Markov Chain Monte Carlo Algorithms

Peadar Coyle

1. General state space Markov chains

Most applications of Markov Chain Monte Carlo algorithms (MCMC) are concerned with continuous random variables, i.e the corresponding Markov chain has a continuous state space S. In this section we will give a brief overview of the theory underlying Markov chains with general state spaces. Although the basic principles are not much different from the discrete case, the study of general state Markov chainns involves many more technicalities and subtleties. Though this section is concerned with general state spaces we will notationally assume that the state space is $S = \mathbb{R}^d$ We need a definition of a Markov chain, to be a stochastic process in which, conditionally on the present, the past and the future are independent. In the discrete case we formalised this idea using the conditional probability of $X_t = j$ given different collections of past events.

. In a general state space it can be that all events of the type $X_t = j$ have probability 0, as it is the case for a process with a continuous state space. A process with a continuous state space spreads the probability so thinly that the probability of hitting one given state is 0 for all states. Thus we have to work with conditional probabilities of sets of states, rather than individual states.

Markov chain . Let X be a stochastic process in discrete time with general state space S. X is called a Markov chain if X satisfies the Markov property

$$\mathbb{P}(X_{t+1} \in A | X_0 = x_0, \cdots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t) \tag{1}$$

for all measurable sets $A \subset S$.

In the following we will assume that the Markov chain is *homogeneous*. i.e. the probabilities $\mathbb{P}\ (X_{t+1} \in A | X_t = x_t)$ are independent of t. For the remainder of this section we shall also assume that we can express the probability from definition 1 using a *transition kernel* $K : S \times S \to \mathbb{R}_0^+$:

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \int_A K(x_t, x_{t+1}) dx_{t+1} \tag{2}$$

where the integration is with respect to a suitable dominating measure, i.e. for example with respect to the Lebesgue measure if S = $\mathbb{R}^d$. The transition kernel K(x,y) is thus just the conditional probability density of X_{t+1} given $X_t = x_t$. We obtain the special case of the definition of a transition kernel.

Definition The matrix $\mathbf{K} = (k_{ij})_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$ is called the transition kernel (or transition matrix) of the homogeneous Markov chain X.

We will see that together with the initial distribution, whcih we might write as a vector $\lambda_{\mathbf{0}} = (\mathbb{P}(X_0 = i)_{(i \in S)}$ the transition kernel $\mathbf{K}$ fully specifies the distribution of a homogeneous Markov chain. However, we start by stating two basic properties of the transition kernel $\mathbf{K}$:

- The entries of the transition kernel are non-negative (they are probabilities).
- Each row of the kernel sums to 1, as

$$\sum_j k_{ij} = \sum_j \mathbb{P}(X_{t+1} = j | X_t = i) = \mathbb{P}(X_{t+1} \in S | X_t = i) = 1 \qquad (3)$$

We obtain the special case of definition 1.8 by setting K(i,j) = k_{ij}, where k_{ij} is the (i,j)-th element of the transition matrix $\mathbb{K}$. For a discrete state space the dominating measure is the counting measure, so integration just corresponds to summation, i.e. equation 3 is equivalent to

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \sum_{x_{t+1} \in A} k_{x_t, x_{t+1}}$$

We have for measurable set $A \subset S$ that

$$\mathbb{P}(X_{t+m} \in A | X_t = x_t) = \int_A \int_S \cdots \int_S K(x_t, x_{t+1} K(x_{t+1}, x_{t+2}) \cdots K(x_{t+m-1}, x_{t+m}) dx_{t+1} \cdots dx_{t+m-1} dx_{t+m},$$

thus the m-step transition kernel is

$$K^{(m)}(x_0, x_m) = \int_S \cdots \int_S K(x_0, x_1) \cdots K(x_{m-1}, x_m) dx_{m-1} \cdots dx_1$$

The m-step transition kernel allows for expressing the m-step transition probabilities more conveniently:

$$\mathbb{P}(X_{t+m} \in A | X_t = x_t) = \int_A K^{(m)}(x_t, x_{t+m}) dx_{t+m}$$

Let us consider an example.

Example Consider the Gaussian random walk on $\mathbb{R}$. Consider the random walk on $\mathbb{R}$ defined by

$$X_{t+1} = X_t + E_t$$

where $E_t \cong N(0,1)$, i.e. the probability density function of E_t is $\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})$. This is equivalent to assuming that $X_{t+1} | X_t = x_t \cong N(x_t, 1)$ We also assume that E_t is independent of $X_0, E_1, \cdots, E_{t-1}$. Suppose that

$X_0 \cong N(0,1)$. In contrast to the random walk on $\mathbb{Z}$ the state space of the Gaussian random walk is $\mathbb{R}$. We have that

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t, \cdots X_0 = x_0) = \mathbb{P}(E_t \in A - x_t | X_t = x_t, \cdots, X_0 = x_0)$$

$$= \mathbb{P}(E_t \in A - x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t), \text{ where A - } x_t = a - x_t : a \in A.$$

Thus X is indeed a Markov chain. Furthermore we have that

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \mathbb{P}(E_t \in A - x_t) = \int_A \phi(x_{t+1} - x_t) dx_{t+1}$$

Thus the transition kernel (which is nothing other than the conditional density of $X_{t+1}|X_t = x_t$) is thus

$$K(x_t, x_{t+1}) = \phi(x_{t+1} - x_t)$$

To find the m-step transition kernel we could use equation 2. However, the resulting integral is difficult to compute. Rather we exploit the fact that

$$X_{t+m} = X_t + \boxed{E_t + \cdots + E_{t+m-1}},$$

where the boxed formula is approximately $N(0, m)$ thus we can write $X_{t+m}|X_t$ $N(x_t, m)$

$$\mathbb{P}(X_{t+m} \in A | X_t = x_t) = \mathbb{P}(X_{t+m} - X_t \in A - x_t) = \int_A \frac{1}{\sqrt{m}} \phi(\frac{x_{t+m} - x_t}{\sqrt{m}}) dx_{t+m}$$

Comparing this with 2 we can identify

$$K^{(m)}(x_t, x_{t+m}) = \frac{1}{\sqrt{m}} \phi(\frac{x_{t+m} - x_t}{\sqrt{m}}$$

as m-step transition kernel □

We need the powerful probabilistic notion of irreducibility.

Irreducibility Given a distribution μ on the states S, a Markov chain is said to be μ-irreducible if for all sets A with $\mu(A) > 0$ and for all $x \in S$, there exists an $m \in \mathbb{N}_0$ such that

$$\mathbb{P}(X_{t+m} \in A | X_t = x) = \int_A K^{(m)}(x, y) dy > 0$$

If the number of steps m=1 then for all A, then the chain is said to be strongly μ-irreducible.

Example In the example above we had that $X_{t+1}|X_t = x_t \cong N(x_t, 1)$. As the range of the Gaussian distribution is $\mathbb{R}$, we have that $\mathbb{P}(X_{t+1} \in A | X_t = x_t)$ ¿ 0 for all sets A of non-zero Lebegue measure. Thus the chain is strongly irreducible with the respect to any continuous distribution. □

We can extend the concepts of periodicity, recurrence, and transience from the discrete case to the general case. However this requires additional technical concepts like *atoms* and *small sets* one can see 'Robert and Casella, 2004' for a rigorous treatment of these concepts. Let us define a recurrent discrete Markov chain.

Definition A discrete Markov chain is recurrent, if all states (on average) are visited inifinitely often.

For more general state spaces, we need to consider the number of visits to a set of states rather than single states. Let $V_A = \sum_{t=0}^{+\infty} \mathbb{1}_{X_t \in A}$ be the number of visits the chain makes to states in the set $A \subset S$. We then define the expected number of visits in $A \subset S$, when we start the chain in $x \in S$:

$$\mathbb{E}(V_A | X_0 = x) = \mathbb{E}(\sum_{t=0}^{+\infty} \mathbb{1}_{X_t \in A} | X_0 = x) = \sum_{t=0}^{+\infty} \int_A K^{(t)}(x, y) dy$$

This allows us to define recurrence for general state spaces. We start with defining recurrence of sets before extending the definition of recurrence of an entire Markov chain.

Definition (a) A set A $\subset S$ is said to be recurrent for a Markov chain X if for all $x \in A$

$$\mathbb{E}(V_A | X_0 = x) = +\infty,$$

(b) A Markov chain to be recurrent, if

- The chain is μ-irreducible for some distribution μ.
- Every measurable set $A \subset S$ with $\mu(A) > 0$ is recurrent.

. According to the definition a set is recurrent if on average it is visited infinitely often. This is already the case if there is a non-zero probability of visiting the set infinitely often. A stronger concept of recurrence can be obtained if we require that the set is visited infinitely often with probability 1. This type of recurrence is referred to as *Harris recurrence.*

Harris Recurrence . (a) A set $A \subset S$ is said to be Harris-recurrent for a Markov chain X if for all $x \in A$ $\mathbb{P}(V_A = +\infty | X_0 = x) = 1$, (b) A Markov chain is said to be Harris-recurrent, if

- The chain is μ-irreducible for some distribution μ.
- Every measurable set $A \subset S$ with $\mu(A) > 0$ is Harris-recurrent.

It is easy to see that Harris-recurrence implies recurrence. For discrete state spaces the two concepts are equivalent. Checking recurrence or Harris recurrence can be very difficult. We will state (without) proof a proposition which establishes that if a Markov chain is irreducible and has a unique invariant distribution, then the chain is also recurrent.

. However, before, we can state this proposition, we need to define invariant distributions for general state spaces.

Definition (Invariant Distribution). A distribution μ with density function f_μ is said to be the invariant distribution of a Markov chain X with transition kernel K if

$$f_\mu(y) = \int_S f_\mu(x)K(x,y)dx$$

for almost all $y \in S$.

Proposition 1.1. *Suppose that X is a μ-irreducible Markov chain having μ as unique invariant distibution. Then X is also recurrent.*

Checking the invariance condition of definition1 requires computing an integral, but this can be cumbersome, so an alternative condition is the simpler (sufficient but not necessary) condition of detailed balance.

Detailed balance . A transition kernel K is said to be in detailed balance with a distribution μ with denisity f_μ if for almost all x,y $\in S$

$$f_\mu(x)K(x,y) = f_\mu(y)K(y,x).$$

In complete analogy with theorem 1.22 one can also show in the general case that if the transition kernel of a Markov chain is in detailed balance with a distribution μ, then the chain is time-reversible and has μ as its invariant distribution.

1.1. Ergodic theorems

In this section we will study the question of whether we can use observations from a Markov chain to make inferences about its invariant distribution. We will see that under some regularity conditions it is even enough to follow a single sample path of the Markov chain.

. For independently identically distributed data the Law of Large Numbers is used to justify estimating the expected value of a functional using empirical averages. A similar result can be obtained for Markov chains. This result is the reason why MCMC methods work: it allows us to set up simulation algorithms to generate a Markov chain, whose sample path we can then use for estimating various quantities of interest.

Theorem 1.2 (Ergodic Theorem). . *Let X be a μ-irreducible, recurrent $\mathbb{R}^d$-valued Markov chain with invariant distribution μ. Then we have for any integrable function $g : \mathbb{R}^d \to \mathbb{R}$ that with probability 1*

$$lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} g(X_i) \to \mathbb{E})\mu(g(X)) = \int_S g(x)f_\mu(x)dx$$

for almost every starting value $X_0 = x$. If X is Harris-recurrent this holds for every starting value x.

Proof For a proof see (Roberts and Rosenthal, 2004, fact 5)

We conclude with an example that illustates that the condition of irreducibility and recurrence are necessary in theorem 1.2. These conditions ensure that the chain is permamently exploring the entire state space, which is a necessary condition for the convergence of ergodic averages.

Example Consider a discrete chain with two states $S = \{1, 2\}$ and transition matrix Any distribution μ on $1, 2$ is an invariant distribution, as

$$\mu' \mathbf{K} = \mu' \mathbf{I} = \mu'$$

for all μ. However the chain is not irreducible (or recurrent): we cannot get from state 1 to state 2 and vice versa. If the inital distribution is $\mu = (\alpha, 1-\alpha)'$ with $\alpha \in [0, 1]$ then for every $t \in \mathbb{N}_0$ we have that

$$\mathbb{P}(X_t = 1) = \alpha \qquad \mathbb{P}(X_t = 2) = 1 - \alpha$$

By observing one sample path (which is either 1,1,1,... or 2,2,2,...) we can make no inference about the distribution of X_t or the parameter α. The reason for this is that the chain fails to explore the whole space space. To clarify the chain fails to switch between the states 1 and 2. In order to estimate the parameter α we would need to look at more than one sample path. □

2. Monte Carlo Methods

2.1. What are Monte Carlo Methods?

This collection of lectures is concerned with Monte Carlo methods, which are sometimes referred to as *stochastic simulation*. Examples of Monte Carlo methods include stochastic integration, where we use a simulation-based method to evaluate an integral, Monte Carlo tests, where we resort to simulation in order to computer the p-value, and Markov-Chain Monte Carlo (MCMC), where we construct a Markov chain which (hopefully) converges to the distribution of interest.

. A formal definition of Monte Carlo methods was given (amongst others) by Halton (1970)[1] He defined a Monte Carlo method as "representing the solution of a problem as a parameter of a hypothetical population, and using a random sequence of numbers to construct a sample of the population, from which statistical estimates of the parameter can be obtained."

[1]Halton, J.H. A retrospective and prospective survey of the Monte Carlo method. SIAM Review, **12**, 1-63.

2.2. Shalizi Notebook on Monte Carlo Methods

Cosma Shalizi obviously has an excellent description of what a Monte Carlo method is. Monte Carlo is an estimation procedure. The basic idea is as follows. You want to know the average value of some random variable. You can't work out what its distribution is, exactly, or you don't want to do integrals numerically, but you can take samples from that distribution. (The random variable may, for instance, be some complicated function of variables with simple distributions, or they distribution may have a hard-to-compute normalizing factor ["partition function" in statistical mechanics].) To estimate it, you simply take samples, independently, to the true value. The central limit theorem says that your average has a Gaussian distribution around the true value.

Here's one of the canonical examples. Say you want to measure the area of a shape with a complicated, irregular outline. The Monte Carlo approach is to draw a square around the shape and measure the square. Now you throw darts into the square, as uniformly as possible. The fraction of darts falling on the shape gives the ratio of the area of the shape to the area of the square. Now, in fact, you can cast almost any integral problem, or any averaging problem, into this form. So you need a good way to tell if you're inside the outline, and you need a good way to figure out how many darts you should throw. Last but not least, you need a good way to throw darts uniformly, i.e., a good random number generator. That's a whole separate art I shan't attempt to describe.

Now, in fact, you don't strictly need to sample independently. You can have dependence, so long as you end up visiting each point just as many times as you would with independent samples. This is useful, since it gives a way to exploit properties of Markov chains in designing your sampling strategy, and even of speeding up the convergence of your estimates to the true averages. (The classic instance of this is the Metropolis-Hastings algorithm, which gives you a way of sampling from a distribution where all you have to know is the ratio of the probability densities at any two points. This is extremely useful when, as in many problems in statistics and statistical mechanics, the density itself contains a complicated normalizing factor; it drops out of the ratio.)

Monte Carlo methods originated in physics, where the integrals desired involved hydrodynamics in complicated geometries with internal heating, i.e., designing nukes. The statisticans were surprisingly slow to pick up on it, though by now they have, especially as "Markov chain Monte Carlo," abbreviated "MC Monte Carlo" (suggesting an gambling rapper) or just "MCMC". Along the way they picked up the odd idea that Monte Carlo had something to do with Bayesianism. In fact it's a general technique for estimating sample distributions and related quantities, and as such it's entirely legitimate for frequentists. Physicists now sometimes use the term for any kind of stochastic estimation or simulation procedure, though I think it's properly reserved for estimating integrals and averages.

2.3. Introductory examples

A raindrop experiment for computing π Assume we want to compute an Monte Carlo estimate of π using a simple experiment. Assume that we could

produce "uniform rain" on the square $[-1,1]\times[-1,1]$, such that the probability of a raindrop falling in to a region $\mathcal{R} \subset [-1,1]^2$ is proportional to the area of $\mathcal{R}$, but independent of the position of $\mathcal{R}$. It is easy to see that this is the case iff the two coordinates X,Y are i.i.d. realisations of uniform distribution on the interval $[-1,1]$(in short $X, Yi.i.d \sim \cup[-1,1]$). Now consider the probability that a raindrop falls into the unit circle. It is

$$\mathbb{P}(drop\ within\ the\ circle) = \frac{area\ of\ the\ unit\ circle}{area\ of\ the\ square} = \frac{\int\int_{x^2+y^2\leq 1} 1dxdy}{\int\int_{-1\leq x,y\leq 1} 1dxdy} = \frac{\pi}{2.2} = \frac{\pi}{4}$$

In other words,

$$\pi = 4.\mathbb{P}(drop\ within\ circle),$$

i.e. we found a way of expressing the desired quantity π as a function of a probability. We can estimate the probability using our raindrop experiment. If we observe n raindrops, then the number of raindrops Z that fall inside the circle is a binomial random variable:

$$Z \sim B(n,p) \qquad with\ p\ =\ \mathbb{P}(drop\ within\ circle).$$

Thus we can estimate p by its maximum -likelihood estimate

$$\hat{p} = \frac{Z}{n}$$

and we can estimate π by

$$\hat{\pi} = 4\hat{p} = 4 \cdot \frac{Z}{n}.$$

Assume we have observed that 77 of the 100 raindrops were inside the circle. In our case our estimate of π is

$$\hat{\pi} = \frac{4 \cdot 77}{100} = 3.08$$

which is relatively poor.

. However the *law of large numbers* guarantees that our estimate $\hat{\pi}$ converges almost surely to π. As n increases, our estimate improves. We can assess the quality of our estimate by computing a confidence interval for π. As we have $Z \sim B(100,p)$ and $\hat{p} = \frac{Z}{n}$, we use the approximation that $Z\ \sim\ N(100p, 100p(1-p))$. Hence, $\hat{p} \sim$ N(p,p(1-p)/100), and we can obtain a 95% confidence interval for p using this normal approximation

$$\left[0.77 - 1.96 \cdot \sqrt{\frac{0.77 \cdot (1-0.77)}{100}}, 0.771.96 \cdot \sqrt{\frac{0.77 \cdot (1-0.77)}{100}}\right]$$

$= [0.6875, 0.8525]$, As our estimate of π is four times the estimate of p, we now also have a confidence interval for π:

$$[2.750, 3.410]$$

Historically, the main drawback of Monte Carlo methods was that they used to be expensive to carry out. Physically random experiments (for example an experiment to examine 'Buffon's Needle' were difficult to perform and so was the numerical processing of their results. This changed fundamentally with the advent of the digital computer. Amongst the first to realize this potential were John von Neuman and Stanislaw Ulam. For any Monte-Carlo simulation we need to be able to reproduce randomness by a deterministic Computer Algorithm. Clearly this is a philosophical paradox, but lots of work has been done on this, and the statistical language R has a lot of 'random number generators' see (?*RNGkind*) in GNU R for further details.

We know that, using importance sampling, we can approximate an expectation $\mathbb{E}_f(h(X))$ without having to sample directly from f. However, finding an instrumental distribution which allows us to *efficiently* estimate $\mathbb{E}_f(h(X))$ can be difficult, especially in large dimensions. In this chapter and the following chapters we will use a somewhat different approach. We will discuss methods that alow obtaining an *approximate* sample from f without having to sample f directly. More mathematically speaking, we will discuss methods that generate a Markov chain whose *stationary distribution* is the distribution of interest f. Such methods are often called MCMC methods. Let us state a few definitions before continuning.

Definition The prior distribution is a key part of Bayesian inference and represents the information about an uncertain parameter θ that is combined with the probability distribution of new data to yield the *posterior distribution.*

Poisson change point model . Assume the following Poisson model of two regimes for n random variables $Y_1, \cdots, Y_n$[2]

$$Y_i \sim \texttt{Poi}(\lambda_1) \; for \; i = 1, \cdots, M$$

$$Y_i \sim \texttt{Poi}(\lambda_2) \; for \; i \; = \; + 1, \cdots, n$$

A suitable (conjugate) prior distibution for λ_j is the $\texttt{Gamma}(\alpha_j, \beta_j)$ distribution with density

$$f(\lambda_j) = \frac{1}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} \beta_j^{\alpha_j} \exp(-\beta_j \lambda_j)$$

The joint distribution of $Y_1, \cdots, Y_n$, λ_1, λ_2, and M is

$$f(y_1, \cdots, y_n, \lambda_1, \lambda_2, M) = \left(\prod_{i=1}^{M} \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} \right) \cdot \left(\prod_{i=1}^{M} \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} \right)$$
$$\cdot \frac{1}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1 - 1} \beta_1^{\alpha_1} \exp(-\beta_1 \lambda_1) \cdot \frac{1}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2 - 1} \beta_2^{\alpha_2} \exp(-\beta_2 \lambda_2)$$

[2]The probability distribution function of the $\texttt{Poi}(\lambda)$ distribution is p(y) = $\frac{\exp(-\lambda)\lambda^y}{y!}$

If M is known, the *posterior distribution* of λ_1 has the density

$$f(\lambda_1|Y_1,\cdots,Y_n,M) \propto \lambda_1^{\alpha_1-1+\sum i=1^M y_i} \exp(-\beta_1+M)\lambda_1),$$

so

$$\lambda_1|Y_1,\cdots,Y_n,M \sim \texttt{Gamma}\left(\alpha_1+\sum_{i=1}^{M} y_i, \beta_1+M\right) \tag{4}$$

$$\lambda_2|Y_1,\cdots,Y_n,M \sim \texttt{Gamma}\left(\alpha_2+\sum_{i=M+1}^{n} y_i, \beta_2+n-M\right) \tag{5}$$

Now assume that we do not know the change point M and that we assume a uniform prior on the set $\{1,\cdots,M-1\}$. It is easy to compute the distribution of M given the observations $Y_1,\cdots,Y_n$, and λ_1 and λ_2. It is a discrete distribution with probability density function proportional to

$$p(M|Y_1,\cdots,Y_n,\lambda_1,\lambda_2) \propto \lambda_1^{\sum_{i=1}^{M} y_i} \cdot \lambda_2^{\sum_{i=M+1}^{n} y_i} \cdot \exp((\lambda_2-\lambda_1)\cdot M) \tag{6}$$

The conditional distributions in (4.1) to (4.3) are all easy to sample from. It is however rather difficult to sample from the joint posterior of (λ_1,λ_2,M). ◁
The example above suggests the strategy of alternately sampling from the (full) conditional distributions(4 to 6 in the example). Ths tentative strategy however raises some questions.

- Is the joint distribution uniquely specified by the conditional distributions?
- Sampling alternately from the conditional distributions yields a Markov chain: the newly proposed values only depend on the present values, not the past values. Will this approach yield a Markov chain with the correct invariant distribution? Will the Markov chain converge to the invariant distribution?

The answer to both questions will turn out to be yes - under certain conditions. The next section will however state the Gibbs sampling algorithm.