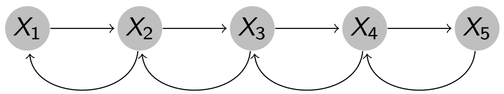
Lecture 8: Markov chain Monte Carlo

- 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.
- However, Markov chain Monte Carlo (MCMC) methods will then be a useful alternative.

en.wikipedia.org/wiki/Markov_chain

Andrey Markov (1856 – 1922), Russian mathematician.

Markov chain:



Given the previous observation X_{i-1}, X_i is independent of the sequence of events that preceded it.

Idea of Markov chain Monte Carlo

Idea

Simulate a Markov chain $X_1, \dots, X_i, \dots$, which is designed in a way such that $\mathrm{P}(X_i = x)$ converges to the target distribution $\pi(x)$, e.g. the posterior distribution.

Properties:

- After convergence, one obtains random samples from the target distribution, which can be used to estimate posterior characteristics.
- Samples will typically be dependent.

Central algorithms:

- Metropolis-Hasting algorithm
- Gibbs sampling

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):

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

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

$$P(x^\star \mid x) = P(X_{i+1} = x^\star \mid X_i = x) \quad [= P_{xx^\star}]$$

Review: Markov chains

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

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

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

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

which gives a time-reversible Markov chain.

Reversible Markov chains

- 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. $P(x^\star \mid x)$ is given, find $\pi(x)$.

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

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

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 $P(x^\star \mid x)$ that solves

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

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)

Metropolis 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 $x^\star$ from a symmetric[1]
    proposal distribution so that $\mathrm{P}(X^\star = x^\star) = Q(x^\star | x_{i-1})$
  Accept $x^\star$ as new state $X_i = x^\star$ with probability

$$\alpha(x^\star \mid x_{i-1}) = \min\left(1, \frac{\tilde{\pi}(x^\star)}{\tilde{\pi}(x_{i-1})}\right)$$

  Otherwise stay at current state and set $X_i = x_{i-1}$.
end for

[1] $Q(y|x) = Q(x|y)$ for each pair of states x and y

Does the detailed balance condition hold?

Metropolis-Hastings algorithm

Use an "asymmetric" proposal distribution (also called proposal kernel)

Init $x_0 \sim g(x_0)$
for $i = 1, 2, \ldots$ **do**
  Generate a proposal $x^\star \sim Q(x^\star | x_{i-1})$
  $u \sim \mathrm{U}(0, 1)$
  if $u < \underbrace{\min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})} \times \underbrace{\frac{Q(x_{i-1}|x^\star)}{Q(x^\star|x_{i-1})}}_{\text{Proposal ratio}}\right)}_{\text{Acceptance probability } \alpha}$ **then**
    $x_i \leftarrow x^\star$
  else
    $x_i \leftarrow x_{i-1}$
  end if
end for

Acceptance step

- In the acceptance step the proposal $x^\star$ 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.

[Exercise: check that detailed balance condition holds].

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

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

Choose proposal kernel

- If $x = 0$

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

- For $x > 0$

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

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 $Q(x^\star \mid x)$ so that this is ok.
- Aperiodic: Sufficient that $P(x \mid x) > 0$ for one $x \in S$, so sufficient that $\alpha(x^\star \mid x) < 1$ for one pair $x^\star, x \in S$.
- 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 $Q(x|x_{i-1})$.
- 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 $\pi(x^\star)/\pi(x_{i-1})$, 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.

Special cases of the Metropolis-Hastings algorithm

Depending on the choice of $Q(x^\star|x)$ different special cases result. In particular, two classes are important

- The independence proposal
- The Metropolis algorithm

Independence proposal

- The proposal distribution does not depend on the current value x_{i-1}

$$Q(x|x_{i-1}) = Q(x).$$

- $Q(x)$ is an approximation to $\pi(x)$.
- The sampler is closer to rejection sampler. However, here if we reject, then we retain the sample.

Experience:

- Performance is either very good or very bad, usually very bad.
- The tails of the proposal distribution should be at least as heavy as the tails of the target distribution.

The Metropolis algorithm

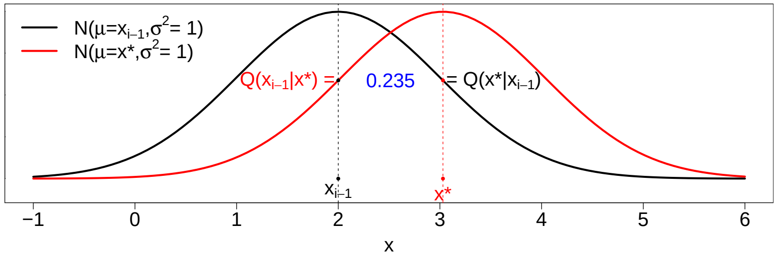
The proposal density is symmetric around the current value, that means

$$Q(x_{i-1}|x^\star) = Q(x^\star|x_{i-1}).$$

Hence,

$$\alpha = \min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})} \times \frac{Q(x_{i-1}|x^\star)}{Q(x^\star|x_{i-1})}\right) = \min\left(1, \frac{\pi(x^\star)}{\pi(x_{i-1})}\right)$$

A particular case is the random walk proposal, defined as the current value x_{i-1} plus a random variate of a 0-centred symmetric distribution.



Examples for random walks proposal

Assume x is scalar.

Then all proposal kernels, which add a random variable generated from a zero-symmetrical distribution to the current value x_{i-1}, are random walk proposals. For example:

$$x^\star \sim \mathcal{N}(x_{i-1}, \sigma^2)$$

$$x^\star \sim t_\nu(x_{i-1}, \sigma^2)$$

$$x^\star \sim \mathcal{U}(x_{i-1} - d, x_{i-1} + d)$$