Stat 535 C - Statistical Computing & Monte Carlo Methods

Lecture 13 - 28 February 2006

Arnaud Doucet

Email: arnaud@cs.ubc.ca

1.1– Outline

- Limitations of Gibbs sampling.
- Metropolis-Hastings algorithm.
- Proof of validity.
- Toy examples.
- Generalizations.

1.2– Gibbs sampler

• If $\theta = (\theta_1, ..., \theta_p)$ where $p > 2$, the Gibbs sampling strategy still applies.

• Initialization:

  • Select deterministically or randomly $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$.

• Iteration i; $i \geq 1$:

  For $k = 1 : p$

    • Sample $\theta_k^{(i)} \sim \pi\left(\theta_k | \, \theta_{-k}^{(i)}\right)$.

  where $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.

1.3– Limitations of the Gibbs sampler

• The Gibbs sampler requires sampling from the full conditional distributions

$$\pi\left(\theta_k \middle| \theta_{-k}\right).$$

• For many complex models, it is impossible to sample from several of these "full" conditional distributions.

• Even if it is possible to implement the Gibbs sampler, the algorithm might be very inefficient because the variables are very correlated or sampling from the full conditionals is extremely expensive/inefficient.

2.1– Metropolis-Hastings algorithm

• The Metropolis-Hastings algorithm is an alternative algorithm to sample from probability distribution $\pi(\theta)$ known up to a normalizing constant.

• This can be interepreted as the basis of all MCMC algorithm:
It provides a generic way to build a Markov kernel admitting $\pi(\theta)$ as an invariant distribution.

• The Metropolis algorithm was named the "Top algorithm of the 20th century" by computer scientists, mathematicians, physicists.

2.2– Description of the algorithm

• Introduce a proposal distribution/kernel $q(\theta, \theta')$, i.e.

$$\int q(\theta, \theta')\, d\theta' = 1 \text{ for any } \theta.$$

• The basic idea of the MH algorithm is to propose a new candidate θ' based on the current state of the Markov chain θ.

• We only accept this algorithm with respect to a probability $\alpha(\theta, \theta')$ which ensures that the invariant distribution of the transition kernel is the target distribution $\pi(\theta)$.

2.3– Metropolis-Hastings algorithm

- Initialization:
  - Select deterministically or randomly $\theta^{(0)}$.
- Iteration i; $i \geq 1$:
  - Sample $\theta^* \sim q\left(\theta^{(i-1)}, \theta^*\right)$ and compute

$$\alpha\left(\theta^{(i-1)}, \theta^*\right) = \min\left(1, \frac{\pi\left(\theta^*\right) q\left(\theta^*, \theta^{(i-1)}\right)}{\pi\left(\theta^{(i-1)}\right) q\left(\theta^{(i-1)}, \theta^*\right)}\right).$$

  - With probability $\alpha\left(\theta^{(i-1)}, \theta^*\right)$, set $\theta^{(i)} = \theta^*$; otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

2.3– Metropolis-Hastings algorithm

• It is not necessary to know the normalizing constant of $\pi(\theta)$ to implement the algorithm.

• This algorithm is extremely general: $q(\theta, \theta')$ can be any proposal distribution. So in practice, we can select it so that it is easy to sample from it.

• There is much more freedom than in the Gibbs sampler where the proposal distributions are fixed.

2.4– Metropolis algorithm

• The original Metropolis algorithm (1953) corresponds to the following choice for $q(\theta, \theta')$

$$\theta' = \theta + Z \text{ where } Z \sim f;$$

i.e. this is a so-called *random walk proposal*.

• The distribution $f(z)$ is the distribution of the random walks increments Z and

$$q(\theta, \theta') = f(\theta' - \theta) \quad \Rightarrow \; \alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta') f(\theta - \theta')}{\pi(\theta) f(\theta' - \theta)}\right).$$

• If $f(\theta' - \theta) = f(\theta - \theta')$ - e.g. $Z \sim \mathcal{N}(0, \Sigma)$- then

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right)$$

2.5– Metropolis algorithm

• The Hastings' generalization (1970) corresponds to the following choice for $q(\theta, \theta')$

$$q(\theta, \theta') = q(\theta');$$

i.e. this is a so-called *independent proposal*.

• In this case, the acceptance probability is given by

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\, q(\theta)}{\pi(\theta)\, q(\theta')}\right) = \min\left(1, \frac{\pi(\theta')}{q(\theta')}\frac{q(\theta)}{\pi(\theta)}\right) = \min\left(1, \frac{\pi^*(\theta')}{q^*(\theta')}\frac{q^*(\theta)}{\pi^*(\theta)}\right)$$

where π^* and q^* are unnormalized versions of π and q.

• The ratio $\pi^*(\theta)/q^*(\theta)$ appearing in the Accept/Reject and Importance Sampling methods also reappears here.

2.6– Properties of the Metropolis-Hastings algorithm

• To establish that the MH chain converges towards the required target, we need to show that

• $\pi(\theta)$ is the invariant distribution of the Markov kernel associated to the MH algorithm.

• The Markov chain is irreducible; i.e. one can reach any set A such that $\pi(A) > 0$.

• The Markov chain is aperiodic; i.e. one does not visit in a periodic way the state-space.

2.7– Invariant distribution of the Metropolis-Hastings algorithm

• The transition kernel associated to the MH algorithm can be rewritten as

$$K(\theta,\theta') = \alpha(\theta,\theta')\, q(\theta,\theta') + \underbrace{\left(1 - \int \alpha(\theta,u)\, q(\theta,u)\, du\right)}_{\text{rejection probability}} \delta_\theta(\theta')$$

Remark: This is a lose notation for

$$K(\theta, d\theta') = \alpha(\theta,\theta')\, q(\theta,\theta')\, d\theta' + \left(1 - \int \alpha(\theta,u)\, q(\theta,u)\, du\right) \delta_\theta(d\theta').$$

• Clearly we have

$$\begin{aligned}\int K(\theta,\theta')\, d\theta' &= \int \alpha(\theta,\theta')\, q(\theta,\theta')\, d\theta' + \left(1 - \int \alpha(\theta,u)\, q(\theta,u)\, du\right) \int \delta_\theta(\theta')\, d\theta' \\ &= 1.\end{aligned}$$

2.7– Invariant distribution of the Metropolis-Hastings algorithm

- We want to show that

$$\int \pi(\theta) K(\theta, \theta') \, d\theta = \pi(\theta') .$$

- Note that this condition is satisfied if the *reversibility property* is satisfied:

For all θ, θ'

$$\pi(\theta) K(\theta, \theta') = \pi(\theta') K(\theta', \theta) ;$$

i.e. the probability of being in A and moving to B is equal to the probability of being in B and moving to A.

2.7– Invariant distribution of the Metropolis-Hastings algorithm

- Indeed the reversibility condition implies that

$$\begin{aligned}\int \pi(\theta) K(\theta, \theta') d\theta &= \int \pi(\theta') K(\theta', \theta) d\theta \\ &= \pi(\theta') \int K(\theta', \theta) d\theta \\ &= \pi(\theta')\end{aligned}$$

2.7– Invariant distribution of the Metropolis-Hastings algorithm

• Be careful: If a kernel is $\pi-$reversible then it is $\pi-$invariant but the reverse is not true.

• The deterministic scan Gibbs sampler is not $\pi-$reversible as

$$\begin{aligned} & \pi\left(\theta_1, \theta_2\right) \pi\left(\theta_2^{\prime} \middle| \theta_1\right) \pi\left(\theta_1^{\prime} \middle| \theta_2^{\prime}\right) \\ \neq \quad & \pi\left(\theta_1^{\prime}, \theta_2^{\prime}\right) \pi\left(\theta_2 \middle| \theta_1^{\prime}\right) \pi\left(\theta_2^{\prime} \middle| \theta_1^{\prime}\right). \end{aligned}$$

2.8– Proof that the MH kernel is reversible

- By definition of the kernel, we have

$$\pi(\theta) K(\theta, \theta') = \pi(\theta) \alpha(\theta, \theta') q(\theta, \theta') + \left(1 - \int \alpha(\theta, u) q(\theta, u) du\right) \delta_\theta(\theta') \pi(\theta).$$

- Then

$$\begin{aligned}
\pi(\theta) \alpha(\theta, \theta') q(\theta, \theta') &= \pi(\theta) \min\left(1, \frac{\pi(\theta') q(\theta', \theta)}{\pi(\theta) q(\theta, \theta')}\right) q(\theta, \theta') \\
&= \min\left(\pi(\theta) q(\theta, \theta'), \pi(\theta') q(\theta', \theta)\right) \\
&= \pi(\theta') \min\left(1, \frac{\pi(\theta) q(\theta, \theta')}{\pi(\theta') q(\theta', \theta)}\right) q(\theta', \theta) \\
&= \pi(\theta') \alpha(\theta', \theta) q(\theta', \theta).
\end{aligned}$$

2.8– Proof that the MH kernel is reversible

- We have obviously

$$\left(1-\int \alpha\left(\theta,u\right)q\left(\theta,u\right)du\right)\delta_{\theta}\left(\theta'\right)\pi\left(\theta\right)=\left(1-\int \alpha\left(\theta',u\right)q\left(\theta',u\right)du\right)\delta_{\theta'}\left(\theta\right)\pi\left(\theta'\right).$$

- It follows that

$$\pi\left(\theta\right)K\left(\theta,\theta'\right)=\pi\left(\theta'\right)K\left(\theta',\theta\right).$$

- Hence, π is the invariant distribution of the transition kernel K.

2.8– Proof that the MH kernel is reversible

• To ensure irreducibility, a sufficient but not necessary condition is that

$$\pi(\theta') > 0 \Rightarrow q(\theta, \theta') > 0.$$

• Aperiodicity is automatically ensured as there is always a strictly positive probability to reject the candidate.

• Theoretically, the MH algorithm converges under very weak assumptions to the target distribution π. In practice, this convergence can be so slow that the algorithm is useless.

2.9– How to select the proposal distribution

• If you are using independent proposals then you would like to have $q(\theta) \simeq \pi(\theta)$.

• In practice, similarly to Rejection sampling or Importance Sampling, you need to ensure that

$$\frac{\pi(\theta)}{q(\theta)} \leq C$$

to obtain good performance.

• If you don't ensure this condition, the algorithm might give you the impression it works well... but it does NOT.

2.10– Example

- **Example**: Consider the case where

$$\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right).$$

- We implement the MH algorithm for

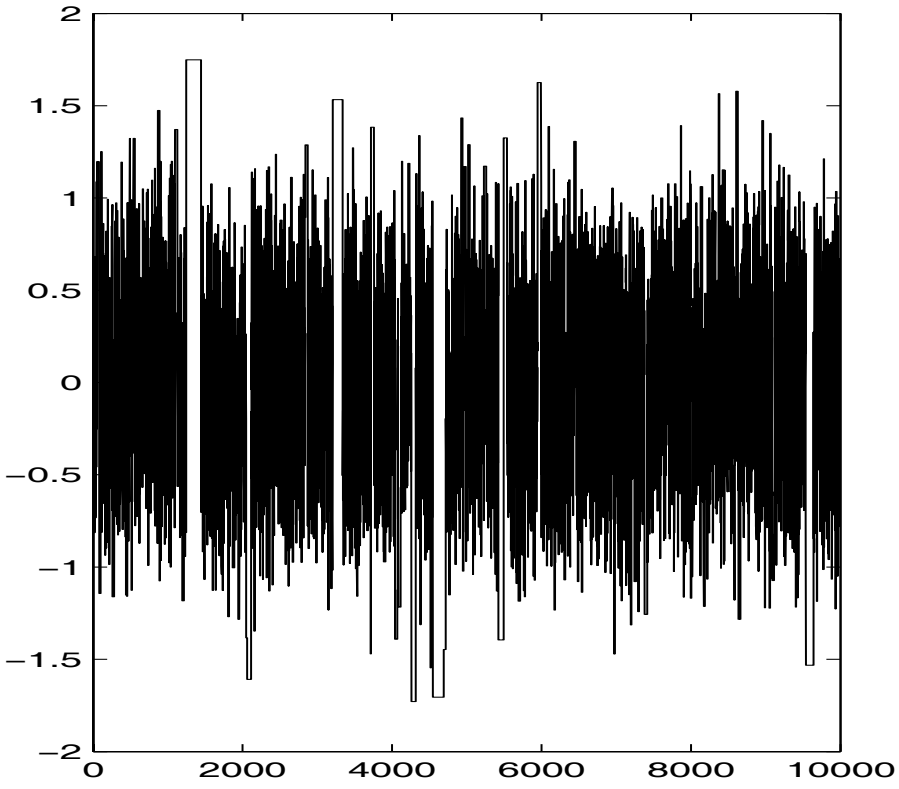
$$q_1(\theta) \propto \exp\left(-\frac{\theta^2}{2(0.2)^2}\right)$$

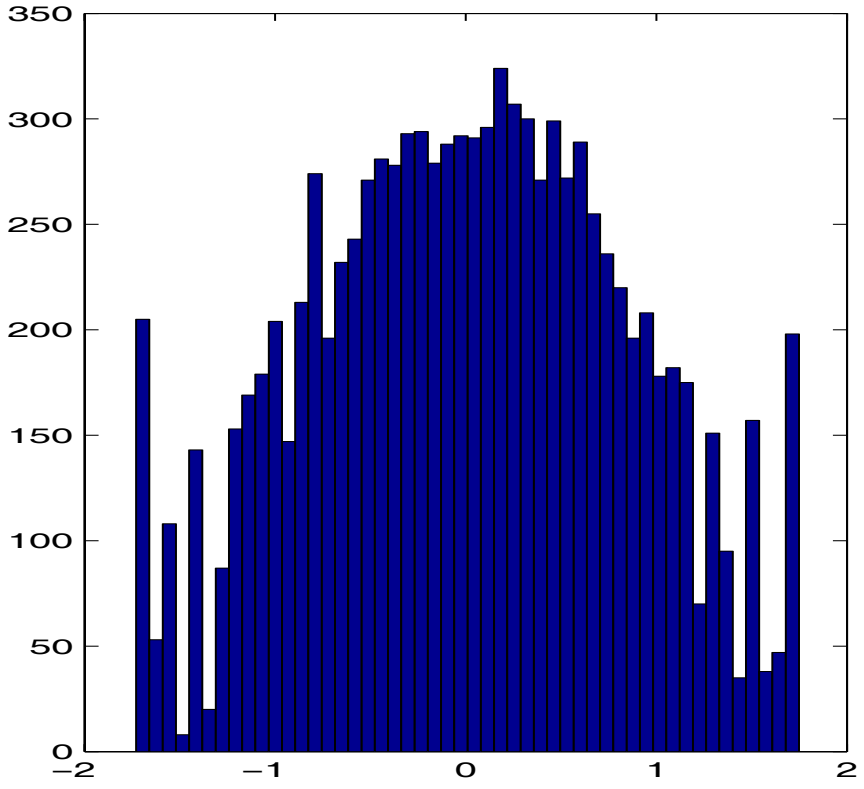
so $\pi(\theta)/q_1(\theta) \to \infty$ as $\theta \to \infty$ and for

$$q_2(\theta) \propto \exp\left(-\frac{\theta^2}{2(5)^2}\right)$$

so $\pi(\theta)/q_2(\theta) \leq C < \infty$ for all θ.

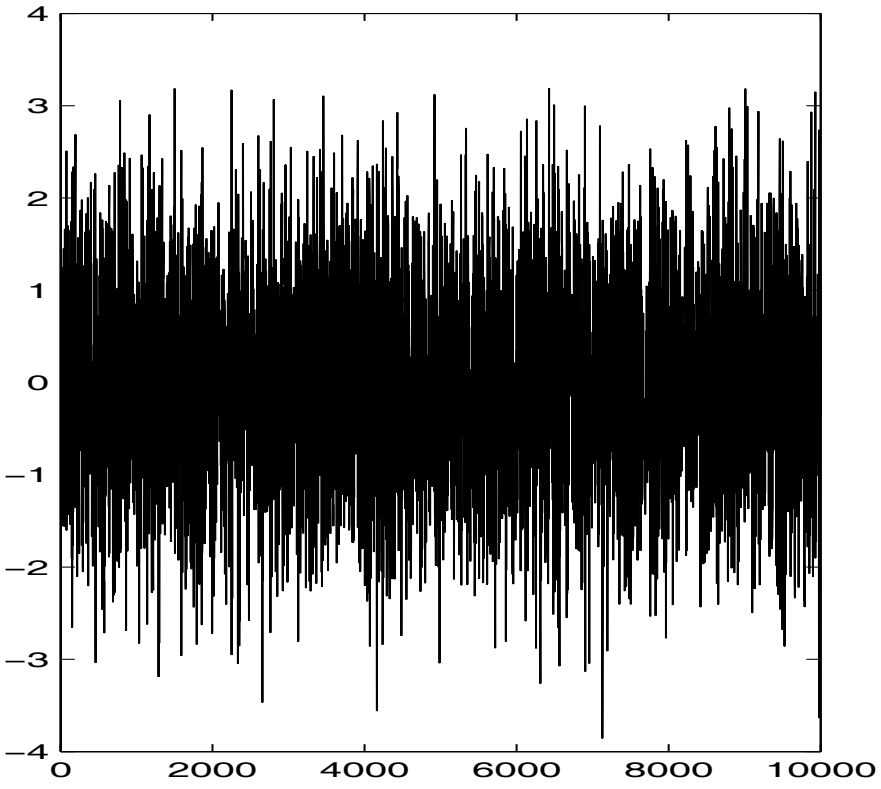
2.10– Example

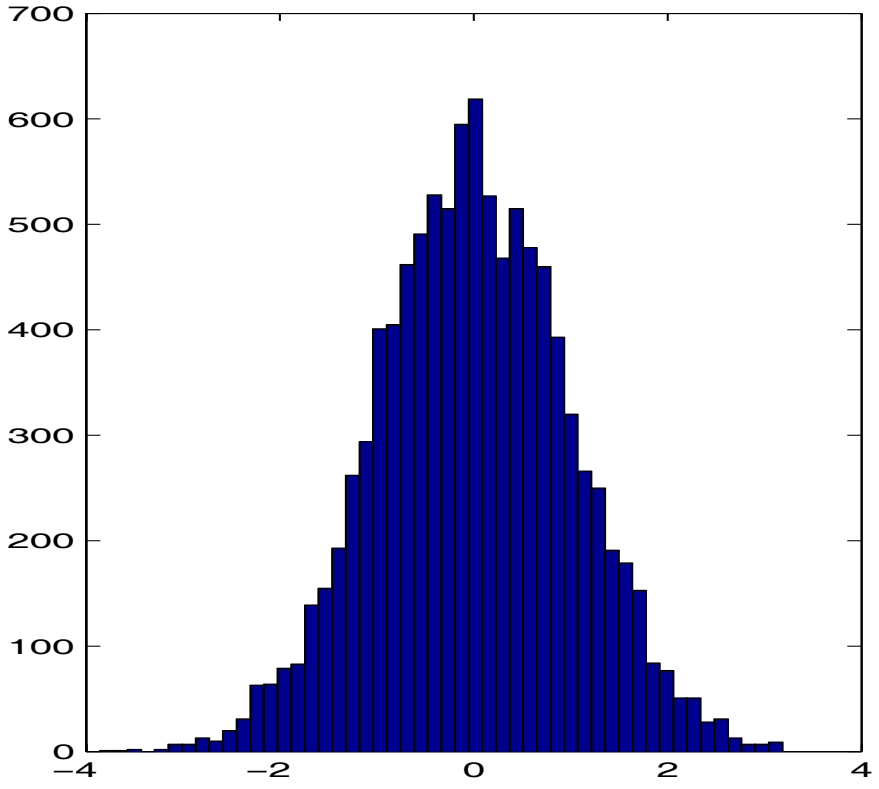




MCMC output for q_1, we estimate $E(\theta) = 0.0206$ and $var(\theta) = 0.83$

2.10– Example





MCMC output for q_2, we estimate $E(\theta) = -0.004$ and $var(\theta) = 1.00$

2.11– How to select the proposal?

• Consider now a random walk move. In this case, there is no clear guideline how to select the proposal distribution.

• When the variance of the random walk increments (if it exists) is very small then the acceptance rate can be expected to be around 0.5-0.7.

• You would like to scale the random walk moves such that it is possible to move reasonably fast in regions of positive probability masses under π.

2.12– Example

- **Example**: Consider the case where

$$\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right).$$

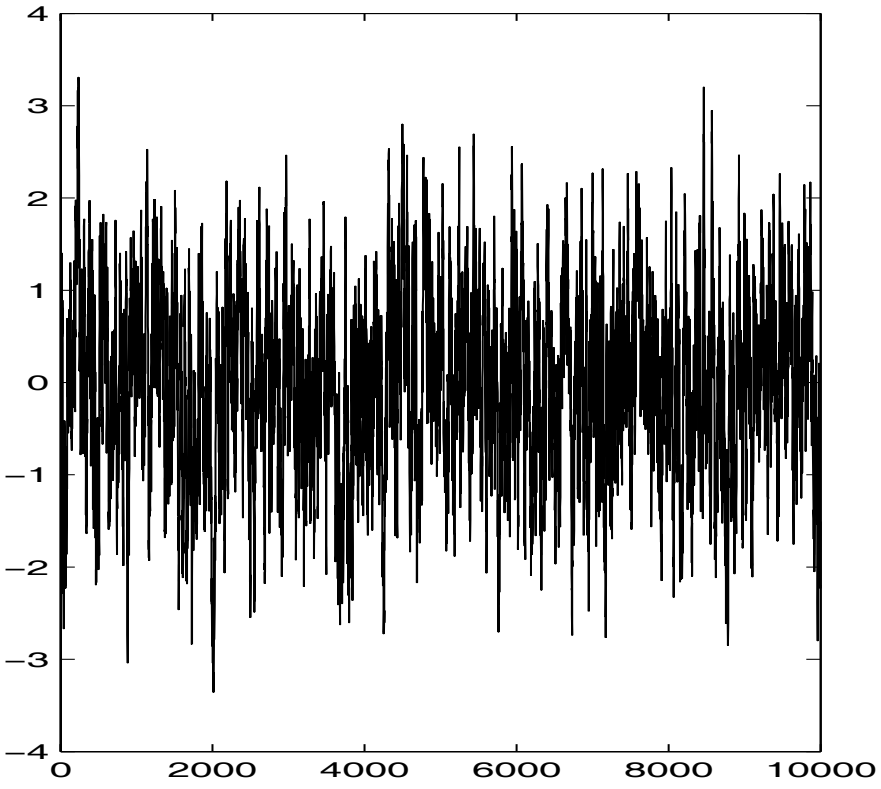
- We implement the MH algorithm for

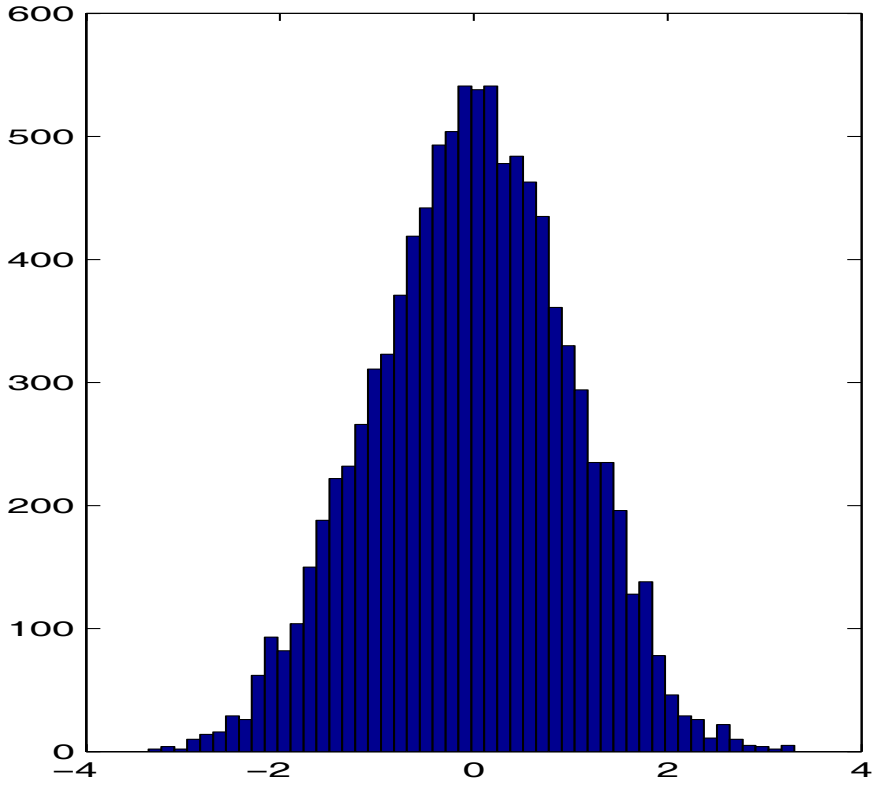
$$q_1(\theta, \theta') \propto \exp\left(-\frac{(\theta' - \theta)^2}{2(0.2)^2}\right).$$

- We implement the MH algorithm for

$$q_2(\theta, \theta') \propto \exp\left(-\frac{(\theta' - \theta)^2}{2(5)^2}\right).$$

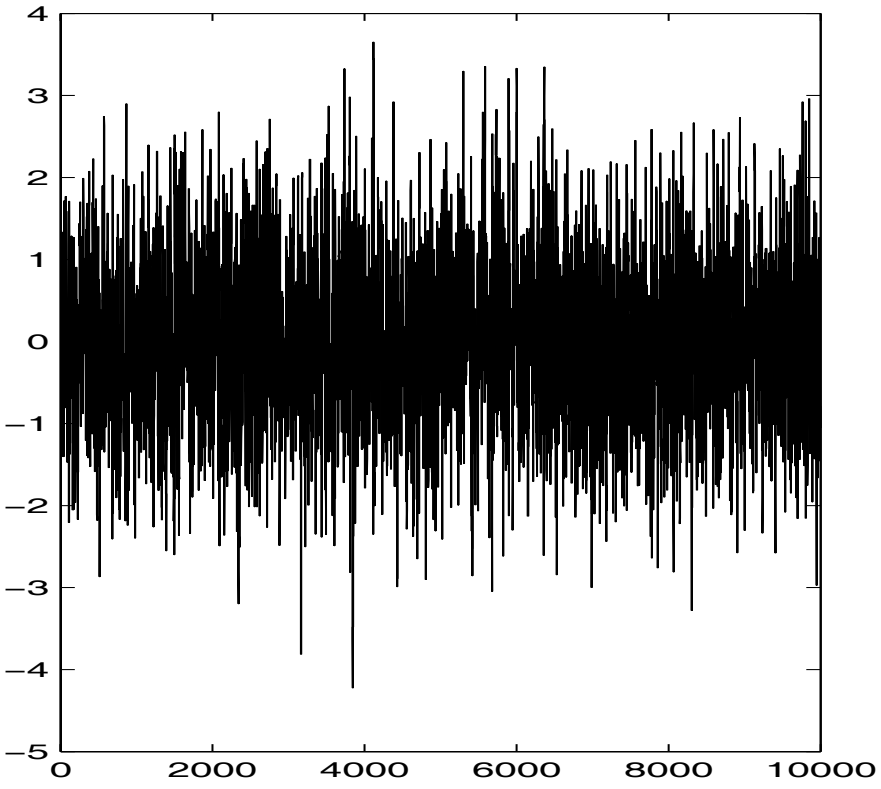
2.12– Example

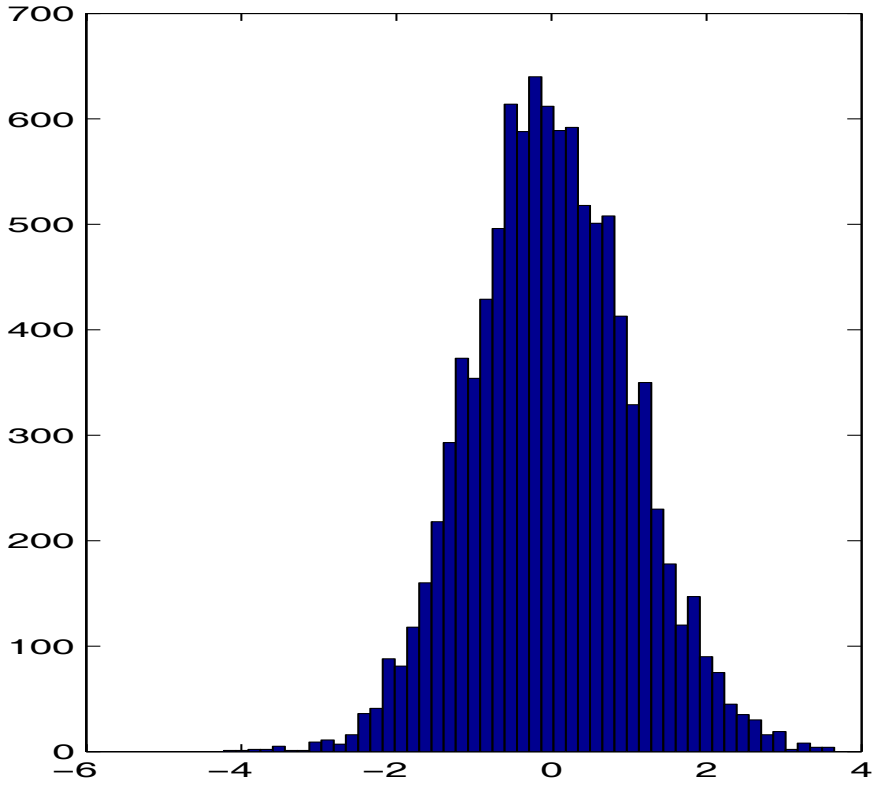




MCMC output for q_1, we estimate $E(\theta) = -0.02$ and $var(\theta) = 0.99$

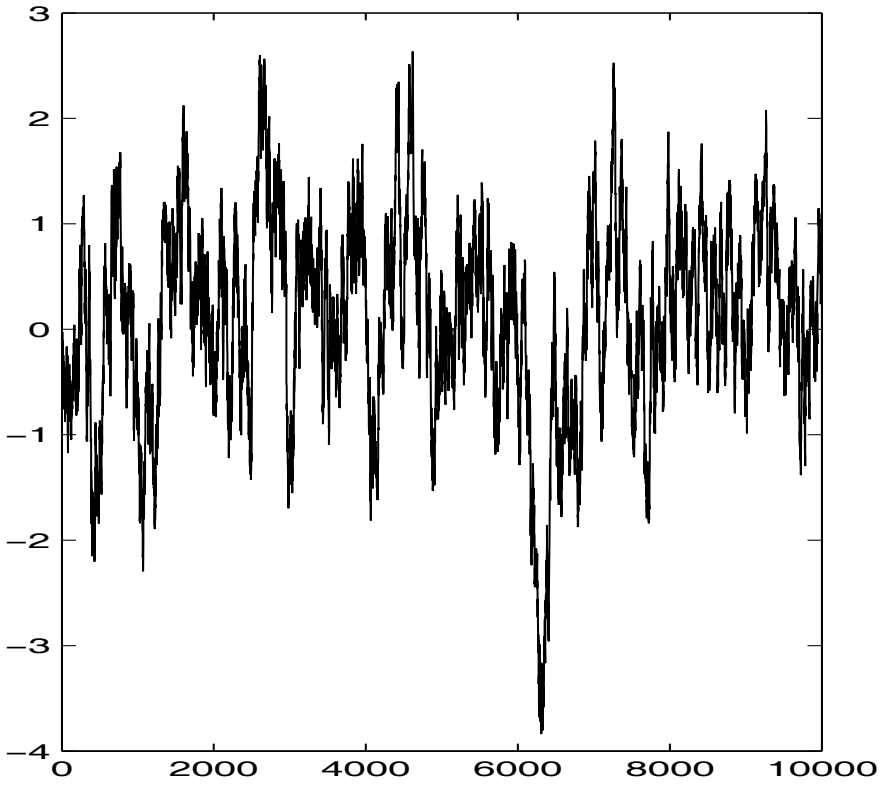
2.12– Example

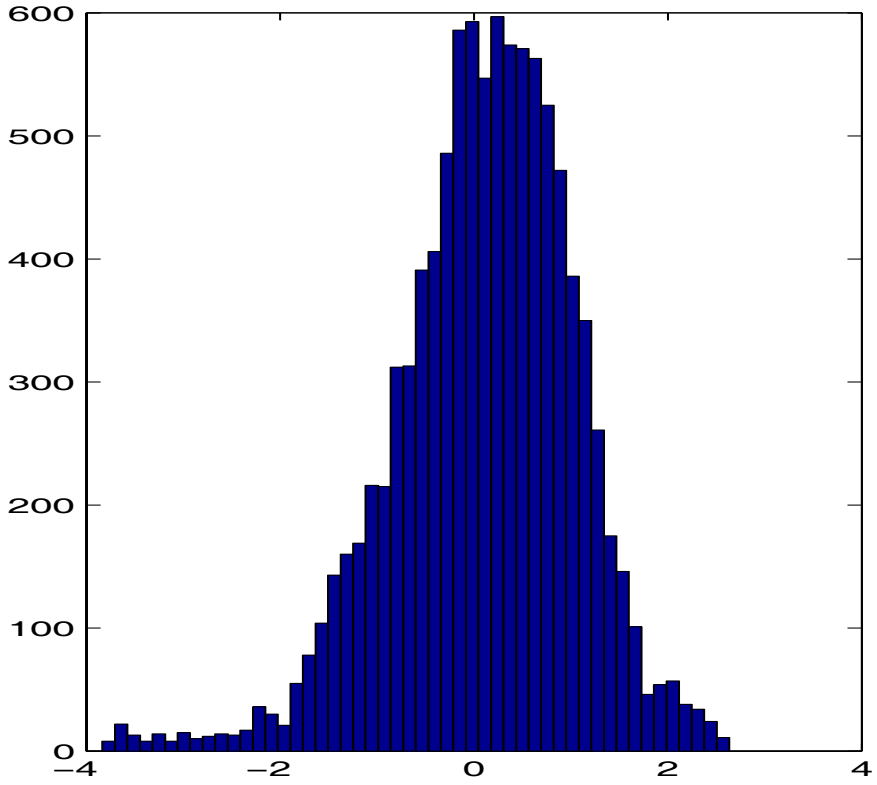




MCMC output for q_2, we estimate $E(\theta) = 0.00$ and $var(\theta) = 1.02$

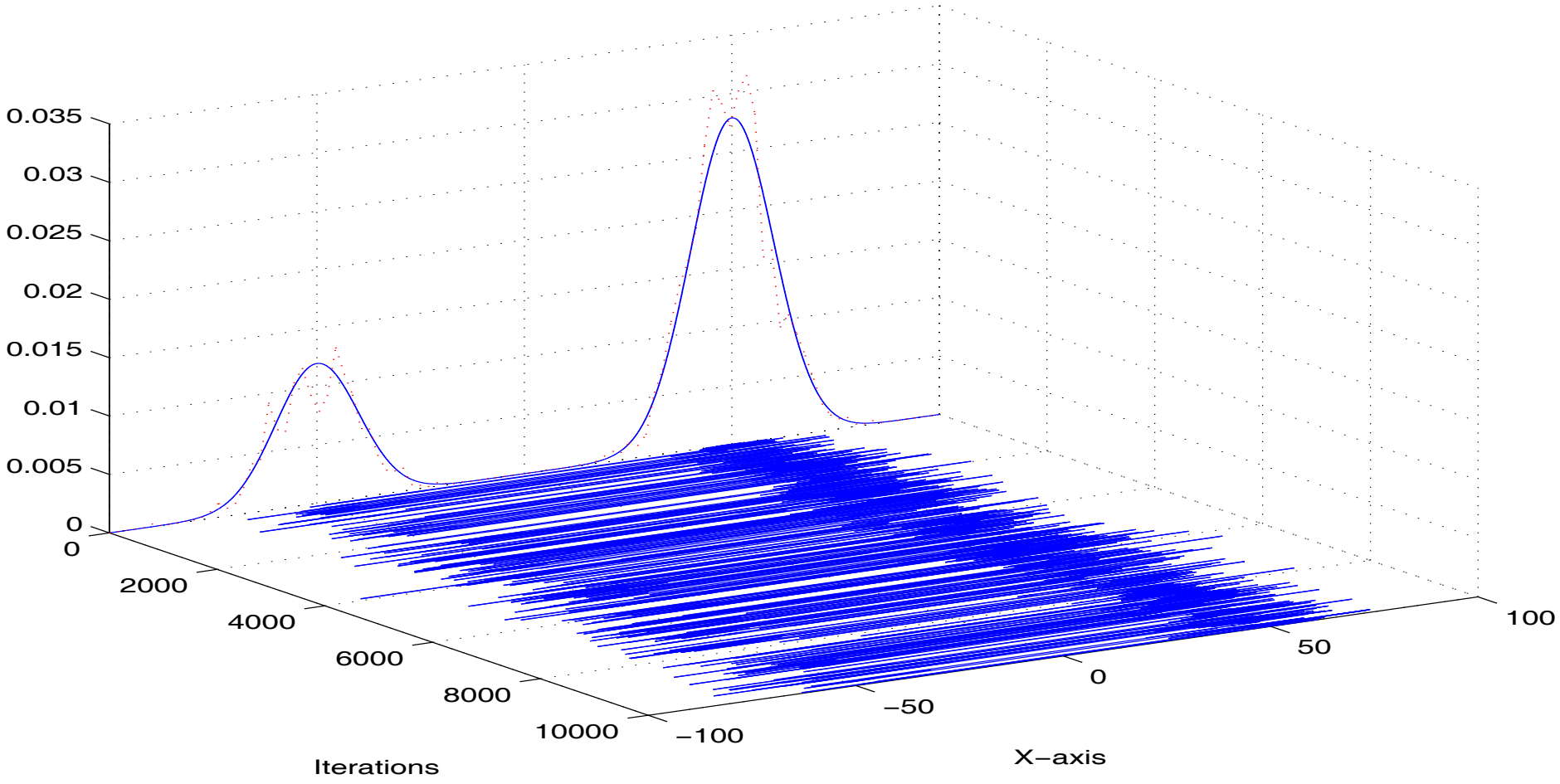
2.12– Example





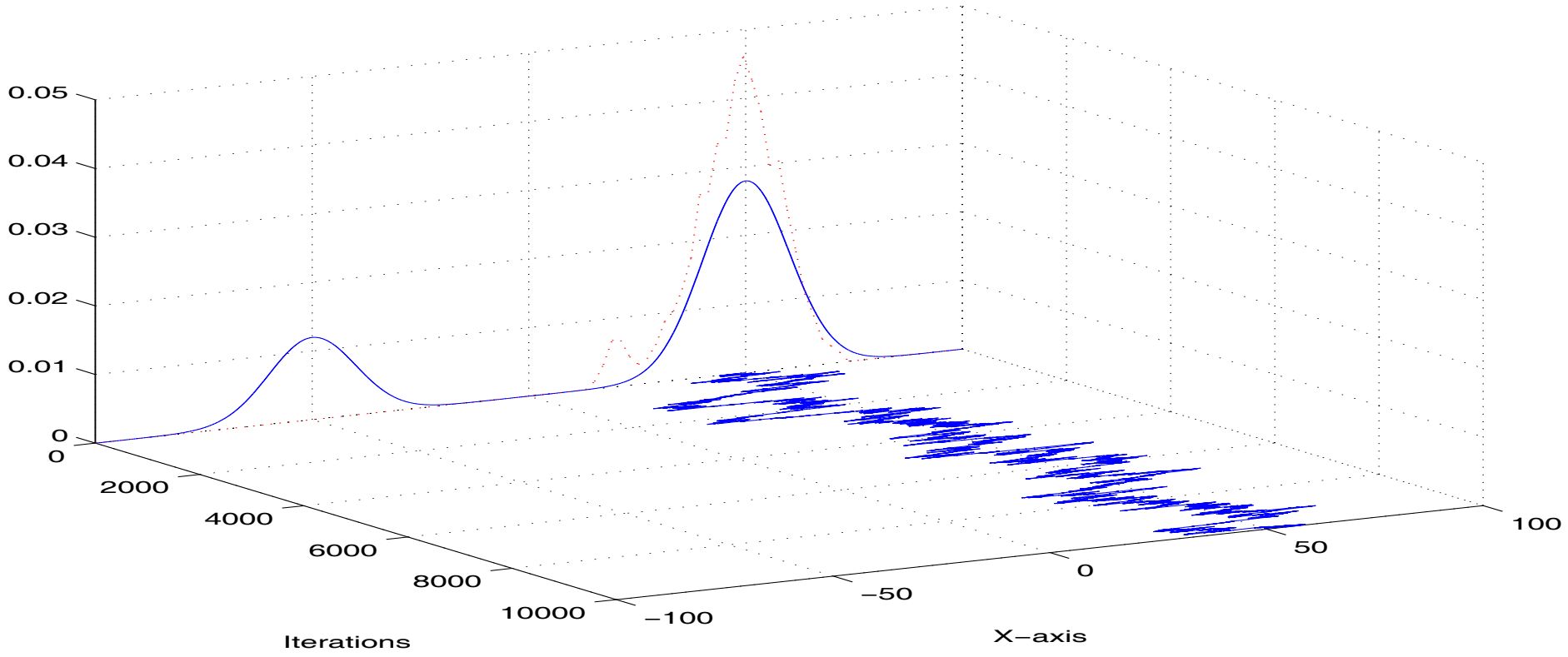
MCMC output for $q_3(\theta, \theta') \propto \exp\left(-\frac{(\theta'-\theta)^2}{2(0.02)^2}\right)$, we estimate $E(\theta) = 0.10$ and $var(\theta) = 0.92$

2.13– A Bimodal Example



Exploration of a bimodal distribution using a random walk MH algorithm

2.13– A Bimodal Example



Bad exploration of a bimodal distribution using a random walk MH algorithm; the variance of the random walk increments is too small.

2.14– Comments

• Heavy tails increments can prevent you from getting trapped in modes.

• It is tempting to adapt the variance of the increments given the simulation output... Unfortunately this breaks the Markov property and biases results if one is not careful.

3.1– Mixture of proposals

• In practice, random walk proposals can be used to explore locally the space whereas independent walk proposals can be used to jump into the space.

• So a good strategy can be to use a proposal distribution of the form

$$q(\theta, \theta') = \lambda q_1(\theta') + (1 - \lambda) q_2(\theta, \theta')$$

where $0 < \lambda < 1$.

• This algorithm is definitely valid as it is just a particular case of the MH algorithm.

• An alternative achieving the same purpose is to use a transition kernel

$$K(\theta, \theta') = \lambda K_1(\theta, \theta') + (1-\lambda) K_2(\theta, \theta')$$

where K_1 (resp. K_2) is an MH algorithm of proposal q_1 (resp. q_2).

• This algorithm is different from using $q(\theta, \theta') = \lambda q_1(\theta') + (1-\lambda) q_2(\theta, \theta')$. It is computationally cheaper and still valid as

$$\begin{aligned}
\int \pi(\theta) K(\theta, \theta') d\theta &= \lambda \int \pi(\theta) K_1(\theta, \theta') d\theta + (1-\lambda) \int \pi(\theta) K_2(\theta, \theta') d\theta \\
&= \lambda \pi(\theta') + (1-\lambda) \pi(\theta') \\
&= \pi(\theta')
\end{aligned}$$

3.3– Using gradient information to build the proposal

- We usually wants to sample candidates in regions of high probability masses.

- We can use

$$\theta' = \theta + \frac{\sigma^2}{2}\nabla \log \pi(\theta) + \sigma V \text{ where } V \sim \mathcal{N}(0,1)$$

where σ^2 is selected such that the acceptance ratio is approximately 0.57.

- The motivation is that, we know that in continuous-time

$$d\theta_t = \frac{1}{2}\nabla \log \pi(\theta) + \sigma dW_t$$

admits π has an invariant distribution.

3.4– Local optimization

• To build $q(\theta, \theta')$, you can use complex deterministic strategies.
Assume you are in θ and you want to propose

$$\theta' \sim \mathcal{N}\left(\varphi(\theta), \sigma^2\right).$$

• You do not need to have an explicit form for the mapping φ!
As long as φ is a *deterministic* mapping, then it is fine. For example $\varphi(\theta)$ could be the local maximum of π closest to θ that has been determined using a gradient algorithm.

• To compute the acceptance probability of the candidate θ', you will need to compute $\varphi(\theta')$ and then you can compute the MH acceptance ratio.

3.5– Alternative acceptance probabilities

- The standard MH algorithm uses the acceptance probability

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\, q(\theta', \theta)}{\pi(\theta)\, q(\theta, \theta')}\right).$$

- This is not necessary and one can also use any function

$$\alpha(\theta, \theta') = \frac{\delta(\theta, \theta')}{\pi(\theta)\, q(\theta, \theta')}$$

which is such that

$$\delta(\theta, \theta') = \delta(\theta', \theta) \text{ and } 0 \leq \alpha(\theta, \theta') \leq 1$$

- Example (Baker, 1965):

$$\alpha(\theta, \theta') = \frac{\pi(\theta')\, q(\theta', \theta)}{\pi(\theta')\, q(\theta', \theta) + \pi(\theta)\, q(\theta, \theta')}.$$

- Indeed one can check that

$$K(\theta,\theta') = \alpha(\theta,\theta')\, q(\theta,\theta') + \left(1 - \int \alpha(\theta,u)\, q(\theta,u)\, du\right) \delta_\theta(\theta')$$

is π-reversible.

- We have

$$\begin{aligned}
\pi(\theta)\,\alpha(\theta,\theta')\,q(\theta,\theta') &= \pi(\theta)\,\frac{\delta(\theta,\theta')}{\pi(\theta)\,q(\theta,\theta')}\,q(\theta,\theta') \\
&= \delta(\theta,\theta') \\
&= \delta(\theta',\theta) \\
&= \pi(\theta')\,\alpha(\theta',\theta)\,q(\theta',\theta)\,.
\end{aligned}$$

- The MH acceptance is favoured as it increases the acceptance probability.

3.5– Alternative acceptance probabilities

• The MH algorithm is a simple and very general algorithm to sample from a target distribution $\pi(\theta)$.

• In practice, the choice of the proposal distribution is absolutely crucial on the performance of the algorithm.

• In high dimensional problems, a simple MH algorithm will be useless. It will be necessary to use a combination of MH kernels.