UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2022
Markov Chain Monte Carlo

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

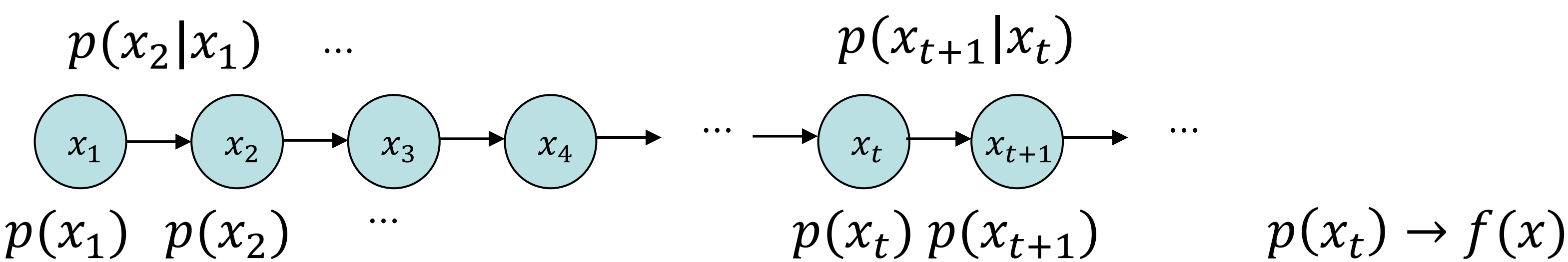
Now

- Exact methods
  - Inversion/transformation methods
  - Rejection sampling
- Approximate methods
  - Sampling importance resampling
  - Sequential Monte Carlo
  - Markov chain Monte Carlo (Chapter 7 and 8)
- Variance reduction methods
  - Importance sampling
  - Antithetic sampling
  - Control variates
  - Rao-blackwellization
  - Common random numbers

Markov chain Monte Carlo

- Previously we computed weights to correct the distribution (or used rejection sampling)

- Now we will create a sequence of samples which will converge to samples from the correct distribution



Markov chain Monte Carlo (McMC)

- Assume now simulating from $f(\mathbf{X})$ is difficult directly
  - $f(\cdot)$ complicated
  - $\mathbf{X}$ high-dimensional
- Markov chain Monte Carlo:
  - Generates $\{\mathbf{X}^{(t)}\}$ sequentially
  - Markov structure: $\mathbf{X}^{(t)} \sim P(\cdot|\mathbf{X}^{(t-1)})$
- Aim now:
  - The distribution of $\mathbf{X}^{(t)}$ converges to $f(\cdot)$ as t increases
  - $\hat{\mu}_{MCMC} = N^{-1}\sum_{t=1}^{N} h(\mathbf{X}^{(t)})$ converges towards $\mu = E^f[h(\mathbf{X})]$ as t increases

Why?
We had problems with weight decay and degeneracy in the direct approach now we can iterate to improve results

Markov chain theory – discrete case

- Assume $\{X^{(t)}\}$ is a Markov chain where $X^{(t)}$ is a discrete random variable

  $$\Pr(X^{(t)} = y|X^{(t-1)} = x) = P(y|x)$$

  giving the transition probabilities
- Assume the chain is
  - irreducible: It is possible to move from any **x** to any **y** in a finite number of steps
  - reccurent: The chain will visit any state infinitely often.
  - aperiodic: Does not go in cycles
- Then there exists a unique distribution $f(x)$ such that

  $$\lim_{t\to\infty} \Pr(X^{(t)} = y|X^{(0)} = x) = f(y)$$

  Limit distribution

  $$\hat{\mu}_{MCMC} \to \mu = E^f[X]$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

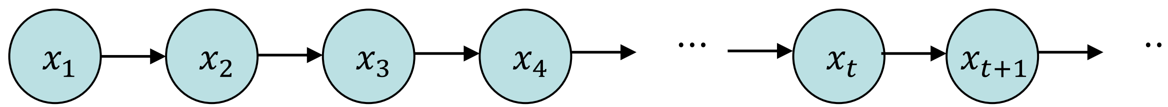
  $$f(y) = \sum_x f(x)P(y|x)$$

  Stationary distribution (fix point)

- Our situation: We have $f(y)$, want to find $P(y|x)$
  - Note: Many possible $P(y|x)$!

Discrete Transition probability

$p(x_2|x_1)$ ... $p(x_{t+1}|x_t)$

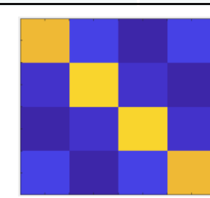


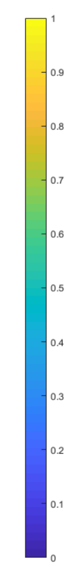
$p(x_1)$ $p(x_2)$ ... $p(x_t)$ $p(x_{t+1})$ $p(x_t) \to f(x)$

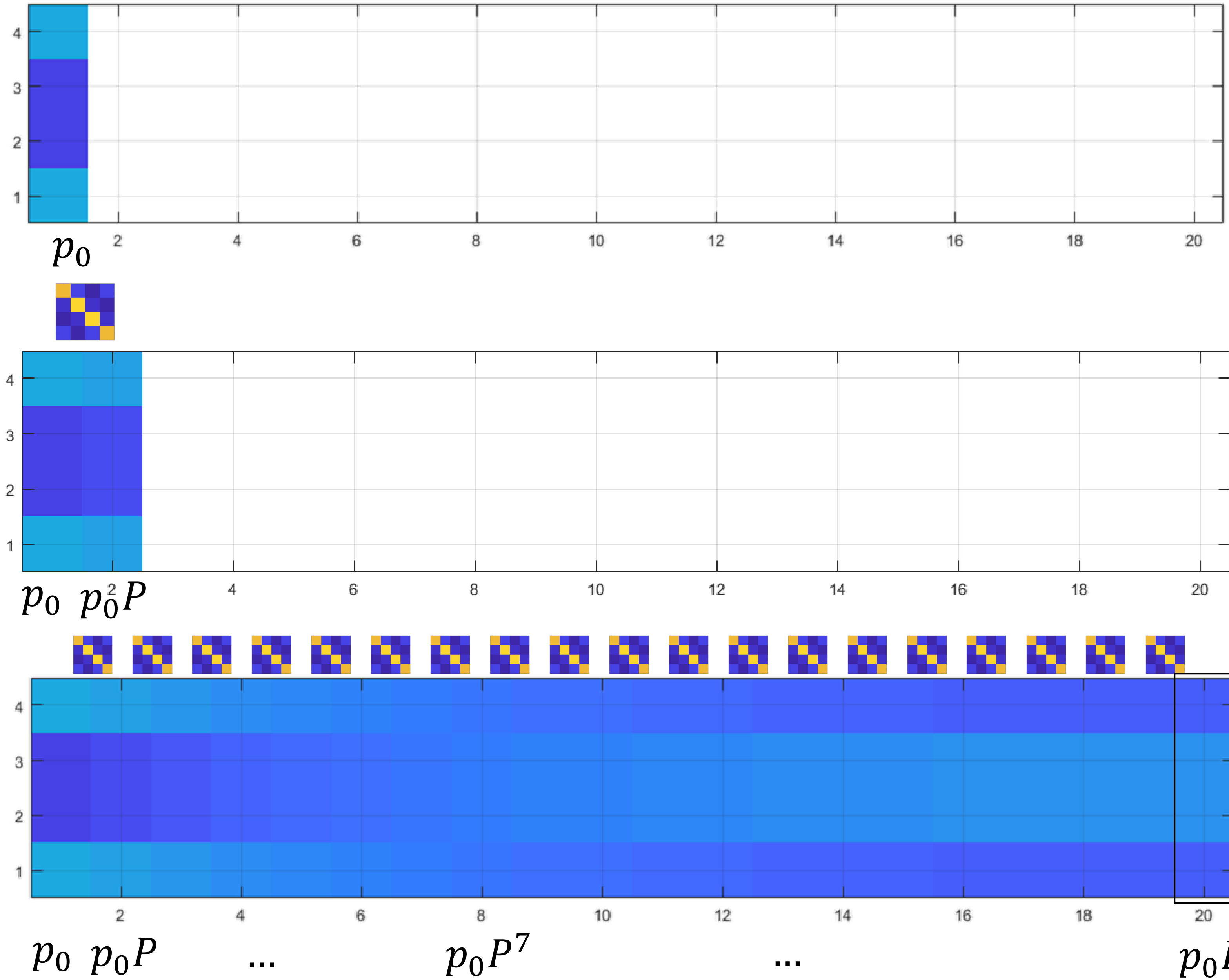
- Need initial distribution $p(x_1)$, say we have 4 possible classes
- and transition probability $p(x_t|x_{t-1})$, we need a transition to each state

		x_2				
x_1	$p(x_1)$		1	2	3	4
1	0.4	$p(x_2\|x_1 = 1)$	0.80	0.10	0.00	0.10
2	0.1	$p(x_2\|x_1 = 2)$	0.05	0.90	0.05	0.00
3	0.1	$p(x_2\|x_1 = 3)$	0.00	0.05	0.90	0.05
4	0.4	$p(x_2\|x_1 = 4)$	0.10	0.00	0.10	0.80

$p_0 =$ 

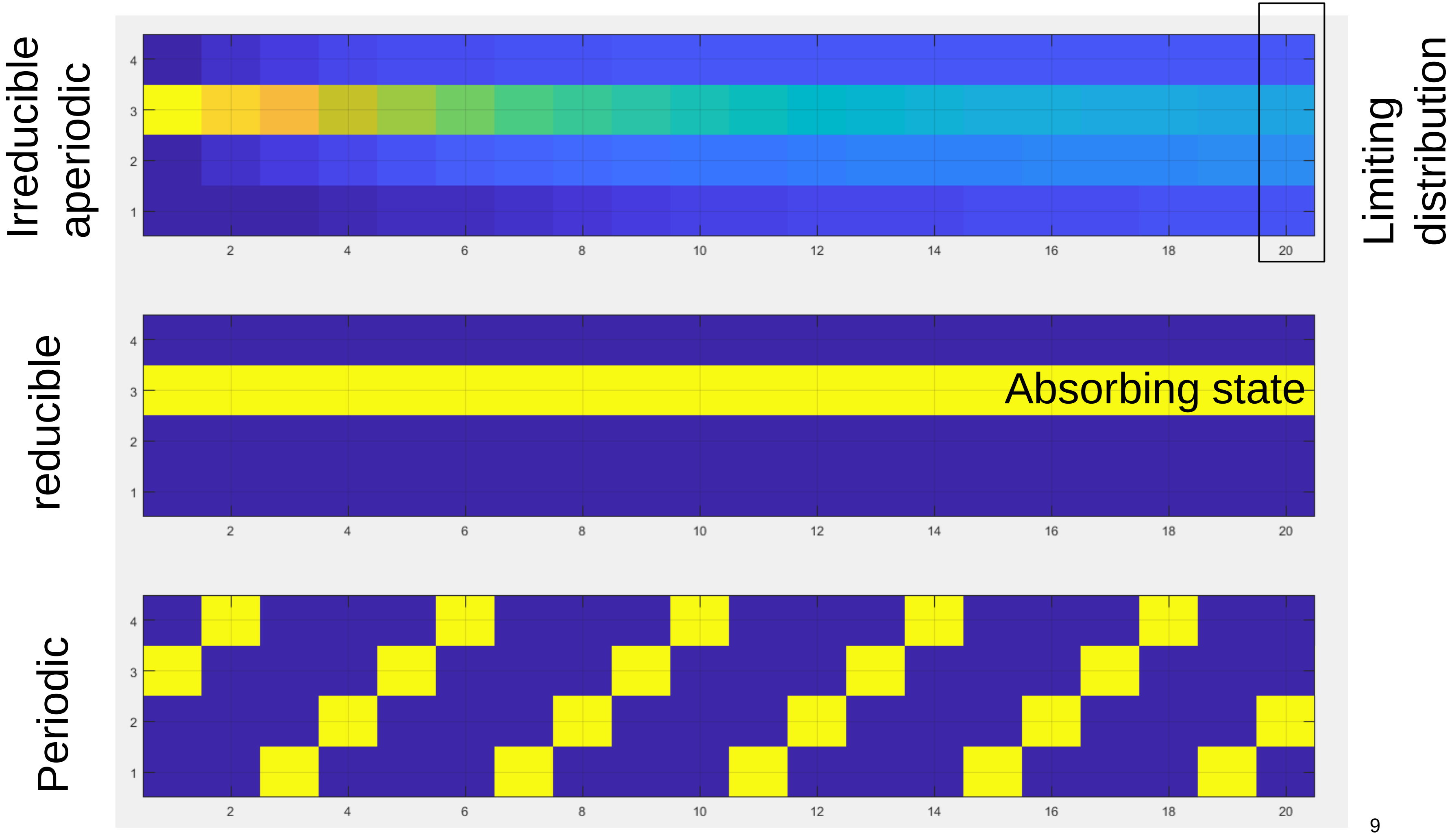
$P =$ 





$p(x_{20})$
0.17
0.33
0.33
0.17

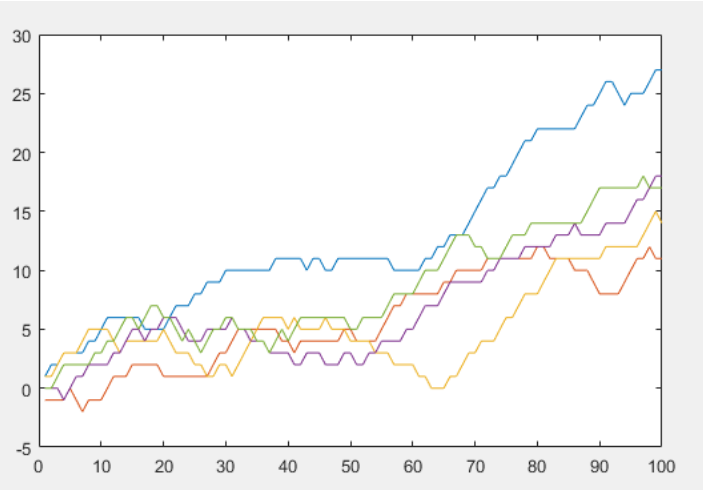
Irreducible/ aperiodic:



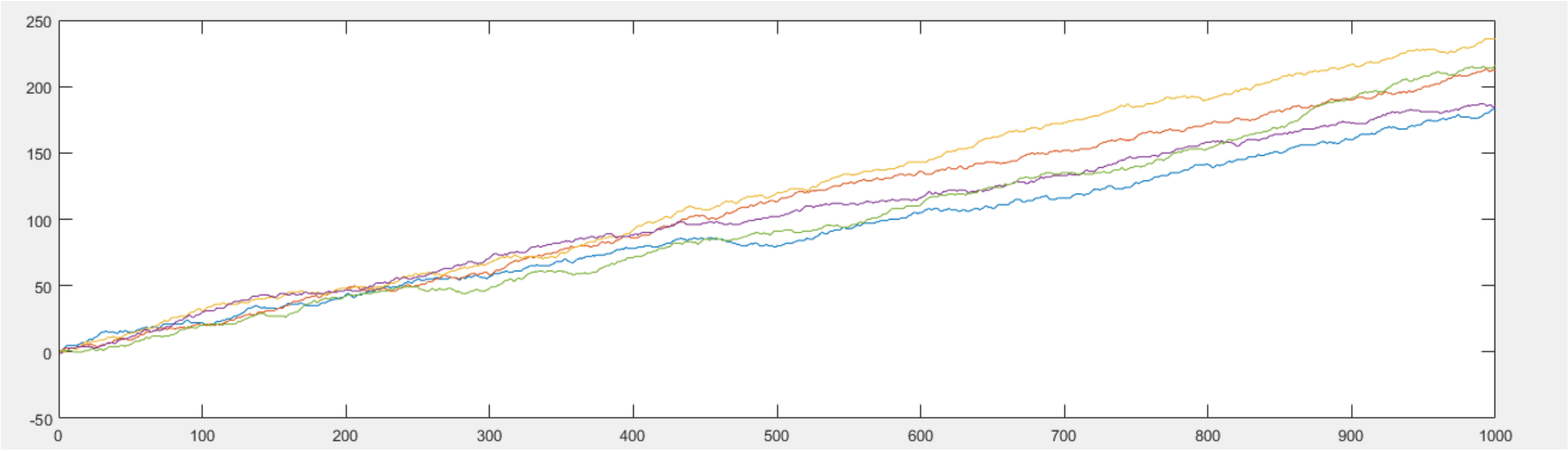
Recurrent (OK if finite and irreducible)

- Problem if countable many discrete classes

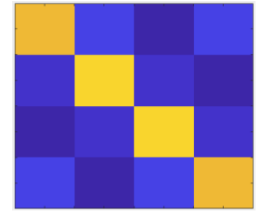
$$P(x_t|x_{t-1}) = \begin{cases} 0.6 & x = x_{t-1} \\ 0.3 & x = x_{t-1} + 1 \\ 0.1 & x = x_{t-1} - 1 \end{cases}$$

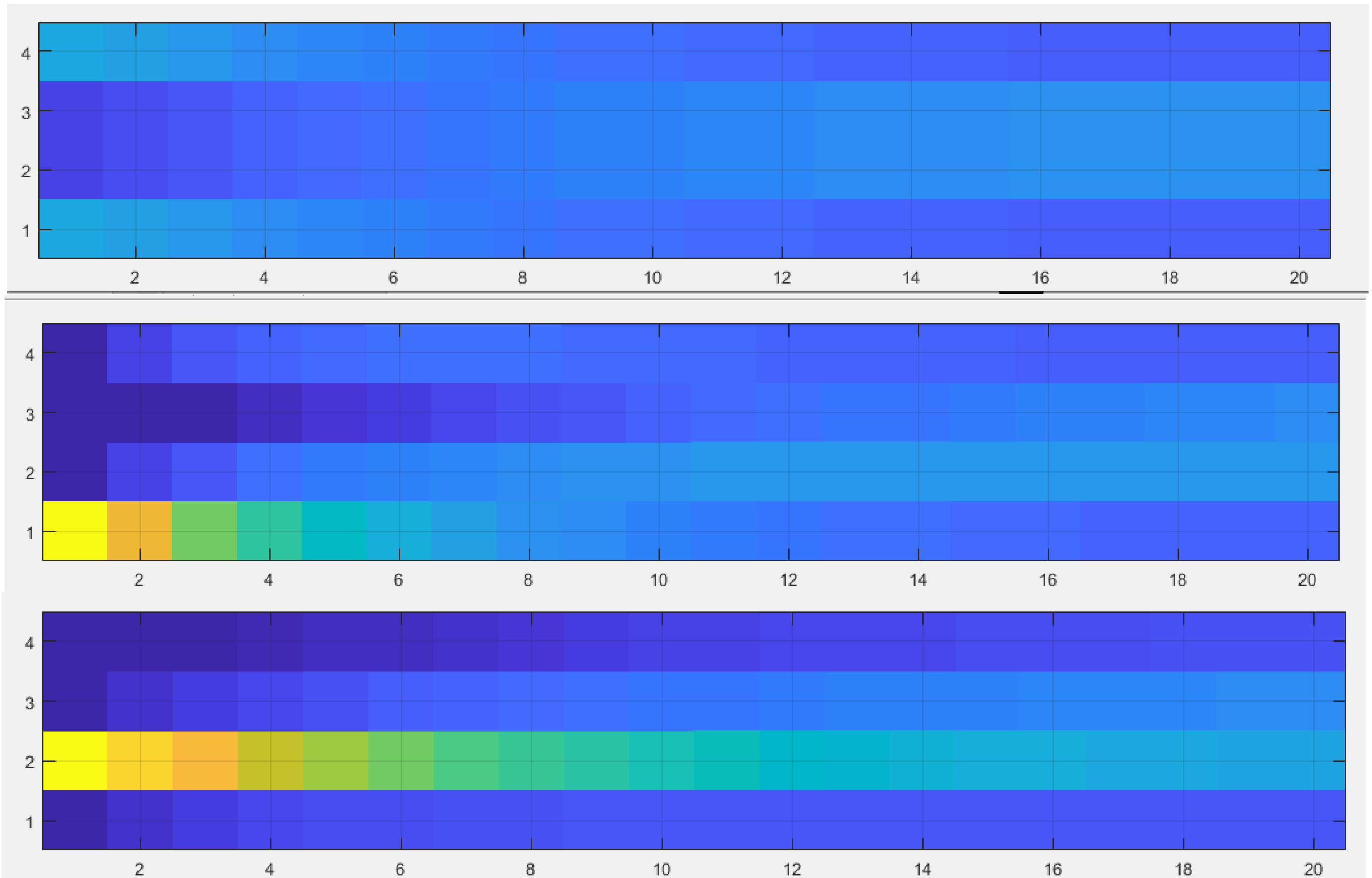


No return



Limiting distribution





When the Markov chain is irreducible / aperiodic /recurrent

- The limiting distribution is equal to the stationary distribution

$$p_s = p_{\mathrm{Lim}}$$

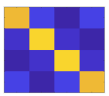
- Stationary distribution is fix point of iteration

$$p_s P = p_s$$

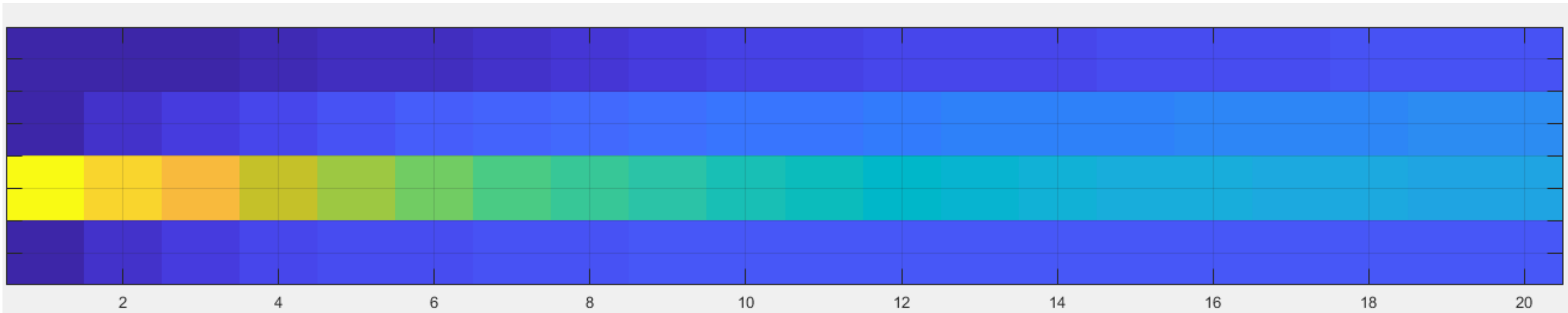
- Limiting distribution (is independent of p_0)

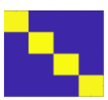
$$\lim_{n\to\infty} p_0 P^n = p_{\mathrm{Lim}}$$

Time to reach limiting distribution n=20

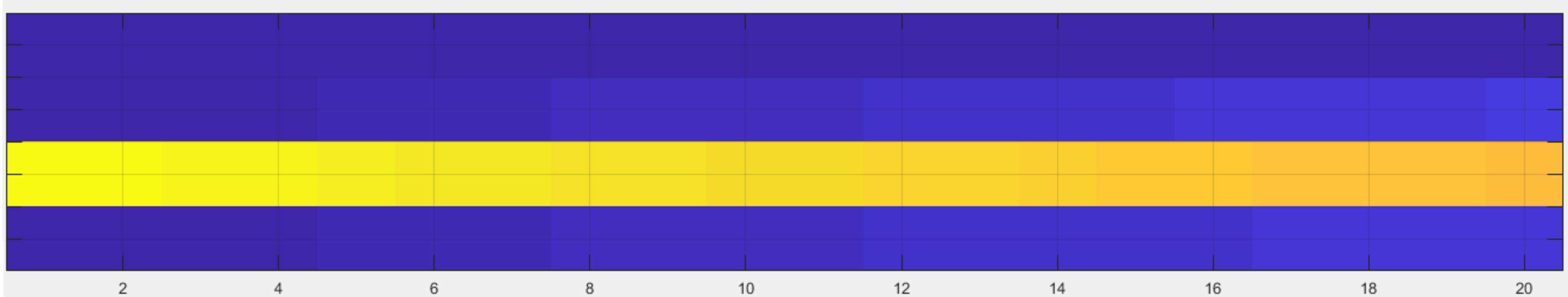


0.8000	0.1000	0	0.1000
0.0500	0.9000	0.0500	0
0	0.0500	0.9000	0.0500
0.1000	0	0.1000	0.8000

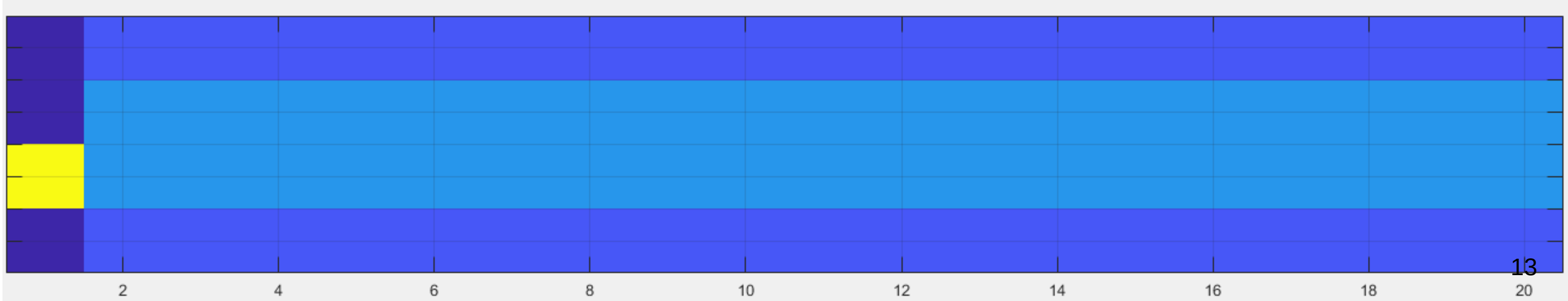




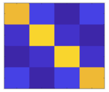
0.9800	0.0100	0	0.0100
0.0050	0.9900	0.0050	0
0	0.0050	0.9900	0.0050
0.0100	0	0.0100	0.9800

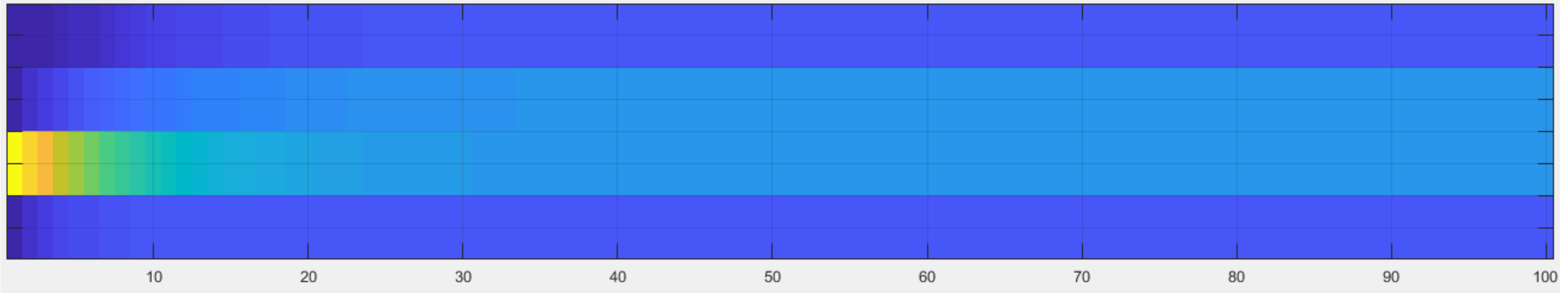


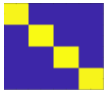
0.1667	0.3333	0.3333	0.1667
0.1667	0.3333	0.3333	0.1667
0.1667	0.3333	0.3333	0.1667
0.1667	0.3333	0.3333	0.1667

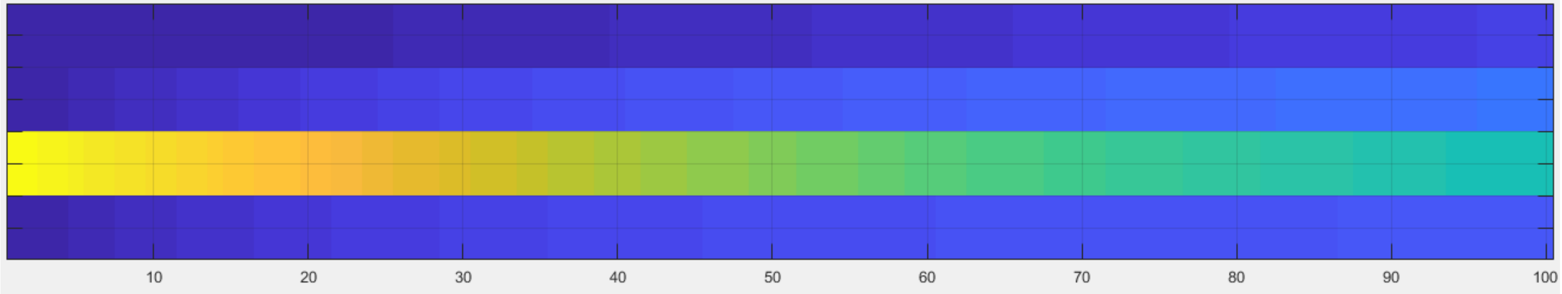


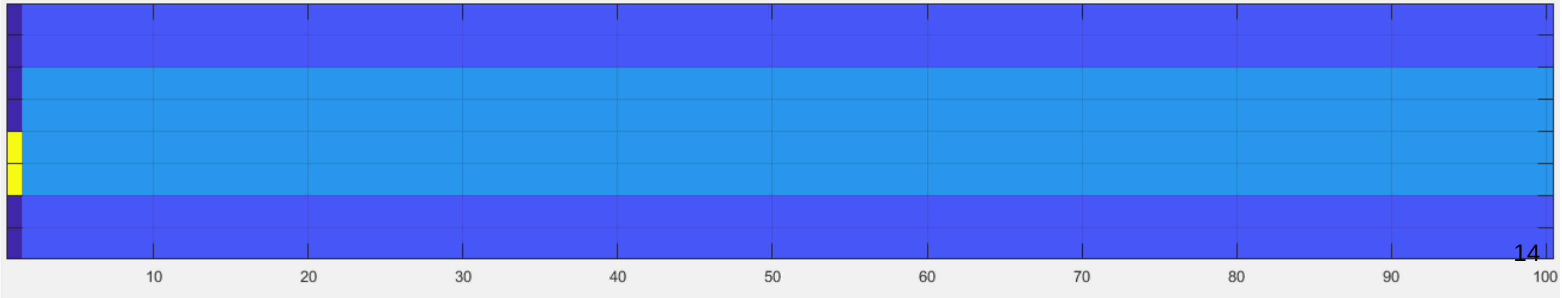
Time to reach limiting distribution n=100



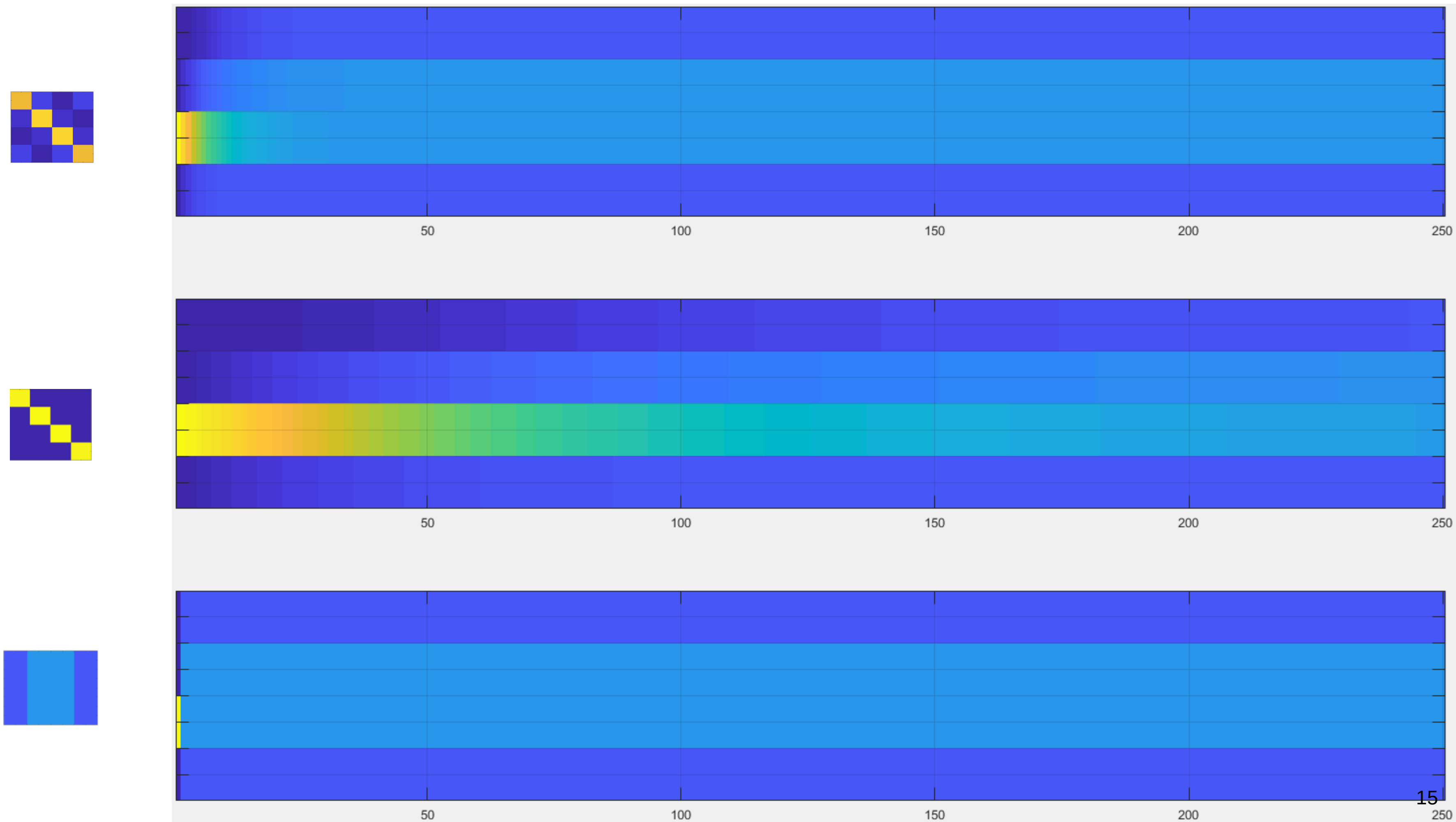








Time to reach limiting distribution n=250



Markov chain theory – discrete case

- Assume $\{X^{(t)}\}$ is a Markov chain where $X^{(t)}$ is a discrete random variable

$$\Pr(X^{(t)} = y|X^{(t-1)} = x) = P(y|x)$$

  giving the transition probabilities
- Assume the chain is
  - irreducible: It is possible to move from any **x** to any **y** in a finite number of steps
  - reccurent: The chain will visit any state infinitely often.
  - aperiodic: Does not go in cycles
- Then there exists a unique distribution $f(x)$ such that

$$\lim_{t\to\infty} \Pr(X^{(t)} = y|X^{(0)} = x) = f(y)$$

$$\hat{\mu}_{MCMC} \to \mu = E^f[X]$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

$$f(y) = \sum_x f(x)P(y|x)$$

- Our situation: We have $f(y)$, want to find $P(y|x)$
  - Note: Many possible $P(y|x)$!

Markov chain theory - general setting

- Assume $\{\mathbf{X}^{(t)}\}$ is a Markov chain where $\mathbf{X}^{(t)} \in S$

$$\Pr(\mathbf{X}^{(t)} \in A | \mathbf{X}^{(t-1)} = \mathbf{x}) = P(\mathbf{x}, A) = \int_{\mathbf{y} \in A} P(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

  giving the transition densities
- Assume the chain is
  - irreducible: It is possible to move from any $\mathbf{x}$ to any $\mathbf{y}$ in a finite number of steps
  - reccurent: The chain will visit any $A \subset S$ infinitely often.
  - aperiodic: Do not go in cycles
- Then there exists a distribution $f(\mathbf{x})$ such that

$$\lim_{t \to \infty} \Pr(\mathbf{X}^{(t)} \in A | \mathbf{X}^{(0)} = \mathbf{x}) = \int_A f(\mathbf{y}) d\mathbf{y}$$

$$\hat{\mu}_{MCMC} \to \mu$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

- Our situation: We have $f(\cdot)$, want to find $P(\mathbf{y}|\mathbf{x})$

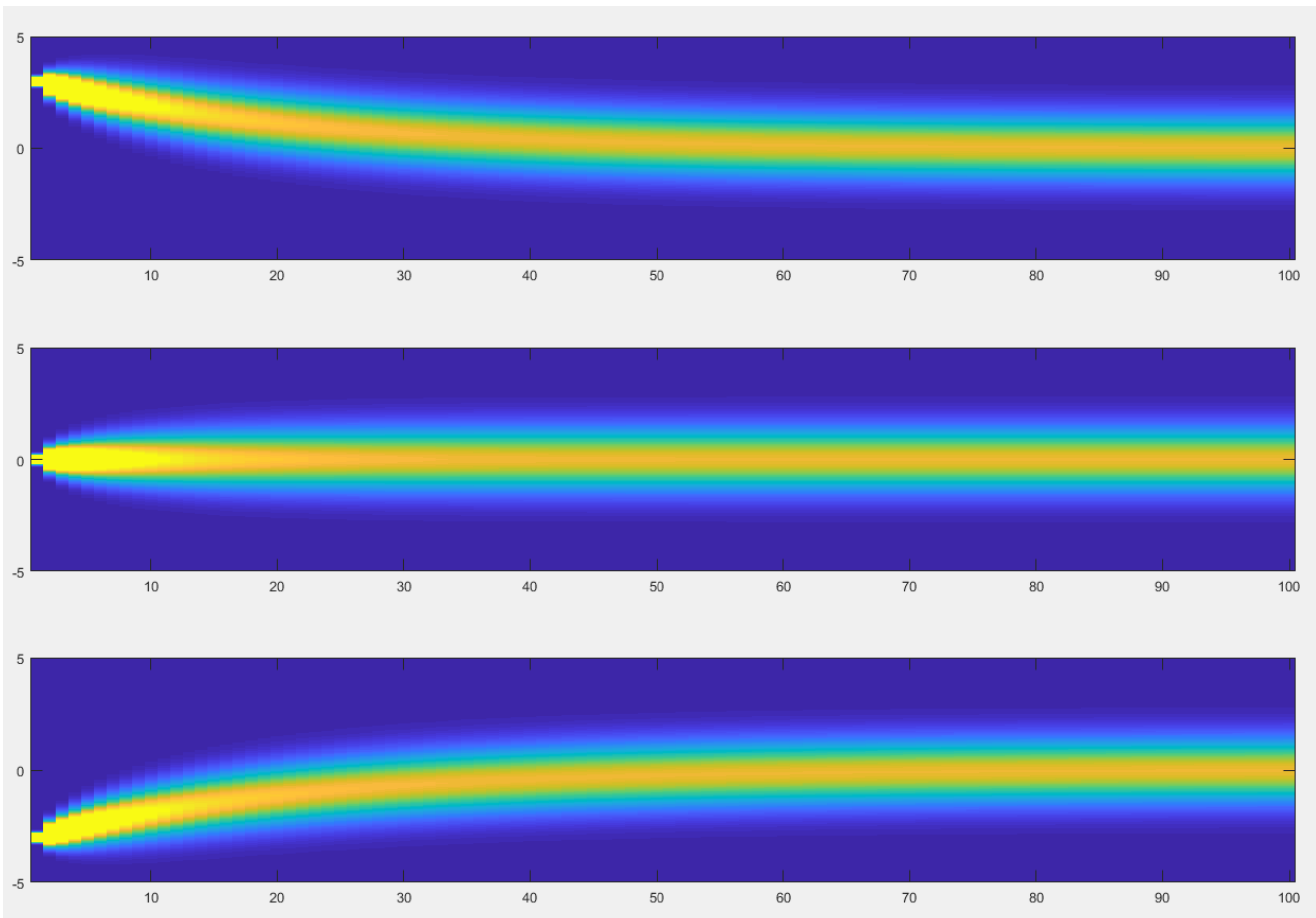
Example of a continuous transition density, AR1 model

$$p(x_t|x_{t-1}) = \phi(ax_{t-1}, \sigma^2(1-a^2))$$

$p(x_1)$ $p(x_n)$

$p(x_1)$ $p(x_n)$

$p(x_1)$ $p(x_n)$



Markov chain theory - general setting

- Assume $\{\mathbf{X}^{(t)}\}$ is a Markov chain where $\mathbf{X}^{(t)} \in S$

$$\Pr(\mathbf{X}^{(t)} \in A | \mathbf{X}^{(t-1)} = \mathbf{x}) = P(\mathbf{x}, A) = \int_{\mathbf{y} \in A} P(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

  giving the transition densities
- Assume the chain is
  - irreducible: It is possible to move from any $\mathbf{x}$ to any $\mathbf{y}$ in a finite number of steps
  - reccurent: The chain will visit any $A \subset S$ infinitely often.
  - aperiodic: Do not go in cycles
- Then there exists a distribution $f(\mathbf{x})$ such that

$$\lim_{t\to\infty} \Pr(\mathbf{X}^{(t)} \in A | \mathbf{X}^{(0)} = \mathbf{x}) = \int_A f(\mathbf{y}) d\mathbf{y}$$

$$\hat{\mu}_{MCMC} \to \mu$$

- How to find $f(\cdot)$ (the stationary distribution): Solve

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

- Our situation: We have $f(\cdot)$, want to find $P(\mathbf{y}|\mathbf{x})$

We want to construct P(x|y) to match our needs

- Need to have good properties
  - Stationary
  - Irreducible
  - Aperiodic
  - Recurrent
- Also need to get our target as a stationary distribution

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x}$$

  - Simplify the hunt by introducing symmertry
  - **detailed balance**

Detailed balance

- The task: Find a transition probability/density $P(\mathbf{y}|\mathbf{x})$ satisfying

$$f(\mathbf{y}) = \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{y}|\mathbf{x})d\mathbf{x}$$

  Can in general be a difficult criterion to check
- Sufficient criterion:

$$f(\mathbf{x})P(\mathbf{y}|\mathbf{x}) = f(\mathbf{y})P(\mathbf{x}|\mathbf{y}) \quad \text{Detailed balance}$$

  We then have

$$\int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{y}|\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}} f(\mathbf{y})P(\mathbf{x}|\mathbf{y})d\mathbf{x}$$
$$= f(\mathbf{y}) \int_{\mathbf{x}} P(\mathbf{x}|\mathbf{y})d\mathbf{x} = f(\mathbf{y})$$

  since $P(\mathbf{x}|\mathbf{y})$ is, for any given $\mathbf{y}$, a density wrt $\mathbf{x}$.
- Note: For $\mathbf{y} = \mathbf{x}$, detailed balance always fulfilled, only necessary to check for $\mathbf{y} \neq \mathbf{x}$.

Metropolis-Hastings algorithm

- $P(\mathbf{y}|\mathbf{x})$ defined through an algorithm:
  1. Sample a candidate value $\mathbf{X}^*$ from a proposal distribution $g(\cdot|\mathbf{x})$.
  2. Compute the Metropolis-Hastings ratio
  $$R(\mathbf{x},\mathbf{X}^*) = \frac{f(\mathbf{X}^*)g(\mathbf{x}|\mathbf{X}^*)}{f(\mathbf{x})g(\mathbf{X}^*|\mathbf{x})}$$
  3. Put
  $$\mathbf{Y} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{1, R(\mathbf{x},\mathbf{X}^*)\} \\ \mathbf{x} & \text{otherwise} \end{cases}$$
- For $\mathbf{y} \neq \mathbf{x}$:
$$P(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}|\mathbf{x}) \min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\}$$
- Note: $P(\mathbf{x}|\mathbf{x})$ somewhat difficult to evaluate in this case.

Either we keep **x** with a certain probability
Or we change to **X*** which have a certain density

Metropolis-Hastings algorithm
Detailed balance

$$
\begin{aligned}
f(\mathbf{x})P(\mathbf{y}|\mathbf{x}) &= f(\mathbf{x})g(\mathbf{y}|x)\min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}\right\} \\
&= \min\{f(\mathbf{x})g(\mathbf{y}|\mathbf{x}), f(\mathbf{y})g(\mathbf{x}|\mathbf{y})\} \\
&= f(\mathbf{y})g(\mathbf{x}|\mathbf{y})\min\left\{\frac{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})}{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}, 1\right\} = f(\mathbf{y})P(\mathbf{x}|\mathbf{y})
\end{aligned}
$$

The probability of a value being repeated is positive

Pf:

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

$$\int_{y\neq \mathrm{x}} P(y|x)dy = \int_{y\neq \mathrm{x}} \underbrace{g(y|x)}_{\substack{\text{Density:} \\ \text{integrates to 1}}} \underbrace{\min\left\{1, \frac{f(y)g(x|y)}{f(x)g(y|x)}\right\}}_{\substack{\text{Positive number:} \\ \leq 1}} dy \leq 1$$

What about unknown scaling and MH

- Assume now $f(\mathbf{x}) = c \cdot q(\mathbf{x})$ with c unknown.

$$R(\mathbf{x},\mathbf{y}) = \frac{f(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{f(\mathbf{x})g(\mathbf{y}|\mathbf{x})} = \frac{c \cdot q(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{c \cdot q(\mathbf{x})g(\mathbf{y}|\mathbf{x})} = \frac{q(\mathbf{y})g(\mathbf{x}|\mathbf{y})}{q(\mathbf{x})g(\mathbf{y}|\mathbf{x})}$$

- Do not depend on c!

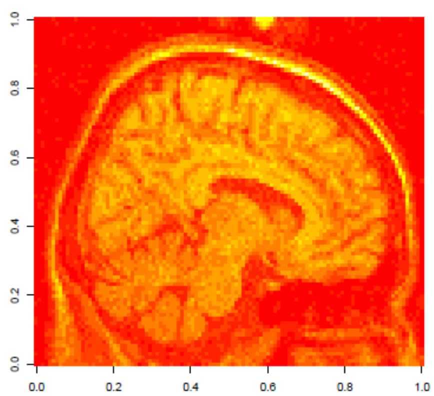
Important for Bayesian analysis Posterior ∝ Likelihood × Prior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \propto p(y|x)p(x)$$

Important for Gibbs type distributions

$$\Pr(\mathbf{C}) = \Pr(C_{11},\ldots.,C_{n_1 n_2}) = \frac{1}{Z} e^{-\beta \sum_{||(i,j)-(i'j')||=1} I(C_{ij} \neq C_{i'j'})}$$

$$\Pr(\mathbf{C}|\mathbf{y}) = \frac{\Pr(\mathbf{C}) \prod_{ij} f(y_{ij}|C_{ij})}{\sum_{\mathbf{C}'} \Pr(\mathbf{C}') \prod_{ij} f(y_{ij}|C'_{ij})}$$



Metropolis Hastings is a general form:

- Specific chains:
  - Random walk chains
  - Independent chains
  - Gibbs sampler
- Tricks to customize sampling
  - Reparametrize
  - Block update
  - Hybrid
  - Griddy Gibbs

Random walk chains

- Popular choice of proposal distribution:

$$\mathbf{X}^* = \mathbf{x} + \boldsymbol{\varepsilon}$$

- $g(\mathbf{x}^*|\mathbf{x}) = h(\mathbf{x}^* - \mathbf{x})$
- Popular choices: Uniform, Gaussian, t-distribution
- Note: If $h(\cdot)$ is symmetric, $g(\mathbf{x}^*|\mathbf{x}) = g(\mathbf{x}|\mathbf{x}^*)$ and

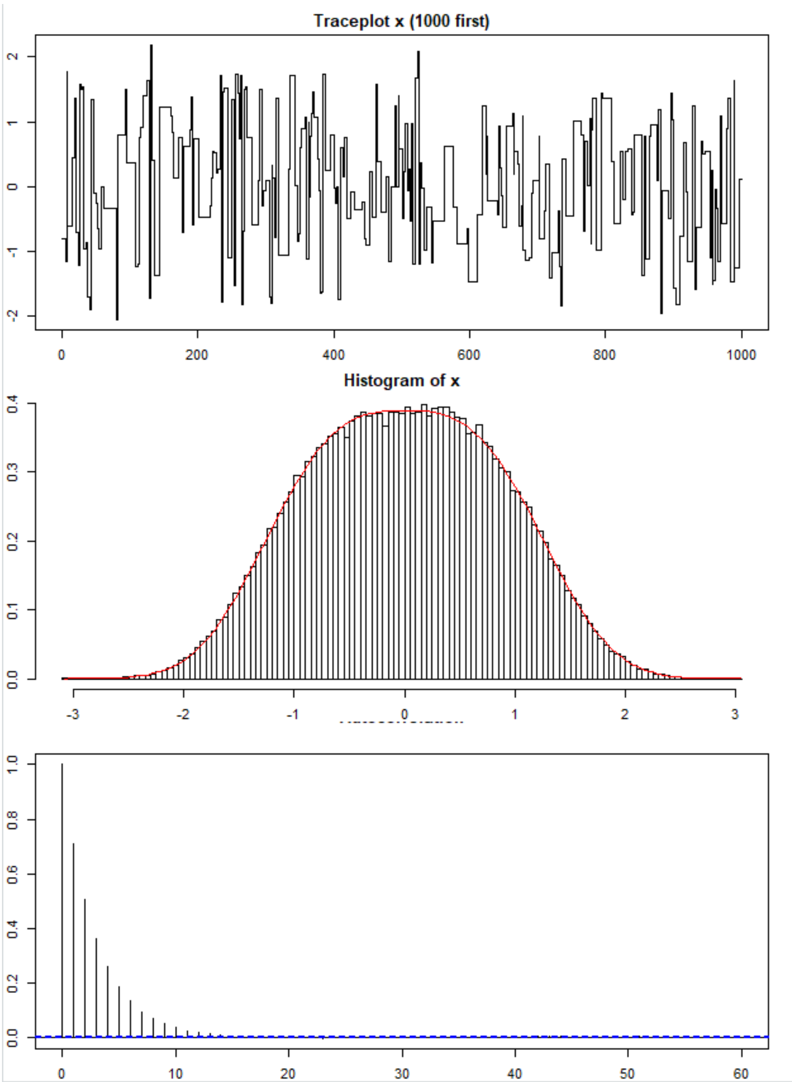
$$R(\mathbf{x}, \mathbf{x}^*) = \frac{f(\mathbf{x}^*)g(\mathbf{x}|\mathbf{x}^*)}{f(\mathbf{x})g(\mathbf{x}^*|\mathbf{x})} = \frac{f(\mathbf{x}^*)}{f(\mathbf{x})}$$

Example

- Assume $f(x) \propto \exp(-|x|^3/3)$
- Proposal distribution $N(x, 4^2)$
- Example_MH_cubic.R

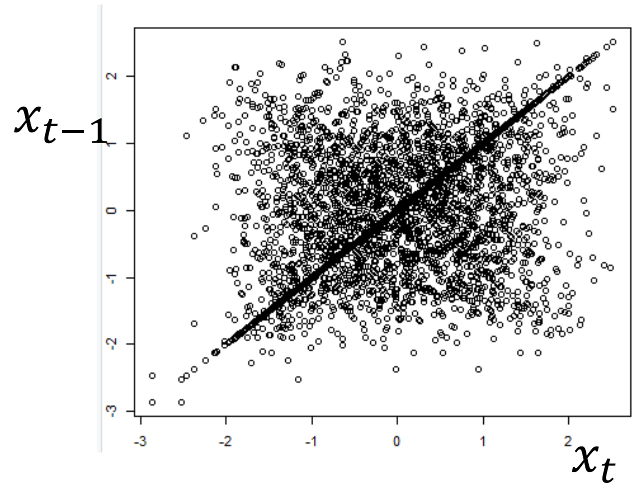
```
#Initial value
x = rnorm(1)
acc = 0
for(i in 2:N)
{
  y = rnorm(1,x[i-1],4) # proposal
  R = f(y)/f(x[i-1])    # acceptance ratio
                        # note that the acceptance rate is min(1,R),
  if(runif(1)<R)        # The syntax her will give that since we allways accept if R>1
  {
    x[i] = y
    acc = acc+1
  }
  else
   x[i] = x[i-1]
}
```

Results random walk




```
Acceptance rate
           = 0.2755276
```

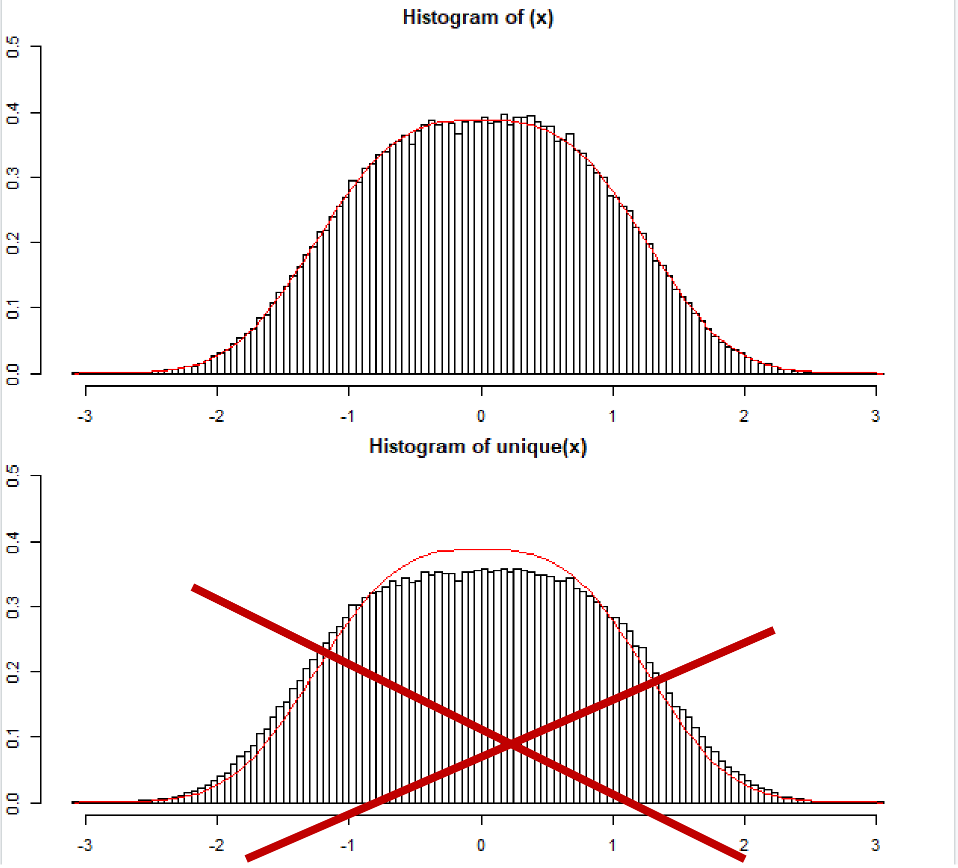
Lag one scatterplot



The repeats of a value is needed to get the correct distribution

Compare histograms to true distribution

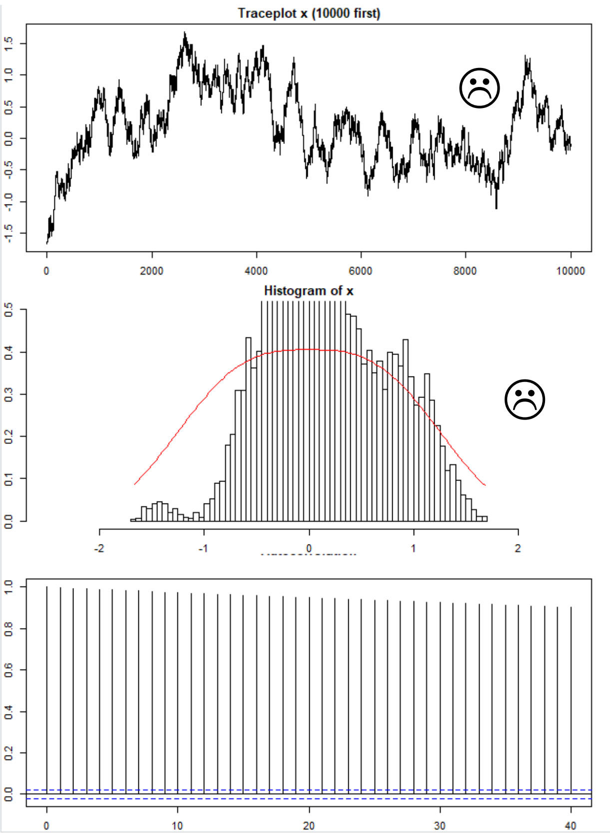
This is kind of similar to what we have for sampling importance resampling (SIR)
If a value is repeated it gets «more weight»



N = 10 000

The effect variance in proposal distribution

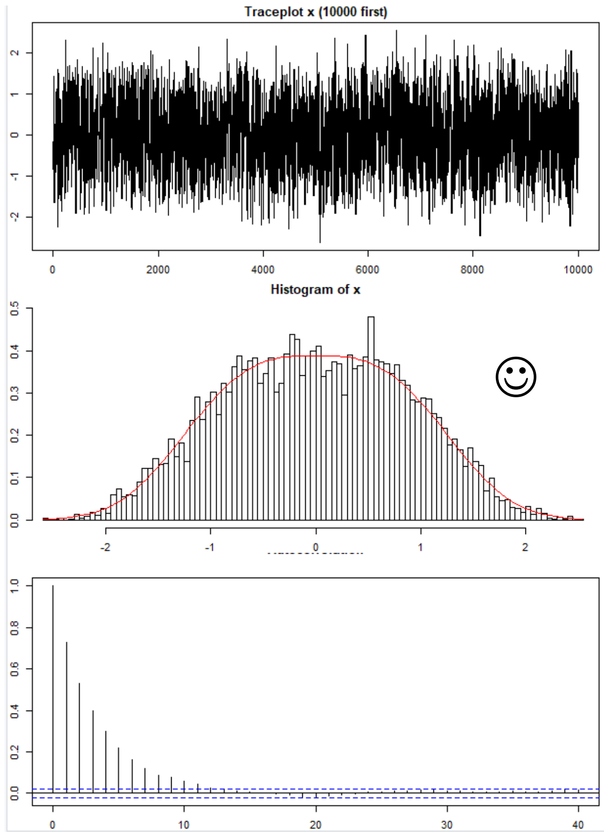
$g(y|x) = \phi(y; x, 0.04^2)$



Acc. rate =0.994

Too small steps,
high acceptance
high correlation ☹

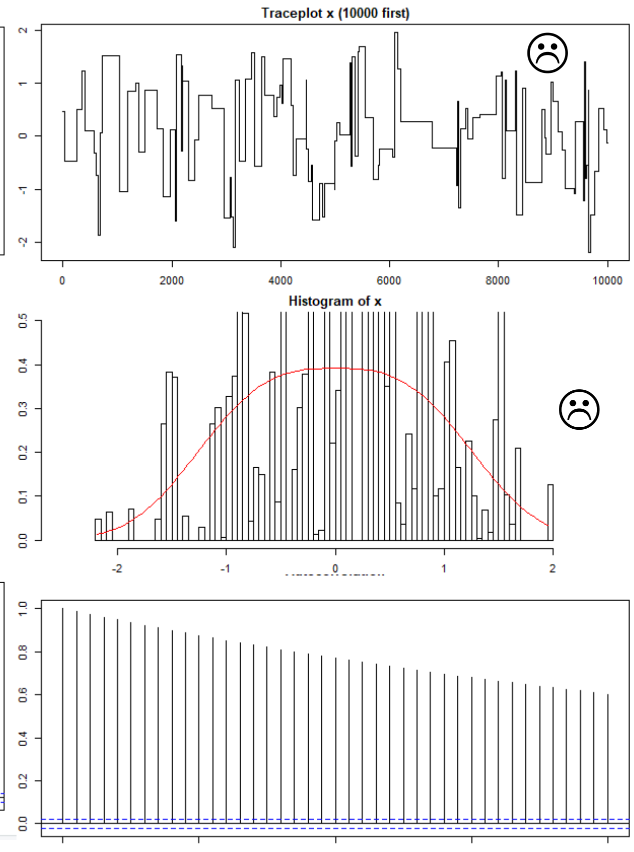
$g(y|x) = \phi(y; x, 1^2)$



Acc. Rate = 0.700

Just about right,
good acceptance
low correlation ☺

$g(y|x) = \phi(y; x, 100^2)$



Acc. Rate = 0.012

Too large changes proposed,
low acceptance
high correlation ☹

Independent chains

- Assume $g(\mathbf{x}^*|\mathbf{x}) = g(\mathbf{x}^*)$. Then

$$R(\mathbf{x}, \mathbf{x}^*) = \frac{f(\mathbf{x}^*)g(\mathbf{x})}{f(\mathbf{x})g(\mathbf{x}^*)} = \frac{\frac{f(\mathbf{x}^*)}{g(\mathbf{x}^*)}}{\frac{f(\mathbf{x})}{g(\mathbf{x})}},$$

  fraction of importance weights!
- Behave very much like importance sampling and SIR
- Difficult to specify $g(\mathbf{x})$ for high-dimensional problems
- Theoretical properties easier to evaluate than for random walk versions.

Challenges similar to what seen in:

- rejection sampling
- importance sampling
- sampling importance resampling

Example

- Assume $f(x) \propto \exp(-|x|^3/3)$

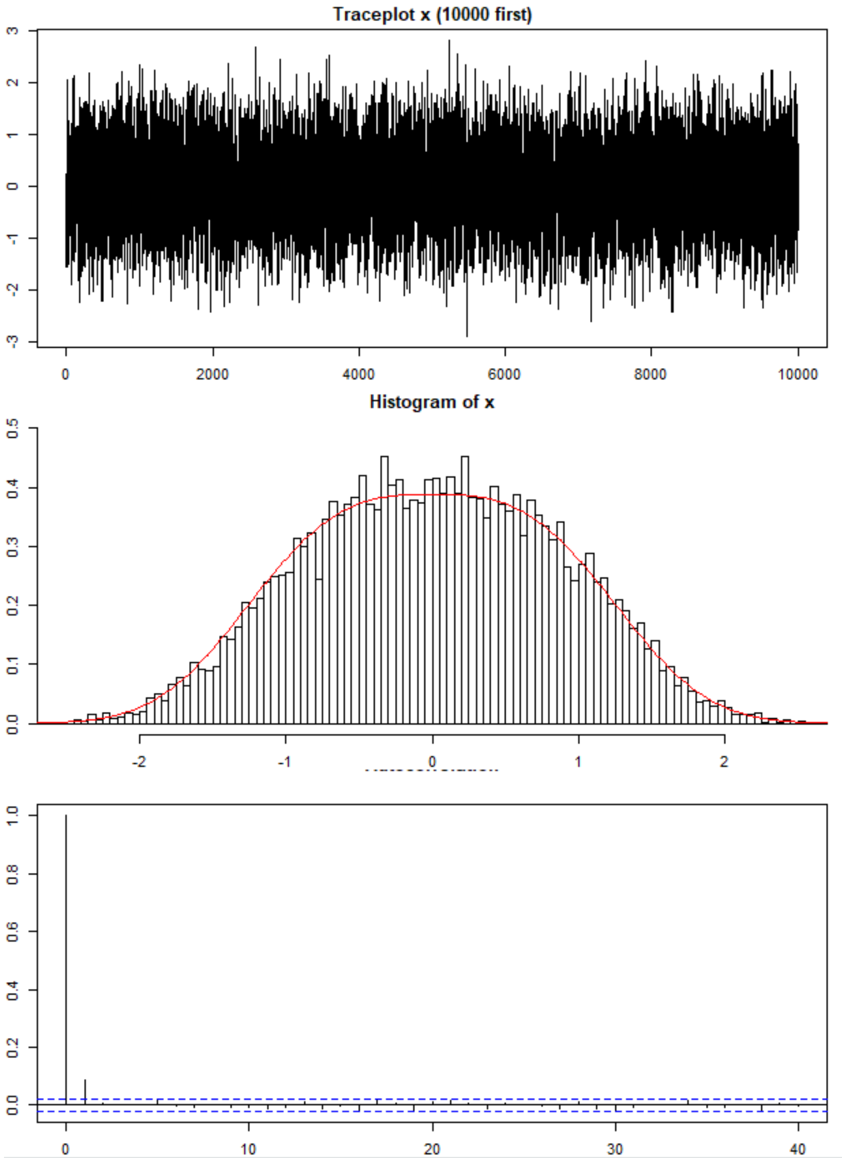
$$g(y|x) = \phi(y; 0, 1^2)$$

Example_MH_cubic_independence.R

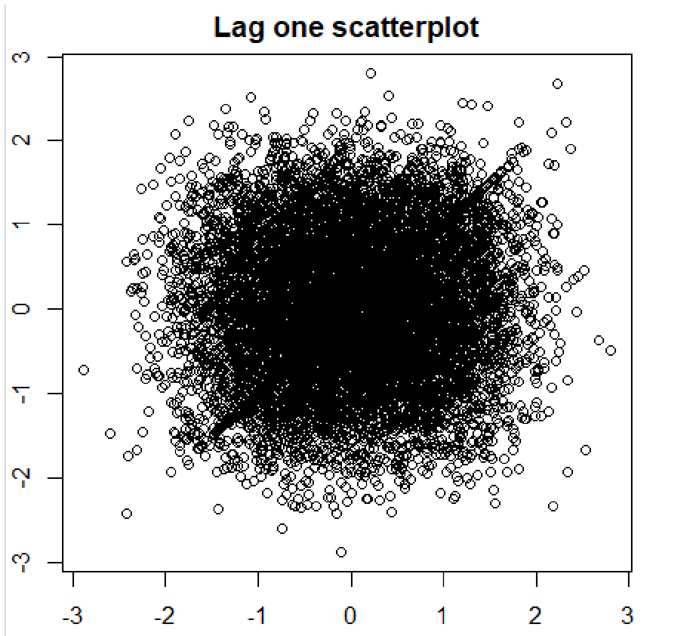
```
N = 10000     # Number of iterations
x = rep(NA,N)
varProp=1^2 # variance of proposal

#Initial value
x =  rnorm(1,0,varProp)
acc = 0
for(i in 2:N)
{
  y = rnorm(1,0,varProp) # proposal
  R = f(y)*dnorm(x[i-1],0,varProp)/(f(x[i-1])*dnorm(y,0,varProp))         # acceptance ratio
                              # note that the acceptance rate is min(1,R),
  if(runif(1)<R)              # The syntax her will give that since we allways accept if R>1
  {
    x[i] = y
    acc = acc+1
  }
  else
   x[i] = x[i-1]
}
```

Results independent



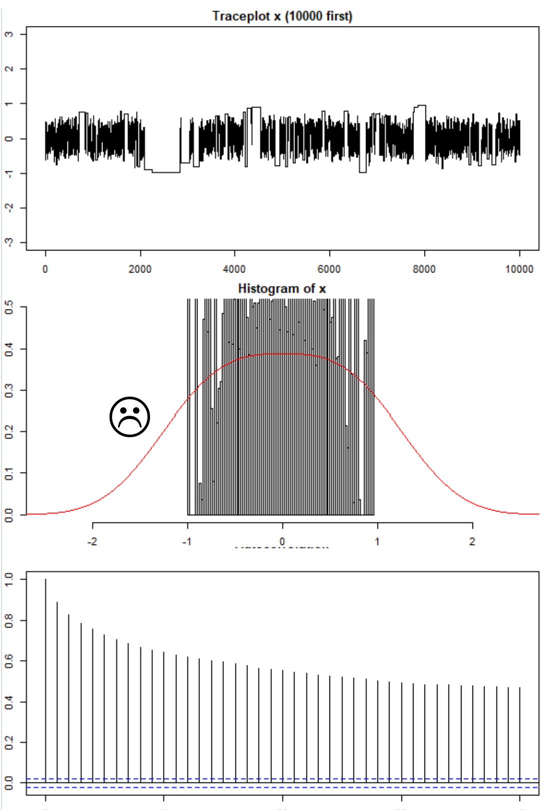
Acceptance rate= 0.9149915



N = 10 000

The effect variance in proposal distribution

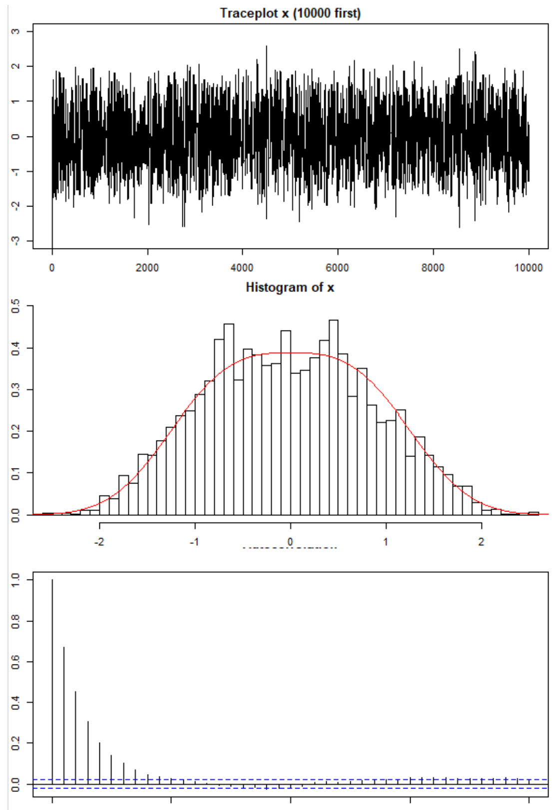
$g(y|x) = \phi(y; 0, 0.25^2)$



Acc. rate =0.419

Too narrow proposal,
good acceptance
high correlation ☹

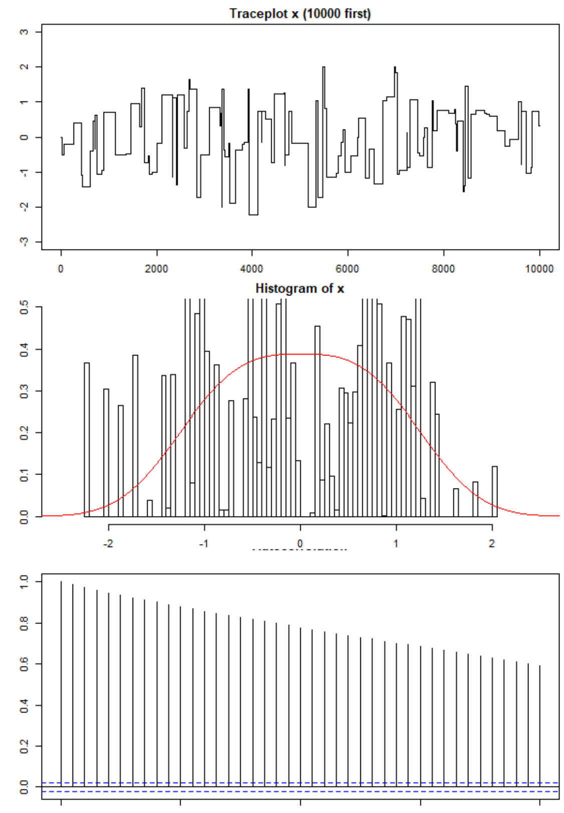
$g(y|x) = \phi(y; 0,4^2)$



Acc. Rate = 0.288

Just about right,
reasonable acceptance
low correlation ☺

$g(y|x) = \phi(y; 0, 100^2)$



Acc. Rate = 0.012

Too large changes proposed,
low acceptance
high correlation ☹

M-H and multivariate settings

- $\mathbf{X} = (X_1, \ldots, X_p)$
- Typical in this case: Only change one or a few components at a time.
  1. Choose index j (randomly)
  2. Sample $X_j^* \sim g_j(\cdot|\mathbf{x})$, put $X_k^* = X_k$ for $k \neq j$
  3. Compute
  $$R(\mathbf{x}, \mathbf{X}^*) = \frac{f(\mathbf{X}^*)g(\mathbf{x}|\mathbf{X}^*)}{f(\mathbf{x})g(\mathbf{X}^*|\mathbf{x})}$$
  4. Put
  $$\mathbf{Y} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{1, R(\mathbf{x}, \mathbf{X}^*)\} \\ \mathbf{x} & \text{otherwise} \end{cases}$$
- Can show that this version also satisfies detailed balance
- Can even go through indexes systematic
  - Should then consider the whole loop through all components as one iteration

Example multivariate with single coordinate update

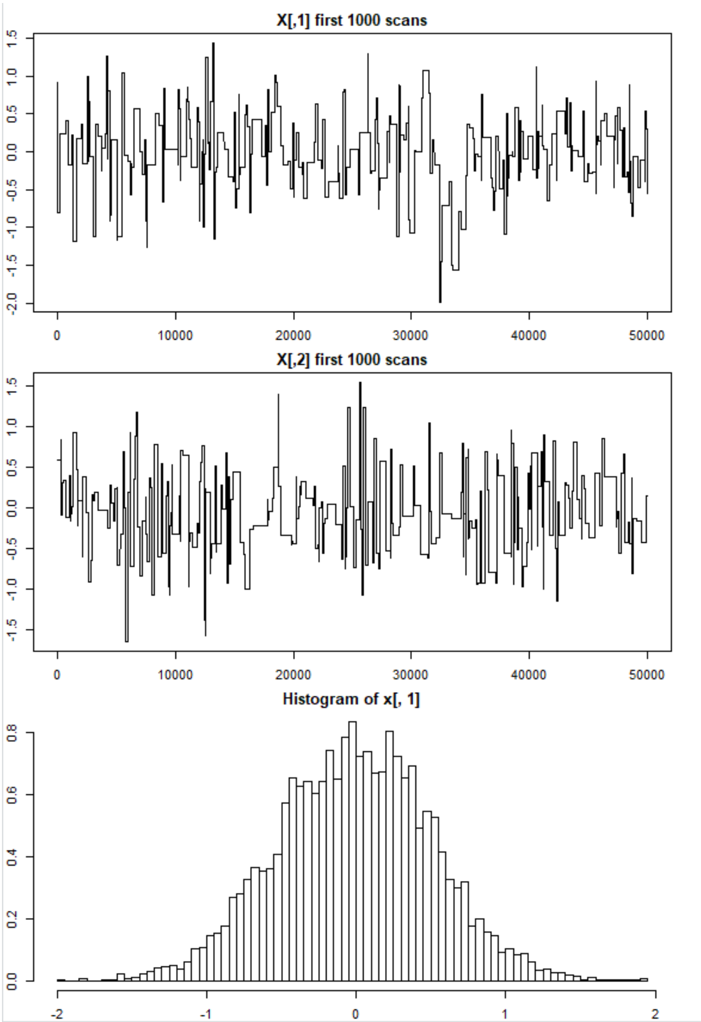
- Assume $f(\mathbf{x}) \propto \exp(-||\mathbf{x}||^3/3) = \exp(-[||\mathbf{x}||^2]^{3/2}/3)$
- Proposal distribution
  1. $j \sim \text{Uniform}[1, 2, ..., p]$
  2. $x_j^* \sim N(x_j, 1)$
- `Example_MH_cubic_multivariate.R`

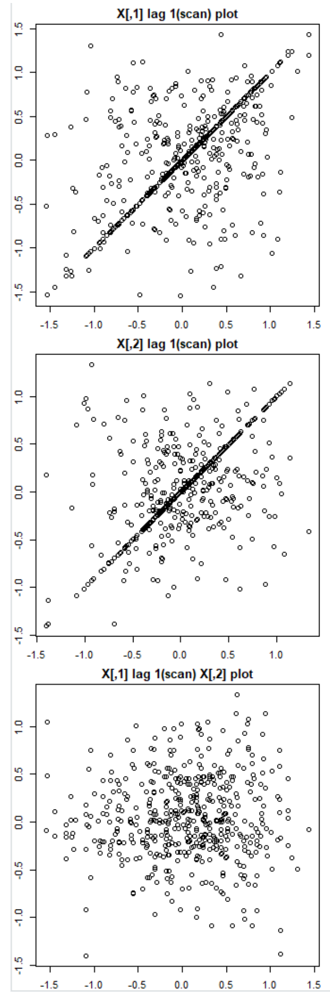
```
#Proposal distribution: Gaussian distribution centered at previous value
p = 50
N = 10000     # Number of iterations
x = matrix(nrow=N,ncol=p)

#Initial value
x[1,] = rnorm(p)
acc = 0
for(i in 2:N)
{
  j = sample(1:p,1)
  y = x[i-1,]
  y[j] = rnorm(1,x[i-1,j],2)
  R = f(y)*dnorm(x[i-1,j],y[j],1)/(f(x[i-1,])*dnorm(y[j],x[i-1,j],1))
  if(runif(1)<R)
  {
    x[i,] = y
    acc = acc+1
  }
  else
   x[i,] = x[i-1,]
}
```


See also fixed scan in: `Example_MH_cubic_multivariate_2.R`

Results independent





Acceptance rate= 0.3053943