APTS Applied Stochastic Processes

(Notes originally produced by Wilfrid Kendall; some material due to Stephen Connor, Christina Goldschmidt, Matt Roberts, Amanda Turner and Nicholas Georgiou)

Hugo Lo[1]

February 14 2026

[1]Department of Statistical Science, UCL, chak.lo@ucl.ac.uk

Contents

Preliminary Material

Chapter 1

Introduction

This module will introduce students to two important notions in stochastic processes — reversibility and martingales — identifying the basic ideas, outlining the main results and giving a flavour of some of the important ways in which these notions are used in statistics.

> Probability provides one of the major underlying languages of statistics, and purely probabilistic concepts often cross over into the statistical world. So statisticians need to acquire some fluency in the general language of probability.

1.1 Learning outcomes

After successfully completing this module an APTS student will be able to:

- describe and calculate with the notion of a reversible Markov chain, both in discrete and continuous time;
- describe the basic properties of discrete-parameter martingales and check whether the martingale property holds;
- recall and apply some significant concepts from martingale theory;
- explain how to use Foster-Lyapunov criteria to establish recurrence and speed of convergence to equilibrium for Markov chains.

> These outcomes interact interestingly with various topics in applied statistics. However the most important aim of this module is to help students to acquire general awareness of further ideas from probability as and when that might be useful in their further research.

Chapter 2

Expectation and probability

For most APTS students most of this material should be well-known:

- Probability and conditional probability;
- Basic expectation and conditional expectation;
- discrete versus continuous (sums and integrals);
- limits versus expectations.

It is set out here, describing key ideas rather than details, in order to establish a sound common basis for the module.

> This material uses a two-part format. The main text presents the theory, often using itemized lists. Indented panels such as this one present commentary and useful exercises (announced by "**Test understanding**"). It is likely that you will have seen most, if not all, of the preliminary material at undergraduate level. However syllabi are not uniform across universities; if some of this material is not well-known to you then:
> - read through it to pick up the general sense and notation;
> - supplement by reading (for example) the first five chapters of Grimmett and Stirzaker [2001].

2.1 Probability

- Sample space Ω of possible outcomes;
- Probability $\mathbb{P}$ assigns a number between 0 and 1 inclusive (the *probability*) to each (sensible) subset $A \subseteq \Omega$ (we say A is an *event*);
- Advanced (measure-theoretic) probability takes great care to specify what *sensible* means: A has to belong to a pre-determined σ-*algebra* $\mathcal{F}$, a family of subsets closed under countable unions and complements, often generated by open sets. We shall avoid these technicalities, though it will later be convenient to speak of σ-algebras $\mathcal{F}_t$ as a shorthand for "information available by time t".

- Rules of probability:
  - **Normalization:** $\mathbb{P}(\Omega) = 1$;
  - **σ-additivity:** if A_1, A_2 ... form a disjoint sequence of events then

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \sum_i \mathbb{P}(A_i)\,.$$

Example of a sample space: $\Omega = (-\infty, \infty)$.
To define a probability $\mathbb{P}$, we could for example start with

$$\mathbb{P}\left((a,b)\right) = \tfrac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\,\mathrm{d}u$$

and then use the rules of probability to determine probabilities for all manner of sensible subsets of $(-\infty, \infty)$.
In this example a "natural" choice for $\mathcal{F}$ is the family of all sets generated from intervals by indefinitely complicated countably infinite combinations of countable unions and complements.
Test understanding: use the rules of probability to explain
a. why $\mathbb{P}\left(\emptyset\right) = 0$,}
b. why $\mathbb{P}\left(A^c\right) = 1 - \mathbb{P}\left(A\right)$ if $A^c = \Omega \setminus A$, and
c. why it makes no sense in general to try to extend σ-additivity to uncountable unions such as $(-\infty, \infty) = \bigcup_x \{x\}$.

2.2 Conditional probability

- We declare the *conditional probability* of A given B to be $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, and declare the case when $\mathbb{P}(B) = 0$ as undefined.

Actually we *often* use limiting arguments to make sense of cases when $\mathbb{P}\left(B\right) = 0$.

- **Bayes:** if $B_1, B_2, \dots$ is an exhaustive disjoint partition of Ω then

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}\,.$$

- Conditional probabilities are clandestine random variables! Let X be the Bernoulli random variable which indicates event B; that is, X takes value 1 if B occurs and value 0 if B does not occur. Consider the conditional probability of A given information of whether or not B occurs: it is random, being $\mathbb{P}(A|B)$ if $X = 1$ and $\mathbb{P}(A|B^c)$ if $X = 0$.

Test understanding: write out an explanation of why Bayes' theorem is a completely obvious consequence of the definitions of probability and conditional probability.
The idea of conditioning is developed in probability theory to the point where this notion (that conditional probabilities are random variables) becomes entirely natural not artificial.
Test understanding: establish the law of inclusion and exclusion: if $A_1, \ldots, A_n$ are potentially overlapping events then

$$\begin{aligned}\mathbb{P}(A_1 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n) \\ - (\mathbb{P}(A_1 \cap A_2) + \ldots + \mathbb{P}(A_i \cap A_j) + \ldots + \mathbb{P}(A_{n-1} \cap A_n)) \\ + \ldots - (-1)^n \mathbb{P}(A_1 \cap \ldots \cap A_n) \ .\end{aligned}$$

2.3 Expectation

Statistical intuition about expectation is based on *properties*:

- If $X \geq 0$ is a non-negative random variable then we can define its (possibly infinite) *expectation* $\mathbb{E}[X]$.
- If $X = X^+ - X^- = \max\{X, 0\} - \max\{-X, 0\}$ is such that $\mathbb{E}[X^{\pm}]$ are both finite then set $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. (The reason that we insist that $\mathbb{E}[X^{\pm}]$ are finite is that we can't make sense of $\infty - \infty$!)
- Familiar properties of expectation follow from
  - **linearity**: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
  - **monotonicity**: $\mathbb{P}(X \geq a) = 1$ implies $\mathbb{E}[X] \geq a$ for constants a, b.

Full definition of expectation takes 3 steps: obvious definition for Bernoulli random variables, finite range random variables by linearity, general case by monotonic limits $X_n \uparrow X$. The hard work lies in proving this is all consistent.
Test understanding: using the properties of expectation,

- deduce $\mathbb{E}[a] = a$ for constant a.
- show *Markov's inequality*:

$$\mathbb{P}(X \geq a) \leq \tfrac{1}{a}\mathbb{E}[X]$$

for $X \geq 0$, $a > 0$.

- Useful notation: for an event A write $\mathbb{E}[X; A] = \mathbb{E}[X\,\mathbf{1}_A]$, where $\mathbf{1}_A$ is the Bernoulli random variable indicating A.

So in the absolutely continuous case

$$\mathbb{E}[X; A] = \int_A x\, f_X(x)\, \mathrm{d}x$$

and in the discrete case

$$\mathbb{E}[X; X = k] = k\, \mathbb{P}(X = k).$$

- If X has countable range then $\mathbb{E}[X] = \sum_x \mathbb{P}(X = x)$.
- If X has *density* f_X then $\mathbb{E}[X] = \int x\, f_X(x)\, \mathrm{d}x$.

In the countable (=discrete) case, expectation is defined exactly when the sum converges absolutely.
When there is a density (=absolutely continuous case), expectation is defined exactly when the integral converges absolutely.

2.4 Independence

Events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

- This definition can be extended to more than two events: $A_1, A_2, \ldots, A_n$ are independent if for any set $J \subseteq \{1, \ldots, n\}$

$$\mathbb{P}(\cap_{j\in J} A_j) = \prod_{j\in J} \mathbb{P}(A_j)\ .$$

Note that it's *not* enough to simply ask for any two events A_i and A_j to be independent (i.e. pairwise independence)!
Test understanding: find a set of three events which are *pairwise* independent, but for which

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) \neq \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\mathbb{P}(A_3).$$

- If A and B and independent, with $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Test understanding: Show that
- if A and B are independent, then events A^c and B^c are independent;
- if $A_1, A_2, \ldots, A_n$ are independent then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - \prod_{i=1}^{n} \mathbb{P}(A_i^c)\ .$$

Random variables X and Y are independent if for all $x, y \in \mathbb{R}$

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

In this case, $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

The definition of independence of random variables is equivalent to

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

for any (Borel) sets $A, B \subset \mathbb{R}$.
Test understanding: Suppose that X and Y are independent, and are either both discrete or both absolutely continuous; show that

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

2.5 Generating functions

We're often interested in expectations of *functions* of random variables (e.g. recall that in the discrete case $\mathbb{E}[g(X)] = \sum_x g(x)\,\mathbb{P}(X = x)$).

Some functions are particularly useful:

- when $g(x) = z^x$ for some $z \geq 0$ we obtain the *probability generating function (pgf)* of X,

$$G_X(z) = \mathbb{E}\left[z^X\right];$$

- when $g(x) = e^{tx}$ we get the *moment generating function (mgf)* of X,

$$m_X(t) = \mathbb{E}\left[e^{tX}\right];$$

- when $g(x) = e^{itx}$, where $i = \sqrt{-1}$, we get the *characteristic function* of X,

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right].$$

Test understanding: Show that

- $\mathbb{E}[X] = G'_X(1)$ and $\mathbb{P}(X = k) = G_X^{(k)}(0)/k!$
  (where $G_X^{(k)}(0)$ means the k^{th} derivative of $G_X(z)$, evaluated at $z = 0$);
- $\mathbb{E}[X] = m'_X(0)$ and

$$m_X(t) = \sum_k \frac{\mathbb{E}\left[X^k\right]}{k!} t^k.$$

2.6 Uses of generating functions

Generating functions are helpful in many ways. In particular:

- They can be used to determine distributions;
- They can often provide an easy route to finding e.g. moments of a distribution (see the two exercises in the previous section!);

- They're useful when working with sums of independent random variables, since the generating function of a *convolution* of distributions is the product of their generating functions. So

$$G_{X+Y}(z) = G_X(z)G_Y(z)$$

etc.

Characteristic functions always uniquely determine distributions (i.e. there is a one-to-one correspondence between a distribution and its characteristic function); the same is true of pgfs and distributions on $\{0, 1, \dots\}$; mgfs are slightly more complicated, but *mostly* they can be used to identify a distribution. See Grimmett and Stirzaker [2001] for more on this.
Test understanding: show that if X and Y are independent random variables then

- $G_{X+Y}(z) = G_X(z)G_Y(z)$;
- $m_{X+Y}(t) = m_X(t)m_Y(t)$;
- $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

(Only one argument is needed to see all three results!)
Use the first of these as a quick method of proving that the sum of two independent Poisson random variables is itself Poisson.

2.7 Conditional Expectation (I): property-based definition

Conventional definitions treat two separate cases (discrete and absolutely continuous:

- $\mathbb{E}[X|Y = y] = \sum_x x\,\mathbb{P}(X = x|Y = y)$,
- $\mathbb{E}[X|Y = y] = \int x\, f_{X|Y=y}(x)\,\mathrm{d}x$.
- ... but what if X is mixed discrete/continuous? or worse?

Focus on *properties* to get unified approach: if $\mathbb{E}[X] < \infty$, we say $Z = \mathbb{E}[X|Y]$ if:

a. $\mathbb{E}[Z] < \infty$;

b. Z is a function of Y;

c. $\mathbb{E}[Z; A] = \mathbb{E}[X; A]$ for events A defined in terms of Y.

This defines $\mathbb{E}[X|Y]$ uniquely, up to events of probability 0.

- "$\mathbb{E}[Z] < \infty$" is needed to get a good definition of *any* kind of expectation;
- We could express "Z is a function of Y" etc. more formally using measure theory if we had to;
- We need (b) to rule out $Z = X$, for example.

Test understanding: verify that the discrete definition of conditional expectation satisfies the three properties (a), (b), (c). Hint: use A running through events $A = \{Y = y\}$ for y in the range of Y.

We can now define $\mathbb{E}[X|Y_1, Y_2, \ldots]$ simply by using "is a function of $Y_1, Y_2, \ldots$" and "defined in terms of $Y_1, Y_2, \ldots$", etc. Indeed we often write $\mathbb{E}[X|\mathcal{G}]$, where ($\sigma$-algebra) $\mathcal{G}$ represents information conveyed by a specified set of random variables and events.

Test understanding: suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed, with finite absolute mean $\mathbb{E}[|X_i|] < \infty$. Use symmetry and linearity to show that

$$\mathbb{E}[X_1|X_1 + \ldots + X_n] = \tfrac{1}{n}(X_1 + \ldots + X_n).$$

2.8 Conditional Expectation (II): some other properties

Many facts about conditional expectation follow easily from this property-based approach. For example:

- Linearity:

$$\mathbb{E}[aX + bY|Z] = a\,\mathbb{E}[X|Z] + b\,\mathbb{E}[Y|Z]\,;$$

- "Tower property":

$$\mathbb{E}[\mathbb{E}[X|Y, Z]\,|Y] = \mathbb{E}[X|Y]\,;$$

- "Taking out what is known":

$$\mathbb{E}[f(Y)X|Y] = f(Y)\,\mathbb{E}[X|Y]\,;$$

- ... and variations involving more than one or two conditioning random variables.

Test understanding: explain how these follow from the property-based definition. Hints:

- Use $\mathbb{E}[aX + bY; A] = a\,\mathbb{E}[X; A] + b\,\mathbb{E}[Y; A]$;
- Take a deep breath and use property (c) of conditional expectation twice. Suppose A is defined in terms of Y. Then

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[X|Y,Z]\,|Y]\,; A] = \mathbb{E}[\mathbb{E}[X|Y,Z]\,; A]$$

and

$$\mathbb{E}[\mathbb{E}[X|Y,Z]\,; A] = \mathbb{E}[X; A]\,.$$

- Just consider when f has finite range, and use the (finite) sum

$$\mathbb{E}[\mathbb{E}[f(Y)X|Y]\,; A] = \sum_t \mathbb{E}[\mathbb{E}[f(Y)X|Y]\,; A \cap \{f(Y) = t\}]\,.$$

But then use

$$\begin{aligned}\mathbb{E}[\mathbb{E}[f(Y)X|Y]\,; A \cap \{f(Y) = t\}] &= \mathbb{E}[\mathbb{E}[tX|Y]\,; A \cap \{f(Y) = t\}] \\ &= \mathbb{E}[t\,\mathbb{E}[X|Y]\,; A \cap \{f(Y) = t\}] \\ &= \mathbb{E}[f(Y)\,\mathbb{E}[X|Y]\,; A \cap \{f(Y) = t\}]\,.\end{aligned}$$

The general case now follows by approximation arguments.

2.9 Conditional Expectation (III): Jensen's inequality

This is powerful and yet rather easy to prove.

Theorem 2.1. *Let ϕ be a convex function ("curves upwards", or $\phi'' \geq 0$ if smooth). Suppose the random variable X is such that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|\phi(X)|] < \infty$. Then*

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]\ ,$$

and the same is true for conditional expectations:

$$\phi(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[\phi(X)|\mathcal{G}]$$

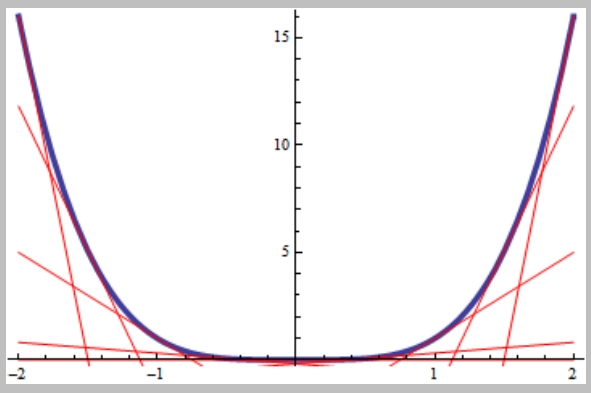
for any conditioning information $\mathcal{G}$.

Clue to proof: *any* convex function can be represented as supremum of all affine functions $ax + b$ lying below it.

Consider the simple convex function $\phi(x) = x^2$. We deduce that if X has finite second moment then

$$(\mathbb{E}[X|\mathcal{G}])^2 \leq \mathbb{E}\left[X^2|\mathcal{G}\right].$$

Here's a picture to illustrate the clue to the proof of Jensen's inequality in case $\phi(x) = x^4$:



2.10 Limits versus expectations

- Often the crux of a piece of mathematics is whether one can exchange limiting operations such as $\lim \sum$ and $\sum \lim$. Here are a few very useful results on this, expressed in the language of expectations (the results therefore apply to both sums and integrals).
- **Monotone Convergence Theorem**: If $\mathbb{P}(X_n \uparrow Y) = 1$ and $\mathbb{E}[X_1] > -\infty$ then

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}\left[\lim_n X_n\right] = \mathbb{E}[Y].$$

Note that the X_n must form an *increasing* sequence. We need $\mathbb{E}[X_1] > -\infty$.
Test understanding: consider case of $X_n = -1/(nU)$ for a fixed Uniform$(0,1)$ random variable.

- **Dominated Convergence Theorem**: If $\mathbb{P}(X_n \to Y) = 1$ and $|X_n| \leq Z$ where $\mathbb{E}[Z] < \infty$ then

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}\left[\lim_n X_n\right] = \mathbb{E}[Y].$$

Note that convergence need not be monotonic here or in the following.
Test understanding: explain why it would be enough to have finite upper and lower bounds $\alpha \leq X_n \leq \beta$.

- **Fubini's Theorem**: If $\mathbb{E}[|f(X,Y)|] < \infty$, X, Y are independent, $g(y) = \mathbb{E}[f(X,y)]$, $h(x) = \mathbb{E}[f(x,Y)]$ then

$$\mathbb{E}[g(Y)] = \mathbb{E}[f(X,Y)] = \mathbb{E}[h(X)].$$

Fubini exchanges expectations rather than an expectation and a limit.

- **Fatou's lemma**: If $\mathbb{P}(X_n \to Y) = 1$ and $X_n \geq 0$ for all n then

$$\mathbb{E}[Y] \leq \lim_n \inf_{m \geq n} \mathbb{E}[X_m].$$

Try Fatou if all else fails. Note that something like $X_n \geq 0$ is essential (any constant lower bound would suffice, though, it doesn't need to be 0).

Chapter 3

Markov chains

- Discrete-time countable-state-space basics:
  - Markov property, transition matrices;
  - irreducibility and aperiodicity;
  - transience and recurrence;
  - equilibrium equations and convergence to equilibrium.
- Discrete-time *countable*-state-space: why 'limit of sum need not always equal sum of limit'.
- Continuous-time countable-state-space: rates and Q-matrices.
- Definition and basic properties of Poisson counting process.

Instead of "countable-state-space" Markov chains, we'll use the shorter phrase "discrete Markov chains" or "discrete space Markov chains".

> If some of this material is not well-known to you, then invest some time in looking over (for example) chapter 6 of Grimmett and Stirzaker [2001].

3.1 Basic properties for discrete time and space case

- Markov chain $X = \{X_0, X_1, X_2, \ldots\}$: we say that X at time t is in state X_t.
- X must have the **Markov property**:

$$p_{xy} = p(x, y) = \mathbb{P}\left(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots\right)$$

  must depend only on x, y, not on rest of past. (Our chains will be *time-homogeneous*, meaning no t dependence either.)

- View states x as integers. More general countable discrete state-spaces can always be indexed by integers.
- We will soon see an example, "Markov's other chain", showing that we need to **insist** on the possibility of conditioning by further past $X_{t-1}, \ldots$ in the definition.

- Chain behaviour is specified by (a) initial state X_0 (could be random) and (b) table of **transition probabilities** p_{xy}.

- Important **matrix** structure: if p_{xy} are arranged in a matrix P then $(i,j)^{\text{th}}$ entry of $P^n = P \times \ldots \times P$ (n times) is $p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i)$.
  Equivalent: **Chapman-Kolmogorov equations**

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

- Note $\sum_y p_{xy} = 1$ by "law of total probability".
- **Test understanding**: show how the Chapman-Kolmogorov equations follow from considerations of conditional probability and the Markov property.

3.2 Example: Models for language following Markov

How to generate "random English" as a Markov chain:

- Take a large book in electronic form, for example Tolstoy's "War and Peace" (English translation).

- Use it to build a table of digram frequencies (digram = pair of consecutive letters).

- Convert frequencies into conditional probabilities of one letter following another, and use these to form a Markov chain to generate "random English".

It is an amusing if substantial exercise to use this as a prior for Bayesian decoding of simple substitution codes.

- To find a large book, try Project Gutenberg.
- Skill is required in deciding *which* characters to use: should one use all, or some, punctuation? Certainly need to use spaces.
- Trigrams would be more impressive than digrams. Indeed, one needs to work at the level of words to simulate something like English. Here is example output based on a children's fable:

> It was able to the end of great daring but which when Rapunzel was a guardian has enjoined on a time, after a faked morning departure more directly; over its days in a stratagem, which supported her hair into the risk of endless figures on supplanted sorrow. The prince's directive, to clamber down would come up easily, and perceived a grudge against humans for a convincing simulation of a nearby robotic despot. But then a computer typing in a convincing simulation of the traditional manner. However they settled in quality, and the prince thought for Rapunzel made its ward's face, that as she then a mere girl.

- Here are some word transition probabilities from the source used for the above example:

$$P(\text{round}|\text{all}) = P(\text{contact}|\text{all}) = 0.50$$

$$P(\text{hearing}|\text{ocean},) = P(\text{first},|\text{go}) = 1.00$$

$$P(\text{As}|\text{up}.) = P(\text{Every}|\text{day}.) = 1.00$$

$$P(\text{woman}|\text{young}) = P(\text{prince}.|\text{young}) = P(\text{man}|\text{young}) = 0.33.$$

3.3 (Counter)example: Markov's other chain

Conditional probability can be subtle. Consider:

- Independent Bernoulli $X_0, X_2, X_4, \ldots$ such that $\mathbb{P}(X_{2n} = \pm 1) = \frac{1}{2}$;
- Define $X_{2n+1} = X_{2n}X_{2n+2}$ for $n = 0, 1, \ldots$; these also form an independent identically distributed sequence.
- $\mathbb{P}(X_{n+1} = \pm 1 | X_n) = \frac{1}{2}$ for any $n \geq 1$.
- Chapman-Kolmogorov equations hold for any $0 \leq k \leq n + k$:

$$\mathbb{P}(X_{n+k} = \pm 1 | X_0) = \sum_{y=\pm 1} \mathbb{P}(X_{n+k} = \pm 1 | X_k = y)\, \mathbb{P}(X_k = y | X_0) \ .$$

- Nevertheless, $\mathbb{P}(X_2 = \pm 1 | X_1 = 1, X_0 = u)$ depends on $u = \pm 1$, so Markov property **fails** for X

This example is taken from Grimmett and Stirzaker [2001].
Note that, although $X_0, X_2, X_4, \ldots$ are independent and $X_1, X_3, X_5, \ldots$ are independent, the entire sequence of random variables $X_0, X_1, X_2, \ldots$ are most certainly *not* independent!
Test understanding by checking the calculations above.
It is usual in stochastic modelling to *start* by specifying that a given random process $X = \{X_0, X_1, X_2, \ldots\}$ is Markov, so this kind of issue is not often encountered in practice. However it is good to be aware of it: conditioning is a subtle concept and should be treated with respect!

3.4 Irreducibility and aperiodicity

- A discrete Markov chain is *irreducible* if for all i and j it has a positive chance of visiting j at some positive time, if it is started at i.

Consider the word game: change "good" to "evil" through other English words by altering just one letter at a time. Illustrative question (compare Gardner [1996]): does your vocabulary of 4-letter English words form an irreducible Markov chain under moves which attempt random changes of letters? You can find an algorithmic approach to this question in Knuth [1993].

- It is *aperiodic* if one cannot divide state-space into non-empty subsets such that the chain progresses through the subsets in a periodic way. Simple symmetric walk (jumps ± 1) is *not* aperiodic.

Equivalent definition: an irreducible chain X is aperiodic if its "independent double" $\{(X_0, Y_0), (X_1, Y_1), \ldots\}$ (for Y an independent copy of X) is irreducible.

- If the chain is not irreducible, then we can compute the chance of it getting from one state to another using *first passage equations*: if

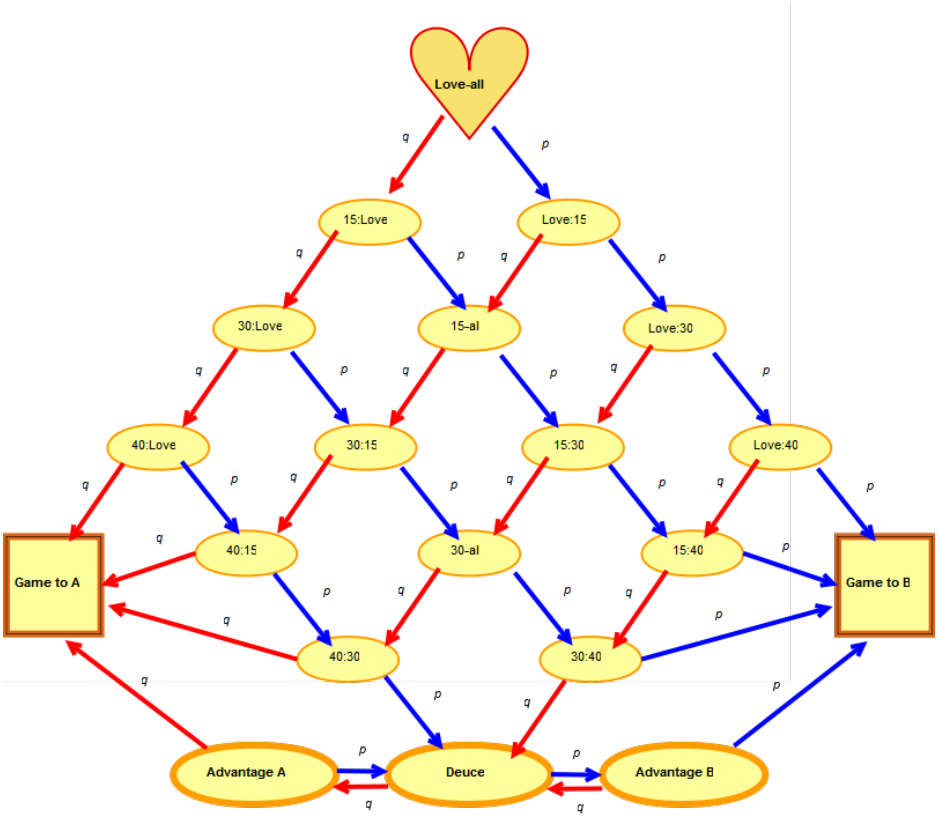
$$f_{ij} = \mathbb{P}\left(X_n = j \text{ for some positive } n | X_0 = i\right)$$

then solve linear equations for the f_{ij}

Because of the connection with matrices noted above, this can be cast in terms of rather basic linear algebra. First passage equations are still helpful in analyzing irreducible chains: for example the chance of visiting j *before* k is the same as computing f_{ij} for the modified chain which stops on hitting k.

3.5 Example: Markov tennis

How does probability of win by B depend on $p = \mathbb{P}\left(B \text{ wins point}\right)$?



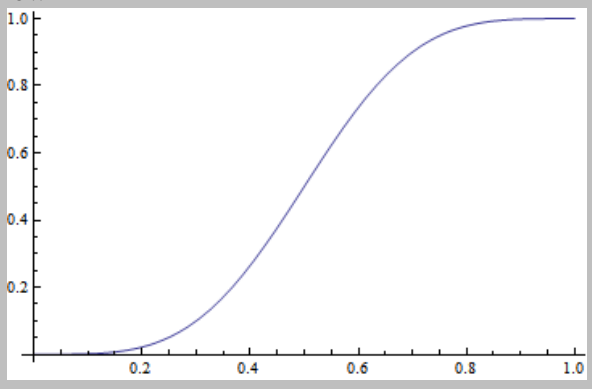
Use *first passage equations*, then solve linear equations for the f_{ij}, noting in particular

$$f_{\text{Game to A,Game to B}} = 0\,, \qquad f_{\text{Game to B,Game to B}} = 1\,.$$

I obtain

$$f_{\text{Love-All,Game to B}} = \frac{p^4(15-34p+28p^2-8p^3)}{1-2p+2p^2}\,,$$

graphed against p below:



3.6 Transience and recurrence

- Is it possible for a Markov chain X never to return to a starting state i? If so then that state is said to be *transient*.
- Otherwise the state is said to be *recurrent*.

- Moreover if the return time T has finite mean then the state is said to be *positive-recurrent*.
- Recurrent states which are not positive-recurrent are called *null-recurrent*.
- States of an irreducible Markov chain are all recurrent if one is, all positive-recurrent if one is.

Because of this last fact, we often talk about recurrent *chains* and transient *chains* rather than recurrent states and transient states.

- Asymmetric simple random walk (jumps ± 1 with prob $\neq 1/2$ of $+1$) is an example of a transient Markov chain: see Cox and Miller [1965] for a pretty explanation using strong law of large numbers.
- Symmetric simple random walk (jumps ± 1 with prob $1/2$ each) is an example of a (null-)recurrent Markov chain.
- We will see later that there exist infinite positive-recurrent chains.
- Terminology is motivated by the limiting behaviour of probability of being found in that state at large time. Asymptotically zero if null-recurrent or transient; tends to $1/\mathbb{E}[T]$ if aperiodic positive-recurrent.
- The fact that either all states are recurrent or all states are transient is based on the criterion for recurrence of state i,

$$\sum_n p_{ii}^{(n)} = \infty,$$

which in turn arises from an application of generating functions. The criterion amounts to asserting, the chain is sure to return to a state i exactly when the *mean number* of returns is infinite.

3.7 Recurrence/transience for random walks on $\mathbb{Z}$

Let X be a random walk on $\mathbb{Z}$ which takes steps of size 1 with probability p and minus one with probability $q = 1 - p$. Define $T_{0,1}$ to be the first time at which X hits 1, if it starts at 0.

Note that it's certainly possible to have $\mathbb{P}(T_{0,1} < \infty) < 1$, that is, for the random variable $T_{0,1}$ to take the value ∞!

The probability generating function for this random variable satisfies

$$G(z) = \mathbb{E}\left[z^{T_{0,1}}\right] = zp + zqG(z)^2$$

Solving this (and noting that we need to take the negative root!) we see that

$$G(z) = \frac{1 - \sqrt{1 - 4pqz^2}}{2qz},$$

and so $\mathbb{P}(T_{0,1} < \infty) = \lim_{z \to 1} G(z) = \min\{p/q, 1\}$. Thus if $p < 1/2$ there is a positive chance that X *never* reaches state 1; by symmetry, X is recurrent iff $p = 1/2$.

- **Test understanding**: Show that the quadratic formula for $G(z)$ holds by considering what can happen at time 1: argue that if $X_1 = -1$ the time taken to get from -1 to 1 has the same distribution as the time taken to get from -1 to 0 plus the time to get from 0 to 1; these random variables are independent, and so the pgf of the sum is easy to work with...
- If we take the positive root then $G(z) \to \infty$ as $z \to 0$, rather than to 0!
- Here we are using the fact that, since our state space is irreducible, state i is recurrent iff $\mathbb{P}(T_{i,j} < \infty) = 1$ for all states j, where $T_{i,j}$ is the first time that X hits j when started from i.

3.8 Equilibrium of Markov chains

- If X is irreducible and positive-recurrent then it has a unique *equilibrium distribution* π: if X_0 is random with distribution given by $\mathbb{P}(X_0 = i) = \pi_i$ then $\mathbb{P}(X_n = i) = \pi_i$ for any n.

In general the chain continues moving, it is just that the marginal probabilities at time n do not change.

- Moreover the equilibrium distribution viewed as a row vector solves the *equilibrium equations*:

$$\pi P = \pi\,, \qquad \text{or} \qquad \pi_j = \sum_i \pi_i p_{ij}\,.$$

Test understanding: Show that the 2-state Markov chain with transition probability matrix

$$\begin{pmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{pmatrix}$$

has equilibrium distribution

$$\pi = (0.470588\ldots, 0.529412\ldots).$$

Note that you need to use the fact that $\pi_1 + \pi_2 = 1$: this is *always* an important extra fact to use in determining a Markov chain's equilibrium distribution!

- If in addition X is aperiodic then the equilibrium distribution is also the limiting distribution (for any X_0):

$$\mathbb{P}(X_n = i) \to \pi_i \quad \text{as } n \to \infty\,.$$

This limiting result is of great importance in MCMC. If aperiodicity fails then it is always possible to sub-sample to convert to the aperiodic case on a subset of state-space.
The note at the end of the previous section shows the possibility of computing mean recurrence time using matrix arithmetic.
NB: π_i can also be interpreted as "mean time in state i".

3.9 Sums of limits and limits of sums

- Finite state-space discrete Markov chains have a useful simplifying property: they are always positive-recurrent if they are irreducible.
- This can be proved by using a result, that for null-recurrent or transient states j we find $p_{ij}^{(n)} \to 0$ as $n \to \infty$, for all other states i. If there were null-recurrent or transient states in a finite state-space discrete Markov chain, this would give a contradiction:

$$\sum_j \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_j p_{ij}^{(n)}$$

  and the right-hand sum equals 1 from "law of total probability", while left-hand sum equals $\sum 0 = 0$ by null-recurrence.
- This argument doesn't give a contradiction for infinite state-space chains as it is incorrect arbitrarily to exchange infinite limiting operations: $\lim \sum \neq \sum \lim$ in general.

> - Some argue that *all* Markov chains met in practice are finite, since we work on finite computers with finite floating point arithmetic. Do you find this argument convincing or not?
> - Recall from the "Limits versus expectations" section the principal theorems which deliver checkable conditions as to when one can swap limits and sums.

3.10 Continuous-time countable state-space Markov chains (a rough guide)

> This is a *very* rough guide, and much of what we will talk about in the course will be in discrete time. However, sometimes the easiest examples in Markov chains are in continuous-time. The important point to grasp is that if we know the transition rates $q(x, y)$ then we can write down differential equations to define the transition probabilities and so the chain. We don't necessarily try to solve the equations...

- Definition of continuous-time (countable) discrete state-space (time-homogeneous) Markov chain $X = \{X_t : t \geq 0\}$: for $s, t > 0$

$$p_t(x, y) = \mathbb{P}\left(X_{s+t} = y | X_s = x, X_u \text{ for various } u \leq s\right)$$

  depends only on x, y, t, not on rest of past.
- Organize $p_t(x, y)$ into matrices $P(t) = \{p_t(x, y) : \text{ states } x, y\}$; as in discrete case $P(t) \cdot P(s) = P(t + s)$ and $P(0)$ is identity matrix.
- (Try to) compute time derivative: $Q = (d/dt)P(t)|_{t=0}$ is matrix of *transition rates* $q(x, y)$.

- For short, we can write

$$p_t(x,y) = \mathbb{P}\left(X_{s+t} = y | X_s = x, \mathcal{F}_s\right)$$

  where $\mathcal{F}_s$ represents all possible information about the past at time s.
- From here on we omit *many* "under sufficient regularity" statements. Norris [1998] gives a careful treatment.
- The row-sums of $P(t)$ all equal 1 ("law of total probability"). Hence the row sums of Q ought to be 0 with non-positive diagonal entries $q(x,x) = -q(x)$ measuring rate of *leaving* x.

- For suitably *regular* continuous-time countable state-space Markov chains, we can use the Q-matrix Q to simulate the chain as follows:
  - rate of leaving state x is $q(x) = \sum_{y \neq x} q(x,y)$ (since row sums of Q should be zero). Time till departure is Exponential$(q(x))$;
  - on departure from x, go straight to state $y \neq x$ with probability $q(x,y)/q(x)$.

Why an exponential distribution? Because an effect of the Markov property is to require the holding time until the first transition to have a memory-less property – which characterizes Exponential distributions.
Here it is relevant to note that "minimum of independent Exponential random variables is Exponential".
Note that there are two strong limitations of continuous-time Markov chains as stochastic models: the Exponential distribution of holding times may be unrealistic; and the state to which a transition is made does not depend on actual length of holding time (this also follows rather directly from the Markov property). Of course, people have worked on generalizations (keyword: semi-Markov processes).

- Compute the s-derivative of $P(s) \cdot P(t) = P(s+t)$. This yields the famous **Kolmogorov backwards equations**:

$$Q \cdot P(t) = P(t)'\,.$$

  The other way round yields the **Kolmogorov forwards equations**:

$$P(t) \cdot Q = P(t)'\,.$$

Test understanding: use calculus to derive

$$\sum_z p_s(x,z)p_t(z,y) = p_{s+t}(x,y) \text{ gives } \sum_z q(x,z)p_t(z,y) = \frac{\partial}{\partial t} p_t(x,y)\,,$$
$$\sum_z p_t(x,z)p_s(z,y) = p_{t+s}(x,y) \text{ gives } \sum_z p_t(x,z)q(z,y) = \frac{\partial}{\partial t} p_t(x,y)\,.$$

Note the shameless exchange of differentiation and summation over potentially infinite state-space...

- If statistical equilibrium holds then the transition probabilities should converge to limiting values as $t \to \infty$: applying this to the forwards equation we expect the equilibrium distribution π to solve

$$\pi \cdot Q = \mathbf{0}\,.$$

Test understanding: applying this idea to the backwards equation gets us nothing, as a consequence of the vanishing of row sums of Q.
In extended form $\pi \cdot Q = \mathbf{0}$ yields the important *equilibrium equations*

$$\sum_z \pi(z) q(z, y) = 0\,.$$

3.11 Example: the Poisson process

We use the above theory to *define* chains by specifying the non-zero rates. Consider the case when X counts the number of people arriving at random at constant rate:

1. Stipulate that the number X_t of people in system at time t forms a Markov chain.
2. Transition rates: people arrive one-at-a-time at constant rate, so $q(x, x+1) = \lambda$.

One can solve the Kolmogorov differential equations in this case:

$$\mathbb{P}\left(X_t = n | X_0 = 0\right) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}\,.$$

For most Markov chains one makes progress *without* solving the differential equations. The interplay between the simulation method above and the distributional information here is exactly the interplay between viewing the Poisson process as a counting process ("Poisson counts") and a sequence of inter-arrival times ("Exponential gaps"). The classic relationships between Exponential, Poisson, Gamma and Geometric distributions are all embedded in this one process.
Two significant extra facts are

- **superposition**: independent sum of Poisson processes is Poisson;
- **thinning**: if arrivals are censored i.i.d. at random then result is Poisson.

3.12 Example: the M/M/1 queue

Consider a queue in which people arrive and are served (in order) at constant rates by a single server.

1. Stipulate that the number X_t of people in system at time t forms a Markov chain.
2. Transition rates (I): people arrive one-at-a-time at constant rate, so $q(x, x+1) = \lambda$.
3. Transition rates (II): people are served in order at constant rate, so $q(x, x-1) = \mu$ **if** $x > 0$.

One can solve the equilibrium equations to deduce: the equilibrium distribution of X exists and is Geometric if and only if $\lambda < \mu$.

Don't try to solve the equilibrium equations at home (unless you enjoy that sort of thing). In this case it is do-able, but during the module we'll discuss a much quicker way to find the equilibrium distribution in favourable cases.
Here is the equilibrium distribution in more explicit form: in equilibrium

$$\mathbb{P}(X = x) = (1 - \rho)\rho^x \qquad \text{for } x = 0, 1, 2, \ldots$$

where $\rho = \lambda/\mu \in (0, 1)$ (the traffic intensity).

Chapter 4

Some useful texts

At increasing levels of mathematical sophistication:

- Häggström [2002] "Finite Markov chains and algorithmic applications".
  Delightful introduction to finite state-space discrete-time Markov chains, from point of view of computer algorithms.
- Grimmett and Stirzaker [2001] "Probability and random processes".
  Standard undergraduate text on mathematical probability. If you are going to buy one book on probability, this is a good choice because it contains so much material.
- Norris [1998] "Markov chains".
  Markov chains at a more graduate level of sophistication, revealing what I have concealed, namely the full gory story about Q-matrices.
- Williams [1991] "Probability with martingales".
  Excellent graduate text for theory of martingales: mathematically demanding.

4.1 Free on the web

- Doyle and Snell [1984] "Random walks and electric networks"
  Available on the web at http://arxiv.org/abs/math/0001057.
  Lays out (in simple and accessible terms) an important approach to Markov chains using relationship to resistance in electrical networks.
- Kindermann and Snell [1980] "Markov random fields and their applications"
  Available on the web at https://doi.org/10.1090/conm/001.
  Sublimely accessible treatment of Markov random fields (Markov property, but in space not time).
- Meyn and Tweedie [1993] "Markov chains and stochastic stability"
  Available on the web at http://probability.ca/MT/.
  The place to go if you need to get informed about theoretical results on rates of convergence for Markov chains (e.g. because you are doing MCMC).
- Aldous and Fill [2001] "Reversible Markov Chains and Random Walks on Graphs"

Only available on the web at http://www.stat.berkeley.edu/~aldous/RWG/book.html.
The best unfinished book on Markov chains known to us.

4.2 Going deeper

- Kingman [1993] "Poisson processes".
  Very good introduction to the wide circle of ideas surrounding the Poisson process.
- Kelly [1979] "Reversibility and stochastic networks".
  We'll cover reversibility briefly in the lectures, but this shows just how powerful the technique is.
- Lindvall [1992] "Lectures on the coupling method".
  We'll also talk briefly about the beautiful concept of coupling for Markov chains; this book gives a very nice introduction.
- Steele [2004] "The Cauchy-Schwarz master class".
  The book to read if you decide you need to know more about (mathematical) inequality.
- Aldous [1989] "Probability approximations via the Poisson clumping heuristic".
  See http://www.stat.berkeley.edu/~aldous/Research/research80.html.
  A book full of what *ought* to be true; hence good for stimulating research problems and also for ways of computing heuristic answers.
- Øksendal [2003] "Stochastic differential equations".
  An accessible introduction to Brownian motion and stochastic calculus, which we do not cover at all.
- Stoyan et al. [1987] "Stochastic geometry and its applications".
  Discusses a range of techniques used to handle probability in geometric contexts.

Bibliography

David J. Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989. ISBN 0-387-96899-7.

David J. Aldous and James A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. Unpublished, 2001. URL http://www.stat.berkeley.edu/~aldous/RWG/book.html.

D. R. Cox and H. D. Miller. *The theory of stochastic processes*. John Wiley & Sons Inc., New York, 1965.

Peter G. Doyle and J. Laurie Snell. *Random walks and electric networks*, volume 22 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, 1984. ISBN 0-88385-024-9.

Martin Gardner. Word ladders: Lewis Carroll's doublets. *The Mathematical Gazette*, 80(487): 195–198, 1996. ISSN 00255572. URL http://www.jstor.org/stable/3620349.

Geoffrey R. Grimmett and David R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001. ISBN 0-19-857223-9.

Olle Häggström. *Finite Markov chains and algorithmic applications*, volume 52 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-81357-3; 0-521-89001-2.

Frank P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons Ltd., Chichester, 1979. ISBN 0-471-27601-4. Wiley Series in Probability and Mathematical Statistics.

Ross Kindermann and J. Laurie Snell. *Markov random fields and their applications*, volume 1 of *Contemporary Mathematics*. American Mathematical Society, Providence, R.I., 1980. ISBN 0-8218-5001-6. URL https://doi.org/10.1090/conm/001.

J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications.

Donald E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. ACM, New York, NY, USA, 1993. ISBN 0-201-54275-7.

Torgny Lindvall. *Lectures on the coupling method*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1992. ISBN 0-471-54025-0.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993. ISBN 3-540-19832-6. URL http://probability.ca/MT/.

J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-48181-3. Reprint of 1997 original.

Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. ISBN 3-540-04758-1. An introduction with applications.

J. Michael Steele. *The Cauchy-Schwarz master class*. MAA Problem Books Series. Mathematical Association of America, Washington, DC, 2004. ISBN 0-521-83775-8; 0-521-54677-X. An introduction to the art of mathematical inequalities.

D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1987. ISBN 0-471-90519-4. With a foreword by D. G. Kendall.

David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6.