1. The Chapman–Kolmogorov equations for Markov chains

Let $X_0, X_1, X_2, \ldots$ be a time stationary Markov chain over some finite state space, say \{1, \ldots, k\}. Its probability mechanism is then governed by the one-step transition probabilities

$$P_{i,j} = \Pr(X_{n+1} = j \mid X_n = i) \quad \text{for all } i, j = 1, \ldots, k,$$

and it is famously convenient to collect these into one matrix

$$P = \begin{pmatrix} P_{1,1} & \cdots & P_{1,k} \\ \vdots & \ddots & \vdots \\ P_{k,1} & \cdots & P_{k,k} \end{pmatrix},$$

termed the transition probability matrix. (A brief typographical comment: One finds both $P_{i,j}$ and $P_{ij}$ in standard literature, i.e. with or without the comma to separate start state $i$ and end state $j$, and Ross’s book prefers not using the comma.)

(a) Let $1 = (1, \ldots, 1)^{\intercal}$ be the column vector of 1s. Show that

$$P1 = 1.$$

In particular, every transition probability matrix has 1 as an eigenvalue.

(b) Show that the two-step transition probabilities can be computed via the one-step transition probabilities, as follows:

$$P_{i,j}^2 = \Pr(X_{n+2} = j \mid X_n = i) = \sum_{\ell} P_{i,\ell}P_{\ell,j}.$$ 

Note that these are identical to the elements of the matrix $P^2$. (In Markov chain literature one often sees $n$-step probabilities given as $P_{i,j}^{(n)}$ rather than as in Ross’s book, where they are given simply as $P_{i,j}^n$. This is fine as long as one knows ‘what is what’; do not confuse $P_{i,j}^{10}$, which I would prefer to denote as $P_{i,j}^{(10)}$, with $(P_{i,j})^{10}$.)

(c) Show more generally that the $(m + n)$-step probabilities may be computed via the $m$-step and $n$-step probabilities:

$$P_{i,j}^{m+n} = \Pr(X_{m+n} = j \mid X_0 = i) = \sum_{\ell} P_{i,\ell}^nP_{\ell,j}^n.$$ 

These are the Chapman–Kolmogorov equations for Markov chains (derived independently by Sydney Chapman and Андрей Колмогоров, с. 1930). In particular, we
have the important and practical result that the \( n \)-step transition probabilities are found simply by computing the \( n \)-power of the one-step probability matrix:

\[
P^{(n)} = (P(X_n = j \mid X_0 = i))_{i,j=1,...,k} = P^n.
\]

In yet other words, the laws of probability for Markov chains fit like the proverbial hand in glove with matrix power multiplication.

### 2. Matrix power computations

This exercise demonstrates how one may perform simple matrix power computations in R (calculations in other software packages may be handled similarly). Consider a Markov chain over the states 1, 2, 3, 4 and with transition probability matrix

\[
P = \begin{pmatrix}
  0.28 & 0.37 & 0.24 & 0.11 \\
  0.20 & 0.24 & 0.26 & 0.30 \\
  0.47 & 0.13 & 0.16 & 0.24 \\
  0.03 & 0.52 & 0.17 & 0.28
\end{pmatrix}.
\]

(a) Compute \( P^2, P^4, P^8, P^{16} \), by repeated multiplications. Note that \( P^{16} \) has ‘almost converged’ to a limit. (Note that matrix multiplication \( AB \) in R is accomplished using \( A \times B \). Asking for \( A * B \) gives something different, namely coordinate-wise multiplication.)

(b) Using the array option and a simple programming loop one may easily enough compute all \( P^n \) powers up to a given limit, say \( n_{\text{max}}=100 \). Write and edit a little programme in your computer, as follows. First you need to read in the \( P \) matrix, e.g. using

\[
P <- \text{rbind(c(0.28, 0.37, 0.24, 0.11), c(0.20, 0.24, 0.26, 0.30), c(0.47, 0.13, 0.16, 0.24), c(0.03, 0.52, 0.17, 0.28))}
\]

Then you create an ‘array’ of dimension \((k,k,n_{\text{max}})\), where \( k=\text{ncol}(P) \) (in this case, equal to 4), and where \( P[ , , n] \) is \( P^n \):

\[
P^{\text{power}} <- \text{array(data=0, dim=c(kk, kk, n_{\text{max}}))}
\]

\[
P^{\text{power}}[ , , 1] <- P
\]

for (\( n \) in 2:n_{\text{max}})

\[
\{ \text{Ppower}[ , , n] <- \text{Ppower}[ , , n-1] \times \times \times P \}
\]

Note that there is relatively quick convergence of \( P^n \) to a limit matrix where each row is the same, say \((\pi_1, \pi_2, \pi_3, \pi_4)\). Convergence to limits may be monitored e.g. via

\[
\text{plot(Ppower[1,3,])}
\]

Experiment a bit with other transition matrices of different sizes, e.g. with a 100-state Markov chain, where you generate a suitable distribution for each of the 100×100 matrix by normalising a 100-sample from some distribution. Is there ‘quick convergence’ of \( P^n \) to its limit?
3. Simulating Markov chains

One interesting and perhaps mildly provocative definition of ‘I understand a certain stochastic process’ is ‘I am able to write down a programme that produces realisations from that stochastic process in my computer’. This exercise is about being able to simulate chains from a given Markov chain mechanism. A key step is the ability to draw a random sample from a given finite probability distribution, which in R is accomplished using

```r
x <- sample(list, m, prob=problist)
```

The outcome is say \(x_1, \ldots, x_m\), a total of \(m\) independent draws from the sample space `list` of possible outcomes, and with probabilities governed by `problist`. Along with a suitable `for` loop, this makes it easy to simulate a short or long Markov chain.

(a) Going back to the \(4 \times 4\) matrix \(P\) of Exercise 2, let us generate say \(X_1, \ldots, X_{1000}\) from that chain mechanism, with initial state say \(X_0 = 3\). Write and edit a suitable programme in your computer, along these lines:

```r
X0 <- 3
sim <- 1000
Xsim <- 0*(1:sim)
Xsim[1] <- sample(1:kk, 1, prob=P[X0, ])
for (n in 2:sim)
  { Xsim[n] <- sample(1:kk, 1, prob=P[Xsim[n-1], ]) }
```

(b) Check that the relative frequencies with which the chain has visited the different states 1, 2, 3, 4 reasonably well match the limits \(\pi_1, \pi_2, \pi_3, \pi_4\) found on the rows of \(P^n\) for large \(n\). Formulate a precise mathematical statement about this.

(c) Use this little machinery to experiment with other Markov chains.

4. The equilibrium distribution for Markov chains

The fundamental limit theorem for Markov chains is that there under mild regularity conditions will be an equilibrium distribution (also called stationary distribution or limiting distribution):

\[
P^n_{i,i} \to \pi_i \quad \text{for all } i.
\]

Ross states his Theorem 4.1 without proof. This and the following exercise reflect a slightly higher ambition level by actually providing proofs of this and some related important results. This takes a bit of linear algebra and matrix decomposition theory.

The framework worked with now is that of a Markov chain on a finite state space, say \(\{1, \ldots, k\}\), with only one class (we say that the chain is irreducible), and where the chain is also aperiodic. This latter assumption is that the period \(d(i)\) is equal to 1, where this quantity is formally defined via

\[
d(i) = \text{lcd}\{n \geq 1: P^n_{i,i} > 0\},
\]
the least common denominator (‘minste felles multiplum’) of all time spans for which
transition from \( i \) back to \( i \) is possible. Thus the period is 1 if e.g. \( P^n_{i,i} \) is positive for
\( n = 4, 5 \) (and many more \( n \)), since 4 and 5 have no common divisor greater than 1.

From linear algebra and matrix theory, recall that \( \lambda \) is an eigenvalue for \( P \) if \( Pu = \lambda u \)
for some non-zero \( u \), which is then called an eigenvector. Thus \( \lambda = 1 \) is an eigenvalue, with
1 and eigenvector, by Exercise 1(a). All eigenvalues are roots of the equation \( Q_n(\lambda) = |P - \lambda I_k| \), with \( I_k = \text{diag}(1, \ldots, 1) \) the identity matrix of size \( k \times k \); this is a polynomial
of degree \( k \), and its roots may be complex. In \( \mathbb{R} \) one computes the eigenvalues \( \lambda_j \) and
eigenvectors \( u_j \), via

\[
\text{lambda} <- \text{eigen(P)}$values \\
A <- \text{eigen(P)}$vectors
\]

Then \( Pu_j = \lambda_j u_j \) for \( j = 1, \ldots, k \), where \( u_j \) is the \( j \)th column of \( A \). This is typically
organised such that \(|\lambda_1| \geq \cdots \geq |\lambda_k|\). The crucial point, in the present connection, is that
we have a spectral decomposition,

\[
A^{-1}PA = D = \text{diag}(\lambda_1, \ldots, \lambda_k).
\]

(a) Show that \( P^n u_j = \lambda^n_j u_j \) for each \( n \). Use this to show that \(|\lambda_j| \leq 1 \) – if a \( \lambda_j \) has a
size bigger than 1, then \( \lambda^n_j \) explodes, which contradicts the property that each row in
\( P^n \) is a probability distribution. Attempt furthermore to show that under the given
assumptions (in particular, \( d(i) = 1 \)), each \( \lambda_j \) has a size strictly smaller than 1, apart
from the first: \( 1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_k| \).
(b) Show that \( P = ADA^{-1} \) and that \( P^n = AD^n A^{-1} \). But

\[
D^n = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & \lambda^n_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda^n_k
\end{pmatrix},
\]

which converges to the simple diagonal matrix \( E_1 = \text{diag}(1, 0, \ldots, 0) \). Hence conclude
that

\[
P^n \rightarrow AE_1A^{-1}.
\]
(c) The formula above may be used to read off the \( \pi_j \) limits in practice, but it is not
yet clear from the mathematics that each row in the limit matrix is the same. Show
however that this is necessarily the case: the limit matrix \( L = AE_1A^{-1} \) has \((i, j)\)-
element

\[
(\text{AE}_1A^{-1})_{i,j} = a_{i,1} a^{1,j} \quad \text{for } i, j = 1, \ldots, k,
\]
where \( a_{i,1} \) are the elements in the very first eigenvector, associated with \( \lambda_1 = 1 \), and
hence this first column vector of \( A \) is proportional to \( 1 \), with all elements equal. Also,
\( a^{1,j} \) are the elements of the first row of \( A^{-1} \). We have now proved that

\[
P^n_{i,j} \rightarrow \pi_j \quad \text{for all } i, j,
\]
with limit independent of starting point.
(d) Flowing from the above is a simple numerical recipe for finding the equilibrium distribution in practice (from a given $P$ matrix), as

$$\pi_j = a_{1,1}a^{1,j} \text{ for } j = 1, \ldots, k.$$  

Carry out these calculations for the case of the four-state Markov chain worked with in Exercise 2. Compare with the other easy recipe which is to compute $P^n$ numerically for moderately large $n$.

(e) Show from the analysis above that there is an explicit formula for $P^n$, in terms of its eigenvalues and eigenvectors:

$$P^n = AE_1A^{-1} + \lambda_2^n AE_2A^{-1} + \cdots + \lambda_k^n AE_kA^{-1},$$

where $E_j$ is the matrix which has zeroes everywhere apart from a 1 at position $(j, j)$. We also learn from this that the convergence is exponentially fast (each $\lambda_j^n$ goes much faster to zero than e.g. $1/n$ or even $1/n^{1000}$). Show more precisely that there is a constant $c$ and a positive number $\rho < 1$ such that

$$|P^n_{i,i} - \pi_i| \leq c\rho^n \text{ for all } n,$$

with neither $c$ nor $\rho$ depending on $i$; we may in fact take $\rho$ equal to the size of the second biggest eigenvalue. This finding has consequences for how quickly a Markov chain forgets where it comes from.

### 5. The ‘Master Theorem’ for Markov chains

Via matrix decomposition methods we were able to prove the essence of Ross’s Theorem 4.1 (which in his book is stated without proof). We now give a more complete ‘Master Theorem’, spelling out also some other consequences that are partly not well highlighted in Ross’s book. The framework is that of Exercise 4, with a Markov chain on a finite state space, aperiodic and irreducible (one class, all states having period equal to 1).

1. There is an equilibrium distribution $\pi_1, \ldots, \pi_k$ to which there is convergence

$$P^n_{i,i} \rightarrow \pi_i \text{ for each } i.$$

2. There is also convergence

$$P^n_{j,i} \rightarrow \pi_i \text{ for all } j, i,$$

i.e. the chain ‘forgets where it started’.

3. Let $B_n(i) = \sum_{j=1}^{n} I(X_j = i)$ be the number of visits in state $i$, in the course of the first $n$ time points. Then the relative frequencies converge,

$$A_n(i) = B_n(i)/n \rightarrow_p \pi_i \text{ for each } i.$$

The convergence here is ‘in probability’, and also takes place in the sense that

$$E A_n(i) \rightarrow \pi_i \text{ and } \text{Var} A_n(i) \rightarrow 0.$$
(4) State $i$ is recurrent and will hence be visited, re-visited, etc., with probability 1. Let $V_{i,1}, V_{i,2}, \ldots$ be the waiting times in state $i$; the visiting times are thus $S_{i,1} = V_{i,1}$, $S_{i,2} = V_{i,1} + V_{i,2}$, $S_{i,3} = V_{i,1} + V_{i,2} + V_{i,3}$, etc. Then the $V_j$ have expected value

$$\xi_i = E V_{i,j} = E(S_{i,1} | X_0 = i) = 1/\pi_i,$$

i.e. the mean return time is inversely proportional to the equilibrium probability.

(5) The limiting probabilities $\pi_i$ are positive and satisfy the equations

$$\pi_i = \sum_j \pi_j P_{j,i} \quad \text{for all } i,$$

along with $\sum_j \pi_j = 1$. The solution to these equations is unique.

We have already seen that (1) and (2) hold true, cf. Exercise 4, and now set out to show (3), (4), (5).

(a) Show that with any convergent sequence, also the sequence of averages converges, to the same limit:

$$c_n \to c \implies \bar{c}_n = n^{-1} \sum_{i=1}^n c_i \to c.$$ 

An ‘$\varepsilon$ and $n_0$’ argument is required.

(b) Regarding statement (3), this holds regardless of initial state $X_0$, but let us for simplicity of presentation assume that the chain starts in the same state $i$, i.e. $X_0 = i$. Show that

$$\mathbb{E} A_n(i) = n^{-1} \sum_{j=1}^n P_{i,j}^j$$

and then also that this sequence tends to $\pi_i$.

(c) We then need to show that

$$\text{Var } A_n(i) = (1/n^2) \text{Var}(D_1 + \cdots + D_n) \to 0,$$

where $D_j = I(X_j = i)$. The crucial step here is to bound the covariances, as

$$\text{Var}(D_1 + \cdots + D_n) = \sum_{j=1}^n \text{Var} D_j + 2 \sum_{j<\ell} \text{cov}(D_j, D_\ell),$$

with the second sum containing $n(n-1)/2$ terms. Show in fact that

$$\text{cov}(D_j, D_\ell) = P_{i,i}^j (P_{i,i}^{\ell-j} - P_{i,i}^{\ell})$$

and that

$$|\text{cov}(D_j, D_\ell)| \leq c \rho^{\ell-j} + c \rho^\ell,$$

with $\rho < 1$, as per Exercise 4(e). Use this to prove that $\text{Var } A_n(i)$ tends to zero, as required.
(d) To prove (4), note first that the waiting times \( V_{i,1}, V_{i,2}, \ldots \) are independent with the same distribution (i.i.d.), by the nature of the Markov chain; it returns an infinite number of times to \( i \), and each time it returns, the process re-starts, with no further memory of the past. Thus

\[
\frac{S_{i,n}}{n} = n^{-1} \sum_{j=1}^{n} V_{i,j} \to \xi_i = E V_{i,j},
\]

by the law of large numbers. Show now that \( B_n(i) = m \) is equivalent to \( S_{i,m} \leq n < S_{i,m+1} \).

Deduce that

\[
\frac{S_{i,B_n(i)}(i)}{B_n(i)} \leq n \frac{S_{i,B_n(i)+1}}{B_n(i)}.
\]

Use a sandwich argument to conclude that \( \xi_i \) must be equal to \( 1/\pi_i \), as claimed.

(e) Use the Chapman–Kolmogorov equations

\[
P^n = PP^{n-1} = P^{n-1}P
\]

to deduce the two identities

\[
P^n_{i,i} = \sum_j P_{i,j}P_{j,i}^{n-1} = \sum_j P_{i,j}^{n-1}P_{j,i}.
\]

Then take the limit as \( n \to \infty \):

\[
\pi_i = \sum_j P_{i,j}\pi_i = \sum_j \pi_jP_{j,i}.
\]

The first result is absolutely true and absolutely uninteresting (in the same class as ‘0 = 0’, a theorem every mathematician regularly rediscover s). The second result is however interesting and crucial.

(f) Show that the equations of (5) can be neatly summarised in

\[
\pi = \pi P,
\]

where \( \pi = (\pi_1, \ldots, \pi_k) \) is seen as a row vector. Show also that \( \pi P^m = \pi \) for every \( m \):

\[
\pi_i = \sum_j \pi_j P_{j,i}^m \quad \text{for all } i.
\]

6. The Gambler’s Ruin

The so-called Gambler’s Ruin model comes in various variants, but the simplest version is that laid out in Ross’s Section 4.5: If \( X_n = i \), with \( 1 \leq i \leq N - 1 \), then

\[
X_{n+1} = \begin{cases} 
  i + 1 & \text{with probability } p, \\
  i - 1 & \text{with probability } q,
\end{cases}
\]

with \( q = 1 - p \). States \( 1, \ldots, N - 1 \) are transient while states 0 and \( N \) are absorbing. You may think of \( X_n \) as your own position in a game (perhaps the number of points, or Kroner, on some scale); if \( X_n \) ends in 0 you’re dead, and if \( X_n \) ends in \( N \) you have won. Below we go through some of the results reached in Ross’s book, but also put in a bit more.
(a) Consider

\[ u_i = P(X_n \text{ ends in } N \mid X_0 = i), \]

so that \( 1 - u_i \) correspondingly is the probability that the game sooner or later ends in zero. Show that

\[ u_i = qu_{i-1} + pu_{i+1} \quad \text{for } 1 \leq i \leq N - 1, \]

and that \( u_0 = 0, u_N = 1 \).

(b) To solve these equations it turns out to be practical to work with the differences, say

\[ \Delta_i = u_{i+1} - u_i \quad \text{for } i = 0, 1, \ldots, N - 1. \]

Show that \( \Delta_i = \alpha \Delta_{i-1} \), with \( \alpha = q/p \), and use this to infer that

\[ \Delta_i = \alpha^i \Delta_0 \quad \text{for } i = 0, 1, \ldots, N - 1. \]

\[ \begin{array}{c}
\text{start position} \\
\text{prob(ending in N)}
\end{array} \]

Figure 1: For \( N = 100 \), curves giving \( u_i = P(X_n \text{ ends in } N \mid X_0 = i) \) are displayed as a function of start position \( X_0 = i \), for \( p \) equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55.

(c) Sum all differences to find

\[ u_i - u_0 = \Delta_0 + \cdots + \Delta_{i-1} = \Delta_0 (1 + \cdots + \alpha^{i-1}) = \Delta_0 \frac{1 - \alpha^i}{1 - \alpha}. \]
Use the side conditions $u_0 = 0$ and $u_N = 1$ to find $\Delta_0$. Conclude that

$$u_i = P(\text{you win in the end} \mid X_0 = i) = \frac{1 - \alpha^i}{1 - \alpha^N} \quad \text{for } i = 0, 1, \ldots, N.$$  

This is valid for the case of $p \neq \frac{1}{2}$. For the symmetric case of $p = q = \frac{1}{2}$ one cannot use the same formula for $1 + \alpha + \cdots + \alpha^{i-1}$, but the arguments may be used and lead to the simpler formula $u_i = i/N$.

(d) Write a computer programme to copy Figure 1, which presents $u_i$ as a function of start position $X_0 = i$, for the case of $N = 100$ and for values of $p$ equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Note that small differences in $p$ lead to rather drastic differences in $u_i$.

(e) We next set out to find

$$v_i = E(T \mid X_0 = i) \quad \text{for } i = 0, 1, \ldots, N,$$

where $T$ denotes the time the process first settles down at one of the endpoints, i.e. the first $n$ where $X_n$ is either 0 or $N$. Show that

$$v_i = 1 + qv_{i-1} + pv_{i-1} \quad \text{for } i = 1, \ldots, N - 1,$$

with side conditions $v_0 = v_N = 0$. 

Figure 2: For $N = 100$, curves giving $v_i = E(T \mid X_0 = i)$ are displayed as a function of start position $X_0 = i$, for $p$ equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Here $T$ is the first time $X_n$ settles at 0 or $N$. 

(e) We next set out to find
(f) As above it is mathematically convenient to transform the problem into equations involving the differences $\delta_i = v_{i+1} - V_i$. Show that

$$q\delta_{i-1} = 1 + p\delta_i \quad \text{for } i = 1, \ldots, N - 1,$$

giving $\delta_i = \alpha\delta_{i-1} - d$, with $\alpha = q/p$ and $d = 1/p$. Show that this leads to

$$\delta_i = \alpha^i \delta_0 - d \frac{1 - \alpha^i}{1 - \alpha} \quad \text{for } i = 0, 1, \ldots, N - 1.$$

(g) Then sum all differences to arrive at

$$v_i - v_0 = \delta_0 + \cdots + \delta_{i-1} = \delta_0 \frac{1 - \alpha^i}{1 - \alpha} - d \left( i - \frac{1 - \alpha^i}{1 - \alpha} \right).$$

Using the side conditions, show that

$$\delta_0 = d \left( \frac{N}{1 - \alpha N} - \frac{1}{1 - \alpha} \right),$$

which gives a formula for $v_i$, the expected time spent until the game ends.

(h) The above assumes $p \neq \frac{1}{2}$, since we were using the geometric sum formula for $1 + \alpha + \cdots + \alpha^{i-1}$. For the symmetric case $p = q = \frac{1}{2}$, use similar but actually simpler arguments to show that $v_i = i(N - i)$.

(i) Finally write a computer programme to copy Figure 2, which presents mean game length time $v_i$ as a function of start position $X_0 = i$, for the case of $N = 100$ and for values of $p$ equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Note again that small differences in $p$ lead to rather drastic differences in $v_i$.

7. MCMC: Markov Chain Monte Carlo

[Exercises 7–8 serve as ‘alternative curriculum’ to Ross’s Sections 4.8–4.9]

The typical modus of thinking and working regarding Markov chains, both generally and inside this course, is that one constructs a Markov chain model to emulate a phenomenon of interest (perhaps with parameters estimated from data etc.), after which one typically finds the equilibrium distribution, perhaps computes some probabilities of interest, etc. A historical footnote of just sufficient relevance to cause my mind to decide to tell you about it in this sentence without waiting for another chance is that Markov chains were invented in 1906 and that Markov himself did the world’s first ever modelling of something real via Markov chains in 1913, with the data being the sequence of the first 20,000 letters of Pushkin’s fantastic Евгений Онегин 1833 poem, unpoetically viewed on this occasion as a string of vowels and consonants. He did just what I mentioned above: (i) modelling something ‘real’ as a Markov chain, with transition probabilities $P_{i,j}$ etc.; (ii) computing the equilibrium distribution $\pi_i$ etc. (along with other aspects).
In this exercise, however, we turn things the other way around – we start with the \( \pi_i \) distribution and then invent a Markov chain \( P_{i,j} \) that happens to have the \( \pi_i \) as its equilibrium distribution! Got it?

This is perhaps a simple point, but turns out to have grandiose consequences, opening up doors to a space of applications where complicated probability distributions are assessed and worked with and simulated from not directly (because it may be too difficult or impossible), but via Markov chains that converge to the desired distribution: the chain \( X_0, X_1, X_2, \ldots \) will after a burn-in period behave as samples from the \( \pi_i \) distribution. This is the MCMC (Markov Chain Monte Carlo) idea.

(a) Let \( \{ \pi_i; i \in I \} \) be some probability distribution on a set \( I \) (finite or countably infinite), assumed for simplicity of presentation to be positive everywhere. Assume that \( P \) is a Markov matrix of \( P_{i,j} \) (that we think of as constructed or chosen after knowledge of the \( \pi_i \)) with the property that

\[
\pi_i P_{i,j} = \pi_j P_{j,i} \quad \text{for all } i, j.
\]

(B)

We also assume that the \( P \) chain is aperiodic and recurrent. Show that the \( \pi_i \) must be the equilibrium distribution of the \( P \) chain.

(b) Just to pause for a minute for you to make sure you know what this means – show that

\[
P^n \to \begin{pmatrix} \pi_1 & \pi_2 & \cdots \\ \pi_1 & \pi_2 & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}
\]

and that

\[
P^n_{i_0,i} = P(X_n = i \mid X_0 = i_0) \to \pi_i \quad \text{for all } i,
\]

regardless of starting point \( i_0 \). Yet another way to look at this is that if \( X_0, X_1, X_2, \ldots \) is a chain drawn from \( P \), then the distribution of \( X_n \) tends to the \( \{ \pi_i; i \in I \} \).

(c) This opens the door to this idea: to simulate realisations from the \( \pi_i \) distribution, you may set up a Markov Chain obeying condition (B) and then simulate \( X_0, X_1, \ldots \) from this chain.

(d) Now let us construct such a Markov chain, with the desired property (B). Suppose there is a ‘proposal distribution’

\[
Q_{i,j} = P(X_{n+1}^{\text{cand}} = j \mid X_n = i)
\]

that draws a candidate value \( j \) after the current \( i \), and that this mechanism is symmetric (i.e. \( Q_{i,j} = Q_{j,i} \) for all \( i, j \)). Assume next that we actually accept the proposed candidate \( j \) with acceptance probability

\[
\text{pracc}(i,j) = \min\left(1, \frac{\pi_j}{\pi_i}\right).
\]
In other words, such a chain has

$$X_{n+1} = \begin{cases} X_{n+1}^{\text{cand}} & \text{if this proposed value is accepted,} \\ X_n & \text{if else.} \end{cases}$$

Show that this corresponds to the Markov chain mechanism with transition probabilities

$$P_{i,j} = Q_{i,j} \text{pracc}(i, j) = Q_{i,j} \min(1, \pi_j / \pi_i) \quad \text{for } j \neq i$$

and

$$P_{i,i} = Q_{i,i} + \sum_{j \neq i} Q_{i,j} \{ 1 - \text{pracc}(i, j) \} = 1 - \sum_{j \neq i} Q_{i,j} \min(1, \pi_j / \pi_i).$$

(e) Show that this Markov chain indeed satisfies condition (B). (There are many other constructions and solutions; this is one type that is easy to implement and often works well.)

Markov, A.A. (1906). Распространение закона больших чисел на величины, зависящие друг от друга. [Extending the law of large numbers for variables that are dependent of each other.] Известия Физико-математического общества при Казанском университете 15 (2-я серия), 124–156.


### 8. MCMC: an example

To give an illustration of the MCMC idea, consider the following moderately strange probability distribution

$$\pi_i = \frac{1}{K} f(i) = \frac{1}{K} \exp(-\beta|i|) |i - \frac{1}{2}|^{\alpha} \exp\{ \gamma \sin(\pi i) \} \quad \text{for } i = 0, \pm 1, \pm 2, \pm 3, \ldots,$$

for certain parameters $\alpha, \beta, \gamma$, where $K = \sum_{i} f(i)$. I invented it just now just to have an example of something we haven’t seen before, and for which there is no free simulation recipe. The task is to produce a sequence of realisations from this distribution.

(a) Show that the $\pi_i$ here really define a proper probability distribution, as long as $\beta > 1$.

What happens if $\beta \leq 1$?

(b) Follow the MCMC recipe of the previous exercise, with symmetric proposal distribution $\left( \frac{1}{2}, \frac{1}{2} \right)$ for $X_{n+1}^{\text{cand}} = j = i \pm 1$ when $X_n = i$, and accept with

$$\text{pracc} = \min(1, \pi_j / \pi_i) = \min\left( 1, \frac{f(X_{n+1}^{\text{cand}})}{f(i)} \right).$$

Explain that $X_n$ converges in distribution to the $\pi_i$ distribution. Note that we do not need the numerical value of $K$ to set this in motion.
Figure 3: These are the first 1,000 steps of the MCMC, started (a bit idiosyncratically) at position $X_1 = 44$. After some 300-400 steps the chain has reached its equilibrium distribution.

(c) Implement this MCMC scheme (don’t {don’t do this at home}) – check the R programme com3d that I have uploaded to the course website. My illustration, see Figures 3 and 4, use $\alpha = 2.222$, $\beta = 1.111$, $\gamma = 7.777$. Check the acceptance rate for this proposal distribution (I had about 75% in my implementation). Check that the chain actually converges even if it started far away from mainstream (but requiring a longer burn-in phase; with $X_1 = 1000$ I need around 1,500 steps before equilibrium is reached).

(d) Display the histogram of the MCMC, after properly trimming away the burn-in period (taken in my illustration to be 1:1000). Compute mean, standard deviation and skewness for the $\pi_i$ distribution based on this.

(e) Use the above type of MCMC to simulate from the Poisson distribution with parameter $\lambda = 10$, say. Again use the symmetric $X^\text{cand}_{n+1} = i \pm 1$ proposal from $X_n = i$, as long as $i \geq 1$, and with the adjustment $Q_{0,0} = \frac{1}{2} = Q_{0,1}$ for the case of $i = 0$. Afterwards, when you are sure your algorithm works, carry out simulations and some distribution assessment for the ‘extended Poisson model’ with

$$
\pi_i = \frac{1}{K(\lambda, \gamma)(i!)^\gamma} \lambda^i \quad \text{for } i = 0, 1, 2, \ldots,
$$

where $\gamma = 1$ corresponds to the special Poisson case.
Figure 4: Histogram of the MCMC $X_{1001}, X_{1002}, \ldots, X_{\text{sim}}$, after having trimmed away the first 1000 iterations as burn-in phase (I used $\text{sim} = 1000 + 10^5$ here).

(f) The above MCMC scheme depends on the proposal distribution being symmetrical, that is, $Q_{i,j} = Q_{j,i}$ (as with proposals $i \pm 1$ being equally likely for the process that led to Figures 3 and 4 here). It is quite fruitful to allow non-symmetric proposal distribution, however, in many situations. Show that the revised scheme that uses

$$\text{pracc}(i, j) = \min \left( 1, \frac{\pi_j Q_{j,i}}{\pi_i Q_{i,j}} \right)$$

works properly, in the required sense that condition (B) of Exercise 7 is met.