Computational statistics, lecture 4

Frank Miller, Department of Computer and Information Science, Linköping University

November 21, 2023

LINKÖPING UNIVERSITY

Markov chain Monte Carlo (MCMC), see GH 7.1, 7.3

- The algorithms considered so far generate sequences of **independent** observations which follow the target distribution exactly
- We will now consider a method which generates a sequence of **dependent** observations which follow the target distribution **approximately**

- The next observation (t+1) will be generated based on a proposal distribution g which depends on the current observation (t), i.e. $g(\cdot|X^{(t)})$
- Since $X^{(t+1)}$ depends on $X^{(t)}$ but not on earlier observations, the sequence $(X^{(t)})$ is a Markov chain

MCMC – Metropolis-Hastings algorithm

- A general method to generate the Markov chain is the Metropolis-Hastings (MH) algorithm
- A starting value $x^{(0)}$ is generated from some starting distribution
- Given observation $x^{(t)}$, generate $x^{(t+1)}$ as follows:

1. Sample a candidate x^* from a proposal distribution $g(\cdot|x^{(t)})$
2. Compute the MH ratio $R(x^{(t)}, x^*) = \frac{f(x^*)\, g(x^{(t)}|x^*)}{f(x^{(t)})\, g(x^*|x^{(t)})}$
3. Sample $x^{(t+1)}$ according to

$$x^{(t+1)} = \begin{cases} x^*, \text{with probability } \min\{R(x^{(t)}, x^*), 1\} \\ x^{(t)}, \text{otherwise} \end{cases}$$

4. If more observations needed, set `t <- t+1`; go to 1

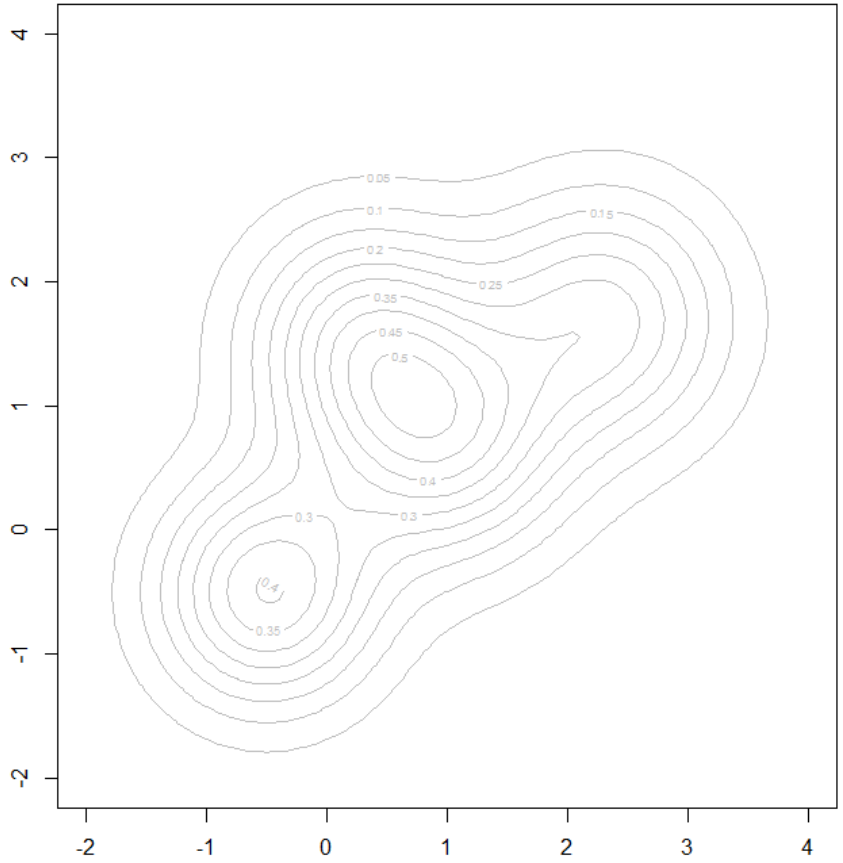
Metropolis algorithm
Special case when g is symmetric:
$g(x^*|x^{(t)}) = g(x^{(t)}|x^*)$

$= \frac{f(x^*)}{f(x^{(t)})}$

Metropolis alg. – Ex.1

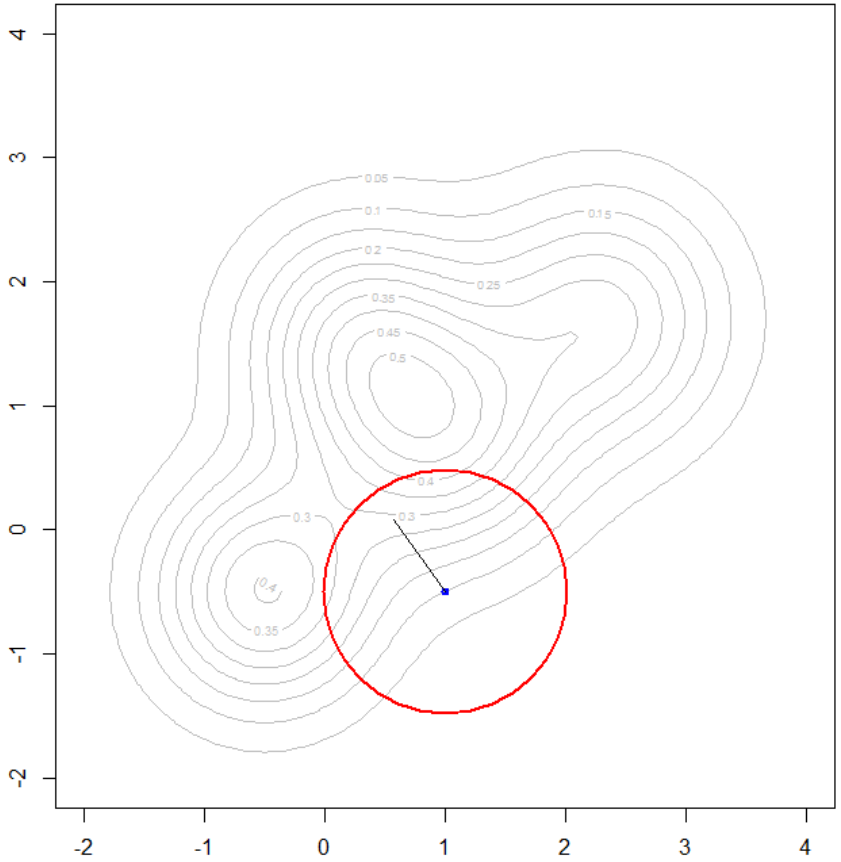
- For illustration, we consider two-dimensional distribution with density f according to contour lines in figure

- Proposal distribution
  $g(x^*|x^{(t)}) = g(x^{(t)}|x^*)$
  $= \frac{1}{\pi r^2}\mathbf{1}\{\|x^{(t)} - x^*\| < r\}$
  for some constant r (here=1)



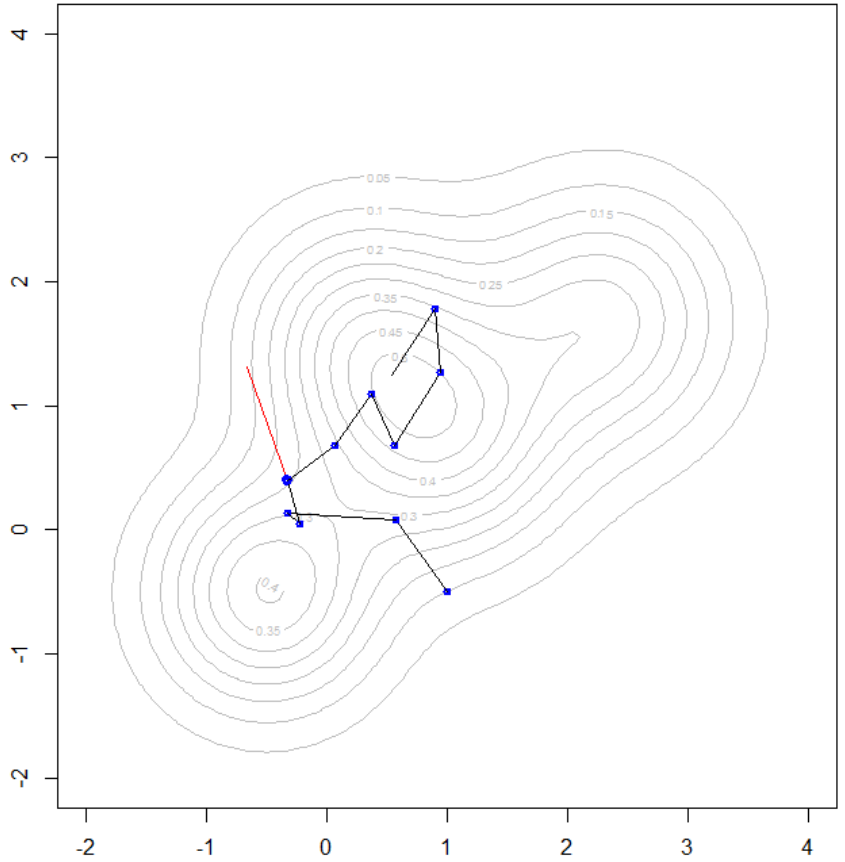
Metropolis alg. – Ex.1

- Proposal distribution
  $g(x^*|x^{(t)}) = g(x^{(t)}|x^*)$
  $= \frac{1}{\pi r^2}\mathbf{1}\{\|x^{(t)} - x^*\| < r\}$
  for some constant r (here=1)
- Start here with $x^{(0)}=(1,-0.5)$
- Randomize uniformly on unit circle around $x^{(0)}$ (proposal distribution); result $x^*=(0.58,0.08)$



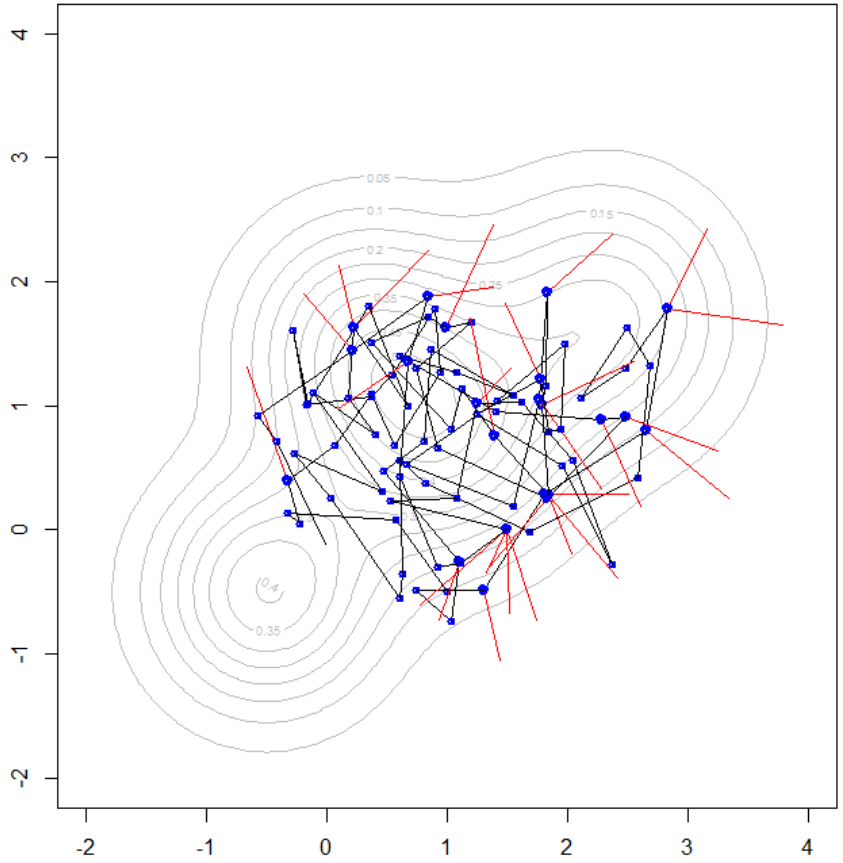
- $f(x^*)=0.296 > f(x^{(0)}) = 0.098$; so this was an uphill step and is automatically accepted ($R\left(x^{(t)}, x^*\right) = \frac{f(x^*)}{f\left(x^{(t)}\right)} > 1$)

Metropolis alg. – Ex.1



- $x^{(0)}=(1,-0.5)$
- Uphill steps: $x^{(1)}=(0.58,0.08)$
- $x^{(2)}=(-0.33,0.13)$
- $x^{(3)}=(-0.23,0.05)$
- Then downhill step proposed: $x^{*}=(-0.32,0.4)$, $R\left(x^{(t)},x^{*}\right)=\frac{f(x^{*})}{f\left(x^{(t)}\right)}=0.774$
- Random Unif(0,1) generated: 0.573 and since this is smaller than R=0.774, $x^{(4)}=x^{*}=(-0.32,0.4)$ is accepted
- Again downhill step proposed: $x^{*}=(-0.67,1.31)$, $R\left(x^{(t)},x^{*}\right)=\frac{f(x^{*})}{f\left(x^{(t)}\right)}=0.560$; random Unif(0,1): 0.890 and rejection of x^{*}
- $x^{(5)}=x^{(4)}=(-0.32,0.4)$

Metropolis alg. – Ex.1

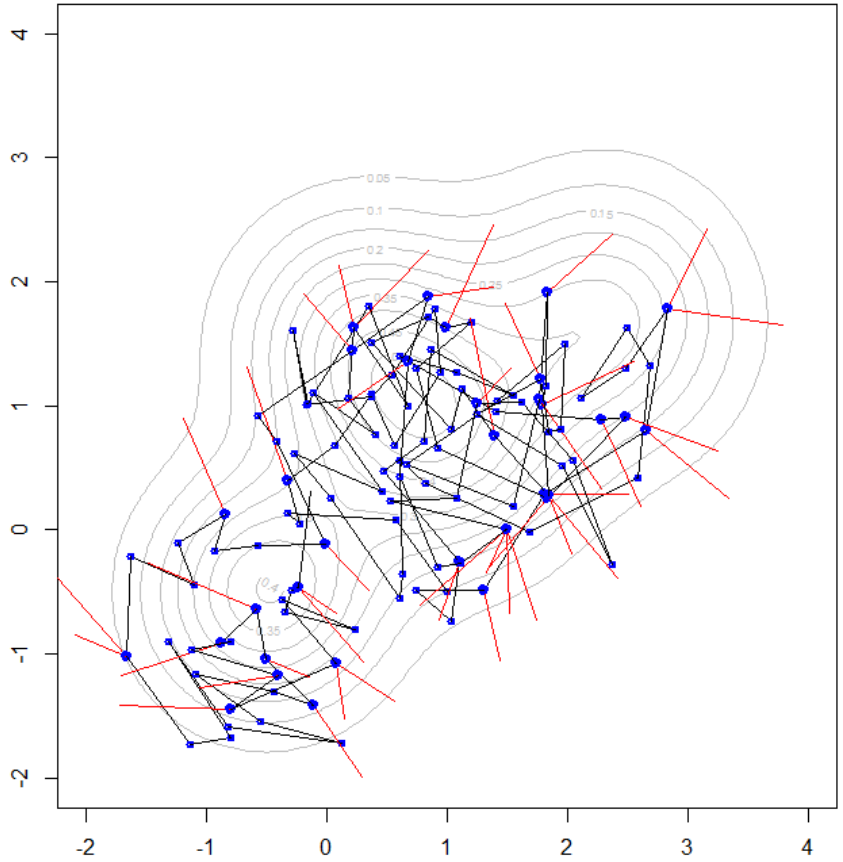


- After several additional iterations (see red lines for rejected proposals), one part of the distribution was explored to a good extend

- Since uphill steps preferred, part of distribution with local maximum at (-0.5,-0.5) is not yet ”detected” at all
- Occasionally, the path will arrive at this part as well

Metropolis alg. – Ex.1

- Now, larger parts of distribution explored



- A couple of animations can be found on: https://chi-feng.github.io/mcmc-demo/app.html#RandomWalkMH,standard (choose Algorithm: RandomWalkMH)

Convergence of Metropolis-Hastings

- If Metropolis-Hastings generated sequence ($X^{(t)}$) is an **irreducible and aperiodic chain** (compare Lecture LM2), the distribution of ($X^{(t)}$) **converges to target distribution**

- For example, if target distribution is uniform distribution on intervals [0, ½] and [3/2, 2], and proposal distribution is uniform distribution on $[X^{(t)}-½, X^{(t)}+½]$, the requirements above are violated

Bayesian analysis

- Data y is collected and assumed that it is generated according to a distribution with density $f(y|\theta)$; θ is a parameter(-vector) to be estimated
- The posterior density is proportional to product of likelihood and prior:

$$f_{\text{posterior}}(\theta|y) = \frac{f(y|\theta) \cdot f_{\text{prior}}(\theta)}{f(y)} \text{ where } f(y) = \int f(y|\theta) f_{\text{prior}}(\theta) d\theta$$

- We would like to generate the posterior distribution $f_{\text{posterior}}$
- We have likelihood $f(y|\theta)$ and an assumption for prior $f_{\text{prior}}(\theta)$
- For Metropolis-Hastings, we do not need the denominator $f(y)$; it cancels out in the MH ratio (see algorithm):

$$R\left(x^{(t)}, x^*\right) = \frac{f_{posterior}(x^*)\, g\left(x^{(t)} \middle| x^*\right)}{f_{posterior}(x^{(t)})\, g\left(x^* \middle| x^{(t)}\right)}$$

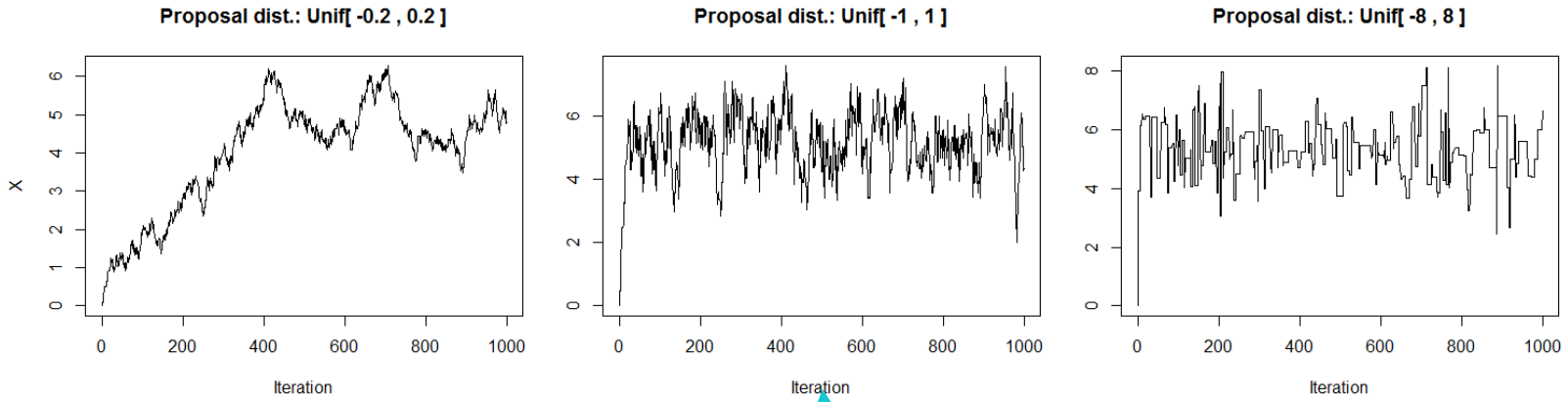
Metropolis algorithm - Example 2

(compare Givens and Hoeting, ex. 5.3)

- Consider Bayesian estimation of μ based on $N(\mu,3^2/7)$ likelihood for μ and Cauchy(5,2) prior; observed mean=5.38
- The posterior density is proportional to product of likelihood and prior
- **Use MCMC to generate random samples following the posterior density**
- Based on these random samples, one can e.g.
  - determine posterior probability that $2 \leq \mu \leq 8$
  - determine mean and variance of posterior

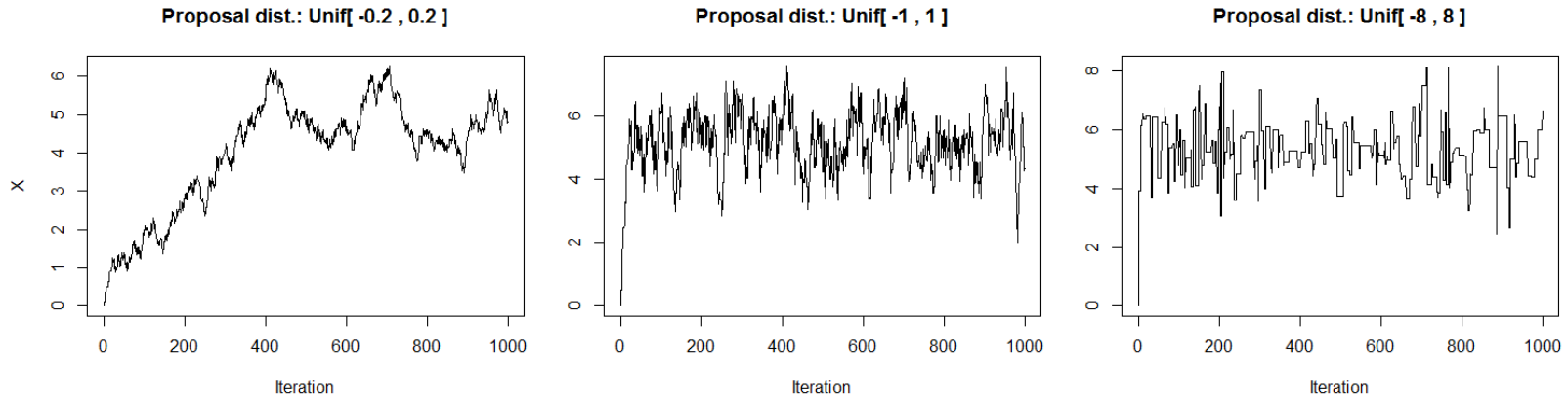
Metropolis algorithm - Example 2

- We use starting value $x^{(0)}=0$, s=1000 iterations and following proposal distributions $g(\cdot|x^{(t)})$:
  $x^{(t)}$+Unif[-0.2,0.2], $x^{(t)}$+Unif[-1,1], $x^{(t)}$+Unif[-8,8]
- **Sample path plots** show simulated values $x^{(t)}$ vs. iteration number t



Best choice

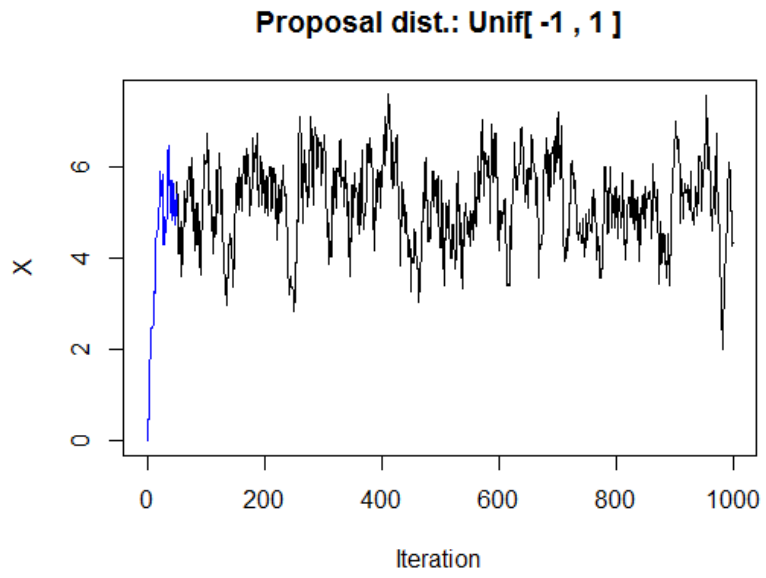
Metropolis algorithm - Example 2

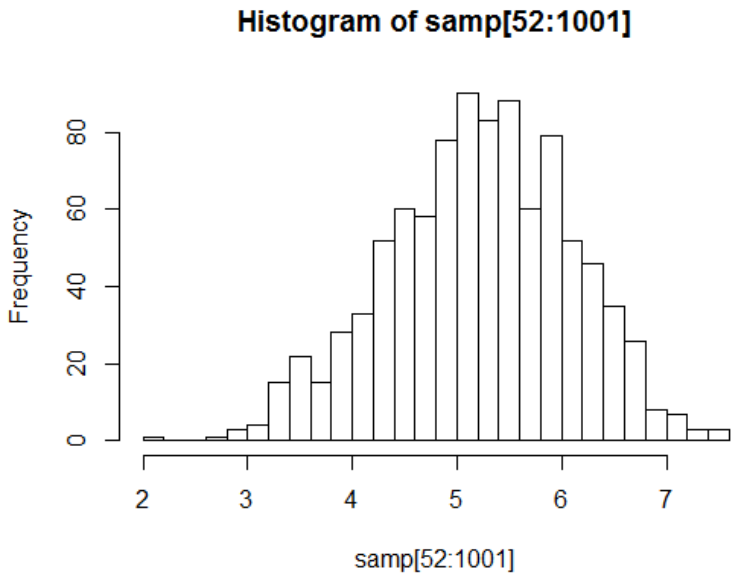


- Count “acceptance rate” (=proportion accepted proposals)
- Here: 98% 78% 18%
- Best results for 44% (uni-dim. case) to 23.4% (high dim. case) acceptance probability (theory based on normal target and proposal functions, see Givens and Hoeting, Chapter 7.3, for references about that)
- For multimodal functions lower acceptance probabilities might be good

Metropolis algorithm - Example 2

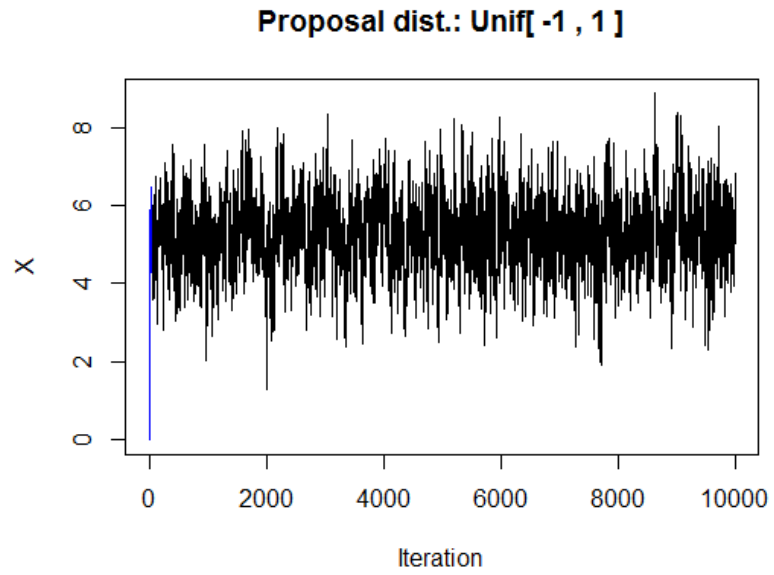
- Based on sample path plots, we might choose Unif[-1,1] as proposal distribution
- Often, one wants to discard initial samples (**burn-in** period) which highly depend on starting value, e.g. 50 values + $x^{(0)}$

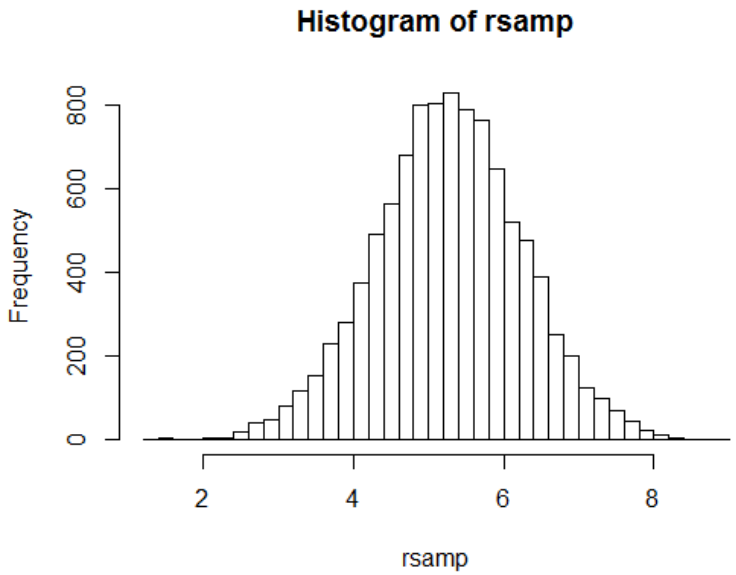




Metropolis algorithm - Example 2

- For s=10 000 iterations and burn-in of 50, we obtain

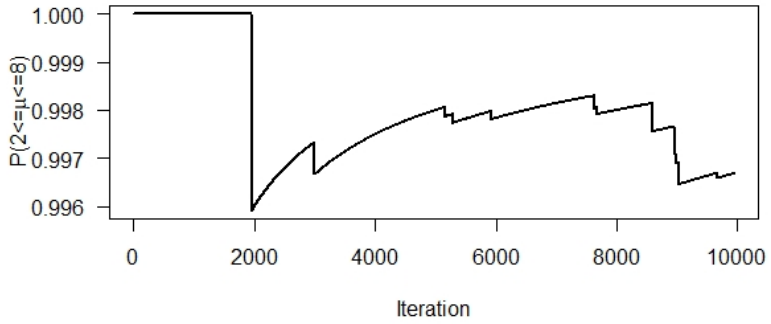




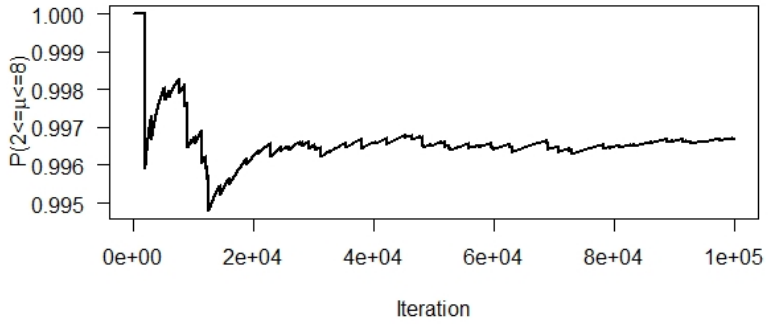
- Monte Carlo estimate for $P(2 \leq \mu \leq 8)$ is 0.9967
  (Monte Carlo standard error=$\sqrt{0.9967 * 0.0033/9950} = 0.0006$)
- Estimated mean = 5.26, standarddeviation =0.99

Metropolis algorithm - Example 2

- Were s=10 000 iterations enough to ensure convergence?
- Can depend on the purpose …
- E.g. for estimating P(2≤μ≤8)
- One can monitor cusum/convergence plots showing estimate versus iterations (see Givens and Hoeting, ch.7.3.1.1)
- After 10 000 iterations



After 100 000 iterations



- After 10 000 iterations, we might not be happy with the left graph; we run longer and are happy with 100 000

Metropolis-Hastings with independent proposals

- Other proposal distributions g possible (not necessarily symmetric), e.g. independent proposals
- Proposal distribution depends not on previous value, $g(\cdot|\mathrm{x}^{(t)}) = g(\cdot)$
- The MH ratio is $R\left(x^{(t)}, x^*\right) = \frac{f(x^*)\, g\left(x^{(t)}|x^*\right)}{f\left(x^{(t)}\right) g\left(x^*|x^{(t)}\right)} = \frac{f(x^*)/g(x^*)}{f\left(x^{(t)}\right)/g(x^{(t)})}$

- A possible application is for Bayesian analysis (f is the posterior) with proposal distribution g being the prior distribution
- f/g is then the likelihood

Gibbs sampling

- Situation:
  - We want to sample a multivariate distribution $f(x_1, \dots, x_d)$ and this density is difficult to sample from
  - The conditional distributions for each single dimension i given fixed values $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d$ are easy to sample from
- Gibbs sampling generates a Markov chain converging to distribution f in this situation
- We sample one dimension in turn

Gibbs sampling

- The algorithm:

0 A starting value $\mathrm{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})$ is chosen

1. Generate $x_1^{(t+1)}$ following $f(x_1 | x_2^{(t)}, \ldots, x_d^{(t)})$
2. Generate $x_2^{(t+1)}$ following $f\left(x_2 \middle| x_1^{(t+1)}, x_3^{(t)}, \ldots, x_d^{(t)}\right)$

…

i. Generate $x_i^{(t+1)}$ following $f(x_i | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_d^{(t)})$

…

d. Generate $x_i^{(t+1)}$ following $f(x_d | x_1^{(t+1)}, \ldots, x_{d-1}^{(t+1)})$

- Go back to 1. if more points needed

Gibbs sampling

- Example:
  Let $f(x_1, x_2) = c \cdot \mathbf{1}\{x_1^2 + 1.8 \cdot x_1 x_2 + x_2^2 < 1\}$
  be the uniform distribution on the ellipse
  $$x_1^2 + 1.8 \cdot x_1 x_2 + x_2^2 < 1$$
- The conditional distribution for x_2 given x_1 is a uniform distribution on the interval
  $(-0.9x_1 - \sqrt{1 - 0.19x_1^2}, -0.9x_1 + \sqrt{1 - 0.19x_1^2})$
  provided that term below root is positive

You can obtain these boundaries by solving $x_1^2 + 1.8 \cdot x_1 x_2 + x_2^2 - 1 = 0$ for x_2

- x_1 given x_2 has a similar distribution with x_2 instead of x_1 in the boundaries