APTS Applied Stochastic Processes
Preliminary material

Stephen Connor[1] & Christina Goldschmidt[2]
stephen.connor@york.ac.uk and goldschm@stats.ox.ac.uk

(Notes originally produced by Wilfrid Kendall)

[1] Department of Mathematics, University of York

[2] Department of Statistics, University of Oxford

16th May 2014

Introduction

Preliminary material
- Expectation and probability
- Markov chains

Some useful texts

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.

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 . . .

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.

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.

Preliminary material

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.

Preliminary material
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.

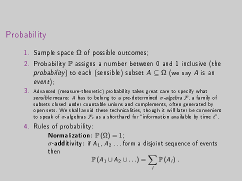
This material uses a two-panel format. Left-hand panels present the theory, often using itemized lists. Right-hand panels present commentary and useful exercises (announced by "Test understanding"). It is likely that you will have seen most, if not all, of the preliminary material at undergraduate level. 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).

Probability

1. Sample space Ω of possible outcomes;
2. 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*);
3. 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".
4. Rules of probability:

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



Probability

1. Sample space Ω of possible outcomes;
2. 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*);
3. 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".
4. Rules of probability:
   Normalization: $\mathbb{P}(\Omega) = 1$;
   σ-additivity: if A_1, A_2 ... form a disjoint sequence of events then
   $$\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \sum_i \mathbb{P}(A_i) .$$

1. Example: $\Omega = (-\infty, \infty)$.
2. We could for example start with $\mathbb{P}((a, b)) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} \,\mathrm{d}\, u$ and then use the rules of probability to determine probabilities for all manner of sensible subsets of $(-\infty, \infty)$.
3. In our 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.
4. Test understanding: use these rules to explain
   (a) why $\mathbb{P}(\emptyset) = 0$,
   (b) why $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ 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 $(-\infty, \infty) = \bigcup_x \{x\}$.

Conditional probability

1. 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.
2. **Bayes:** if $B_1, B_2, \ldots$ is an exhaustive disjoint partition of Ω then
$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}\,.$$
3. Conditional probabilities are clandestine random variables! Let X be the Bernoulli[1] random variable which indicates[2] event B. 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 $X = 0$.

[1] Taking values only 0 or 1.
[2] $X = 1$ exactly when B happens.

Conditional probability

1. 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.
2. **Bayes:** if $B_1, B_2, \ldots$ is an exhaustive disjoint partition of Ω then
$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\,\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}.$$
3. Conditional probabilities are clandestine random variables! Let X be the Bernoulli[1] random variable which indicates[2] event B. 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 $X = 0$.

[1] Taking values only 0 or 1.
[2] $X = 1$ exactly when B happens.

1. Actually we *often* use limiting arguments to make sense of cases when $\mathbb{P}(B) = 0$.
2. Hence all of Bayesian statistics ...
   Test understanding: write out an explanation of why Bayes' theorem is a completely obvious consequence of the definitions of probability and conditional probability.
3. 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

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

   Hint: represent RHS as expectation of expansion of $1 - (1 - X_1) \ldots (1 - X_n)$ for suitable Bernoulli random variables X_i indicating various A_i.

Expectation

Statistical intuition about expectation is based on *properties*:

1. If $X \geq 0$ is a non-negative random variable then we can define its (possibly infinite) *expectation* $\mathbb{E}[X]$.
2. If $X = X^+ - X^- = \max\{X, 0\} - \max\{-X, 0\}$ is such that $\mathbb{E}[X^\pm]$ are both finite[3] then set $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$.
3. Familiar properties of expectation follow from **linearity** ($\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$) and **monotonicity** ($\mathbb{P}(X \geq a) = 1$ implies $\mathbb{E}[X] \geq a$) for constants a, b.
4. Useful notation: for an event A write $\mathbb{E}[X; A] = \mathbb{E}[X\,1_A]$, where 1_A is the Bernoulli random variable indicating A.

   We can then consider specific *constructions*:
5. If X has countable range then $\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X = x)$.
6. If X has probability *density* f_X then $\mathbb{E}[X] = \int x\,f_X(x)\,\mathrm{d}\,x$.

[3] We wish to avoid having to make sense of $\infty - \infty$!

Expectation

Statistical intuition about expectation is based on *properties*:

1. If $X \geq 0$ is a non-negative random variable then we can define its (possibly infinite) *expectation* $\mathbb{E}[X]$.
2. If $X = X^+ - X^- = \max\{X, 0\} - \max\{-X, 0\}$ is such that $\mathbb{E}[X^\pm]$ are both finite[3] then set $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$.
3. Familiar properties of expectation follow from **linearity** ($\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$) and **monotonicity** ($\mathbb{P}(X \geq a) = 1$ implies $\mathbb{E}[X] \geq a$) for constants a, b.
4. Useful notation: for an event A write $\mathbb{E}[X; A] = \mathbb{E}[X\, 1_A]$, where 1_A is the Bernoulli random variable indicating A.

   We can then consider specific *constructions*:

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

[3] We wish to avoid having to make sense of $\infty - \infty$!

1. 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 . . . .
2. Any decomposition as difference of *integrable* random variables will do.
3. Test understanding: using these properties
   - deduce $\mathbb{E}[a] = a$ for constant a.
   - show *Markov's inequality* $\mathbb{P}(X \geq a) \leq \frac{1}{a}\,\mathbb{E}[X]$ for $X \geq 0$, $a > 0$.
4. So in absolutely continuous case $\mathbb{E}[X; A] = \int_A x\, f_X(x)\,\mathrm{d}\,x$ and in discrete case $\mathbb{E}[X; X = k] = k\,\mathbb{P}(X = k)$.
5. Countable [=discrete] case: expectation defined exactly when sum converges absolutely.
6. Density [=(absolutely) continuous] case: expectation defined exactly when integral converges absolutely.

Generating functions

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

Some functions are particularly useful:

1. when $g(x) = z^x$ for some $z \geq 0$ we obtain the *probability generating function (pgf)* of X, $G_X(z) = \mathbb{E}\left[z^X\right]$;
2. when $g(x) = e^{tx}$ we get the *moment generating function (mgf)*, $m_X(t) = \mathbb{E}\left[e^{tX}\right]$;
3. when $g(x) = e^{itx}$, where $i = \sqrt{-1}$, we get the *characteristic function* of X, $\phi_X(t)$.

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:

1. when $g(x) = z^x$ for some $z \geq 0$ we obtain the *probability generating function (pgf)* of X, $G_X(z) = \mathbb{E}\left[z^X\right]$;
2. when $g(x) = e^{tx}$ we get the *moment generating function (mgf)*, $m_X(t) = \mathbb{E}\left[e^{tX}\right]$;
3. when $g(x) = e^{itx}$, where $i = \sqrt{-1}$, we get the *characteristic function* of X, $\phi_X(t)$.

1. The pgf is only really useful when X takes values in $\{0, 1, 2, \ldots\}$. The mgf and characteristic function are more generally applicable.
2. 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, evaluated at $z = 0$).
3. Test understanding: Show that $\mathbb{E}[X] = m'_X(0)$ and

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

Uses of generating functions

Generating functions are helpful in many ways. In particular:

1. They can be used to determine distributions
2. They can often provide an easy route to finding e.g. moments of a distribution
3. 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

$$G_{X+Y}(z) = G_X(z)G_Y(z) \quad \text{etc.}$$

Uses of generating functions

Generating functions are helpful in many ways. In particular:

1. They can be used to determine distributions
2. They can often provide an easy route to finding e.g. moments of a distribution
3. 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

$$G_{X+Y}(z) = G_X(z)G_Y(z) \quad \text{etc.}$$

1. 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 $\{0, 1, \ldots\}$; mgfs are slightly more complicated, but *mostly* they can be used to identify a distribution. See Grimmett and Stirzaker (2001) for more on this.
2. See the two exercises on the previous notes slide!
3. Test understanding: show that if X and Y are independent random variables then $G_{X+Y}(z) = G_X(z)G_Y(z)$, $m_{X+Y}(t) = m_X(t)m_Y(t)$ and $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. (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.

Conditional Expectation (I): property-based definition

1. 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:
2. 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 prob 0.
3. 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.

Conditional Expectation (I): property-based definition

1. 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:
2. 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 prob 0.
3. 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.

Conditional expectation needs careful definition to capture all cases. But focus on *properties* to build intuitive understanding.

1. Notice that conditional expectation is also properly viewed as a random variable.
2. – "$\mathbb{E}[Z] < \infty$" 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.
3. 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 $\mathbb{E}[X_1|X_1 + \ldots + X_n] = \frac{1}{n}(X_1 + \ldots + X_n)$.

Conditional Expectation (II): some other properties

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

1. Linearity: $\mathbb{E}\left[aX + bY|Z\right] = a\,\mathbb{E}\left[X|Z\right] + b\,\mathbb{E}\left[Y|Z\right]$;
2. "Tower property": $\mathbb{E}\left[\mathbb{E}\left[X|Y, Z\right]|Y\right] = \mathbb{E}\left[X|Y\right]$;
3. "Taking out what is known": $\mathbb{E}\left[f(Y)X|Y\right] = f(Y)\,\mathbb{E}\left[X|Y\right]$;

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

Conditional Expectation (II): some other properties

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

1. Linearity: $\mathbb{E}[aX+bY|Z] = a\,\mathbb{E}[X|Z] + b\,\mathbb{E}[Y|Z]$;
2. "Tower property": $\mathbb{E}[\mathbb{E}[X|Y,Z]|Y] = \mathbb{E}[X|Y]$;
3. "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:

1. Use $\mathbb{E}[aX+bY;A] = a\,\mathbb{E}[X;A] + b\,\mathbb{E}[Y;A]$.
2. 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 $\mathbb{E}[\mathbb{E}[X|Y,Z];A] = \mathbb{E}[X;A]$.
3. 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 $\mathbb{E}[\mathbb{E}[f(Y)X|Y];A\cap\{f(Y)=t\}] = \mathbb{E}[\mathbb{E}[tX|Y];A\cap\{f(Y)=t\}] = \mathbb{E}[t\,\mathbb{E}[X|Y];A\cap\{f(Y)=t\}] = \mathbb{E}[f(Y)\,\mathbb{E}[X|Y];A\cap\{f(Y)=t\}]$.
   General case now follows by approximation arguments.

Conditional Expectation (III): Jensen's inequality

This is powerful and yet rather easy to prove.

Theorem

Let ϕ be a convex function ("curves upwards", $\phi'' \geq 0$ 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 some conditioning information $\mathcal{G}$.

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

THE UNIVERSITY of York
UNIVERSITY OF OXFORD

Conditional Expectation (III): Jensen's inequality

This is powerful and yet rather easy to prove.

Theorem

Let ϕ be a convex function ("curves upwards", $\phi'' \geq 0$ 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:

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

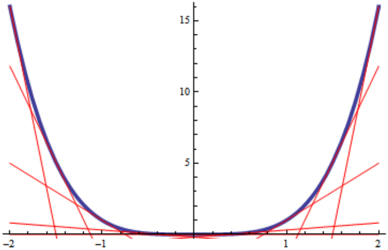
for some conditioning information $\mathcal{G}$.

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, 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 $\phi(x) = x^4$:



Limits versus expectations

1. Often the crux of a piece of mathematics is whether one can exchange limiting operations such as $\lim \sum \leftrightarrow \sum \lim$. Here are a few very useful results on this, expressed in the language of expectations.
2. Monotone Convergence Theorem: If $\mathbb{P}\left(X_n \uparrow Y\right) = 1$ and $\mathbb{E}\left[X_1\right] > -\infty$ then $\lim_n \mathbb{E}\left[X_n\right] = \mathbb{E}\left[\lim_n X_n\right] = \mathbb{E}\left[Y\right]$.
3. Dominated Convergence Theorem: If $\mathbb{P}\left(X_n \to Y\right) = 1$ and $|X_n| \leq Z$ where $\mathbb{E}\left[Z\right] < \infty$ then $\lim_n \mathbb{E}\left[X_n\right] = \mathbb{E}\left[\lim_n X_n\right] = \mathbb{E}\left[Y\right]$.
4. Fubini's Theorem: If $\mathbb{E}\left[|f(X,Y)|\right] < \infty$, X, Y are independent, $g(y) = \mathbb{E}\left[f(X,y)\right]$, $h(x) = \mathbb{E}\left[f(x,Y)\right]$ then $\mathbb{E}\left[g(Y)\right] = \mathbb{E}\left[f(X,Y)\right] = \mathbb{E}\left[h(X)\right]$.
5. Fatou's lemma: If $\mathbb{P}\left(X_n \to Y\right) = 1$ and $X_n \geq 0$ for all n then $\mathbb{E}\left[Y\right] \leq \lim_n \inf_{m \geq n} \mathbb{E}\left[X_m\right]$.

Limits versus expectations

1. Often the crux of a piece of mathematics is whether one can exchange limiting operations such as $\lim \sum \leftrightarrow \sum \lim$. Here are a few very useful results on this, expressed in the language of expectations.
2. Monotone Convergence Theorem: If $\mathbb{P}(X_n \uparrow Y) = 1$ and $\mathbb{E}[X_1] > -\infty$ then $\lim_n \mathbb{E}[X_n] = \mathbb{E}[\lim_n X_n] = \mathbb{E}[Y]$.
3. Dominated Convergence Theorem: If $\mathbb{P}(X_n \to Y) = 1$ and $|X_n| \leq Z$ where $\mathbb{E}[Z] < \infty$ then $\lim_n \mathbb{E}[X_n] = \mathbb{E}[\lim_n X_n] = \mathbb{E}[Y]$.
4. 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 $\mathbb{E}[g(Y)] = \mathbb{E}[f(X,Y)] = \mathbb{E}[h(X)]$.
5. Fatou's lemma: If $\mathbb{P}(X_n \to Y) = 1$ and $X_n \geq 0$ for all n then $\mathbb{E}[Y] \leq \lim_n \inf_{m \geq n} \mathbb{E}[X_m]$.

---

1. As we formulate this in expectation language, our results apply equally to sums and integrals.
2. 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.
3. Note that convergence need not be monotonic here or in following. Test understanding: explain why it would be enough to have finite upper and lower bounds $\alpha \leq X_n \leq \beta$.
4. Fubini exchanges expectations rather than an expectation and a limit.
5. Try Fatou if all else fails. Note that something like $X_n \geq 0$ is essential (a constant lower bound would suffice, though).

Preliminary material

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.

Preliminary material
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.

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).

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

Basic properties for discrete time and space case

1. Markov chain $X = \{X_0, X_1, X_2, \ldots\}$: X at time t is in state $X_t = x$. View states x as integers.
2. X must have the Markov property: $p_{xy} = p(x, y) = \mathbb{P}(X_{t+1} = y | X_t = x, X_{t-1}, \ldots)$ must depend only on x, y, not on rest of past. (Our chains will be *time-homogeneous*, meaning no t dependence either.)
3. Chain behaviour is specified by (a) initial state X_0 (could be random) and (b) table of transition probabilities p_{xy}.
4. Important matrix structure: if p_{xy} are arranged in matrix P then $(i, j)^{\text{th}}$ entry of $P^n = P \cdot \ldots \cdot P$ (n times) is $p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i)$.
   Equivalent: Chapman-Kolmogorov equations

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

Basic properties for discrete time and space case

1. Markov chain $X = \{X_0, X_1, X_2, \ldots\}$: X at time t is in state $X_t = x$. View states x as integers.
2. X must have the Markov property: $p_{xy} = p(x,y) = \mathbb{P}(X_{t+1} = y | X_t = x, X_{t-1}, \ldots)$ must depend only on x, y, not on rest of past. (Our chains will be *time-homogeneous*, meaning no t dependence either.)
3. Chain behaviour is specified by (a) initial state X_0 (could be random) and (b) table of transition probabilities p_{xy}.
4. Important matrix structure: if p_{xy} are arranged in matrix P then $(i,j)^{\text{th}}$ entry of $P^n = P \cdot \ldots \cdot P$ (n times) is $p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i)$.
   Equivalent: Chapman-Kolmogorov equations
   $$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(n)} p_{kj}^{(m)}$$

---

1. More general countable discrete state-spaces can always be indexed by integers
2. The example of "Markov's other chain" below shows we need to insist on the possibility of conditioning by further past $X_{t-1}, \ldots$ in this definition.
   Note $\sum_y p_{xy} = 1$ by "law of total probability".
3. Example: some word transition probabilities arising in the "random English" example given immediately below:

P("round"|"all")=0.50	P("contact"|"all")=0.50	P("hearing"|"ocean,")=1.00
P("first,"|"go")=1.00	P("As"|"up.")=1.00	P("Every"|"day.")=1.00
P("woman"|"young")=0.33		P("prince."|"young")=0.33
P("man"|"young")=0.33		P("on"|"enjoined")=1.00 . . .

4. Test understanding: show how the Chapman-Kolmogorov equations follow from considerations of conditional probability and the Markov property.

Example: Models for language following Markov

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

1. Take a large book in electronic form, for example Tolstoy's "War and Peace" (English translation).
2. Use it to build a table of digram frequencies (digram = pair of consecutive letters).
3. 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.

Example: Models for language following Markov

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

1. Take a large book in electronic form, for example Tolstoy's "War and Peace" (English translation).
2. Use it to build a table of digram frequencies (digram = pair of consecutive letters).
3. 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.

1. The World-Web Web has made this part much easier: try Project Gutenberg (`www.gutenberg.org/etext/2600`).
2. Skill is required in deciding *which* letters to use: should one use all, or some, punctuation? Certainly need to use spaces.
3. Trigrams would be more impressive. 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.*

(Counter)example: Markov's other chain

Conditional probability can be subtle. Consider:

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

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

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

(Counter)example: Markov's other chain

Conditional probability can be subtle. Consider:

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

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

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

Example taken from Grimmett and Stirzaker (2001).

Note that the entirety of random variables $X_0, X_1, X_2, \ldots$ are most certainly *not* independent!
Test understanding by checking these calculations.

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 as well to be aware of it: conditioning is a subtle concept and should be treated with respect!

Irreducibility and aperiodicity

1. 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.
2. 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.
3. 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}.

Irreducibility and aperiodicity

1. 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.
2. 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.
3. 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

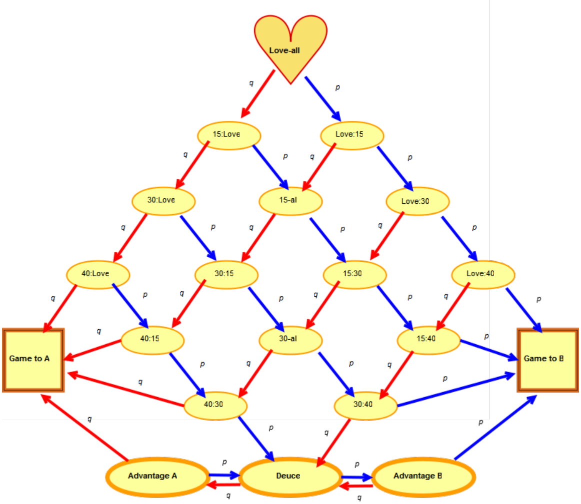
$$f_{ij} = \mathbb{P}(X_n = j \text{ for some positive } n | X_0 = i)$$

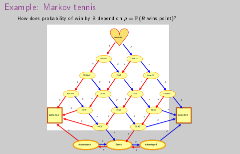
then solve linear equations for the f_{ij}.

1. 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).
2. 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.
3. 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.

Example: Markov tennis

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





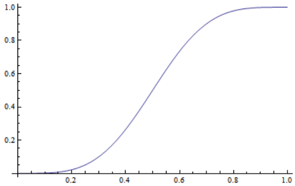
Use first passage equations, then solve linear equations for the f_{ij}, noting in particular

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

I obtain

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

graphed against p below:



Transience and recurrence

1. 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*.
2. Otherwise the state is said to be *recurrent*.
3. Moreover if the return time T has finite mean then the state is said to be *positive-recurrent*.
4. Recurrent states which are not positive-recurrent are called *null-recurrent*.
5. States of an irreducible Markov chain are all recurrent if one is, all positive-recurrent if one is.

Transience and recurrence

1. 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*.
2. Otherwise the state is said to be *recurrent*.
3. Moreover if the return time T has finite mean then the state is said to be *positive-recurrent*.
4. Recurrent states which are not positive-recurrent are called *null-recurrent*.
5. States of an irreducible Markov chain are all recurrent if one is, all positive-recurrent if one is.

1. Example: asymmetric simple random walk (jumps ± 1): see Cox and Miller (1965) for a pretty explanation using strong law of large numbers.
2. Example: symmetric simple random walk (jumps ± 1).
3. As we will see, there exist infinite positive-recurrent chains (eg, "discrete AR(1)").
4. Why "null", "positive"? 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.)
5. This 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.

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

Let X be a random walk on $\mathbb{Z}$ which takes steps of size 1 with prob p and minus one with prob $q = 1 - p$. Define $T_{0,1}$ to be the first time at which X hits 1, if it starts at 0. The probability generating function for this RV 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 $p = 1/2$.

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

Let X be a random walk on $\mathbb{Z}$ which takes steps of size 1 with prob p and minus one with prob $q = 1 - p$. Define $T_{0,1}$ to be the first time at which X hits 1, if it starts at 0. The probability generating function for this RV 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 $p = 1/2$.

1. 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 ∞!
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...
3. If we take the positive root then $G(z) \to \infty$ as $z \to 0$, rather than to 0!
4. 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.

Equilibrium of Markov chains

1. 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.
2. Moreover the equilibrium distribution viewed as a row vector solves the *equilibrium equations*:

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

3. If in addition X is aperiodic then the equilibrium distribution is also the limiting distribution:

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

Equilibrium of Markov chains

1. 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.
2. Moreover the equilibrium distribution viewed as a row vector solves the *equilibrium equations*:
$$\pi P = \pi\,, \qquad \text{or} \qquad \pi_j = \sum_i \pi_i p_{ij}\,.$$
3. If in addition X is aperiodic then the equilibrium distribution is also the limiting distribution:
$$\mathbb{P}(X_n = i) \to \pi_i \quad \text{as } n \to \infty\,.$$

1. In general the chain continues moving, but the marginal probabilities at time n do not change.
2. Test understanding: Show that the 2-state Markov chain with transition probability matrix $\left(\begin{smallmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{smallmatrix}\right)$ 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!
3. 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.
Note 4 of previous segment shows possibility of computing mean recurrence time using matrix arithmetic.
NB: π_i can also be interpreted as "mean time in state i".

Sums of limits and limits of sums

1. Finite state-space discrete Markov chains have a useful simplifying property: they are always positive-recurrent if they are irreducible.
2. 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. Hence 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.
3. This argument fails for infinite state-space as it is incorrect arbitrarily to exchange infinite limiting operations: $\lim \sum \neq \sum \lim$ in general.

Sums of limits and limits of sums

1. Finite state-space discrete Markov chains have a useful simplifying property: they are always positive-recurrent if they are irreducible.
2. 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. Hence 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.
3. This argument fails for infinite state-space as it is incorrect arbitrarily to exchange infinite limiting operations: $\lim \sum \neq \sum \lim$ in general.

1. 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?
2. The result used here puts the "null" in *null-recurrence*.
3. We have earlier summarized the principal theorems which deliver checkable conditions as to when one can make this exchange.

   Note that the simple random walk (irreducible but null-recurrent or transient) is the simplest practical example of why one must not carelessly exchange infinite limiting operations!

Continuous-time countable state-space Markov chains

(a rough guide)

1. 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.

2. 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.
3. (Try to) compute time derivative: $Q = (d/dt)P(t)|_{t=0}$ is matrix of *transition rates* $q(x,y)$.

Continuous-time countable state-space Markov chains

(a rough guide)

1. 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.
2. 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.
3. (Try to) compute time derivative: $Q = (d/dt)P(t)|_{t=0}$ is matrix of *transition rates* $q(x,y)$.

This is a *very* rough guide: I pondered for a while whether to add this to prerequisites, since most of what I want to talk about will be in discrete time. I decided to add it in the end because 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 . . . .

1. For short, 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.
2. From here on I omit *many* "under sufficient regularity" statements. Norris (1998) gives a careful treatment.
3. 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.

Continuous-time countable state-space Markov chains

(a rough guide continued)

For suitably *regular* continuous-time countable state-space Markov chains, we can use the Q-matrix Q to simulate the chain as follows:

1. 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($q(x)$);
2. on departure from x, go straight to state $y \neq x$ with probability $q(x, y)/q(x)$.

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

For suitably *regular* continuous-time countable state-space Markov chains, we can use the Q-matrix Q to simulate the chain as follows:

1. 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($q(x)$);
2. on departure from x, go straight to state $y \neq x$ with probability $q(x,y)/q(x)$.

1. 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".
2. This also follows rather directly from the Markov property. Note that this shows 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. Of course, people have worked on generalizations (keyword: semi-Markov processes).

Continuous-time countable state-space Markov chains

(a rough guide continued)

1. 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":

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

2. 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

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

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

1. 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":
$$P(t) \cdot Q = P(t)' .$$
2. 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
$$\pi \cdot Q = \mathbf{0} .$$

1. 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 . . . .
2. 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*
$$\sum_z \pi(z)q(z,y) = 0 .$$

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:

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

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.

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 $\lambda < \mu$.

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

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

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

Some useful texts (I)

At increasing levels of mathematical sophistication:

1. Häggström (2002) “Finite Markov chains and algorithmic applications”.
2. Grimmett and Stirzaker (2001) “Probability and random processes”.
3. Norris (1998) “Markov chains”.
4. Williams (1991) “Probability with martingales”.

Some useful texts (I)

At increasing levels of mathematical sophistication:

1. Häggström (2002) “Finite Markov chains and algorithmic applications”.
2. Grimmett and Stirzaker (2001) “Probability and random processes”.
3. Norris (1998) “Markov chains”.
4. Williams (1991) “Probability with martingales”.

1. Delightful introduction to finite state-space discrete-time Markov chains, from point of view of computer algorithms.
2. Standard undergraduate text on mathematical probability. This is the book I advise my students to buy, because it contains so much material.
3. Markov chains at a more graduate level of sophistication, revealing what I have concealed, namely the full gory story about Q-matrices.
4. Excellent graduate text for theory of martingales: mathematically demanding.

Some useful texts (II): free on the web

1. Doyle and Snell (1984) “Random walks and electric networks” available on web at http://arxiv.org/abs/math/0001057.
2. Kindermann and Snell (1980) “Markov random fields and their applications” available on web at http://www.ams.org/online_bks/conm1/.
3. Meyn and Tweedie (1993) “Markov chains and stochastic stability” available on web at http://probability.ca/MT/.
4. Aldous and Fill (2001) “Reversible Markov Chains and Random Walks on Graphs” *only* available on web at http://www.stat.berkeley.edu/~aldous/RWG/book.html.

Some useful texts (II): free on the web

1. Doyle and Snell (1984) “Random walks and electric networks” available on web at http://arxiv.org/abs/math/0001057.
2. Kindermann and Snell (1980) “Markov random fields and their applications” available on web at http://www.ams.org/online_bks/conm1/.
3. Meyn and Tweedie (1993) “Markov chains and stochastic stability” available on web at http://probability.ca/MT/.
4. Aldous and Fill (2001) “Reversible Markov Chains and Random Walks on Graphs” *only* available on web at http://www.stat.berkeley.edu/~aldous/RWG/book.html.

1. Lays out (in simple and accessible terms) an important approach to Markov chains using relationship to resistance in electrical networks.
2. Sublimely accessible treatment of Markov random fields (Markov property, but in space not time).
3. The place to go if you need to get informed about theoretical results on rates of convergence for Markov chains (eg, because you are doing MCMC).
4. The best unfinished book on Markov chains known to me.

Some useful texts (III): going deeper

1. Kingman (1993) “Poisson processes”.
2. Kelly (1979) “Reversibility and stochastic networks”.
3. Steele (2004) “The Cauchy-Schwarz master class”.
4. Aldous (1989) “Probability approximations via the Poisson clumping heuristic” see `www.stat.berkeley.edu/~aldous/Research/research80.html`.
5. Øksendal (2003) “Stochastic differential equations”.
6. Stoyan, Kendall, and Mecke (1987) “Stochastic geometry and its applications”.

Some useful texts (III): going deeper

1. Kingman (1993) "Poisson processes".
2. Kelly (1979) "Reversibility and stochastic networks".
3. Steele (2004) "The Cauchy-Schwarz master class".
4. Aldous (1989) "Probability approximations via the Poisson clumping heuristic" see www.stat.berkeley.edu/~aldous/Research/research80.html.
5. Øksendal (2003) "Stochastic differential equations".
6. Stoyan, Kendall, and Mecke (1987) "Stochastic geometry and its applications".

Here are a few of the many texts which go much further

1. Very good introduction to the wide circle of ideas surrounding the Poisson process.
2. We'll cover reversibility briefly in the lectures, but this shows just how powerful the technique is.
3. The book to read if you decide you need to know more about (mathematical) inequality.
4. A book full of what *ought* to be true; hence good for stimulating research problems and also for ways of computing heuristic answers.
5. An accessible introduction to Brownian motion and stochastic calculus, which we do not cover at all.
6. Discusses a range of techniques used to handle probability in geometric contexts.

Aldous, D. J. (1989).
Probability approximations via the Poisson clumping heuristic, Volume 77 of *Applied Mathematical Sciences*.
New York: Springer-Verlag.

Aldous, D. J. and J. A. Fill (2001).
Reversible Markov Chains and Random Walks on Graphs.
Unpublished.

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

Doyle, P. G. and J. L. Snell (1984).
Random walks and electric networks, Volume 22 of *Carus Mathematical Monographs*.
Washington, DC: Mathematical Association of America.

Gardner, M. (1996).
Word ladders: Lewis Carroll's doublets.
The Mathematical Gazette 80(487), 195–198.

Grimmett, G. R. and D. R. Stirzaker (2001).
Probability and random processes (Third ed.).
New York: Oxford University Press.

Häggström, O. (2002).
Finite Markov chains and algorithmic applications, Volume 52 of *London Mathematical Society Student Texts*.
Cambridge: Cambridge University Press.

Kelly, F. P. (1979).
Reversibility and stochastic networks.
Chichester: John Wiley & Sons Ltd.
Wiley Series in Probability and Mathematical Statistics.

Kindermann, R. and J. L. Snell (1980).
Markov random fields and their applications, Volume 1 of *Contemporary Mathematics*.
Providence, R.I.: American Mathematical Society.

Kingman, J. F. C. (1993).
Poisson processes, Volume 3 of *Oxford Studies in Probability*.
New York: The Clarendon Press Oxford University Press.
Oxford Science Publications.

Knuth, D. E. (1993).
The Stanford GraphBase: a platform for combinatorial computing.
New York, NY, USA: ACM.

Meyn, S. P. and R. L. Tweedie (1993).
Markov chains and stochastic stability.
Communications and Control Engineering Series. London: Springer-Verlag London Ltd.

Norris, J. R. (1998).
Markov chains, Volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*.
Cambridge: Cambridge University Press.
Reprint of 1997 original.

Øksendal, B. (2003).
Stochastic differential equations (Sixth ed.).
Universitext. Berlin: Springer-Verlag.
An introduction with applications.

Steele, J. M. (2004).
The Cauchy-Schwarz master class.
MAA Problem Books Series. Washington, DC: Mathematical Association of America.
An introduction to the art of mathematical inequalities.

Stoyan, D., W. S. Kendall, and J. Mecke (1987).
Stochastic geometry and its applications.
Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Chichester: John Wiley & Sons Ltd.
With a foreword by D. G. Kendall.

Williams, D. (1991).
Probability with martingales.
Cambridge Mathematical Textbooks. Cambridge: Cambridge University Press.