Plan for today

- (very) short summary of Part1
- More on Bayesian statistics
  - Conjugate priors
  - Hierarchical Models

What have we done in Part 1 - Simulation

- Given a distribution $f(x)$
  - x may be a discrete or continuous stochastic variable
  - x may be a scalar or a vector
- Want to generate a sample $x \sim f(x)$,or iid $x_1, x_2, ..., x_n \sim f(x)$

What have we done in Part 1 - Simulation

- Given a distribution $f(x)$
  - x may be a discrete or continuous stochastic variable
  - x may be a scalar or a vector
- Want to generate a sample $x \sim f(x)$,or iid $x_1, x_2, ..., x_n \sim f(x)$
- We have discussed several simulation techniques:
  - probability integral transform (inversion method)
  - bivariate transformation (Box-Muller)
  - ratio-of-uniforms method
  - method based on mixtures
  - rejection sampling
  - (Importance sampling)

Why do we want to sample?

For a given function $g(x)$ we want to find:

$$\mu = \mathrm{E}[g(x)] = \int g(x)f(x)dx$$

- if we can find the integral analytically, we should do so
- if we can't solve the integral analytically we can estimate μ
  - generate iid $x_1, x_2, \ldots, x_n \sim f(x)$
  - estimate μ by

    $$\hat{\mu} = \frac{1}{2}\sum_{i=1}^{n} g(x_i)$$

  - then

    $$\mathrm{E}(\mu) = \mu \text{ and } \mathrm{Var}(\mu) = \frac{\mathrm{Var}(g(x))}{n}$$

  - so by choosing n large enough we may estimate μ with the precision we want

Why do we want to sample?

For a given function $g(x)$ we want to find:

$$\mu = \mathrm{E}[g(x)] = \int g(x)f(x)dx$$

- if we can find the integral analytically, we should do so
- if we can't solve the integral analytically we can estimate μ
  - generate iid $x_1, x_2, \ldots, x_n \sim f(x)$
  - estimate μ by

    $$\hat{\mu} = \frac{1}{2}\sum_{i=1}^{n} g(x_i)$$

  - then

    $$\mathrm{E}(\mu) = \mu \text{ and } \mathrm{Var}(\mu) = \frac{\mathrm{Var}(g(x))}{n}$$

  - so by choosing n large enough we may estimate μ with the precision we want

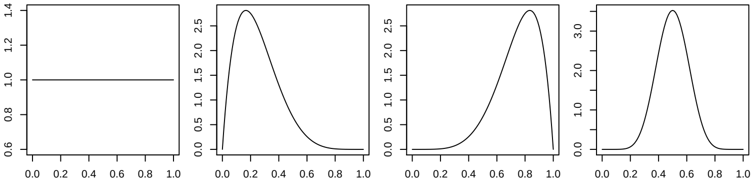
Can we sample from any $f(x)$ now??

What have we done in Part 1 -Bayesian Statistics

- Bayesian modelling: consider parameters as stochastic variables also when their value is not the result of a stochastic experiment
- A (toy) example:
  - I have a dice, let p: probability of getting a six
  - Consider p as a stochastic variable, you don't know it is a proper dice
  - what distribution would you assign to p?

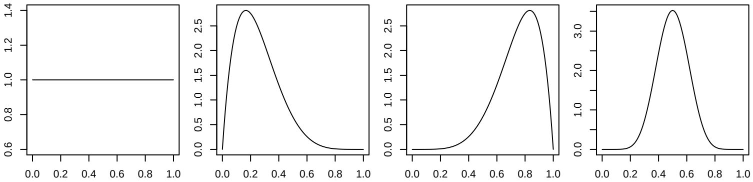
What have we done in Part 1 -Bayesian Statistics

- Bayesian modelling: consider parameters as stochastic variables also when their value is not the result of a stochastic experiment
- A (toy) example:
  - I have a dice, let p: probability of getting a six
  - Consider p as a stochastic variable, you don't know it is a proper dice
  - what distribution would you assign to p?



What have we done in Part 1 -Bayesian Statistics

- Bayesian modelling: consider parameters as stochastic variables also when their value is not the result of a stochastic experiment
- A (toy) example:
  - I have a dice, let p: probability of getting a six
  - Consider p as a stochastic variable, you don't know it is a proper dice
  - what distribution would you assign to p?



- We roll the dice n times, let x be the number of six

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

What have we done in Part 1 -Bayesian Statistics

- Likelihood Model:

$$f(x|p) = P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Prior Model:

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- Posterior Model:

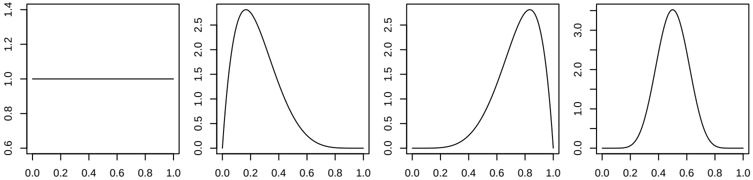
$$f(p|x) = \frac{f(x|p)f(p)}{\int f(x|p)f(p)\ dp} \propto f(x|p)f(p)$$

  - In this case:

$$f(p|x) \propto p^{\alpha+x-1}(1-p)^{\beta+n-x-1} = B(\alpha + x, \beta + n - x)$$

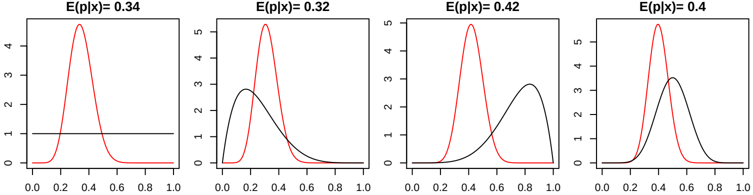
What have we done in Part 1 -Bayesian Statistics

- Before we observe x



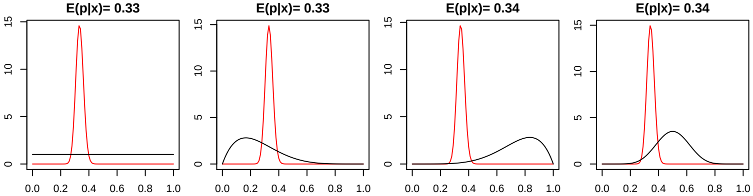
What have we done in Part 1 -Bayesian Statistics

- After observing $n = 30$ and $x = 10$



What have we done in Part 1 -Bayesian Statistics

- After observing $n = 300$ and $x = 100$



Interpretation of probability

- Frequentist (objective): Probability of event A is

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

  where m:number of times A occurres in n identical and independent trials.

- Bayesian (subjective): Probability of event A, $P(A)$, is a measure of someone's degree of belief in the occurrence of A.
  - different persons may have different $P(A)$

Prior and Posterior Distribution

- Prior distribution: $f(\theta)$
  - a measure of our belief about the value of θ before we have observed the data
  - based on prior information/experience
- Observation and Likelihood: $f(x|\theta)$
  - observed value x, and its probability distribution given θ
- Posterior distribution: $f(\theta|x)$
  - a measure of our belief about the of value of θ after we have observed the data x
  - based on prior information/experience and the observed data x

Prior and Posterior Distribution

- Prior distribution: $f(\theta)$
  - a measure of our belief about the value of θ before we have observed the data
  - based on prior information/experience
- Observation and Likelihood: $f(x|\theta)$
  - observed value x, and its probability distribution given θ
- Posterior distribution: $f(\theta|x)$
  - a measure of our belief about the of value of θ after we have observed the data x
  - based on prior information/experience and the observed data x
- Bayes theorem

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta)$$

Choice of prior distributions

- Under a uniform prior the posterior mode equals the MLE, as

$$f(\theta|x) \propto f(x|\theta)$$

- The prior distribution has to be chosen appropriately, which often causes concerns to practitioners.
- It should reflect the knowledge about the parameter of interest (e.g. a relative risk parameter in an epidemiological study).
- Ideally it should be elicited from experts.
- In the absence of expert opinions, simple informative prior distributions may still be a reasonable choice.

There have been various attempts to specify "non-informative" or "reference" priors to lessen the influence of the prior distribution.

Conjugate prior

Conjugate priors makes analytical evaluations easier...

Conjugate prior distribution

Let $\mathsf{L}_x(\theta) = f(x|\theta)$ denote a likelihood function based on the observation $X = x$. A class $\mathcal{G}$ of distributions is called conjugate with respect to $\mathsf{L}_x(\theta)$ if the posterior distribution $\mathsf{p}(\theta|x)$ is in $\mathcal{G}$ for all x whenever the prior distribution $\mathsf{p}(\theta)$ is in $\mathcal{G}$.

Conjugate prior - Example

- Binomial conjugate prior
  - $x|p \sim \text{Binom}(n, p)$
  - $p \sim \text{Beta}(\alpha, \beta)$
  - $p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$

Conjugate prior - Example

- Binomial conjugate prior
  - $x|p \sim \text{Binom}(n, p)$
  - $p \sim \text{Beta}(\alpha, \beta)$
  - $p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$
- Normal (mean) conjugate prior
  - $x_1, \ldots, x_n|p \sim \mathcal{N}(\mu, \sigma_0^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \tau^2)$
  - $\mu|x_1, \ldots, x_n \sim \mathcal{N}(\cdot, \cdot)$

Conjugate prior - Example

- Binomial conjugate prior
  - $x|p \sim \text{Binom}(n, p)$
  - $p \sim \text{Beta}(\alpha, \beta)$
  - $p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$
- Normal (mean) conjugate prior
  - $x_1, \ldots, x_n|p \sim \mathcal{N}(\mu, \sigma_0^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \tau^2)$
  - $\mu|x_1, \ldots, x_n \sim \mathcal{N}(\cdot, \cdot)$
- Normal (variance) conjugate prior
  - $x_1, \ldots, x_n|p \sim \mathcal{N}(\mu_0, \sigma^2)$
  - $\sigma^2 \sim (IG)(\alpha, \beta)$
  - $\sigma^2|x_1, \ldots, x_n \sim (IG)(\cdot, \cdot)$

List of conjugate prior distributions

Likelihood	Conjugate prior	Posterior distribution
$X \mid p \sim \text{Bin}(n, p)$	$p \sim \text{Be}(\alpha, \beta)$	$p \mid x \sim \text{Be}(\alpha + x, \beta + n - x)$
$X \mid p \sim \text{Geom}(p)$	$p \sim \text{Be}(\alpha, \beta)$	$p \mid x \sim \text{Be}(\alpha + 1, \beta + x - 1)$
$X \mid \lambda \sim \text{Po}(e \cdot \lambda)$	$\lambda \sim \text{G}(\alpha, \beta)$	$\lambda \mid x \sim \text{G}(\alpha + x, \beta + e)$
$X \mid \lambda \sim \text{Exp}(\lambda)$	$\lambda \sim \text{G}(\alpha, \beta)$	$\lambda \mid x \sim \text{G}(\alpha + 1, \beta + x)$
$X \mid \mu \sim \mathcal{N}(\mu, \sigma_\star^2)$	$\mu \sim \mathcal{N}(\nu, \tau^2)$	$\mu \mid x \sim \mathcal{N}\left[(A)^{-1}\left(\frac{x}{\sigma^2} + \frac{\nu}{\tau^2}\right), (A)^{-1}\right]$
$X \mid \sigma^2 \sim \mathcal{N}(\mu_\star, \sigma^2)$	$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\sigma^2 \mid x \sim \text{IG}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$

$_\star$: known.

$A = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$

Conditional Conjugacy

The use of conjugate priors become difficult when the models gets more complex....

Hierarchical Bayesian models

Hierarchical models are an extremely useful tool in Bayesian model building.

Three parts:

- Observation model $\boldsymbol{y}|\boldsymbol{x}$: Encodes information about observed data.
- The latent model $\boldsymbol{x}|\boldsymbol{\theta}$: The unobserved process.
- Hyperpriors for $\boldsymbol{\theta}$: Models for all of the parameters in the observation and latent processes.

Note: here we indicate the observed data by $\boldsymbol{y}$ while $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are parameters

Hierarchical Bayesian models - A simple example

Example from George et al. (1993) regarding the analysis of 10 power plants.

- y_i number of observed failures of pump $i = 1, \ldots, 10$
- t_i length of operation time of pump $i = 1, \ldots, 10$ (in 1000 hours)

Hierarchical Bayesian models - A simple example

Example from George et al. (1993) regarding the analysis of 10 power plants.

- y_i number of observed failures of pump $i = 1, \ldots, 10$
- t_i length of operation time of pump $i = 1, \ldots, 10$ (in 1000 hours)

Model:

$$y_i \mid \lambda_i \sim \text{Po}(\lambda_i t_i)$$

Hierarchical Bayesian models - A simple example

Example from George et al. (1993) regarding the analysis of 10 power plants.

- y_i number of observed failures of pump $i = 1, \dots, 10$
- t_i length of operation time of pump $i = 1, \dots, 10$ (in 1000 hours)

Model:

$$y_i \mid \lambda_i \sim \mathrm{Po}(\lambda_i t_i)$$

Conjugate prior for λ_i:

$$\lambda_i \mid \alpha, \beta \sim \mathrm{G}(\alpha, \beta)$$

Hierarchical Bayesian models - A simple example

Example from George et al. (1993) regarding the analysis of 10 power plants.

- y_i number of observed failures of pump $i = 1, \ldots, 10$
- t_i length of operation time of pump $i = 1, \ldots, 10$ (in 1000 hours)

Model:

$$y_i \mid \lambda_i \sim \mathrm{Po}(\lambda_i t_i)$$

Conjugate prior for λ_i:

$$\lambda_i \mid \alpha, \beta \sim \mathrm{G}(\alpha, \beta)$$

Hyper-prior on α and β:

$$\alpha \sim \mathrm{Exp}(1.0) \qquad \beta \sim \mathrm{G}(0.1, 1)$$

Hierarchical Bayesian models - A simple example

Example from George et al. (1993) regarding the analysis of 10 power plants.

- y_i number of observed failures of pump $i = 1, \ldots, 10$
- t_i length of operation time of pump $i = 1, \ldots, 10$ (in 1000 hours)

Model:

$$y_i \mid \lambda_i \sim \text{Po}(\lambda_i t_i)$$

Conjugate prior for λ_i:

$$\lambda_i \mid \alpha, \beta \sim \text{G}(\alpha, \beta)$$

Hyper-prior on α and β:

$$\alpha \sim \text{Exp}(1.0) \qquad \beta \sim \text{G}(0.1, 1)$$

Posterior of interest:

$$f(\alpha, \beta, \lambda_1, \ldots, \lambda_{10} | y_1, \ldots, y_{10})$$

Hierarchical Bayesian models - A simple example

Posterior of Interest

$$f(\alpha, \beta, \lambda_1, \ldots, \lambda_{10} | y_1, \ldots, y_{10}) \propto$$

$$\left[\prod_{i=1}^{10} (\lambda_i t_i)^{y_i} e^{-\lambda_i t_i}\right] \times \left[\prod_{i=1}^{10} \frac{\beta^\alpha}{\Gamma(\beta)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i}\right] \times \alpha e^{-\alpha} \times \beta^{-0.9} e^{-\beta}$$

Hierarchical Bayesian models - A simple example

Posterior of Interest

$$f(\alpha, \beta, \lambda_1, \ldots, \lambda_{10} | y_1, \ldots, y_{10}) \propto$$
$$\left[\prod_{i=1}^{10} (\lambda_i t_i)^{y_i} e^{-\lambda_i t_i}\right] \times \left[\prod_{i=1}^{10} \frac{\beta^\alpha}{\Gamma(\beta)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i}\right] \times \alpha e^{-\alpha} \times \beta^{-0.9} e^{-\beta}$$

Can we sample from this distribution?

Markov chain Monte Carlo

- Goal: Generation of samples or approximation of an expected value for a (possibly high-dimensional) density $\pi(x)$.
- Application of ordinary Monte Carlo methods is difficult.
- Idea: Use Markov chain theory to build a process that converges to our target distribution!

Idea of Markov chain Monte Carlo

- Contruct a Markov chain $\{X_i\}_{i=0}^{\infty}$ such that

$$\lim_{i \to \infty} P(X_i = x_i) = f(x)$$

- Simulate the Markov chain for many iterations
- For large enough m the samples $x_{m+1}, x_{m+2}, \ldots$ are (essentially) samples from $f(x)$
- Estimate $\mu = \mathrm{E}_f[g(x)] = \int g(x) f(x) dx$ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=m}^{m+n} g(x_i)$$

we have that $E[\hat{\mu}] = \mu$ and $\mathrm{Var}\, \hat{\mu} = ?$

Idea of Markov chain Monte Carlo

- Contruct a Markov chain $\{X_i\}_{i=0}^{\infty}$ such that

$$\lim_{i \to \infty} P(X_i = x_i) = f(x)$$

How do we contruct such Markov Chain?

- Simulate the Markov chain for many iterations
- For large enough m the samples $x_{m+1}, x_{m+2}, \ldots$ are (essentially) samples from $f(x)$
- Estimate $\mu = \mathrm{E}_f[g(x)] = \int g(x) f(x) dx$ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=m}^{m+n} g(x_i)$$

we have that $E[\hat{\mu}] = \mu$ and $\mathrm{Var}\,\hat{\mu} = ?$

Idea of Markov chain Monte Carlo

- Contruct a Markov chain $\{X_i\}_{i=0}^{\infty}$ such that

$$\lim_{i \to \infty} P(X_i = x_i) = f(x)$$

- Simulate the Markov chain for many iterations How do we simulate frm such Markov Chain?
- For large enough m the samples $x_{m+1}, x_{m+2}, \ldots$ are (essentially) samples from $f(x)$
- Estimate $\mu = \mathrm{E}_f[g(x)] = \int g(x) f(x) dx$ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=m}^{m+n} g(x_i)$$

we have that $E[\hat{\mu}] = \mu$ and $\mathrm{Var}\, \hat{\mu} = ?$

Idea of Markov chain Monte Carlo

- Contruct a Markov chain $\{X_i\}_{i=0}^{\infty}$ such that

$$\lim_{i\to\infty} P(X_i = x_i) = f(x)$$

- Simulate the Markov chain for many iterations
- For large enough m the samples $x_{m+1}, x_{m+2}, \ldots$ are (essentially) samples from $f(x)$
- Estimate $\mu = \mathrm{E}_f[g(x)] = \int g(x)f(x)dx$ as

$$\hat{\mu} = \frac{1}{n}\sum_{i=m}^{m+n} g(x_i)$$

How do we know m is large enough?

we have that $E[\hat{\mu}] = \mu$ and $\mathrm{Var}\,\hat{\mu} =?$

Review: Discrete-time Markov chains

A Markov chain is a discrete-time stochastic process $\{X_i\}_{i=0}^{\infty}$, $X_i \in S$, where given the present state, past and future states are independent (Markov assumption):

$$P(X_{i+1} = x_{i+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_i = x_i) = P(X_{i+1} = x_{i+1} \mid X_i = x_i)$$

Review: Markov chains

A Markov chain with stationary transition probabilities can be specified by:

- the initial distribution $P(X_0 = x_0) = g(x_0)$
- the transition matrix

$$P(y \mid x) = P(X_{i+1} = y \mid X_i = x) \quad [= P_{xy}]$$

Review: Markov chains

Theorem: A Markov chain has a unique limiting distribution $\pi(x)$ if the chain is irreducible, aperiodic, and positive recurrent.

If so, the limiting distribution $\pi(x) = \lim_{i \to \infty} P(X_i = x)$ is given by

$$\begin{aligned} \pi(y) &= \sum_{x \in S} \pi(x) P(y \mid x) \quad \text{for all } y \in S \\ \sum_{x \in S} \pi(x) &= 1 \end{aligned} \tag{1}$$

Detailed Balance

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

$$\pi(x)P(y \mid x) = \pi(y)P(x \mid y) \quad \text{for all } x, y \in S \tag{2}$$

Proof: on blackboard

Detailed Balance

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

$$\pi(x)P(y \mid x) = \pi(y)P(x \mid y) \quad \text{for all } x, y \in S \tag{2}$$

Proof: on blackboard

This gives a time-reversible Markov chain.

- In a reversible MC we cannot distinguish the direction of simulation from inspecting a realisation of the chain (even if we know the transition matrix).
- Most MCMC algorithms are based on reversible Markov chains.

Problem statement

In stochastic processes course: The Markov chain is given, i.e. $P(y \mid x)$ is given, find $\pi(x)$.

Problem statement

In stochastic processes course: The Markov chain is given, i.e. $P(y \mid x)$ is given, find $\pi(x)$.

Now: $\pi(x)$, $x \in S$ is given, want to find $P(y \mid x)$, $x, y \in S$ so that

$$\pi(y) = \sum_{x \in S} \pi(x) P(y \mid x) \quad \text{for all } y \in S$$

$$\sum_{x \in S} \pi(x) = 1$$

Problem statement

In stochastic processes course: The Markov chain is given, i.e. $P(y \mid x)$ is given, find $\pi(x)$.

Now: $\pi(x)$, $x \in S$ is given, want to find $P(y \mid x)$, $x, y \in S$ so that

$$\pi(y) = \sum_{x \in S} \pi(x) P(y \mid x) \quad \text{for all } y \in S$$

$$\sum_{x \in S} \pi(x) = 1$$

However, # unknowns: $|S| \cdot (|S| - 1)$; # equations: $|S|$.

Problem statement

In stochastic processes course: The Markov chain is given, i.e. $P(y \mid x)$ is given, find $\pi(x)$.

Now: $\pi(x)$, $x \in S$ is given, want to find $P(y \mid x)$, $x, y \in S$ so that

$$\pi(y) = \sum_{x \in S} \pi(x) P(y \mid x) \quad \text{for all } y \in S$$

$$\sum_{x \in S} \pi(x) = 1$$

However, # unknowns: $|S| \cdot (|S| - 1)$; # equations: $|S|$.

$\Rightarrow$ many solutions exist – we want one!

(Note: $|S|$ can be huge, so solving this as a matrix equation is not possible.)

Idea

Focus on (2) the detailed balance condition instead. We want to find $P(y \mid x)$ that solves

$$\pi(x)P(y \mid x) = \pi(y)P(x \mid y) \quad \text{for all } x, y \in S$$

Here, we still have many solutions. However, we do not need a general solution, one (good) solution is enough.

We show how to generate an irreducible, aperiodic and pos. recurrent Markov chain with arbitrary limiting distribution $\pi(x)$. (never as good as iid samples but much wider applicability)

A possible solution

Let's see if this work:

$$P(y|x) = \begin{cases} Q(y|x)\ \alpha(y|x) & \text{if } y \neq x \\ 1 - \sum_{y \neq x} Q(y|x)\ \alpha(y|x) & \text{if } y = x \end{cases}$$

where :

- $Q(y|x)$ is a proposal density
- $\alpha(y|x)$ is the probability of accepting the move

Metropolis-Hastings algorithm

Setting: We want to sample from some distribution

$$\pi(x) = \frac{\tilde{\pi}(x)}{c}$$

where c is the normalising constant. How about this?

Draw initial state $X_0 \sim g(x_0)$
for $i = 0, 1, \ldots$ **do**
    Propose a potential new state y from $Q(y|x_{i-1})$
    Compute the acceptance probability $\alpha(y|x_{i-1})$
    Draw $u \sim \text{Unif}(0, 1)$
    if $u < \alpha(y|x_{i-1})$ **then**
        Set $x_i = y$ (ie accept y)
    else
        Set $x_i = x_{i-1}$ (ie reject y)
    end if
end for

How to choose α so that the detailed balance condition hold?

- Assume we have a proposal $Q(y|x)$
- What should $\alpha(y|x)$ be for the detailed balance condition to hold?

See Blackboard!

Acceptance step

- In the acceptance step the proposal y is accepted with probability α as new value of the Markov chain.
- This is similar to rejection sampling. However, here no constant c needs to be determined.
- Further, if we reject, then we retain the sample.

History of Metropolis-Hastings

- The algorithm was presented 1953 by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller from the Los Alamos group. It is named after the first author Nicholas Metropolis.
- W. Keith Hastings extended it to the more general case in 1970.
- It was then ignored for a long time.
- Since 1990 it has been used more intensively.

Toy example

We consider the Poisson distribution

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

Choose proposal kernel

- If $x = 0$

$$Q(y|0) = \begin{cases} \frac{1}{2} & \text{for} \quad y \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

- For $x > 0$

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

Toy example

- If $x = 0$

$$\alpha(0|0) = \min\{1, 1\} = 1$$
$$\alpha(1|0) = \min\{1, 10\} = 1$$

- If $x > 0$

$$\alpha(x-1|x) = \min\left\{1, \frac{\frac{10^{x-1}}{(x-1)!}e^{-10}}{\frac{10^{x}}{(x)!}e^{-10}} \cdot \frac{\frac{1}{2}}{\frac{1}{2}}\right\} = \min\left\{1, \frac{x}{10}\right\} \tag{3}$$

$$\alpha(x+1|x) = \min\left\{1, \frac{\frac{10^{x+1}}{(x+1)!}e^{-10}}{\frac{10^{x}}{(x)!}e^{-10}} \cdot \frac{\frac{1}{2}}{\frac{1}{2}}\right\} = \min\left\{1, \frac{10}{x+1}\right\} \tag{4}$$

From (3) we see that $\alpha = 1$ if $x > 9$ and $x/10$ else.

From (4) we see that $\alpha = 1$ if $x \leq 9$ and $10/(x+1)$ else.

Toy example

Note this gives for $x > 0$:

$$P(x-1|x) = \frac{1}{2}\min\left\{1, \frac{x}{10}\right\} = \begin{cases} \frac{x}{20} & \text{for} \quad x \leq 9 \\ \frac{1}{2} & \text{for} \quad x > 9 \end{cases}$$

$$P(x+1|x) = \frac{1}{2}\min\left\{1, \frac{10}{x+1}\right\} = \begin{cases} \frac{1}{2} & \text{for} \quad x \leq 9 \\ \frac{5}{x+1} & \text{for} \quad x > 9 \end{cases}$$

$P(x|x)$ follows directly.

(For $x = 0$ we have $P(0|0) = 1/2$ and $P(1|0) = 1/2$).

However, we do not have to compute these values! (Show R-code `demo_toyMCMC2.R`)

What about

- Irreducible: Must be checked in each case. Must choose $Q(y \mid x)$ so that this is ok.

What about

- Irreducible: Must be checked in each case. Must choose $Q(y \mid x)$ so that this is ok.
- Aperiodic: Sufficient that $P(x \mid x) > 0$ for one $x \in S$, so sufficient that $\alpha(y \mid x) < 1$ for one pair $y, x \in S$.

What about

- Irreducible: Must be checked in each case. Must choose $Q(y \mid x)$ so that this is ok.
- Aperiodic: Sufficient that $P(x \mid x) > 0$ for one $x \in S$, so sufficient that $\alpha(y \mid x) < 1$ for one pair $y, x \in S$.
- Positive recurrent: for finite S, irreducibility is sufficient. More difficult in general, but if Markov chain is not recurrent we will see this as drift in the simulations. (In practice usually no problem).

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(y|x)$.

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(y|x)$.
- However, the speed of convergence and the dependence between the successive samples depends strongly on the proposal distribution.

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(y|x)$.
- However, the speed of convergence and the dependence between the successive samples depends strongly on the proposal distribution.
- Since we only need to compute the ratio $\pi(y)/\pi(x)$, the proportionality constant is irrelevant.

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(y|x)$.
- However, the speed of convergence and the dependence between the successive samples depends strongly on the proposal distribution.
- Since we only need to compute the ratio $\pi(y)/\pi(x)$, the proportionality constant is irrelevant.
- Similarly, we only care about $Q(.)$ up to a constant.

Remarks on the Metropolis-Hastings algorithm

- Under some regularity conditions it can be shown that the Metropolis-Hasting algorithm converges to the target distribution regardless of the specific choice of $Q(y|x)$.
- However, the speed of convergence and the dependence between the successive samples depends strongly on the proposal distribution.
- Since we only need to compute the ratio $\pi(y)/\pi(x)$, the proportionality constant is irrelevant.
- Similarly, we only care about $Q(.)$ up to a constant.
- Often it is advantageous to calculate the acceptance probability on log-scale, which makes the computations more stable.