Decision-Making & Modelling Under Uncertainty (DMU)

Oliver Maclaren (oliver.maclaren@auckland.ac.nz)

[10 lectures/tutorials]

- Decision-making under uncertainty [5/10]
  - Basic concepts
  - Risk, probability, utility
  - Statistical: extended setup
    - formulation & empirical risk approx.
    - minimax & Bayes
  - Tutorial sheet
- Modelling under uncertainty { models of risk & intervention [5/10]
  - probability, graphical models, & independence
  - causal interpretations of graphical models
  - stochastic process models (esp. Markov)
  - simulation & estimation tools
  - Tutorial sheet

Lecture 7 : Markov models in time

- DAGs & template DAGs
- Stochastic processes
- Markov chains
- Key concepts & definitions for Markov chains

Consider the following DAG for RVs $X_0, X_1, \ldots X_n, \ldots$

$$\boxed{X_0 \to X_1 \to \quad \ldots \; X_n \to \ldots}$$

In one sense this is quite a simple DAG!

However, we actually now have a potentially infinite sequence of random variables. Requires extra tools!

Above represents a dynamical system (see eg 7(1)) evolving over (eg) time

→ In the graphical* model literature this can be thought of as a $\boxed{\text{template}}$ graphical model of the form:

$$\boxed{X_n \to X_{n+1}} \qquad (\text{or } X_t \to X_{t+1})$$

This can be instantiated an arbitrary number of times, 'unrolling' it

(* see eg Koller & Friedman 2009)

Stochastic processes: math of temporal (etc) systems

In the probability & stochastic processes literature, this is an example of a stochastic process

In general, a $\boxed{\text{stochastic process}}$ is an $\boxed{\text{indexed collection}}$ (or sequence) of random variables:

$$\boxed{(X_t \mid t \in T)}$$

also written eg

$$\boxed{\{X_t : t \in T\}}$$

or

$$\boxed{(X_t)_{t \in T}}$$

etc.

for 'time' variable t in index set T.

The variables X_t each take values in the state space $\mathcal{X}$, $X_t \in \mathcal{X}$

Examples?

- IID trials
- Weather
- Stock prices
- Epidemic models

etc!

→ see tutorial 2

Classifications of stochastic processes

The state space X can be $\begin{cases} \text{discrete} \\ \text{or} \\ \text{continuous} \end{cases}$

The index (time) set T can be $\begin{cases} \text{discrete} \\ \text{or} \\ \text{continuous} \end{cases}$

Here we will just consider $\boxed{\text{discrete}}$ time, $\boxed{\text{discrete}}$ state stochastic processes for simplicity.

A realisation of a stochastic process is a particular value for the $\boxed{\text{whole}}$ sequence eg $(X_1, X_2, .., X_n, ..) = (1, 3, -1, ..)$

or (sunny, rainy, cloudy, ...) etc.

Markov stochastic processes

A Markov process is a stochastic process for which the future only depends on the current state & not the rest of the past, eg, for a discrete time process $(X_t)_{t \in T}$:

$$\boxed{X_{t+1} \perp\!\!\!\perp X_{(\dots,t-1)} \mid X_t} \text{ for all } t,$$

where $X_{(\dots,t-1)}$ denotes all the RVs up to $t-1$

ie future $\perp\!\!\!\perp$ past | present.

→ "The Markov property"

(just like for DAGs, though now a template $X_{t-1} \to X_t \to X_{t+1} \cdots$ to handle infinite number of nodes)

Markov chains

A discrete-state, discrete-time Markov process is called a $\boxed{\text{Markov chain}}$

The Markov property for Markov chains is, for $\boxed{\text{all } n}$:

$$\boxed{\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})}$$

ie

$$\boxed{p(x_n \mid x_{n-1}, \dots, x_0) = p(x_n \mid x_{n-1})}$$

(recall $\mathbb{P}(X = x_n) = p(x_n)$ notation)

This also implies, for all n

$$\boxed{p(x_n, x_{n-1}, \dots x_0) = \prod_{i=0}^{n-1} p(x_{i+1} \mid x_i)}$$

ie $X_i \to X_{i+1}$ as expected.

Homogeneous Markov chains

In principle, $P(x_{n+1} | x_n)$ might be different for each n....

→ A common further simplification is to assume this is not the case!

ie

$$\boxed{P(x_{n+1} | x_n) = P(x_1 | x_0) \quad \text{for all } n}$$

→ these are called $\boxed{\text{homogeneous}}$ Markov chains

we will restrict attention to homogenous Markov chains in this course

Transition probabilities

Using $\boxed{i, j}$ etc to $\boxed{\text{label}}$ possible states, we can then define the $\boxed{\text{transition probabilities}}$

eg {sunny, rainy, ...} ↕ {1, 2, ...}

$$\boxed{P_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i) = P(j|i)}$$

& the $\boxed{\text{transition matrix}}$ $\underline{P}$ as the matrix with ijth element $(\underline{P})_{ij}$:

$$\boxed{(\underline{P})_{ij} = P_{ij}}$$

eg

$$\begin{array}{cc} & \text{to} \\ & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \text{from} \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.2 & 0.7 & 0.1 \\ 0.3 & 0.5 & 0.2 \end{pmatrix} \end{array}$$

← rows sum to 1 (why?)

(could write $P_{i \to j}$)

Transition diagrams vs DAGs

- We often use a $\boxed{\text{different}}$ type of $\boxed{\text{graphical}}$ diagram to represent Markov chains:

→ $\boxed{\text{transition diagrams}}$ ←

- while $\boxed{\text{DAGs}}$ represent conditional independence relations between nodes representing $\boxed{\text{random variables}}$, $\boxed{\text{transition diagrams}}$ use nodes to represent possible $\boxed{\text{values}}$ of the $\boxed{\text{state}}$ random variable & allowable $\boxed{\text{transitions}}$ for $n \to n+1$ &, typically, the $\boxed{\text{transition probabilities}}$.

→ Also usually $\boxed{\text{cyclic!}}$

→ example

Example

Given homog. Markov chain: $\underline{P} = $

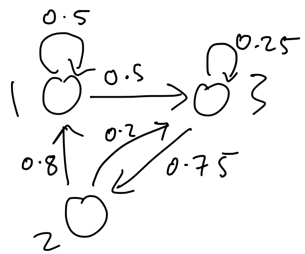
$$\begin{array}{c|ccc} & & \text{to} & \\ \text{from} & 1 & 2 & 3 \\ 1 & 0.5 & 0 & 0.5 \\ 2 & 0.8 & 0 & 0.2 \\ 3 & 0 & 0.75 & 0.25 \end{array}$$

$\boxed{\text{DAG}}$: $X_n \longrightarrow X_{n+1}$ } same for all MCs!

where X_n take values in $\mathcal{X} = \{1,2,3\}$, while nodes are $\boxed{\text{RVs}}$, graph acyclic

$\boxed{\text{VS}}$

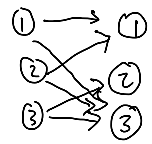
$\boxed{\text{Transition diagram}}$:



] represents $\underline{P}$ matrix in detail.

nodes are values of $\mathcal{X} = \{1,2,3\}$ ie poss. $\boxed{\text{values}}$ of each X_n. Graph cyclic in general*

(*can also do:



still values

X_n $\quad$ X_{n+1})

Simulating a Markov chain

→ we want to draw a sample of a full 'trajectory' from 0 to n ie a sample from

$$\boxed{p(x_n, x_{n-1}, \ldots x_0).} \quad \text{(joint)}$$

This factorises for any Markov chain (homogeneous or not) as

$$\boxed{\underbrace{p(x_n|x_{n-1})p(x_{n-1}|x_{n-2})\cdots}_{\text{repeated transition prob.}}\underbrace{p(x_0)}_{\text{initial prob.}}}$$

→ we can use the same 'direct sampling' algorithm from the appendix of L6 to sample trajectories (here homog.) eg

```
sample i value from P(X_0), record i
For k = 1 to n
    sample j value from P(X_{n+1}|X_n = i) *homog.
    record j
    set i ← j
```

→ recorded values give realisation.

n-step transitions

The previous algorithm gives a sample of a full trajectory

Sometimes we just want to know the probability we will be in some state after n steps, not worrying about (= averaging over!) any intermediate steps.

This leads to the n-step transition probabilities (assuming homogeneous*):

$$\boxed{p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i)}$$

For homogeneous Markov chains this gives an $\boxed{\text{n-step transition matrix}}$ $\underline{P}_n$ with elements:

$$\boxed{(\underline{P}_n)_{ij} = p_{ij}(n)}$$

(* hence no dependence on m)

Chapman-Kolmogorov equations

The n-step transition probabilities satisfy

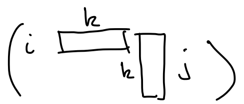
$$\boxed{p_{ij}(m+n) = \sum_k p_{ik}(m)\, p_{kj}(n)}$$

which is just

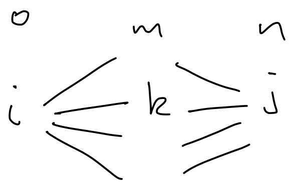
$$\boxed{\underline{P}_{m+n} = \underline{P}_m \cdot \underline{P}_n}$$

in terms of matrices.

Matrix multiplication!



→ First go $i \to k$ in m
then go $k \to j$ in n



sum over

Note: $\underline{P}_1 = \underline{P}$

$\underline{P}_2 = \underline{P}_1 \cdot \underline{P}_1 = \underline{P}^2$

$\vdots$

$$\boxed{\underline{P}_n = \underline{P}^n}$$

← multiply single-step $\underline{P}$ by itself n times!

Proof

In general, in probability theory:

- $p(x,y) = p(y|x)p(x)$
- $p(x,y|z) = p(y|x,z)p(x|z)$
- $p(x) = \sum_y p(x,y)$ (Here discrete vars. need $\int$ for cont.)

(also $\mathbb{P}(X=x) = p(x)$ etc – discrete vars)

Now $p_{ij}(m+n) = \mathbb{P}(X_{m+n}=j \,|\, X_0 = i)$

$$= \sum_k \mathbb{P}(X_{m+n}=j, X_m = k \,|\, X_0 = i)$$

$$= \sum_k \mathbb{P}(X_{m+n}=j \,|\, X_m = k, X_0 = i)\,\mathbb{P}(X_m = k \,|\, X_0 = i)$$

Markov property

$$= \sum_k \underbrace{\mathbb{P}(X_{m+n}=j \,|\, X_m = k)}_{p_{kj}}\,\underbrace{\mathbb{P}(X_m = k \,|\, X_0 = i)}_{p_{ik}(m)}$$

$$= \sum_k p_{ik}(m)\, p_{kj}(m+n) \qquad \square$$

Example $\underline{P}^n$

$$\underline{P} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

$$\underline{P}^1 = \underline{P} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

$$\underline{P}^2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

$$= \begin{bmatrix} 3/8 & 5/8 \\ 5/16 & 11/16 \end{bmatrix}$$

$$\underline{P}^3 = \begin{bmatrix} 22/64 & 42/64 \\ 21/64 & 43/64 \end{bmatrix}$$

$\vdots$

$$\underline{P}^n \underset{(n\to\infty)}{\longrightarrow} \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix}$$

Initial & marginal distributions

Recall our simulation started from an initial distribution $\mathbb{P}(X_0)$

Define μ_0 with $\boxed{\mu_0(i) = \mathbb{P}(X_0 = i)}$ to be a $\boxed{\text{row vector}}$ * representing the initial distribution

Then $\boxed{\mu_1 = \mu_0 \underline{P}}$ is a row vector giving the marginal distribution $\boxed{\mu_1(i) = \mathbb{P}(X_1 = i)}$ for X_1

In general have marginal distributions!

$$\boxed{\begin{aligned} \mu_n &= \mu_0 \underline{P}^n, \text{ where} \\ \mu_n(i) &= \mathbb{P}(X_n = i) \end{aligned}}$$

(* due to $\underline{P}_{\text{from, to}}$ convention)

Example

$$\underline{P} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

If $\mu_0 = [1 \quad 0]$ start at state 1 with prob $= 1$

$$\Rightarrow \mu_1 = [1/2 \quad 1/2]$$

If $\mu_0 = [1/2 \quad 1/2]$

$$\Rightarrow \mu_1 = [1/4 + 1/8 \quad 1/4 + 3/8]$$

$$= [3/8 \quad 5/8]$$

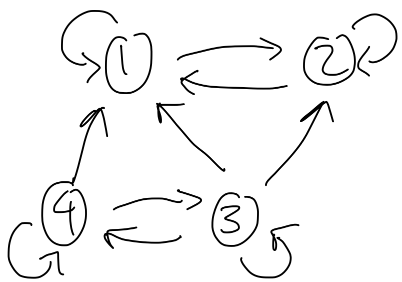
etc.

Exercise: find μ_2 for each of above μ_0.

Further properties/classifications of states

- We say a state i $\boxed{\text{communicates}}$ with a state j if $p_{ij}(n) > 0$ for some n (or 'i reaches j', 'j is accessible from i' etc) & write $\boxed{i \rightsquigarrow j}$
- if $\boxed{i \rightsquigarrow j \;\&\; j \rightsquigarrow i}$ then we say i & j $\boxed{\text{communicate}}$ & write $\boxed{i \leftrightsquigarrow j}$
- if $\boxed{i \rightsquigarrow j}$ & $\boxed{j \rightsquigarrow k}$ then $\boxed{i \rightsquigarrow k}$

$\Rightarrow$ hence can determine if $i \rightsquigarrow k$ by finding a directed path in the state transition diagram:



$1 \rightsquigarrow 2$

$1 \not\rightsquigarrow 3$

$4 \rightsquigarrow 2$

$2 \not\rightsquigarrow 4$

etc.

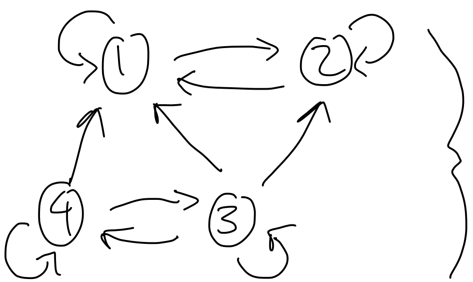
Communication classes

- If $\boxed{i \leftrightsquigarrow j \;\; \& \; j \leftrightsquigarrow k}$ then $\boxed{i \leftrightsquigarrow k}$

- Can determine if $i \leftrightsquigarrow j$ by finding two paths, one giving $i \rightsquigarrow j$ & one $j \rightsquigarrow i$

→ The set of states of a Markov chain can be divided into $\boxed{\text{disjoint}}$ $\boxed{\text{classes}}$ (sets) of states such that i & j are in the $\boxed{\text{same class}}$ if & only if $\boxed{i \leftrightsquigarrow j}$

Eg



Classes:

$\{1, 2\}, \{3, 4\}$

If all states communicate with all others, call Markov chain 'irreducible'

Trapping sets of states

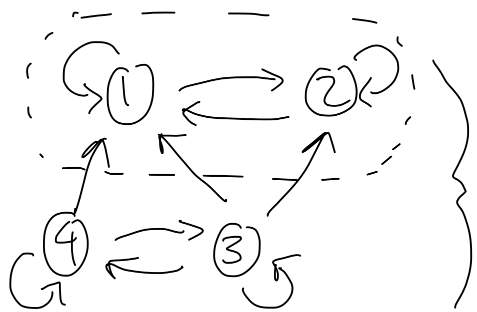
- A set of states is called a $\boxed{\text{trapping set}}$ (or closed set) if once you enter you never leave:

→ only have inwards arrows 'into' the set

→ can overlap! see example below

- A single single state that is a trapping set is called $\boxed{\text{absorbing}}$

Eg



Trapping sets:

$\{1, 2\}$

$\{1, 2, 3, 4\}$ ← whole state space counts

Other concepts

- Recurrent (persistent) & transient states
- Periodic & aperiodic states
- Ergodic states

→ not covered but

see Wasserman (2004)

Exercises

1. Given $\underline{P} = \begin{bmatrix} 1/3 & 2/3 \\ 1/4 & 3/4 \end{bmatrix}$

   calculate the 2- & 3-step transition matrices ($\underline{P}^2$ & $\underline{P}^3$)

2. Given $\underline{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

   Calculate $\underline{P}^2$

3. Given $\underline{P}$ from the previous prob. (2)

   & $\mu_0 = (1/2, 1/4, 1/4)$

   - find μ_1
   - find μ_2 via $\begin{cases} \mu_2 = \mu_0 \underline{P}^2 \\ \& \\ \mu_2 = \mu_1 \underline{P} \end{cases}$
   - are these the same? what justifies this? (eg a property or eqn etc)

Exercises

4. Consider

$$\underline{P} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 & ? \end{bmatrix}$$

- fill in the missing entry
- draw a state transition diagram labelling states as 1, 2, 3, 4
- determine any trapping sets & absorbing states
- determine the communication classes
- which states does state 4 communicate with?
- which states communicate with state 4?

Wasserman (2004) 'All of statistics'

23

Probability Redux: Stochastic Processes

23.1 Introduction

Most of this book has focused on IID sequences of random variables. Now we consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on one day is not independent of the temperature on the previous day.

A **stochastic process** $\{X_t : t \in T\}$ is a collection of random variables. We shall sometimes write $X(t)$ instead of X_t. The variables X_t take values in some set $\mathcal{X}$ called the **state space**. The set T is called the **index set** and for our purposes can be thought of as time. The index set can be discrete $T = \{0, 1, 2, \ldots\}$ or continuous $T = [0, \infty)$ depending on the application.

23.1 Example (IID observations). A sequence of IID random variables can be written as $\{X_t : t \in T\}$ where $T = \{1, 2, 3, \ldots, \}$. Thus, a sequence of IID random variables is an example of a stochastic process. ■

23.2 Example (The Weather). Let $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. A typical sequence (depending on where you live) might be

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \text{cloudy}, \cdots$$

This process has a discrete state space and a discrete index set. ■

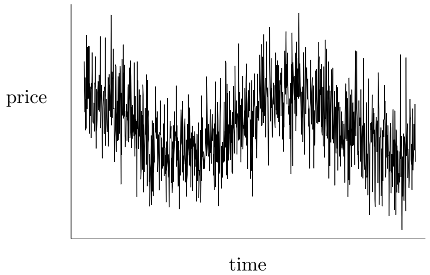


FIGURE 23.1. Stock price over ten week period.

23.3 Example (Stock Prices). Figure 23.1 shows the price of a fictitious stock over time. The price is monitored continuously so the index set T is continuous. Price is discrete but for all practical purposes we can treat it as a continuous variable. ■

23.4 Example (Empirical Distribution Function). Let $X_1, \ldots, X_n \sim F$ where F is some CDF on [0,1]. Let

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t)$$

be the empirical CDF. For any fixed value t, $\widehat{F}_n(t)$ is a random variable. But the whole empirical CDF

$$\left\{ \widehat{F}_n(t) : \; t \in [0,1] \right\}$$

is a stochastic process with a continuous state space and a continuous index set. ■

We end this section by recalling a basic fact. If $X_1, \ldots, X_n$ are random variables, then we can write the joint density as

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} f(x_i|\text{past}_i) \end{aligned} \qquad (23.1)$$

where $\text{past}_i = (X_1, \ldots, X_{i-1})$.

23.2 Markov Chains

A Markov chain is a stochastic process for which the distribution of X_t depends only on X_{t-1}. In this section we assume that the state space is discrete, either $\mathcal{X} = \{1, \ldots, N\}$ or $\mathcal{X} = \{1, 2, \ldots, \}$ and that the index set is $T = \{0, 1, 2, \ldots\}$. Typically, most authors write X_n instead of X_t when discussing Markov chains and I will do so as well.

23.5 Definition. *The process* $\{X_n : \; n \in T\}$ *is a* **Markov chain** *if*

$$\mathbb{P}(X_n = x \mid X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1}) \qquad (23.2)$$

for all n *and for all* $x \in \mathcal{X}$.

For a Markov chain, equation (23.1) simplifies to

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2) \cdots f(x_n|x_{n-1}).$$

A Markov chain can be represented by the following DAG:

$$X_0 \longrightarrow X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_n \longrightarrow \cdots$$

Each variable has a single parent, namely, the previous observation.

The theory of Markov chains is a very rich and complex. We have to get through many definitions before we can do anything interesting. Our goal is to answer the following questions:

1. When does a Markov chain "settle down" into some sort of equilibrium?
2. How do we estimate the parameters of a Markov chain?
3. How can we construct Markov chains that converge to a given equilibrium distribution and why would we want to do that?

We will answer questions 1 and 2 in this chapter. We will answer question 3 in the next chapter. To understand question 1, look at the two chains in Figure 23.2. The first chain oscillates all over the place and will continue to do so forever. The second chain eventually settles into an equilibrium. If we constructed a histogram of the first process, it would keep changing as we got more and more observations. But a histogram from the second chain would eventually converge to some fixed distribution.

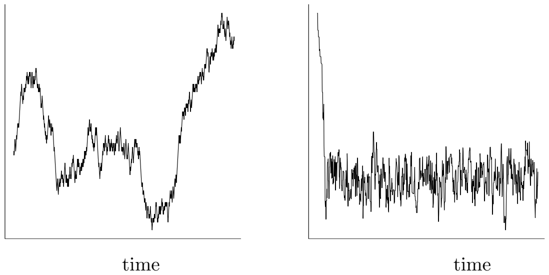


FIGURE 23.2. Two Markov chains. The first chain does not settle down into an equilibrium. The second does.

TRANSITION PROBABILITIES. The key quantities of a Markov chain are the probabilities of jumping from one state into another state. A Markov chain is **homogeneous** if $\mathbb{P}(X_{n+1} = j | X_n = i)$ does not change with time. Thus, for a homogeneous Markov chain, $\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i)$. We shall only deal with homogeneous Markov chains.

23.6 Definition. *We call*

$$p_{ij} \equiv \mathbb{P}(X_{n+1} = j | X_n = i) \tag{23.3}$$

the **transition probabilities.** *The matrix* $\mathbf{P}$ *whose* (i, j) *element is* p_{ij} *is called the* **transition matrix.**

We will only consider homogeneous chains. Notice that $\mathbf{P}$ has two properties: (i) $p_{ij} \geq 0$ and (ii) $\sum_i p_{ij} = 1$. Each row can be regarded as a probability mass function.

23.7 Example (Random Walk With Absorbing Barriers)**.** Let $\mathcal{X} = \{1, \ldots, N\}$. Suppose you are standing at one of these points. Flip a coin with $\mathbb{P}(\text{Heads}) = p$ and $\mathbb{P}(\text{Tails}) = q = 1 - p$. If it is heads, take one step to the right. If it is tails, take one step to the left. If you hit one of the endpoints, stay there. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad \blacksquare$$

23.8 Example. Suppose the state space is $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. Then X_1, X_2, ... represents the weather for a sequence of days. The weather today clearly depends on yesterday's weather. It might also depend on the weather two days ago but as a first approximation we might assume that the dependence is only one day back. In that case the weather is a Markov chain and a typical transition matrix might be

	Sunny	Cloudy
Sunny	0.4	0.6
Cloudy	0.8	0.2

For example, if it is sunny today, there is a 60 per cent chance it will be cloudy tomorrow. ■

Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i) \tag{23.4}$$

be the probability of of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i, j) element is $p_{ij}(n)$. These are called the **n-step transition probabilities.**

23.9 Theorem (The Chapman-Kolmogorov equations)**.** *The n-step probabilities satisfy*

$$p_{ij}(m + n) = \sum_k p_{ik}(m) p_{kj}(n). \tag{23.5}$$

PROOF. Recall that, in general,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y | X = x).$$

This fact is true in the more general form

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z)\mathbb{P}(Y = y | X = x, Z = z).$$

Also, recall the law of total probability:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y).$$

Using these facts and the Markov property we have

$$\begin{aligned} p_{ij}(m+n) &= \mathbb{P}(X_{m+n}=j|X_0=i) \\ &= \sum_k \mathbb{P}(X_{m+n}=j, X_m=k|X_0=i) \\ &= \sum_k \mathbb{P}(X_{m+n}=j|X_m=k, X_0=i)\mathbb{P}(X_m=k|X_0=i) \\ &= \sum_k \mathbb{P}(X_{m+n}=j|X_m=k)\mathbb{P}(X_m=k|X_0=i) \\ &= \sum_k p_{ik}(m)p_{kj}(n). \quad \blacksquare \end{aligned}$$

Look closely at equation (23.5). This is nothing more than the equation for matrix multiplication. Hence we have shown that

$$\mathbf{P}_{m+n} = \mathbf{P}_m\mathbf{P}_n. \tag{23.6}$$

By definition, $\mathbf{P}_1 = \mathbf{P}$. Using the above theorem, $\mathbf{P}_2 = \mathbf{P}_{1+1} = \mathbf{P}_1\mathbf{P}_1 = \mathbf{PP} = \mathbf{P}^2$. Continuing this way, we see that

$$\mathbf{P}_n = \mathbf{P}^n \equiv \underbrace{\mathbf{P}\times\mathbf{P}\times\cdots\times\mathbf{P}}_{\text{multiply the matrix n times}}. \tag{23.7}$$

Let $\mu_n = (\mu_n(1), \ldots, \mu_n(N))$ be a row vector where

$$\mu_n(i) = \mathbb{P}(X_n = i) \tag{23.8}$$

is the marginal probability that the chain is in state i at time n. In particular, μ_0 is called the **initial distribution.** To simulate a Markov chain, all you need to know is μ_0 and $\mathbf{P}$. The simulation would look like this:

Step 1: Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = i) = \mu_0(i)$.
Step 2: Denote the outcome of step 1 by i. Draw $X_1 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_1 = j|X_0 = i) = p_{ij}$.
Step 3: Suppose the outcome of step 2 is j. Draw $X_2 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_2 = k|X_1 = j) = p_{jk}$.
And so on.

It might be difficult to understand the meaning of μ_n. Imagine simulating the chain many times. Collect all the outcomes at time n from all the chains. This histogram would look approximately like μ_n. A consequence of theorem 23.9 is the following:

23.10 Lemma. *The marginal probabilities are given by*

$$\mu_n = \mu_0\mathbf{P}^n.$$

PROOF.

$$\begin{aligned} \mu_n(j) &= \mathbb{P}(X_n = j) \\ &= \sum_i \mathbb{P}(X_n = j|X_0 = i)P(X_0 = i) \\ &= \sum_i \mu_0(i)p_{ij}(n) = \mu_0\mathbf{P}^n. \quad \blacksquare \end{aligned}$$

Summary of Terminology

1. Transition matrix: $\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j|X_n = i) = p_{ij}$.
2. n-step matrix: $\mathbf{P}_n(i,j) = \mathbb{P}(X_{n+m} = j|X_m = i)$.
3. $\mathbf{P}_n = \mathbf{P}^n$.
4. Marginal: $\mu_n(i) = \mathbb{P}(X_n = i)$.
5. $\mu_n = \mu_0\mathbf{P}^n$.

STATES. The states of a Markov chain can be classified according to various properties.

23.11 Definition. *We say that i* **reaches** *j (or j is* **accessible** *from i) if $p_{ij}(n) > 0$ for some n, and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that i and j* **communicate.**

23.12 Theorem. *The communication relation satisfies the following properties:*

1. *$i \leftrightarrow i$.*
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of* **classes** *$\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \cdots$ where two states i and j communicate with each other if and only if they are in the same class.*

If all states communicate with each other, then the chain is called **irreducible**. A set of states is **closed** if, once you enter that set of states you never leave. A closed set consisting of a single state is called an **absorbing state.**

23.13 Example. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}, \{3\}$ and $\{4\}$. State 4 is an absorbing state. ∎

Suppose we start a chain in state i. Will the chain ever return to state i? If so, that state is called persistent or recurrent.

23.14 Definition. *State i is* **recurrent** *or* **persistent** *if*

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1.$$

Otherwise, state i is **transient**.

23.15 Theorem. *A state i is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty. \tag{23.9}$$

A state i is transient if and only if

$$\sum_n p_{ii}(n) < \infty. \tag{23.10}$$

PROOF. Define

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases}$$

The number of times that the chain is in state i is $Y = \sum_{n=0}^{\infty} I_n$. The mean of Y, given that the chain starts in state i, is

$$\mathbb{E}(Y|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{E}(I_n|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i|X_0 = i) = \sum_{n=0}^{\infty} p_{ii}(n).$$

Define $a_i = \mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i)$. If i is recurrent, $a_i = 1$. Thus, the chain will eventually return to i. Once it does return to i, we argue again that since $a_i = 1$, the chain will return to state i again. By repeating this argument, we conclude that $\mathbb{E}(Y|X_0 = i) = \infty$. If i is transient, then $a_i < 1$. When the chain is in state i, there is a probability $1 - a_i > 0$ that it will never return to state i. Thus, the probability that the chain is in state i exactly n times is $a_i^{n-1}(1 - a_i)$. This is a geometric distribution which has finite mean. ∎

23.16 Theorem. *Facts about recurrence.*

1. *If state i is recurrent and $i \leftrightarrow j$, then j is recurrent.*
2. *If state i is transient and $i \leftrightarrow j$, then j is transient.*
3. *A finite Markov chain must have at least one recurrent state.*
4. *The states of a finite, irreducible Markov chain are all recurrent.*

23.17 Theorem (Decomposition Theorem)**.** *The state space $\mathcal{X}$ can be written as the disjoint union*

$$\mathcal{X} = \mathcal{X}_T \bigcup \mathcal{X}_1 \bigcup \mathcal{X}_2 \cdots$$

where $\mathcal{X}_T$ are the transient states and each $\mathcal{X}_i$ is a closed, irreducible set of recurrent states.

23.18 Example (Random Walk)**.** Let $\mathcal{X} = \{\ldots, -2, -1, 0, 1, 2, \ldots, \}$ and suppose that $p_{i,i+1} = p$, $p_{i,i-1} = q = 1 - p$. All states communicate, hence either all the states are recurrent or all are transient. To see which, suppose we start at $X_0 = 0$. Note that

$$p_{00}(2n) = \binom{2n}{n} p^n q^n \tag{23.11}$$

since the only way to get back to 0 is to have n heads (steps to the right) and n tails (steps to the left). We can approximate this expression using Stirling's formula which says that

$$n! \sim n^n \sqrt{n} e^{-n} \sqrt{2\pi}.$$

Inserting this approximation into (23.11) shows that

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{n\pi}}.$$

It is easy to check that $\sum_n p_{00}(n) < \infty$ if and only if $\sum_n p_{00}(2n) < \infty$. Moreover, $\sum_n p_{00}(2n) = \infty$ if and only if $p = q = 1/2$. By Theorem (23.15), the chain is recurrent if $p = 1/2$ otherwise it is transient. ∎