Stochastic Simulation
Markov Chain Monte Carlo

Bo Friis Nielsen

Institute of Mathematical Modelling

Technical University of Denmark

2800 Kgs. Lyngby – Denmark

Email: bfni@dtu.dk

Explanation: What is the problem with the Pareto distribution

- Moment distributions
- For nonnegative valued random variables

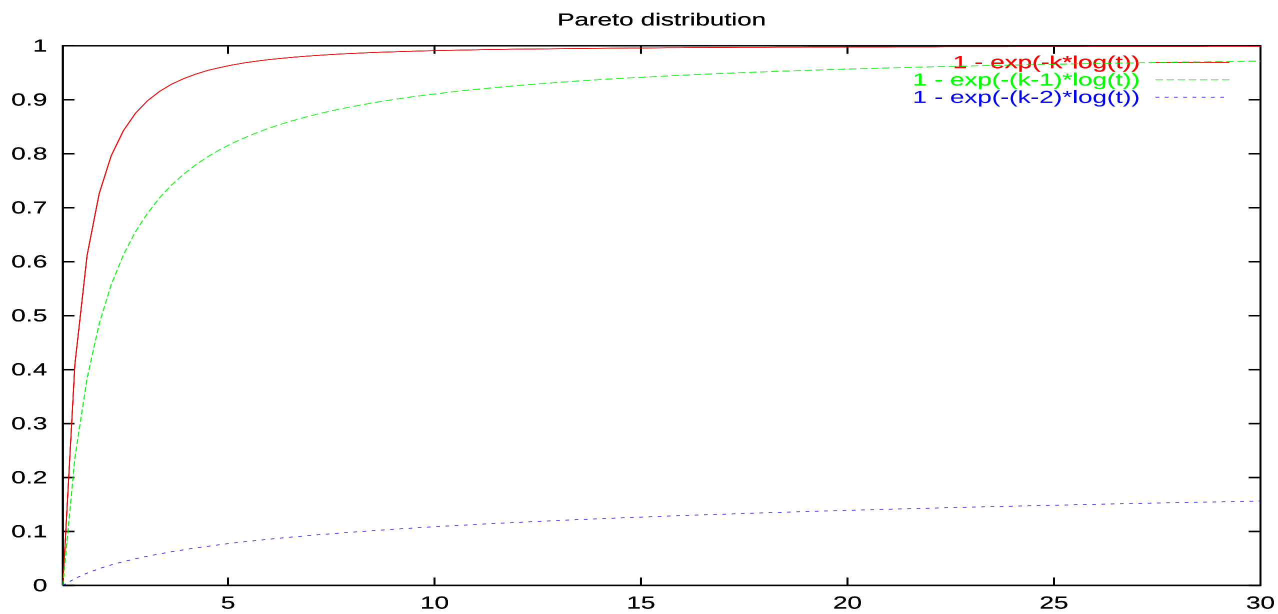
$$G_j(x) = \frac{\int_0^x t^j f(t)dt}{\int_0^\infty t^j f(t)dt} = \frac{\int_0^x t^j f(t)dt}{\mathbb{E}\left(X^j\right)}$$

The contribution to the j'th moment from values $\leq x$.

$$\int_0^x t^1 f(t)dt = \int_\beta^x t\frac{k}{\beta}\left(\frac{t}{\beta}\right)^{-k-1} dt = \int_\beta^x k\left(\frac{t}{\beta}\right)^{-k} dt$$

$$= \beta\frac{k}{k-1}\int_\beta^x \frac{k-1}{\beta}\left(\frac{t}{\beta}\right)^{-k} dt = \frac{\beta k}{k-1}\left[1 - \left(\frac{x}{\beta}\right)^{-k+1}\right]$$

Explanation: What is the problem with the Pareto distribution



- The first moment distribution for the Pareto distribution (green)
- The second moment distribution for the Pareto distribution (blue)

Some numbers $\beta = 1$

$$F(t) = 1 - t^{-k} \qquad f(t) = kt^{-k-1}$$

$$G_1(t) = 1 - t^{-k+1} \qquad G_2(t) = 1 - t^{-k+2}$$

For $k = 2.05$

t	$F(t)$	$G_1(t)$	$G_2(t)$
2	0.7585	0.5170	0.0341
10	0.9911	0.9109	0.1190
100	0.9999	0.9921	0.2057
844.5	$1 - 10^{-6}$	0.9992	0.2860

- Even when if we simulate 10^6 values we can not expect to get a decent estimate of the variance!

What to learn:

- Care is needed when using simulation
- Especially if one wants to study strange or rare phenomena.
- Always use your practical, theoretical and intuitive understanding of the system to support the analysis by simulation.

The queueing example

We simulated the system until “stochastic steady state”.

We were then able to describe this steady state:

- What is the distribution of occupied servers
- What is the rejection probability

The model was a “state machine”, i.e. a Markov Chain.

To obtain steady-state statistics, we used stochastic simulation, i.e. Monte Carlo.

Discrete time Markov chains

- We observe a sequence of X_ns taking values in some sample space
- The Next value in the sequence X_{n+1} is determined from some decision rule depending on the value of X_n only.
- For discrete sample space we can express the decision rule as a matrix of transition probabilities $P = \{p_{ij}\}$, $p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$
- Under some technical assumptions we can find a stationary and limiting distribution $\boldsymbol{\pi}$.$\pi_j = \mathbb{P}(X_\infty = j)$.
- This distribution can be analytically found by solving

$$\boldsymbol{\pi} = \boldsymbol{\pi} P \qquad \text{(equilibrium distribution)}$$

Markov chains continued

- The theory can be extended to:
  - ◇ Continuous sample space or
  - ◇ Continuous time: exercise 4 is an example of a Continuous time Markov chain

The probability of X_n

- The behaviour of the process itself - X_n
- The behaviour conditional on $X_0 = i$ is $(p_{ij}(n))$
- Define $\mathbb{P}(X_n = j) = \mu_j^{(n)}$ with $\mathbb{P}(X_0 = j) = \mu_j^{(0)}$
- with $\vec{\mu}^{(n)} = \{\mu_j^{(n)}\}$ we find

$$\vec{\mu}^{(n)} = \vec{\mu}^{(n-1)} P = \vec{\mu}^{(0)} P_n = \vec{\mu}^{(0)} P^n$$

Small example

$$P = \begin{bmatrix} 1-p & p & 0 & 0 \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ 0 & 0 & q & 1-q \end{bmatrix}$$

with $\vec{\mu}^{(0)} = \left(\frac{1}{3}, 0, 0, \frac{2}{3}\right)$ we get

$$\vec{\mu}^{(1)} = \left(\frac{1}{3}, 0, 0, \frac{2}{3}\right) \begin{bmatrix} 1-p & p & 0 & 0 \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ 0 & 0 & q & 1-q \end{bmatrix} = \left(\frac{1-p}{3}, \frac{p}{3}, \frac{2q}{3}, \frac{2(1-q)}{3}\right)$$

and

$$\vec{\mu}^{(0)} = \left(\frac{1}{3}, 0, 0, \frac{2}{3}\right),$$

$$P^2 = \begin{bmatrix} (1-p)^2 + pq & (1-p)p & p^2 & 0 \\ q(1-p) & 2qp & 0 & p^2 \\ q^2 & 0 & 2qp & p(1-q) \\ 0 & q^2 & (1-q)q & (1-q)^2 + qp \end{bmatrix}$$

$$\vec{\mu}^{(2)} = \left(\frac{1}{3}, 0, 0, \frac{2}{3}\right) \cdot$$

$$\begin{bmatrix} (1-p)^2 + pq & (1-p)p & p^2 & 0 \\ q(1-p) & 2qp & 0 & p^2 \\ q^2 & 0 & 2qp & p(1-q) \\ 0 & q^2 & (1-q)q & (1-q)^2 + qp \end{bmatrix}$$

$$= \left(\frac{(1-p)^2 + pq}{3}, \frac{(1-p)p}{3}, \frac{4qp}{3}, \frac{2p(1-q)}{3}\right)$$

MCMC: What we aim to achieve

We have a variable X with a “complicated” distribution.

We cannot sample X directly.

We aim to generate a sequence of X_i's

- which each has the same distribution as X
- but we allow them to be interdependent.

This is an **inverse problem** relative to the queueing exercise: We start with the distribution of X, and aim to design a state machine which has this steady-state distribution.

MCMC example from Bayesian statistics

Prior distribution of parameter

$$P \sim U(0,1) \qquad : \qquad f_P(p) = \mathbf{1}(0 \leq p \leq 1)$$

Distribution of data, conditional on parameter

$$X for given P = p \text{ is } \mathrm{Binomial}(n, P)$$

i.e. the data has the conditional probabilities

$$\mathbb{P}(X = i|P) = \begin{pmatrix} n \\ i \end{pmatrix} P^i (1-P)^{n-i}$$

The posterior distribution of P

Conditional density of parameter, given observed data $X = i$:

$$f_{P|X=i}(p) = f_P(p)\frac{\mathbf{P}(X = i|P = p)}{\mathbf{P}(X = i)}$$

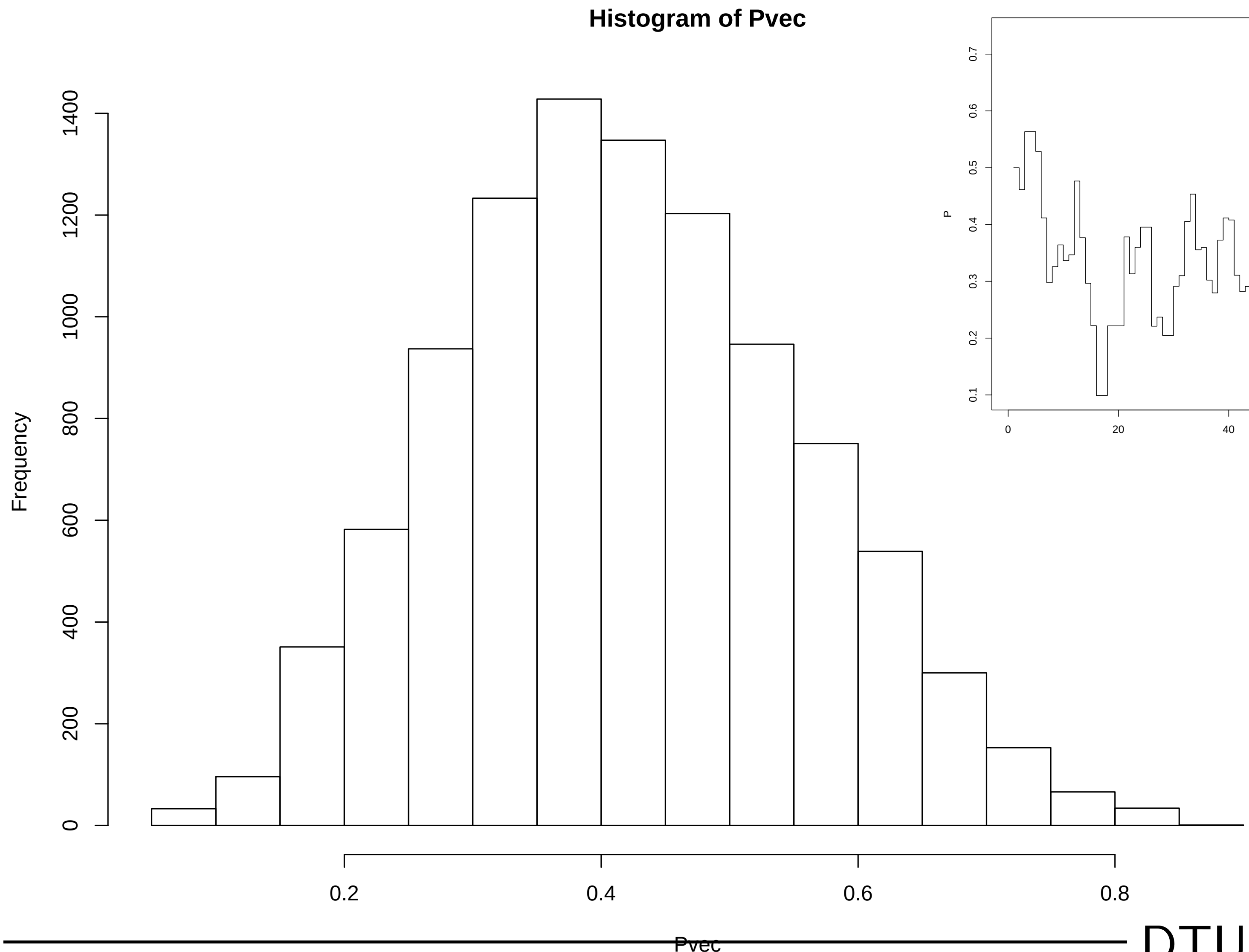
We need the unconditional probability of the observation:

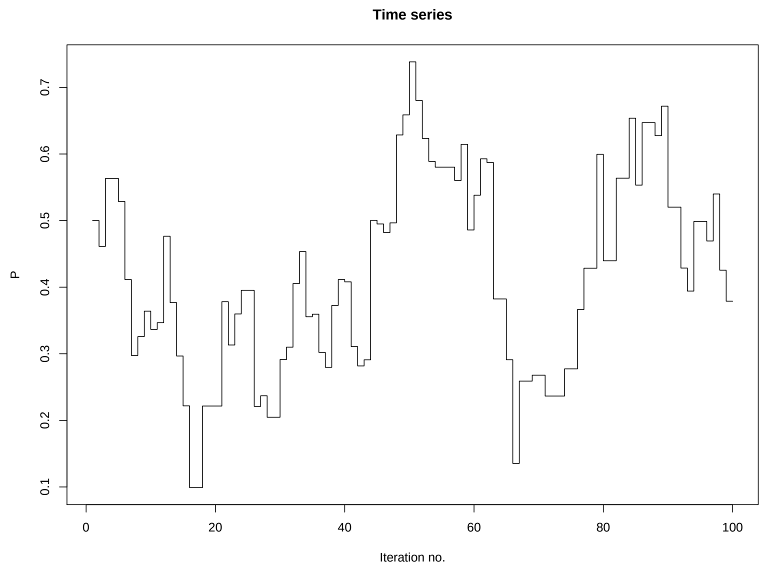
$$\mathbf{P}(X = i) = \int_0^1 f_P(p) \begin{pmatrix} n \\ i \end{pmatrix} p^i(1-p)^{n-i}\, dp$$

We *can* evaluate this; in more complex models we could not.

AIM: To sample from $f_{P|X=i}$, without evaluating $\mathbf{P}(X = i)$.

The posterior distribution





When to apply MCMC?

The distribution is given by

$$f(x) = c \cdot g(x)$$

where the *unnormalized density* g can be evaluated, *but* the normalising constant c cannot be evaluated (easily).

$$c = \frac{1}{\int_{\mathbf{X}} g(x)\, dx}$$

This is frequently the case in Bayesian statistics - the posterior density is proportional to the likelihood function

Note (again) the similarity between simulation and evaluation of integrals

Metropolis-Hastings algorithm

- Proposal distribution $h(\boldsymbol{x}, \boldsymbol{y})$
- Acceptance of solution? The solution will be accepted with probability

$$\min\left(1, \frac{f(\boldsymbol{y})h(\boldsymbol{y},\boldsymbol{x})}{f(\boldsymbol{x})h(\boldsymbol{x},\boldsymbol{y})}\right) = \min\left(1, \frac{g(\boldsymbol{y})h(\boldsymbol{y},\boldsymbol{x})}{g(\boldsymbol{x})h(\boldsymbol{x},\boldsymbol{y})}\right)$$

$$\left(= \min\left(1, \frac{g(\boldsymbol{y})}{g(\boldsymbol{x})}\right) \text{ for } h(\boldsymbol{y},\boldsymbol{x}) = h(\boldsymbol{x},\boldsymbol{y})\right)$$

- Avoiding the troublesome constant K!
- Frequently we apply a symmetric proposal distribution $h(\boldsymbol{y}, \boldsymbol{x}) = h(\boldsymbol{y}, \boldsymbol{x})$ Metropolis algorithm
- It can be shown that this Markov chain will have $f(\boldsymbol{x})$ as stationary distribution.

Random Walk Metropolis-Hastings

Sampling from p.d.f. $c \cdot g(x)$ where c is unknown.

1. At iteration i, the state is X_i
2. *Propose* to jump from X_i to $Y_i = X_i + \Delta X_i$ where ΔX_i is sampled indepedently from a symmetric distribution
   - If $g(Y) \geq g(X_i)$, accept
   - If $g(Y) \leq g(X_i)$, accept w.p. $g(Y)/g(X_i)$
3. On accept: Set $X_{i+1} = Y_i$ and goto 1.
4. On reject: Set $X_{i+1} = X_i$ and goto 1.

Note that knowing c is not necessary!

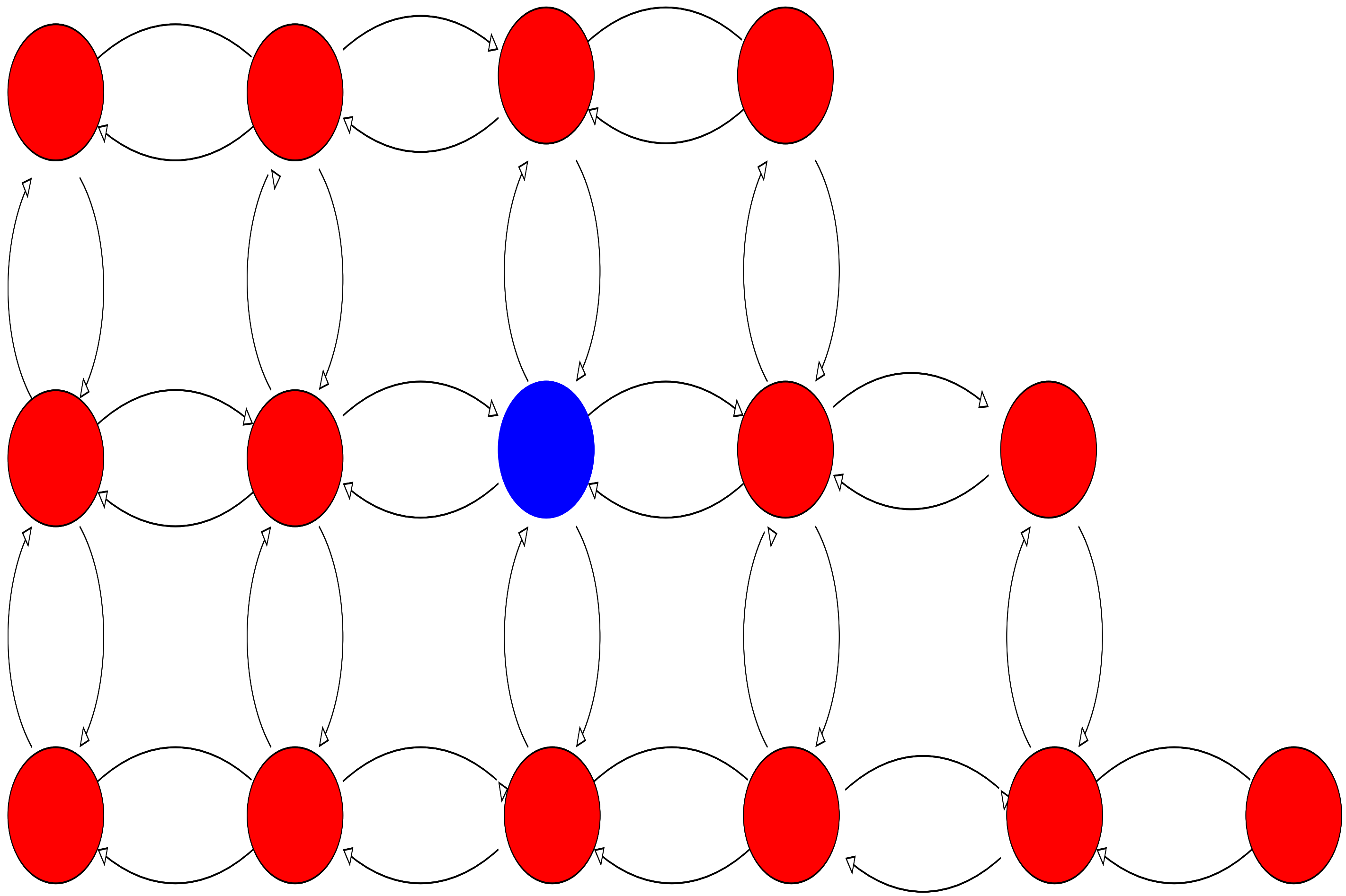
Proposal distribution (Gelman 1998)

- A good proposal distribution has the following properties
  - ◇ For any $\boldsymbol{x}$, it is easy to sample from $h(\boldsymbol{x}, \boldsymbol{y})$
  - ◇ It is easy to compute the accpetance probability
  - ◇ Each jump goes a reasonable distance in the parameter space
  - ◇ The proposals are not rejected too frequently

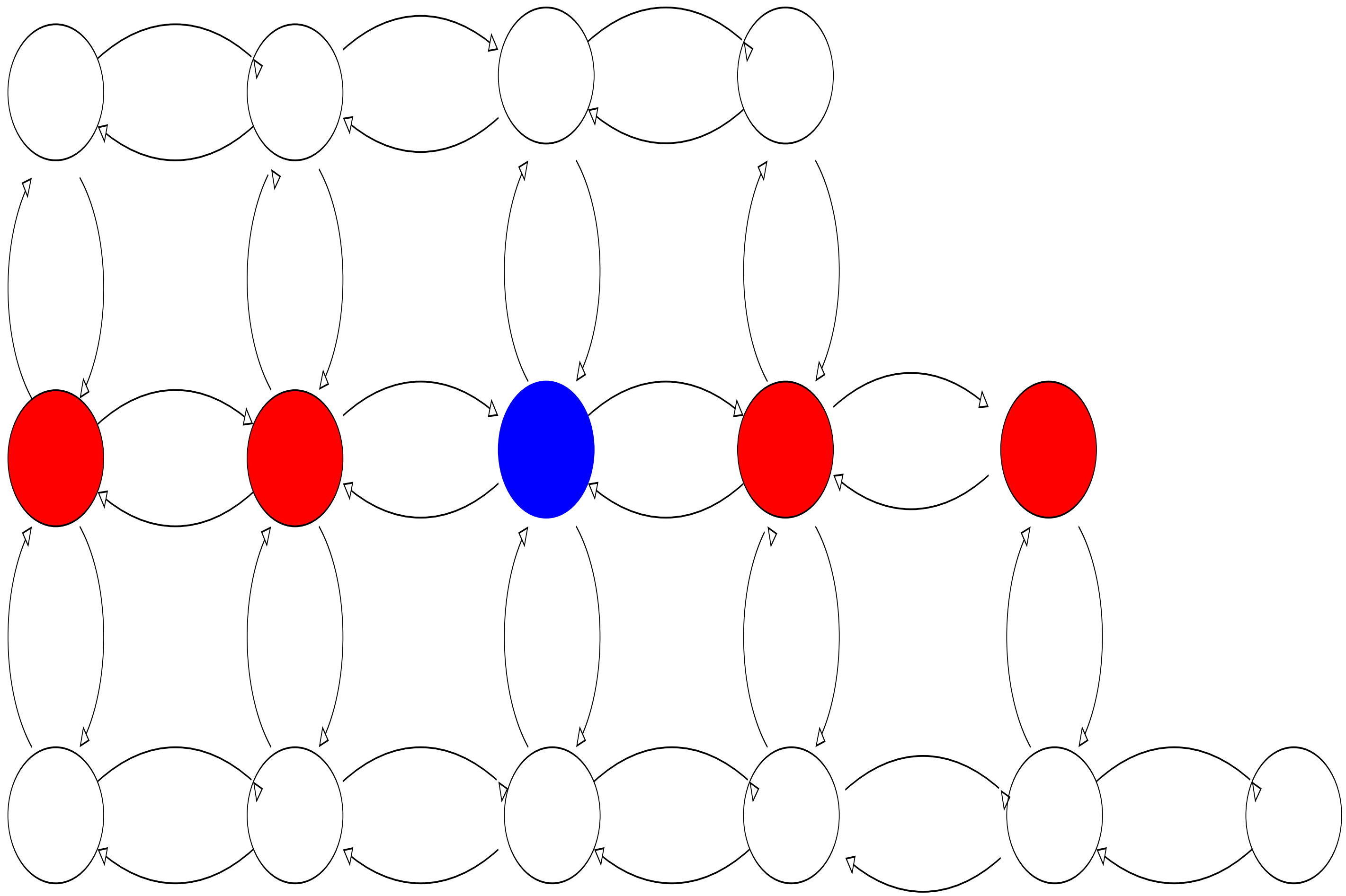
Gibss sampling

- Applies in multivariate cases where the conditional distribution among the coordinates are known.
- For a multidimensional distribution $\boldsymbol{x}$ the Gibss sampler will modify only one coordinate at a time.
- Typically d-steps in each iteration, where d is the dimension of the parameter space , that is of $\boldsymbol{x}$

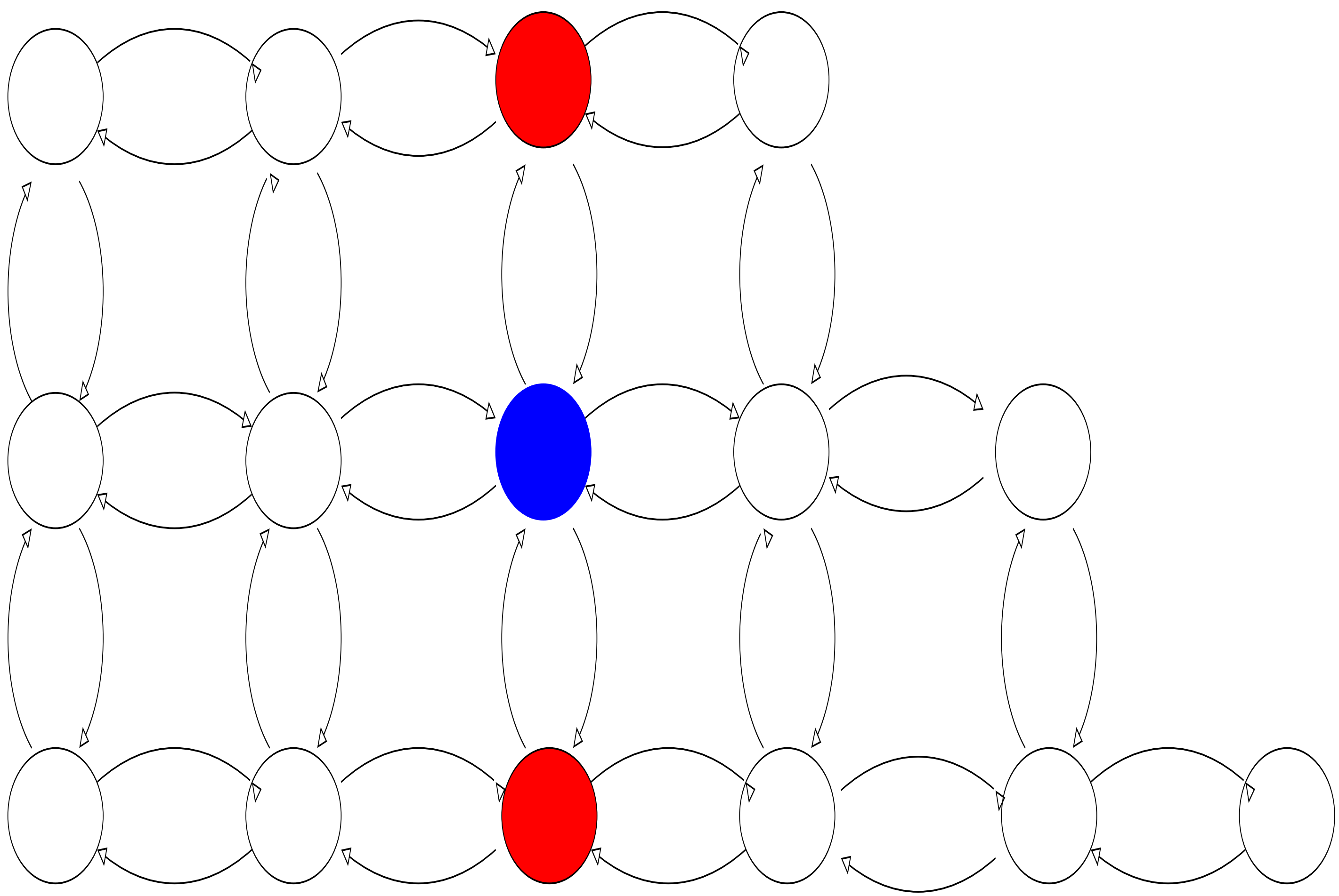
Illustration of ordinary and MCMC sampling



Gibbs sampling - first dimension



Gibbs sampling - second dimension



Direct Markov chain as oppposed to MCMC

- For an ordinary Markov chain we know P and find $\boldsymbol{\pi}$ - analytically or by simulation
- When we apply MCMC
  - ◇ For a discrete distribution we know $K\boldsymbol{\pi}$ construct P which has no physical interpretation in general and obtain $\boldsymbol{\pi}$ by simulation
  - ◇ For a continuous distribution we know $g(\boldsymbol{x})$ construct a transition kernel $P(\boldsymbol{x},\boldsymbol{y})$ and get $f(\boldsymbol{x})$ by simulation.

Remarks

- The method is computer intensive
- It is hard to verify the assumptions (Read: impossible)
- Warmup period strongly recommended (necessary indeed!)
- The samples are correlated
- Should be run several times with different starting conditions
  - ◇ Comparing within run variance with between run variance
- Check the BUGS site:
  http://www.mrc-bsu.cam.ac.uk/bugs/and/or links given at the BUGS site

Further reading

- A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin: Bayesian Data Analysis, Chapmann & Hall 1998, ISBN 0 412 03991 5
- W.R. Gilks, S. Richarson, D.J. Spiegelhalter: Markov chain Mone Carlo in practice, Chapmann & Hall 1996, ISBN 0 412 05551 1

Beyond Random Walk Metropolis-Hastings

- Proposed points Y_i can be generated with other schemes - this would change the acceptance probabilities.
- In mulitvariate situations, we can process one co-ordinate at a time (Gibbs sampling)
- This is well suited for *graphical models* with many variables, which each interact only with a few others
- (Decision support systems is a big area of application)
- Many hybrids and specialized versions exist
- Very active research area, both theory and applications

Exercise 6: Markov Chain Monte Carlo simulation

- The number of busy lines in a trunk group (Erlang system) is given by a truncated Poisson distribution

$$P(i) = \frac{\frac{A^i}{i!}}{\sum_{j=0}^{n} \frac{A^j}{j!}}$$

- Generate values from this distribution by applying the Metropolis-Hastings algorithm, verify with a χ^2-test. You can use the parameter values from exercise 4.

Exercise 6 continued

- For two different call types the joint number of occupied lines is given by

$$P(i,j) = \frac{1}{K}\frac{A_1^i}{i!}\frac{A_2^j}{j!}$$

- Use Metropolis-Hastings, directly and coordinate wise to generate variates from this distribution. You can use $A_1, A_2 = 4$ og $n = 10$.
- Test the distribution with a χ^2 test
- Optional: Redo the coordinate wise solution using Gibbs sampling. You will need to find the conditional distributions analytically.
- Optional: Redo the exercise with BUGS or other available software
- The system can be extended to an arbitrary dimension, and we

can add restrictions on the different call types.