Monte Carlo Methods

Lecture notes for MAP001169
Based on Script by Martin Sköld

adopted by Krzysztof Podgórski

Contents

Part I

Simulation and Monte-Carlo Integration

Chapter 1

Simulation and Monte-Carlo integration

1.1 Issues in simulation

1.2 Raw ingredients

Chapter 2

Simulating from specified distributions

2.1 Transforming uniforms

2.2 Transformation methods

2.3 Rejection sampling

2.4 Conditional methods

Chapter 3

Monte-Carlo integration

3.1 Generic Monte Carlo integration

3.2 Bias and the Delta method

3.3 Variance reduction by rejection sampling

3.4 Variance reduction by importance sampling

3.4.1 Unknown constant of proportionality

Chapter 4

Markov Chain Monte-Carlo

Today, the most-used method for simulating from complicated and/or high-dimensional distributions is *Markov Chain Monte Carlo* (MCMC). The basic idea of MCMC is to construct a Markov Chain that has f as stationary distribution, where f is the distribution we want to simulate from. In this chapter we introduce the algorithms, more applications will be given later.

4.1 Markov chains - basic concepts

The sequences of random values, say X_n's, that we have obtained so far were obtained by independent sampling from a certain distribution. In our context this type of sampling was referred to as Monte Carlo sampling. The simplest but important case of this was a sequence of independent Bernoulli variables that models a random flip of a not necessarily symmetric coin. The limiting results of probability theory such as the law of large numbers or the central limit theorem have been used to establish some fundamental asymptotic properties (approximation errors) of the Monte Carlo method. Markov chains can be viewed as simplest models for obtained sequence of random observations that does not involve direct independent samples. The dependence in a sequence of experiments affecting the next value is only through the most recent value. Simplest Markov chains are those that takes values in a discrete (finite or countable) state-space.

More specifically, we take a sequence X_n's such that the distribution of X_{n+1} given that we obtained $X_n = x^{(n)}, \dots, X_0 = x^{(0)}$ depends only on the value $x^{(n)}$ and not on $x^{(i)}$'s for $i < n$. The transition probabilities from the state i to j are given by

$$q(j|i) = P(X_{n+1} = j | X_n = i).$$

They together with the initial distribution distribution X_0 given by $\pi(i) = P(X_n = i)$ on the states i's fully described distributions of the model.

Example 4.1. For a simple example of a Markov chain, let us consider a simple case of three states -1,0,1 and the following matrix $\mathbf{P} = (p_{ij})$

representing the transition probabilities $p_{ij} = q_{j|i}$

$$\mathbf{P} = \begin{bmatrix} 1-2p & 2p & 0 \\ p & 1-2p & p \\ 0 & 2p & 1-2p \end{bmatrix}.$$

The following program simulates from this Markov chain that start from a state x_0.

```
SMC=function(n,p,x0){
  x=vector("numeric",n)
  x[1]=x0
  for(i in 2:n)
    {
   z=rmultinom(1,1,prob=c(p,1-2*p,p))
   if(x[i-1]==0){
     x[i]=z[1,1]-z[3,1]
   }else{
     if(x[i-1]==1){
       x[i]=x[i-1]-z[1,1]-z[3,1]
     }else{
       x[i]=x[i-1]+z[1,1]+z[3,1]
     }
   }
  }
  SMC=x
}
```

An example of sample can be obtained by running

```
n=100
p=1/4
x0=0
x=SMC(n,p,0)
plot(x)
```

and is shown in Figure 4.1 *Left*.

The theory of Markov chains demonstrates that much of asymptotics observed for independent samples are still valid for Markov chains. For example, in Figure 4.1 *Right* it is observed that a sort of law of large numbers should be valid for the Markov chain in hand as the asymptotic frequency of observing the state "1" is evidently converging. One can utilize the above program to observe the asymptotics

```
n=2000
p=1/4
x=SMC(n,p,1)
P1=cumsum(x==1)/cumsum(rep(1,n))
plot(P1,type='l')
```

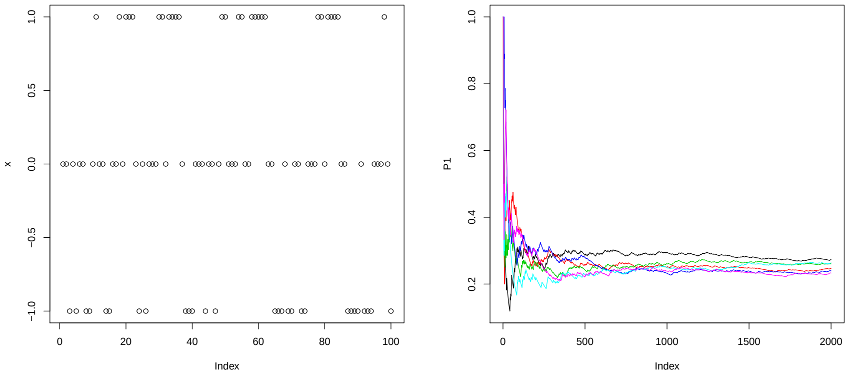

Figure 4.1: A simple 3-state Markov chain. *Left:* a trajectory of size 100 starting from $x^{(0)} = 0$, *Right:* asymptotic frequency of state ”1” based on several trajectories of size 2000.

Markov Chains for which the law of large numbers holds are called ergodic.

Exercise 4.1. *Using the provided example of a Markov chain, make some claims about asymptotical values of the frequencies of states and provided with analysis of error of your claims based on Monte Carlo study.*

Markov Chains can serve often as simple models of real phenomena occuring in time. The following exercise can lead the reader through an attempt to model weather in her/his town. For this one needs a definition of a stationary state.

Definition 4.1. A distribution π_0 on the state space is called stationary if the process starting from that distribution remains in this distribution over the entire time, or more technically the row vector of probabilities given by π_0 satisfies the equation

$$\pi_0 \mathbf{P} = \pi_0.$$

Exercise 4.2. *Consider the following simplistic model for certain aspect of the weather that assumes the lack of memory property, i.e. that cloudeness and rain depends only on the next day depends only on what it was on the previous day. We consider five states: sunny (S) or partly sunny (P), cloudy (C), rainy (R), heavy rain (H). Because of the lack of memory property, this weather model is fully described by providing the matrix of transition probabilities, that describe what are chances for tomorrow to be in one of the five states under the conditions that today we observe one of these states.*

1. *Propose the values of the transition probabilities for summer weather in your town (use your own judgement, not necessarily scientific evidence).*

2. *Using the proposed values generate a 90 days long trajectory of weather pattern.*

3. *Based on your sample estimate the probabilities that on a randomly chosen day in summer the weather will be in one of its five states.*

4. *Check if your estimated values of probabilities satisfies the stationary state equation for the proposed Markov process.*

5. *Compare the estimated values with the evaluated theoretical values for the stationary state (the latter can be found by solving a proper linear equation).*

It should be remembered that not always a Markov chain leads to the asymptotics observed for independent sampling and the conditions for this have to be examined. An interested reader can do the following exercise to see possible problems.

Exercise 4.3 (*Random Walk)***.** *Consider a Markov Chain with infinite but discrete state space*

$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$$

and with transition probabilities given by

$$p_{ij} = \begin{cases} p & : \ j = i+1; \\ 1-p-q & : \ j = i; \\ q & : \ j = i-1; \\ 0 & : \ \textit{otherwise} \end{cases}$$

Generate sample paths of such a Markov chain. By analyzing sample paths, discuss if such a Markov chain is ergodic.

4.2 Markov chains with continuous state-space

The theory for Markov chains that take values in a continuous state-space is complex. Much more so than the theory for finite/countable state-spaces, that you might have already been exposed to. We will here not pretend to give a full account of the underlying theory, but be happy with formulating a number of sufficient conditions under which the methods work. Links to more advanced material will be provided on the course web-page.

In this chapter, we will consider Markov-Chains $X_1, X_2, \dots, X_N$ defined through a starting value x_0 and a transition density $\tilde{q}$. The transition density $x_i \mapsto \tilde{q}(x_{i-1}, x_i) = g(x_i|x_{i-1})$ is the density of $X_i|X_{i-1} = x_{i-1}$, and this determines the evolution of the Markov-chain across the state-space. We would ideally like to draw the starting value x_0 and choose $\tilde{q}$ in such a way that the following realisations $x_1, x_2, \dots, x_N$ are independent draws from f, but this is a too ambitious task in general. In contrast to the previous chapter, here our draws will neither be independent nor *exactly* distributed according to f. What we will require is that f is the *asymptotic distribution* of the chain, i.e. if X_i takes values in $\mathcal{X}$,

$$P(X_n \in A) \to \int_A f(x)\, dx, \tag{4.1}$$

as $n \to \infty$, for all subsets A of $\mathcal{X}$ and independently of the starting value $x_0 \in \mathcal{X}$.

A condition on the transition density that ensures the Markov chain has *a* stationary distribution f, is that it satisfies the *global balance condition*

$$f(y)\tilde{q}(y,x) = f(x)\tilde{q}(x,y). \tag{4.2}$$

This says roughly that the flow of probability mass is equal in both directions (i.e. from x to y and vice versa). Global balance is not sufficient for the stationary distribution to be unique (hence the Markov chain might converge to a different stationary distribution). Sufficient conditions for uniqueness are irreducibility and aperiodicity of the chain, a simple sufficient condition for this is that the support of $f(y)$ is contained in the support of $y \mapsto \tilde{q}(x,y)$ for all x (minimal *necessary* conditions for $\tilde{q}$ to satisfy (4.1) are not known in general, however our *sufficient* conditions are far from the weakest known in literature).

4.3 Markov chain Monte-Carlo integration

Before going into the details of how to construct a Markov chain with specified asymptotic distribution, we will look at Monte-Carlo integration under the new assumption that draws are neither independent not exactly from the target distribution. Along this line, we need a Central Limit Theorem for Markov chains. First we observe that if $X_1, X_2, \ldots, X_n$ is a Markov chain on $\mathbf{R}^d$ and $\phi : \mathbf{R}^d \mapsto \mathbf{R}$, then with $Z_i = \phi(X_i)$, the sequence $Z_1, Z_2, \ldots, Z_n$ forms a Markov chain on $\mathbf{R}$. As before, we want to approximate $\tau = E(Z) = E(\phi(X))$ by $t_N = N^{-1}\sum_{i=1}^N z^{(i)}$, for a sequence $z^{(i)} = \phi(x^{(i)})$ of draws from the chain.

4.3.1 Burn-in

An immediate concern is that, unless we can produce a single starting value x^0 with the correct distribution, our sampled random variables will only *asymptotically* have the correct distribution. This will induce a bias in our estimate of τ.

In general, given some starting value $x^{(0)}$, there will be iterations $x^{(i)}$, $i = 1, \ldots, k$, before the distribution of $x^{(i)}$ can be regarded as "sufficiently close" to the stationary distribution $f(x)$ in order to be useful for further analysis (i.e. the value of $x^{(i)}$ is still strongly influenced by the choice of $x^{(0)}$ when $i \le k$). The values $x^{(0)}, \ldots, x^{(k)}$ are then referred to as the *burn-in* of the chain, and they are usually discarded in the subsequent output analysis.

For example when estimating $\int \phi(x) f(x)\, dx$, it is common to use

$$\frac{1}{N-k} \sum_{i=k+1}^{N} \phi(x^{(i)}).$$

An illustration is given in Figure 4.2. We now provide the CLT for Markov Chains.

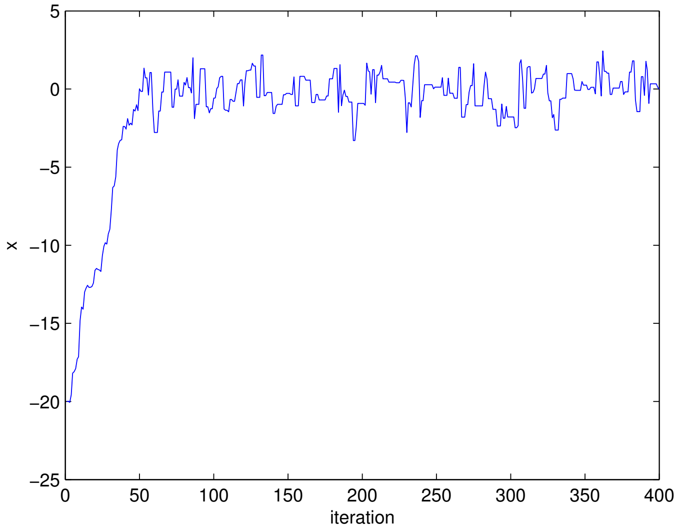

Figure 4.2: A Markov chain starting from $x^{(0)} = -20$ that seems to have reached its stationary distribution after roughly 70 iterations

Theorem 4.1. *Suppose a geometrically ergodic Markov chain X_i, $i = 1, \ldots, n$, on $\mathbf{R}^d$ with stationary distribution f and a real-valued function $\phi : \mathbf{R}^d \mapsto \mathbf{R}$ satisfies $E(\phi^{2+\epsilon}(X)) \leq \infty$ for some $\epsilon > 0$, then with $\tau = E(\phi(X))$ and $T_n = n^{-1} \sum_{i=1}^{n} \phi(X_i)$*

$$P\Big(\frac{\sqrt{n}(T_n - \tau)}{\sigma} \leq x\Big) \to \Phi(x) \tag{4.3}$$

where Φ is the distribution function of the $N(0,1)$ distribution and

$$\sigma^2 = r(0) + 2\sum_{i=1}^{\infty} r(i) \tag{4.4}$$

where $r(i) = \lim_{k\to\infty} Cov\{\phi(X_k), \phi(X_{k+i})\}$, the covariance function of a stationary version of the chain.

Note that the above holds for an arbitrary starting value x_0. The error due to “wrong” starting value, i.e. the bias $E(T_n) - \tau$, is of $O(n^{-1})$. Hence squared bias is dominated by variance, and asymptotically negligible. This does not suggest that it is unnecessary to discard burn-in, but rather that it is unnecessary to put too much effort into deciding on how long it should be. A visual inspection usually suffices.

4.3.2 After burn-in

If we manage to simulate our starting value $x^{(0)}$ from the target distribution f, all subsequent values will also have the correct distribution. Would our problems then be solved if we were given that single magic starting value with the correct distribution? As the above discussion suggests, the answer is no. The “big problem” of MCMC is that (4.4) can be very large, especially in problems where the dimension d is large. It is important to understand that converging to the stationary distribution (or getting close enough) is just the very beginning of the analysis. After that we need to do sufficiently many iterations in order for the variance σ^2/n of T_n to be small.

It is actually a good idea to choose starting values $x^{(0)}$ that are *far away* from the main support of the distribution. If the chain then takes a long time to stabilize you can expect that even after reaching stationarity, the autocorrelation of the chain will be high and σ^2 in (4.4) large. Here we give a rough guide for estimating σ^2, which is needed want we to produce a confidence interval:

1. We assume your draws $x^{(i)}$, $i = 1, \ldots, N$ are d-variate vectors. First make a visual inspection of each of the d trajectories, and decide a burn-in $1, \ldots, k$ where *all* of them seems to have reached stationarity. Throw away the first k sample vectors.

2. Now you want to estimate $E(\phi(X))$. Compute $z_i = \phi(x^{(i+k)})$, $i = 1, \ldots, N-k$ and estimate the autocorrelation function of the sequence z_i, i.e. the function $\rho(t) = \mathrm{Corr}(Z_1, Z_t)$ over a range of values $t = 1, \ldots, L$, this can be done with R-function `acf` (though it uses the range $-L, \ldots, L$). A reasonable choice of L is around $(N-k)/50$ (estimates of $\rho(L)$ tends to be too unreliable for larger values). If the estimated autocorrelation function does not reach zero in the range $1, \ldots, L$, go back to the first step and choose a larger value for N.

3. Divide your sample into m batches of size l, $ml = N - k$. Here l should be much larger than the time it took for the autocorrelation to reach zero in the previous step. The batches are $(z_1, \ldots, z_l)$,$(z_{l+1}, \ldots, z_{2l})$, . . ., $(z_{m(l-l)+1}, \ldots, z_{ml})$. Now compute the arithmetic mean $\bar{z}_j$ of each batch, $j = 1, \ldots, m$. Estimate τ by the mean of the batch means and σ^2 by s^2/m, where s^2 is the empirical variance of the batch means.

This procedure is based on the fact that if batches are sufficiently large, their arithmetic means should be approximately uncorrelated. Hence, since our estimate is the mean of m approximately independent batch means it should have variance σ_b^2/m, where σ_b^2 is the variance of a batch mean.

We now turn to the actual construction of the Markov chains.

4.4 Two simple continuous time Markov chain models

4.4.1 Autoregressive model

Probably, the simplest continuous state is an autoregressive time series X_n that is given by

$$X_{n+1} = \rho X_n + \epsilon_n,$$

where ϵ_n are iid normal random variables with the mean zero and variance σ^2. One can easily argue that this is a Markov chain. Derivation of the transition densities is left for the reader, who can also study the model through simulations as suggested in the following exercise.

Exercise 4.4. *Propose a simulator of the autoregressive model. By its means simulate trajectories of from such a model and consider for which values of the parameter ρ the model is stable. Using autocorrelation facilities of R approximate the autocorrelation function of X_n as well as its variance.*

4.4.2 Modeling cloud coverage

Daily cloud coverage is an important characteristics in weather studies. It can be expressed in the percentage of sunlight passing through the clouds relatively to the amount recorded when the sky is completely clear. Modeling such a process is extension of the simple model of Exercise 4.2. For modeling percentages the beta distribution is a natural family of distributions. The densities of this distributions are given up to a proportionality constant by

$$f(x; \alpha, \beta) \sim x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0, 1].$$

We leave to the reader to develop a Markov model that would model cloud coverage using continuous state space and beta distributions in the spirit of Exercise 4.2 and Example 4.1

Exercise 4.5. *Propose a Markov chain approach to modeling cloud coverage using beta distributions. For the model develop programs and study its stability and asymptotic behavior.*