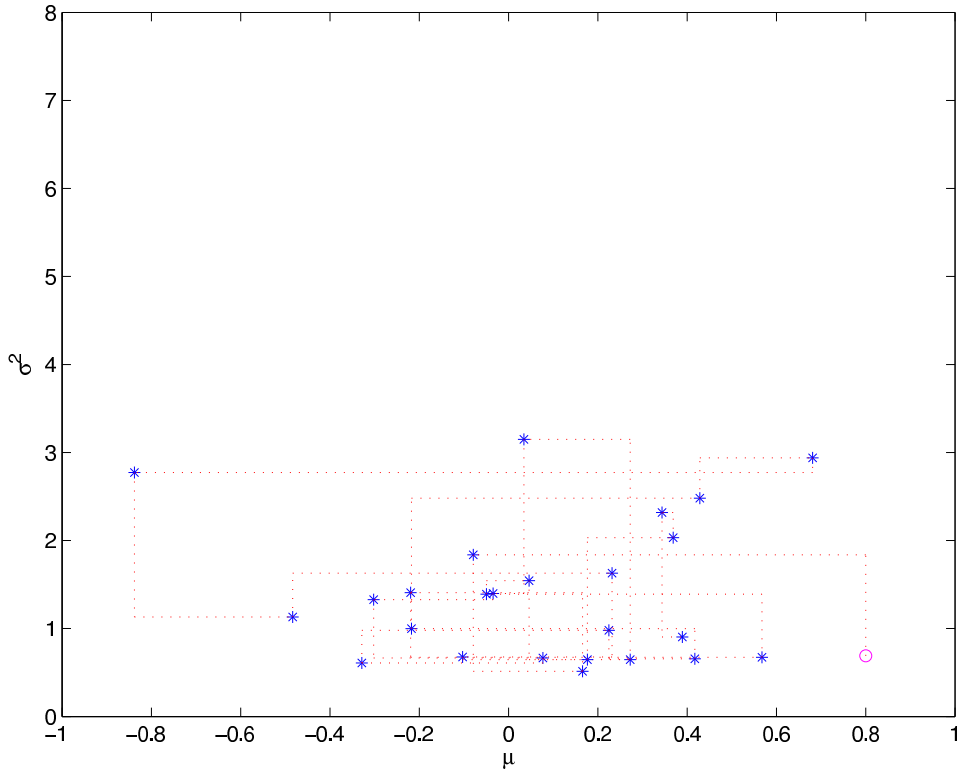
Number of Iterations 1, 2, 3, 4, 5

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



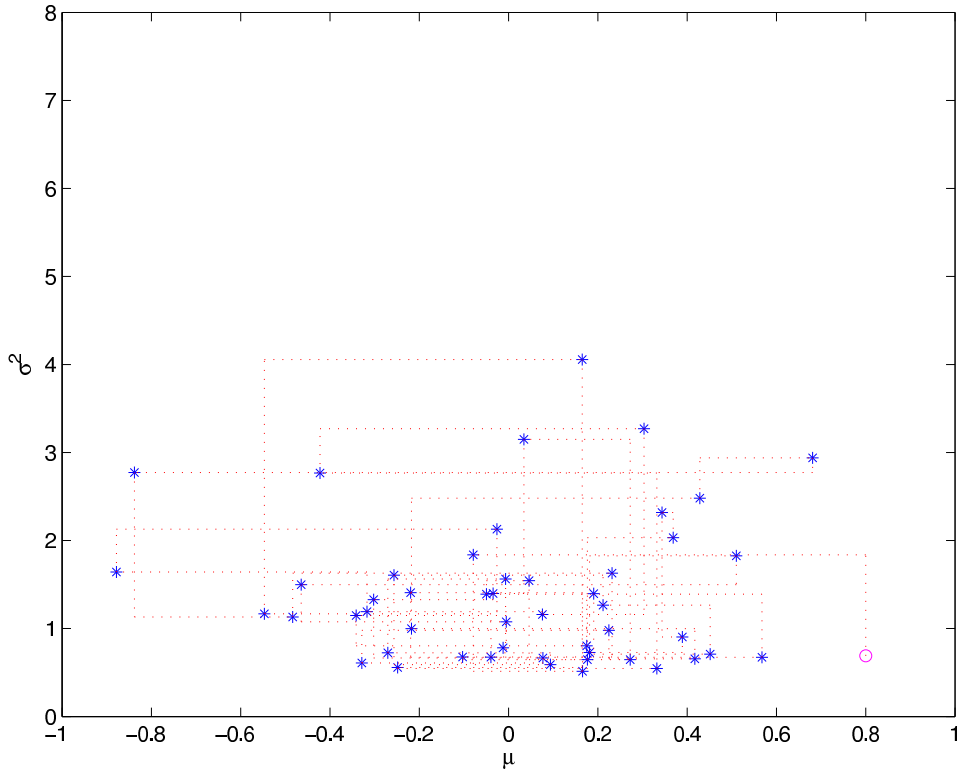
Number of Iterations 1, 2, 3, 4, 5, 10

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



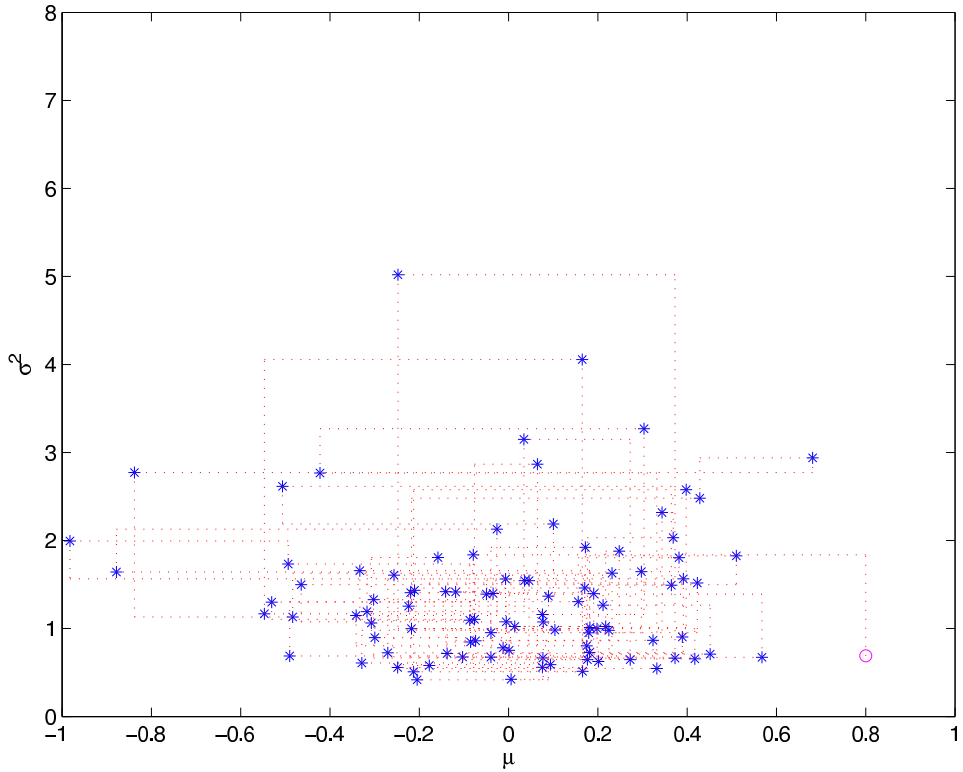
Number of Iterations 1, 2, 3, 4, 5, 10, 25

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



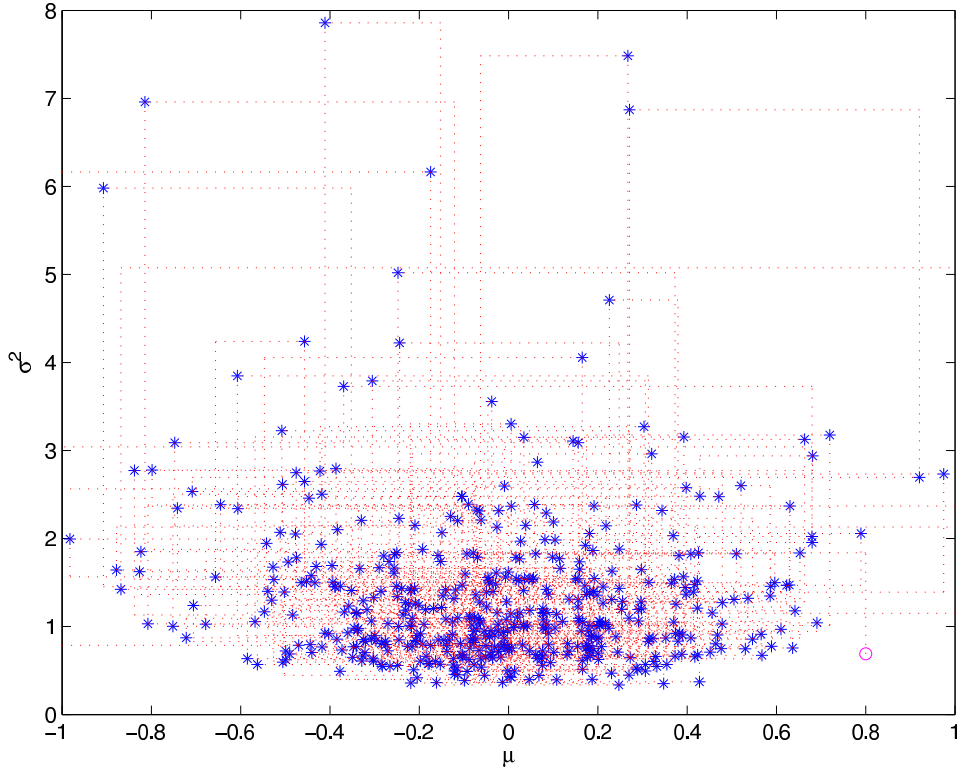
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100

Example of results with $n = 10$ observations from the $\mathcal{N}(0, 1)$ distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

Latent variables

The Gibbs sampler can be generalized in much wider generality
A density g is a completion of f if

$$\int_{\mathscr{Z}} g(x, z)\, dz = f(x)$$

Latent variables

The Gibbs sampler can be generalized in much wider generality
A density g is a completion of f if

$$\int_{\mathcal{Z}} g(x, z)\, dz = f(x)$$

Note

The variable z may be meaningless for the problem

Purpose

g should have full conditionals that are easy to simulate for a Gibbs sampler to be implemented with g rather than f

For $p > 1$, write $y = (x, z)$ and denote the conditional densities of $g(y) = g(y_1, \ldots, y_p)$ by

$$\begin{aligned} Y_1|y_2, \ldots, y_p &\sim g_1(y_1|y_2, \ldots, y_p), \\ Y_2|y_1, y_3, \ldots, y_p &\sim g_2(y_2|y_1, y_3, \ldots, y_p), \\ &\ldots, \\ Y_p|y_1, \ldots, y_{p-1} &\sim g_p(y_p|y_1, \ldots, y_{p-1}). \end{aligned}$$

Purpose

The move from $Y^{(t)}$ to $Y^{(t+1)}$ is defined as follows:

Algorithm (Completion Gibbs sampler)

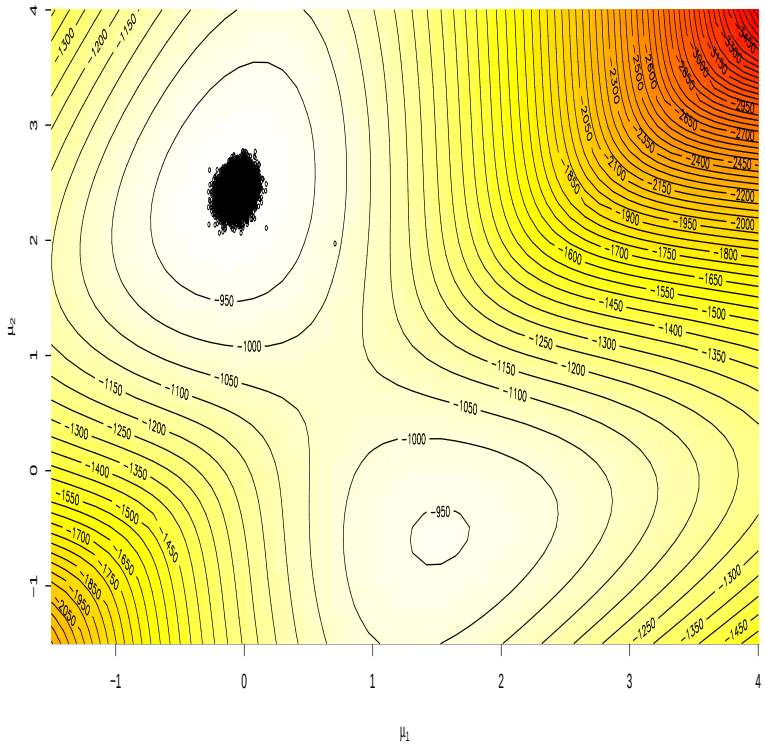
Given $(y_1^{(t)}, \ldots, y_p^{(t)})$, simulate

1. $Y_1^{(t+1)} \sim g_1(y_1 | y_2^{(t)}, \ldots, y_p^{(t)})$,
2. $Y_2^{(t+1)} \sim g_2(y_2 | y_1^{(t+1)}, y_3^{(t)}, \ldots, y_p^{(t)})$,

$\ldots$

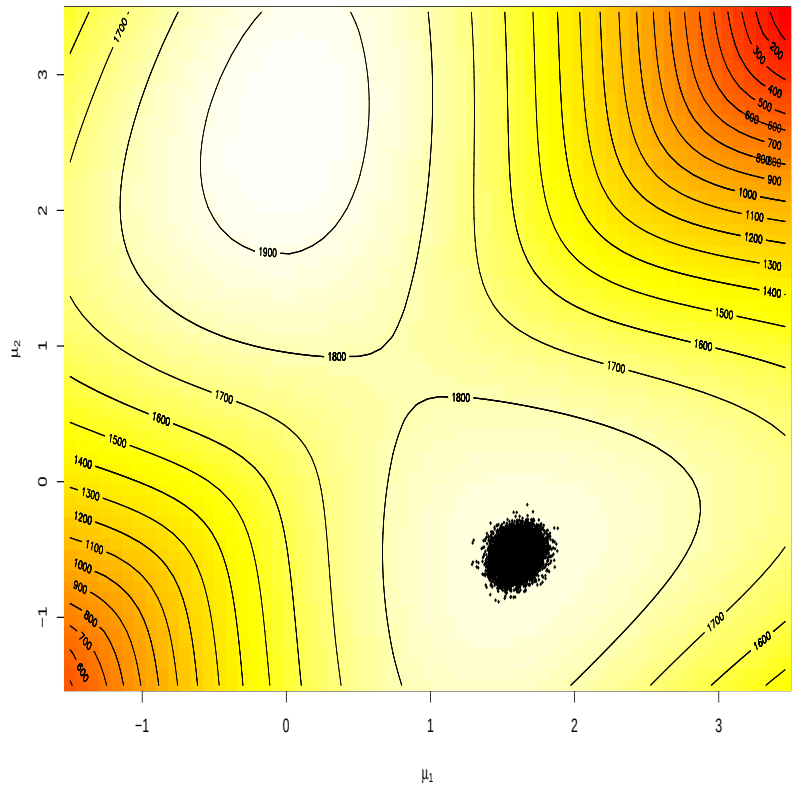
p. $Y_p^{(t+1)} \sim g_p(y_p | y_1^{(t+1)}, \ldots, y_{p-1}^{(t+1)})$.

A wee problem



Gibbs started at random

Gibbs stuck at the wrong model



Slice sampler as generic Gibbs

If $f(\theta)$ can be written as a product

$$\prod_{i=1}^{k} f_i(\theta),$$

Slice sampler as generic Gibbs

If $f(\theta)$ can be written as a product

$$\prod_{i=1}^{k} f_i(\theta),$$

it can be completed as

$$\prod_{i=1}^{k} \mathbb{I}_{0 \le \omega_i \le f_i(\theta)},$$

leading to the following Gibbs algorithm:

Slice sampler

Algorithm (Slice sampler)

Simulate

1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,f_1(\theta^{(t)})]}$;

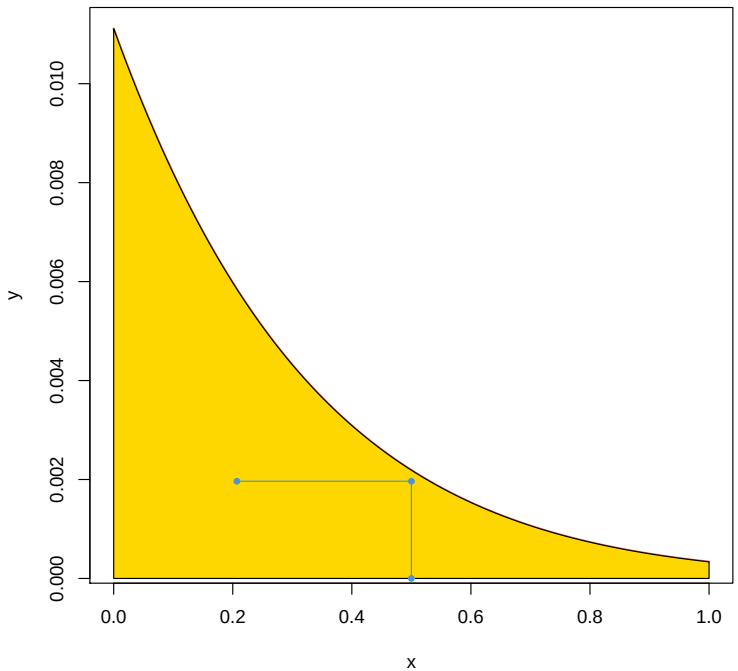
$\ldots$

k. $\omega_k^{(t+1)} \sim \mathcal{U}_{[0,f_k(\theta^{(t)})]}$;

k+1. $\theta^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$, with

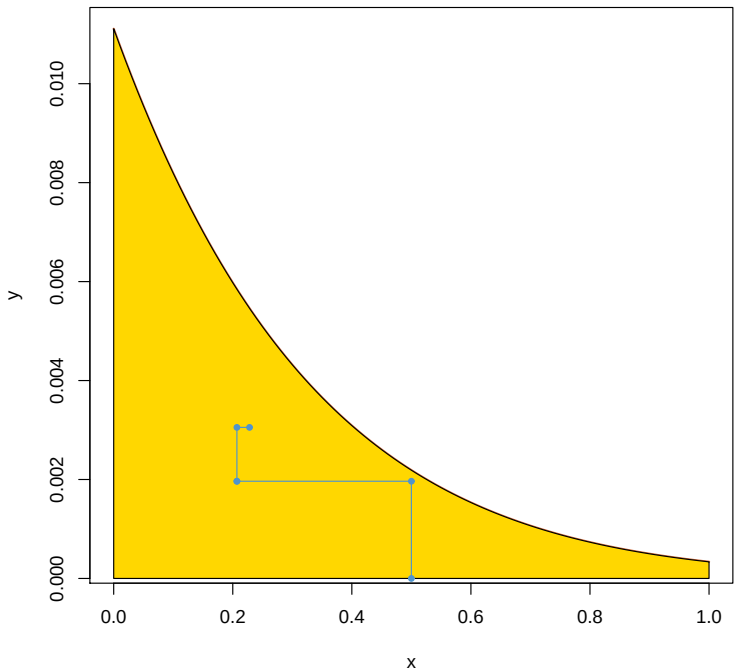
$$A^{(t+1)} = \{y;\ f_i(y) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\}.$$

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



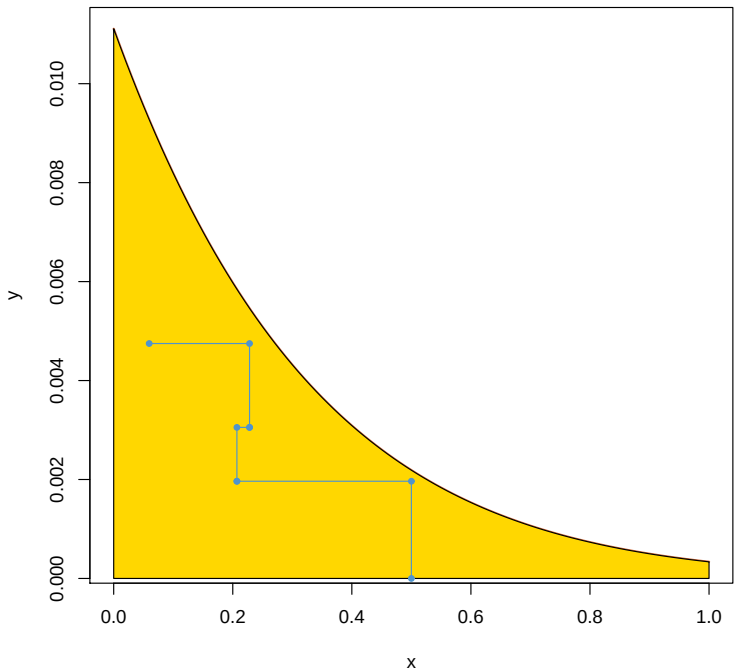
Number of Iterations 2

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



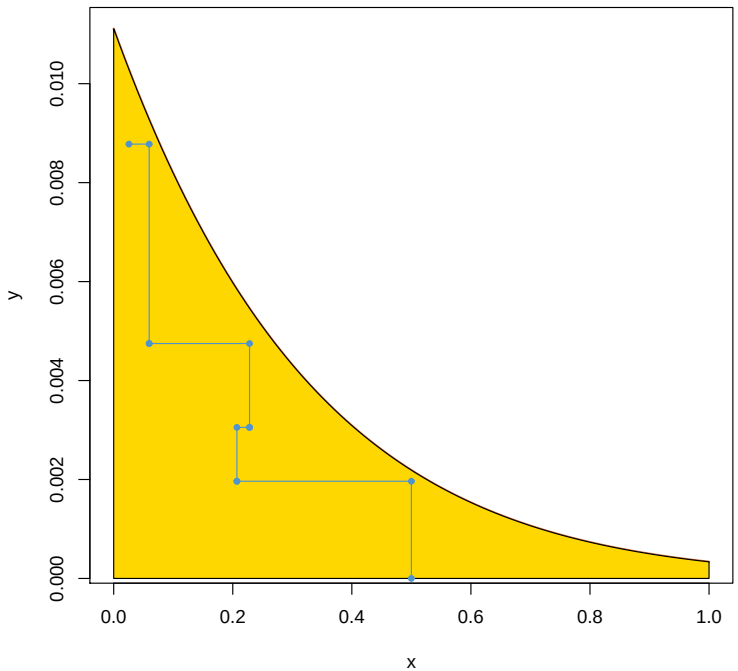
Number of Iterations 2, 3

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



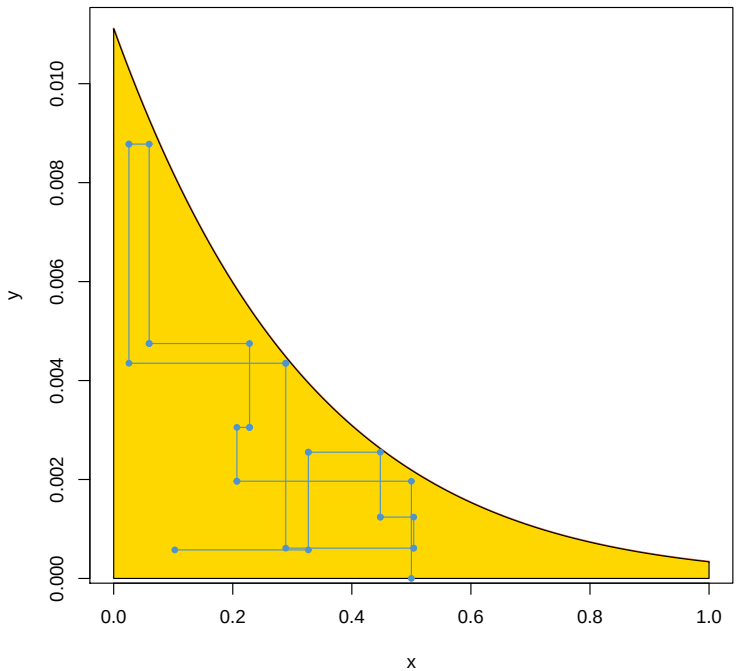
Number of Iterations 2, 3, 4

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



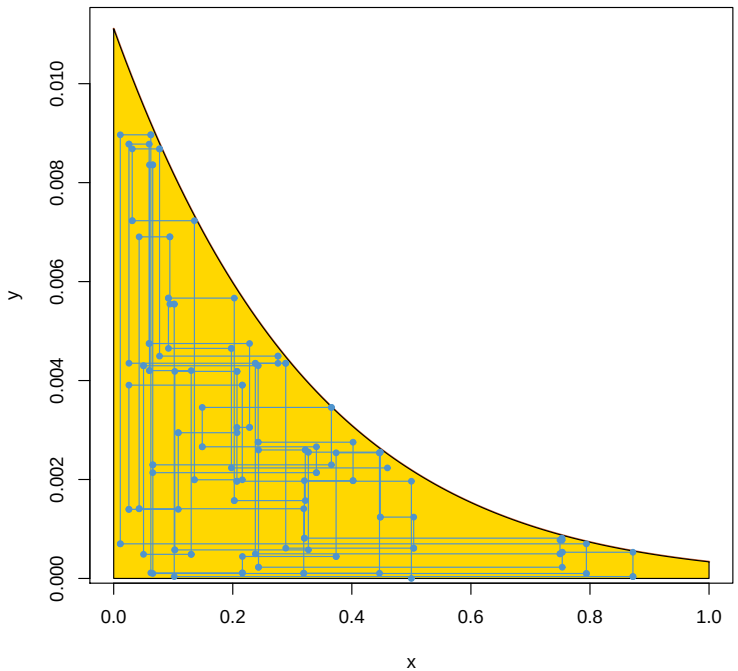
Number of Iterations 2, 3, 4, 5

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



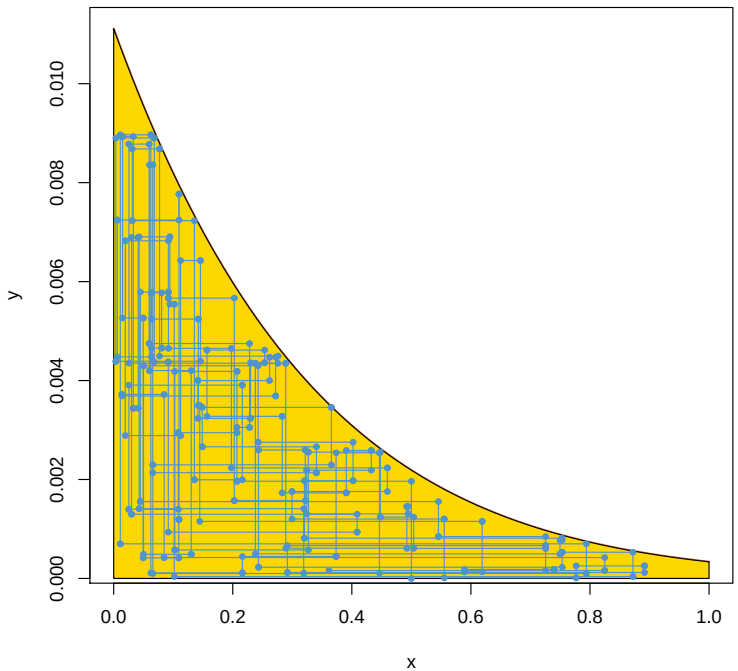
Number of Iterations 2, 3, 4, 5, 10

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



Number of Iterations 2, 3, 4, 5, 10, 50

Example of results with a truncated $\mathcal{N}(-3, 1)$ distribution



Number of Iterations 2, 3, 4, 5, 10, 50, 100

Good slices

The slice sampler usually enjoys good theoretical properties (like geometric ergodicity and even uniform ergodicity under bounded f and bounded $\mathscr{X}$).
As k increases, the determination of the set $A^{(t+1)}$ may get increasingly complex.

Slice sampler: illustration

Example (Stochastic volatility core distribution)

Difficult part of the stochastic volatility model

$$\pi(x) \propto \exp - \left\{ \sigma^2 (x-\mu)^2 + \beta^2 \exp(-x) y^2 + x \right\} / 2 \,,$$

simplified in $\exp - \left\{ x^2 + \alpha \exp(-x) \right\}$

Slice sampler: illustration

Example (Stochastic volatility core distribution)

Difficult part of the stochastic volatility model

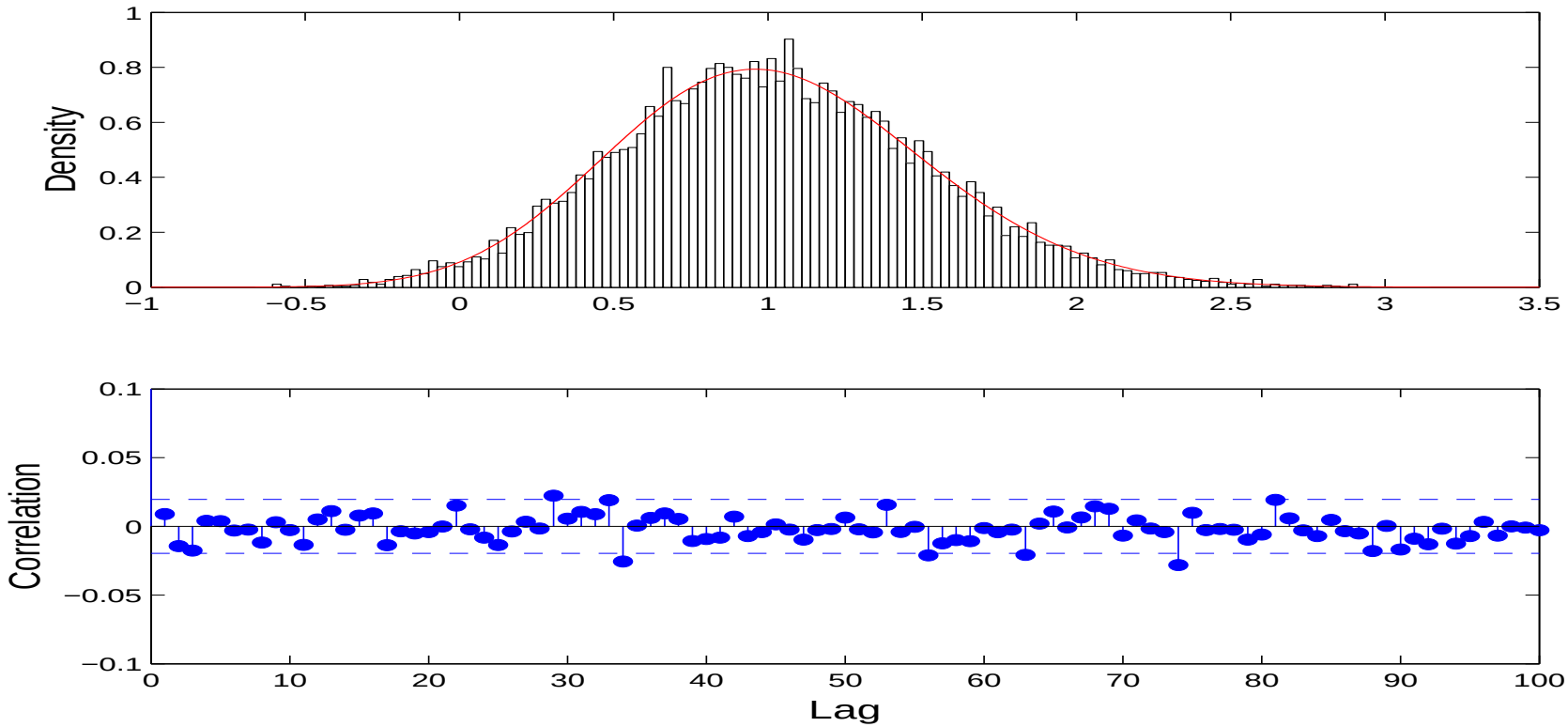
$$\pi(x) \propto \exp - \left\{\sigma^2(x-\mu)^2 + \beta^2 \exp(-x) y^2 + x\right\}/2\,,$$

simplified in $\exp - \left\{x^2 + \alpha \exp(-x)\right\}$
Slice sampling means simulation from a uniform distribution on

$$\begin{aligned} \mathfrak{A} &= \left\{x; \exp - \left\{x^2 + \alpha \exp(-x)\right\}/2 \geq u\right\} \\ &= \left\{x; x^2 + \alpha \exp(-x) \leq \omega\right\} \end{aligned}$$

if we set $\omega = -2 \log u$.
Note Inversion of $x^2 + \alpha \exp(-x) = \omega$ needs to be done by trial-and-error.

Slice sampler: illustration



Histogram of a Markov chain produced by a slice sampler and target distribution in overlay.

Properties of the Gibbs sampler

Theorem (Convergence)

For

$$(Y_1, Y_2, \cdots, Y_p) \sim g(y_1, \ldots, y_p),$$

if either

[Positivity condition]

(*i*) $g^{(i)}(y_i) > 0$ *for every* $i = 1, \cdots, p$, *implies that* $g(y_1, \ldots, y_p) > 0$, *where* $g^{(i)}$ *denotes the marginal distribution of* Y_i, *or*

(*ii*) *the transition kernel is absolutely continuous with respect to* g,

then the chain is irreducible and positive Harris recurrent.

Properties of the Gibbs sampler (2)

Consequences

(i) If $\int h(y)g(y)dy < \infty$, then

$$\lim_{nT\to\infty} \frac{1}{T}\sum_{t=1}^{T} h_1(Y^{(t)}) = \int h(y)g(y)dy \text{ a.e. } g.$$

(ii) If, in addition, $(Y^{(t)})$ is aperiodic, then

$$\lim_{n\to\infty} \left\| \int K^n(y,\cdot)\mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution μ.

Slice sampler

fast on that slice

For convergence, the properties of X_t and of $f(X_t)$ are identical

Theorem (Uniform ergodicity)

If f is bounded and $\mathrm{supp} f$ is bounded, the simple slice sampler is uniformly ergodic.

[Mira & Tierney, 1997]

Hammersley-Clifford theorem

An illustration that conditionals determine the joint distribution

Theorem

If the joint density $g(y_1, y_2)$ *have conditional distributions* $g_1(y_1|y_2)$ *and* $g_2(y_2|y_1)$*, then*

$$g(y_1, y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)\, dv}.$$

[Hammersley & Clifford, circa 1970]

General HC decomposition

Under the positivity condition, the joint distribution g satisfies

$$g(y_1, \ldots, y_p) \propto \prod_{j=1}^{p} \frac{g_{\ell_j}(y_{\ell_j} | y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}{g_{\ell_j}(y'_{\ell_j} | y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}$$

for every permutation ℓ on $\{1, 2, \ldots, p\}$ and every $y' \in \mathscr{Y}$.

Rao-Blackwellization

If $(y_1, y_2, \ldots, y_p)^{(t)}, t = 1, 2, \ldots T$ is the output from a Gibbs sampler

$$\delta_0 = \frac{1}{T} \sum_{t=1}^{T} h\left(y_1^{(t)}\right) \rightarrow \int h(y_1) g(y_1) dy_1$$

and is unbiased.

Rao-Blackwellization

If $(y_1, y_2, \ldots, y_p)^{(t)}, t = 1, 2, \ldots T$ is the output from a Gibbs sampler

$$\delta_0 = \frac{1}{T} \sum_{t=1}^{T} h\left(y_1^{(t)}\right) \rightarrow \int h(y_1) g(y_1) dy_1$$

and is unbiased.
The Rao-Blackwellization replaces δ_0 with its conditional expectation

$$\delta_{rb} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\left[h(Y_1)|y_2^{(t)}, \ldots, y_p^{(t)}\right].$$

Rao-Blackwellization (2)

Then

- Both estimators converge to $\mathbb{I}[h(Y_1)]$
- Both are unbiased,

Rao-Blackwellization (2)

Then

- Both estimators converge to $\mathbb{I}[h(Y_1)]$
- Both are unbiased,
- and

$$\operatorname{var}\left(\mathbb{I}\left[h(Y_1)|Y_2^{(t)},\ldots,Y_p^{(t)}\right]\right) \leq \operatorname{var}(h(Y_1)),$$

so δ_{rb} is uniformly better (for Data Augmentation)

Examples of Rao-Blackwellization

Example

Bivariate normal Gibbs sampler

$$\begin{aligned} X \mid y &\sim \mathcal{N}(\rho y,\ 1-\rho^2) \\ Y \mid x &\sim \mathcal{N}(\rho x,\ 1-\rho^2). \end{aligned}$$

Then

$$\delta_0 = \frac{1}{T}\sum_{i=1}^{T} X^{(i)} \quad \text{and} \quad \delta_1 = \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}[X^{(i)}|Y^{(i)}] = \frac{1}{T}\sum_{i=1}^{T} \varrho Y^{(i)},$$

estimate $\mathbb{I}[X]$ and $\sigma^2_{\delta_0}/\sigma^2_{\delta_1} = \frac{1}{\rho^2} > 1$.

Examples of Rao-Blackwellization (2)

Example (Poisson-Gamma Gibbs cont'd)

Naïve estimate

$$\delta_0 = \frac{1}{T} \sum_{t=1}^{T} \lambda^{(t)}$$

and Rao-Blackwellized version

$$\begin{aligned} \delta^{\pi} &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[\lambda^{(t)} | x_1, x_2, \ldots, x_5, y_1^{(i)}, y_2^{(i)}, \ldots, y_{13}^{(i)}] \\ &= \frac{1}{360T} \sum_{t=1}^{T} \left(313 + \sum_{i=1}^{13} y_i^{(t)} \right), \end{aligned}$$

back to graph

NP Rao-Blackwellization & Rao-Blackwellized NP

Another substantial benefit of Rao-Blackwellization is in the approximation of densities of different components of y without nonparametric density estimation methods.

NP Rao-Blackwellization & Rao-Blackwellized NP

Another substantial benefit of Rao-Blackwellization is in the approximation of densities of different components of y without nonparametric density estimation methods.

Lemma

The estimator

$$\frac{1}{T}\sum_{t=1}^{T} g_i(y_i|y_j^{(t)}, j \neq i) \longrightarrow g_i(y_i),$$

is unbiased.

Improper priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:

Improper priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:

It may happen that

- all conditional distributions are well defined,
- all conditional distributions may be simulated from, but...

Improper priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:
It may happen that

- all conditional distributions are well defined,
- all conditional distributions may be simulated from, but...
- the system of conditional distributions may not correspond to any joint distribution

Improper priors

↯ Unsuspected danger resulting from careless use of MCMC algorithms:
It may happen that

- all conditional distributions are well defined,
- all conditional distributions may be simulated from, but...
- the system of conditional distributions may not correspond to any joint distribution

Warning The problem is due to careless use of the Gibbs sampler in a situation for which the underlying assumptions are violated

Improper posteriors

Example (Conditional exponential distributions)

For the model

$$X_1|x_2 \sim \mathcal{E}xp(x_2)\ , \quad X_2|x_1 \sim \mathcal{E}xp(x_1)$$

the only candidate $f(x_1, x_2)$ for the joint density is

$$f(x_1, x_2) \propto \exp(-x_1 x_2),$$

but

$$\int f(x_1, x_2) dx_1 dx_2 = \infty$$

© **These conditionals do not correspond to a joint probability distribution**

Improper posteriors

Example (Improper random effects)

Consider

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J,$$

where

$$\alpha_i \sim \mathcal{N}(0, \sigma^2) \text{ and } \varepsilon_{ij} \sim \mathcal{N}(0, \tau^2),$$

the Jeffreys (improper) prior for the parameters μ, σ and τ is

$$\pi(\mu, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2} \; .$$

Improper posteriors

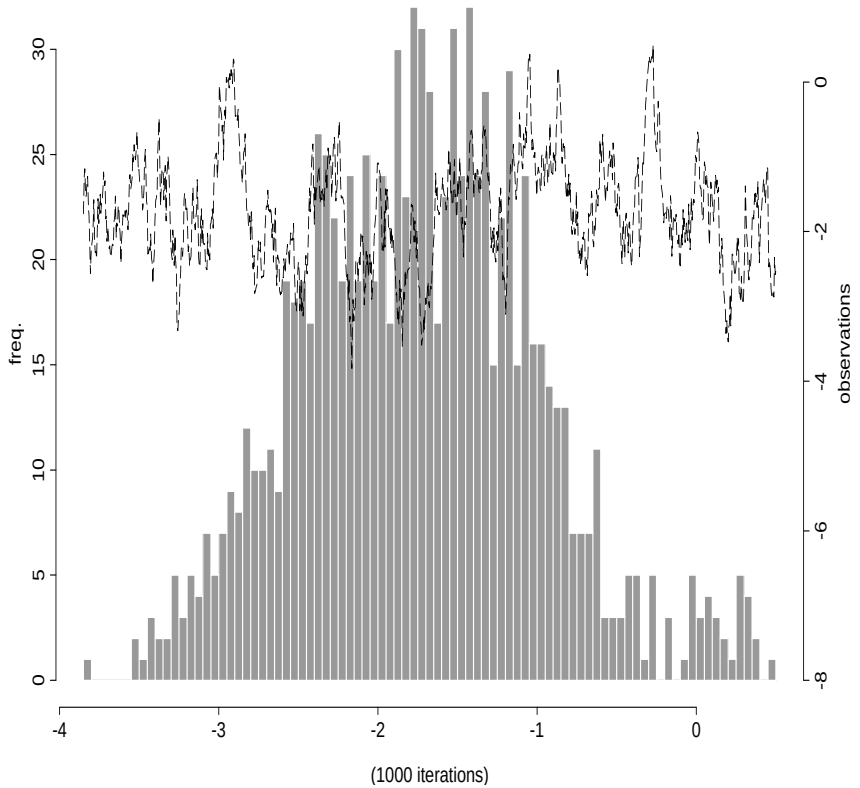
Example (Improper random effects 2)

The conditional distributions

$$\begin{aligned}
\alpha_i|y,\mu,\sigma^2,\tau^2 &\sim \mathcal{N}\left(\frac{J(\bar{y}_i-\mu)}{J+\tau^2\sigma^{-2}}, (J\tau^{-2}+\sigma^{-2})^{-1}\right) , \\
\mu|\alpha,y,\sigma^2,\tau^2 &\sim \mathcal{N}(\bar{y}-\bar{\alpha}, \tau^2/JI) , \\
\sigma^2|\alpha,\mu,y,\tau^2 &\sim \mathcal{IG}\left(I/2, (1/2)\sum_i \alpha_i^2\right) , \\
\tau^2|\alpha,\mu,y,\sigma^2 &\sim \mathcal{IG}\left(IJ/2, (1/2)\sum_{i,j}(y_{ij}-\alpha_i-\mu)^2\right) ,
\end{aligned}$$

are well-defined and a Gibbs sampler can be easily implemented in this setting.

Improper posteriors



Example (Improper random effects 2)

The figure shows the sequence of $\mu^{(t)}$'s and its histogram over $1,000$ iterations. They both **fail to** indicate that the corresponding "joint distribution" **does not exist**

Final notes on impropriety

The improper posterior Markov chain cannot be positive recurrent

Final notes on impropriety

The improper posterior Markov chain cannot be positive recurrent

The major task in such settings is to find indicators that flag that something is wrong. However, the output of an "improper" Gibbs sampler may not differ from a positive recurrent Markov chain.

Final notes on impropriety

The improper posterior Markov chain cannot be positive recurrent

The major task in such settings is to find indicators that flag that something is wrong. However, the output of an "improper" Gibbs sampler may not differ from a positive recurrent Markov chain.

Example

The random effects model was initially treated in Gelfand & al (1990) as a legitimate model

Hamiltonian Monte Carlo

The Metropolis-Hastings Algorithm

Gibbs Sampling

Hamiltonian Monte Carlo

Piecewise Deterministic Versions

Continuous time Markov process

Hamiltonian (or hybrid) Monte Carlo (HMC) auxiliary variable technique that takes advantage of a continuous time Markov process to sample from target $\pi(\theta)$

Auxiliary variable $\vartheta \in \mathbb{R}^d$ introduced along with a density $\varpi(\vartheta|\theta)$ so that the joint distribution of (θ, ϑ) enjoys $\pi(\theta)$ as its marginal

$$\pi(\theta) = \int \pi(\theta)\varpi(\vartheta|\theta)\mathrm{d}\vartheta$$

Continuous time Markov process

Based on representation of joint distribution

$$\omega(\theta, \vartheta) = \pi(\theta)\varpi(\vartheta|\theta) \propto \exp\{-H(\theta, \vartheta)\},$$

where $H(\cdot)$ called *Hamiltonian*

Continuous time Markov process

Based on representation of joint distribution

$$\omega(\theta, \vartheta) = \pi(\theta)\varpi(\vartheta|\theta) \propto \exp\{-H(\theta, \vartheta)\}\,,$$

where $H(\cdot)$ called *Hamiltonian*
Hamiltonian Monte Carlo (HMC) associated with the continuous time process (θ_t, ϑ_t) generated by the so-called *Hamiltonian equations*

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = \frac{\partial H}{\partial \vartheta}(\theta_t, \vartheta_t) \qquad \frac{\mathrm{d}\vartheta_t}{\mathrm{d}t} = -\frac{\partial H}{\partial \theta}(\theta_t, \vartheta_t)\,,$$

Continuous time Markov process

Based on representation of joint distribution

$$\omega(\theta, \vartheta) = \pi(\theta)\varpi(\vartheta|\theta) \propto \exp\{-H(\theta, \vartheta)\}\,,$$

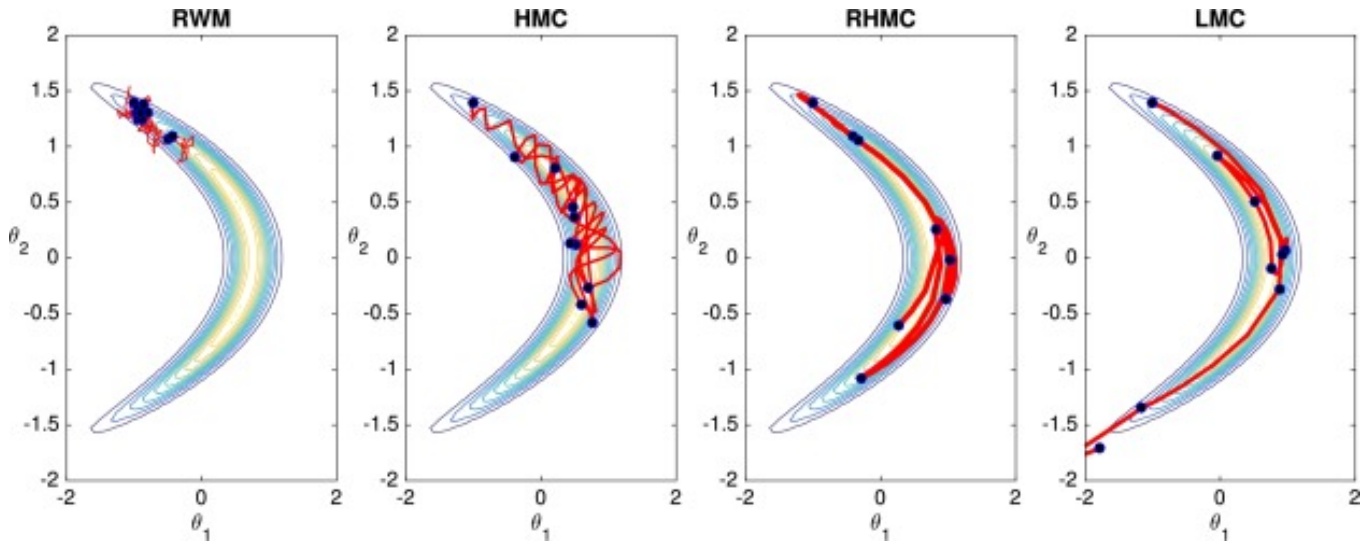
where $H(\cdot)$ called *Hamiltonian*

Keep Hamiltonian target stable over time, as

$$\frac{\mathrm{d}H(\theta_t, \vartheta_t)}{\mathrm{d}t} = \frac{\partial H}{\partial \vartheta}(\theta_t, \vartheta_t)\,\frac{\mathrm{d}\vartheta_t}{\mathrm{d}t} + \frac{\partial H}{\partial \theta}(\theta_t, \vartheta_t)\,\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = 0\,.$$

Background

Approach from physics (Duane et al., 1987) popularised in statistics by Neal (1996, 2002)



[Lan et al., 2016]

Background

- Above continuous time Markov process is deterministic
- Only explores single given level set

$$\{(\theta, \vartheta) : H(\theta, \vartheta) = H(\theta_0, \vartheta_0)\},$$

  instead of the whole augmented state space $\mathbb{R}^{2\times d}$
- Meaning lack of irreducibility
- Solution out is to refresh momentum,

$$\vartheta_t \sim \varpi(\vartheta|\theta_{t-})$$

  at random times τ_n with $\{\tau_n - \tau_{n-1}\}$ exponential variates
- Specific piecewise deterministic Markov process using Hamiltonian dynamics

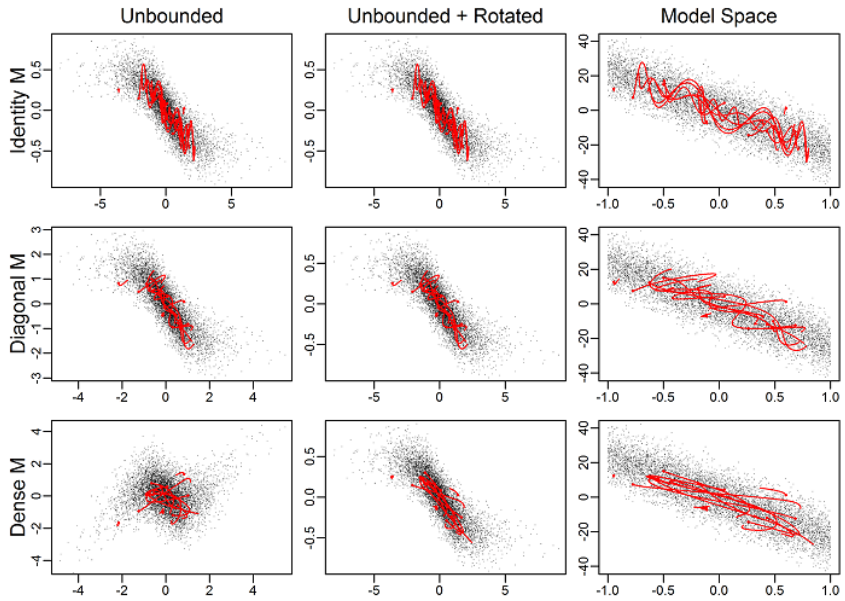
Practical implementation

Free conditional density $\varpi(\vartheta|\theta)$, usually chosen as Gaussian with either a constant covariance matrix M corresponding to target covariance or as local curvature depending on θ in Riemannian HMC (Girolami and Calderhead, 2011)
For fixed covariance matrix M, Hamiltonian equations

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = M^{-1}\vartheta_t \qquad \frac{\mathrm{d}\vartheta_t}{\mathrm{d}t} = \nabla\mathcal{L}(\theta_t)\,,$$

equal to the score function

Practical implementation



For fixed covariance matrix M, Hamiltonian equations

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = M^{-1}\vartheta_t \qquad \frac{\mathrm{d}\vartheta_t}{\mathrm{d}t} = \nabla\mathcal{L}(\theta_t)\,,$$

equal to the score function

Leapfrog integrator

Discretisation simulation technique: symplectic integrator

Leapfrog integrator

One version in the independent case with constant covariance M made of leapfrog steps

$$\begin{aligned}\vartheta_{t+\epsilon/2} &= \vartheta_t + \epsilon\nabla\mathcal{L}(\theta_t)/2,\\ \theta_{t+\epsilon} &= \theta_t + \epsilon M^{-1}\vartheta_{t+\epsilon/2},\\ \vartheta_{t+\epsilon} &= \vartheta_{t+\epsilon/2} + \epsilon\nabla\mathcal{L}(\theta_{t+\epsilon})/2,\end{aligned}$$

where ϵ is time-discretisation step

Using proposal on ϑ_0 drawn from Gaussian auxiliary target and deciding on acceptance of the value of $(\theta_{T\epsilon}, \vartheta_{T\epsilon})$ by a Metropolis–Hastings step

Leapfrog integrator

One version in the independent case with constant covariance M made of leapfrog steps

$$\begin{aligned}\vartheta_{t+\epsilon/2} &= \vartheta_t + \epsilon\nabla\mathcal{L}(\theta_t)/2,\\ \theta_{t+\epsilon} &= \theta_t + \epsilon M^{-1}\vartheta_{t+\epsilon/2},\\ \vartheta_{t+\epsilon} &= \vartheta_{t+\epsilon/2} + \epsilon\nabla\mathcal{L}(\theta_{t+\epsilon})/2,\end{aligned}$$

where ϵ is time-discretisation step

Note that first two leapfrog steps induce a Langevin move on θ_t:

$$\theta_{t+\epsilon} = \theta_t + \epsilon^2 M^{-1}\nabla\mathcal{L}(\theta_t)/2 + \epsilon M^{-1}\vartheta_t\,,$$

Leapfrog integrator

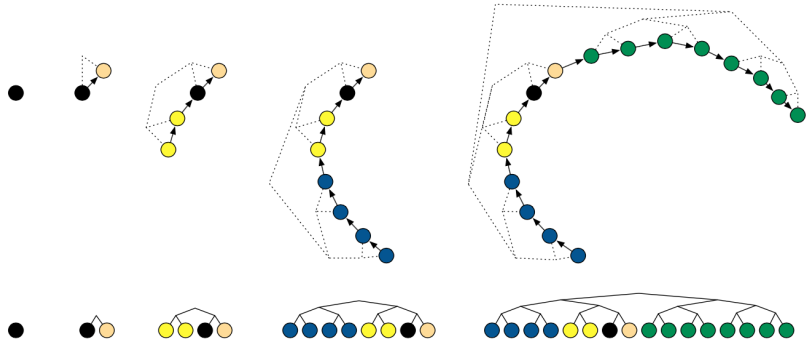
One version in the independent case with constant covariance M made of leapfrog steps

$$\begin{aligned}\vartheta_{t+\epsilon/2} &= \vartheta_t + \epsilon\nabla\mathcal{L}(\theta_t)/2,\\ \theta_{t+\epsilon} &= \theta_t + \epsilon M^{-1}\vartheta_{t+\epsilon/2},\\ \vartheta_{t+\epsilon} &= \vartheta_{t+\epsilon/2} + \epsilon\nabla\mathcal{L}(\theta_{t+\epsilon})/2,\end{aligned}$$

where ϵ is time-discretisation step

Discretising Hamiltonian dynamics introduces two free parameters, step size ϵ and trajectory length $T\epsilon$, both to be calibrated.

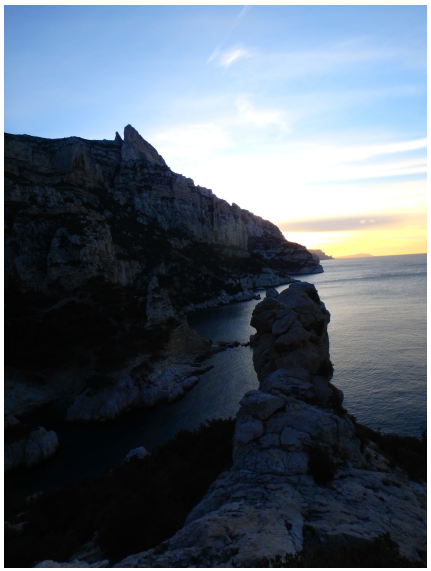
no U turns



- empirically successful and popular version of HMC: "no-U-turn sampler" (NUTS) adapts value of ϵ based on primal-dual averaging
- and eliminates need to choose trajectory length T via a recursive algorithm that builds a set of candidate proposals for a number of forward and backward leapfrog steps, stopping automatically when simulated path traces back

[Hoffman and Gelman, 2014]

Piecewise Deterministic Versions



Generic issue

Goal: sample from a target known up to a constant, defined over $\mathbb{R}^d$,

$$\pi(\mathbf{x}) \propto \gamma(\mathbf{x})$$

with energy $U(\mathbf{x}) = -\log \pi(\mathbf{x})$, $U \in \mathcal{C}^1$.

Marketing arguments

Current default workhorse: reversible MCMC methods

Non-reversible MCMC algorithms based on piecewise deterministic Markov processes perform well empirically

Marketing arguments

Non-reversible MCMC algorithms based on piecewise deterministic Markov processes perform well empirically
Quantitative convergence rates and variance now available

- Physics (Peters & De With, 2012; Krauth et al., 2009, 2015, 2016) roots
- Mesquita and Hespanha (2010) show geometric ergodicity for exponentially decaying tail targets
- Monmarch (2016) gives sharp results for compact state-spaces
- Bierkens et al. (2016a,b) show ergdicity targets on the real line

Motivation: piecewise deterministic Markov process

PDMP sampler is a (new?) continuous-time, non-reversible MCMC method based on auxiliary variables

1. particle physics simulation
   [Peters et al., 2012]
2. empirically state-of-the-art performances
   [Bouchard et al., 2017]
3. exact subsampled in big data
   [Bierkens et al., 2017]
4. geometric ergodicity for a large class of distribution
   [Deligiannidis et al., 2017, Bierkens et al., 2017]
5. Ability to deal with intractable potential $U(x) = \int U_\omega(x)\mu(d\omega)$
   [Pakman et al., 2016]

Older versions

Use of alternative methodology based on Birth–&-Death (point) process

Older versions

Use of alternative methodology based on Birth–&-Death (point) process

Idea: Create Markov chain in *continuous time*, i.e. a *Markov jump process*

Time till next modification **(jump)** exponentially distributed with intensity $q(\theta, \theta')$ depending on current and future states.

[Preston, 1976; Ripley, 1977; Geyer & Møller, 1994; Stevens, 1999]

Older versions

Difference with MH-MCMC: Whenever jump occurs, corresponding move *always accepted*. Acceptance probabilities replaced with holding times.

Implausible configurations

$$L(\theta)\pi(\theta) \ll 1$$

die quickly.

Older versions

Difference with MH-MCMC: Whenever jump occurs, corresponding move *always accepted*. Acceptance probabilities replaced with holding times.
Implausible configurations

$$L(\theta)\pi(\theta) \ll 1$$

die quickly.
Sufficient to have **detailed balance**

$$L(\theta)\pi(\theta)q(\theta,\theta') = L(\theta')\pi(\theta')q(\theta',\theta) \quad \textbf{for all } \theta,\theta'$$

for $\tilde{\pi}(\theta) \propto L(\theta)\pi(\theta)$ to be stationary.

[Cappé et al., 2000]

Setup

All MCMC schemes presented here target an extended distribution on $\mathfrak{Z} = \mathbb{R}^d \times \mathbb{R}^d$

$$\rho(z) = \pi(x) \times \psi(v) = \exp(-H(z))$$

where $z = (x, v)$ extended state and $\Psi(v)$ [by default] multivariate standard Normal
Physics takes v as velocity or momentum variables allowing for a deterministic dynamics on $\mathbb{R}^d$
Obviously sampling from ρ provides samples from π

Piecewise deterministic Markov process

Piecewise deterministic Markov process $\{\mathbf{z}_t \in \mathcal{Z}\}_{t\in[0,\infty)}$, with three ingredients,

1. **Deterministic dynamics**: between events, deterministic evolution based on ODE

$$d\mathbf{z}_t/dt = \Phi(\mathbf{z}_t)$$

2. **Event occurrence rate**: $\lambda(t) = \lambda(\mathbf{z}_t)$
3. **Transition dynamics**: At event time, τ, state prior to τ denoted by $\mathbf{z}_{\tau-}$, and new state generated by $\mathbf{z}_\tau \sim Q(\cdot|\mathbf{z}_{\tau-})$.

[Davis, 1984, 1993]

Implementation

Algorithm 2 Simulation of PDMP

Input: Starting point $\mathbf{z}_0$, $\tau_0 \leftarrow 0$.

for $k = 1, 2, 3, \cdots$

Sample inter-event time η_k from distribution

$$\mathbb{P}(\eta_k > t) = \exp\left\{ - \int_0^t \lambda(\mathbf{z}_{\tau_{k-1}+s}) \mathrm{d}s \right\}.$$

$\tau_k \leftarrow \tau_{k-1} + \eta_k$, $\mathbf{z}_{\tau_{k-1}+s} \leftarrow \Psi_s(\mathbf{z}_{\tau_{k-1}})$, for $s \in (0, \eta_k)$, where Ψ ODE flow of Φ.

$\mathbf{z}_{\tau_k -} \leftarrow \Psi_{\eta_k}(\mathbf{z}_{\tau_{k-1}})$, $\mathbf{z}_{\tau_k} \sim Q(\cdot | \mathbf{z}_{\tau_k -})$.

Simulation of PDMP: constraints

- requires being able to compute exactly flow $z_t = \Phi_t(z_0)$

  existing algorithms use $\Phi(z) = (v; 0_d)$ so thatt $\Phi(z_0) = (x_0 + v_0 t; v_0)$

  except for Hamiltonian BPS that uses the Hamiltonian dynamics for a proxy Gaussian Hamiltonian (Vanetti et al., 2017).
- Requires ability to simulate event times (Inversion, thinning, superposition, Devroye, 1986)
- Requires being able to simulate from Q

Basic bouncy particle sampler

Simulation of continuous-time piecewise linear trajectory $(x_t)_t$ with each segment in trajectory specified by

- initial position x
- length τ
- velocity v

[Bouchard et al., 2017]

Basic bouncy particle sampler

Simulation of continuous-time piecewise linear trajectory $(x_t)_t$ with each segment in trajectory specified by

- initial position x
- length τ
- velocity v

length specified by inhomogeneous Poisson point process with intensity function

$$\lambda(x, v) = \max\{0, < \nabla U(x), v >\}$$

[Bouchard et al., 2017]

Basic bouncy particle sampler

Simulation of continuous-time piecewise linear trajectory $(x_t)_t$ with each segment in trajectory specified by

- initial position x
- length τ
- velocity v

new velocity after bouncing given by Newtonian elastic collision

$$R(x)v = v - 2\frac{< \nabla U(x), v >}{||\nabla U(x)||^2}\nabla U(x)$$

[Bouchard et al., 2017]

Basic bouncy particle sampler

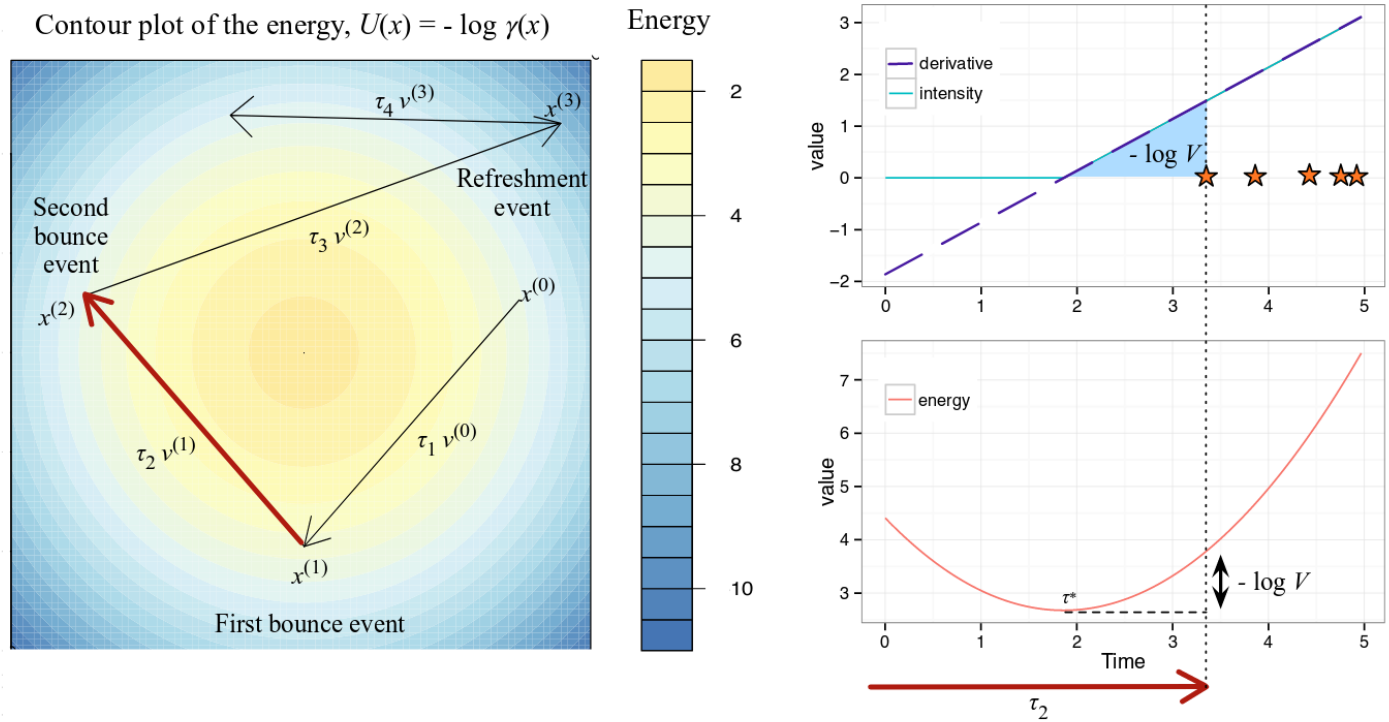


Figure 1: Illustration of BPS on a standard bivariate Gaussian distribution. Left and top right: see Section 2.2; bottom right: see Example 1.

[Bouchard et al., 2017]

Implementation hardships

Generally speaking, the main difficulties of implementing PDMP come from

1. Computing the ODE flow Ψ: linear dynamic, quadratic dynamic
2. Simulating the inter-event time η_k: many techniques of superposition and thinning for Poisson processes

[Devroye, 1986]

Poisson process on $\mathbb{R}_+$

Definition (Poisson process)

Poisson process with rate λ on $\mathbb{R}_+$ is sequence

$$\tau_1, \tau_2, \cdots$$

of rv's when intervals

$$\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \cdots$$

are iid with

$$\mathbb{P}(\tau_i - \tau_{i-1} > T) = \exp\left\{ - \int_{\tau_{i-1}}^{\tau_{i-1}+T} \lambda(t) dt \right\}, \qquad \tau_0 = 0$$

Poisson process on $\mathbb{R}_+$

Definition (Poisson process)

Poisson process with rate λ on $\mathbb{R}_+$ is sequence

$$\tau_1, \tau_2, \cdots$$

of rv's when intervals

$$\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \cdots$$

are iid with **a rarely available cdf**

Simulation by thinning

Theorem (Lewis et al., 1979)

Let

$$\lambda, \Lambda : \mathbb{R}^+ \to \mathbb{R}^+$$

be continuous functions such that $\lambda(\cdot) \leq \Lambda(\cdot)$. *Let*

$$\tau_1, \tau_2, \cdots,$$

be the increasing sequence of a Poisson process with rate $\Lambda(\cdot)$. *For all* i, *if* τ_i *is removed from the sequence with probability*

$$1 - \lambda(t)/\Lambda(t)$$

then the remaining $\tilde{\tau}_1, \tilde{\tau}_2, \cdots$ *form a non-homogeneous Poisson process with rate*$\lambda(\cdot)$

Simulation by thinning

Theorem (Lewis et al., 1979)

Let

$$\lambda, \Lambda : \mathbb{R}^+ \to \mathbb{R}^+$$

be continuous functions such that $\lambda(\cdot) \leq \Lambda(\cdot)$. *Let*

$$\tau_1, \tau_2, \cdots,$$

be the increasing sequence of a Poisson process with rate $\Lambda(\cdot)$.

Simulation from upper bound

Simulation by superposition theorem

Theorem (Kingman,1992)

Let $\Pi_1, \Pi_2, \cdots,$ be countable collection of independent Poisson processes on $\mathbb{R}^+$ with resp. rates $\lambda_n(\cdot)$. If $\sum_{n=1}^{\infty} \lambda_n(t) < \infty$ for all t's, then superposition process

$$\Pi = \bigcup_{n=1}^{\infty} \Pi_n$$

is Poisson process with rate

$$\lambda(t) = \sum_{n=1}^{\infty} \lambda_n(t)$$

Simulation by superposition theorem

Theorem (Kingman,1992)

Let $\Pi_1, \Pi_2, \cdots$, be countable collection of independent Poisson processes on $\mathbb{R}^+$ with resp. rates $\lambda_n(\cdot)$. If $\sum_{n=1}^{\infty} \lambda_n(t) < \infty$ for all t's, then superposition process is Poisson process with rate $lambda(t)$

Decomposition of $U = \sum_j U_j$ plus thinning

Simulation by superposition plus thinning

For(z) = (x + vAlmost all implementations of discrete-time schemes of consist of sampling a Bernoulli of parameter $\alpha(z)$
For $Phi(z) = (x + v\epsilon, v)$ and $\alpha(z) = 1 \wedge \pi(x + v\epsilon)/\pi(x)$, sampling inter-event time for strictly convex U can be obtained by solving $t^\star = \arg\min U(x + vt)$ and additional randomization

- thinning: if there exists $\bar{\alpha}$ such that $\alpha(\Phi^k(z)) \geq \bar{\alpha}(x, k)$, accept-reject
- superposition and thinning: when $\alpha(z) = 1 \wedge \rho(\Phi(z))/\rho(z)$ and $\rho(\cdot) = \prod_i \rho_i(\cdot)$ then $\bar{\alpha}(z, k) = \prod_i \bar{\alpha}_i(z, k)$

Extended generator

Definition

For $\mathcal{D}(\mathcal{L})$ set of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ such that there exists a measurable function $h : \mathcal{Z} \to \mathbb{R}$ with $t \to h(\mathbf{z}_t)$ $P_{\mathbf{z}}$-a.s. for each $\mathbf{z} \in \mathcal{Z}$ and the process

$$C_t^f = f(\mathbf{z}_t) - f(\mathbf{z}_0) - \int_0^t h(\mathbf{z}_s) ds$$

a local martingale. Then we write $h = \mathcal{L}f$ and call $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$ the extended generator of the process $\{\mathbf{z}_t\}_{t \geq 0}$.

Extended Generator of PDMP

Theorem (Davis, 1993)

The generator, $\mathcal{L}$, of above PDMP is, for $f \in \mathcal{D}(\mathcal{L})$

$$\mathcal{L}f(\mathbf{z}) = \nabla f(\mathbf{z}) \cdot \Phi(\mathbf{z}) + \lambda(\mathbf{z}) \int_{\mathbf{z}'} \left[f(\mathbf{z}') - f(\mathbf{z}) \right] Q(d\mathbf{z}'|\mathbf{z})$$

Furthermore, $\mu(d\mathbf{z})$ is an invariant distribution of above PDMP, if

$$\int \mathcal{L}f(\mathbf{z})\mu(d\mathbf{z}) = 0, \quad \text{for all } f \in \mathcal{D}(\mathcal{L})$$

PDMP-based sampler

PDMP-based sampler is an auxiliary variable technique

Given target $\pi(\mathbf{x})$,

1. introduce auxiliary variable $\mathbf{V} \in \mathcal{V}$ along with a density $\pi(\mathbf{v}|\mathbf{x})$,
2. choose appropriate Φ, λ and Q

for $\pi(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})$ to be unique invariant distribution of Markov process

Bouncy Particle Sampler (Bouchard et al., 2017)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(\mathbf{v}|\mathbf{x}) = \varphi(\mathbf{v})$ for $\mathcal{N}(0, I_d)$

1. **Deterministic dynamics**:

$$d\mathbf{x}_t/dt = \mathbf{v}_t, d\mathbf{v}_t/dt = \mathbf{0}$$

2. **Event occurrence rate**: $\lambda(\mathbf{x}, \mathbf{v}) = \langle \mathbf{v}, \nabla U(\mathbf{x}) \rangle_+ + \lambda^{\text{ref}}$
3. **Transition dynamics**:

$$\begin{aligned} &Q((d\mathbf{x}', d\mathbf{v}')|(\mathbf{x}, \mathbf{v})) \\ &= \frac{\langle \mathbf{v}, \nabla U(\mathbf{x}) \rangle_+}{\lambda(\mathbf{x}, \mathbf{v})} \delta_{\mathbf{x}}(d\mathbf{x}') \delta_{R_{\nabla U(\mathbf{x})}\mathbf{v}}(d\mathbf{v}') + \frac{\lambda^{\text{ref}}}{\lambda(\mathbf{x}, \mathbf{v})} \delta_{\mathbf{x}}(d\mathbf{x}') \varphi(d\mathbf{v}') \end{aligned}$$

where $R_{\nabla U(\mathbf{x})}\mathbf{v} = \mathbf{v} - 2\frac{\langle \nabla U(\mathbf{x}), \mathbf{v} \rangle}{\langle \nabla U(\mathbf{x}), \nabla U(\mathbf{x}) \rangle} \nabla U(\mathbf{x})$

Zig-Zag Sampler (Bierkens et al., 2016)

Zig-Zag Sampler (Bierkens et al., 2016)

$\mathcal{V} = \{+1, -1\}^d$, and $\pi(\mathbf{v}|\mathbf{x}) \sim \text{Uniform}(\{+1, -1\}^d)$

1. **Deterministic dynamics**:

$$d\mathbf{x}_t/dt = \mathbf{v}_t, d\mathbf{v}_t/dt = \mathbf{0}$$

2. **Event occurrence rate**:
$\lambda(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^d \lambda_i(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^d \left[\{v_i \nabla_i U(\mathbf{x})\}_+ + \lambda_i^{\text{ref}}\right]$
3. **Transition dynamics**:

$$Q((d\mathbf{x}', d\mathbf{v}')|(\mathbf{x}, \mathbf{v})) = \sum_{i=1}^d \frac{\lambda_i(\mathbf{x}, \mathbf{v})}{\lambda(\mathbf{x}, \mathbf{v})} \delta_{\mathbf{x}}(d\mathbf{x}') \delta_{F_i \mathbf{v}}(d\mathbf{v}')$$

where F_i operator that flips i-th component of $\mathbf{v}$ and keep others unchanged

Continuous-time Hamiltonian Monte Carlo (Neal, 1999)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(\mathbf{v}|\mathbf{x}) = \varphi(\mathbf{v}) \sim \mathcal{N}(0, I_d)$

1. **Deterministic dynamics**:

$$d\mathbf{x}_t/dt = \mathbf{v}_t, d\mathbf{v}_t/dt = -\nabla U(\mathbf{x}_t)$$

2. **Event occurrence rate**: $\lambda(\mathbf{x}, \mathbf{v}) = \lambda_0(\mathbf{x})$
3. **Transition dynamics**:

$$Q((d\mathbf{x}', d\mathbf{v}')|(\mathbf{x}, \mathbf{v})) = \delta_{\mathbf{x}}(d\mathbf{x}')\varphi(d\mathbf{v}')$$

Continuous-time Riemannian Manifold HMC (Girolami & Calderhead, 2011)

$\mathcal{V} = \mathbb{R}^d$, and $\pi(\mathbf{v}|\mathbf{x}) = \mathcal{N}(0, G(\mathbf{x}))$, the Hamiltonian is

$$H(\mathbf{x}, \mathbf{v}) = U(\mathbf{x}) + {}^1\!/\!_2 \mathbf{v}^T G(\mathbf{x})^{-1} \mathbf{v} + {}^1\!/\!_2 \log(|G(\mathbf{x})|)$$

1. **Deterministic dynamics**:

$$d\mathbf{x}_t/dt = \partial H/\partial \mathbf{v}(\mathbf{x}_t, \mathbf{v}_t), d\mathbf{v}_t/dt = -\partial H/\partial \mathbf{x}(\mathbf{x}_t, \mathbf{v}_t)$$

2. **Event occurrence rate**: $\lambda(\mathbf{x}, \mathbf{v}) = \lambda_0(\mathbf{x})$
3. **Transition dynamics**:

$$Q((d\mathbf{x}', d\mathbf{v}')|(\mathbf{x}, \mathbf{v})) = \delta_{\mathbf{x}}(d\mathbf{x}')\varphi(d\mathbf{v}'|\mathbf{x}')$$

Randomized BPS

Define

$$\mathbf{a} = \frac{\langle \mathbf{v}, \nabla U(\mathbf{x}) \rangle}{\langle \nabla U(\mathbf{x}), \nabla U(\mathbf{x}) \rangle} \nabla U(\mathbf{x}), \quad \mathbf{b} = \mathbf{v} - \mathbf{a}$$

Regular BPS, move $\mathbf{v}' = -\mathbf{a} + \mathbf{b}$

Alternatives

1. (Fearnhead et al., 2016):
   $\mathbf{v}' \sim Q_{\mathbf{x}}(d\mathbf{v}'|\mathbf{v}) = \max\{0, \langle -\mathbf{v}', \nabla U(\mathbf{x}) \rangle\} \, d\mathbf{v}'$
2. (Wu & X, 2017): $\mathbf{v}' = -\mathbf{a} + \mathbf{b}'$, where $\mathbf{b}'$ Gaussian variate over the space orthogonal to $\nabla U(\mathbf{x})$ in $\mathbb{R}^d$.

HMC-BPS (Vanetti et al., 2017)

$\rho(\mathbf{x}) \propto \exp\{-V(\mathbf{x})\}$ is a Gaussian approximation of the target $\pi(\mathbf{x})$.

$$\hat{H}(\mathbf{x}, \mathbf{v}) = V(\mathbf{x}) + {}^1\!/\!_2 \mathbf{v}^T \mathbf{v}, \quad \tilde{U}(\mathbf{x}) = U(\mathbf{x}) - V(\mathbf{x})$$

1. **Deterministic dynamics**:

$$d\mathbf{x}_t/dt = \mathbf{v}_t, d\mathbf{v}_t/dt = -\nabla V(\mathbf{x}_t)$$

2. **Event occurrence rate**: $\lambda(\mathbf{x}, \mathbf{v}) = \langle \mathbf{v}, \nabla \tilde{U}(\mathbf{x}) \rangle_+ + \lambda^{\text{ref}}$
3. **Transition dynamics**:

$$Q((d\mathbf{x}', d\mathbf{v}')|(\mathbf{x}, \mathbf{v}))$$
$$= \frac{\langle \mathbf{v}, \nabla \tilde{U}(\mathbf{x}) \rangle_+}{\lambda(\mathbf{x}, \mathbf{v})} \delta_{\mathbf{x}}(d\mathbf{x}') \delta_{R_{\nabla \tilde{U}(\mathbf{x})}\mathbf{v}}(d\mathbf{v}') + \frac{\lambda^{\text{ref}}}{\lambda(\mathbf{x}, \mathbf{v})} \delta_{\mathbf{x}}(d\mathbf{x}') \varphi(d\mathbf{v}')$$

Discretisation

1. Sherlock & Thiery (2017) considers delayed rejection approach with only point-wise evaluations of target, by making speed flip move once proposal involving flip in speed and drift in variable of interest rejected. Also add random perturbation for eergodicity, plus another perturbation based on a Brownian argument. Requires calibration
2. Vanetti et al. (2017)

Benefit: bypassing the generation of inter-event time of inhomogeneous Poisson processes.

Discretisation

1. Sherlock & Thiery (2017)
2. Vanetti et al. (2017) unifies many threads and relates PDMP, HMC, and discrete versions, with convergence results. Main idea improves upon existing deterministic methods by accounting for target. Borrows from earlier slice sampler idea of Murray et al. (AISTATS, 2010), exploiting exact Hamiltonian dynamics for approximation to true target. Except that bouncing avoids the slice step. Eight discrete BPS both correct against target and do not simulating event times.

Benefit: bypassing the generation of inter-event time of inhomogeneous Poisson processes.