Applied Stochastic Processes

Nicholas Georgiou and Matt Roberts
Durham University and University of Bath
APTS Oxford, 20-24 March 2023

Outline (clickable links)

- Introduction
- Markov chains and reversibility
- Renewal processes and stationarity
- Martingales
- Martingale convergence
- Ergodicity and general state spaces
- Foster-Lyapunov criteria
- Cutoff

Introduction

Two notions in probability

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.

The notes are arranged in a "two-panel" format: the main slides, with a white background, contain the main content of the course. The alternating slides with a grey background contain extra useful information and some exercises that will help you to cement your understanding of the material.

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.

The notes illustrate typical features of probability: the interplays between theory and practice, between rigour and intuition.

These notes were originally written by Wilfrid Kendall. Some material has since been added by Stephen Connor, Christina Goldschmidt and Amanda Turner.

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.

Some useful books

At increasing levels of mathematical sophistication:

- Finite Markov chains and algorithmic applications by Olle Häggström.
- Probability and random processes by Geoffrey Grimmett and David Stirzaker.
- Probability by Leo Breiman.
- Markov chains by James Norris.
- Stochastic processes by Sheldon Ross.
- Probability with martingales by David Williams.

- "Finite Markov chains and algorithmic applications" is a delightful introduction to finite state-space discrete-time Markov chains, starting from the point of view of computer algorithms.
- "Probability and random processes" is the standard undergraduate text on mathematical probability, and contains a huge amount of material.
- "Probability" by Breiman is a first-rate graduate-level introduction to probability.
- "Markov chains" by Norris presents the theory of Markov chains at a more graduate level of sophistication, revealing what we have concealed, namely the full gory story about Q-matrices.
- "Stochastic processes" by Ross makes use of the renewal process approach to convergence to equilibrium for a Markov chain which we will exploit here.
- "Probability with martingales" provides an excellent if mathematically demanding graduate treatment of the theory of martingales.

Free texts on the web

- Random walks and electric networks by Peter Doyle and Laurie Snell.
- Reversibility and stochastic networks by Frank Kelly.
- Markov chains and stochastic stability by Sean Meyn and Richard Tweedie.
- Reversible Markov Chains and Random Walks on Graphs by David Aldous and Jim Fill.

- Random walks and electric networks lays out (in simple and accessible terms) an important approach to Markov chains using relationship to resistance in electrical networks.

- We'll cover reversibility briefly in the lectures, but Reversibility and stochastic networks shows just how powerful the technique can be.

- Consult Markov chains and stochastic stability if you need to get informed about theoretical results on rates of convergence for Markov chains, e.g. because you are doing MCMC.

- Reversible Markov Chains and Random Walks on Graphs is the best unfinished book on Markov chains known to me (this was written by Wilfrid Kendall many years ago but remains true to the present lecturers).

Markov chains and reversibility

Reminder: convergence to equilibrium

Recall from the preliminary notes that if a Markov chain X on a countable state space (in discrete time) is

- irreducible
- aperiodic (only an issue in discrete time)
- positive recurrent (only an issue for infinite state spaces)

then

$$\mathbb{P}(X_n = i | X_0 = j) \to \pi_i$$

as $n \to \infty$ for all states i.

π is the unique solution to $\pi P = \pi$ such that $\sum_i \pi_i = 1$.

- Periodic cases, and continuous-time chains, can be considered with appropriate changes.

- Notice that if there are N states then $\pi P = \pi$ is (potentially) a system of N simultaneous equations in N unknowns, so that there is a complexity issue here if N is very large.

- We will see that for chains which have the property of time-reversibility, the calculation of π becomes much easier. Along the way, we will encounter some subtle but significant dependence issues.

A simple example

Consider simple symmetric random walk X on $\{0, 1, \ldots, k\}$, with "prohibition" boundary conditions: 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 $\pi = (\pi_0, \pi_1, \ldots, \pi_k)$.
2. The equilibrium equations $\pi P = \pi$ are solved by $\pi_i = \frac{1}{k+1}$ for all i.
3. Consider X in equilibrium:

$$\begin{aligned}\mathbb{P}(X_{n-1} = x,\, X_n = y) &= \mathbb{P}(X_{n-1} = x)\, \mathbb{P}(X_n = y | X_{n-1} = x) \\ &= \pi_x p_{x,y}\end{aligned}$$

and

$$\mathbb{P}(X_n = x,\, X_{n-1} = y) = \pi_y p_{y,x} = \pi_x p_{x,y}.$$

4. In equilibrium, the chain looks the same forwards and backwards. We say that the chain is reversible.

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

- Test understanding: verify the solution of the equilibrium equations.

- Test understanding: show that the chain run backwards satisfies the Markov property.

Reversibility

Definition

Suppose that $(X_{n-k})_{0 \leq k \leq n}$ and $(X_k)_{0 \leq k \leq n}$ have the same distribution for every n. Then we say that X is reversible.

- So X "looks the same" whether we run it backwards or forwards in time.
- A little thought shows that $X_0, X_1, \ldots$ must all have the same marginal distribution in order for this to be true. So, in particular, X must start in equilibrium in order for it to be reversible.

Detailed balance

1. Generalising the calculation we did for the random walk shows that a discrete-time Markov chain is reversible if it starts from equilibrium and the detailed balance equations hold: $\pi_x p_{x,y} = \pi_y p_{y,x}$.

2. If one can solve for π in $\pi_x p_{x,y} = \pi_y p_{y,x}$, then it is easy to show that $\pi P = \pi$.

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

4. In continuous time we instead require $\pi_x q_{x,y} = \pi_y q_{y,x}$, and if we can solve this system of equations then $\pi Q = 0$.

5. From a computational point of view, it is usually worth trying to solve the (easier) detailed balance equations first; if these are insoluble then revert to the more complicated $\pi P = \pi$ or $\pi Q = 0$.

- Test understanding: check that the calculation we did for the random walk generalizes to show point 1 on the previous slide.

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

- Even in the simple example of simple symmetric random walk, 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's book goes into more detail about this.

- Test understanding: show that detailed balance doesn't hold for the 3-state chain with transition probabilities $\frac{1}{3}$ for $0 \to 1$, $1 \to 2$, $2 \to 0$ and $\frac{2}{3}$ for $2 \to 1$, $1 \to 0$, $0 \to 2$. (Draw a picture.) We could have guessed that this chain is not reversible. Consider starting the chain in equilibrium and let T_n be the number of clockwise jumps before time n. For large n, this is approximately $n/3$. For the time-reversed chain, we would get $2n/3$ so that *statistically* we can tell the difference between the forwards and backwards chains.

- Test understanding: show that detailed balance *does* work for doubly reflected *asymmetric* simple random walk.

- We will see later that there are still major computational issues for more general Markov chains, connected with determining the normalizing constant for π.

Detailed balance and reversibility

Definition

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 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! We will soon see some examples where it is not immediately apparent that the time-reversed process in equilibrium should look statistically the same as the original process.

- 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. Invariant measures do have probabilistic signficance, but we won't touch on it here.

We will now consider examples of progressively more complicated Markov chains:

- the $M/M/1$ queue;
- a discrete-time chain on a 8×8 state space;
- Gibbs samplers;
- and Metropolis-Hastings samplers (briefly).

The $M/M/1$ queue

Here is a continuous-time example, the $M/M/1$ queue. We have

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

Detailed balance gives $\mu\pi_x = \lambda\pi_{x-1}$ 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).

Reversibility is more than a computational device: it tells us that if a stable $M/M/1$ queue is in equilibrium then people *leave* according to a Poisson process of rate λ. (This is known as Burke's theorem.)

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.

- Queueing processes are examples of *generalized birth-death processes*. The only possible transitions are from x to $x \pm 1$ and so the detailed balance equations are easily solved.

- Detailed balance is a subtle and important tool for the study of Markovian queueing networks. See e.g. Kelly's book.

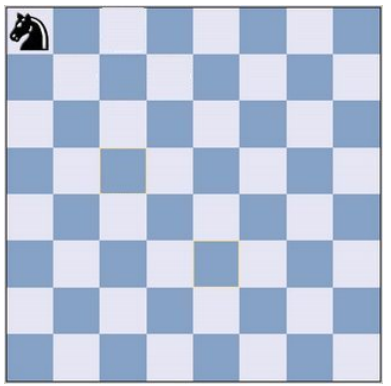
- Burke's theorem has deep consequences, with surprising applications (for example in the theory of random matrices).

Random chess

Now we turn to a multi-dimensional and less generic example, taken from the book of Aldous and Fill.

Example: a mean knight's tour

Place a chess knight at the corner of a standard 8x8 chessboard. Move it randomly, at each move choosing uniformly from the available legal chess moves independently of the past.

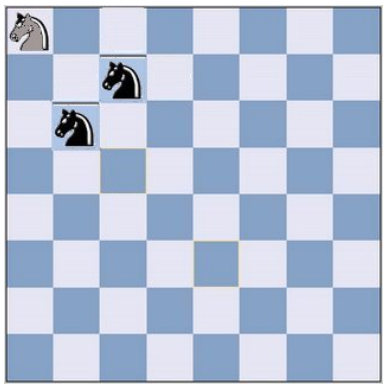


Random chess

Now we turn to a multi-dimensional and less generic example, taken from the book of Aldous and Fill.

Example: a mean knight's tour

Place a chess knight at the corner of a standard 8x8 chessboard. Move it randomly, at each move choosing uniformly from the available legal chess moves independently of the past.

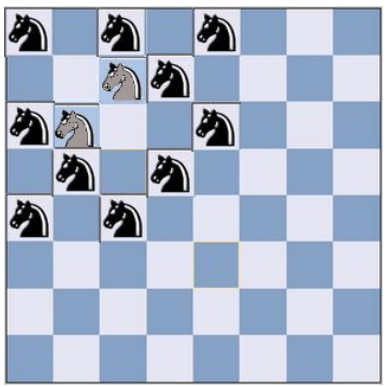


Random chess

Now we turn to a multi-dimensional and less generic example, taken from the book of Aldous and Fill.

Example: a mean knight's tour

Place a chess knight at the corner of a standard 8x8 chessboard. Move it randomly, at each move choosing uniformly from the available legal chess moves independently of the past.



1. Is the resulting Markov chain periodic?
2. What if you sub-sample at even times?
3. What is the equilibrium distribution? (Use detailed balance.)
4. What is the mean time till the knight returns to its starting point? (Inverse of equilibrium probability.)

- The chain is finite and irreducible.
- It is periodic of period 2, white versus black. It is necessary in computation to take care about this. In particular, a unique equilibrium distribution exists, but the chain does not converge to it in distribution. Sub-sampling at even times makes chain aperiodic on squares of one colour.
- Note: we didn't require aperiodicity for any of our detailed balance/reversibility arguments. Reversibility is pretty plausible for this chain, so detailed balance is a good bet. It's helpful to think about the *degree* of a square, i.e. the number of other squares to which the knight can jump from it. We will do the calculations on the next page.

- We have $\pi_v/d_v = \pi_u/d_u = c$ if $u \sim v$, where d_u is the degree of u. Consider the vertex degrees for the top-left quarter of the board:

$$\begin{matrix} 2 & 3 & 4 & 4 \\ 3 & 4 & 6 & 6 \\ 4 & 6 & 8 & 8 \\ 4 & 6 & 8 & 8 \end{matrix}$$

- The total degree is

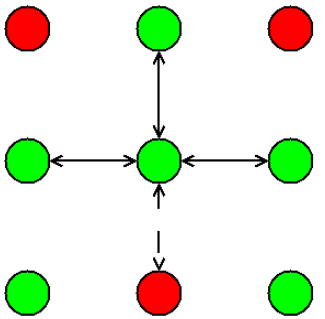
$$336 = (1 \times 2 + 2 \times 3 + 5 \times 4 + 4 \times 6 + 4 \times 8) \times 4$$

and so $1 = c \sum_v d_v = 336c$ and thus the equilibrium probability at a corner is $\pi_{\text{corner}} = 2/336 = 1/168$.

- Inverse of equilibrium probability shows that mean return time to corner is 168.

The Ising model

Pattern of spins $S_i = \pm 1$ on (finite fragment of) lattice. So here i is a vertex of the lattice. We will write $i \sim j$ to mean that i is a neighbour of j in the lattice.



Probability mass function:

$$\mathbb{P}(S_i = s_i \text{ all } i) \propto \exp\left(J \sum_{i,\, j:\, i \sim j} s_i s_j\right)$$

- The Ising model was introduced as an idealized model for magnetism. Other applications include modelling a simple binary image with noise.
- If $J = 0$ then we get the uniform distribution on spin configurations, which means spins behave independently. If $J < 0$, neighbouring spins like to differ (ferromagnetic case); if $J > 0$, neighbouring spins like to agree (antiferromagnetic case). In the physics setting, we're interested in expanding the fragment to fill the whole lattice. Changing the strength of the interaction J yields a phase transition (i.e. a large quantitative change in behaviour of the system, in this case, in the existence of large-scale order or not).
- The Ising model is the nexus for a whole variety of scientific approaches, each bringing their own rather different questions.
- Note, physics treatments use a (physically meaningful) over-parametrization $J \to \frac{J}{kT}$ where T is temperature and k is the Boltzmann constant. The $H \sum_i s_i \tilde{s}_i$ term can be interpreted physically as modelling an external magnetic field. We'll come back to a statistical interpretation for it in a moment.
- For a simulation physics view of the Ising model, see the expository article by David Landau in "Markov chain Monte Carlo: Innovations and Applications".

Gibbs sampler (or heat-bath) for the Ising model

Note that actually computing the normalizing constant for the Ising distribution is *hard* (in the sense of complexity theory). So how can we actually sample from the Ising model?

For a configuration $\mathbf{s}$, let $\mathbf{s}^{(i)}$ be the configuration obtained from $\mathbf{s}$ by flipping spin i. Let $\mathbf{S}$ be a configuration distributed according to the Ising measure.

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

1. Suppose the current configuration is $\mathbf{s}$.
2. Choose a site i in the lattice uniformly at random.
3. Flip the spin at i with probability $\mathbb{P}\left(\mathbf{S} = \mathbf{s}^{(i)} \middle| \mathbf{S} \in \{\mathbf{s}, \mathbf{s}^{(i)}\}\right)$; otherwise, leave it unchanged.

Gibbs sampler for the Ising model

Noting that $s_i^{(i)} = -s_i$, careful calculation yields

$$\mathbb{P}\Big(\mathbf{S} = \mathbf{s}^{(i)} \Big| \mathbf{S} \in \{\mathbf{s}, \mathbf{s}^{(i)}\}\Big) = \frac{\exp\Big(-J \sum_{j:j\sim i} s_i s_j\Big)}{\exp\Big(J \sum_{j:j\sim i} s_i s_j\Big) + \exp\Big(-J \sum_{j:j\sim i} s_i s_j\Big)}.$$

We have transition probabilities

$$p(\mathbf{s}, \mathbf{s}^{(i)}) = \frac{1}{n^2}\, \mathbb{P}\Big(\mathbf{S} = \mathbf{s}^{(i)} \Big| \mathbf{S} \in \{\mathbf{s}, \mathbf{s}^{(i)}\}\Big), \quad p(\mathbf{s}, \mathbf{s}) = 1 - \sum_i p(\mathbf{s}, \mathbf{s}^{(i)})$$

and simple calculations then show that

$$\sum_i \mathbb{P}\Big(\mathbf{S} = \mathbf{s}^{(i)}\Big) p(\mathbf{s}^{(i)}, \mathbf{s}) + \mathbb{P}(\mathbf{S} = \mathbf{s}) p(\mathbf{s}, \mathbf{s}) = \mathbb{P}(\mathbf{S} = \mathbf{s}),$$

so the chain has the Ising model as its equilibrium distribution.

- This is really a completely general computation! Theoretically it's straightforward.
- But note that the equilibrium equations are complicated to solve: n^2 equations, each with n^2 terms on left-hand side.
- This algorithm fits into the general pattern for Gibbs samplers: update individual random variables *sequentially* using conditional distributions given all other random variables.
- We only need to use *conditional* distributions, which means we can calculate with ratios, and so normalizing constants cancel out.

Detailed balance for the Gibbs sampler

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

$$\mathbb{P}(\mathbf{S} = \mathbf{s}) p(\mathbf{s}, \mathbf{s}^{(i)}) = \mathbb{P}\left(\mathbf{S} = \mathbf{s}^{(i)}\right) p(\mathbf{s}^{(i)}, \mathbf{s})$$

for all $\mathbf{s}$.

Test understanding: check the detailed balance calculations.

It turns out that detailed balance also holds for processes obtained from:

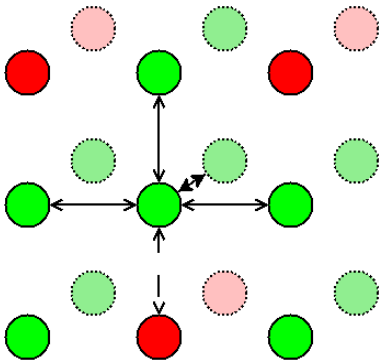
- systematic scans
- coding ("simultaneous updates on alternate colours of a chessboard")

but *not* for wholly simultaneous updates.

The Ising model with an external field

Probability mass function for the Ising model with an external field $\{\tilde{s}_i\}$:

$$\mathbb{P}(S_i = s_i \text{ all } i) \propto \exp\left(J \sum_{i,\, j:\, i \sim j} s_i s_j + H \sum_i s_i \tilde{s}_i\right).$$



We could do exactly the same as we did for the model without an external field. The calculations all work without any problems.

Image reconstruction using the Gibbs sampler

Suppose that we have a black and white image that has been corrupted by some noise. Let $\tilde{\mathbf{s}}$ represent the noisy image (e.g. $\tilde{s}_i = 1$ if pixel i is black, and -1 if white), and use it as an external field, with $J, H > 0$. H here measures the "noisiness".

Bayesian interpretation: we observe the noisy signal $\widetilde{\mathbf{S}}$ and want to make inference about the true signal. We obtain posterior distribution $\mathbb{P}\left(\mathbf{S} = \mathbf{s} \middle| \widetilde{\mathbf{S}} = \tilde{\mathbf{s}}\right) \propto \exp\left(J \sum_{i \sim j} s_i s_j + H \sum_i s_i \tilde{s}_i\right)$ from which we would like to sample. In order to do this, we run the Gibbs sampler to equilibrium (with $\tilde{\mathbf{s}}$ fixed), starting from the noisy image.

Metropolis-Hastings sampler

Suppose again that we want to sample from a distribution π whose form we know only up to a normalising constant.

An important alternative to the Gibbs sampler, even more closely connected to detailed balance, is Metropolis-Hastings:

- Suppose that $X_n = x$.
- Pick y using a transition probability kernel $\kappa(x, y)$ (the proposal kernel).
- Accept the proposed transition $x \to y$ with probability

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

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

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

- Actually the Gibbs sampler is a special case of the Metropolis-Hastings sampler.
- Test understanding: write down the transition probability kernel for X.
- Test understanding: check that π solves the detailed balance equations.
- Common proposal kernels include:
  - *independence sampler*: $\kappa(x, y) = f(y)$;
  - *random-walk sampler*: $\kappa(x, y) = f(y - x)$;
  - *Langevin sampler*: replace random-walk shift by shift depending on $\operatorname{grad} \log \pi$.
- Importantly, the acceptance probability depends only on the *ratio* $\pi(x)/\pi(y)$, so normalizing constants cancel out.

Renewal processes and stationarity

We spent the first lecture exploring a particular way to go about finding equilibrium distributions for Markov chains. In this lecture, we're going to delve a little deeper into the structure of Markov chains and into their convergence to stationarity. A key tool is a simple class of processes called renewal processes. The approach we take here will be the one we build upon towards the end of the course when we consider Markov chains on general state-spaces.

Stopping times

Let $(X_n)_{n\geq 0}$ be a stochastic process and write $\mathcal{F}_n$ for the collection of events "which can be determined from $X_0, X_1, \ldots, X_n$."For example,

$$\left\{ \min_{0\leq k\leq n} X_k = 5 \right\} \in \mathcal{F}_n$$

but

$$\left\{ \min_{0\leq k\leq n+1} X_k = 5 \right\} \notin \mathcal{F}_n.$$

Definition

A random variable T taking values in $\{0, 1, 2, \ldots\} \cup \{\infty\}$ is said to be a stopping time (for the process X) if, for all n, $\{T \leq n\}$ is determined by the information available at time n, i.e. $\{T \leq n\} \in \mathcal{F}_n$.

- Intuitively, a stopping time is one such that "we know when it has occurred".

- Note that we need to have a clear notion of exactly what might be $\mathcal{F}_n$, the 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.)

Random walk example

Let X be a random walk started from 0.

- The random time $T = \inf\{n > 0 : X_n \geq 10\}$ is a stopping time.
- Indeed $\{T \leq n\}$ is clearly determined by the information available at time n:

$$\{T \leq n\} = \{X_1 \geq 10\} \cup \ldots \cup \{X_n \geq 10\}\,.$$

- On the other hand, the random time $S = \sup\{0 \leq n \leq 100 : X_n \geq 10\}$ is not a stopping time.

Note that the minimum of two stopping times is a stopping time!

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

- Note that the stopping time T could take the value ∞, since it is possible that X_n never goes above 10. In particular, if X has a negative drift, this occurs with positive probability.

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

- In the case of the hitting time of A,

$$\{T_A \leq n\} = \{X_1 \in A\} \cup \ldots \cup \{X_n \in A\}$$

  so $\{T_A \leq n\}$ is determined by information at time n, so T_A is a stopping time.

- Let T_1 and T_2 be stopping times for X, and let $T = \min\{T_1, T_2\}$. Then

$$\{T \leq n\} = \{T_1 \leq n\} \cup \{T_2 \leq n\} \in \mathcal{F}_n\,.$$

Strong Markov property

Suppose that T is a stopping time for the Markov chain $(X_n)_{n\geq 0}$.

Theorem

Conditionally on $\{T < \infty\}$ and $X_T = i$, the process $(X_{T+n})_{n\geq 0}$ has the same distribution as $(X_n)_{n\geq 0}$ started from $X_0 = i$. Moreover, given $\{T < \infty\}$, $(X_{T+n})_{n\geq 0}$ and $(X_n)_{0\leq n<T}$ are conditionally independent given X_T.

This is called the strong Markov property.

- The strong Markov property says that "the future and the past are independent given the present" remains true even at random times T, as long as they are stopping times.

- This is, in general, not true for non-stopping times. For example, if X is a simple random walk and $T = \sup\{0 \leq n \leq 100 : X_n = 10\}$ then, if $T < \infty$, we know the path of X between times $T + 1$ and 100 cannot hit 10. So it cannot be a new copy of the original Markov chain.

Hitting times and the Strong Markov property

Consider an irreducible recurrent Markov chain on a discrete state-space S. Fix $i \in S$ and let

$$H_0^{(i)} = \inf\{n \geq 0 : X_n = i\}.$$

For $m \geq 0$, recursively let

$$H_{m+1}^{(i)} = \inf\{n > H_m^{(i)} : X_n = i\}.$$

It follows from the strong Markov property that the random variables

$$H_{m+1}^{(i)} - H_m^{(i)}, m \geq 0$$

are independent and identically distributed and also independent of $H_0^{(i)}$.

- Test understanding: show that $H_0^{(i)}, H_1^{(i)}, \ldots$ is a sequence of stopping times.

- Notice that the MC is recurrent iff these random variables are all almost surely finite. It is positive recurrent iff they all have finite mean.

- Test understanding: check that you can see why the increments

$$H_{m+1}^{(i)} - H_m^{(i)}, m \geq 0$$

are independent and identically distributed.

Suppose we start our Markov chain from $X_0 = i$. Then $H_0^{(i)} = 0$.

Consider the number of visits to state i which have occurred by time n (not including the starting point!) i.e.

$$N^{(i)}(n) = \# \left\{ k \geq 1 : H_k^{(i)} \leq n \right\}.$$

This is an example of a renewal process.

Renewal processes

Definition Let $Z_1, Z_2, \ldots$ be i.i.d. integer-valued random variables such that $\mathbb{P}(Z_1 > 0) = 1$. Let $T_0 = 0$ and, for $k \geq 1$, let

$$T_k = \sum_{j=1}^{k} Z_j$$

and, for $n \geq 0$,

$$N(n) = \#\{k \geq 1 : T_k \leq n\}.$$

Then $(N(n))_{n \geq 0}$ is a (discrete) renewal process.

- Draw a picture!
- $T_1, T_2, \ldots$ are the times of incidents (say, times when we need to replace a lightbulb) and $N(n)$ counts the number of incidents which have occurred by time n.
- Note that the times $Z_1, Z_2, \ldots$ between incidents must be strictly positive!
- Note that $\{N(n) = k\} = \{T_k \leq n < T_{k+1}\}$.
- We can make a more general definition where Z_1 doesn't have to be integer-valued, but we won't need that here. But note that if Z_1 is exponentially distributed then N is a Poisson process.

Here is a special example. Suppose that $Z_1, Z_2, \ldots$ are i.i.d. $\mathrm{Geom}(p)$ i.e.

$$\mathbb{P}(Z_1 = k) = (1-p)^{k-1}p, \quad k \geq 1.$$

Then we can think of Z_1 as the number of independent coin tosses required to first see a head, if heads has probability p.

So $N(n)$ has the same distribution as the number of heads in n independent coin tosses i.e. $N(n) \sim \mathrm{Bin}(n, p)$ and, moreover,

$$\begin{aligned}&\mathbb{P}(N(k+1) = n_k + 1 | N(0) = n_0, N(1) = n_1, \ldots, N(k) = n_k) \\ &= \mathbb{P}(N(k+1) = n_k + 1 | N(k) = n_k) = p\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(N(k+1) = n_k | N(0) = n_0, N(1) = n_1, \ldots, N(k) = n_k) \\ &= \mathbb{P}(N(k+1) = n_k | N(k) = n_k) = 1 - p.\end{aligned}$$

So, in this case, $(N(n))_{n \geq 0}$ is a Markov chain.

However, renewal processes are not normally Markov. The example on the previous slide is essentially the only example of a discrete renewal process which is Markov.

Why? Because the geometric distribution has the memoryless property:

$$\mathbb{P}(Z_1 - r = k | Z_1 > r) = (1-p)^{k-1}p, \quad k \geq 1.$$

So, regardless of what I know about the process up until the present time, the distribution of the remaining time until the next renewal is again geometric. The geometric is the only discrete distribution with this property.

For example, take $Z_1 \sim 1 + \mathrm{Ber}(p)$. Then

$$\mathbb{P}(N(k+1) = N(k) + 1 | N(k) = N(k-1)) = 1$$

and

$$\mathbb{P}(N(k+1) = N(k) + 1 | N(k) = N(k-1) + 1) = 1 - p.$$

Delayed renewal processes

Definition

Let Z_0 be a non-negative integer-valued random variable and, independently, let $Z_1, Z_2, \ldots$ be independent strictly positive and identically distributed integer-valued random variables.
For $k \geq 0$, let

$$T_k = \sum_{j=0}^{k} Z_j$$

and, for $n \geq 0$,

$$N(n) = \#\{k \geq 0 : T_k \leq n\}.$$

Then $(N(n))_{n \geq 0}$ is a (discrete) delayed renewal process, with delay Z_0.

- In general, we may wish to start our Markov chain X from a general initial distribution. Then we do not necessarily have $X_0 = i$ and so, in particular, $H_0^{(i)}$ may be non-zero. This motivates a more general definition of a renewal process.

- The distribution of $H_0^{(i)}$ depends on the initial distribution of the Markov chain X.

- In the lightbulb setting: you might not know when the bulb present at time 0 was installed! So its lifetime is represented by Z_0.

- Note that this allows for the possibility that $Z_0 = 0$, unlike the other inter-renewal times which must be strictly positive.

Strong law of large numbers

Suppose that $\mu := \mathbb{E}[Z_1] < \infty$. Then the SLLN tells us that

$$\frac{T_k}{k} = \frac{1}{k}\sum_{j=0}^{k} Z_j \to \mu \text{ almost surely as } k \to \infty.$$

One can use this to show that

$$\frac{N(n)}{n} \to \frac{1}{\mu} \text{ almost surely as } n \to \infty$$

which tells us that we see renewals at a long-run average rate of $1/\mu$.

- Observe that T and N are effectively inverses of one another.
- Test understanding: see if you can write down the argument to show what is claimed in the previous slide. Hint: $N(n) \to \infty$ as $n \to \infty$.

Probability of a renewal

Think back to our motivating example of hitting times of state i for a Markov chain. Suppose we want to think in terms of convergence to equilibrium: we would like to know what is the probability that at some large time n there is a renewal (i.e. a visit to i). We have $N(n) \approx n/\mu$ for large n (where μ is the expected return time to i), so as long as renewals are evenly spread out, the probability of a renewal at a particular large time should look like $1/\mu$.

This intuition turns out to be correct as long as every sufficiently large integer time is a possible renewal time. In particular, let

$$d = \gcd\{n : \mathbb{P}(Z_1 = n) > 0\}.$$

If $d = 1$ then this is fine; if we are interpreting renewals as returns to i for our Markov chain, this says that the chain is aperiodic.

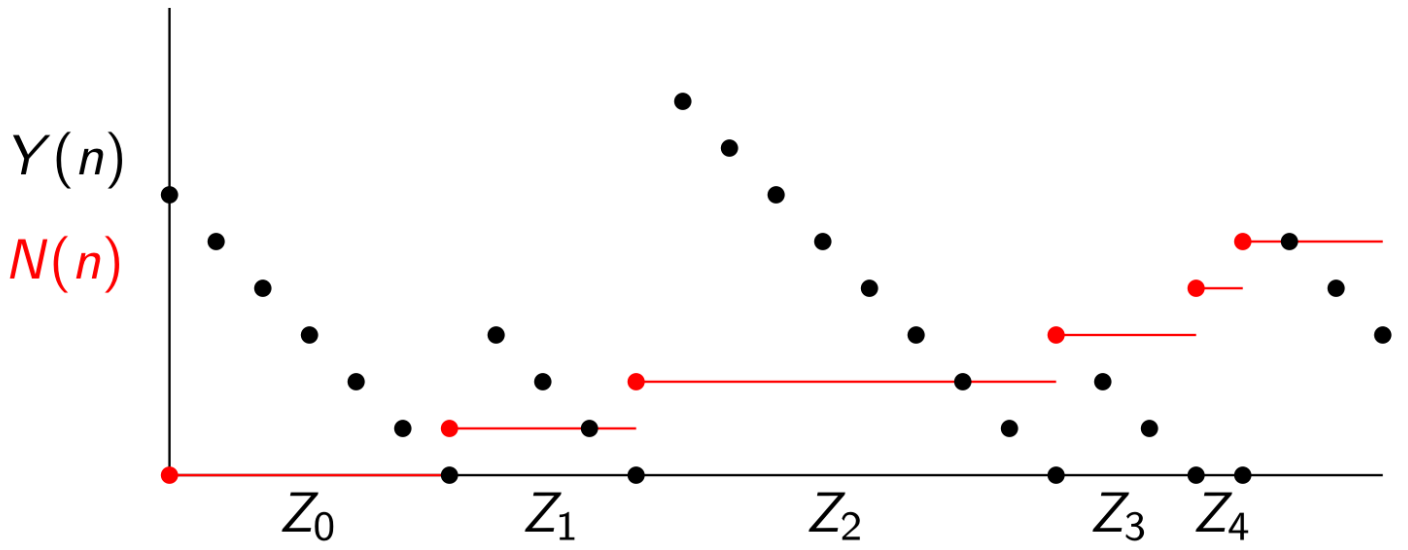
- Recall that $Z_1 \geq 1$ and so $1/\mu \leq 1$!
- If $d \geq 2$ and we are in the non-delayed setting, then renewals are only possible at times which are multiples of d and so the probability of a renewal at one of these times ends up tending to d/μ (note $\mathbb{E}[Z_1] \geq d$ necessarily in this case!).

An auxiliary Markov chain

We saw that a delayed renewal process $(N(n))_{n\geq 0}$ is not normally itself Markov. But we can find an auxiliary process which is. For $n \geq 0$, let

$$Y(n) := T_{N(n-1)} - n.$$

This is the time until the next renewal.



Transition probabilities for the time to next renewal

For $n \geq 0$,

$$Y(n) := T_{N(n-1)} - n.$$

$(Y(n))_{n\geq 0}$ has very simple transition probabilities: if $k \geq 1$ then

$$\mathbb{P}(Y(n+1) = k-1 | Y(n) = k) = 1$$

and

$$\mathbb{P}(Y(n+1) = i | Y(n) = 0) = \mathbb{P}(Z_1 = i+1) \text{ for } i \geq 0.$$

A stationary version

Recall that $\mu = \mathbb{E}[Z_1]$, where $(Z_n)_{n\geq 0}$ is our original Markov chain. Then the stationary distribution for the auxiliary Markov chain is

$$\nu_i = \frac{1}{\mu}\,\mathbb{P}(Z_1 \geq i+1), \quad i \geq 0.$$

If we start a delayed renewal process $(N(n))_{n\geq 0}$ with $Z_0 \sim \nu$ then the time until the next renewal is always distributed as ν. We call such a delayed renewal process stationary.

Notice that the stationary probability of being at a renewal time is $\nu_0 = 1/\mu$.

- Test understanding: check that ν defines a probability mass function.

- Test understanding: demonstrate that ν is stationary for Y. Note: Y is clearly not reversible, so there's no point trying detailed balance!

- Test understanding: check that in the case $Z_1 \sim \mathrm{Geom}(p)$, the stationary distribution ν is also $\mathrm{Geom}(p)$. So the renewal process is already stationary.

Size-biasing and inter-renewal intervals

The stationary distribution

$$\nu_i = \frac{1}{\mu}\,\mathbb{P}(Z_1 \geq i+1), \quad i \geq 0$$

has an interesting interpretation.

Let Z^* be a random variable with probability mass function

$$\mathbb{P}(Z^* = i) = \frac{i\,\mathbb{P}(Z_1 = i)}{\mu}, \qquad i \geq 1.$$

We say that Z^* has the size-biased distribution associated with the distribution of Z_1.

Now let $L \sim \mathrm{U}\{0, 1, \ldots, Z^* - 1\}$. Then $L \sim \nu$.

- Test understanding: check that the definition of Z^* is indeed a probability mass function.

- Test understanding: check the last line of the previous slide.

Interpretation

This isn't a coincidence! We are looking at a large time n and want to know how much time there is until the next renewal. Intuitively, n has more chance to fall in a longer interval. Indeed, it is i times more likely to fall in an interval of length i than an interval of length 1. So the inter-renewal time that n falls into is size-biased.

Again intuitively, it is equally likely to be at any position inside that renewal interval, and so the time until the next renewal should be uniform on $\{0, 1, \ldots, Z^* - 1\}$ i.e. it should have the same distribution as L.

Convergence to stationarity

Theorem (Blackwell's renewal theorem)

Suppose that the distribution of Z_1 in a delayed renewal process is such that $\gcd\{n : \mathbb{P}(Z_1 = n) > 0\} = 1$ and $\mu := \mathbb{E}[Z_1] < \infty$. Then

$$\mathbb{P}(\text{renewal at time } n) = \mathbb{P}(Y(n) = 0) \to \frac{1}{\mu}$$

as $n \to \infty$.

(Recall that the stationary probability of being at a renewal time is $\nu_0 = 1/\mu$.)

The idea behind the proof: coupling

- We start two independent versions of our renewal process: one with a general delay distribution, and one with delay distribution ν. The second version of Y is therefore in equilibrium.

- We show that, if we wait long enough, there will be a (random but finite) time τ such that both renewal processes have a renewal at τ.

- After τ, the renewal times of both versions have the same law. Therefore the difference between the probability that the first version has a renewal at time n and the probability that the second version has a renewal at time n must be smaller than $\mathbb{P}(\tau > n)$.

The idea behind the proof: coupling (cont.)

- Since the second version is in equilibrium, the probability it has a renewal at time n is $\nu_0 = 1/\mu$ for any n. Thus we deduce that

$$|\,\mathbb{P}(\text{renewal of first version at time } n) - 1/\mu| \leq \mathbb{P}(\tau > n) \to 0.$$

This is only a sketch proof (and not examinable), but it is not too difficult to make rigorous. Most of the work is in showing that τ is almost surely finite.

Convergence to stationarity

We have shown:

Theorem (Blackwell's renewal theorem)
Suppose that the distribution of Z_1 in a delayed renewal process is such that $\gcd\{n : \mathbb{P}(Z_1 = n) > 0\} = 1$ and $\mu := \mathbb{E}[Z_1] < \infty$. Then

$$\mathbb{P}(\text{renewal at time } n) = \mathbb{P}(Y(n) = 0) \to \frac{1}{\mu}$$

as $n \to \infty$.

Convergence to stationarity

From Blackwell's renewal theorem we can immediately deduce the usual convergence to stationarity for a Markov chain.

Theorem

Let X be an irreducible, aperiodic, positive recurrent Markov chain (i.e. $\mu_i = \mathbb{E}\left[H_1^{(i)} - H_0^{(i)}\right] < \infty$). Then, whatever the distribution of X_0,

$$\mathbb{P}(X_n = i) \to \frac{1}{\mu_i}$$

as $n \to \infty$.

Note the interpretation of the stationary probability of being in state i as the inverse of the mean return time to i.

- You may have seen a direct proof of convergence to stationarity that is very similar to the proof of Blackwell's renewal theorem.

- Irreducibility tells that we can reach i from any other state. Aperiodicity gives us the necessary non-lattice condition in Blackwell's theorem. Positive recurrence tells us precisely that the inter-renewal times have finite expectation.

Decomposing a Markov chain

Consider an irreducible, aperiodic, positive recurrent Markov chain X, fix some state i and let $H_m = H_m^{(i)}$ for all $m \geq 0$.

Recall that $(H_{m+1} - H_m, m \geq 0)$ is a collection of i.i.d. random variables, by the Strong Markov property.

More generally, it follows that the collection of pairs

$$\left(H_{m+1} - H_m,\ (X_{H_m+n})_{0 \leq n \leq H_{m+1}-H_m}\right), m \geq 0,$$

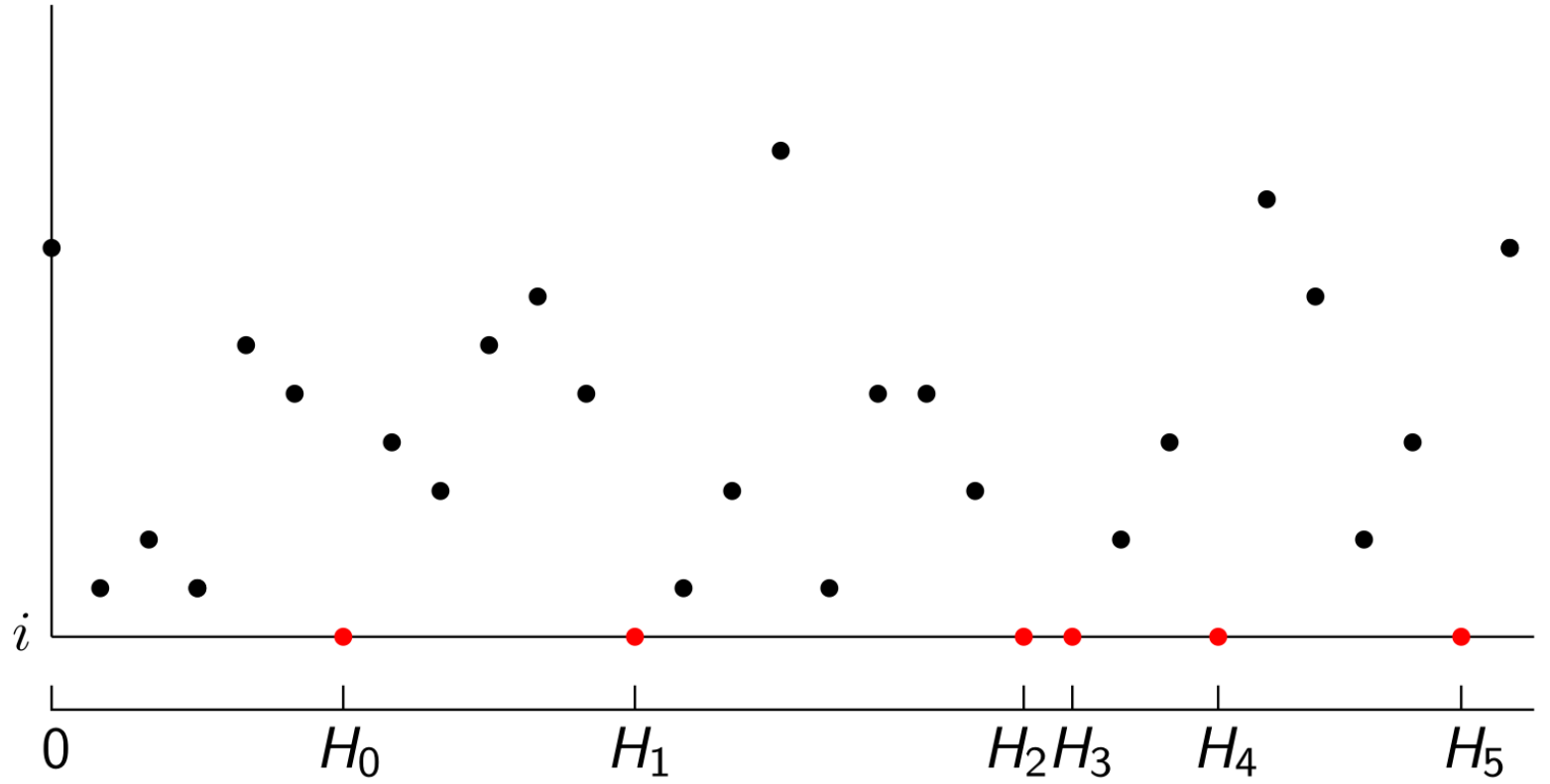
(where the first element of the pair is the time between the mth and $(m+1)$st visits to i, and the second element is a path which starts and ends at i and doesn't touch i in between) are independent and identically distributed.

Decomposing a Markov chain (cont.)

Conditionally on $H_{m+1} - H_m = k$, $(X_{H_m+n})_{0 \leq n \leq k}$ has the same distribution as the Markov chain X started from i and conditioned to first return to i at time k.

So we can split the path of a recurrent Markov chain into independent chunks ("excursions"), between successive visits to i. The renewal process of times when we visit i becomes stationary. To get back the whole Markov chain, we just need to "paste in" pieces of conditioned path.

Decomposing a Markov chain (cont.)



Essentially the same picture will hold true when we come to consider general state-space Markov chains in the last three lectures.

Martingales

This is the second major theme of this course: 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 unfair to your opponent.)

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 pervade modern probability

1. We say the random process $X = (X_n : n \geq 0)$ is a martingale if it satisfies the martingale property:

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

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

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

- For a conversation with the inventor, see https://chance.dartmouth.edu/Doob/conversation.html
- In words: expected future level of X is current level.
- We use $\mathcal{F}_n$ notation without comment in future, usually representing conditioning by $X_0, X_1, \ldots, X_n$ (if X is martingale in question). *Sometimes* further conditioning will be added; but $\mathcal{F}_{n+1}$ has at least as much conditioning as $\mathcal{F}_n$. Crucially, the "Tower property" of conditional expectation then applies:

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

- Test understanding: deduce $\mathbb{E}[X_{n+k}|\mathcal{F}_n] = X_n\,.$
- 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...

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 amartingale if you do."

3. Result of playing Thackeray's martingale system and stopping on first win: 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 $M_n = 1$.

- This is the "doubling" strategy. The equestrian meaning resembles the probabilistic definition to some extent.
- 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.
- 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.)

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 $Z_{n+1,1} + \ldots + Z_{n+1,Y_n}$ of Y_n independent copies of a non-negative integer-valued family-size r.v. Z. The formal definition requires the $Z_{n+1,i}$ 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.

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.

4. If ζ is the smallest non-negative root of the equation $G(s) = s$, then ζ^{Y_n} defines a martingale.

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

- New Yorker's definition of branching process:

> 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.

- Test understanding: check Example 2. Note, X measures relative deviation from the deterministic Malthusian model of growth.

- Test understanding: check Example 3.

- Test understanding: check Example 4. It can be shown that ζ is the probability that the population eventually becomes extinct.

- Indeed, we can also generalize to general Y_0.

Definition of a martingale

Formally:

Definition

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

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

- It is useful to have a general definition of expectation here (see the section on conditional expectation in the preliminary notes).
- 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 $\mathcal{F}_n$:

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

Supermartingales and submartingales

Two associated definitions.

Definition

$(X_n : n \geq 0)$ is a supermartingale if $\mathbb{E}[|X_n|] < \infty$ for all n and

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

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

Definition

$(X_n : n \geq 0)$ is a submartingale if $\mathbb{E}[|X_n|] < \infty$ for all n and

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

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

Supermartingales:

- 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.

Submartingales:

- 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!

Examples of super/submartingales

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 its 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 the branching process (Y_n) and think about 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), and a martingale if $\mu = 1$ (critical case).

5. By the conditional form of Jensen's inequality, if X is a martingale then $|X|$ is a submartingale.

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.

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

More martingale examples

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

$$\left(\frac{1-p}{p}\right)^{X_n}$$

   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.

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

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

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

An example of importance in finance

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

2. Then the following is a martingale:

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

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

- Here modifications of Y_n provide the simplest model for market price fluctuations appropriately discounted.
- In fact (S_n) is a martingale, though this is not the point here.
- Test understanding: Prove that (Y_n) is a martingale! Hint: $\mathbb{E}[\exp(N_1)] = e^{\sigma^2/2}$.
- Test understanding: figure out the modification!
- A continuous-time variation on this example (using Brownian motion) is an important baseline model in mathematical finance.
- Note that the martingale can be expressed as

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

Martingales and likelihood

1. Suppose that a random variable X has a distribution which depends on a parameter θ. Independent copies $X_1, X_2, \ldots$ of X are observed at times $1, 2, \ldots$. The likelihood of θ at time n is

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

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

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

- Simple case of normal data with unknown mean θ:

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

- Idea to have in mind: $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. So taking the expectation under H_0 is assuming that θ_0 is the true value.

- Test understanding: check that this is a martingale.

- 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 $N(\theta, \sigma^2)$.

Martingales for Markov chains

To connect to the first theme of the course, Markov chains provide us with a large class of examples of martingales.

1. Let X be a Markov chain with countable state-space S and transition probabilities $p_{x,y}$. Let $f : S \to \mathbb{R}$ be any bounded function.
2. Take $\mathcal{F}_n$ to contain all the information about $X_0, X_1, \ldots, X_n$.
3. Then

$$M_n^f = f(X_n) - f(X_0) - \sum_{i=0}^{n-1} \left[\sum_{y \in S} (f(y) - f(X_i)) p_{X_i,y} \right]$$

   defines a martingale.
4. In fact, if M^f is a martingale for all bounded functions f then X is a Markov chain with transition probabilities $p_{x,y}$.

- We need some condition on f to ensure that the resulting process is integrable.

- Note that $\mathcal{F}_n$ contains more information than just $M_0, M_1, \ldots, M_n$.

- Test understanding: show that M^f is indeed a martingale. First note that $\sum_{y \in S}(f(y) - f(X_i))p_{X_i,y} = \mathbb{E}[f(X_{i+1}) - f(X_i)|X_i]$. Using this and the Markov property, check that $\mathbb{E}\left[M^f_{n+1} - M^f_n | \mathcal{F}_n\right] = 0$.

- See Section 4.1 of Norris' book for a proof and discussion.

Martingales for Markov chains: harmonic functions

Call a function $f : S \to \mathbb{R}$ harmonic if $f(x) = \sum_{y \in S} f(y) p_{x,y}$ for all $x \in S$.

We defined

$$M_n^f = f(X_n) - f(X_0) - \sum_{i=0}^{n-1} \left[\sum_{y \in S} (f(y) - f(X_i)) p_{X_i,y} \right]$$

and so we see that if f is harmonic then $f(X_n)$ is itself a martingale.

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

Martingale convergence

It's often not enough just to know about stochastic processes at fixed, deterministic times; for many applications, it's also natural to think about what happens at random times. For example, suppose you are 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.

The martingale property at random times
The big idea

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

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

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

1. T must not "look ahead";
2. T must not be "too big".

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

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

- Example of "looking ahead": Set $S = \sup\{X_n : 0 \leq n \leq 10\}$ and set $T = \inf\{n : X_n = S\}$. Then $\mathbb{E}[X_T] \geq \mathbb{P}(S > 0) > 0 = \mathbb{E}[X_0]$.

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

- 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.

Stopping times

We have already seen what we mean by a random time "not looking ahead": such a time T is more properly called a *stopping time*.

Example

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

- The random time $T = \inf\{n : Y_n = 0\} = \inf\{n : X_n = 0\}$ is a stopping time.
- Indeed $\{T \leq n\}$ is clearly determined by the information available at time n:

$$\{T \leq n\} = \{Y_n = 0\},$$

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

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

- For a more interesting example, consider

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

- In the case of S,

$$\{S \leq n\} = A_1 \cup A_2 \cup \ldots \cup A_n$$

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.

- 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.

Stopping times aren't enough

However, even if T is a stopping time, we clearly need a stronger condition in order to say that $\mathbb{E}[M_T|\mathcal{F}_0] = M_0$.

e.g. let X be a random walk on $\mathbb{Z}$, started at 0.

- $T = \inf\{n > 0 : X_n \geq 10\}$ is a stopping time
- T is typically "too big": so long as it is almost surely finite, $X_T \geq 10$ and we deduce that $0 = \mathbb{E}[X_0] < \mathbb{E}[X_T]$.

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

- T being almost surely finite means that, with probability 1, X does eventually exceed 10. (This happens if $\mathbb{E}[X_1] > 0$ or if $\mathbb{E}[X_1] = 0$ and $\mathbb{P}(X_1 > 0) > 0$.)

- Another example: let Y be a branching process of mean-family-size μ with $Y_0 = 1$ — recall that $X_n = Y_n/\mu^n$ determines a martingale.

The random time $T = \inf\{n : Y_n = 0\} = \inf\{n : X_n = 0\}$ is a stopping time but, again, T here is "too big": so long as it is almost surely finite then $1 = \mathbb{E}[X_0] > \mathbb{E}[X_T]$. (T is almost surely finite if $\mu < 1$, or if $\mu = 1$ and there is positive chance of zero family size.)

Optional stopping theorem

Theorem

Suppose M is a martingale and T is a bounded stopping time. Then

$$\mathbb{E}[M_T|\mathcal{F}_0] = M_0\,.$$

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

Note we can take expectation of a single random variable X (i.e., X is integrable) exactly when $\mathbb{E}[|X|; |X| > n] \equiv \mathbb{E}\left[|X|\mathbf{1}_{|X|>n}\right] \to 0$ as $n \to \infty$. (This fails when $\mathbb{E}[|X|; |X| > n] = \infty$!).

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

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

Examples:

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

The optional stopping theorem also holds for stopping times $S \le T$, in that $\mathbb{E}[M_T|\mathcal{F}_S] = M_S$, but this requires more serious measure theory to carefully define what we mean by $\mathcal{F}_S$ (events determined by time S)...

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

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

Contrast with 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.

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

Strategies proposed for other games to seem less convincing; for example, the Labouchère system favoured by Ian Fleming en.wikipedia.org/wiki/Labouchère_system:

> The Labouchère 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.

Exit from an interval

Here's an elegant application of the optional stopping theorem.

- Suppose that X is a simple symmetric random walk started from 0. Then X is a martingale.
- Let $T = \inf\{n : X_n = a \text{ or } X_n = -b\}$. (T is almost surely finite.) Suppose we want to find $\mathbb{P}(X \text{ hits } a \text{ before } -b) = \mathbb{P}(X_T = a)$.
- On the (random) time interval $[0, T]$, X is bounded, and so we can apply the optional stopping theorem to see that

$$\mathbb{E}[X_T] = \mathbb{E}[X_0] = 0.$$

- But then

$$\begin{aligned} 0 = \mathbb{E}[X_T] &= a\,\mathbb{P}(X_T = a) - b\,\mathbb{P}(X_T = -b) \\ &= a\,\mathbb{P}(X_T = a) - b(1 - \mathbb{P}(X_T = a)). \end{aligned}$$

Solving gives $\mathbb{P}(X \text{ hits } a \text{ before } -b) = \frac{b}{a+b}$.

Martingales and hitting times

Suppose that $X_1, X_2, \ldots$ are i.i.d. $N(-\mu, 1)$ random variables, where $\mu > 0$. Let $S_n = X_1 + \ldots + X_n$ and let T be the time when S first exceeds level $\ell > 0$.

Then $\exp\left(\alpha(S_n + \mu n) - \frac{\alpha^2}{2} n\right)$ determines a martingale (for any $\alpha \geq 0$), and the optional stopping theorem can be applied to show

$$\mathbb{E}[\exp(-pT)] \sim e^{-(\mu + \sqrt{\mu^2 + 2p})\ell}, \quad p > 0.$$

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

Note $T = \inf\{n : S_n \geq \ell\}$. Use the optional stopping theorem on the bounded stopping time $\min\{T, n\}$:

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

and then take $n \to \infty$. (For this example, $T = \infty$ with positive probability, so extra care must be taken!)

On $\{T = \infty\}$, we have $\min\{n, T\} = n$ for all $n \geq 0$. Also, $S_n \to -\infty$ a.s. (by SLLN), so for $\alpha > 2\mu$,

$$\mathbb{E}\Big[\exp\Big(\alpha S_n + \alpha(\mu - \frac{\alpha}{2})n\Big)\Big] \to 0 \text{ as } n \to \infty.$$

On $\{T < \infty\}$, we have S_T relatively close to ℓ; letting $n \to \infty$, for $\alpha > 2\mu$ and large ℓ,

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

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

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

Improvement to equality arises since Brownian motion is continuous in time and so cannot jump over the level ℓ without hitting it.

Martingale convergence

Theorem

Suppose X is a non-negative supermartingale. Then there exists a random variable Z such that $X_n \to Z$ a.s.; moreover, $\mathbb{E}[Z|\mathcal{F}_n] \leq X_n$.

Theorem

Suppose X is a bounded martingale (or, more generally, uniformly integrable). Then $Z = \lim_{n\to\infty} X_n$ exists a.s.; moreover, $\mathbb{E}[Z|\mathcal{F}_n] = X_n$.

Theorem

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 $Z = \lim_{n\to\infty} X_n$ exists a.s.; moreover, $\mathbb{E}[Z|\mathcal{F}_n] = X_n$.

At the heart of the argument here is Doob's famous "upcrossing lemma": 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 X 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 $0 = \mathbb{E}[Z|\mathcal{F}_n] \leq X_n$ since $X_n \geq 0$.

- Thus simple symmetric random walk X begun at 0 and stopped on hitting a or $-b$ must converge to a limiting value Z. Evidently Z can only take the values a or $-b$. Moreover, since $\mathbb{E}[Z|\mathcal{F}_n] = X_n$ we deduce $\mathbb{P}(Z = a|\mathcal{F}_n) = \frac{X_n+b}{a+b}$.

- Sketch argument: from martingale property

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

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

Birth-death process

Suppose Y is a discrete-time birth-and-death process started at $y > 0$ and *absorbed at zero*:

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

Y is a non-negative supermartingale and so $\lim_{n\to\infty} Y_n$ exists.

Y is a biased random walk with a single absorbing state at 0. Let $T = \inf\{n : Y_n = 0\}$; then $T < \infty$ a.s., and so the only possible limit for Y is 0.

- This is the discrete-time analogue of the $M/M/1$ queue seen in Lecture 1.
- Test understanding: show that Y is a supermartingale.

Birth-death process

Now let

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

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

Thus, recalling that $Y_0 = X_0 = y$ and using the martingale convergence theorem,

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

- Test understanding: show that X is a martingale.
- The convergence theorem tells us that, $\frac{\mu-\lambda}{\mu+\lambda}\,\mathbb{E}[T] = \mathbb{E}[Z] = \mathbb{E}[Z|\mathcal{F}_0] \leq X_0 = y$.
- Markov's inequality then implies that

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

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 $f(\theta; x)$. Suppose also that $\mathbb{E}[|\log(f(\theta; X_1))|] < \infty$. 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 $n \geq 1$.

The SLLN and Jensen's inequality show that

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

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

- Test understanding: check that this argument works.
- 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 $\theta = \theta_0$.
- 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.

Sequential hypothesis testing

In the setting above, suppose that we want to satisfy

$$\mathbb{P}(\text{reject } H_0 | H_0) \leq \alpha \quad \text{and} \quad \mathbb{P}(\text{reject } H_1 | H_1) \leq \beta\,.$$

How large a sample size do we need?

Let

$$T = \inf\{n : M_n \geq \alpha^{-1} \text{ or } M_n \leq \beta\}$$

and consider observing $X_1, \ldots, X_T$ and then rejecting H_0 iff $M_T \geq \alpha^{-1}$.

- Here H_0 is the hypothesis that $X_1, X_2, \ldots$ have density $f(\theta_0; \cdot)$ etc.
- (A slight variant of) this strategy was originally suggested by Wald.
- Note that $\mathbb{P}(T = \infty \mid H_0) = 0$, since we've seen that in this case $M_n \to 0$.

Sequential hypothesis testing continued

On the (random) time interval $[0, T]$, M is a bounded martingale, and so

$$\mathbb{E}[M_T] = \mathbb{E}[M_0] = 1$$

(where we are computing the expectation using $\theta = \theta_0$).

So

$$1 = \mathbb{E}[M_T] \geq \alpha^{-1}\, \mathbb{P}\big(M_T \geq \alpha^{-1} \,|\, \theta_0\big) = \alpha^{-1}\, \mathbb{P}(\textsf{reject } H_0 | H_0)\,.$$

Interchanging the roles of H_0 and H_1 we also obtain $\mathbb{P}(\textsf{reject } H_1 | H_1) \leq \beta$.

The attraction here is that on average, fewer observations are needed than for a fixed-sample-size test.

- Test understanding: check that $\mathbb{P}(\textbf{reject } H_1 | H_1) \leq \beta$.
- For $X \sim N(\theta, 1)$ with $H_0 : \theta = 0$ vs $H_1 : \theta = 0.4$ with $\alpha = 0.05$ and $\beta = 0.10$, the sequential test should need (roughly) 29 observations, whereas classical sample size calculations suggest 54. See (Williams 2001) for details and another example.

Ergodicity

and general state spaces

We have a well-understood theory for discrete state space Markov chains.

For example, recall that X is irreducible if for all states i and j it has a positive chance of visiting j at some positive time, if it is started at i. And X is positive recurrent if $\mathbb{E}[T_i|X_0 = i] < \infty$ for all states i.

But what if the state space is not discrete? How should we generalise these ideas? And how can we describe the speed of convergence to equilibrium? The purpose of this lecture is to give you a taster of what's possible.

Ergodicity

We already know that 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;
2. X is aperiodic;
3. X is positive-recurrent.

In this case, we call the chain ergodic.

What can we say quantitatively, in general, about the speed at which convergence to equilibrium occurs? And what if the state-space is not discrete?

- Irreducible: the state space of X cannot be divided into regions some of which are inaccessible from others;
- Aperiodic: the state space of X cannot be broken into regions that can only be accessed at different times;
- Positive recurrent: the mean time for X to return to its starting point is finite.

Measuring speed of convergence to equilibrium (I)

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

- Simple possibility: total variation distance. Let $\mathcal{X}$ be the state-space. For $A \subseteq \mathcal{X}$, find the maximum discrepancy between $\mathcal{L}\left(X_n | X_0 = x\right)(A) = \mathbb{P}(X_n \in A | X_0 = x)$ and $\pi(A)$:

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

- Alternative expression in the case of a discrete state-space:

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

- There are *many* other possible measures of distance.

- $\mathcal{L}\left(X_n | X_0 = x\right)(A)$ is the probability that X_n belongs to A, given that $X_0 = x$.

- Test understanding: in the definition of total variation distance, why is it not necessary to take $|\,\mathbb{P}(X_n \in A | X_0 = x) - \pi(A)|$? (Hint: consider $\mathbb{P}(X_n \in A^c | X_0 = x) - \pi(A^c)$.)

- Test understanding: prove the alternative expression for the total variation distance by considering $A = \{y : \mathbb{P}(X_n = y | X_0 = x) > \pi_y\}$.

- It is not even clear that total variation is the "best" notion of distance (if there is such a thing).

Measuring speed of convergence to equilibrium (II)

Definition

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 $\gamma \in (0, 1)$,

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

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.

- In fact, uniform ergodicity is a consequence of the apparently weaker assertion that as $n \to \infty$,

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

- Any *finite* ergodic Markov chain is automatically uniformly ergodic.

- Much depends on the size of C and on how small γ is. Typically theoretical estimates of C and γ are *very* conservative.

- 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)

Definition 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 $\gamma \in (0, 1)$ (not depending on the starting point),

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

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!

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:

- Proposals are normal, with mean given by the current location x, and fixed variance-covariance matrix.
- We use the Hastings ratio to accept/reject proposals.
- We end up with a Markov chain X which has a transition mechanism which mixes a density with staying at the starting 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 state-space theory cannot apply.

- Draw a picture. $\mathbb{P}(\text{hit } y | X_0 = x) = 0$ for all $y \neq x$.
- This example is actually quite well-behaved!
- Clearly the discrete-chain theory needs major rehabilitation if it is to be helpful in the continuous state space case!

Measures

A measure is a way of assigning sizes to subsets of a space. Think of probability measures, but not necessarily with total mass 1 (in fact, not necessarily with finite total mass).

- For example, length is a measure on $\mathbb{R}$, area is a measure on $\mathbb{R}^2$ and volume is a measure on $\mathbb{R}^3$. These are all special cases of Lebesgue measure on $\mathbb{R}^d$.
- And counting measure is a measure on the integers: if $A \subset \mathbb{Z}$ then $c(A) = |A|$.

These examples, and probability measures, are the only ones we will use in this course, but there are many more.

Non-examinable definition:

A measure ϕ on a space E is a function from (some) subsets of E to $\mathbb{R}$ such that:

- $\phi(A) \geq 0$, i.e. sets can only have positive measure;
- $\phi(\emptyset) = 0$, i.e. the empty set has measure zero;
- if $A_1, A_2, \ldots$ are *pairwise disjoint* then

$$\phi\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \phi(A_i)\,.$$

(There is some subtlety in deciding which subsets of E the measure ϕ is defined on, but we won't go into that.)

Irreducibility for general chains

- As we have seen, the discrete theory fails to apply directly even to well-behaved chains on non-discrete state-spaces. (The chain will never hit "most" points in the space, so we can't ask it to be irreducible in the discrete sense.)

- 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*.

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

(That is, letting $T_B = \min\{n \geq 1 : X_n \in B\}$, if $\phi(B) > 0$ we have $\mathbb{P}(T_B < \infty | X_0 = x) > 0$ for all x.)

- We are skating over the issue of periodicity, which is largely technical.
- Recall the example of 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.
- But it will hit any set of positive length with positive probability! So it is length-irreducible, or in technical jargon, Lebesgue-irreducible.
- Indeed, if a set has $\mathrm{Leb}(B) > 0$, then $\mathbb{P}(N(0,1) \in B - x) > 0$ and so $\mathbb{P}(X_1 \in B | X_0 = x) > 0$.
- Irreducible chains on a discrete state-space are c-irreducible where c is counting measure. So ϕ-irreducibility generalizes the discrete notion of irreducibility.

Example of ϕ-irreducibility

- Let $U_1, U_2, \ldots$ be i.i.d. $U(-1,1)$ random variables, and define a Markov chain X on the state space $(-1,1)$ by setting

$$X_{n+1} = \frac{X_n + U_{n+1}}{2}.$$

- Claim that X is ϕ-irreducible, where ϕ is uniform on $(-1,1)$ (or Lebesgue measure - this is equivalent).

- Need to check that for any $x \in (-1,1)$ and for any interval (a,b) with $-1 \leq a < b \leq 1$, the chain X has positive chance of reaching (a,b) if begun at x.

- Idea of proof: show that

$$\mathbb{P}\Big(T_{[1-2^{-n},1)} \leq n+1 \,\Big|\, X_0 = x\Big) > 0 \quad \forall x \in (-1,1).$$

- It's enough to check intervals of the form $[1 - 2^{-n}, 1)$ for any n because (a) there is symmetry, so the same will be true for $(1, 1 - 2^{-n}]$ and (b) any interval (a, b) with $-1 \leq a < b \leq 1$ intersects either $(1, 1 - 2^{-n}]$ or $[1 - 2^{-n}, 1)$ for some $n \geq 0$.

- What if the state space was $\mathbb{R}$ instead of $(-1, 1)$?

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 a question about renewal processes, with renewals occurring each time the chain visits a specified state. (See lecture 2.)

We want to extend this idea by thinking in terms of renewals when visiting sets instead.

Regeneration and small sets (II)

Definition

Suppose that X is a ϕ-irreducible Markov chain. A set E of positive ϕ-measure is a small set for X if there is $\alpha \in (0,1)$ and a probability measure ν such that for all $x \in E$ the following minorisation condition is satisfied:

$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \nu(A) \quad \text{for all } A\,.$$

(The terminology "small" might be confusing. There is no reason why E necessarily has to be "small" in the traditional sense.)

Non-examinable, more general definition:

A set E of positive ϕ-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 minorisation condition is satisfied:

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

Example of a small set

- Recall the example from a few slides ago: let $U_1, U_2, \ldots$ be i.i.d. $U(-1,1)$ random variables, and define a Markov chain X on the state space $(-1,1)$ by setting $X_{n+1} = \frac{X_n + U_{n+1}}{2}$.
- We decided X is ϕ-irreducible, where ϕ is uniform on $(-1,1)$.
- Claim that $[-1/2, 1/2]$ is a small set with $\alpha = 1/2$ and ν the uniform distribution on $[-1/4, 1/4]$.
- Proof: for any x and any $A \subset (-1,1)$,

$$\mathbb{P}(X_1 \in A | X_0 = x) = \int_{A \cap [x/2 - 1/2, x/2 + 1/2]} dy$$

(check this!). If $x \in [-1/2, 1/2]$ then this is at least

$$\frac{1}{2} \int_{A \cap [-1/4, 1/4]} 2dy = \frac{1}{2} \nu(A).$$

Are there any other small sets for this chain? Is $[-3/4, 3/4]$ a small set? What about $(-1, 1)$?

Regeneration and small sets (III)

Why is this useful? Consider a small set E, so that for $x \in E$,

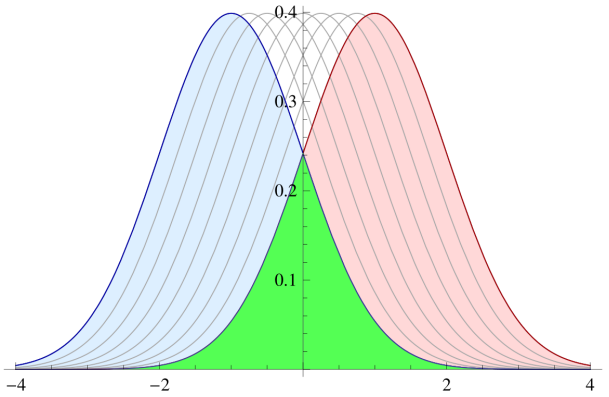
$$p(x, A) = \mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \nu(A) \quad \text{for all } A\,.$$

This means that, given $X_0 = x$, we can think of sampling X_1 as a two-step procedure. With probability α, sample X_1 from ν. With probability $1 - \alpha$, sample X_1 from the probability distribution $\frac{p(x,\cdot) - \alpha\nu(\cdot)}{1-\alpha}$.

Non- examinable: if instead we have a small set of lag k for $k > 1$, we can sub-sample X every k time-steps and then, every time it visits E, there is probability α that X forgets its entire past and starts again, using probability measure ν.

Regeneration and small sets (IV)

Consider the Gaussian random walk described earlier in the lecture. Any bounded set is small. For example, take $E = [-1, 1]$.



The green region represents the overlap of all the Gaussian densities centred at all points in E.

Let α be the area of the green region and let f be its upper boundary. Then $f(x)/\alpha$ is a probability density and, for any $x \in E$,

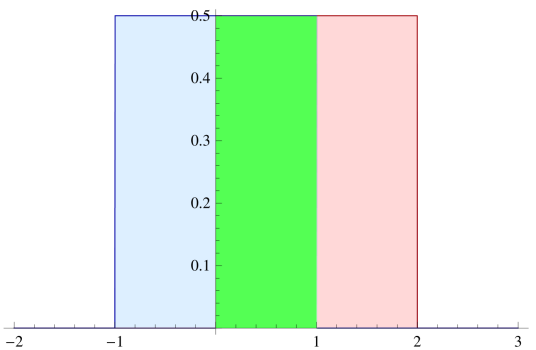
$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \int_A \frac{f(x)}{\alpha} dx = \alpha \nu(A).$$

Regeneration and small sets (V)

Let X be a random walk with transition density

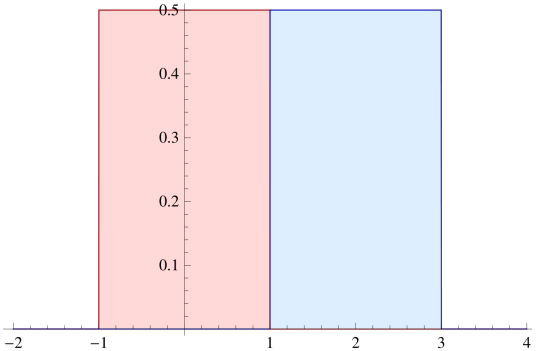
$$p(x,y) = \frac{1}{2}\mathbf{1}_{|x-y|<1}.$$

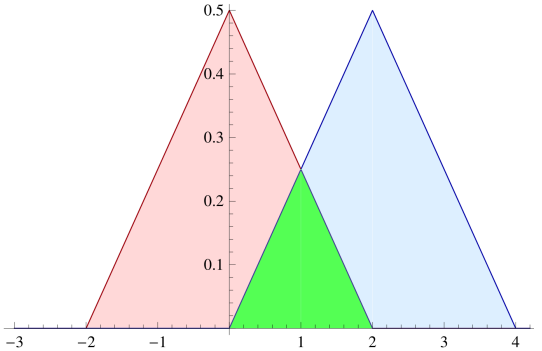
Consider the set $[0,1]$: this is small, with $\alpha = 1/2$ and ν the uniform distribution on $[0,1]$.



Take the same random walk as the previous slide,with transition density $p(x, \mathrm{d}\, y) = \frac{1}{2}\mathbf{1}_{|x-y|<1}$.

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





(Non-examinable.)

Small sets would not be very interesting except that:

1. All ϕ-irreducible Markov chains X possess small sets of some lag;
2. Consider chains X with continuous transition density kernels. They possess *many* small sets of lag 1 (This is very useful! Test understanding: try seeing why this is 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;
4. Given just one small set, X can be represented using a chain which has a single recurrent atom.

In a word, small sets discretize Markov chains.

Example: a random walk on $[0,1]$

Transition density

$$p(x,y) = 2\min\left\{\frac{y}{x}, \frac{1-y}{1-x}\right\},$$

i.e. transitions from x according to a triangular density around x.

For any $A \subset [0,1]$ and all $x \in [0,1]$,

$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \frac{1}{2}\nu(A),$$

where $\nu(A) = 2\int_A \min\{x, 1-x\}\,\mathrm{d}\,x$. Hence, the whole state-space is a small set.

Detailed balance equations (in terms of densities):
$\pi(x)p(x,y) = \pi(y)p(y,x)$; we can spot an invariant probability density, $\pi(x) = 6x(1-x)$.

Regeneration and small sets (VII)

Here is an indication of how we can use the discretization provided by small sets.

Theorem

Suppose that π is a stationary distribution for X. Suppose that the whole state-space $\mathcal{X}$ is a small set i.e. there exists a probability measure ν and $\alpha \in (0,1)$ such that

$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \nu(A) \text{ for all } x \in \mathcal{X}.$$

Then

$$\sup_{x \in \mathcal{X}} \text{dist}_{\text{TV}}(\mathcal{L}(X_n | X_0 = x), \pi) \leq (1-\alpha)^n$$

and so X is uniformly ergodic.

Non-examinable sketch proof

We use a coupling argument: we create two copies, $X^{(1)}$ and $X^{(2)}$, of the Markov chain X such that $X_0^{(1)} = x$ and $X_0^{(2)} \sim \pi$. We construct them so that the two meet after a random time and, since $X^{(2)}$ is stationary, this will entail that $X^{(1)}$ converges.

Given that $X_n^{(1)} = x_1$ and $X_n^{(2)} = x_2$, to generate step $n+1$:

(a) with probability α, choose the next position for both chains according to distribution ν, and then run them so that $X^{(1)} = X^{(2)}$ forever after.

(b) with probability $1-\alpha$, sample $X_{n+1}^{(1)}$ from $\frac{1}{1-\alpha}(P(x_1, \cdot) - \alpha\nu)$ and sample $X_{n+1}^{(2)}$ independently from $\frac{1}{1-\alpha}(P(x_2, \cdot) - \alpha\nu)$.

Non-examinable sketch proof (cont.)

Then $X^{(1)}$ and $X^{(2)}$ are Markov chains with the same transition mechanism P (Test understanding: check this!) and $X_n^{(2)} \sim \pi$ for all $n \geq 0$.

Let T be the first time that option (a) is used. T has a Geometric distribution with success probability α and so $\mathbb{P}(T > k) = (1-\alpha)^k$. Since the total variation distance between two probability measures is always bounded above by 1, we have

$$\mathrm{dist}_{\mathrm{TV}}(\mathcal{L}(X_n^{(1)} | X_0^{(1)} = x), \pi) \leq \mathbb{P}(T > n) = (1-\alpha)^n.$$

Note that this also proves that π is unique (otherwise the same argument produces a contradiction).

Foster-Lyapunov criteria

Recap from previous lecture

Let X be a Markov chain and let $T_B = \inf\{n \geq 1 : X_n \in B\}$. Let ϕ be a measure on the state-space.

- X is ϕ-irreducible if $\mathbb{P}(T_B < \infty | X_0 = x) > 0$ for all x whenever $\phi(B) > 0$.
- A set E of positive ϕ-measure is a small set for X if there is $\alpha \in (0, 1)$ and a probability measure ν such that for all $x \in E$,

$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \nu(A) \quad \text{for all } A\,.$$

Renewal and regeneration

Suppose C is a small set for X: for $x \in C$,

$$\mathbb{P}(X_1 \in A | X_0 = x) \geq \alpha \nu(A)\,.$$

If we can ensure that X hits C with probability 1, then we can identify regeneration events: X regenerates at $x \in C$ with probability α and then makes a transition with distribution ν; otherwise it makes a transition with distribution $\frac{p(x,\cdot)-\alpha\nu(\cdot)}{1-\alpha}$.

The regeneration events occur as a renewal sequence. Set

$$p_j \;=\; \mathbb{P}(\text{next regeneration at time } j \mid \text{regeneration at time } 0).$$

If the renewal sequence is non-defective (i.e. $\sum_j p_j = 1$) and positive-recurrent (i.e. $\sum_j j p_j < \infty$) then there exists a stationary version. This is the key to equilibrium theory whether for discrete or continuous state-space.

- If we have a small set with lag $k > 1$ then sub-sample every k steps!
- Non-defective: So there will always be a next regeneration.
- Positive-recurrent: So mean time to next regeneration is finite.

Positive recurrence

Here is the Foster-Lyapunov criterion for positive recurrence of a ϕ-irreducible Markov chain X on a state-space $\mathcal{X}$.

Theorem

Suppose that there exist a function $\Lambda : \mathcal{X} \to [0, \infty)$, strictly positive constants a, b, c, and a small set $C = \{x : \Lambda(x) \leq c\} \subseteq \mathcal{X}$ such that

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \Lambda(X_n) - a + b\mathbf{1}_{X_n \in C}.$$

Then $\mathbb{E}[T_A|X_0 = x] < \infty$ for all $x \in \mathcal{X}$ and any A such that $\phi(A) > 0$ (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.

- 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!
- Λ is called the "scale function". Choosing a good Λ is a real art form, but once we have Λ the choice of a, b, c is mechanical. In fact we can always renormalise Λ to get $a = 1$.
- In fact, if the criterion holds then it can be shown that *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.

Example of a positive recurrent chain

- Recall an example from the previous lecture: let $U_1, U_2, \ldots$ be i.i.d. $U(-1,1)$ random variables, and define a Markov chain X on the state space $(-1,1)$ by setting $X_{n+1} = \frac{X_n + U_{n+1}}{2}$.
- We decided X is ϕ-irreducible, where ϕ is uniform on $(-1,1)$, and that $[-1/2, 1/2]$ is a small set with $\alpha = 1/2$ and ν the uniform distribution on $[-1/4, 1/4]$.
- Let $\Lambda(x) = x^2$ and take $c = 1/4$ so $C = \{x : \Lambda(x) \leq c\} = [-1/2, 1/2]$.
- Exercise 1: $\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] = \frac{X_n^2}{4} + \frac{1}{12}$.
- Exercise 2: check that if we choose $a = 5/48$ and $b = 3/16$ then we satisfy the Foster-Lyapunov criterion for positive recurrence.

Solution to Exercise 2: if $X_n \notin [-1/2, 1/2]$ then

$$\frac{X_n^2}{4} = X_n^2 - \frac{3}{4} X_n^2 \leq X_n^2 - \frac{3}{16}$$

so whenever $X_n \notin [-1/2, 1/2]$ we have

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \Lambda(X_n) - \frac{5}{48}$$

and this is why we choose $a = -5/48$.

Then if $X_n \in [-1/2, 1/2]$ we have

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] = \frac{X_n^2}{4} + \frac{1}{12} \leq X_n^2 + \frac{1}{12} = X_n^2 - \frac{5}{48} + \frac{3}{16}$$

which is why we choose $b = 3/16$.

Positive recurrence

Before we give a sketch proof, we recall the theorem:

Theorem: Foster-Lyapunov criterion for positive recurrence
Suppose that there exist a function $\Lambda : \mathcal{X} \to [0, \infty)$, strictly positive constants a, b, c, and a small set $C = \{x : \Lambda(x) \leq c\} \subseteq \mathcal{X}$ such that

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \Lambda(X_n) - a + b\mathbf{1}_{X_n \in C}.$$

Then $\mathbb{E}[T_A|X_0 = x] < \infty$ for all $x \in \mathcal{X}$ and any A such that $\phi(A) > 0$ (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.

Sketch of proof (non-examinable)

1. Suppose $X_0 \notin C$. Then $Y_n = \Lambda(X_n) + an$ is non-negative supermartingale up to time $T_C = \inf\{m \geq 1 : X_m \in C\}$: if $T_C > n$ then

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] \leq (\Lambda(X_n) - a) + a(n+1) = Y_n \,.$$

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

2. So $\mathbb{P}(T_C < \infty) = 1$ (otherwise $\Lambda(X_n) > c$, $Y_n > c + an$ and so $Y_n \to \infty$). Moreover, $\mathbb{E}[Y_{T_C}|X_0] \leq \Lambda(X_0)$ (martingale convergence theorem) so $a\,\mathbb{E}[T_C|X_0] \leq \Lambda(X_0)$.

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

4. ϕ-irreducibility: X has a positive chance of hitting A between regenerations in C. Hence, $\mathbb{E}[T_A|X_0] < \infty$.

- 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 $\mathbb{E}[T_C|X_0 = x] < \infty$ for all starting points x, where C is some small set.

The Foster-Lyapunov criterion for positive recurrence follows for $\Lambda(x) = \mathbb{E}[T_C|X_0 = x]$ as long as $\mathbb{E}[T_C|X_0 = x]$ is bounded for $x \in C$.

- ϕ-irreducibility follows automatically from the condition.
- To see why the converse holds, note that

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

  where $b = \sup_{x \in C} \mathbb{E}[T_C | X_0 = x] < \infty$.
- 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.

Example: general reflected random walk

Let

$$X_{n+1} = \max\{X_n + Z_{n+1}, 0\}\,,$$

for $Z_1, Z_2, \ldots$ i.i.d. with continuous density $f(z)$, $\mathbb{E}[Z_1] < 0$ and $\mathbb{P}(Z_1 > 0) > 0$. Then

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

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

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

(a) $\mathbb{E}[Z_1] < 0$ so by SLLN $\frac{1}{n}(Z_1 + \ldots + Z_n) \to -\infty$, so X hits 0 for any X_0.

$\mathbb{P}(Z_1 > 0) > 0$ so $f(z) > 0$ for $a < z < a(1 + \frac{1}{m})$, some a, $m > 0$. So if $X_0 = 0$ then the 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 $A \cap (na, na + \frac{n}{m}a)$ for some $n \geq m$ is of positive measure so $\mathbb{P}(X \text{ hits } A | X_0 = 0) > 0$.

$\mathbb{E}[Z_1] < 0$ 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 $\Lambda(x) = x$.

Geometric ergodicity

Here is the Foster-Lyapunov criterion for geometric ergodicity of a ϕ-irreducible Markov chain X on a state-space $\mathcal{X}$.

Theorem

Suppose that there exist a function $\Lambda : \mathcal{X} \to [1, \infty)$, positive constants $\gamma \in (0, 1)$, b, $c \geq 1$, and a small set $C = \{x : \Lambda(x) \leq c\} \subseteq \mathcal{X}$ with

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \gamma\Lambda(X_n) + b\mathbf{1}_{X_n \in C}.$$

Then $\mathbb{E}\left[\gamma^{-T_A}|X_0 = x\right] < \infty$ for any A such that $\phi(A) > 0$ (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.

Uniform ergodicity follows if the function Λ is bounded above.

- 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!
- 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 $\mathbb{E}[\Lambda(X_{n+m})|\mathcal{F}_n]$.

Example of geometric ergodicity

- Return to our favourite example on $(-1, 1)$ where $X_{n+1} = \frac{X_n + U_{n+1}}{2}$ with $U_{n+1} \sim U(-1, 1)$.
- This time use $\Lambda(x) = e^{|x|}$. Previously showed that $[-1/2, 1/2]$ is small; in fact $[-3/4, 3/4]$ is also small (and so is $[-\alpha, \alpha]$ for any $0 < \alpha < 1$).
- Use $C = [-3/4, 3/4]$ i.e. $c = e^{3/4}$.
- Exercise: show that X satisfies the Foster-Lyapunov criterion for geometric ergodicity with $\gamma = e^{-3/8} \cdot 2(e^{1/2} - 1)$ and $b = (2(e^{1/2} - 1) - \gamma)e^{3/4}$ (or some other γ and b if easier).

Solution to exercise:

$$\begin{aligned}\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] &= \mathbb{E}\left[e^{|X_{n+1}|}\middle|\mathcal{F}_n\right] \\ &= \mathbb{E}\left[e^{|X_n+U_{n+1}|/2}\middle|\mathcal{F}_n\right] \\ &\leq e^{|X_n|/2}\,\mathbb{E}\left[e^{|U_{n+1}|/2}\right] \\ &= e^{|X_n|/2}\cdot 2(e^{1/2}-1).\end{aligned}$$

If $X_n \notin [-3/4, 3/4]$ then

$$e^{|X_n|/2} = e^{|X_n|}\cdot e^{-|X_n|/2} \leq e^{|X_n|-3/8}$$

so

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \Lambda(X_n)e^{-3/8}\cdot 2(e^{1/2}-1).$$

We can therefore choose $\gamma = e^{-3/8}\cdot 2(e^{1/2}-1) \approx 0.89 < 1$.

Continued...

If $X_n \in [-3/4, 3/4]$ then

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \gamma\Lambda(X_n) + (2(e^{1/2} - 1) - \gamma)e^{3/4}.$$

We can therefore choose $b = (2(e^{1/2} - 1) - \gamma)e^{3/4}$.

Geometric ergodicity

Before we give a sketch proof, we recall the theorem:

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

$$\mathbb{E}[\Lambda(X_{n+1})|\mathcal{F}_n] \leq \gamma\Lambda(X_n) + b\mathbf{1}_{X_n \in C}.$$

Then $\mathbb{E}\left[\gamma^{-T_A}|X_0 = x\right] < \infty$ for any A such that $\phi(A) > 0$ (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.

Uniform ergodicity follows if the function Λ is bounded above.

Sketch of proof (non-examinable)

1. Suppose $X_0 \notin C$. Then $Y_n = \Lambda(X_n)/\gamma^n$ defines a non-negative supermartingale up to time T_C: if $T_C > n$ then

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

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

2. So $\mathbb{P}(T_C < \infty) = 1$ (otherwise $\Lambda(X_n) > c$ and so $Y_n > c/\gamma^n$ does not converge). Moreover,

$$\mathbb{E}\left[\gamma^{-T_C}|X_0\right] \leq \Lambda(X_0).$$

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

4. From ϕ-irreducibility there is a positive chance of hitting A between regenerations in C. Thus $\mathbb{E}\left[\gamma^{-T_A}|X_0\right] < \infty$.

- Geometric/uniform ergodicity follows by a coupling argument which we 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!

A converse

Suppose, on the other hand, that $\mathbb{E}\left[\gamma^{-T_C}|X_0\right] < \infty$ for all starting points X_0 (and fixed $\gamma \in (0,1)$), where C is some small set and T_C is the first time for X to return to C.

The Foster-Lyapunov criterion for geometric ergodicity then follows for $\Lambda(x) = \mathbb{E}\left[\gamma^{-T_C}|X_0 = x\right]$ as long as $\mathbb{E}\left[\gamma^{-T_C}|X_0 = x\right]$ is bounded for $x \in C$.

Markov's 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.

Strikingly, for Harris-recurrent Markov chains (a ϕ-irreducible chain X is Harris recurrent if, for all starting points x and any set B with $\phi(B) > 0$, when started at x the chain X hits B eventually with probability 1) the existence of a geometric Foster-Lyapunov condition is equivalent to the property of geometric ergodicity.

Example: reflected simple asymmetric random walk

Let $X_{n+1} = \max\{X_n + Z_{n+1}, 0\}$, where $Z_1, Z_2, \ldots$ are i.i.d. with $\mathbb{P}(Z_1 = 1) = p < 1/2$ and $\mathbb{P}(Z_1 = -1) = 1 - p = q$.

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

(b) Foster-Lyapunov criterion for positive recurrence applies, using $\Lambda(x) = x$ and $C = \{0\}$:

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

(c) Foster-Lyapunov criterion for geometric ergodicity applies, using $\Lambda(x) = e^{ax}$ and $C = \{0\} = \Lambda^{-1}(\{1\})$.

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

(b) Test understanding: check this calculation!

(c) Test understanding: show this works! Note that

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

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.

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?

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.

The theory for this is developing fast, but many questions remain open. In this section we describe a a few interesting results, and look in detail at a specific easy example.

Cutoff: first example

Consider repeatedly shuffling a pack of n cards using a riffle shuffle.

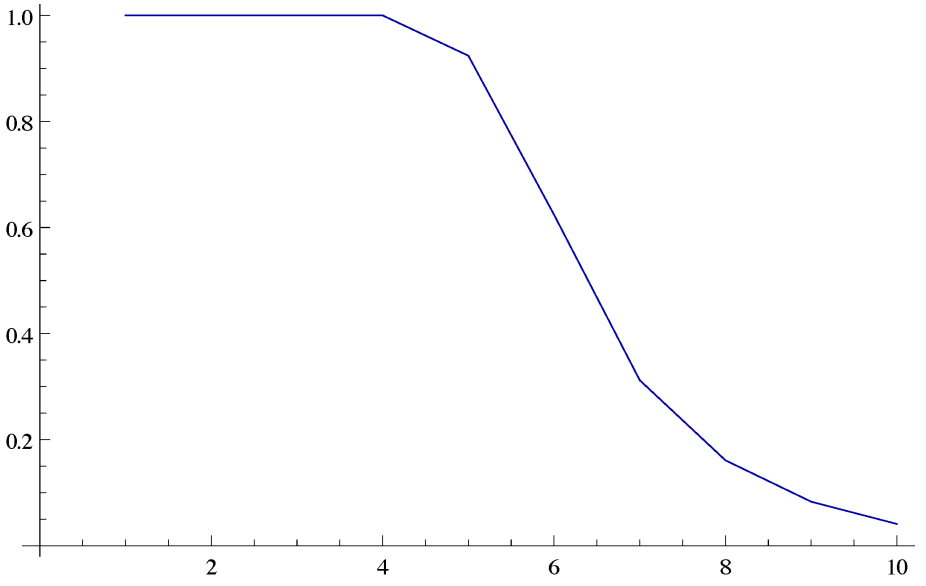
Write P_n^t for the distribution of the cards at time t.

This shuffle can be viewed as a random walk on S_n with uniform equilibrium distribution π_n.

- S_n is the symmetric group on n elements. The fact that the equilibrium distribution is uniform means that repeated shuffles really *do* mix up the cards as we would wish!

Cutoff: first example

With $n = 52$, the total variation distance $\mathrm{dist}_{\mathrm{TV}}(P_n^t, \pi_n)$ of P_n^t from equilibrium decreases like this:



- Notice that it takes about 7 shuffles for the total variation distance to get small — the decay beyond this point is pretty fast, and so one could argue that there's not a lot of point in shuffling more than this.

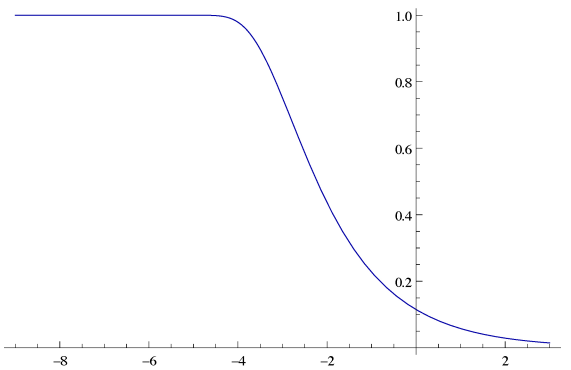
Riffle shuffle: sharp result (Bayer and Diaconis, 1992)

Let $\tau_n(\theta) = \frac{3}{2}\log_2 n + \theta\,.$

Then

$$\mathrm{dist}_{\mathrm{TV}}(P_n^{\tau_n(\theta)}, \pi_n) = 1 - 2\Phi\left(\frac{-2^{-\theta}}{4\sqrt{3}}\right) + O(n^{-1/4})\,.$$

As a function of θ this looks something like:

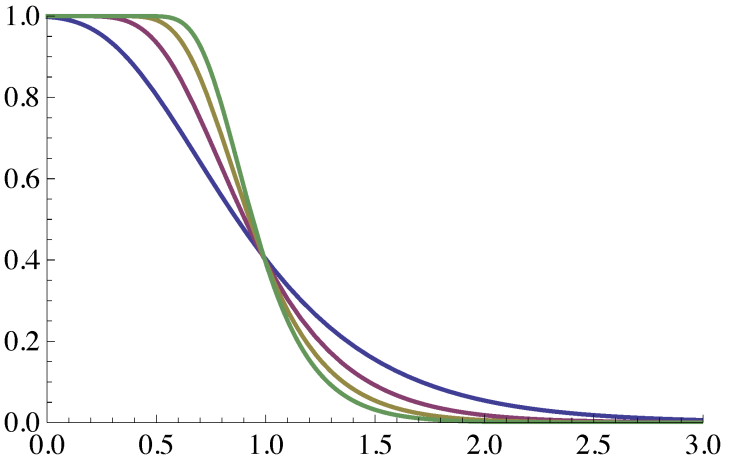


So as n gets large, convergence to uniform happens quickly after about $(3/2)\log_2 n$ shuffles (≈ 7 when $n = 52$).

- Here Φ is the standard normal distribution function.

Cutoff: the general picture

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.

- This says that convergence really does take place quickly around the cutoff time τ_n.
- The speed of convergence of the re-normalised graphs to a step function depends on the size of the 'time window' over which the cutoff takes place.

Cutoff: more examples

There are *many* examples of this type of behaviour:

$\mathcal{X}_n$	**Chain**	τ_n
S_n	Riffle shuffle	$\frac{3}{2}\log_2 n$
S_n	Top-to-random	??
S_n	Random transpositions	??
$\mathbb{Z}_2^n$	Symmetric random walk	$\frac{1}{4}n\log n$

- Methods of proving cutoff include coupling theory, eigenvalue-analysis, group representation theory, ...

In general, expect cutoff when there are large numbers of "second" eigenvalues of the transition matrix P.

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, due to David Aldous (Levin, Peres and Wilmer, 2009). So the conjecture needs to be refined!

An example in more detail: the top-to-random shuffle

Let us show how to prove cutoff in a very simple case: the top-to-random shuffle. This is another random walk X on the symmetric group S_n: each 'shuffle' consists of removing the top card and replacing it into the pack uniformly at random.

Hopefully it's not too hard to believe that the equilibrium distribution of X is again the *uniform distribution* π_n on S_n (i.e., $\pi_n(\sigma) = 1/n!$ for all permutations $\sigma \in S_n$).

Theorem (Aldous & Diaconis, 1986)
Let $\tau_n(\theta) = n \log n + \theta n$. Then

1. $\mathrm{dist}_{\mathrm{TV}}(P_n^{\tau_n(\theta)}, \pi_n) \leq e^{-\theta}$ for $\theta \geq 0$ and $n \geq 2$;
2. $\mathrm{dist}_{\mathrm{TV}}(P_n^{\tau_n(\theta)}, \pi_n) \to 1$ as $n \to \infty$, for $\theta = \theta(n) \to -\infty$.

- Note that this random walk is certainly not reversible!
- Test understanding: prove that π_n really is stationary for X, and check that the distribution of X_k really will converge to π_n as $k \to \infty$.
- This theorem shows that there is a cutoff (in total variation distance) at time $n \log n$. If we do an extra θn shuffles, with $\theta \geq 0$, then the first part of the theorem says that the distance from stationarity decreases exponentially fast in θ. Alternatively, if we perform only $n \log n + \theta n$ shuffles, with $\theta < 0$, then as $\theta \to -\infty$ the distance between $P_n^{\tau_n(\theta)}$ and π_n tends to 1. So as n gets large the time taken for the pack to randomise concentrates more and more around $n \log n$.

Strong uniform times

Recall from earlier lectures that a stopping time is a non-negative integer-valued random variable T, with $\{T \leq k\} \in \mathcal{F}_k$ for all k. Let X be a random walk on a group G, with uniform equilibrium distribution π.

Definition

A strong uniform time T is a stopping time such that for each $k < \infty$ and $\sigma \in G$,

$$\mathbb{P}(X_k = \sigma \,|\, T = k) = \pi(\sigma) = 1/|G| \,.$$

Strong uniform times (SUT's) are useful for the following reason...

- If you're not familiar with the idea of a group, don't worry! Just think of G as the set of possible permutations of n cards.
- Note that this definition is equivalent to saying that $X_T \sim \pi$ and that X_T is independent of T.
- Test understanding: show that the definition is equivalent to

$$\mathbb{P}(X_k = \sigma, T = k) = \pi(\sigma)\,\mathbb{P}(T = k),$$

or to

$$\mathbb{P}(X_k = \sigma \,|\, T \leq k) = \pi(\sigma).$$

Lemma (Aldous & Diaconis, 1986)

Let X be a random walk on a group G, with uniform stationary distribution π, and let T be a SUT for X. Then for all $k \geq 0$,

$$\mathrm{dist}_{\mathrm{TV}}(P^k, \pi) \leq \mathbb{P}(T > k)\,.$$

Proof.

For any set $A \subseteq G$,

$$\begin{aligned}
\mathbb{P}(X_k \in A) &= \sum_{j \leq k} \mathbb{P}(X_k \in A, T = j) + \mathbb{P}(X_k \in A, T > k) \\
&= \sum_{j \leq k} \pi(A)\,\mathbb{P}(T = j) + \mathbb{P}(X_k \in A \,|\, T > k)\,\mathbb{P}(T > k) \\
&= \pi(A) + \big(\mathbb{P}(X_k \in A \,|\, T > k) - \pi(A)\big)\,\mathbb{P}(T > k)\,.
\end{aligned}$$

So $|P^k(A) - \pi(A)| \leq \mathbb{P}(T > k)$.

- This lemma, and the definition of SUT's, can be generalised to cases where the stationary distribution of X is not the uniform distribution. In this case, T is called a strong *stationary* time.)
- Note where we have used in the proof the fact that T is a SUT...
- Recall the definition of total variation distance to see why this implies the required result!

Back to shuffling: the upper bound

Consider the card originally at the bottom of the deck (suppose for convenience that it's $Q\heartsuit$). Let

- $T_1 =$ time until the 1st card is placed below $Q\heartsuit$;
- $T_2 =$ time until a 2nd card is placed below $Q\heartsuit$;
- $\ldots$
- $T_{n-1} =$ time until $Q\heartsuit$ reaches the top of the pack.

Then note that:

- at time T_2, the 2 cards below $Q\heartsuit$ are equally likely to be in either order;
- at time T_3, the 3 cards below $Q\heartsuit$ are equally likely to be in any order;
- $\ldots$

... so at time T_{n-1}, the $n-1$ cards below $Q\heartsuit$ are uniformly distributed.

Hence, at time $T = T_{n-1} + 1$, $Q\heartsuit$ is inserted uniformly at random, and now the cards are *all* uniformly distributed!

Since T is a SUT, we can use it in our Lemma to upper bound the total variation distance between π_n and the distribution of the pack at time k.

Note first of all that

$$T = T_1 + (T_2 - T_1) + \cdots + (T_{n-1} - T_{n-2}) + (T - T_{n-1}),$$

and that

$$T_{i+1} - T_i \overset{\text{ind}}{\sim} \mathsf{Geom}\left(\frac{i+1}{n}\right).$$

- Convince yourself that T really is a SUT! (An inductive argument along the lines of the previous slide should suffice.)

- Instead of considering T, we could instead work with $T^* =$ the time at which the card originally *second* from bottom is inserted into the pack for the first time. Show that this is also a SUT.

  (Clearly this is a faster SUT than the time T that we're using. In fact, T^* is a *fastest SUT*:

  $$\mathbb{P}(T^* > k) \leq \mathbb{P}\big(\tilde{T} > k\big)$$

  for all $k \geq 0$ and for *any other* SUT $\tilde{T}$!)

- Test understanding: ensure that you understand why

$$T_{i+1} - T_i \overset{\text{ind}}{\sim} \mathsf{Geom}\left(\frac{i+1}{n}\right),$$

i.e., that

$$\mathbb{P}(T_{i+1} - T_i = k) = \left(\frac{i+1}{n}\right)\left(1 - \frac{i+1}{n}\right)^{k-1}, \quad k \geq 1\,.$$

We can find the distribution of T by turning to the coupon collector's problem. Consider a bag with n distinct balls - keep sampling (with replacement) until each ball has been seen at least once.

Let $W_i =$ number of draws needed until i distinct balls have been seen. Then

$$W_n = (W_n - W_{n-1}) + (W_{n-1} - W_{n-2}) + \cdots + (W_2 - W_1) + W_1 \,,$$

where

$$W_{i+1} - W_i \overset{\text{ind}}{\sim} \text{Geom}\left(\frac{n-i}{n}\right) .$$

Thus, $T \overset{d}{=} W_n$.

This is a classic probability problem; maybe you have seen it before!

Test understanding: check that the distribution of T really does agree with that of W_n.

Now let A_d be the event that ball d has not been seen in the first k draws.

$$\begin{aligned}\mathbb{P}(W_n > k) = \mathbb{P}\big(\cup_{d=1}^n A_d\big) &\leq \sum_{d=1}^n \mathbb{P}(A_d) \\ &= n\left(1 - \frac{1}{n}\right)^k \leq ne^{-k/n}.\end{aligned}$$

Plugging in $k = \tau_n(\theta) = n\log n + \theta n$, we get

$$\mathbb{P}(W_n > \tau_n(\theta)) \leq e^{-\theta}.$$

Now use the fact that T and W_n have the same distribution, the important information that T is a SUT for the chain, and the Lemma above to deduce part 1 of our cutoff theorem.

Test understanding: check that you understand every step of this argument!

The lower bound

To prove lower bounds of cutoffs, a frequent trick is to find a set B such that $|P_n^{\tau_n(\theta)}(B) - \pi_n(B)|$ is large, where $\tau_n(\theta)$ is now equal to $n \log n + \theta(n)n$, with $\theta(n) \to -\infty$.

So let

$$B_i = \{\sigma \,:\, \text{bottom } i \text{ original cards remain in original relative order}\}.$$

This satisfies $\pi_n(B_i) = 1/i!$. Furthermore, we can argue that, for any fixed i, with $\theta = \theta(n) \to -\infty$,

$$P_n^{\tau_n(\theta)}(B_i) \to 1 \quad \text{as } n \to \infty.$$

Therefore,

$$\mathrm{dist}_{\mathrm{TV}}(P_n^{\tau_n(\theta)}, \pi_n) \geq \max_i \left(P_n^{\tau_n(\theta)}(B_i) - \pi_n(B_i) \right) \to 1\,.$$

- This is often sufficient, thanks to the definition of total variation distance.
- Test understanding: show that $\pi_n(B_i) = 1/i!$.
- Note that, for any $k \geq 0$,

$$P_n^k(B_i) \geq \mathbb{P}(T - T_{i-1} > k)\,,$$

since $T - T_{i-1}$ is distributed as the time for the card initially i^{th} from bottom to come to the top and be inserted - if this has not occurred by time k then the original bottom i cards must still be in their original relative order.

- Test understanding: estimate $\mathbb{P}(T - T_{i-1} > \tau_n(\theta))$ using Chebychev's inequality applied to the random variable $T - T_{i-1}$. Now complete the proof of the lower bound!

Final comments...

So how does this shuffle compare to others?

$\mathcal{X}_n$	**Chain**	τ_n
S_n	Top-to-random	$n \log n$
S_n	Riffle shuffle	$\frac{3}{2}\log_2 n$
S_n	Random transpositions	$\frac{1}{2} n \log n$
S_n	Overhand shuffle	$\Theta(n^2 \log n)$

- So shuffling using random transpositions, or even the top-to-random shuffle, is *much* faster than the commonly used overhand shuffle!