APTS Applied Stochastic Processes

Lecture notes

Stephen B Connor and Wilfrid S Kendall

These lecture notes have been compiled directly from the slides for the lectures without further editing. They are designed to be used while attending the lectures, rather than as a substitute! In particular this means that the text contains occasional notes like "ANIMATION", which in the slides proper serve as buttons for the animations used to illustrate the lectures, but serve no purpose in the text.

The last section (Section 8 on cut-off) will not be covered in the lectures, in order to make room for a special lecture by Professor Sir David Cox.

Contents

Introduction

Two notions in probability

> "…you never learn anything unless you are willing to take a risk and tolerate a little randomness in your life." – Heinz Pagels, *The Dreams of Reason*, 1988.

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 and to build their own mental map of the subject. The *Applied Stochastic Processes* module aims to contribute towards this end.

This module is intended to 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 significant ways in which these notions are used in statistics.

These notes outline the content of the module; they represent work-in-progress and will grow, be corrected, and be modified as time passes.

The notes illustrate typical features of probability: the interplays between theory and practice, between rigour and intuition. The following quote is not intended to imply that we are great probabilists, but it nicely illustrates the point that the perceived conflicts between these interplays are often illusory:

> Some people complain that the great probabilists have no intuition. But they do have intuition. The real problem is, they have so much of it. – David Williams

Corrections and suggestions are of course welcome! Email `stephen.connor@york.ac.uk` or `w.s.kendall@warwick.ac.uk`.

Every image in these notes has been either constructed by the author or released into the public domain.

Learning outcomes

What you should be able to do after working through this module

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.

An important instruction

First of all, read the preliminary notes . . .

They provide notes and examples concerning a basic framework covering:

- Probability and conditional probability;
- Expectation and conditional expectation;
- Discrete-time countable-state-space Markov chains;
- Continuous-time countable-state-space Markov chains;
- Poisson processes.

The purpose of the preliminary notes is not to provide all the information you might require concerning probability, but to serve as a prompt about material you may need to revise, and to introduce and to establish some basic choices of notation.

The polish sausage syndrome, or when all else fails read the manual
`www.hci.com.au/hcisite3/journal/Whenallelsefails.htm`

> PLEASE READ THIS OWNER'S MANUAL BEFORE UNPACKING THE DEVICE.
> You've already unpacked it, haven't you? You've unpacked it and plugged it in and turned it on and fiddled with the knobs, and now your four-year old child, the same child who once shoved a Polish sausage into your new VCR and pressed fast forward, this child is also fiddling with the knobs, right? We might as well just break these devices right at the factory before we ship them out, you know that?

Books

Some useful texts (I)

> "There is no such thing as a moral or an immoral book. Books are well written or badly written." – Oscar Wilde (1854–1900), *The Picture of Dorian Gray*, 1891, preface

The next three slides list various useful textbooks.
At increasing levels of mathematical sophistication:

1. Häggström (2002) "Finite Markov chains and algorithmic applications".

   Häggström (2002) is a delightful introduction to finite state-space discrete-time Markov chains, starting from the point of view of computer algorithms.

2. Grimmett and Stirzaker (2001) "Probability and random processes".

   Grimmett and Stirzaker (2001) is the standard undergraduate text on mathematical probability. This is the book I advise my undergraduate students to buy, because it contains so much material.

3. Breiman (1992) "Probability".

   Breiman (1992) is a first-rate graduate-level introduction to probability.

4. Norris (1998) “Markov chains”.

   Norris (1998) presents the theory of Markov chains at a more graduate level of sophistication, revealing what I have concealed, namely the full gory story about Q-matrices.

5. Williams (1991) “Probability with martingales”.

   Williams (1991) provides an excellent if mathematically demanding graduate treatment of the theory of martingales.

Some useful texts (II): free on the web

The moon belongs to everyone;
The best things in life are free.
The stars belong to everyone;
They gleam there for you and me.
“The Best Things in Life are Free” – George Gard “Buddy” DeSylva (27/01/1895 – 11/07/1950)

1. Doyle and Snell (1984) “Random walks and electric networks” available on web at
   `www.arxiv.org/abs/math/0001057`.

   Doyle and Snell (1984) lay out (in simple and accessible terms) an important approach to Markov chains using relationship to resistance in electrical networks.

2. Kindermann and Snell (1980) “Markov random fields and their applications” available on web at
   `www.ams.org/online_bks/conm1/`.

   Kindermann and Snell (1980) is a sublimely accessible treatment of Markov random fields (Markov property, but in space not time).

3. Meyn and Tweedie (1993) “Markov chains and stochastic stability” available on web at
   `www.probability.ca/MT/`.

   Consult Meyn and Tweedie (1993) if you need to get informed about theoretical results on rates of convergence for Markov chains (*eg*, because you are doing MCMC).

4. Aldous and Fill (2001) “Reversible Markov Chains and Random Walks on Graphs” *only* available on web at
   `www.stat.berkeley.edu/~aldous/RWG/book.html`.

   Aldous and Fill (2001) is the best unfinished book on Markov chains known to me (at the time of writing these notes).

Some useful texts (III): going deeper

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

1. Kingman (1993) “Poisson processes”.

   Kingman (1993) gives a very good introduction to the wide circle of ideas surrounding the Poisson process.

2. Kelly (1979) “Reversibility and stochastic networks”.

   We’ll cover reversibility briefly in the lectures, but Kelly (1979) shows just how powerful the technique can be.

3. Steele (2004) “The Cauchy-Schwarz master class”.

   Steele (2004) is the book to read if you decide you need to know more about (mathematical) inequality.

4. Aldous (1989) “Probability approximations via the Poisson clumping heuristic”.

   Aldous (1989) is a book full of what *ought* to be true; hence good for stimulating research problems and also for ways of computing heuristic answers. See `www.stat.berkeley.edu/~aldous/Research/research80.html`.

5. Øksendal (2003) “Stochastic differential equations”.

   Øksendal (2003) provides an accessible introduction to Brownian motion and stochastic calculus, which we do not cover at all.

6. Stoyan, Kendall, and Mecke (1987) “Stochastic geometry and its applications”.

   Stoyan et al. (1987) discuss a range of techniques used to handle probability in geometric contexts.

1 Markov chains and reversibility

We begin our module with the important, simple and subtle idea of a *reversible* Markov chain, and the associated notion of *detailed balance*; we will return to these ideas periodically through the module. This first major theme isolates a class of Markov chains for which computation of the equilibrium distribution is relatively straightforward. (Remember from the pre-requisites: if a chain is irreducible and positive-recurrent then it has an equilibrium distribution $\underline{\pi}$; and if it is aperiodic then $\underline{\pi}$ is also the limiting long-time empirical distribution. Moreover $\underline{\pi} \cdot \underline{\underline{P}} = \underline{\pi}$. However if there are k states then these matrix equation presents k equations each potentially involving all k unknowns ... a complexity issue if k is large!) We will see a delicate interplay of

1. time-reversibility;
2. (greater) accessibility of equilibrium calculations;
3. subtle but significant dependence considerations.

In this section we'll discuss many examples together with animations.

Markov chains and reversibility

> "People assume that time is a strict progression of cause to effect, but actually from a non-linear, non-subjective viewpoint, it's more like a big ball of wibbly-wobbly, timey-wimey ... stuff." The Tenth Doctor, *Doctor Who*, in the episode "Blink", 2007

1.1 Introduction to reversibility

In a nutshell ...

We dive straight in, presuming prerequisite knowledge.

Here is detailed balance in a nutshell: Suppose we could solve (nontrivially, please!) for $\underline{\pi}$ in $\pi_x p_{xy} = \pi_y p_{yx}$ (discrete-time) or $\pi_x q_{xy} = \pi_y q_{yx}$ (continuous-time). In both cases simple algebra then shows $\underline{\pi}$ solves the equilibrium equations.

NOTICE: the trivial solution $\pi_x \equiv 0$ won't do, as we also need $\sum_x \pi_x = 1$.

So on a prosaic level it is always worth trying this easy route; if the detailed balance equations are insoluble then revert to the more complicated equilibrium equations $\underline{\pi} \cdot \underline{\underline{P}} = \underline{\pi}$, respectively $\underline{\pi} \cdot \underline{\underline{Q}} = \underline{0}$. We will consider reversibility of Markov chains in both discrete and continuous time, the computation of equilibrium distributions for such chains, and discuss applications to some illustrative examples.

We will consider progressively more and more complicated Markov chains:

- simple symmetric random walk;
- the birth-death-immigration process;
- the $M/M/1$ queue;
- a discrete-time chain on a 8×8 state space;
- Gibbs' samplers (briefly);
- and Metropolis-Hastings samplers (briefly).

"Test understanding": a phrase signalling good questions you should ask yourself, to check you know what is going on.

Test understanding: show the detailed balance equations (discrete-case) lead to equilibrium equations by applying them to $\sum_x \pi_x p_{xy}$ and then using $\sum_x p_{yx} = 1$.

Simplest non-trivial example (I)

Consider *doubly-reflected simple symmetric random walk* X on $\{0, 1, \ldots, k\}$, with reflection "by prohibition": moves $0 \to -1$, $k \to k+1$ are replaced by $0 \to 0$, $k \to k$.

1. X is *irreducible* and *aperiodic*, so there is a unique equilibrium distribution $\underline{\pi} = (\pi_0, \pi_1, \ldots, \pi_k)$.

   Test understanding: explain why X is aperiodic when *non-reflected* simple symmetric random walk has period 2. Getting boundary conditions right is crucial both for this and for reversibility.

2. The *equilibrium equations* $\underline{\pi} \cdot \underline{\underline{P}} = \underline{\pi}$ are solved by $\pi_i = \frac{1}{k+1}$ for all i.

   Test understanding: verify solution of equilibrium equations.

3. Consider X in equilibrium and *run backwards in time*. Calculation: ANIMATION $\mathbb{P}\left[X_{n-1} = x | X_n = y\right] = \pi_x \mathbb{P}\left[X_n = y | X_{n-1} = x\right] / \pi_y = \mathbb{P}\left[X_n = y | X_{n-1} = x\right]$ so here *by symmetry of the kernel* the equilibrium chain has the same transition kernel (so looks the same) whether run forwards or backwards.

   - Develop Markov property to deduce $X_0, X_1, \ldots, X_{n-1}$ is conditionally independent of $X_{n+1}, X_{n+2}, \ldots$ given X_n. Hence reversed Markov chain is *still* Markov (though not necessarily time-homogeneous in more general circumstances). Suppose the reversed chain has kernel $\overline{p}_{y,x}$.
   - Use definition of conditional probability to compute $\overline{p}_{y,x} = \mathbb{P}\left[X_{n-1} = x\,,\, X_n = y\right] / \mathbb{P}\left[X_n = y\right]$,
   - then $\mathbb{P}\left[X_{n-1} = x\,,\, X_n = y\right] / \mathbb{P}\left[X_n = y\right] = \mathbb{P}\left[X_{n-1} = x\right] p_{x,y} / \mathbb{P}\left[X_n = y\right]$.
   - now substitute, using $\mathbb{P}\left[X_n = i\right] = \frac{1}{k+1}$ for all i so $\overline{p}_{y,x} = p_{x,y}$.
   - Symmetry of kernel ($p_{x,y} = p_{y,x}$) then shows backwards kernel $\overline{p}_{y,x}$ is same as forwards kernel $\overline{p}_{y,x} = p_{y,x}$.

   The construction generalizes …so link between reversibility and detailed balance holds generally. In particular, above still works even if random walk is asymmetric: the $p = q = \frac{1}{2}$ symmetry is *not* the point here!

Simplest non-trivial example (II)

There is a computational aspect to this.

1. Even in more general cases, if the π_i depend on i then above computations show reversibility holds if equilibrium distribution exists and *equations of detailed balance* hold: $\pi_x p_{x,y} = \pi_y p_{y,x}$.

   Test understanding: check this.

2. Moreover *if* one can solve for π_i in $\pi_x p_{x,y} = \pi_y p_{y,x}$ then it is easy to show $\underline{\pi} \cdot \underline{\underline{P}} = \underline{\pi}$.

   Test understanding: check this.

3. Consequently if one can solve the equations of detailed balance, and if the solution can be normalized to have unit total probability, then the result also solves the equilibrium equations.

Even in this simple example reversibility helps us deal with complexity. Detailed balance involves k equations each with two unknowns, easily "chained together". The equilibrium equations involve k equations of which $k-2$ involve three unknowns.

In general the detailed balance equations can be solved unless "chaining together by different routes" delivers inconsistent results. Kelly (1979) goes into more detail about this.

Test understanding: show detailed balance doesn't work for 3-state chain with transition probabilities $\frac{1}{3}$ for $0 \to 1$, $1 \to 2$, $2 \to 0$ and $\frac{2}{3}$ for $2 \to 1$, $1 \to 0$, $0 \to 2$.

Test understanding: show detailed balance *does* work for doubly reflected *asymmetric* simple random walk. We will see there are still major computational issues for more general Markov chains, connected with determining the normalizing constant to ensure $\sum_i \pi_i = 1$.

1.2 Population transitions

Birth-death-immigration process

The same idea works for continuous-time Markov chains: replace transition probabilities $p_{x,y}$ by rates $q_{x,y}$ and equilibrium equation $\underline{\pi} \cdot \underline{\underline{P}} = \underline{\pi}$ by differentiated variant using Q-matrix: $\underline{\pi} \cdot \underline{\underline{Q}} = \underline{0}$. (Recall: $\underline{\underline{Q}} = \frac{\mathrm{d}}{\mathrm{d}t}\, \underline{\underline{P}}_t$.)

Definition 1. The birth-death-immigration process has transitions:

- Birth ($X \to X+1$ at rate λX);
- Death ($X \to X-1$ at rate μX);
- plus an extra Immigration term ($X \to X+1$ at rate α).

Note that for this population process the rates $q_{x,x\pm1}$ make sense and are defined only for $x = 0, 1, 2, \ldots$.

Reversibility here is decidedly non-trivial. We need $0 \leq \lambda < \mu$ and $\alpha > 0$.

Hence $q_{x,x+1} = \lambda x + \alpha$; $q_{x,x-1} = \mu x$. ANIMATION

Equilibrium is easily derived from detailed balance:

$$\pi_x \quad = \quad \tfrac{\lambda(x-1)+\alpha}{\mu x} \cdot \tfrac{\lambda(x-2)+\alpha}{\mu(x-1)} \cdot \ldots \cdot \tfrac{\alpha}{\mu} \cdot \pi_0 \,.$$

Detailed balance equations:

$$\pi_x \times \mu x \quad = \quad \pi_{x-1} \times (\lambda(x-1)+\alpha)\,.$$

Normalizing constant can be computed exactly when $\lambda < \mu$ *via* generalized Binomial theorem:

$$\pi_0^{-1} \quad = \quad \sum_{x=0}^{\infty} \tfrac{\lambda(x-1)+\alpha}{\mu x} \cdot \tfrac{\lambda(x-2)+\alpha}{\mu(x-1)} \cdot \ldots \cdot \tfrac{\alpha}{\mu} \quad = \quad \left(\frac{\mu}{\mu-\lambda}\right)^{\frac{\alpha}{\lambda}}\,.$$

If the condition $\lambda < \mu$ is not satisfied then the sum does not converge and therefore there can be *no* equilibrium!

Note carefully: if $\alpha = 0$ then equilibrium = extinction.

Poisson process: $\lambda = \mu = 0$.

1.3 A key theorem

Detailed balance and reversibility

We summarise the notion of detailed balance in a definition and a theorem.

Definition 2. The Markov chain X satisfies *detailed balance* if

Discrete time: there is a non-trivial solution of $\pi_x p_{x,y} = \pi_y p_{y,x}$;

Continuous time: there is a non-trivial solution of $\pi_x q_{x,y} = \pi_y q_{y,x}$.

Theorem 3. *The irreducible Markov chain X satisfies detailed balance and the solution $\{\pi_x\}$ can be normalized by $\sum_x \pi_x = 1$ if and only if $\{\pi_x\}$ is an equilibrium distribution for X and X started in equilibrium is statistically the same whether run forwards or backwards in time.*

Proof of the theorem is routine: see example of random walk above.

The reversibility phenomenon has surprisingly deep ramifications! Consider birth-death-immigration example above and ask yourself whether it is immediately apparent that the time-reversed process in equilibrium should look statistically the same as the original process. (Note: both immigrations *and* births convert to deaths, and vice versa ....)

In general, if $\sum_x \pi_x < \infty$ is not possible then we end up with an *invariant measure* rather than an invariant probability distribution.

1.4 Queuing for insight

$M/M/1$ queue

We recall the $M/M/1$ queue example discussed in the preliminary notes.

Here we have

- Arrivals: $X \to X + 1$ at rate λ;
- Departures: $X \to X - 1$ at rate μ *if* $X > 0$.

Hence detailed balance: $\mu\pi_x = \lambda\pi_{x-1}$ and therefore when $\lambda < \mu$ (stability) the equilibrium distribution is $\pi_x = \rho^x(1-\rho)$ for $x = 0, 1, \ldots$, where $\rho = \frac{\lambda}{\mu}$ (the traffic intensity). ANIMATION

Reversibility/detailed balance is more than a computational device: consider Burke's theorem, if a stable $M/M/1$ queue is in equilibrium then people *leave* according to a Poisson process of rate λ.

Hence if a stable $M/M/1$ queue feeds into another stable $\cdot/M/1$ queue then in equilibrium the second queue on its own behaves as an $M/M/1$ queue in equilibrium.

Birth-death-immigration processes and queueing processes are both examples of *generalized birth-death processes*; only $X \to X \pm 1$ transitions, hence detailed balance equations easily solved.

Note: the $M/M/1$ queue is *non-linear*. Linearity allows solution of forwards equations: we do not discuss this here.

Detailed balance is also a subtle and important tool for the study of Markovian queueing networks (e.g. Kelly 1979).

The argument connecting reversibility to detailed balance runs both ways. If detailed balance equations can be solved to derive equilibrium then the process is reversible if run in equilibrium. Hence a one-line proof of Burke's theorem: if queue is run backwards in time then departures become arrivals.

Burke's theorem has deep consequences, with surprising applications (for example in the theory of random matrices).

Test understanding: use Burke's theorem for a feed-forward $\cdot/M/1$ queueing network (no loops) to show that in equilibrium each queue viewed in isolation is $M/M/1$. This uses the fact that independent thinnings and superpositions of Poisson processes are still Poisson . . . .

1.5 A simple multidimensional example

Random chess (Aldous and Fill 2001, Ch1, Ch3§2)

Now we turn to a multi-dimensional and less generic example.

Example 4 (A mean Knight's tour). Place a chess Knight at the corner of a standard 8×8 chessboard. Move it randomly, at each move choosing uniformly from available legal chess moves independently of the past.

This chain is periodic of period 2, and it is necessary in computation to take care about this. One can sub-sample the chain at even times to obtain an aperiodic chain, or (alternative approach) establish detailed balance between π^{even} and π^{odd}.

1. What is the equilibrium distribution?

   (use detailed balance)

   Use $\pi_v/d_v = \pi_u/d_u = c$ if $u \sim v$, where d_u is the degree of u. Also use fact, there are $168 = (1 \times 2 + 2 \times 3 + 5 \times 4 + 4 \times 6 + 4 \times 8) \times 4/2$ different edges. So total degree is 2×168, $1 = c \sum_{\text{black}} d_v = 168c$ and thus equilibrium probability at corner is $2c = 2/168$.

2. Is the resulting Markov chain periodic?

   (what if you sub-sample at even times?)

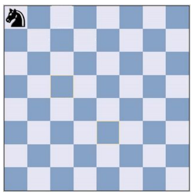
   Period 2 (white *versus* black). Sub-sampling at even times makes chain aperiodic on squares of one colour.

3. What is the mean time till the Knight returns to its starting point?

   (inverse of equilibrium probability)

   Inverse of equilibrium probability shows that mean return time to corner (allowing for periodicity!) is 168. Test understanding: follow through these calculations to check that you really do understand detailed balance. The following table gives a clue:

2	3	4	4
3	4	6	6
4	6	8	8
4	6	8	8



1.6 Ising model

Gibbs' sampler for Ising model (I) Ising model

- Pattern of spins $S_i = \pm 1$ on (finite fragment of) lattice (here i is typical node of lattice).

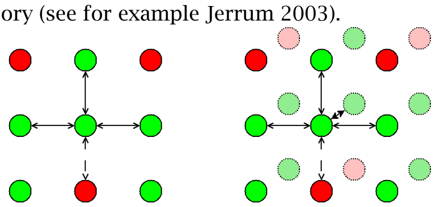
  Sample applications: idealized model for magnetism, simple binary image. Physics: interest in fragment expanding to fill whole lattice: cases of zero-interaction, sub-critical, critical ($\frac{kT}{J} = 2.269185$), super-critical. The Ising model is the nexus for a whole variety of scientific approaches, each bringing their own rather different questions.

- Probability mass function

$$\mathbb{P}\left[S_i = s_i \text{ all } i\right] \quad \propto \quad \begin{cases} \exp\left(J \sum \sum_{i \sim j} s_i s_j\right) \,, \\ \exp\left(J \sum \sum_{i \sim j} s_i s_j + H \sum_i s_i \tilde{s}_i\right) \\ \qquad \text{if external field } \tilde{s}_i \,. \end{cases}$$

  $i \sim j$ if i and j are lattice neighbours. Note, physics treatments use a (physically meaningful) over-parametrization $J \to \frac{J}{kT}$, $H \to mH$. The $H \sum_i s_i \tilde{s}_i$ term can be interpreted physically as modelling an external magnetic field, or statistically as a noisy image conditioning the image. In the latter case, $\tilde{s}_i$ can be viewed as a noisy image pixel, interacting with the true image pixel s_i but constrained to be fixed. Then H measures the "noisiness". For a simulation physics view of the Ising model, see the expository article by David Landau in Kendall et al. (2005).

Actually computing the normalizing constant here is *hard* in the sense of complexity theory (see for example Jerrum 2003).



Gibbs' sampler for Ising model (II) Gibbs' sampler (or heat-bath)

A *particular* example of the Gibbs' sampler in the special context of Ising models.

- Consider Markov chain with states which are Ising configurations on an $n \times n$ lattice, moving as follows:

  View configurations as vectors of ±1's listing spins at different sites.

  - Set $\underline{s}$ to be a given configuration, with $\underline{s}^{(i)}$ obtained by flipping spin i,
  - Choose a site i in the lattice at random;

- Compute the conditional probability $\mathbb{P}\left[\underline{s}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right]$ of current configuration given configuration at other sites;

  $\{\underline{s}^{(i)},\underline{s}\}$ is the event that we see configuration $\underline{s}$ except perhaps at state i.
- Flip the current value of S_i with probability $\mathbb{P}\left[\underline{s}^{(i)}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right]$, otherwise leave unchanged.

In case of the Ising model, noting that $s_i^{(i)} = -s_i$, careful calculation yields

$$\mathbb{P}\left[\underline{s}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right] \quad = \quad \frac{\exp\left(J\sum_{j:j\sim i} s_i s_j\right)}{\exp\left(J\sum_{j:j\sim i} s_i s_j\right) + \exp\left(-J\sum_{j:j\sim i} s_i s_j\right)}\,.$$

(Obvious changes if external field.)

- Simple general calculations show,

$$\sum_i \frac{1}{n^2}\,\mathbb{P}\left[\underline{s}^{(i)}\right] \times \mathbb{P}\left[\underline{s}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right] \quad = \quad \mathbb{P}\left[\underline{s}\right]$$

  so chain has Ising model as equilibrium distribution.

  This is really a completely general computation! Note that the equilibrium equations are complicated: n^2 equations, each with n^2 terms on left-hand side.

General pattern for Gibbs sampler: update individual random variables *sequentially* using conditional distributions given all other random variables.
Conditional distributions, so ratios, so normalizing constants cancel out.

Gibbs' sampler for Ising model (III) Detailed balance

- Detailed balance calculations provide a much easier justification: merely check

$$\frac{1}{n^2}\,\mathbb{P}\left[\underline{s}^{(i)}\right] \times \mathbb{P}\left[\underline{s}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right] \quad = \quad \frac{1}{n^2}\,\mathbb{P}\left[\underline{s}\right] \times \mathbb{P}\left[\underline{s}^{(i)}\,\middle|\,\{\underline{s}^{(i)},\underline{s}\}\right]\,.$$

  Test understanding: check the detailed balance calculations. This also works for processes obtained from:

  - systematic scans
  - coding ("simultaneous updates on alternate colours of a chessboard")

  but *not* for wholly simultaneous updates.

- Here is an animation of a Gibbs' sampler producing an Ising model conditioned by a noisy image, produced by systematic scans: 128 × 128, with 8 neighbours. Noisy image to left, draw from Ising model to right.

ANIMATION

The example is taken from a discussion of "perfect simulation", but that is another story! See

```
www.warwick.ac.uk/go/wsk/ising-animations
```

for more on perfect sampling for the Ising model.

There is an immediate Bayesian interpretation: if the noisy pixels aren't held fixed then we get the underlying joint distribution of signal and noise as equilibrium distribution. If they are held fixed then the equilibrium has (by reversibility) to be proportional to the joint distribution subject to the restriction, and this is exactly the posterior distribution.

1.7 Metropolis-Hastings sampler

Metropolis-Hastings

1. An important alternative to the Gibbs' sampler, even more closely connected to detailed balance:

   Actually the Gibbs' sampler is a special case of the Metropolis-Hastings sampler.

   - Suppose $X_n = x$;
   - Pick y using a transition probability kernel $q(x, y)$ (the *proposal kernel*);
   - *accept* the proposed transition $x \to y$ with probability

   $$\alpha(x, y) \quad = \quad \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} .$$

   - if transition accepted, set $X_{n+1} = y$; otherwise set $X_{n+1} = x$.

     Test understanding: write down the transition probability kernel for X. Test understanding: check that π solves the detailed balance equations.

2. Since π satisfies detailed balance therefore π is an equilibrium distribution (if the chain converges to a unique equilibrium!).

   Common variations on choice of proposal kernel:

   - *independence sampler*: $q(x, y) = f(y)$;
   - *random-walk sampler*: $q(x, y) = f(y - x)$;
   - *Langevin sampler*: replace random-walk shift by shift depending on $\operatorname{grad}\log\pi$.

Ratio $\pi(x)/\pi(y)$, so normalizing constants cancel out.

Slice sampler

Here is an attractive special case of MCMC.

- Suppose we wish to draw from a continuous density of bounded range.
- This is the same as drawing from the uniform distribution over the region between density and x-axis.
- Given a point (X, Y) in this region, first discard X and draw uniformly from segment of fixed Y under density.
- Then discard Y, and draw uniformly from vertical segment between x-axis and density.
- Repeat: result is a Gibbs sampler from the uniform distribution under the density. ANIMATION

This univariate case is rather trivial, but there are useful multivariate generalizations. It turns out that the slice sampler is very amenable to calculation!

2 Martingales

This is the second major theme of these notes: *martingales* are a class of random processes which are closely linked to ideas of conditional expectation.

Briefly, martingales model your fortune if you are playing a fair game. (There are associated notions of “supermartingale”, for a game unfair to you, and “submartingale”, for a game fair to you.) But martingales can do so much more! They are fundamental to the theory of how one’s predictions should evolve as time progresses.

In this section we discuss a wide range of different martingales.

Martingales

> “One of these days . . . a guy is going to come up to you and show you a nice brand-new deck of cards on which the seal is not yet broken, and this guy is going to offer to bet you that he can make the Jack of Spades jump out of the deck and squirt cider in your ear. But, son, do not bet this man, for as sure as you are standing there, you are going to end up with an earful of cider.” Frank Loesser, *Guys and Dolls* musical, 1950, script

2.1 Simplest possible example

Martingales pervade modern probability

We use X as a convenient abbreviation for the stochastic process $\{X_n : n \geq 0\}$, et cetera.

1. We say the random process X is a martingale if it satisfies the martingale property:

   $$\mathbb{E}\left[X_{n+1}|X_n, X_{n-1}, \ldots\right] = \mathbb{E}\left[X_n \text{ plus jump at time } n|X_n, X_{n-1}, \ldots\right] \quad = \quad X_n\,.$$

   For a conversation with the inventor, see `www.dartmouth.edu/~chance/Doob/conversation.html`.

2. Simplest possible example: simple symmetric random walk $X_0 = 0$, X_1, X_2, ... The martingale property follows from independence and distributional symmetry of jumps.

   Expected future level of X is current level.

3. For convenience and brevity, we often replace $\mathbb{E}\left[\ldots|X_n, X_{n-1}, \ldots\right]$ by $\mathbb{E}\left[\ldots|\mathcal{F}_n\right]$ and think of “conditioning on $\mathcal{F}_n$” as “conditioning on all events which can be determined to have happened by time n”.

   We use $\mathcal{F}_n$ notation without comment in future, usually representing conditioning by X_0, X_1, . . . , X_n (if X is martingale in question). *Sometimes* further conditioning will be added; but $\mathcal{F}_{n+1}$ always represents at least as much conditioning as $\mathcal{F}_n$. Crucially, the “Tower property” of conditional expectation then applies:

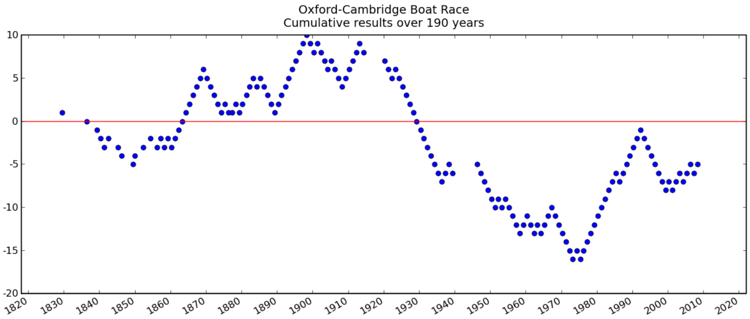
   $$\mathbb{E}\left[\mathbb{E}\left[Z|\mathcal{F}_{n+1}\right]|\mathcal{F}_n\right] \quad = \quad \mathbb{E}\left[Z|\mathcal{F}_n\right]\,.$$

   Test understanding: deduce

   $$\mathbb{E}\left[X_{n+k}|\mathcal{F}_n\right] \quad = \quad X_n\,.$$

There is an extensive theory about the notion of a *filtration of σ-algebras* (also called *σ-fields*), $\{\mathcal{F}_n : n \geq 0\}$. We avoid going into details . . . .

University Boat Race results over 190 years



Could this represent a martingale?

WSK first became aware of the boat race in about 1970, at which time the martingale property would have seemed not to apply.

There is now a much more satisfactory balance (especially for WSK, who studied at Oxford!), but one might still doubt validity of martingale property here . . .

Suppose $\mathbb{P}[\text{Oxford win}] = p \neq \frac{1}{2}$. Where can one find a martingale in this asymmetric case? Set X_n to be difference of wins minus losses and consider

$$\mathbb{E}[X_{2020}|\mathcal{F}_n] \quad = \quad (2020 - n)(2p - 1) + X_n ,$$

so (iterated expectations) $Y_n = (2020 - n)(2p - 1) + X_n$ *is* a martingale for $1830 \leq n \leq 2020$ (if you forgive the occasional intermissions).

2.2 Thackeray's martingale

Thackeray's martingale

1. MARTINGALE:
   - spar under the bowsprit of a sailboat;
   - a harness strap that connects the nose piece to the girth; prevents the horse from throwing back its head.

2. MARTINGALE in gambling: The original sense is given in the OED: “a system in gambling which consists in doubling the stake when losing in the hope of eventually recouping oneself.” The oldest quotation is from 1815 but the nicest is from 1854: Thackeray in *The Newcomes* I. 266 “You have not played as yet? Do not do so; above all avoid a martingale if you do.”

   This is the “doubling” strategy. The equestrian meaning resembles the probabilistic definition to some extent.

   Another nice quotation is the following:

   > “I thought there was something of wild talent in him, mixed with a due leaven of absurdity, – as there must be in all talent, let loose upon the world, without a martingale.” Lord Byron, Letter 401. to Mr Moore. Dec. 9. 1820, writing about an Irishman Muley Moloch.

   Notice how the randomness of Thackeray’s martingale is the same as for a simple symmetric random walk.
   Test understanding: compute the expected value of M_n from first principles.

3. Result of playing Thackeray’s martingale system and stopping on first win: ANIMATION set fortune at time n to be M_n. If $X_1 = -1, \ldots, X_n = -n$ then $M_n = -1 - 2 - \ldots - 2^{n-1} = 1 - 2^n$, otherwise $M_n = 1$.

   Test understanding: what should be the value of $\mathbb{E}\left[\widetilde{M}_n\right]$ if $\widetilde{M}$ is computed as for M but stopping play if M hits level $1 - 2^N$? (Think about this, but note that a satisfactory answer has to await discussion of optional stopping theorem in next section.)

2.3 Populations

Martingales and populations

1. Consider a branching process Y: population at time n is Y_n, where $Y_0 = 1$ (say) and Y_{n+1} is the sum $Z_{n,1} + \ldots + Z_{n,Y_n}$ of Y_n independent copies of a non-negative integer-valued *family-size r.v.* Z.

   New Yorker’s definition of branching process (to be read out aloud in strong New York accent): “You are born. You live a while. You have a random number of kids. You die. Your children are completely independent of you, but behave in exactly the same way.” The formal definition requires the $Z_{n,i}$ to be independent of $Y_0, \ldots, Y_n$.

2. Suppose $\mathbb{E}[Z] = \mu < \infty$. Then $X_n = Y_n/\mu^n$ defines a martingale.

   Test understanding: check this example. Note, X measures relative deviation from the deterministic Malthusian model of growth.

3. Suppose $\mathbb{E}\left[s^Z\right] = G(s)$. Let $H_n = Y_0 + \ldots + Y_n$ be total of all populations up to time n. Then $s^{H_n}/(G(s)^{H_{n-1}})$ defines a martingale.

   Test understanding: check this example. What interpretation can you put on s^{H_n}?

4. In all these examples we can use $\mathbb{E}[\ldots|\mathcal{F}_n]$, representing conditioning by all $Z_{m,i}$ for $m \leq n$.

   Indeed, we can also generalize to general Y_0.

2.4 Definitions

Definition of a martingale

It is useful to have a general definition of expectation here (see the section on conditional expectation in the preliminary notes).

Formally:

Definition 5. X is a *martingale* if $\mathbb{E}[|X_n|] < \infty$ (for all n) and

$$X_n \quad = \quad \mathbb{E}[X_{n+1}|\mathcal{F}_n] \ .$$

It is important that the X_n are integrable.

It is a consequence that X_n is part of the conditioning expressed by $\mathcal{F}_n$.
Sometimes we expand the reference to $\mathcal{F}_n$:

$$X_n \quad = \quad \mathbb{E}[X_{n+1}|X_n, X_{n-1}, \ldots, X_1, X_0] \ .$$

Supermartingales and submartingales

Two associated definitions

Definition 6. $\{X_n\}$ is a *supermartingale* if $\mathbb{E}[|X_n|] < \infty$ (for all n) and

$$X_n \quad \geq \quad \mathbb{E}[X_{n+1}|\mathcal{F}_n] \ ,$$

(and X_n forms part of conditioning expressed by $\mathcal{F}_n$).

It is important that the X_n are integrable. It is now *not* automatic that X_n forms part of the conditioning expressed by $\mathcal{F}_n$, and it is therefore important that this requirement is part of the definition.

Definition 7. $\{X_n\}$ is a *submartingale* if $\mathbb{E}[|X_n|] < \infty$ (for all n) and

$$X_n \quad \leq \mathbb{E}[X_{n+1}|\mathcal{F}_n] \ ,$$

(and X_n forms part of conditioning expressed by $\mathcal{F}_n$).

It is important that the X_n are integrable. Again it is important that X_n forms part of the conditioning expressed by $\mathcal{F}_n$. How to remember the difference between "sub-" and "super-"? Suppose $\{X_n\}$ measures your fortune in a casino gambling game. Then "sub-" is bad and "super-" is good *for the casino!*

Wikipedia: life is a supermartingale, as one's expectations are always no greater than one's present state.

Examples of supermartingales and submartingales

Test understanding: check all these examples.

In each case the general procedure is as follows: compare $\mathbb{E}[X_{n+1}|\mathcal{F}_n]$ to X_n.

1. Consider asymmetric simple random walk: supermartingale if jumps have negative expectation, submartingale if jumps have positive expectation.

2. This holds even if the walk is stopped on first return to 0.

3. Consider Thackeray's martingale based on asymmetric random walk. This is a supermartingale or a submartingale depending on whether jumps have negative or positive expectation.

4. Consider branching process $\{Y_n\}$ and consider Y_n on its own instead of Y_n/μ^n. This is a supermartingale if $\mu < 1$ (sub-critical case), a submartingale if $\mu > 1$ (super-critical case), a martingale if $\mu = 1$ (critical case).

   Note that all martingales are automatically both sub- and supermartingales, and moreover they are the *only* processes to be both sub- and supermartingales.

2.5 More martingale examples

More martingale examples

Test understanding: check both of these examples.

It is instructive to try to figure out why it is "obvious" that the second example is a martingale. (Hint: it's about symmetry . . .)

On the other hand, the first example yields a martingale because

$$p \times \left(\frac{1-p}{p}\right) + (1-p) \times \left(\frac{1-p}{p}\right)^{-1} \quad = \quad 1\,.$$

After some training, one can often spot martingales like this almost on sight.

1. Repeatedly toss a coin, with probability of heads equal to p: each Head earns £1 and each Tail loses £1. Let X_n denote your fortune at time n, with $X_0 = 0$. Then

   $$\left(\frac{1-p}{p}\right)^{X_n} \qquad \text{defines a martingale.}$$

2. A shuffled pack of cards contains b black and r red cards. The pack is placed face down, and cards are turned over one at a time. Let B_n denote the number of black cards left *just before* the n^{th} card is turned over:

   $$\frac{B_n}{r+b-(n-1)}\,,$$

   the proportion of black cards left just before the n^{th} card is revealed, defines a martingale.

2.6 Finance example

An example of importance in finance

Here (modifications of) Y_n provides the simplest model for market price fluctuations appropriately discounted.

1. Suppose N_1, N_2, ... are independent identically distributed normal random variables of mean 0 and variance σ^2, and put $S_n = N_1 + \ldots + N_n$.

   In fact $\{S_n\}$ is a martingale, though this is not the point here.

2. Then the following is a martingale:

$$Y_n \quad = \quad \exp\left(S_n - \tfrac{n}{2}\sigma^2\right) .$$

ANIMATION

   Test understanding: Prove this! *Hint:* $\mathbb{E}\left[\exp(N_1)\right] = e^{\sigma^2/2}$.

3. A modification exists for which the N_i have non-zero mean μ. *Hint:* $S_n \to S_n - n\mu$.

   Test understanding: figure out the modification!

   A continuous-time variation on this (using Brownian motion) is an important baseline model in mathematical finance. Note that the martingale can be expressed as

$$Y_{n+1} = Y_n \exp\left(N_{n+1} - \tfrac{\sigma^2}{2}\right) .$$

2.7 Martingales and likelihood

Martingales and likelihood

- Suppose independent random variables X_1, X_2, ... are observed at times 1, 2, ... Write down likelihood at time n:

$$L(\theta; X_1, \ldots, X_n) \quad = \quad p(X_1, \ldots, X_n | \theta) .$$

  Simple case of normal data with unknown mean θ:

$$L(\theta; X_1, \ldots, X_n) \quad \propto \quad \exp\left(-\frac{1}{2\sigma^2}\sum_1^n (X_i - \theta)^2\right) .$$

- If θ_0 is "true" value then (computing expectation with $\theta = \theta_0$)

$$\mathbb{E}\left[\frac{L(\theta_1; X_1, \ldots, X_{n+1})}{L(\theta_0; X_1, \ldots, X_{n+1})} \middle| \mathcal{F}_n\right] \quad = \quad \frac{L(\theta_1; X_1, \ldots, X_n)}{L(\theta_0; X_1, \ldots, X_n)}$$

  Hence likelihood ratios are really the same thing as martingales.

  The martingale in the finance example can also arise in this way, as the likelihood ratio between two different values of θ if the model is that the X_i are independent identically distributed $N(\theta, \sigma^2)$.

2.8 Chicken Little

The “Chicken Little” example

1. A comet may or may not collide with Earth in n days time. Chaotic dynamics: model by supposing comet may follow one of n possible paths, of which just one leads to collision at day n.

   Considerable simplification of chaotic dynamics, but not unreasonable.

2. Each day, new observations eliminate exactly one of possible paths: path to be eliminated on day r is chosen from $n - r + 1$ surviving paths uniformly at random and independently of the past.

   This models the fact that observations are hard to come by, and do not provide much information. For example, see `http://en.wikipedia.org/wiki/99942_Apophis`. This near-Earth asteroid, when discovered in 2004, was estimated to have a 2.7% chance of hitting the Earth in 2029. Further observations have reduced this figure, and as of 16/04/08, the impact probability for April 13 2036 (the most likely collision date) fell to 1 in 45,000. More recently (07/10/09) risk has been downgraded to 4×10^{-6}. See `http://neo.jpl.nasa.gov/news/news146.html` for an informative contemporary account.

3. Compute conditional collision probability at day r, supposing collision path is not yet eliminated. Deduce that conditional collision probabilities at days $r = 0, 1, \ldots, n$ form a martingale.

   Let D be indicator random variable indicating event that collision occurs, and compute $\mathbb{E}[D|\mathcal{F}_r]$ where $\mathcal{F}_r$ captures information of whether or not collision occurs by day r. Probability of collision grows more and more rapidly ($\frac{1}{n-r}$ on day r) till either it suddenly falls to zero (if collision path eliminated before n) or collision actually occurs (if collision path not eliminated before day n). Therefore collision probability increases day by day (engendering increasing despair), until it (hopefully) falls to zero (engendering mass relief).

3 Stopping times

Playing a fair game, what happens if you adopt a strategy of leaving the game at a random time? For "reasonable" random times, this should offer you no advantage. Here we seek to make sense of the term "reasonable". Note that the gambling motivation is less frivolous than it might appear. Mathematical finance is about developing trading strategies (complex gambles!) aimed at controlling uncertainty.

Stopping times

> "Hurry please it's time." T. S. Eliot, *The Waste Land*, 1922

Stopping times

Martingales M stopped at "nice" times are still martingales. In particular, for a "nice" random T,

$$\mathbb{E}[M_T] \quad = \quad \mathbb{E}[M_0] \ .$$

How can T fail to be "nice"? Consider simple symmetric random walk X begun at 0.

For a random time T to be "nice", two things are required:

1. T must not "look ahead";

   Example of "looking ahead": Set $S = \sup\{X_n : 0 \leq n \leq 10\}$ and set $T_2 = \inf\{n : X_n = S\}$. Then $\mathbb{E}\left[X_{T_2}\right] \geq \mathbb{P}[S > 0] > 0 = \mathbb{E}[X_0]$

2. T must not be "too big". ANIMATION

   Example of being "too big": $T_1 = \inf\{n : X_n = 1\}$ so (assuming T_1 is almost surely finite) $\mathbb{E}\left[X_{T_1}\right] = 1 > 0 = \mathbb{E}[X_0]$. This is the nub of the matter for the Thackeray example.

3. Note that random times T turning up in practice often have positive chance of being infinite.

   Example of possibly being infinite: asymmetric simple random walk X begun at 0, $\mathbb{E}[X_1] < 0$, $T_1 = \inf\{n : X_n = 1\}$ as above.

3.1 "No-look-ahead" condition

Non-obvious "no-look-ahead" condition

Definition 8. A non-negative integer-valued random variable T is said to be a stopping time if (equivalently) for all n

- $[T \leq n]$ is determined by information at time n;
- or $[T \leq n] \in \mathcal{F}_n$

- or we can write *rules* (Bernoulli random variables) ζ_0, ζ_1, ... with ζ_n in $\mathcal{F}_n$, such that
$$[\zeta_n = 1] \quad = \quad [T \leq n]\,.$$

Note that we need to have a clear notion of exactly what might be $\mathcal{F}_n$, information revealed by time n.

Here is a poetical illustration of a *non-stopping time*, due to David Kendall:

> There is a rule for timing toast, You never need to guess; Just wait until it starts to smoke, And then ten seconds less.

(Adapted from a "grook" by Piet Hein, *Grooks II* MIT Press, 1968.)

Recall the example on previous slide of T being the time to hit 1 for a negatively-biased simple random walk begun at 0: stopping times can have positive chance of being infinite.

3.2 Random walk example

Example using random walks

Let X be a random walk begun at 0.

X need not be symmetric, need not be simple. Indeed a Markov chain or even a general random process would do.

- The random time $T = \inf\{n > 0 : X_n \geq 10\}$ is a stopping time.

  We could replace $n > 0$ by $n \geq 0$, $X \geq 10$ by $X \in A$ for some subset A of state-space, ...: thus we could have $T_A = \inf\{n > 0 : X_n \in A\}$ (the "hitting time on A").

- Indeed $[T \leq n]$ is clearly determined by information at time n:
$$[T \leq n] \quad = \quad [X_1 \geq 10] \cup \ldots \cup [X_n \geq 10]\,.$$
  In case of hitting time on A,
$$[T_A \leq n] \quad = \quad [X_1 \in A] \cup \ldots \cup [X_n \in A]$$
  so $[T_A \leq n]$ is determined by information at time n, so T_A is a stopping time.

- Finally, T is typically "too big": so long as it is almost surely finite, we find that $0 = \mathbb{E}[X_0] < \mathbb{E}[X_T]$.
  Finiteness is the case if $\mathbb{E}[X_1] > 0$ or if $\mathbb{E}[X_1] = 0$ and $\mathbb{P}[X_1 > 0] > 0$.

  General hitting times T_A need not be "too big": example if X is simple symmetric random walk begun at 0 and $A = \{\pm 10\}$.

3.3 Branching process example

Example using branching processes

Let Y be a branching process of mean-family-size μ (so $X_n = Y_n/\mu^n$ determines a martingale), with $Y_0 = 1$.

So $Y_n = Z_{n-1,1} + \ldots + Z_{n-1,Y_{n-1}}$ for independent family sizes $Z_{m,j}$.

- The random time $T = \inf\{n : Y_n = 0\} = \inf\{n : X_n = 0\}$ is a stopping time.

  For a more interesting example, consider

  $$S = \inf\{n : \text{ at least one family of size 0 before } n\}$$

- Indeed $[T \leq n]$ is clearly determined by information at time n:

  $$[T \leq n] = [Y_n = 0]$$

  since $Y_{n-1} = 0$ implies $Y_n = 0$ et cetera.

  In case of S, consider

  $$[S \leq n] = A_0 \cup A_1 \cup \ldots \cup A_{n-1}$$

  where $A_i = [Z_{i,j} = 0 \text{ for some } j \leq Y_i]$. Thus $[S \leq n]$ is determined by information at time n, so S is a stopping time.

- Again T here is "too big": so long as it is almost surely finite then $1 = \mathbb{E}[X_0] > \mathbb{E}[X_T]$.

  Finiteness occurs if $\mu < 1$, or if $\mu = 1$ and there is positive chance of zero family size.

  It is important to be clear about what *is* information provided at time n. Here we suppose it to be made up only of the sizes of families produced by individuals in generations $0, 1, \ldots, n-1$. Other choices are possible, of course.

3.4 Events revealed by stopping time

Events revealed by the time of a stopping time T

Suppose T is a stopping time.

Definition 9. The "pre-T σ-algebra" $\mathcal{F}_T$ is composed of events which, if T does not occur later than time n, are themselves determined at time n. Thus:

$$A \in \mathcal{F}_T \quad \text{if} \quad A \cap [T \leq n] \in \mathcal{F}_n \text{ for all } n\,.$$

Here we are very close to having to take measure theory seriously .... Measure theory is required to unwrap the meaning of the word "determined" in the definition.

Consider random walk X begun at 0 and the stopping time $T = \inf\{n : X_n \geq 10\}$. Then the event $[X_{15} < 5 \text{ and } T > 15]$ is in the pre-T σ-algebra $\mathcal{F}_T$.

Definition 10. Random variables Z are said to be "$\mathcal{F}_T$-measurable" if events made up from them ($[Z \leq z], \ldots$) are in pre-T σ-algebra $\mathcal{F}_T$.

The random variable $X_{\min\{15,T\}}$ is $\mathcal{F}_T$-measurable.

Consider the branching process example with S being the time at which a zero-size family is first encountered. Then

$$Y_0 + Y_1 + \ldots + Y_S \in \mathcal{F}_S\,.$$

3.5 Optional Stopping Theorem

Optional stopping theorem

Theorem 11. *Suppose M is a martingale and $S \leq T$ are two bounded stopping times. Then*

$$\mathbb{E}\left[M_T | \mathcal{F}_S\right] \quad = \quad M_S\,.$$

We can generalize to general stopping times $S \leq T$ either if M is bounded or (more generally) if M is "uniformly integrable".

Uniform integrability: note we can take expectation of a single random variable X exactly when $\mathbb{E}\left[|X|; |X| > n\right] \to 0$ as $n \to \infty$. (This fails when $\mathbb{E}\left[|X|; |X| > n\right] = \infty$!).

Uniform integrability requires this to hold *uniformly* for a whole collection of random variables X_i:

$$\lim_{n\to\infty} \sup_i \mathbb{E}\left[|X_i|; |X_i| > n\right] \quad = \quad 0\,.$$

Examples: if the X_i are bounded; if there is a single non-negative random variable Z with $\mathbb{E}\left[Z\right] < \infty$ and $|X_i| \leq Z$ for all i; if the p-moments $\mathbb{E}\left[X_i^p\right]$ are bounded for some $p > 1$.

3.6 Application to gambling

Gambling: you shouldn't expect to win

Suppose your fortune in a gambling game is X, a martingale begun at 0 (for example, a simple symmetric random walk). If N is the maximum time you can spend playing the game, and if $T \leq N$ is a *bounded* stopping time, then

$$\mathbb{E}\left[X_T\right] \quad = \quad 0\,.$$

There are exceptions, for example Blackjack (using card-counting: `en.wikipedia.org/wiki/Card_counting`).

I find strategies proposed for other games to be less convincing, for example the Labouchére system favoured by Ian Fleming (`en.wikipedia.org/wiki/Labouch\%C3\%A8re_system`):

> The Labouchére system, also called the cancellation system, is a gambling strategy used in roulette. The user of such a strategy decides before playing how much money they want to win, and writes down a list of positive numbers that sum to the predetermined amount. With each bet, the player stakes an amount equal to the sum of the first and last numbers on the list. If only one number remains, that number is the amount of the stake. If bet is successful, the two amounts are removed from the list. If the bet is unsuccessful, the amount lost is appended to the end of the list. This process continues until either the list is completely crossed out, at which point the desired amount of money has been won, or until the player runs out of money to wager.

Contrast Fleming (1953):

> "Then the Englishman, Mister Bond, increased his winnings to exactly three million over the two days. He was playing a progressive system on red at table five. . . . It seems that he is persevering and plays in maximums. He has luck."

3.7 Hitting times

Martingales and hitting times

Suppose $X_1, X_2, \ldots$ are independent Gaussian random variables of mean $-\mu < 0$ and variance 1. Let $S_n = X_1 + \ldots + X_n$ and let T be the time when S first exceeds level $\ell > 0$.

So $T = \inf\{n : S_n \geq \ell\}$.

Then $\exp\left(\alpha(S_n + \mu n) - \frac{\alpha^2}{2} n\right)$ determines a martingale, and the optional stopping theorem can be applied to show

$$\mathbb{E}\left[\exp\left(-pT\right)\right] \quad \sim \quad e^{-(\mu+\sqrt{\mu^2+2p})\ell}\,.$$

Use the optional stopping theorem on the bounded stopping time $\min\{T, n\}$:

$$\mathbb{E}\left[\exp\left(\alpha S_{\min\{T,n\}} + \alpha(\mu - \frac{\alpha}{2})\min\{T, n\}\right)\right] \quad = \quad 1\,.$$

Use careful analysis of the left-hand side, letting $n \to \infty$, large ℓ,

$$\mathbb{E}\left[\exp\left(\alpha\ell + \alpha(\mu - \frac{\alpha}{2})T\right)\right] \quad \sim \quad 1\,.$$

($S_{\min\{T,n\}}$ is relatively close to ℓ, $\min\{T, n\}$ is relatively close to T)

Now set $\alpha = \mu + \sqrt{\mu^2 + 2p} > 0$, so $\alpha(\mu - \frac{\alpha}{2}) = -p$:

$$\mathbb{E}\left[\exp\left(-pT\right)\right] \quad \sim \quad \exp\left(-(\mu + \sqrt{\mu^2 + 2p})\ell\right)\,.$$

This improves to an equality, at the expense of using more advanced theory, if we replace the Gaussian random walk S by Brownian motion.

Improvement: Brownian motion is continuous in time and so cannot jump over the level ℓ without hitting it.

3.8 Martingale convergence

Martingale convergence

Theorem 12. *Suppose X is a non-negative supermartingale. Then $Z = \lim X_n$ exists, moreover $\mathbb{E}\left[Z|\mathcal{F}_n\right] \leq X_n$.*

ANIMATION

At the heart of the argument here is the famous "upcrossings" result ...: use the supermartingale property and non-negativity to control the number of times a supermartingale can cross up from a fixed low level to a fixed high level.

Consider symmetric simple random walk begun at 1 and *stopped at* 0: $X_n = Y_{\min\{n,T\}}$ if $T = \inf\{n : Y_n = 0\}$ and Y is symmetric simple random walk. Clearly X_n is non-negative; clearly $X_n = Y_{\min\{n,T\}} \to Z = 0$, since Y will eventually hit 0; clearly $0 = \mathbb{E}\left[Z|\mathcal{F}_n\right] \leq X_n$ since $X_n \geq 0$.

Theorem 13. *Suppose X is a bounded martingale (or, more generally, uniformly integrable). Then $Z = \lim X_n$ exists, moreover $\mathbb{E}\left[Z|\mathcal{F}_n\right] = X_n$.*

Thus symmetric simple random walk Y begin at 0 and stopped at ± 10 must converge to a limiting value Z. Evidently $Z = \pm 10$. Moreover since $\mathbb{E}\left[Z|\mathcal{F}_n\right] = Y_n$ we deduce $\mathbb{P}\left[Z = 10|\mathcal{F}_n\right] = \frac{Y_n+10}{20}$.

Theorem 14. *Suppose X is a martingale and $\mathbb{E}\left[X_n^2\right] \leq K$ for some fixed constant K. Then one can prove directly that $Z = \lim X_n$ exists, moreover $\mathbb{E}\left[Z|\mathcal{F}_n\right] = X_n$.*

Sketch argument: from martingale property

$$0 \leq \mathbb{E}\left[(X_{m+n} - X_n)^2 \middle| \mathcal{F}_n\right] \quad = \quad \mathbb{E}\left[X_{m+n}^2 \middle| \mathcal{F}_n\right] - X_n^2\,;$$

hence $\mathbb{E}\left[X_n^2\right]$ is non-decreasing; hence it converges to a limiting value; hence $\mathbb{E}\left[(X_{m+n} - X_n)^2\right]$ tends to 0.

Birth-death process revisited

Y is a discrete-time birth-death process *absorbed at zero*:

$$p_{k,k+1} = \frac{\lambda}{\lambda+\mu}, \quad p_{k,k-1} = \frac{\mu}{\lambda+\mu}, \qquad \text{for } k > 0, \text{ with } 0 < \lambda < \mu.$$

This is the discrete-time analogue of the birth-death-immigration process of Section 1 with $\alpha = 0$ (so no immigration).

This is a non-negative supermartingale and so $\lim Y_n$ exists.

Test understanding: show that Y is a supermartingale, and use the SLLN to show that $Y_n \to 0$ almost surely as $n \to \infty$. In Section 1 we computed the equilibrium distribution and concluded that

$$\pi_0^{-1} = \left(\frac{\mu}{\mu-\lambda}\right)^{\frac{\alpha}{\lambda}},$$

and so with $\alpha = 0$ the equilibrium distribution is simply extinction of the process, in agreement with what you have just shown.

Now let $T = \inf\{n : Y_n = 0\}$: $T < \infty$ a.s. Then

$$X_n = Y_{n\wedge T} + \left(\frac{\mu-\lambda}{\mu+\lambda}\right)(n \wedge T)$$

is a non-negative (super)martingale converging to $Z = \frac{\mu-\lambda}{\mu+\lambda}T$.

Here we have written $n \wedge T$ for $\min\{n, T\}$.

Test understanding: show that X is a martingale.

Thus (recalling that $X_0 = Y_0$)

$$\mathbb{E}[T] \leq \left(\frac{\mu+\lambda}{\mu-\lambda}\right) X_0 .$$

Markov's inequality then implies that

$$\mathbb{P}[T > k] \leq \left(\frac{\mu+\lambda}{\mu-\lambda}\right)\frac{X_0}{k}.$$

Likelihood revisited

Suppose i.i.d. random variables $X_1, X_2, \ldots$ are observed at times $1, 2, \ldots,$ and suppose the common density is $f(\theta; x)$. Recall that, if the "true" value of θ is θ_0, then

$$M_n = \frac{L(\theta_1; X_1, \ldots, X_n)}{L(\theta_0; X_1, \ldots, X_n)}$$

is a martingale, with $\mathbb{E}[M_n] = 1$ for all $n \geq 1$.

Test understanding: The result is still true even if the random variables are neither independent nor identically distributed. Show this is true!

Remember that the expectation is computed using $\theta = \theta_0$.

The SLLN and Jensen's inequality show that

$$\frac{1}{n}\log M_n \to -c \quad \text{as } n \to \infty,$$

moreover if $f(\theta_0;\cdot)$ and $f(\theta_1;\cdot)$ differ as densities then $c > 0$, and so $M_n \to 0$.

Jensen's inequality for *concave* functions is opposite to that for convex functions: if ψ is concave then $\mathbb{E}[\psi(X)] \leq \psi(\mathbb{E}[X])$. Moreover if X is non-deterministic and ψ is strictly concave then the inequality is strict.

The rate of convergence of M_n is geometric if the difference between θ_0 and θ_1 is identifiable.

Note that this is in keeping with hypothesis testing: as more information is gathered, so we would expect the evidence against θ_1 to accumulate, and the likelihood ratio to tend to zero.

3.9 Harmonic functions

Martingales and bounded harmonic functions

- Consider a discrete state-space Markov chain X with transition kernel p_{ij}. Suppose $f(i)$ is a bounded harmonic function: a function for which $f(i) = \sum_j f(j)p_{ij}$. Then $f(X)$ is a bounded martingale, hence must converge as time increases to infinity.

  The terminology supermartingale/submartingale was actually chosen to mirror the potential-theoretic terminology superharmonic/subharmonic.

- The simplest example: consider simple random walk X absorbed at boundaries $a < b$. Then $f(x) = \frac{x-a}{b-a}$ is a bounded harmonic function, and can be shown to satisfy

  $$f(x) \quad = \quad \mathbb{P}[X \text{ hits } b \text{ before } a | X_0 = x] \; .$$

  Use martingale convergence theorem and optional stopping theorem.

- Another example: given branching process Y and family size generating function $G(s)$, suppose ζ is smallest non-negative root of $\zeta = G(\zeta)$. Set $f(y) = \zeta^y$. Check this is a non-negative martingale (and therefore harmonic).

  We'd like to say, therefore $f(y) = \mathbb{P}[Y \text{ becomes extinct } | Y_0 = y]$. Since $\zeta \leq 1$, it follows f is bounded, so this follows as before.

  Further significant examples come from, for example, multidimensional random walk absorbed at boundary of a geometric region. (Relationship to "discrete Laplacian" and hence to partial differential equation theory.)

4 Counting and compensating

We can now make a connection between martingales and Markov chains. We start with the Poisson process, viewed as a process used for counting incidents, and show how martingales can be used to describe much more general counting processes.

Counting and compensating

> "It is a law of nature we overlook, that intellectual versatility is the compensation for change, danger, and trouble." H. G. Wells, *The Time Machine*, 1896

4.1 Simplest example: Poisson process

Simplest example: Poisson process

Consider birth-death-immigration process from above, with birth and death rates set to zero: $\lambda = \mu = 0$. The result is a Poisson process of rate α as described before:

This has a claim to be the simplest possible continuous-time Markov chain. Its state-space is *very* reducible, so it does not supply good examples for questions of equilibrium!

Definition 15. A continuous-time Markov chain N is a Poisson process of rate $\alpha > 0$ if the only transitions are $N \to N + 1$ of rate α.

In one approach to stochastic processes this serves as a fundamental building block for more complicated processes.

Theorem 16. *If N is Poisson process of rate α then $N_{t+s} - N_t$ is independent of past at time t and*

$$\mathbb{P}\left[N_t = k\right] \quad = \quad \mathbb{P}\left[\textit{Poisson}(\alpha t) = k\right] \quad = \quad \frac{(\alpha t)^k}{k!} e^{-\alpha t} \,.$$

Times of transitions often referred to as *incidents.*

Times between consecutive incidents are independent Exponential(α). Thence a whole wealth of distributional relationships between Exponential, Poisson, and indeed Gamma, Geometric, Hypergeometric, ...

A more general result is suggestive about how to generalize to Poisson point patterns: if $A \subset [0, \infty)$ has length measure a then

$$\mathbb{P}\left[k \text{ incidents in } A\right] \quad = \quad \mathbb{P}\left[\text{Poisson}(\alpha a) = k\right] \,.$$

A significant converse: given a random point pattern such that

$$\mathbb{P}\left[\text{No incidents in } A\right] \quad = \quad \exp(-\alpha a)$$

for any A of length measure a, the point pattern marks the incidents of a Poisson counting process of rate α.

Poisson process directions

There are ways to extend the Poisson process idea:

- view as a pattern of points:
  - Slivnyak's theorem: condition on t being a transition / incident. Then remaining incidents form transitions of Poisson process of same rate.

    Slivnyak's theorem generalizes directly to Poisson point patterns. The trick is, of course, to make sense of conditioning on an event of probability 0.
  - PASTA principle: if a Markov chain has "arrivals" following a Poisson distribution, then in statistical equilibrium *P*oisson *A*rrivals *S*ee *T*ime *A*verages.

    PASTA: That is to say, at "just before" the arrival time, the probability that the system is in state k is π_k the equilibrium probability. Easy consequence of Slivnyak's theorem.
  - How to make points "interact"?
  - Generalize to Poisson patterns of geometric objects.

    The following is crucial for calculations for Poisson patterns of geometric objects: the chance of seeing no object of given kind in given region is $\exp(-\mu)$ where μ is mean number of such objects.
- view as counting process and generalize:
  - varying "hazard rate";

    The hazard rate here is "infinitesimal chance of seeing an incident right now given that one hasn't seen anything since the last incident". For Poisson processes the times between incidents are exponentially distributed, with rate parameter α say. If the time since the last incident is u then this is $f(u)/\overline{F}(u)$ for $f(u) = \alpha\exp(-\alpha u)$ and $\overline{F}(u) = \exp(-\alpha u)$. Hence the hazard rate is $\alpha\exp(-\alpha u)/\exp(-\alpha u) = \alpha$. This suggests generalizations if the times between incidents are no longer exponentially distributed.
  - relate to martingales?

Here we follow the second direction.

4.2 Compensators

Hazard rate and compensators

Starting point: if N is Poisson process of rate α then

- ("mean") $N_t - \alpha t$ determines a martingale;

  Carry out these calculations!

  Calculation based on $\mathbb{E}\left[N_{t+s} - N_s | \mathcal{F}_s\right] = \alpha t$.

- ("variance") $(N_t - \alpha t)^2 - \alpha t$ determines a martingale;

  Calculation based on $\mathrm{Var}\left[N_{t+s} - N_s | \mathcal{F}_s\right] = \alpha t$. **HINT:** Expand $(N(t+s) - \alpha(t+s))^2 = ((N(t+s) - \alpha(t+s)) - (N(t) - \alpha t))^2 + 2((N(t+s) - \alpha(t+s)) - (N(t) - \alpha t))(N(t) - \alpha t) + (N(t) - \alpha t)^2$.

Consider processes which "count" incidents:

Later we will also briefly consider population processes counting births $+1$ and deaths -1.

Definition 17. A counting process is a continuous-time process—not necessarily Markov—changing by single jumps of $+1$.

Try to subtract something to turn it into a martingale.

Definition 18. We say $\int_0^t \ell(s)\,\mathrm{d}s$ compensates a counting process N if

- the (possibly random) $\ell(s)$ is in $\mathcal{F}_s$;
- $N_t - \int_0^t \ell(s)\,\mathrm{d}s$ determines a martingale.

It is possible to make a more general definition which replaces $\int_0^t \ell(s)\,\mathrm{d}s$ by a non-decreasing process Λ_t, but then we have to require "$\Lambda_t \in \mathcal{F}_{t-}$", and need measure theory to make sense of this.

It can then be shown that

- compensators always exist
- and are essentially unique.

Compensators generalize the notion of hazard rate.

4.3 Examples

Example: random sample of lifetimes

Suppose $X_1, \ldots, X_n$ are independent and identically distributed non-negative random variables (lifetimes) with common density f.

Note that $h(t) = f(t)/\overline{F}(t)$, where $\overline{F}(t) = 1 - F(t)$.

- Set $\mathbb{P}\left[X_i > t\right] = 1 - \int_0^t f(s)\,\mathrm{d}s = \exp\left(-\int_0^t h(s)\,\mathrm{d}s\right)$.
- Counting process $N_t = \#\{i : X_i \leq t\}$ increases by $+1$ jumps in continuous time.
- Observe:
  - $N_t - \int_0^t h(s)(n - N_s)\,\mathrm{d}s$ is a martingale.

    Resolves to showing the following is a martingale:

    $$\mathbb{I}_{[X_i \leq t]} - \int_0^{\min\{t, X_i\}} h(u)\,\mathrm{d}u\,.$$

    Key calculation: the expectation of the above is

    $$\mathbb{P}\left[X_i \leq t\right] - \int_0^t h(u)\,\mathbb{P}\left[X_i > u\right]\mathrm{d}u\,,$$

which vanishes if we substitute in $\mathbb{P}[X_i > u] = \exp\left(-\int_0^u h(s)\,\mathrm{d}s\right)$. This of course is computation of an absolute probability: Test understanding: make changes to get the relevant conditional probability calculation.

- $(N_t - \int_0^t h(s)(n - N_s)\,\mathrm{d}s)^2 - \int_0^t h(s)(n - N_s)\,\mathrm{d}s$ is a martingale.

  This follows most directly by noting independence of the $\mathbb{I}_{[X_i \leq t]} - \int_0^{\min\{t,X_i\}} h(s)\,\mathrm{d}s$. However it is actually true for a more general reason ... see later.

Example: pure birth process

Example 19 (Pure birth process). If the pure birth process N makes transitions $N \to N+1$ at rate λN then

$$N_t - \int_0^t \lambda N_s\,\mathrm{d}s \quad \text{is a martingale.}$$

A direct proof can be obtained by computing the distribution of N_t given N_0. Alternatively here is a plausibility argument: in a small period of time $[t, t+\Delta t)$ it is most likely no transition will occur; the chance of one transition is about $\lambda N_t \Delta t$, and the chance of more is infinitesimal. So the conditional mean increment is $\lambda N_t \Delta t$ which is exactly matched by the compensator.

The measure-theoretic approach to martingales makes sense of this plausibility argument, at the same time showing how it generalizes to its proper full scope.

Here again one can check that the expression of variance type $(N_t - \int_0^t \lambda N_s\,\mathrm{d}s)^2 - \int_0^t \lambda N_s\,\mathrm{d}s$ also determines a martingale.

Direct computations would permit a direct proof; but a similar plausibility argument also applies. The conditional variance of the increment is about $\lambda N_t \Delta t(1 - \lambda N_t \Delta t) \approx \lambda N_t \Delta t$, again matching the compensator.

4.4 Variance of compensated counting process

Variance of compensated counting process

The above expression of variance type holds more generally:

Theorem 20. *Suppose N is a counting process compensated by $\int \ell(s)\,\mathrm{d}s$. Then*

$$\left(N_t - \int_0^t \ell(s)\,\mathrm{d}s\right)^2 - \int_0^t \ell(s)\,\mathrm{d}s \quad \textit{is a martingale.}$$

Rigorous proof, or heuristic limiting argument ....

The key point of the rigorous proof, which we omit, is that "$\Lambda_t = \int_0^t \ell(s)\,\mathrm{d}s \in \mathcal{F}_{t-}$".

But again one can argue plausibly, starting with the comment that the increment over $(t, t+\Delta t)$ has conditional expectation $\int_t^{t+\Delta t} \ell(s)\,\mathrm{d}s$ and takes values 0 or 1. Hence we can deduce the conditional probability of a $+1$-jump as being $\int_t^{t+\Delta t} \ell(s)\,\mathrm{d}s$, and so argue as above.

4.5 Counting processes and Poisson processes

Counting processes and Poisson processes

The compensator of a counting process can be used to tell whether the counting process is Poisson:

Theorem 21. *Suppose N is a counting process which has compensator αt. Then N is a Poisson process of rate α.*

Again there is a plausibility argument: the increment over $(t, t + \Delta t)$ has conditional probability $\alpha\Delta t$, hence is approximately independent of past; hence N_t is approximately the sum of many Bernoulli random variables each of the same small mean, hence is approximately approximately Poisson ....

Better still, counting processes with compensators approximating αt are approximately Poisson of rate α. Here is a nice way to see this:

Theorem 22. *Suppose N is a counting process with compensator $\Lambda = \int \ell(s)\,\mathrm{d}s$. Consider the random time change $\tau(t) = \inf\{s : \Lambda_s = t\}$. Then the time-changed counting process $N_{\tau(t)}$ is Poisson of unit rate.*

Begs the question, is $N_{\tau(t)}$ a counting process? (Yes, but needs proof.)

There is an amazing multivariate generalization of this time-change result, related to Cox's proportional hazards model.

If the compensator approximates αt then it is immediate that $\tau(t)$ approximates t, and hence good approximation results can be derived!

The above gives a good pay-off for this theory.

Compensators and likelihoods

Here is an even bigger pay-off.

Theorem 23. *Suppose N is a counting process with compensator $\Lambda = \int \ell(s)\,\mathrm{d}s$. Then its* likelihood with respect to a unit-rate Poisson point process *over the time interval $[0, T]$ is proportional (for fixed T) to*

$$\exp\left(\int_0^T (\log \ell(t)\,\mathrm{d}N(t) - \ell(t)\,\mathrm{d}t)\right),$$

where $\int_0^T \log \ell(t)\,\mathrm{d}N(t) = \sum_{0 \leq t \leq T} \log \ell(t)\,\mathbb{I}_{[\Delta N(t)=1]}$ simply sums $\log \ell(t)$ over the times of N-incidents.

Why is this true?

Consider the case when N is Poisson of rate α. Then the required likelihood is a ratio of probabilities: if $N(T) = n$ then it equals

$$\frac{(\alpha T)^n e^{-\alpha T}/n!}{T^n e^{-T}/n!} = \exp(n \log \alpha - (\alpha - 1)T)$$

which agrees with the result stated in the theorem (note: up to a constant of proportionality namely e^T).

The case of varying intensities ℓ (time-varying, random) follows by an approximation argument.

We can use this to build likelihoods for epidemics, by viewing them as streams of incidents of different kinds.

4.6 Compensation of population processes

Compensation of population processes

The notion of compensation works for much more general processes, such as population processes:

Example 24 (Birth-death-immigration process). If the birth-death-immigration process X makes transitions $X \to X+1$ at rate $\lambda X + \alpha$ and $X \to X-1$ at rate μX then

$$X_t - \int_0^t ((\lambda - \mu) X_s + \alpha)\,\mathrm{d}s \quad \text{is a martingale.}$$

Plausibility argument much as before.

But we now need something other than the compensator to convert $(X_t - \int_0^t ((\lambda - \mu) X_s + \alpha)\,\mathrm{d}s)^2$ into a martingale.

The plausibility argument fails for the variance case! However it *is* possible to use a slightly different integral here. In fact

$$(X_t - \int_0^t ((\lambda - \mu) X_s + \alpha)\,\mathrm{d}s)^2 - \int_0^t ((\lambda + \mu) X_s + \alpha)\,\mathrm{d}s \quad \text{is a martingale.}$$

This is best understood using ideas of *stochastic integrals* (of rather simple form), which we will not explore here.

More generally a continuous-time Markov chain X relates to martingales obtained from $f(X)$ (for given functions f) by compensation using the rates of X.

This is the heart of the famous “Stroock-Varadhan martingale formulation”, which allows one to use martingales to study and to define very general Markov chains.

A multivariate version of the likelihood result above now allows us to convert specification of rates into a likelihood.

5 Central Limit Theorem

The Central Limit Theorem is one of the jewels of classical probability theory, with a huge literature developing such questions as, how may the assumptions be relaxed? and at what speed does the convergence actually occur? Before discussing this, review notions of almost sure convergence, convergence in probability, convergence in distribution, and weak convergence.

Central Limit Theorem

> "Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation" Lippmann, quoted in E. T. Whittaker and G. Robinson, *Normal Frequency Distribution.* Ch. 8 in *The Calculus of Observations: A Treatise on Numerical Mathematics*, 1967.

5.1 Classical Central Limit Theorem

The classical Central Limit Theorem

Definition 25. Random variables Y_n are said to converge in distribution to a random variable Z (or its distribution) if

$$\mathbb{P}\left[Y_n \leq y\right] \;\rightarrow\; \mathbb{P}\left[Z \leq y\right] \quad \textit{whenever } \mathbb{P}\left[Z \leq y\right] \text{is continuous at } y\,.$$

Theorem 26. *Suppose $X_1, \ldots, X_n$ are independent and identically distributed, with finite mean μ and finite variance σ^2. Then*

$$Y_n = \frac{(X_1 + \ldots + X_n) - n\mu}{\sqrt{n}\sigma} \quad \overset{\mathcal{D}}{\rightarrow} \quad N(0,1)\,,$$

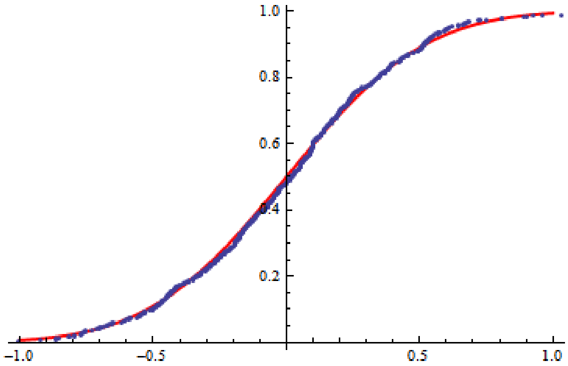
where convergence is in distribution.

$N(0,1)$ denotes a random variable with standard normal distribution.

Common notations: $Y_n \overset{d}{\rightarrow} Z$ or $Y_n \overset{\mathcal{D}}{\rightarrow} Z$ or $Y_n \Rightarrow Z$.

Cleanest proof involves *characteristic functions* $\mathbb{E}\left[\exp(iuY_n)\right]$, $\mathbb{E}\left[\exp(iuZ)\right] = e^{-\frac{1}{2}u^2}$ and hence complex numbers. A Taylor series expansion shows $\mathbb{E}\left[\exp(iuX_n)\right] \approx \exp(iu\mu)(1 - \frac{u^2}{2}\sigma^2)$; hence $\mathbb{E}\left[\exp(iuY_n)\right] \approx \left(1 - \frac{u^2}{2n}\right)^n \rightarrow e^{-u^2/2}$. Result follows from theory of characteristic function transform.

Example



Empirical CDF of 500 draws from mean of 10 independent Student t on 5 df, with limiting normal CDF graphed in red.

It is appropriate to use the CDF (cumulative distribution function) here, because that is the approximation which the CLT describes.

Note there is good agreement!

Questions arising

In this section we address the following questions:

1. Do we really need "identically distributed"?

   No we don't need exactly "identically distributed", and we can produce a useful generalization.

2. How fast does the convergence happen?

   Something really rather definite can be said about rate of convergence.

3. Do we really need "independent"?

   No we do not need exactly "independent", and we can produce a useful generalization.

In particular we can produce a satisfying answer to items 1 and 3 in terms of martingales.

(Our answers to items 1 and 3 are satisfying though not as good as possible!)

5.2 Lindeberg's Central Limit Theorem

Lindeberg's Central Limit Theorem

Strongest result about non-identically distributed case:

Theorem 27. *Suppose $X_1, \ldots, X_n$ are independent and not identically distributed, with X_i having finite mean μ_i and finite variance σ_i^2. Set $m_n = \mu_1 + \ldots + \mu_n$ and $s_n^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Suppose further that $\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\left[(X_i - \mu_i)^2\,;\,(X_i - \mu_i)^2 > \varepsilon^2 s_n^2\right] \to 0$ for every $\varepsilon > 0$. Then*

$$Y_n = \frac{X_1 + \ldots + X_n - m_n}{s_n} \quad \overset{\mathcal{D}}{\to} \quad N(0,1)\,.$$

The beauty of the Lindeberg condition is that it simply requires that relatively large components do not contribute too much to the total variance relative to the intended limit. Put this way, it is rather easy to remember the final result!

However the Lindeberg condition can be tricky to check. The Lyapunov condition is easier, and implies the Lindeberg condition: a useful special case of this condition is that the sum of the third central moments $r_n^3 = \sum_{i=1}^n \mathbb{E}\left[|X_i - \mu_i|^3\right]$ is finite and satisfies $r_n/s_n \to 0$.

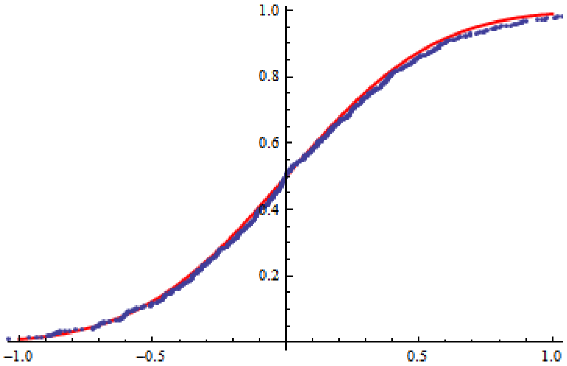
Proof is by a more careful development of the characteristic function proof of the classical Central Limit Theorem.

Very recently it has been noticed that there is a remarkable generalization to the vector-valued case. If $X_1, X_2, \ldots$ are independent zero-mean vector-valued random variables, and

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[\|X_i\|^2 \,;\, \|X_i\|^2 > \varepsilon^2 s_n^2\right] \to 0$$

where s_n^2 is the trace of the variance-covariance matrix of $X_1 + \ldots + X_n$, then $\frac{1}{s_n}(X_1 + \ldots + X_n)$ need not converge, but will get closer and closer to the corresponding sequence of matching multivariate normal distributions (Kendall and Le 2011).

Example, distributions not identical (I)

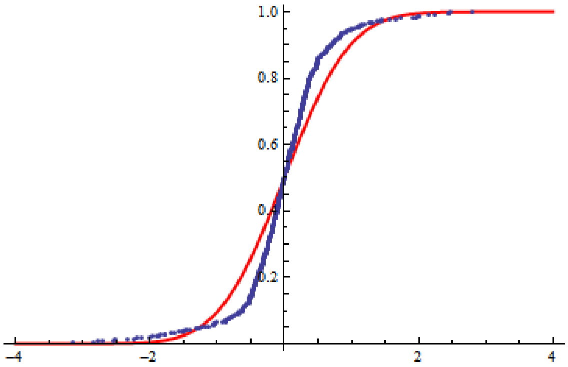


Empirical CDF of 500 draws from mean of 10 independent Student t on 5 df together with 100 draws from mean of 10 independent Student t on 3 df, with limiting normal CDF graphed in red.

Here the distributions are not all the same.

There is still reasonably good agreement!

Example, distributions not identical (II)



Empirical CDF of 500 draws from mean of 10 independent Student t on 5 df together with 100 draws from mean of 10 independent Student t on 3 df scaled by a factor of 3, with limiting normal CDF graphed in red.

Now agreement is rather poorer.

5.3 Rates of convergence

Rates of convergence

Remarkably, we can capture how fast convergence occurs if we are given some extra information about the X_i. Reverting to the classical conditions (identically distributed, finite mean and variance), using above notation, suppose $\rho^{(3)} = \mathbb{E}\left[|X_i - \mu|^3\right] < \infty$. Let $F_n(x)$ be the distribution function of $\frac{(X_1+...+X_n)-n\mu}{\sqrt{n}\sigma}$, and let $\Phi(x)$ be the standard normal distribution function. Then there is a universal constant $C > 0$ such that

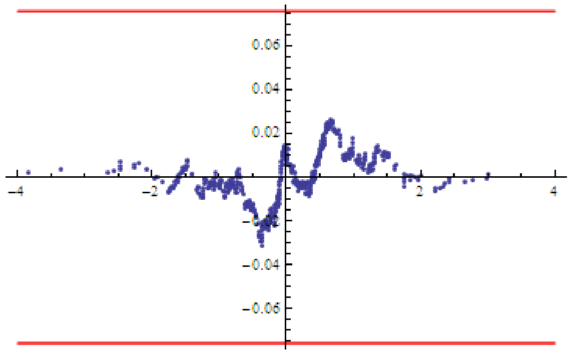
$$|F_n(x) - \Phi(x)| \quad \leq \quad \frac{C\rho^{(3)}}{\sigma^3\sqrt{n}} .$$

Note the re-appearance of the third moment condition.

There are many variants and many improvements on this result, whose proof requires much detailed mathematical analysis. For example, what is C? (Latest: we can take $C = 0.7655$.) And what can we say about the tails of the distribution?

And so forth ..., leading back to the material discussed in the Statistical Asymptotics module.

Example



Plot of difference between limiting normal CDF of empirical CDF of 500 draws from mean of 10 independent Student t on 5 df, together with upper and lower bounds.

It is apparent that the bound on CLT discrepancy is not too bad . . . at least according to this particular measure of discrepancy.

However statisticians are likely to be more interested in relative error out in the tails . . . .

5.4 Martingale case

Martingale case

There are central limit theorems for martingales, typically close in spirit to the Lindeberg theorem. Namely: the total variance needs to be nearly constant, and there must be no relatively large contributions to the variance.

Theorem 28. *Suppose $X_0 = 0$, X_1, . . . is a martingale for which $\mathbb{E}\left[X_n^2\right]$ is finite for each n. Set $s_n^2 = \mathbb{E}\left[X_n^2\right]$ and suppose $s_n^2 \to \infty$. The following two conditions taken together imply that X_n/s_n converges to a standard normal distribution:*

$$\frac{1}{s_n^2}\sum_{m=0}^{n-1}\mathbb{E}\left[|X_{m+1}-X_m|^2|\mathcal{F}_m\right] \to 1\,,$$

$$\frac{1}{s_n^2}\sum_{m=0}^{n-1}\mathbb{E}\left[|X_{m+1}-X_m|^2;|X_{m+1}-X_m|^2\geq\varepsilon^2 s_n^2\right] \to 0 \text{ for each } \varepsilon>0\,.$$

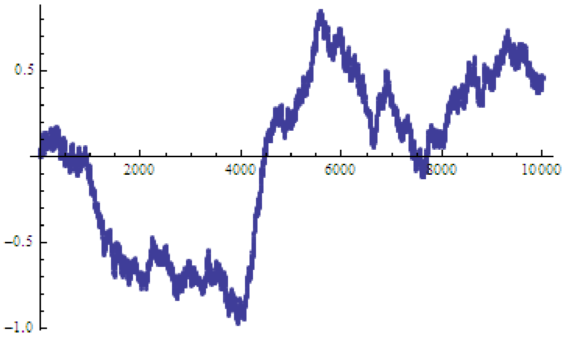
In fact $s_n^2 \to \infty$ is forced by the second (Lindeberg-type) condition.

Even more is true! the linear interpolation of the X_n, suitably rescaled, then converges to a Brownian motion.

There are *many* references, and many variations and generalizations. See for example Brown (1971). (Practical remarks about contrast between theory and practice . . . .)

Convergence to Brownian motion

Plot of $X_1/\sqrt{n}, \ldots, X_n/\sqrt{n}$ for n = 10, 100 ,1000, 10000.



Central-limit scaled (simple symmetric) random walk converges to *Brownian motion B*, characterized by independent increments, $\mathbb{E}\left[B_{t+s} - B_s\right] = 0$ (so martingale) and $\text{Var}\left[B_{t+s} - B_s\right] = t$, *continuous paths.*

If paths weren't continuous, then the compensated Poisson process would produce another example of a process with independent increments and these mean and variance properties!

In fact any random walk with jumps of zero mean and finite variance also converges to Brownian motion under central-limit scaling.

There are also similar theorems for martingales .... Classical probability deals well with central limit theorems and discrete-time martingales. If we want to deal well with continuous-time processes such as Brownian motion then *stochastic calculus* becomes very useful. From what we have said here, it should be plain that such continuous-time processes can be viewed as particular limits of discrete-time processes.

6 Recurrence

We have a theory of recurrence for discrete state space Markov chains (does $\sum_n p_{ii}^{(n)}$ diverge?). But what if the state space is not discrete? and how can we describe speed of convergence?

Recurrence

> “A bad penny always turns up” Old English proverb.

Motivation from MCMC

> Given a probability density $p(x)$of interest, for example a Bayesian posterior, we could address the question of drawing from $p(x)$ by using for example Gaussian random-walk Metropolis-Hastings.
>
> Thus proposals are normal, mean the current location x, fixed variance-covariance matrix.
>
> Using the Hastings ratio to accept/reject proposals, we end up with a Markov chain X which has transition mechanism which mixes a density with staying at the start-point.
>
> Evidently the chain almost surely *never* visits specified points other than its starting point. Thus it can never be irreducible in the classical sense, and the discrete-chain theory cannot apply ....

Clearly the discrete-chain theory needs major rehabilitation if it is to be helpful in the continuous state space case!

Recurrence

We already know, if X is a Markov chain on a discrete state-space then its transition probabilities converge to a unique limiting equilibrium distribution if:

1. X is irreducible;

   the state space of X cannot be divided into substantial regions some of which are inaccessible from others;

2. X is aperiodic;

   the state space of X cannot be broken into periodic cycles;

3. X is positive-recurrent.

   the mean time for X to return to its starting point is finite.

How in general can one be quantitative about the speed at which convergence to equilibrium can occur? and what if the state-space is not discrete?

6.1 Speed of convergence

Measuring speed of convergence to equilibrium (I) Total variation distance

- Speed of convergence of a Markov chain X to equilibrium can be measured as discrepancy between two probability measures: $\mathcal{L}\left(X_t|X_0 = x\right)$ (distribution of X_t) and π (equilibrium measure).

  $\mathcal{L}\left(X_t|X_0 = x\right)(A)$ is probability that X_t belongs to A.

- Simple possibility: total variation distance. Let $\mathcal{X}$ be state-space, for $A \subseteq \mathcal{X}$ maximize discrepancy between $\mathcal{L}\left(X_t|X_0 = x\right)(A) = \mathbb{P}\left[X_t \in A|X_0 = x\right]$ and $\pi(A)$:

$$\text{dist}_{\text{TV}}(\mathcal{L}\left(X_t|X_0 = x\right), \pi) \quad = \quad \sup_{A\subseteq\mathcal{X}}\{\mathbb{P}\left[X_t \in A|X_0 = x\right] - \pi(A)\} .$$

  Test understanding: why is it not necessary to consider $|\,\mathbb{P}\left[X_t \in A|X_0 = x\right] - \pi(A)|$? (Hint: consider $\mathbb{P}\left[X_t \in A^c|X_0 = x\right] - \pi(A^c)$.)

- Alternative expression in case of discrete state-space:

$$\text{dist}_{\text{TV}}(\mathcal{L}\left(X_t|X_0 = x\right), \pi) \quad = \quad \frac{1}{2}\sum_{y\in\text{X}}|\,\mathbb{P}\left[X_t = y|X_0 = x\right] - \pi_y| .$$

  Test understanding: prove this by considering $A = \{y : \mathbb{P}\left[X_t = y|X_0 = x\right] > \pi_y\}$.

  (*Many* other possible measures of distance ....)

It is not even clear that total variation is best notion: in the case of MCMC one might consider a spectral approach (which we will pick up again when we come to consider cutoff):

$$\sup_{f:\int|f(x)|^2\pi(\mathrm{d}x)<\infty}\left(\mathbb{E}\left[f(X_t)|X_0 = x\right] - \int f(x)\pi(\mathrm{d}x)\right)^2 .$$

Nevertheless the concept of total variation isolates a desirable kind of rapid convergence.

Measuring speed of convergence to equilibrium (II) Uniform ergodicity

Definition 29. The Markov chain X is uniformly ergodic if its distribution converges to equilibrium in total variation *uniformly in the starting point* $X_0 = x$: for some fixed $C > 0$ and for fixed $\gamma \in (0,1)$,

$$\sup_{x\in\mathcal{X}}\text{dist}_{\text{TV}}(\mathcal{L}\left(X_n|X_0 = x\right), \pi) \quad \leq \quad C\gamma^n .$$

Any finite ergodic Markov chain is automatically uniformly ergodic.
In fact this is a consequence of the apparently weaker assertion, as $n \to \infty$ so

$$\sup_{x\in\mathcal{X}}\text{dist}_{\text{TV}}(\mathcal{L}\left(X_t|X_0 = x\right), \pi) \quad \to \quad 0 .$$

In theoretical terms, for example when carrying out MCMC, this is a very satisfactory property. No account need be taken of the starting point, and accuracy improves in proportion to the length of the simulation.

Much depends on size of C and on how small is γ.

Typically theoretical estimates of C and γ are *very* conservative.

Uniform ergodicity is tantamount to "boundedness" for one's Markov chain.

Other things being equal(!), given a choice, consider choosing a uniformly ergodic Markov chain for your MCMC algorithm.

Measuring speed of convergence to equilibrium (III) Geometric ergodicity

Definition 30. The Markov chain X is geometrically ergodic if its distribution converges to equilibrium in total variation for some $C(x) > 0$ *depending on the starting point* x and for fixed $\gamma \in (0, 1)$,

$$\text{dist}_{\text{TV}}(\mathcal{L}\,(X_t | X_0 = x)\,, \pi) \quad \leq \quad C(x)\gamma^n\,.$$

Here account does need to be taken of the starting point, but still accuracy improves in proportion to the length of the simulation.

A significant question is, how might one get a sense of whether a specified chain *is* indeed geometrically ergodic (because at least that indicates the rate at which the distribution of X_t gets closer to equilibrium) and how one might obtain upper bounds on γ.

We shall see later on that even given good information about γ and C, and even if total variation is of primary interest, geometric ergodicity still leaves important phenomona untouched!

6.2 Irreducibility for general chains

ϕ-irreducibility (I)

We make two observations about Markov chain irreducibility:

We are skating over the issue of periodicity, which is largely technical.

1. The discrete theory fails to apply directly even to well-behaved chains on non-discrete state-space.

   Consider the Gaussian random walk X (jumps have standard normal distribution): if $X_0 = 0$ then we can assert that with probability one X *never* returns to its starting point.

2. Suppose ϕ is a measure on the state-space: then we could ask for the chain to be irreducible *on sets of positive ϕ measure.*

   "measure": like a probability measure, but not necessarily of finite total mass. Think of length, area, or volume as examples. Also, counting measure.

   Definition 31. The Markov chain X is *ϕ-irreducible* if for any state x and for any subset B of state-space of positive ϕ-measure $\phi(B) > 0$ we find that X has positive chance of reaching B if begun at x.

   The Gaussian random walk is Lebesgue-measure-irreducible! (Here Lebesgue measure is just length measure.)

ϕ-irreducibility (II)

1. We call ϕ an *irreducibility measure.* It is possible to modify ϕ to construct a *maximal irreducibility measure* ψ; one such that any set B of positive measure under some irreducibility measure for X is of positive measure for ψ.

   Lebesgue measure is a maximal irreducibility measure for the Gaussian random walk.

2. Irreducible chains on countable state-space are c-irreducible where c is counting measure ($c(A) = |A|$).

   So ϕ-irreducibility simply generalizes the original notion of irreducibility.

3. If a chain has unique equilibrium measure π then π will serve as a maximal irreducibility measure.

   Note that ϕ can be replaced by any other measure which is "measure-equivalent" (has the same null-sets). So while π will serve as a maximal irreducibility measure, we can use any alternative measure which has the same sets of measure zero.

6.3 Regeneration and small sets

Regeneration and small sets (I)

The discrete-state-space theory works because (a) the Markov chain *regenerates* each time it visits individual states, and (b) it has a positive chance of visiting specified individual states.

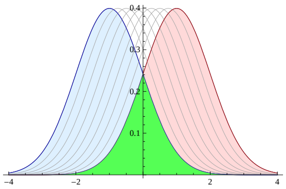
In effect this reduces the theory of convergence to equilibrium to a chapter in the theory of renewal processes, with renewals occurring each time the chain visits a specified state.

So it is natural to consider regeneration when visiting sets.

Definition 32. A set E of ϕ-positive measure is a small set of lag k for X if there is $\alpha \in (0, 1)$ and a probability measure ν such that for all $x \in E$ the following minorization condition is satisfied

$$\mathbb{P}\left[X_k \in A | X_0 = x\right] \quad \geq \quad \alpha\nu(A) \qquad \text{for all } A\,.$$

In effect, if we *sub-sample* X every k time-steps then, every time it visits E, there is a chance α that X forgets its entire past and starts again, using probability measure ν. Consider the Gaussian random walk described above. *Any* bounded set is small of lag 1.

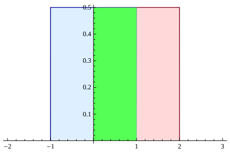


In general α can be very small—reducing practical impact, but still helping theoretically.

Regeneration and small sets (II)

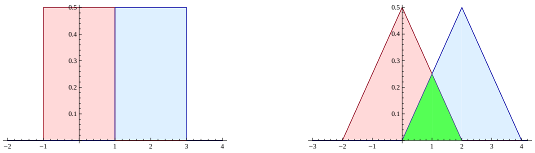
Let X be a RW with transition density $p(x, \mathrm{d}y) = \frac{1}{2}\mathbb{I}_{[|x-y|<1]}$.

Consider the set $[0, 1]$: this is small of lag 1, with $\alpha = 1/2$ and ν the uniform distribution on $[0, 1]$:



This can be seen by looking at the common overlap of the transition densities from all points $x \in [0, 1]$. This overlap is shaded here in green.

The set $[0, 2]$ is *not* small of **lag 1**, but *is* small of **lag 2**.



However, the common overlap of all one-step transition kernels from $x \in [0, 2]$ is the empty set, and so $[0, 2]$ is not a small set of lag 1. If we look at the two-step transition kernels however (the triangular kernels on the right), then there *is* a common overlap: now $\alpha = 1/4$ (the area of the green triangle) and ν is the triangular density supported on $[0, 2]$.

Regeneration and small sets (III)

Small sets would not be very interesting except that:

1. all ϕ-irreducible Markov chains X possess small sets;

   This is a very old result: see Nummelin (1984) for a recent treatment.

2. consider chains X with continuous transition density kernels. They possess *many* small sets of lag 1;

   Exercise: try seeing why this is obviously true!

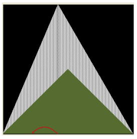
3. consider chains X with measurable transition density kernels. They need possess *no* small sets of lag 1, but will possess many sets of lag 2;

   Kendall and Montana (2002): so measurable transition density kernels lead to chains which possess latent discretizations.

4. given just one small set, X can be represented using a chain which has a single recurrent atom.

   “Split-chain construction” (Athreya and Ney 1978; Nummelin 1978).

   In a word, small sets *discretize* Markov chains.



ANIMATION

6.4 Harris-recurrence

Harris-recurrence

Now it is evident what we should mean by recurrence for non-discrete state spaces. Suppose X is ϕ-irreducible and ϕ is a maximal irreducibility measure.

Definition 33. X is (ϕ-)recurrent if, for ϕ-*almost* all starting points x and any subset B with $\phi(B) > 0$, when started at x the chain X is almost sure eventually to hit B.

So the irreducibility measure is used to focus attention on sets rather than points.

Definition 34. X is Harris-recurrent if we can drop "ϕ-almost" in the above.

And in fact we don't even then need ϕ to be *maximal*.

6.5 Examples

Examples of ϕ-irreducibility

- Random walks with continuous jump densities. And in fact measurable jump densities suffice.

  Convolutions of measurable densities are continuous!

- Chains with continuous or even measurable transition densities with exception that chain may stay put.

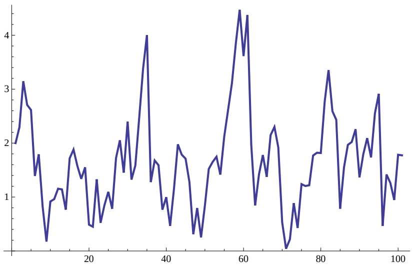
  Many examples of Metropolis-Hastings samplers.

- Vervaat perpetuities:

  $$X_{n+1} \quad = \quad U_{n+1}^{\alpha}(X_n + 1)$$

  where $U_1, U_2, \ldots$ are independent Uniform$(0,1)$.

  Test understanding: find a small set for the Vervaat perpetuity example (a simulation of which is graphed below)!



- Volatility models:

  $$\begin{aligned} X_{n+1} &= X_n + \sigma_n Z_{n+1} \\ \sigma_{n+1} &= f(\sigma_n, U_{n+1}) \end{aligned}$$

  for suitable f, and independent Gaussian Z_{n+1}, U_{n+1}.

7 Foster-Lyapunov criteria

Geometric and uniform ergodicity make sense for general Markov chains: how to find out whether they hold? and how to find out whether equilibrium distributions exist? We want simple criteria, and we can capture these using the language of martingales. Lyapunov, an account from Wikipedia:

> His student and collaborator, Vladimir Steklov, recalled his first lecture in the following way: "A handsome young man, almost of the age of the other students, came before the audience, where there was also the old Dean, Professor Levakovsky, who was respected by all students. After the Dean had left, the young man with a trembled voice started to lecture a course on the dynamics of material points, instead of a course on dynamical systems. This subject was already known to the students from the lectures of professor Delarue. But what Lyapunov taught us was new to me and I had never seen this material in any textbook. All antipathy to the course was immediately blown to dust. From that day students would show Lyapunov a special respect."

Foster-Lyapunov criteria

> "Even for the physicist the description in plain language will be the criterion of the degree of understanding that has been reached." Werner Heisenberg, *Physics and philosophy: The revolution in modern science*, 1958

7.1 Renewal and regeneration

Renewal and regeneration

Suppose C is a small set for ϕ-recurrent X, with lag 1:

If lag is $k > 1$ then sub-sample every k steps!

$$\mathbb{P}\left[X_1 \in A | X_0 = x \in C\right] \quad \geq \quad \alpha\nu(A)\,.$$

Identify *regeneration events*:

This is a coupling construction, linked to the split-chain construation (Athreya and Ney 1978; Nummelin 1978) and the Murdoch and Green (1998) approach to CFTP.

X regenerates at $x \in C$ with probability α and then makes transition with distribution ν; otherwise it makes transition with distribution $\frac{p(x,\cdot)-\alpha\nu(\cdot)}{1-\alpha}$.

This is just the appropriate compensating distribution

$$\frac{p(x,\cdot) - \alpha\nu(\cdot)}{p(x,X) - \alpha\nu(X)} \quad = \quad \frac{p(x,\cdot) - \alpha\nu(\cdot)}{1-\alpha}\,.$$

Test understanding: check that this really *is* a probability distribution!

The regeneration events occur as a *renewal sequence.* Set

$$p_k \;=\; \mathbb{P}\left[\text{next regeneration at time } k \mid \text{regeneration at time } 0\right]\,.$$

If the renewal sequence is *non-defective* (if $\sum_k p_k = 1$)

Non-defective: So there will always be a next regeneration.

and *positive-recurrent* (if $\sum_k kp_k < \infty$)

Positive-recurrent: So mean time to next regeneration is finite.

then there exists a stationary version. This is the key to equilibrium theory whether for discrete or continuous state-space.

Richard Tweedie at WRASS 1998: “continuous is no harder than discrete!”

7.2 Positive recurrence

Positive recurrence

The Foster-Lyapunov criterion for positive recurrence of a ϕ-irreducible Markov chain X on a state-space $\mathcal{X}$:

There is a delicate balance between all these conditions on Λ and C. Each one is absolutely essential!

In words, we can find a non-negative $\Lambda(X)$ such that $\Lambda(X_n) + an$ determines a supermartingale until $\Lambda(X)$ becomes small enough for X to belong to a small set!

Theorem 35 (Foster-Lyapunov criterion for positive recurrence)**.** *Given* $\Lambda : \mathcal{X} \to [0,\infty)$*, positive constants* a*,* b*,* c*, and a small set* $C = \{x : \Lambda(x) \leq c\} \subseteq \mathcal{X}$ *with*

$$\mathbb{E}\left[\Lambda(X_{n+1})|\mathcal{F}_n\right] \quad \leq \quad \Lambda(X_n) - a + b\,\mathbb{I}_{[X_n \in C]}\,;$$

then $\mathbb{E}\left[T_A|X_0 = x\right] < \infty$ *for any* A *with* $\phi(A) > 0$*, where* $T_A = \inf\{n \geq 0 : X_n \in A\}$ *is the time when* X *first hits* A*, and moreover* X *has an equilibrium distribution.*

We can re-scale Λ so that $a = 1$.

In fact if the criterion holds then it can be shown, *any* sub-level set of Λ is small.

It is evident from the verbal description that reflected simple asymmetric random walk (negatively biased) is an example for which the criterion applies.

Sketch of proof

Supplementary:

1. $Y_n = \Lambda(X_n) + an$ is non-negative supermartingale up to time $T = \inf\{m \geq 0 : X_m \in C\} > n$:

   $$\mathbb{E}\left[Y_{\min\{n+1,T\}}|\mathcal{F}_n, T > n\right] \quad \leq \quad (\Lambda(X_n) - a) + a(n+1) \;=\; Y_n\,.$$

   Hence $Y_{\min\{n,T\}}$ converges.

2. So $\mathbb{P}\left[T < \infty\right] = 1$ (otherwise $\Lambda(X) > c$, $c + an < Y_n$ so $Y_n \to \infty$). Moreover $\mathbb{E}\left[Y_T|X_0\right] \leq \Lambda(X_0)$ (Fatou argument) so $a\,\mathbb{E}\left[T\right] \leq \Lambda(X_0)$.

3. Now use finiteness of b to show $\mathbb{E}\left[T^*|X_0\right] < \infty$, where T^* first regeneration in C.

4. ϕ-irreducibility: positive chance of hitting A before first regeneration in C. Hence $\mathbb{E}\left[T_A|X_0\right] < \infty$.

There is a stationary version of the renewal process of successive regenerations on C.

One can construct a “bridge” of X conditioned to regenerate on C at time 0, and then to regenerate again on C at time n.

Hence one can sew these together to form a stationary version of X, which therefore has the property that X_t has the equilibrium distribution for all time t.

A converse ...

Suppose on the other hand that $\mathbb{E}\left[T|X_0\right] < \infty$ for all starting points X_0, where C is some small set and T is the first time for X to return to C.

ϕ-irreducibility then follows automatically.

The Foster-Lyapunov criterion for positive recurrence follows for $\Lambda(x) = \mathbb{E}\left[T|X_0 = x\right]$ if $\mathbb{E}\left[T|X_0\right]$ is bounded on C.

Indeed, (supposing lag 1 for simplicity)

$$\mathbb{E}\left[\Lambda(X_{n+1})|\mathcal{F}_n\right] \quad \leq \quad \Lambda(X_n) - 1 + b\,\mathbb{I}_{[X_n \in C]}\,,$$

where b is the mean value of $\mathbb{E}\left[Y_T|x\right]$ if x is chosen using the regeneration probability measure for C.

Moreover if the renewal process of successive regenerations on C is aperiodic then a coupling argument shows general X will converge to equilibrium.

If the renewal process of successive regenerations on C is not aperiodic then one can sub-sample ....

Showing that X has an equilibrium is then a matter of probabilistic constructions using the renewal process of successive regenerations on C.

7.3 Geometric ergodicity

Geometric ergodicity

The Foster-Lyapunov criterion for geometric ergodicity of a ϕ-irreducible Markov chain X on a state-space $\mathcal{X}$:

In words, we can find a $\Lambda(X) \geq 1$ such that $\Lambda(X_n)/\gamma^n$ determines a supermartingale until $\Lambda(X)$ becomes small enough for X to belong to a small set!

Theorem 36 (Foster-Lyapunov criterion for geometric ergodicity)**.** *Given $\Lambda : \mathcal{X} \to [1, \infty)$, positive constants $\gamma \in (0,1)$, b, $c \geq 1$, and a small set $C = \{x : \Lambda(x) \leq c\} \subseteq \mathcal{X}$ with*

$$\mathbb{E}\left[\Lambda(X_{n+1})|\mathcal{F}_n\right] \quad \leq \quad \gamma\Lambda(X_n) + b\,\mathbb{I}_{[X_n \in C]}\,;$$

then $\mathbb{E}\left[\gamma^{-T_A}|X_0 = x\right] < \infty$ for any A with $\phi(A) > 0$, where $T_A = \inf\{n \geq 0 : X_n \in A\}$ is the time when X first hits A, and moreover (under suitable periodicity conditions) X is geometrically ergodic.

We can rescale Λ so that $b = 1$.

The criterion for positive-recurrence is implied by this criterion.

We can enlarge C and alter b so that the criterion holds simultaneously for all $\mathbb{E}\left[\Lambda(X_{n+m})|\mathcal{F}_n\right]$.

Sketch of proof

1. $Y_n = \Lambda(X_n)/\gamma^n$ defines non-negative supermartingale up to time T when X first hits C:

$$\mathbb{E}\left[Y_{\min\{n+1,T\}}|\mathcal{F}_n, T > n\right] \quad \leq \quad \gamma \times \Lambda(X_n)/\gamma^{n+1} \quad = \quad Y_n\,.$$

Hence $Y_{\min\{n,T\}}$ converges.

2. $\mathbb{P}[T < \infty] = 1$, for otherwise $\Lambda(X) > c$ and so $Y_n > c/\gamma^n$ does not converge. Moreover $\mathbb{E}[\gamma^{-T}] \leq \Lambda(X_0)$.

3. Finiteness of b shows $\mathbb{E}\left[\gamma^{-T^*}|X_0\right] < \infty$, where T^* is time of regeneration in C.

4. From ϕ-irreducibility there is positive chance of hitting A before regeneration in C. Hence $\mathbb{E}[\gamma^{-T_A}|X_0] < \infty$.

   Geometric ergodicity follows by a coupling argument which I do not specify here.

   The constant γ here provides an upper bound on the constant γ used in the definition of geometric ergodicity. *However* it is not necessarily a very good bound!

Two converses

This was used in Kendall 2004 to provide perfect simulation *in principle*. The Markov inequality can be used to convert the condition on $\Lambda(X)$ into the existence of a Markov chain on $[0,\infty)$ whose exponential dominates $\Lambda(X)$. The chain in question turns out to be a kind of queue (in fact, $D/M/1$). For $\gamma \geq e^{-1}$ the queue will not be recurrent; however one can sub-sample X to convert the situation into one in which the dominating queue will be positive-recurrent. In particular, geometric ergodicity forces a useful partial ordering on the state-space.

1. Suppose on the other hand that $\mathbb{E}[\gamma^{-T}|X_0] < \infty$ for all starting points X_0 (and fixed $\gamma \in (0,1)$), where C is some small set and T is the first time for X to return to C. The Foster-Lyapunov criterion for geometric ergodicity then follows for $\Lambda(x) = \mathbb{E}[\gamma^{-T}|X_0 = x]$ if $\mathbb{E}[\gamma^{-T}|X_0]$ is bounded on C.

   Uniform ergodicity follows if the Λ function is bounded above.

   But more is true. Strikingly,

2. For Harris-recurrent Markov chains the existence of a geometric Foster-Lyapunov condition is *equivalent* to the property of geometric ergodicity.

7.4 Examples

Examples

It is instructive to notice that the criteria continue to apply to a considerable variety of appropriately modified Markov chains.

1. General reflected random walk: $X_{n+1} = \max\{X_n + Z_{n+1}, 0\}$ with independent Z_{n+1} of continuous density $f(z)$, $\mathbb{E}[Z_{n+1}] < 0$, $\mathbb{P}[Z_{n+1} > 0] > 0$. Then

   (a) X is Lebesgue-irreducible on $[0,\infty)$;

   (b) Foster-Lyapunov criterion for positive recurrence applies.

Similar considerations often apply to Metropolis-Hastings Markov chains based on random walks.

(a) $\mathbb{E}[Z_{n+1}] < 0$ so by SLLN $\frac{1}{n}(Z_1 + \ldots + X_n) \to -\infty$, so X hits 0 for any X_0. $\mathbb{P}[Z_{n+1} > 0] > 0$ so $f(z) > 0$ for $a < z < a(1 + \frac{1}{m})$, some a, $m > 0$. So if $X_0 = 0$ then density of X_n is positive on $(na, na + \frac{n}{m}a)$. If $A \subset (ma, \infty)$ is of positive measure then one of $A \cap (na, na + \frac{n}{m}a)$ $(n \geq m)$ is of positive measure so $\mathbb{P}[X \text{ hits } A | X_0 = 0] > 0$. $\mathbb{E}[Z_{n+1}] < 0$ so $f(z) > 0$ for $-b - \frac{1}{k} < z < -b$, some b, $k > 0$. Start X at some x in $(nb - \frac{1}{k}, nb)$ (positive chance of hitting this interval if $nb - \frac{1}{k} > ma$). Then X_n has positive density over $(\max\{0, x - nb\}, x - nb + \frac{n}{m})$ which includes $(0, \frac{n-1}{k})$. By choosing n large enough, we now see we can get anywhere.

(b) Test understanding: Check Foster-Lyapunov criterion for positive recurrence for $\Lambda(x) = x$.

2. Reflected Simple Asymmetric Random Walk: $X_{n+1} = \max\{X_n + Z_{n+1}, 0\}$ with independent Z_{n+1} such that $\mathbb{P}[Z_{n+1} = -1] = q = 1 - p = 1 - \mathbb{P}[Z_{n+1} = +1] > \frac{1}{2}$.

   (a) X is counting-measure-irreducible on non-negative integers;

   (b) Foster-Lyapunov criterion for geometric ergodicity applies.

   Aim for $\mathbb{E}\left[e^{aZ_{n+1}}\right] < 1$ for some positive a.

(a) Test understanding: this is the same as ordinary irreducibility for discrete-state-space Markov chains!

(b) Test understanding: Check Foster-Lyapunov criterion for geometric ergodicity for $\Lambda(x) = e^{ax}$ for small positive a.

Reflected Simple asymmetric random walk (II)

- Positive recurrence criterion: check for $\Lambda(x) = x$, $C = \{0\}$:

$$\mathbb{E}[\Lambda(X_1) | X_0 = x_0] = \begin{cases} \Lambda(x_0) - (q - p) & \text{if } x_0 \notin C\,, \\ 0 + p & \text{if } x_0 \in C\,. \end{cases}$$

- Geometric ergodicity criterion: check for $\Lambda = e^{ax}$, $C = \{0\} = \Lambda^{-1}(\{1\})$:

$$\mathbb{E}[\Lambda(X_1) | X_0 = x_0] = \begin{cases} \Lambda(x_0) \times (pe^a + qe^{-a}) & \text{if } x_0 \notin C\,, \\ 1 \times (p + qe^{-a}) & \text{if } x_0 \in C\,. \end{cases}$$

This works when $pe^a + qe^{-a} < 1$; equivalently when $0 < a < \log(q/p)$ (solve the quadratic in e^a!).

One may ask, does this kind of argument show that *all* positive-recurrent random walks can be shown to be geometrically ergodic simply by moving from $\Lambda(x) = x$ to $\Lambda(x) = e^{ax}$? The answer is no, essentially because there exist random walks whose jump distributions have negative mean but fail to have exponential moments ....

8 Cutoff

In what way does a Markov chain converge to equilibrium? Is it a gentle exponential process? Or might most of the convergence happen relatively quickly? Once again we focus on reversible Markov chains, as these make computations simpler.

Cutoff

> "I have this theory of convergence, that good things always happen with bad things."
> Cameron Crowe, *Say Anything* film, 1989

8.1 The cutoff phenomenon

Convergence: cutoff or geometric decay?

What we have so far said about convergence to equilibrium will have left the misleading impression that the distance from equilibrium for a Markov chain is characterized by a gentle and rather geometric decay. It is true that this is typically the case after an extremely long time, and it can be the case over all time. However it is entirely possible for "most" of the convergence to happen quite suddenly at a specific threshold.

Random walk wrapped around a circle exhibits a gentle and rather geometric decay. Famously (Bayer and Diaconis 1992) the riffle shuffle does not! (For a pack of 52 cards, 7 shuffles suffice for essentially all practical purposes. Compare this to the commonly-used overhand shuffle, which takes > 1000 shuffles to randomize a deck of 52 cards! (Pemantle 1989).)

The theory for this is developing fast, but many questions remain open. In this section we describe a specific easy example.

8.2 Cutoff and eigenvalues

Cutoff (I): Markov chains and matrices

We need to understand something about eigenvalues for Markov chains.

Finite-state-space reversible Markov chains and (weighted) euclidean spaces.

Fix attention on a finite state space $\mathcal{X}$, with reversible aperiodic Markov chain of transition kernel $p_{x,y}$ and equilibrium distribution π.

The vector space of functions on $\mathcal{X}$ can be given a weighted Euclidean norm:

$$\|f\|_\pi^2 \quad = \quad \sum_{x\in\mathcal{X}} |f(x)|^2\pi(x)$$

and hence an inner product $\langle f,g\rangle_\pi$.

$$\langle f,g\rangle_\pi \quad = \quad \sum_y f(y)g(y)\pi(y)\,.$$

View transition kernel as linear operator $Pf(x) = \sum_y p_{x,y}f(y)$: by reversibility this is $\langle\cdot,\cdot\rangle_\pi$ symmetric.

Test understanding: use detailed balance to show

$$\langle f, Pg\rangle_\pi \quad = \quad \sum_x f(x)\sum_y p_{x,y} g(y)\pi(x) \quad = \quad \langle Pf, g\rangle_\pi$$

Adam Willis (MMORSE student at Warwick, 2004-2008) recently wrote an excellent Integrated Masters project on this subject.

The vector space of functions on a finite state space is finite-dimensional!

Cutoff (II): eigenvalues and eigenfunctions

So P can be viewed as a symmetric matrix and thus has a full set of eigenvalues $-1 \leq \lambda_k \leq \ldots \leq \lambda_1 \leq 1$ (if $\mathcal{X}$ has k elements) and corresponding normalized eigenfunctions $V_1, \ldots, V_k$.

Normalized: $\|V_i\|^2_\pi = 1$; eigen property: $PV_i = \lambda_i V_i$.

Because of symmetry of P we may take the V_i to be an orthonormal basis, so

$$\sum |f(y)|^2\pi(y) \quad = \quad \sum_{i=1}^k \langle f, V_i\rangle^2_\pi \,.$$

The law of total probability implies $\lambda_1 = 1$ and $V_1 \equiv 1$, and irreducibility implies $\lambda_2 < \lambda_1$.

In fact all eigenvalues cannot exceed 1 in absolute value, by an inequality argument. Two eigenvalues equal to 1 would allow us to split state space into 2 components which violates irreducibility.

Aperiodicity implies $-1 < \lambda_k$.

In passing, there is a useful analysis of rate of convergence of *expectations* of functions of Markov chains based on this spectral analysis. Good when you know *a priori* what you want to estimate ....

8.3 Two metrics

Cutoff (III): metrics

We need to relate total variation distance to the weighted Euclidean distance. Recall

$$\text{dist}_{\text{TV}}(P_x^{(n)}, \pi) \;=\; \frac{1}{2}\sum_y |P_x^{(n)}(y) - \pi(y)| \;=\; \frac{1}{2}\sum_y \left|\tfrac{P_x^{(n)}(y)}{\pi(y)} - 1\right|\pi(y)\,.$$

But this relates to weighted Euclidean distance by using the Cauchy-Schwartz inequality and $\sum_y \pi(y) = 1$:

$$2\,\text{dist}_{\text{TV}}(P_x^{(n)}, \pi) \;\leq\; \sqrt{\|\tfrac{P_x^{(n)}(\cdot)}{\pi(\cdot)} - 1\|^2_\pi}\sqrt{\sum_y \pi(y)} \;=\; \sqrt{\|\tfrac{P_x^{(n)}(\cdot)}{\pi(\cdot)} - 1\|^2_\pi}\,.$$

Now expand using orthonormal eigenfunctions and $V_1 \equiv 1$:

$$\|\tfrac{P_x^{(n)}(\cdot)}{\pi(\cdot)} - 1\|^2_\pi \;=\; \sum_{i=2}^k \langle \tfrac{P_x^{(n)}(\cdot)}{\pi(\cdot)}, V_i\rangle^2_\pi \;=\; \sum_{i=2}^k (P_x^n V_i)^2 \;=\; \sum_{i=2}^k \lambda_i^{2n} V_i(x)^2\,.$$

The key here is the Cauchy-Schwartz inequality:

$$(\mathbb{E}[XY])^2 \quad \leq \quad \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right].$$

Applied probabilists and statisticians may be more comfortable with this if they recognize that it is proved in the same way as the statement that correlations are always bounded between ±1.

Miss $i = 1$ since $V_1 \equiv 1$, so

$$\langle \frac{P_x^{(n)}(\cdot)}{\pi(\cdot)} - 1, V_1 \rangle_\pi \quad = \quad \sum_y P_x^{(n)}(y) - \langle V_1, V_1 \rangle_\pi \quad = \quad 1 - 1 \; = \; 0\,.$$

Miss -1 in other terms by orthogonality, since for $i > 1$

$$\langle -1, V_i \rangle_\pi \quad = \quad -\langle V_1, V_i \rangle_\pi \quad = \quad 0\,.$$

Bear in mind that in this finite-state-space context eigenfunctions are the same as eigenvectors!

8.4 A special case

Cutoff (IV): upper bound in special case Gibbs' sampler for zero-interaction Ising model

Model for Gibb's sampler. Consider $N \times N$ array of ±1. At each step choose entry at random, flip sign.

As above, identify $\binom{N^2}{r}$ eigenfunctions of eigenvalue $1 - \frac{2r}{N^2}$, for $0 \leq r \leq N^2$. Set $n = \frac{N^2}{4}(\log(N^2) + \theta)$.

$$\begin{aligned} \| \tfrac{P_x^{(n)}(\cdot)}{\pi(\cdot)} - 1 \|_\pi^2 &= \sum_{r=1}^{N^2} \binom{N^2}{r} \left(1 - \tfrac{2r}{N^2}\right)^{2n} \\ &\leq \sum_{r=1}^{N^2} \binom{N^2}{r} \exp\left(-\tfrac{2r}{N^2}(\tfrac{N^2}{2}(\log(N^2) + \theta))\right) \\ &= \sum_{r=1}^{N^2} \binom{N^2}{r} (N^2)^{-r} e^{-r\theta} \;\leq\; \sum_{r=1}^{N^2} \tfrac{1}{r!} e^{-r\theta} \;\leq\; \exp(e^{-\theta}) - 1\,. \end{aligned}$$

Eigenfunctions are just products $X_{i_1} \ldots X_{i_k}$ of spin variables $X_r = \pm 1$. Test understanding: check this! In particular, note $PX_1 = \frac{1}{N^2}(-X_1) + (1 - \frac{1}{N^2})X_1 = (1 - \frac{2}{N^2})X_1, \ldots$.

Note, $1 - x \leq e^{-x}$ always.

Cutoff (V): lower bound in special case

The upper bound suggests a cutoff:

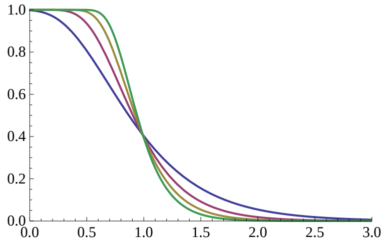
$$\mathrm{dist}_{\mathrm{TV}}(P_x^{(n)}, \pi) \quad \leq \quad \frac{1}{2}\sqrt{\exp(e^{-\theta}) - 1}$$

Since $n = \frac{N^2}{4}(\log(N^2) + \theta)$, the cutoff occurs at around $\frac{N^2}{4}\log(N^2)$ and lasts of order $\frac{N^2}{4}$.

However to make *sure* this works, we also need a lower bound on $\text{dist}_{\text{TV}}(P_x^{(n)}, \pi))$. Achieve this by comparing means and variances of $Z = \sum_{i=1}^{N^2} X_i$, where X_i is spin at site i. Simple estimates confirm that there is still substantial total variation distance at $\frac{N^2}{4}\log(N^2)$, so this is a real cutoff.

At any fixed time Z has a (scaled and shifted) Binomial distribution, and π is also of this form. We can then use Markov's inequality to convert mean and variance comparisons into inequalities.

Scaling the x-axis by the cutoff time, we see that the total variation distance drops more and more rapidly towards zero as N becomes larger: the curves in the graph below tend to a step function as $N \to \infty$.



Moral: *effective* convergence can be much faster than one realizes, and occur over a fairly well defined period of time.

The graph shows $\text{dist}_{\text{TV}}(P_x^{(cn)}, \pi)$ for $c \in (0, 3)$, for four increasing values of N, for the (closely related) simple random walk on $\mathbb{Z}_2^N$.

Calculations for other cases can be much harder, but cutoffs are known to occur for a large number of *random walks on groups*. These include a number of card-shuffles, such as the riffle shuffle, random transpositions and top-in-at-random shuffle. There are very interesting links here to group representation theory ....

In general, expect cutoff when there are large numbers of "second" eigenvalues. Should one expect cutoff for the case of an Ising model with weak interaction? Probably ....

The famous *Peres conjecture* says cutoff is to be expected for a chain with transitive symmetry if $(1-\lambda_2)\tau \to \infty$, where λ_2 is the second largest eigenvalue (so $1-\lambda_2$ is the "spectral gap"), and τ is the (deterministic) time at which the total variation distance to equilibrium becomes smaller than $\frac{1}{2}$. However there is a counterexample to Peres' conjecture as expressed above, (P. Diaconis, personal communication). So the conjecture needs to be refined!

Photographs used in text

- Police phone box en.wikipedia.org/wiki/Image:Earls_Court_Police_Box.jpg
- The standing martingale en.wikipedia.org/wiki/Image:Hunterhorse.jpg
- Boat Race: en.wikipedia.org/wiki/Image:Boat_Race_Finish_2008_-_Oxford_winners.jpg
- Impact site of fragment G of Comet Shoemaker-Levy 9 on Jupiter en.wikipedia.org/wiki/Image:Impact_site_of_fragment_G.gif
- The cardplayers en.wikipedia.org/wiki/Image:Paul_C\%C3\%A9zanne\%2C_Les_joueurs_de_carte_\%281892-95\%29.jpg
- Chinese abacus en.wikipedia.org/wiki/Image:Boulier1.JPG
- Error function en.wikipedia.org/wiki/Image:Error_Function.svg
- Boomerang en.wikipedia.org/wiki/Image:Boomerang.jpg
- Alexander Lyapunov en.wikipedia.org/wiki/Image:Alexander_Ljapunow_jung.jpg
- Riffle shuffle (photo by Johnny Blood) en.wikipedia.org/wiki/Image:Riffle_shuffle.jpg