Introduction to Markov chain Monte Carlo — with examples from Bayesian statistics

First winter school in eScience
Geilo, Wednesday January 31st 2007

Håkon Tjelmeland
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Introduction

- Mixed audience
  - some with (almost) no knowledge about (Markov chain) Monte Carlo
  - some know a little about (Markov chain) Monte Carlo
  - some have used (Markov chain) Monte Carlo a lot
- Please ask questions/give comments!
- I will discuss topics also discussed by Morten and Laurant
  - Metropolis–Hastings algorithm and Bayesian statistics
  - will use different notation/terminology
- My goal: Everyone should understand
  - allmost all I discuss today
  - much of what I discuss tomorrow
  - the essence of what I talk about on Friday
- You should
  - understand the mathematics
  - get intuition
- The talk will be available on the web next week
- Remember to ask questions: We have time for it

Plan

- **The Markov chain Monte Carlo (MCMC) idea**
- **Some Markov chain theory**
- **Implementation of the MCMC idea**
  - **Metropolis–Hastings algorithm**
- **MCMC strategies**
  - **independent proposals**
  - **random walk proposals**
  - **combination of strategies**
  - **Gibbs sampler**
- **Convergence diagnostics**
  - **trace plots**
  - **autocorrelation functions**
  - **one chain or many chains?**
- **Typical MCMC problems — and some remedies**
  - **high correlation between variables**
  - **multimodality**
  - **different scales**

Plan (cont.)

- **Bayesian statistics — hierarchical modelling**
  - **Bayes (1763) example**
  - **what is a probability?**
  - **Bayesian hierarchical modelling**

- **Examples**
  - **analysis of microarray data**
  - **history matching — petroleum application**

- **More advanced MCMC techniques/ideas**
  - **reversible jump**
  - **adaptive Markov chain Monte Carlo**
  - **mode jumping proposals**
  - **parallelisation of MCMC algorithms**
  - **perfect simulation**

Why (Markov chain) Monte Carlo?

- **Given a probability distribution of interest**

$$\pi(x), x \in \mathbb{R}^N$$

- **Usually this means: have a formula for** $\pi(x)$
- **But normalising constant is often not known**

$$\pi(x) = ch(x)$$

  - **have a formula for** $h(x)$

- **Want to**
  - **want to "understand"** $\pi(x)$
  - **generates realisations from** $\pi(x)$ **and look at them**
  - **compute mean values**

$$\mu_f = \mathbf{E}[f(x)] = \int f(x)\pi(x)\mathbf{d}x$$

- **Note: most things of interest in a stochastic model can be expressed as an expectation**
  - **probabilities**
  - **distributions**

The Monte Carlo idea

- **Probability distribution of interest** $\pi(x), x \in \mathbb{R}^N$
- $\pi(x)$ **is a high dimensional, complex distribution**
- **Analytical calculations on** $\pi(x)$ **is not possible**
- **Monte Carlo idea**
  - **generate iid samples** $x_1, \ldots, x_n$ **from** $\pi(x)$**.**
  - **estimate interesting quantities about** $\pi(x)$

$$\mu_f = \mathbf{E}[f(x)] = \int f(x)\pi(x)\mathbf{d}x$$

$$\widehat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

  - **unbiased estimator**

$$\mathbf{E}[\widehat{\mu}_f] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[f(x_i)] = \frac{1}{n}\sum_{i=1}^{n} \mu_f = \mu_f$$

  - **estimation uncertainty**

$$\mathbf{Var}[\widehat{\mu}_f] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{Var}[f(x_i)] = \frac{\mathbf{Var}[f(x)]}{n}$$

$$\Rightarrow \mathbf{SD}[\widehat{\mu}_f] = \frac{\mathbf{SD}[f(x)]}{\sqrt{n}}$$

The Markov chain Monte Carlo idea

- **Probability distribution of interest:** $\pi(x), x \in \mathbb{R}^N$
- $\pi(x)$ **is a high dimensional, complex distribution**
- **Analytical calculations on** $\pi(x)$ **is not possible**
- **Direct sampling from** $\pi(x)$ **is not possible**
- **Markov chain Monte Carlo idea**
  - **construct a Markov chain,** $\{X_i\}_{i=0}^{\infty}$**, so that**

$$\lim_{i \to \infty} \mathbf{P}(X_i = x) = \pi(x)$$

  - **simulate the Markov chain for many iterations**
  - **for** m **large enough,** $x_m, x_{m+1}, x_{m+2}, \ldots$ **are (essentially) samples from** $\pi(x)$
  - **estimate interesting quantities about** $\pi(x)$

$$\mu_f = \mathbf{E}[f(x)] = \int f(x)\pi(x)\mathbf{d}x$$

$$\widehat{\mu}_f = \frac{1}{n} \sum_{i=m}^{m+n-1} f(x_i)$$

  - **unbiased estimator**

$$\mathbf{E}[\widehat{\mu}_f] = \frac{1}{n} \sum_{i=m}^{m+n-1} \mathbf{E}[f(x_i)] = \frac{1}{n} \sum_{i=m}^{m+n-1} \mu_f = \mu_f$$

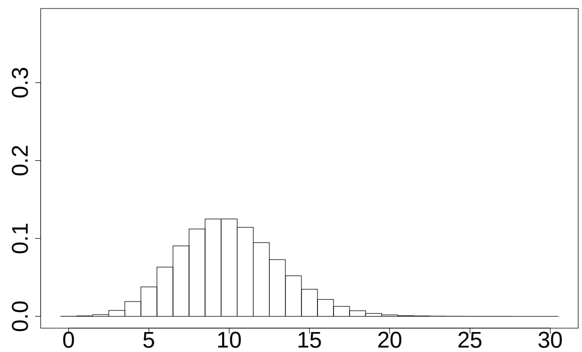
  - **what about the variance?**

A (very) simple MCMC example

- **Note: This is just for illustration, you should never never use MCMC for this distribution!**

- **Let**

$$\pi(x) = \frac{10^x}{x!}e^{-10} \text{ , } \quad x = 0, 1, 2, \ldots$$



- **Set x_0 to 0, 1 or 2 with probability $1/3$ for each**

- **Markov chain kernel**

$$\mathbf{P}(x_{i+1} = x_i - 1 | x_i) = \begin{cases} x_i/20 & \textbf{if } x_i \leq 9, \\ 1/2 & \textbf{if } x_i > 9 \end{cases}$$

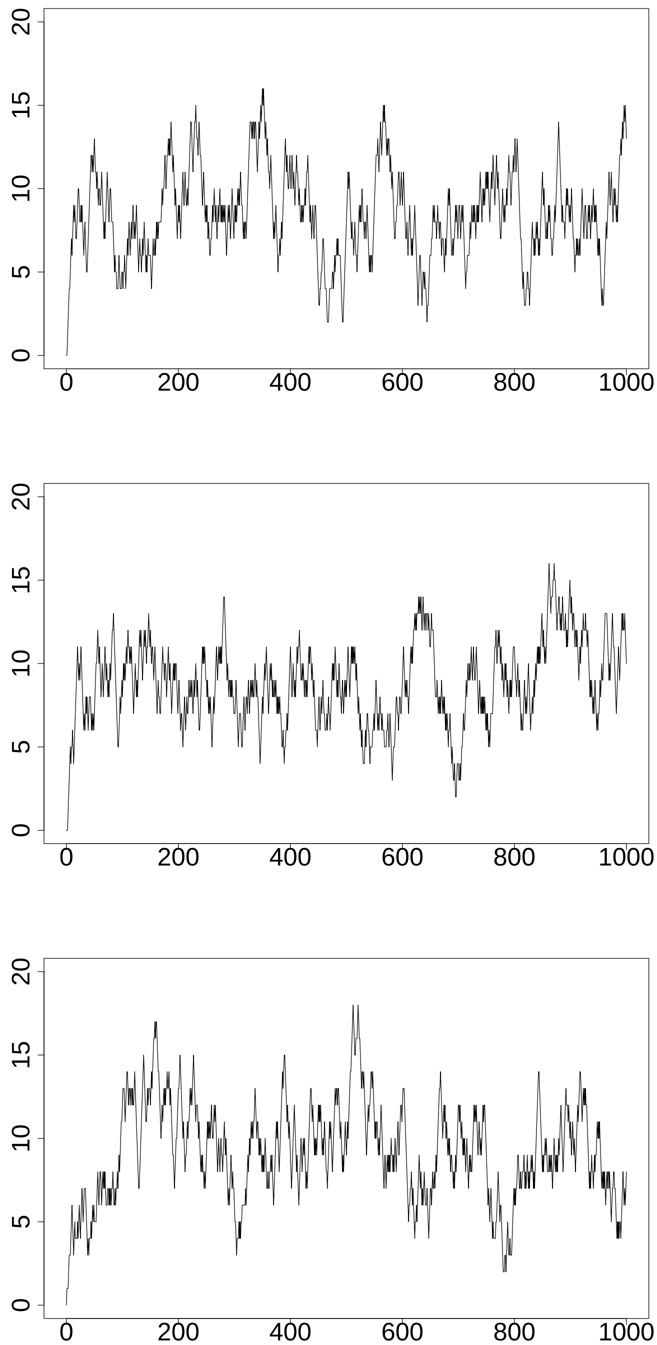
$$\mathbf{P}(x_{i+1} = x_i | x_i) = \begin{cases} (10 - x_i)/20 & \textbf{if } x_i \leq 9, \\ (x_i - 9)/(2(x_i + 1)) & \textbf{if } x_i > 9 \end{cases}$$

$$\mathbf{P}(x_{i+1} = x_i + 1 | x_i) = \begin{cases} 1/2 & \textbf{if } x_i \leq 9, \\ 5/(x_i + 1) & \textbf{if } x_i > 9 \end{cases}$$

- **This Markov chain has limiting distribution $\pi(x)$**
  - **will explain why later**

A (very) simple MCMC example (cont.)

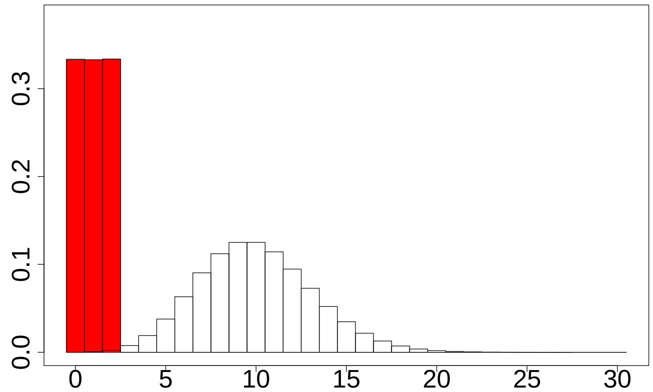
- **Trace plots of three runs**



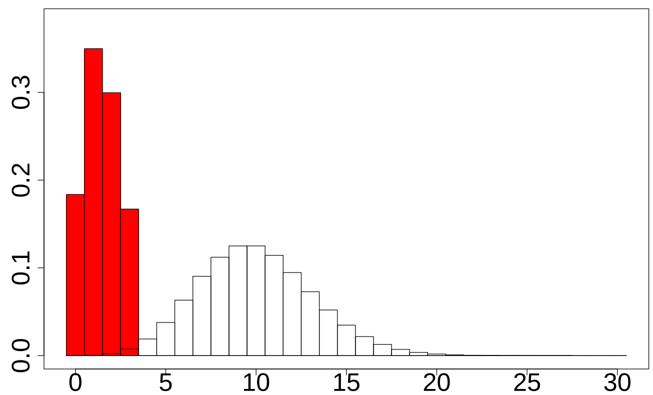
A (very) simple MCMC example (cont.)

- **Convergence to the target distribution**

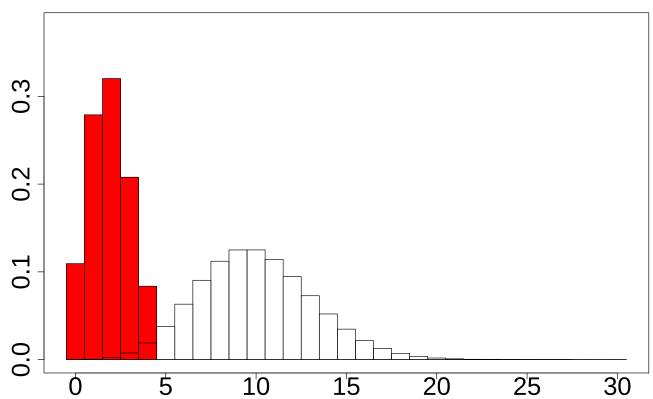
0 **iterations**



1 **iterations**



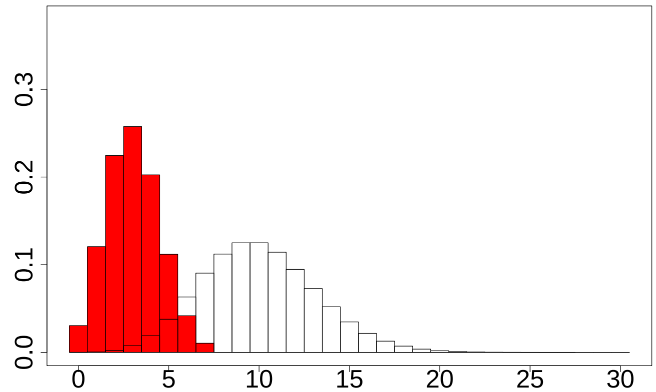
2 **iterations**



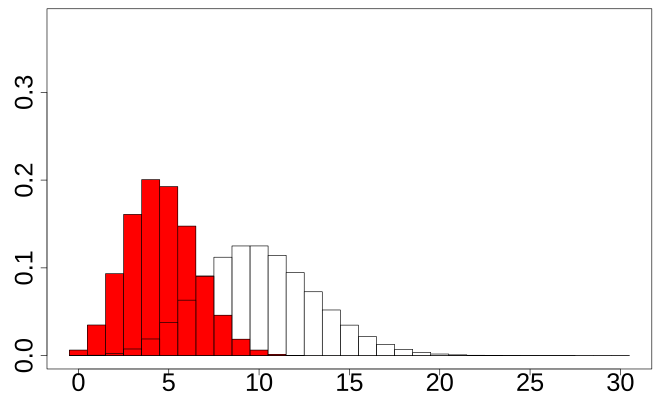
A (very) simple MCMC example (cont.)

- **Convergence to the target distribution**

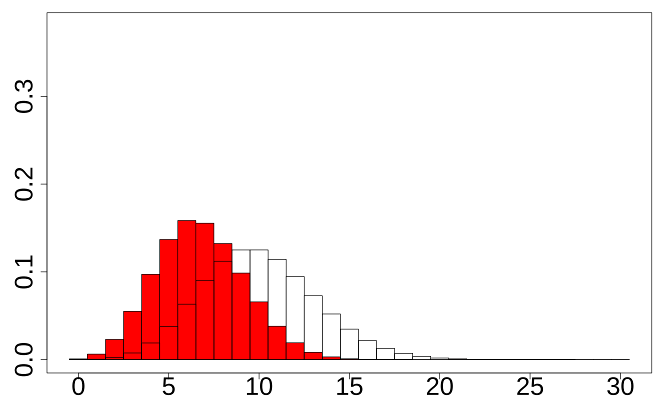
5 iterations



10 iterations



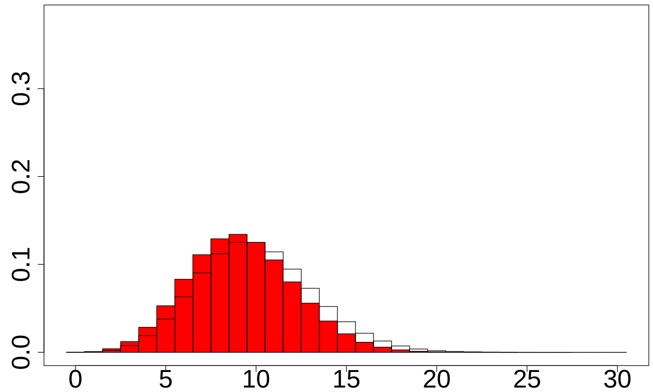
20 iterations



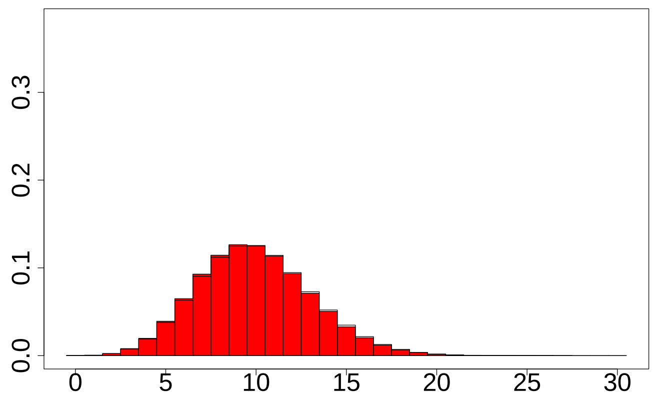
A (very) simple MCMC example (cont.)

- **Convergence to the target distribution**

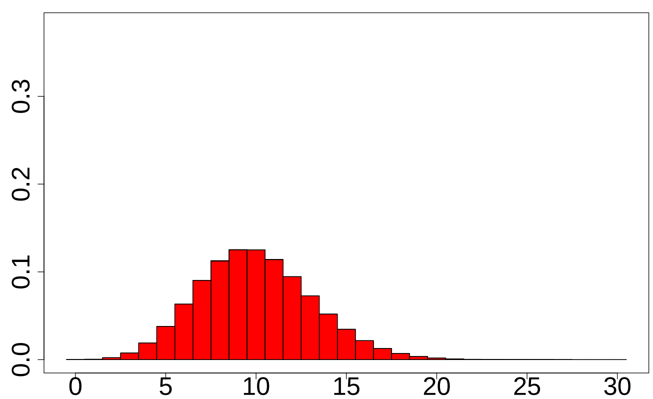
50 iterations



100 iterations



150 iterations



Markov chain Monte Carlo

- **Note:**
  - **the chain $x_0, x_1, x_2, \ldots$ is not converging!**
  - **the distribution $\mathbf{P}(X_n = x)$ is converging**
  - **we simulate/observe only the chain $x_0, x_1, x_2, \ldots$**

- **Need a (general) way to construct a Markov chain for a given target distribution $\pi(x)$.**

- **To simulate the Markov chain must be easy (or at least possible)**

- **Need to decide when (we think) the chain has converged (well enough)**

Some Markov chain theory

- **A Markov chain ($x \in \Omega$ discrete) is a discrete time stochastic process $\{X_i\}_{i=0}^{\infty}, x_i \in \Omega$ which fulfils the *Markov assumption***

$$\mathbf{P}\{X_{i+1} = x_{i+1} | X_0 = x_0, \ldots, X_i = x_i\} = \mathbf{P}\{X_{i+1} = x_{i+1} | X_i = x_i\}$$

- **Thus: a Markov chain can be specified by**
  - **the initial distribution $\mathbf{P}\{X_0 = x_0\} = g(x_0)$**
  - **the transition kernel/matrix**

$$\mathbf{P}(y|x) = \mathbf{P}(X_{i+1} = y | X_i = x)$$

- **Different notations are used**

$$\mathbf{P}_{ij} \qquad \mathbf{P}_{xy} \qquad \mathbf{P}(x, y) \qquad \mathbf{P}(y|x)$$

Some Markov chain theory

- **A Markov chain ($x \in \Omega$ discrete) is defined by**
  - **initial distribution:** $f(x_0)$
  - **transition kernel:** $\mathbf{P}(y|x)$**, note:** $\sum_{y \in \Omega} P(y|x) = 1$
- **Unique limiting distribution** $\pi(x) = \lim_{i \to \infty} f(x_i)$ **if**
  - **chain is irreducible, aperiodic and positive recurrent**
  - **if so, we have**

$$\pi(y) = \sum_{x \in \Omega} \pi(x)\mathbf{P}(y|x) \text{ for all } y \in \Omega \tag{1}$$

- **Note: A sufficient condition for (1) is the detailed balance condition**

$$\pi(x)\mathbf{P}(y|x) = \pi(y)\mathbf{P}(x|y) \text{ for all } x, y \in \Omega$$

  - **proof:**

$$\begin{aligned} \sum_{x \in \Omega} \pi(x)\mathbf{P}(y|x) &= \sum_{x \in \Omega} \pi(y)\mathbf{P}(x|y) \\ &= \pi(y) \sum_{x \in \Omega} \mathbf{P}(x|y) = \pi(y) \end{aligned}$$

- **Note:**
  - **in a stochastic modelling setting:** $\mathbf{P}(y|x)$ **is given, want to find** $\pi(x)$
  - **in an MCMC setting:** $\pi(x)$ **is given, need to find a** $\mathbf{P}(y|x)$

Implementation of the MCMC idea

- **Given a (limiting distribution)** $\pi(x), x \in \Omega$
- **Want a transition kernel so that**
$$\pi(y) = \sum_{x \in \Omega} \pi(x)\mathbf{P}(y|x) \text{ for all } y \in \Omega$$
- **Any solutions?**
  - **# of unknowns:** $|\Omega|(|\Omega| - 1)$**;**
  - **# of equations:** $|\Omega| - 1$
- **Difficult to construct** $\mathbf{P}(y|x)$ **from the above**
- **Require the detailed balance condition**
$$\pi(x)\mathbf{P}(y|x) = \pi(y)\mathbf{P}(x|y) \text{ for all } x, y \in \Omega$$
- **Any solutions:**
  - **# of unknowns:** $|\Omega|(|\Omega| - 1)$
  - **# of equations:** $|\Omega|(|\Omega| - 1)/2$
- **Still many solutions**
- **Recall: don't need all solutions, one is enough!**
- **General (and easy) construction strategy for** $\mathbf{P}(y|x)$ **is available → Metropolis–Hastings algorithm**

Metropolis–Hastings algorithm

- **Detailed balance condition**

$$\pi(x)\mathbf{P}(y|x) = \pi(y)\mathbf{P}(x|y) \text{ for all } x, y \in \Omega$$

- **Choose**

$$\mathbf{P}(y|x) = Q(y|x)\alpha(y|x) \text{ for } y \neq x,$$

  where

  – $Q(y|x)$ **is a *proposal* kernel, we can choose this**

  – $\alpha(y|x) \in [0,1]$ **is an *acceptance probability*, need to find a formula for this**

- **Recall: must have**

$$\sum_{y \in \Omega} \mathbf{P}(y|x) = 1 \text{ for all } x \in \Omega$$

  so then

$$\mathbf{P}(x|x) = 1 - \sum_{y \neq x} \mathbf{Q}(y|x)\alpha(y|x)$$

- **Simulation algorithm**

  – **generate initial state** $x_0 \sim f(x_0)$

  – **for** $i = 1, 2, \ldots$

    ∗ **propose potential new state** $y_i \sim \mathbf{Q}(y_i|x_{i-1})$

    ∗ **compute acceptance probability** $\alpha(y_i|x_{i-1})$

    ∗ **draw** $u_i \sim$ **Uniform**$(0,1)$

    ∗ **if** $u_i \leq \alpha(y_i|x_{i-1})$ **accept** y_i**, i.e. set** $x_i = y_i$**, otherwise reject** y_i **and set** $x_i = x_{i-1}$

The acceptance probability

- **Recall: detailed balance condition**

$$\pi(x)\mathbf{P}(y|x) = \pi(y)\mathbf{P}(x|y) \textbf{ for all } x, y \in \Omega$$

  – **Proposal kernel**

$$\mathbf{P}(y|x) = \mathbf{Q}(y|x)\alpha(y|x) \textbf{ for } y \neq x$$

- **Thus, must have**

$$\pi(x)\mathbf{Q}(y|x)\alpha(y|x) = \pi(y)\mathbf{Q}(x|y)\alpha(x|y) \textbf{ for all } x \neq y$$

- **General solution**

$$\alpha(y|x) = r(x, y)\pi(y)\mathbf{Q}(x|y) \textbf{ where } r(x, y) = r(y, x)$$

- **Recall: must have**

$$\alpha(y|x) = r(x, y)\pi(y)\mathbf{Q}(x|y) \leq 1 \Rightarrow r(x, y) \leq \frac{1}{\pi(y)Q(x|y)}$$

$$\alpha(x|y) = r(x, y)\pi(x)\mathbf{Q}(y|x) \leq 1 \Rightarrow r(x, y) \leq \frac{1}{\pi(x)Q(y|x)}$$

- **Choose $r(x, y)$ as large as possible**

$$r(x, y) = \min\left\{\frac{1}{\pi(x)\mathbf{Q}(y|x)}, \frac{1}{\pi(y)\mathbf{Q}(x|y)}\right\}$$

- **Thus**

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)\mathbf{Q}(x|y)}{\pi(x)\mathbf{Q}(y|x)}\right\}$$

Metropolis–Hastings algorithm

- **Recall: For convergence it is sufficient with**
  - **detailed balance**
  - **irreducible**
  - **aperiodic**
  - **positive recurrent**

- **Detailed balance: ok by construction**

- **Irreducible: must be checked in each case**
  - **usually easy**

- **Aperiodic: sufficient that $\mathbf{P}(x|x) > 0$ for one $x \in \Omega$**
  - **for example by $\alpha(y|x) < 1$ for one set $x, y \in \Omega$**

- **Positive recurrent: in discrete state space, irreducibility and finite state space is sufficient**
  - **more difficult in general, but Markov chain drifts if it is not recurrent**
  - **usually not a problem in practice**

Metropolis–Hastings algorithm

- **Building blocks:**
  - **target distribution $\pi(x)$ (given by problem)**
  - **proposal distribution $\mathbf{Q}(y|x)$ (we choose)**
  - **acceptance probability**

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)\mathbf{Q}(x|y)}{\pi(x)\mathbf{Q}(y|x)}\right\}$$

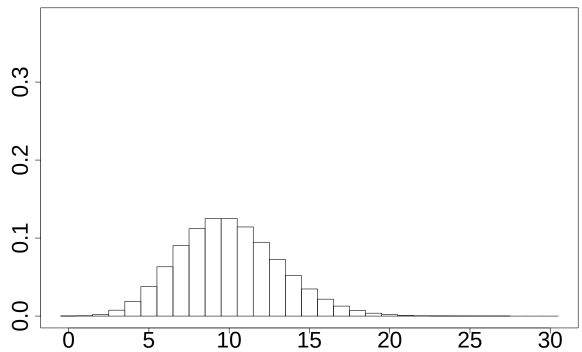
- **Note: unknown normalising constant in $\pi(x)$ ok**

- **A little history**
  - **Metropolis et al. (1953). Equations of state calculations by fast computing machines. J. of Chemical Physics.**
  - **Hastings (1970). Monte Carlo simulation methods using Markov chains and their applications. Biometrika.**
  - **Green (1995). Reversible jump MCMC computation and Bayesian model determination. Biometrika.**

A simple MCMC example (revisited)

- **Let**

$$\pi(x) = \frac{10^x}{x!} e^{-10} \ , \quad x = 0, 1, 2, \ldots$$



- **Proposal distribution**

$$\mathbf{Q}(y|x) = \begin{cases} 1/2 & \textbf{for } y \in \{x-1, x+1\}, \\ 0 & \textbf{otherwise} \end{cases}$$

- **Acceptance probability**

$$y = x - 1 \ : \ \alpha(x-1|x) = \min\left\{1, \frac{\frac{10^{x-1}}{(x-1)!}e^{-10}}{\frac{10^x}{x!}e^{-10}}\right\} = \min\left\{1, \frac{x}{10}\right\}$$

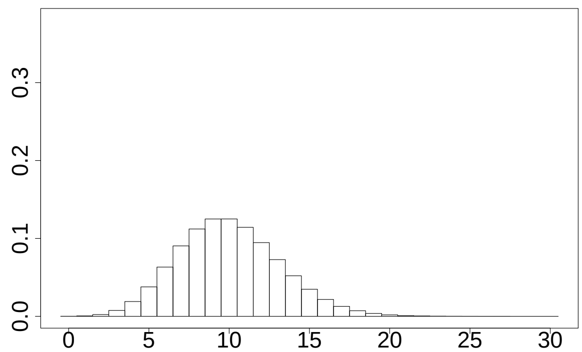
$$y = x + 1 \ : \ \alpha(x+1|x) = \min\left\{1, \frac{\frac{10^{x+1}}{(x+1)!}e^{-10}}{\frac{10^x}{x!}e^{-10}}\right\} = \min\left\{1, \frac{10}{x+1}\right\}$$

- **P**$(y|x)$ **then becomes as specified before**

A (very) simple MCMC example

- **Note: This is just for illustration, you should never use MCMC for this distribution!**
- **Let**

$$\pi(x) = \frac{10^x}{x!} e^{-10} \;\; , \quad x = 0, 1, 2, \ldots$$



- **Set x_0 to 0, 1 or 2 with probability $1/3$ for each**
- **Markov chain kernel**

$$\mathbf{P}(x_{i+1} = x_i - 1 | x_i) \;=\; \begin{cases} x_i/20 & \textbf{if } x_i \le 9, \\ 1/2 & \textbf{if } x_i > 9 \end{cases}$$

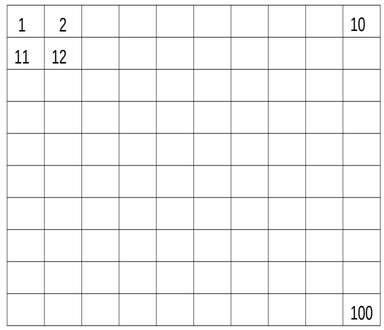
$$\mathbf{P}(x_{i+1} = x_i | x_i) \;=\; \begin{cases} (10 - x_i)/20 & \textbf{if } x_i \le 9, \\ (x_i - 9)/(2(x_i + 1)) & \textbf{if } x_i > 9 \end{cases}$$

$$\mathbf{P}(x_{i+1} = x_i + 1 | x_i) \;=\; \begin{cases} 1/2 & \textbf{if } x_i \le 9, \\ 5/(x_i + 1) & \textbf{if } x_i > 9 \end{cases}$$

- **This Markov chain has limiting distribution $\pi(x)$**
  - **will explain why later**

Another MCMC example — Ising

- **2D rectangular lattice of nodes**
- **Number the nodes from 1 to N**



- $x^i \in \{0, 1\}$**: value (colour) in node i,** $x = (x^1, \ldots, x^N)$
- **First order neighbourhood**

- **Probability distribution**

$$\pi(x) = c \cdot \exp\left\{ -\beta \sum_{i \sim j} I(x^i \neq x^j) \right\}$$

β**: parameter;** c**: normalising constant,**

$$c = \left[\sum_x \exp\left\{ -\beta \sum_{i \sim j} I(x^i \neq x^j) \right\} \right]^{-1}$$

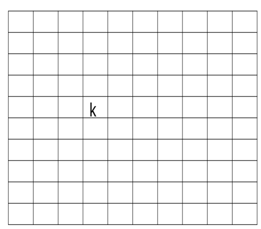
Ising example (cont.)

- **Probability distribution**

$$\pi(x) = c \cdot \exp\left\{ -\beta \sum_{i \sim j} I(x^i \neq x^j) \right\}$$

- **Proposal algorithm**
  - **current state:** $x = (x^1, \ldots, x^N)$
  - **draw a node** $k \in \{1, \ldots, n\}$ **at random**
  - **propose to revers the value of node** k**, i.e.**

$$y = (x^1, \ldots, x^{k-1}, 1 - x^k, x^{k+1}, \ldots, x^N)$$



- **Proposal kernel**

$$Q(y|x) = \begin{cases} \frac{1}{N} & \textbf{if } x \textbf{ and } y \textbf{ differ in (exactly) one node}, \\ 0 & \textbf{otherwise} \end{cases}$$

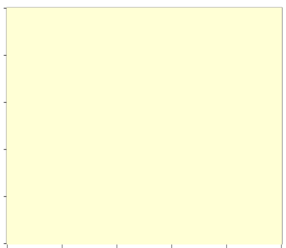
- **Acceptance probability**

$$\begin{aligned} \alpha(y|x) &= \min\left\{ 1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)} \right\} \\ &= \min\left\{ 1, \exp\left\{ -\beta \sum_{j \sim k} \left[I(x^j \neq 1 - x^k) - I(x^j \neq x^k) \right] \right\} \right\} \end{aligned}$$

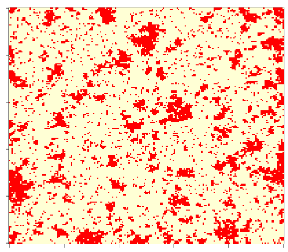
Ising example (cont.)

- $\beta = 0.87$
- $x^0 = 0$

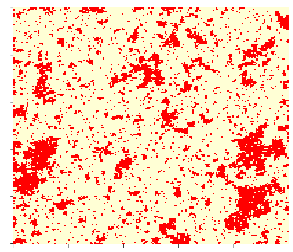
$0n$



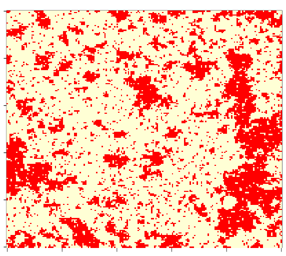
$200n$ **iterations**



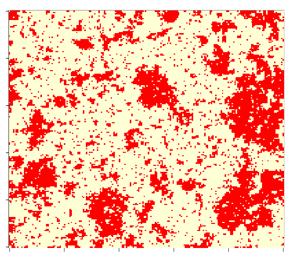
$400n$ **iterations**



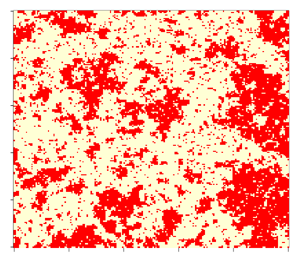
$600n$



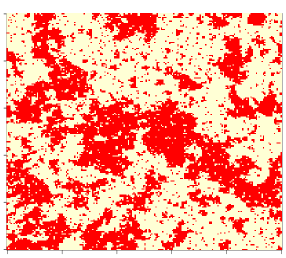
$800n$ **iterations**



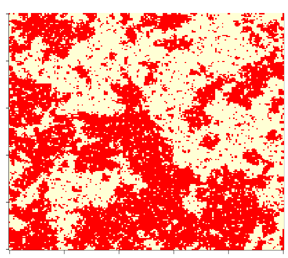
$1000n$ **iterations**



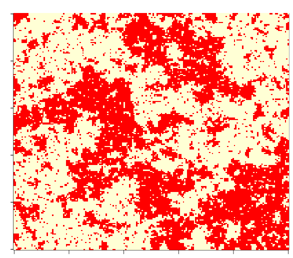
$5000n$



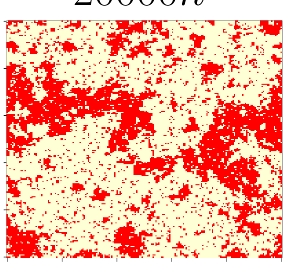
$10000n$ **iterations**



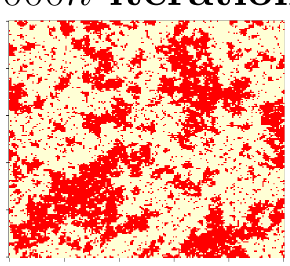
$15000n$ **iterations**



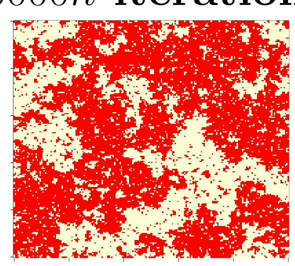
$20000n$



$25000n$ **iterations**

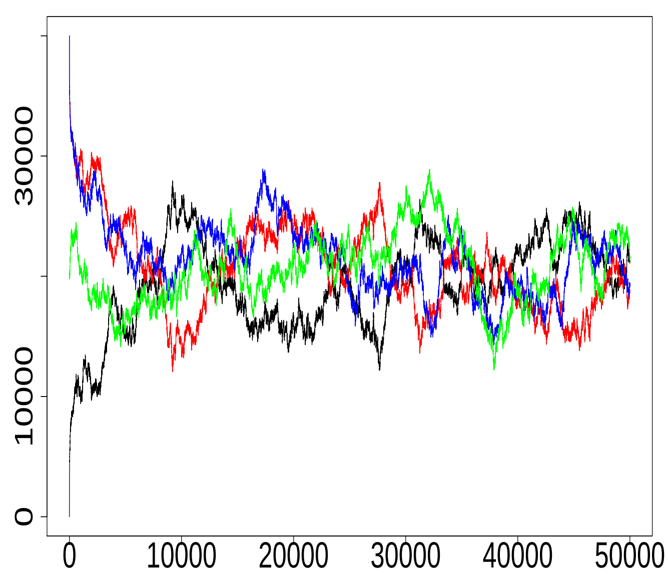


$30000n$ **iterations**



Ising example (cont.)

- trace plot of number of 1's
  - three runs
  - different initial state:
    * all 0's
    * all 1's
    * independent random in each node



Continuous state space

- **Target distribution**
  - **discrete:** $\pi(x), x \in \Omega$
  - **continuous:** $\pi(x), x \in \mathbb{R}^N$

- **Proposal distribution**
  - **discrete:** $Q(y|x)$
  - **continuous:** $Q(y|x)$

- **Acceptance probability**
  - **discrete:** $\alpha(y|x)$
  - **continuous:** $\alpha(y|x)$

$$\alpha(y|x) = \min\left\{1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}\right\}$$

- **Rejection probability**
  - **discrete:**

$$r(x) = 1 - \sum_{y \neq x} Q(y|x)\alpha(y|x)$$

  - **continuous:**

$$r(x) = 1 - \int_{\mathbb{R}^N} Q(y|x)\alpha(y|x)\mathbf{d}y$$

Plan

- The Markov chain Monte Carlo (MCMC) idea
- Some Markov chain theory
- Implementation of the MCMC idea
  - Metropolis–Hastings algorithm
- MCMC strategies
  - independent proposals
  - random walk proposals
  - combination of strategies
  - Gibbs sampler
- Convergence diagnostics
  - trace plots
  - autocorrelation functions
  - one chain or many chains?
- Typical MCMC problems — and some remedies
  - high correlation between variables
  - multimodality
  - different scales