1

Discrete-time Markov chains

1.1 The Markov property and its immediate consequences

Mathematics cannot be learned by lectures alone, anymore than piano playing can be learned by listening to a player.
C. Runge (1856–1927), German applied mathematician

Typically, the subject of Markov chains represents a logical continuation from a basic course of probability. We will study a class of *random processes* describing a wide variety of systems of theoretical and practical interest (and sometimes simply amusing). The fact that deep insight into the subject is possible without using sophisticated mathematical tools may also be an explanation of why Markov chains are popular in so many different disciplines which are seemingly remote from pure mathematics.

The basic model for the first half of the book will be a system which changes state in *discrete* time, according to some random mechanism. The collection of states is called a *state space* and throughout the whole book will be assumed finite or countable; we will denote it by I. Each $i \in I$ is called a state; our system will always be in one of these states. Sometimes we will know what state the system occupies and sometimes only that the system is in state i with some probability. Therefore it makes sense to introduce a *probability measure* or *probability distribution* (or, more simply, a distribution) on I. A probability distribution λ on I is simply a countable collection $(\lambda_i, i \in I)$ of non-negative numbers of total sum 1:

$$\lambda_i \geq 0, \ \sum_{i \in I} \lambda_i = 1. \tag{1.1}$$

We can think of a unit 'mass' spread over the set I where point i has mass λ_i. For that reason it is sometimes convenient to speak of a probability mass function $i \in I \mapsto \lambda_i$. Then the probability of a set $J \subseteq I$ is $\lambda(J) = \sum_{j\in J} \lambda_j$.

If $\lambda_i = 1$ for some $i \in I$ and $\lambda_j = 0$ when $j \neq i$, the distribution is 'concentrated' at point i. Then the state of our system becomes 'deterministic'. We will denote such a distribution by δ_i (the Dirac measure being an extreme case).

Sometimes the condition $\sum_{i\in I} \lambda_i = 1$ is not fulfilled; then we simply say that λ is a *measure* on I. If the total mass $\sum_{i\in I} \lambda_i < \infty$, the measure is called finite and can be transformed into a probability distribution by the normalisation: $\widetilde{\lambda}_i = \lambda_i / \sum_{j\in I} \lambda_j$ which defines a probability measure on I, since $\sum_{i\in I} \widetilde{\lambda}_i = \sum_{i\in I} \lambda_i / \sum_{j\in I} \lambda_j = 1$. But even if $\sum_{i\in I} \lambda_i = \infty$ (i.e. the total mass is infinite), we still can assign a finite value $\lambda(J) = \sum_{i\in J} \lambda_i$ to finite subsets $J \subset I$.

The random mechanism through which a change of state occurs is described by a *transition matrix* P, with entries p_{ij}, $i, j \in I$. Entry p_{ij} gives the probability that the system will change state i to j in a unit of time. That is, p_{ij} is the conditional probability that the system will occupy state j at the next time-step given that it is currently in state i. Hence, we have that each entry in P is non-negative but not greater than 1, and the sum of entries along every row equals 1:

$$0 \leq p_{ij} \leq 1 \text{ for all } i, j \in I \text{ and } \sum_{j\in I} p_{ij} = 1 \text{ for all } i \in I. \tag{1.2}$$

A matrix P with these properties is called *stochastic*. By analogy, a probability distribution (λ_i) on I is often called a *stochastic vector*. Then a stochastic matrix is one where every row is a stochastic vector.

Example 1.1.1 The simplest case is 2×2 (a two-state space). Without loss of generality, we may think that the states are 0 and 1: then the entries will be p_{ij}, $i, j = 0, 1$. Here, the stochastic matrix has the form

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

where $0 \leq \alpha, \beta \leq 1$. In particular, $\alpha = \beta = 0$ gives the identity matrix $\mathbf{I}$ and $\alpha = \beta = 1$ the anti-diagonal matrix:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A system with the identity transition matrix remains in the initial state forever; in the anti-diagonal case it flips state every time, from 0 to 1 and *vice versa*.

On the other hand, $\alpha = \beta = 1/2$ gives the matrix

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

In this case the system may stay in its state or change it with equal probabilities.

It is convenient to represent the transition matrix by a diagram where arrows show possible transitions and are labelled with the corresponding transition probabilities (arrows leading back to their own origin are often omitted as well as labels for deterministic transitions). See Figure 1.1, top.

La Dolce Beta
(From the series *'Movies that never made it to the Big Screen'*.)

Example 1.1.2 The 4×4 matrix

$$\begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is represented in Figure 1.1, bottom.

The time will take values $n = 0, 1, 2, \ldots$. To complete the picture, we have to specify in what state our system is at the initial time $n = 0$. Typically, we will assume that the system at time $n = 0$ is in state i with probability λ_i for some given 'initial' distribution λ on I.

Denote by X_n the state of our system at time n. The rules specifying a Markov chain with initial distribution λ and transition matrix P are that

(i) X_0 has distribution λ:

$$\mathbb{P}(X_0 = i) = \lambda_i, \text{ for all } i \in I,$$

(ii) more generally, for all n and $i_0, \ldots, i_n \in I$, the probabilities $\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n)$ that the system occupies states i_0, i_1, …, i_n at times 0, 1, …, n is written as a product

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}. \tag{1.3}$$

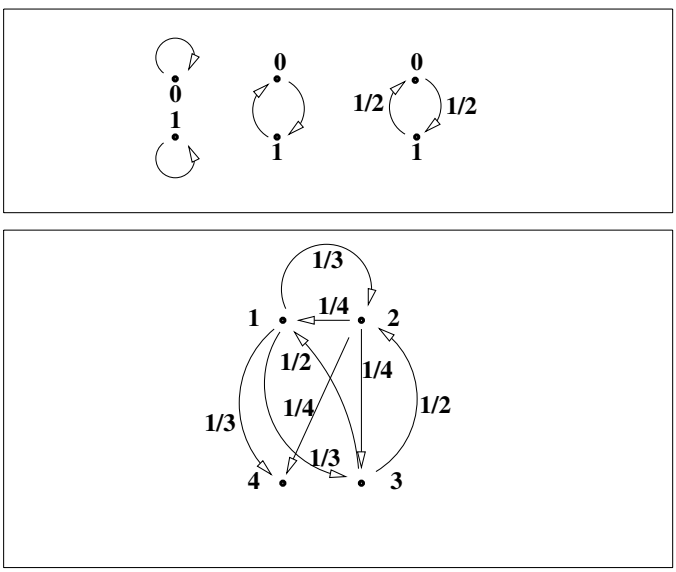

Fig. 1.1

Of course, (i) is a particular case of (ii), with $n = 0$.

An important corollary of (1.3) is the equation for the conditional probability $\mathbb{P}(X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i)$ that the state at time $n+1$ is j, given states $i_0, \ldots, i_{n-1}$ and $i_n = i$ at times $0, \ldots, n-1, n$:

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) \\
&\quad = \frac{\mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i, X_{n+1} = j)}{\mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i)} \\
&\quad = \frac{\lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i} p_{ij}}{\lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i}} = p_{ij}. \qquad (1.4)
\end{aligned}$$

That is, conditional on $X_0 = i_0, \ldots, X_{n-1} = i_{n-1}$ and $X_n = i$, we see X_{n+1} has the distribution $(p_{ij}, j \in I)$. In particular, the conditional distribution of X_{n+1} does not depend on $i_0, \ldots, i_{n-1}$, i.e., depends only on the state i at the last preceding time n.

Formula (1.4) illustrates the 'no memory' property of a Markov chain (only the current state counts for determining probabilities of future states).

Another consequence of (1.3) is an elegant formula involving matrix multiplication for the marginal probability distribution of X_n. Here we ask the question: what is the probability $\mathbb{P}(X_n = j)$ that at time n our system is in state j? For example, for $n = 1$ we can write:

$$\mathbb{P}(X_1 = j) = \sum_{i \in I} \mathbb{P}(X_0 = i, X_1 = j),$$

by considering all possible initial states i. In fact, the events

$$\{\text{state } i \text{ at time } 0,\ \text{state } j \text{ at time } 1\}$$

do not intersect for different $i \in I$ and their union gives the event

$$\{\text{state } j \text{ at time } 1\}.$$

Now use (1.3) and recall the rules of matrix algebra:

$$\sum_{i\in I}\mathbb{P}(X_0=i,X_1=j)=\sum_{i\in I}\lambda_i p_{ij}=(\lambda P)_j.$$

By a direct calculation, this formula is extended to a general n:

$$\begin{aligned}\mathbb{P}(X_n=j) &= \sum_{i_0,\dots,i_{n-1}}\mathbb{P}(X_0=i_0,\dots,X_{n-1}=i_{n-1},X_n=j)\\ &= \sum_{i_0,\dots,i_{n-1}}\lambda_{i_0}p_{i_0i_1}\cdots p_{i_{n-1}j}=(\lambda P^n)_j,\end{aligned}\tag{1.5}$$

where P^n is the nth power of the matrix P. That is, the stochastic vector describing the distribution of X_n is obtained by applying the matrix P^n to the initial stochastic vector λ.

Then, similarly,

$$\begin{aligned}&\mathbb{P}(X_n=i,X_{n+1}=j)\\ &\quad=\sum_{i_0,\dots,i_{n-1}}\mathbb{P}(X_0=i_0,\dots,X_{n-1}=i_{n-1},X_n=i,X_{n+1}=j)\\ &\quad=\sum_{i_0,\dots,i_{n-1}}\lambda_{i_0}p_{i_0i_1}\cdots p_{i_{n-1}i}p_{ij}=(\lambda P^n)_i\,p_{ij},\end{aligned}$$

and, hence

$$\mathbb{P}(X_{n+1}=j|X_n=i)=\frac{\mathbb{P}(X_n=i,X_{n+1}=j)}{\mathbb{P}(X_n=i)}=\frac{(\lambda P^n)_i\,p_{ij}}{(\lambda P^n)_i}=p_{ij}.\tag{1.6}$$

In other words, the entry p_{ij} is the conditional probability that the state at the next time-step is j given that at the preceding one it is i.

Moreover,

$$\begin{aligned}&\mathbb{P}(X_0=i,X_n=j)\\ &\quad=\sum_{i_1,\dots,i_{n-1}}\mathbb{P}(X_0=i,X_1=i_1,\dots,X_{n-1}=i_{n-1},X_n=j)\\ &\quad=\sum_{i_1,\dots,i_{n-1}}\lambda_i p_{ii_1}\cdots p_{i_{n-1}j}=\lambda_i(P^n)_{ij},\end{aligned}$$

and

$$\mathbb{P}(X_n=j|X_0=i)=\frac{\mathbb{P}(X_0=i,X_n=j)}{\mathbb{P}(X_0=i)}=\frac{\lambda_i(P^n)_{ij}}{\lambda_i}=(P^n)_{ij}.\tag{1.7}$$

That is, the entry $(P^n)_{ij}$ of matrix P^n gives the n-step transition probability from state i to j. We also denote it sometimes by $p_{ij}^{(n)}$.

More generally,

$$\mathbb{P}(X_k = i, X_{n+k} = j) = (\lambda P^k)_i (P^n)_{ij}$$

and

$$\mathbb{P}(X_{k+n} = j | X_k = i) = \frac{\mathbb{P}(X_k = i, X_{k+n} = j)}{\mathbb{P}(X_k = i)} = \frac{(\lambda P^k)_i (P^n)_{ij}}{(\lambda P^k)_i} = (P^n)_{ij}. \tag{1.8}$$

A corollary of this observation is that the power P^n of a stochastic matrix is again stochastic, viz. $\sum_{j\in I} p_{ij}^{(n)} = 1$ for all $i \in I$. Of course, this fact can be verified directly:

$$\sum_{j\in I} p_{ij}^{(n)} = \sum_{i_1,\dots,i_{n-1},j} p_{ii_1}\cdots p_{i_{n-1}j} = \sum_{i_1} p_{ii_1} \cdots \sum_{j} p_{i_{n-1}j} = 1$$

as at each step (beginning with $\sum_j$) we get the sum 1, owing to (1.2).

Another consequence is that if we apply to a stochastic vector a stochastic matrix (P or more generally P^n), we obtain another stochastic vector. Again, direct calculation confirms this:

$$\sum_j (\lambda P^n)_j = \sum_{i,j} \lambda_i (P^n)_{ij} = \sum_i \lambda_i \sum_j (P^n)_{ij} = \sum_i \lambda_i = 1.$$

An ultimate generalisation of (1.3) is the formula

$$\begin{aligned} &\mathbb{P}\big(X_{k_1} = i_1, X_{k_2} = i_2, \dots, X_{k_n} = i_n\big) \\ &\quad = (\lambda P^{k_1})_{i_1} \big(P^{k_2-k_1}\big)_{i_1 i_2} \cdots \big(P^{k_n - k_{n-1}}\big)_{i_{n-1} i_n} \end{aligned} \tag{1.9}$$

valid for all times $0 \le k_1 < k_2 < \cdots < k_n$ and states $i_1, \dots, i_n \in I$.

It is now time to summarise our findings. Suppose that $\lambda = (\lambda_i)$ is a stochastic vector and $P = (p_{ij})$ a transition matrix on I. The random state X_n at time n is considered as a random variable with values in I.

Definition 1.1.3 A sequence of random variables X_n with values in a finite or countable set I is a *discrete-time Markov chain* (DTMC), or a *Markov chain* for short, with the initial distribution λ and transition matrix P if, for all $i_0, \dots, i_n \in I$, the joint probability $\mathbb{P}(X_0 = i_0, \dots, X_n = i_n)$ is given by formula (1.3). In this case we also say that (X_n) is Markov (λ, P) or call it a (λ, P) Markov chain.

Theorem 1.1.4 *If (X_n) is Markov (λ, P), then:*

(i) *the conditional probability*

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i)$$

is equal to the conditional probability $\mathbb{P}(X_{n+1} = j|X_n = i)$ *and coincides with* p_{ij}. *In particular, the conditional distribution of* X_{n+1} *given that* $X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i$ *does not depend on* $i_0, \dots, i_{n-1}$ *and coincides with* $(p_{ij}, j \in I)$, *i.e. with row* i *of* P;

(ii) *the probability* $\mathbb{P}(X_n = i)$ *that the state at time* n *is* i *equals* $(\lambda P^n)_i$;

(iii) *the entry* $p_{ij}^{(n)}$ *of matrix* P^n *corresponds to the conditional probability* $\mathbb{P}(X_{k+n} = j|X_k = i)$, *i.e. gives the* n*-step transition probability from* i *to* j;

(iv) *the general probability*

$$\mathbb{P}\big(X_{k_1} = i_1, X_{k_2} = i_2, \dots, X_{k_n} = i_n\big)$$

is given by (1.9).

Example 1.1.5 Suppose that all rows of P are the same, i.e. $p_{ij} = p_j$ does not depend on i. In addition, suppose that $\lambda_j = p_j$, i.e. λ coincides with the row of P. Then, by (1.3)

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = p_{i_0} p_{i_1} \cdots p_{i_n}.$$

Also, in this example $P^n = P$, as

$$p_{ij}^{(n)} = \sum_{i_1, \dots, i_{n-1}} p_{i_1} \cdots p_{i_{n-1}} p_j = \sum_{i_1} p_{i_1} \sum_{i_2} p_{i_2} \cdots \sum_{i_{n-1}} p_{i_{n-1}} p_j = p_j,$$

owing to the fact that $\sum_{l \in I} p_l = 1$. Hence,

$$\mathbb{P}(X_n = j) = (\lambda P^n)_j = \sum_{i \in I} p_i p_{ij}^{(n)} = \sum_{i \in I} p_i p_j = p_j.$$

We see that

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \mathbb{P}(X_0 = i_0)\mathbb{P}(X_1 = i_1) \cdots \mathbb{P}(X_n = i_n).$$

That is (X_n) is a sequence of independent, identically distributed (IID) random variables.

Example 1.1.6 If P is diagonal then it must coincide with the identity matrix $\mathbf{I}$ where row i is given by the stochastic vector δ_i:

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

In this case, every power P^n again equals the identity matrix (this property is called idempotency; correspondingly, such a matrix P is called idempotent). Hence, by (1.5), $\mathbb{P}(X_n = i) = \lambda_i$. That is, the distribution of X_n is the same as X_0. In other words, the initial distribution is preserved in time.

Example 1.1.7 For a two-state DTMC, $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$, the entries of P^n can be found by a straightforward calculation. In fact, $P^n = P^{n-1}P$, which for entry $p_{00}^{(n)}$ yields

$$\begin{aligned} p_{00}^{(n)} &= p_{00}^{(n-1)}(1-\alpha) + p_{01}^{(n-1)}\beta \\ &= p_{00}^{(n-1)}(1-\alpha) + \left(1 - p_{00}^{(n-1)}\right)\beta = \beta + (1-\alpha-\beta)p_{00}^{(n-1)}. \end{aligned}$$

This is a recursion in n, with $p_{00}^{(0)} = 1$ and $p_{00}^{(1)} = 1-\alpha$. Hence,

$$p_{00}^{(n)} = A + B(1-\alpha-\beta)^n,$$

with

$$A + B = 1,\ A + B(1-\alpha-\beta) = 1-\alpha,$$

and, clearly,

$$p_{00}^{(n)} = \begin{cases} \dfrac{\beta}{\alpha+\beta} + \dfrac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n, & \text{if } \alpha+\beta > 0, \\ 1, & \text{if } \alpha = \beta = 0. \end{cases}$$

Entry $p_{11}^{(n)}$ is obtained by swapping α and β, and entries $p_{01}^{(n)}$ and $p_{10}^{(n)}$ as complements to 1.

Example 1.1.8 In the general case, we can use the eigenvalues and eigenvectors of P to find elements of P^n. Consider a 3×3 example

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

The eigenvalues are solutions to the characteristic equation:

$$\begin{aligned} \det \begin{pmatrix} -\mu & 1 & 0 \\ 0 & 2/3-\mu & 1/3 \\ 1/3 & 0 & 2/3-\mu \end{pmatrix} &= -\mu^3 + \frac{4}{3}\mu^2 - \frac{4}{9}\mu + \frac{1}{9} \\ &= -(\mu-1)\left(\mu^2 - \frac{1}{3}\mu + \frac{1}{9}\right) = 0, \end{aligned}$$

whence

$$\mu_0 = 1, \quad \mu_\pm = \frac{1 \pm i\sqrt{3}}{6}.$$

As the eigenvalues are distinct, matrix P is diagonalisable: there exists an invertible matrix D such that

$$D^{-1}PD = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (1+i\sqrt{3})/6 & 0 \\ 0 & 0 & (1-i\sqrt{3})/6 \end{pmatrix},$$

i.e.

$$P = D \begin{pmatrix} 1 & 0 & 0 \\ 0 & (1+i\sqrt{3})/6 & 0 \\ 0 & 0 & (1-i\sqrt{3})/6 \end{pmatrix} D^{-1}.$$

Then

$$P^n = D \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left[(1+i\sqrt{3})/6\right]^n & 0 \\ 0 & 0 & \left[(1-i\sqrt{3})/6\right]^n \end{pmatrix} D^{-1},$$

and each entry of P^n is a sum of the form

$$A + B\left(\frac{1+i\sqrt{3}}{6}\right)^n + C\left(\frac{1-i\sqrt{3}}{6}\right)^n.$$

The coefficients A, B and C may be complex; they vary from entry to entry and are found from the initial values $n = 0, 1, 2$. For $n = 0$, P^0 is the identity matrix (just as in the scalar case $p^0 = 1$ for any p ($p = 0$ included!)); for $n = 1$, we use the matrix P and for $n = 2$ we have to square it, to obtain P^2. For instance, suppose that the states are 1, 2 and 3; then the entries are $p_{ij}^{(n)}$, $i, j = 1, 2, 3$. Then, for $p_{12}^{(n)}$:

$$p_{12}^{(0)} = A + B + C = 0, \; p_{12}^{(1)} = A + B\frac{1+i\sqrt{3}}{6} + C\frac{1-i\sqrt{3}}{6} = 1,$$

and

$$p_{12}^{(2)} = A + B\left(\frac{1+i\sqrt{3}}{6}\right)^2 + C\left(\frac{1-i\sqrt{3}}{6}\right)^2 = \frac{2}{3}.$$

The calculations may be simplified if we get rid of imaginary parts (as all entries $p_{ij}^{(n)}$ of P^n are real non-negative). To this end, observe that $\mu_\pm$ are complex conjugate roots and write

$$\frac{1 \pm i\sqrt{3}}{6} = \frac{1}{3}\left(\frac{1 \pm i\sqrt{3}}{2}\right) = \frac{1}{3}\,\mathrm{e}^{\pm i\pi/3} = \frac{1}{3}\left(\cos\frac{\pi}{3} \pm i \sin\frac{\pi}{3}\right).$$

Then

$$\left(\frac{1\pm i\sqrt{3}}{6}\right)^n = \left(\frac{1}{3}\right)^n \mathrm{e}^{\pm in\pi/3} = \left(\frac{1}{3}\right)^n \left(\cos\frac{\pi n}{3} \pm i\sin\frac{\pi n}{3}\right),$$

and

$$p_{ij}^{(n)} = \alpha + \left(\frac{1}{3}\right)^n \left(\beta\cos\frac{\pi n}{3} + \gamma\sin\frac{\pi n}{3}\right),$$

where $\alpha = A$, $\beta = (B+C)$ and $\gamma = i(B-C)$ must be real. Again, we have the equations for $n = 0,1,2$; for $p_{12}^{(n)}$ they are

$$\alpha+\beta = 0,\ \alpha + \frac{1}{3}\left(\frac{1}{2}\beta + \frac{\sqrt{3}}{2}\gamma\right) = 1,\ \alpha + \frac{1}{9}\left(-\frac{1}{2}\beta + \frac{\sqrt{3}}{2}\gamma\right) = \frac{2}{3},$$

whence

$$\alpha = \frac{3}{7},\ \beta = -\frac{3}{7},\ \gamma = \frac{9}{7}\sqrt{3}.$$

In particular, $\lim_{n\to\infty} p_{12}^{(n)} = 3/7$.

Example 1.1.9 Consider another 3×3 matrix

$$P = \begin{pmatrix} 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

Here the characteristic equation is:

$$-\mu^3 + \frac{4}{3}\mu^2 - \frac{1}{3}\mu = -(\mu-1)\left(\mu - \frac{1}{3}\right)\mu = 0,$$

with the eigenvalues

$$\mu_0 = 1,\ \mu_1 = \frac{1}{3},\ \mu_2 = 0.$$

Hence, the entries $p_{ij}^{(n)}$ have a simple form

$$p_{ij}^{(n)} = A + B\left(\frac{1}{3}\right)^n + C\cdot 0^n.$$

Again we use three initial conditions, with P^0, P and P^2. For instance, for $p_{11}^{(n)}$:

$$A+B+C = 1,\ A + \frac{1}{3}B = \frac{1}{3},\ A + \left(\frac{1}{3}\right)^2 B = \frac{1}{3},$$