APTS Applied Stochastic Processes

Christina Goldschmidt[1]
goldschm@stats.ox.ac.uk

(Slides originally produced by Wilfrid Kendall,
modified and extended by Stephen Connor)

[1]Department of Statistics, University of Oxford

APTS Warwick, 29th July–2nd August 2013

Two notions in probability

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

This module is intended to introduce students to two important notions in stochastic processes — reversibility and martingales — identifying the basic ideas, outlining the main results and giving a flavour of some significant ways in which these notions are used in statistics. These notes outline the content of the module; they

represent work-in-progress and will grow, be corrected, and be modified as time passes.

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

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

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

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.

Some useful texts (I)

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

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

1. Häggström (2002) "Finite Markov chains and algorithmic applications".
2. Grimmett and Stirzaker (2001) "Probability and random processes".
3. Breiman (1992) "Probability".
4. Norris (1998) "Markov chains".
5. Williams (1991) "Probability with martingales".

Some useful texts (II): free on the web

1. Doyle and Snell (1984) “Random walks and electric networks” available on web at
   `www.arxiv.org/abs/math/0001057`.
2. Kelly (1979) “Reversibility and stochastic networks” available on web at
   `http://www.statslab.cam.ac.uk/~frank/BOOKS/kelly_book.html`.
3. Kindermann and Snell (1980) “Markov random fields and their applications” available on web at
   `www.ams.org/online_bks/conm1/`.
4. Meyn and Tweedie (1993) “Markov chains and stochastic stability” available on web at
   `www.probability.ca/MT/`.
5. Aldous and Fill (2001) “Reversible Markov Chains and Random Walks on Graphs” *only* available on web at
   `www.stat.berkeley.edu/~aldous/RWG/book.html`.

Some useful texts (III): going deeper

1. Kingman (1993) “Poisson processes”.
2. Steele (2004) “The Cauchy-Schwarz master class”.
3. Aldous (1989) “Probability approximations via the Poisson clumping heuristic”.
4. Øksendal (2003) “Stochastic differential equations”.
5. Stoyan, Kendall, and Mecke (1987) “Stochastic geometry and its applications”.

Markov chains and reversibility

"People assume that time is a strict progression of cause to effect, but actually from a non-linear, non-subjective viewpoint, it's more like a big ball of wibbly-wobbly, timey-wimey . . . stuff."

The Tenth Doctor,

Doctor Who, in the episode "Blink", 2007

Detailed balance in a nutshell

Suppose we could solve (non-trivially) for π in $\pi_x p_{x,y} = \pi_y p_{y,x}$ (discrete-time) or $\pi_x q_{x,y} = \pi_y q_{y,x}$ (continuous-time).

In both cases, simple algebra then shows that π solves the equilibrium equations.

So on a prosaic level it is always worth trying this easy route; if the detailed balance equations are insoluble then revert to the more complicated equilibrium equations $\pi P = \pi$ and $\pi Q = 0$ respectively.

We will consider reversibility of Markov chains in both discrete and continuous time, the computation of equilibrium distributions for such chains, and discuss applications via some illustrative examples.

We will consider progressively more and more complicated Markov chains:

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

Simplest non-trivial example (I)

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

1. X is **irreducible** and **aperiodic**, so there is a unique equilibrium distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_k)$.
2. The **equilibrium equations** $\pi P = \pi$ are solved by $\pi_i = \frac{1}{k+1}$ for all i.
3. Consider X in equilibrium and **run backwards in time**. Calculation:
   ▸ ANIMATION
   $\mathbb{P}\left[X_{n-1} = x | X_n = y\right] = \pi_x \, \mathbb{P}\left[X_n = y | X_{n-1} = x\right] / \pi_y = \mathbb{P}\left[X_n = y | X_{n-1} = x\right]$ so here *by symmetry of the kernel* the equilibrium chain has the same transition kernel (so looks the same) whether run forwards or backwards.
4. We say X is **reversible** in equilibrium.

Simplest non-trivial example (II)

There is a computational aspect to this.

1. Even in more general cases, if the π_x depend on x then the above computations show that reversibility holds if an equilibrium distribution exists and the **detailed balance equations** hold: $\pi_x p_{x,y} = \pi_y p_{y,x}$.
2. Moreover, **if** one can solve for π_x in $\pi_x p_{x,y} = \pi_y p_{y,x}$ then it is easy to show that $\pi P = \pi$.
3. Consequently, 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.

Birth-death-immigration process

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

Definition

The birth-death-immigration process has transitions:

- birth ($x \to x + 1$ at rate λx);
- death ($x \to x - 1$ at rate μx);
- plus an extra immigration term ($x \to x + 1$ at rate α).

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

▸ ANIMATION

Equilibrium is easily derived from detailed balance:

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

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.*

$M/M/1$ queue

Here we have

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

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

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

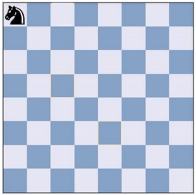
Hence, if a stable $M/M/1$ queue feeds into another stable $\cdot/M/1$ queue then in equilibrium the second queue on its own behaves as an $M/M/1$ queue in equilibrium.

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

Example (A mean knight's tour)

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

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



ANIMATION

Gibbs sampler for the Ising model

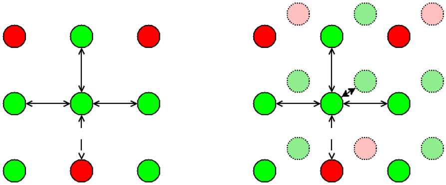
(I) Ising model

Pattern of spins $S_i = \pm 1$ on (finite fragment of) lattice (here i is a node of the lattice). Probability mass function:

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

or, if there is an external field,

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



Gibbs sampler for the Ising model

(II) Gibbs sampler (or heat-bath)

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.

- Suppose the current configuration is $\mathbf{s}$.
- Choose a site i in the lattice uniformly at random.
- 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.

We have transition probabilities

$$p(\mathbf{s}, \mathbf{s}^{(i)}) = \frac{1}{n^2}\,\mathbb{P}\left[\mathbf{S} = \mathbf{s}^{(i)} \middle| \mathbf{S} \in \{\mathbf{s}, \mathbf{s}^{(i)}\}\right].$$

Gibbs sampler for the Ising model

(II) Gibbs sampler (or heat-bath)

Simple general calculations show that

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

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

Gibbs sampler for the Ising model

(III) Detailed balance

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

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

  for all $\mathbf{s}$.

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

ANIMATION

Metropolis-Hastings

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

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

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

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

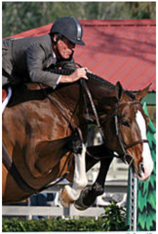
Since π satisfies detailed balance, π is an equilibrium distribution (if the chain converges to a unique equilibrium!).

Martingales

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

Frank Loesser,

Guys and Dolls musical, 1950, script



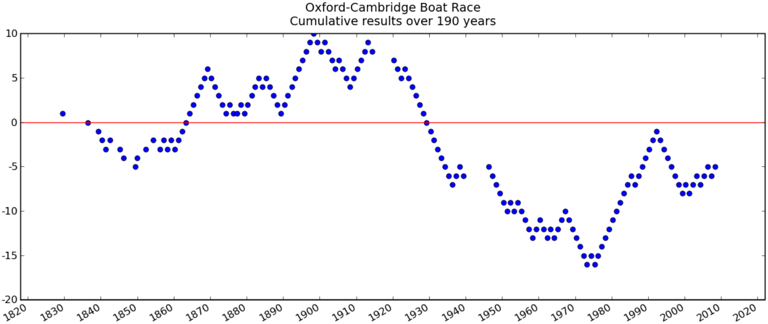
Martingales pervade modern probability

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

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

2. Simplest possible example: simple symmetric random walk $X_0 = 0$, X_1, X_2, ... The martingale property follows from independence and distributional symmetry of jumps.
3. For convenience and brevity, we often replace $\mathbb{E}\left[\ldots | X_n, X_{n-1}, \ldots\right]$ by $\mathbb{E}\left[\ldots | \mathcal{F}_n\right]$ and think of "conditioning on $\mathcal{F}_n$" as "conditioning on all events which can be determined to have happened by time n".

University Boat Race results over 190 years



Could this represent a martingale?

Thackeray's martingale

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

   The original sense is given in the OED: "a system in gambling which consists in doubling the stake when losing in the hope of eventually recouping oneself." The oldest quotation is from 1815 but the nicest is from 1854: Thackeray in *The Newcomes* I. 266 "You have not played as yet? Do not do so; above all avoid a martingale if you do."
3. Result of playing Thackeray's martingale system and stopping on first win: ▸ ANIMATION

   set fortune at time n to be M_n.
   If $X_1 = -1, \ldots, X_n = -n$ then
   $M_n = -1 - 2 - \ldots - 2^{n-1} = 1 - 2^n$, otherwise $M_n = 1$.

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

Definition of a martingale

Formally:

Definition

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

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

Supermartingales and submartingales

Two associated definitions.

Definition

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

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

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

Definition

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

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

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

Examples of supermartingales and 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.

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

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

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

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

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

An example of importance in finance

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

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

▸ ANIMATION

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

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, ... of X are observed at times 1, 2, ... The likelihood of θ at time n is
$$L(\theta; X_1, \ldots, X_n) = p(X_1, \ldots, X_n | \theta)\,.$$
2. If θ_0 is the "true" value then (computing expectation with $\theta = \theta_0$)
$$\mathbb{E}\left[\frac{L(\theta_1; X_1, \ldots, X_{n+1})}{L(\theta_0; X_1, \ldots, X_{n+1})}\middle| \mathcal{F}_n\right] = \frac{L(\theta_1; X_1, \ldots, X_n)}{L(\theta_0; X_1, \ldots, X_n)}.$$

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

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

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 $p_{x,y}$.

Martingales for Markov chains: harmonic functions

Call a function $f : S \to \mathbb{R}$ **harmonic** if

$$f(x) = \sum_{y \in S} f(y) p_{x,y} \text{ for all } x \in S.$$

We defined

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

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

Stopping times and martingale convergence

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

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

Non-obvious "no-look-ahead" condition

It turns out that the right random times to think about are those which "only depend on what we have seen so far".

Definition

A non-negative integer-valued random variable T is said to be a **stopping time** if, for all n, $\{T \leq n\}$ is determined by the information available at time n i.e. $\{T \leq n\} \in \mathcal{F}_n$.

Random walk example

Let X be a random walk begun at 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:

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

- Finally, T is typically "too big": so long as it is almost surely finite (so that, with probability 1, X does eventually exceed 10) then $X_T \geq 10$ and we deduce that $0 = \mathbb{E}\left[X_0\right] < \mathbb{E}\left[X_T\right]$.

T is almost surely finite if $\mathbb{E}\left[X_1\right] > 0$ or if $\mathbb{E}\left[X_1\right] = 0$ and $\mathbb{P}\left[X_1 > 0\right] > 0$.

Branching process 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.

- 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.

Events revealed by the time of a stopping time T

Suppose that T is a stopping time.

Definition

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

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

Definition

Random variables Z are said to be $\mathcal{F}_T$-**measurable** if events made up from them (e.g. $\{Z \leq z\}$) are in $\mathcal{F}_T$.

Optional stopping theorem

Theorem

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

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

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

Gambling: you shouldn't expect to win

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

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

▸ ANIMATION

Contrast Fleming (1953):

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

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, ... 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, and the optional stopping theorem can be applied to show

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

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.

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. and, moreover, $\mathbb{E}[Z|\mathcal{F}_n] \leq X_n$.

ANIMATION

Theorem

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

Birth-death process revisited

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

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

This is a non-negative supermartingale and so $\lim_{n\to\infty} Y_n$ exists. Since Y is a transient Markov chain with a single absorbing state at 0, the only possible limit is 0. Let $T = \inf\{n : Y_n = 0\}$; then $T < \infty$ a.s.

Birth-death process revisited

Now let

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

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 Fatou's lemma,

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

Likelihood revisited

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

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

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

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 $M_n \to 0$.

Counting and compensating

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

Simplest example: Poisson process

Definition

A continuous-time Markov chain N is a **Poisson process of rate** $\lambda > 0$ if $N_0 = 0$ and, conditional on $N_t = n$ (where $n \in \{0, 1, 2, \ldots\}$), the only possible transition is to $n + 1$, which occurs at rate λ.

For small Δt, we can think of this as

$$\mathbb{P}\left[N_{t+\Delta t} - N_t = 1\right] \approx \lambda \Delta t, \quad \mathbb{P}\left[N_{t+\Delta t} - N_t = 0\right] \approx 1 - \lambda \Delta t.$$

Simplest example: Poisson process

Theorem

If N is Poisson process of rate λ then $N_{t+s} - N_t$ is independent of past at time t, has same distribution as N_s, and

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

Times of transitions often referred to as incidents.

Poisson process directions

The Poisson process idea can be very considerably extended.

- Suppose that we still have a continuous-time Markov chain but that it's no longer time-homogeneous: instead of a constant rate of incidents, the rate varies with the time-parameter. For small Δt,

$$\mathbb{P}\left[N_{t+\Delta t} - N_t = 1\right] \approx \lambda(t)\Delta t,$$
$$\mathbb{P}\left[N_{t+\Delta t} - N_t = 0\right] \approx 1 - \lambda(t)\Delta t.$$

- The number of incidents $N_{t+s} - N_t$ occurring between times t and $t + s$ has Poisson distribution with parameter

$$\int_t^{t+s} \lambda(u)\,\mathrm{d}\,u.$$

- This is called an **inhomogeneous Poisson process**.

Poisson process directions

We can extend these ideas even further to general patterns of points in an arbitrary set S (so far, S has been the time-set $[0, \infty)$).

Take a (measurable) space S with a measure μ. For any subset $A \subset S$, let $N(A)$ denote the number of points falling in the set A.

- Suppose that $N(A)$ has a Poisson distribution with mean $\mu(A)$ for every $A \subset S$.
- Suppose that for any collection $A_1, A_2, \ldots, A_n$ of disjoint subsets of S, the numbers $N(A_1), N(A_2), \ldots, N(A_n)$ are independent.

Then we call N a **Poisson point process of intensity measure** μ.

Counting processes

We will stick to the one-dimensional situation and consider processes which “count” incidents.

Definition

A **counting process** is a continuous-time stochastic process (not necessarily Markov) which changes by single jumps of $+1$.

Example: renewal processes

Definition

Let $S_1, S_2, \ldots$ be i.i.d. random variables with density f on $(0, \infty)$. Let $T_0 = 0$ and, for $n \geq 1$, let

$$T_n = \sum_{i=1}^{n} S_i$$

and, for $t \geq 0$,

$$N_t = \#\{n \geq 1 : T_n \leq t\} = \sum_{n=1}^{\infty} \mathbb{1}_{\{T_n \leq t\}}\,.$$

Then $(N_t)_{t \geq 0}$ is a **renewal process**.

Renewal processes in Markov chains

One reason for studying renewal processes is that they come up very naturally in the context of Markov chains.

Suppose that X is a continuous-time Markov chain on a countable state-space S.

- Suppose that S is irreducible and recurrent. Let $i \in S$ be fixed.
- Set $X_0 = i$ and consider the times $T_1, T_2, \ldots$ at which the chain returns to state i.
- These are stopping times and, moreover, by the strong Markov property, the times $T_2 - T_1, T_3 - T_2, \ldots$ between successive returns are independent and have the same distribution as T_1.
- So the number N_t of returns to state i by time t forms a renewal process.

Continuous-time martingales

We introduced martingales in discrete time, but the concept makes perfect sense in continuous time too.

Definition

X is a **martingale** if $\mathbb{E}\left[|X_t|\right] < \infty$ for all $t \geq 0$ and for all $0 \leq s < t$,

$$\mathbb{E}\left[X_t | \mathcal{F}_s\right] = X_s.$$

$\mathcal{F}_s$ again represents the information we have about the process X up until time s.

Martingales for Poisson processes

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

- ("mean") $(N_t - \alpha t)_{t \geq 0}$ is a martingale;
- ("variance") $((N_t - \alpha t)^2 - \alpha t)_{t \geq 0}$ is a martingale.

Compensators for counting processes

Idea: try to subtract something from your process in order to turn it into a martingale.

Definition

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

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

Hazard rate

Compensators generalise the notion of hazard rate. Intuitively, this is the "infinitesimal rate of seeing an incident right now given that one hasn't seen anything since the last incident".

If the time between incidents is some random variable X, then time t after the last incident the hazard rate should be something like

$$\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \mathbb{P}\left[X \in [t, t+\epsilon) | X \geq t\right] = \frac{\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \mathbb{P}\left[X \in [t, t+\epsilon)\right]}{\mathbb{P}\left[X \geq t\right]}.$$

So if X has probability density f on $(0, \infty)$, the associated hazard rate is well-defined and equal to

$$h(t) = \frac{f(t)}{\int_t^\infty f(s) \, \mathrm{d}\, s} = \frac{f(t)}{\overline{F}(t)}.$$

Example: random sample of lifetimes

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

- Counting process $N_t = \#\{i : X_i \leq t\} = \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}$ increases by $+1$ jumps in continuous time.
- Observe:
  - $N_t - \int_0^t h(s)(n - N_s) \, \mathrm{d}\, s$ is a martingale.
  - $(N_t - \int_0^t h(s)(n - N_s) \, \mathrm{d}\, s)^2 - \int_0^t h(s)(n - N_s) \, \mathrm{d}\, s$ is a martingale.

Example: pure birth process

If the pure birth process N makes transitions $n \to n+1$ at rate λn then

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

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

Variance of compensated counting process

The above expression of variance type holds more generally.

Theorem

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

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

Counting processes and Poisson processes

The compensator of a counting process can be used to tell whether or not the counting process is Poisson.

Theorem

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

Counting processes and Poisson processes

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

Theorem

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

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

Compensators and likelihoods

Here is an even bigger pay-off.

Theorem

Suppose N is a counting process with compensator $\Lambda_t = \int_0^t \ell(s)\,\mathrm{d}\,s$. Then its likelihood with respect to a unit-rate Poisson point process over the time interval $[0, t]$ is proportional (for fixed t) to

$$\left(\prod_{i=1}^{N_t} \ell(T_i)\right) \exp\left(-\int_0^t \ell(s)\,\mathrm{d}\,s\right),$$

where $T_1, T_2, \ldots$ are the incident times of N.

ℓ needs to be left-continuous for this to work.

Compensation of population processes

Suppose a birth-death-immigration process X makes transitions $n \to n+1$ at rate $\lambda n + \alpha$ and $n \to n-1$ at rate μn. Then

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

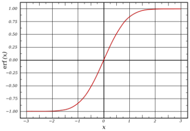
(But we now need something other than the compensator to convert $(X_t - \int_0^t ((\lambda - \mu) X_s + \alpha) \, \mathrm{d}\, s)^2$ into a martingale.)

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

Central Limit Theorem

"Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation"

Lippmann in letter to Poincaré,
quoted in E. T. Whittaker and G. Robinson,
Normal Frequency Distribution. Ch. 8 in
The Calculus of Observations:
A Treatise on Numerical Mathematics, 1967.



Notions of convergence

Definition

Random variables X_n are said to **converge in distribution** to a random variable Y (or its distribution) if

$$\mathbb{P}\left[X_n \leq x\right] \to \mathbb{P}\left[Y \leq x\right]$$

whenever the function $F(x) = \mathbb{P}\left[Y \leq x\right]$ is continuous at x.

Some other notions of convergence:

- **Convergence almost surely**: $\mathbb{P}\left[X_n \to Y\right] = 1$.
- **Convergence in probability**: $\mathbb{P}\left[|X_n - Y| > \varepsilon\right] \to 0$ for all $\varepsilon > 0$.
- **Convergence in mean** (or in L^1): $\mathbb{E}\left[|X_n - Y|\right] \to 0$.

The interaction of probability and size of discrepancy enforces a wide variety of different kinds of convergence.

The strong law of large numbers

The classic almost sure convergence result is the strong law of large numbers.

Theorem

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with finite mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

almost surely as $n \to \infty$.

There is a very beautiful martingale proof of this result.

Random variables not taking values in $\mathbb{R}$

1. We sometimes need to make sense of convergence for random variables which are not real-valued. For example, we might have random vectors, or random functions, or all sorts of exotic objects!
2. In general, we can cope with any random object X as long as it takes values in a nice metric space (S, d).
3. It's fairly clear what the analogues of convergence almost surely and convergence in probability should be:

$$\mathbb{P}\left[d(X_n, Y) \to 0\right] = 1 \text{ and } \mathbb{P}\left[d(X_n, Y) > \epsilon\right] \to 0 \text{ for all } \epsilon > 0.$$

Random variables not taking values in $\mathbb{R}$

What about convergence in distribution, though?
$\mathbb{P}[X_n \leq x]$ doesn't make much sense now.

It turns out that the right generalisation is the following.

Definition

Suppose that $X_1, X_2, \ldots$ are random variables taking values in a metric space (S, d). Then X_n **converges in distribution** (or **converges weakly**) to Y if

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(Y)]$$

for all bounded continuous functions $f : S \to \mathbb{R}$.

The Cramér-Wold device

A neat trick in case of random vectors $X_n = (X_{n,1}, X_{n,2}, \dots, X_{n,k})$ is that the weak convergence of random vectors follows from weak convergence of all linear combinations i.e.

$$\sum_{i=1}^{k} \lambda_i X_{n,i} \xrightarrow{d} \sum_{i=1}^{k} \lambda_i Y_i$$

as $n \to \infty$ for every possible choice of $\lambda_1, \dots, \lambda_k \in \mathbb{R}$.

The classical Central Limit Theorem

Consider the **standardized sum** for sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables with finite mean μ and variance σ^2:

$$\frac{(X_1 + \ldots + X_n) - n\mu}{\sqrt{n}\sigma}.$$

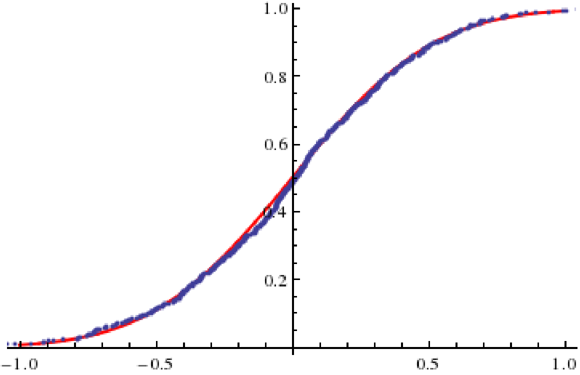
The standardized sum has zero mean and unit variance.

Theorem

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed, with finite mean μ and finite variance σ^2. Then

$$Y_n = \frac{(X_1 + \ldots + X_n) - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0, 1).$$

Example



Empirical c.d.f. of 500 draws of the sample mean of $n = 10$ independent random variables with the Student's t-distribution with 5 d.f. Limiting normal c.d.f. graphed in red.

Questions arising

In this lecture we address the following questions:

1. Do we really need "identically distributed"?
2. How fast does the convergence happen?
3. Do we really need "independent"?
4. What happens if we don't have finite variance? Or even a finite mean?

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

Lindeberg's Central Limit Theorem

Here's the strongest result for the non-identically distributed case.

Theorem

Suppose $X_1, \ldots, X_n$ are independent and **not** *identically distributed, with X_i having finite mean μ_i and finite variance σ_i^2. Set $m_n = \mu_1 + \ldots + \mu_n$ and $s_n^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Suppose further that*

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu_i)^2 \,;\, (X_i - \mu_i)^2 > \varepsilon^2 s_n^2\right] \to 0 \text{ as } n \to \infty,$$

for every $\varepsilon > 0$. Then

$$\frac{X_1 + \ldots + X_n - m_n}{s_n} \xrightarrow{d} N(0,1).$$

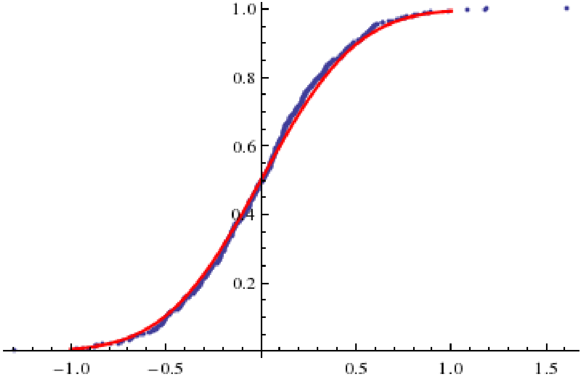
Proof is by a more careful development of the characteristic function proof of the classical Central Limit Theorem.

The Lindeberg condition

For independent random variables X_i the Lindeberg condition implies the CLT. Moreover (Feller, 1935), it is a necessary condition for the CLT under the overarching condition

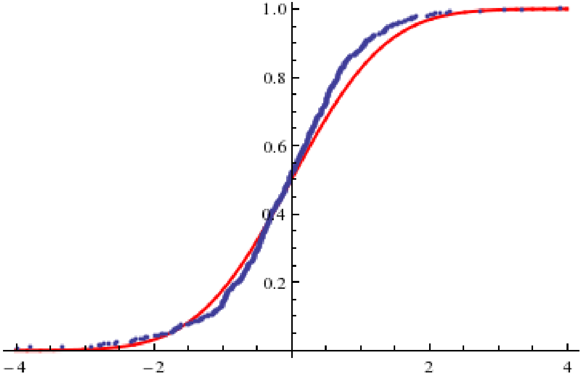
$$\mathbb{E}\left[(X_n - \mu_n)^2\right] / s_n^2 \to 0 \,.$$

Example: distributions not identical (I)



Empirical c.d.f. of 500 draws of the sample mean of 12 independent random variables, of which 10 have the Student's t-distribution with 5 d.f. and 2 have the Student's t-distribution with 3 d.f. Limiting normal c.d.f. graphed in red.

Example: distributions not identical (II)



Empirical c.d.f. of 500 draws of the sample mean of 12 independent random variables, of which 10 have the Student's t-distribution with 5 d.f. and 2 have the Student's t-distribution with 3 d.f. scaled by a factor of 5. Limiting normal c.d.f. graphed in red.

Rates of convergence: the Berry–Esseen theorem

Remarkably, we can capture how fast convergence occurs if we are given some extra information about the random variables.

Theorem

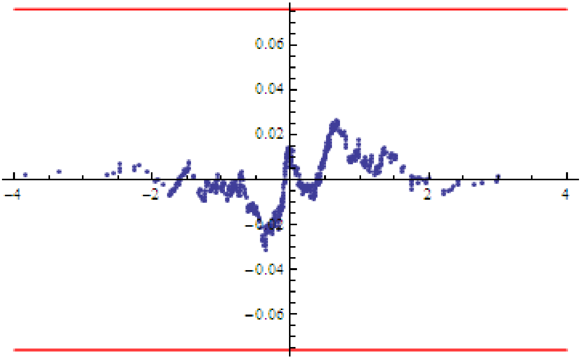
Suppose that $X_1, X_2, \ldots$ *are i.i.d. random variables with mean* μ *and finite variance* σ^2*. Suppose in addition that*

$$\rho^{(3)} = \mathbb{E}\left[|X_1 - \mu|^3\right] < \infty.$$

Write $F_n(x)$ *for the distribution function of* $\frac{(X_1 + \ldots + X_n) - n\mu}{\sqrt{n}\sigma}$*, and* $\Phi(x)$ *be the standard normal distribution function. Then there is a universal constant* $C > 0$ *such that*

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

Example



Plot of the difference between the limiting normal c.d.f. and the empirical c.d.f. of 500 draws from sample mean of 10 independent random variables with the Student's t-distribution with 5 d.f., together with upper and lower bounds.

Martingale case

Theorem

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

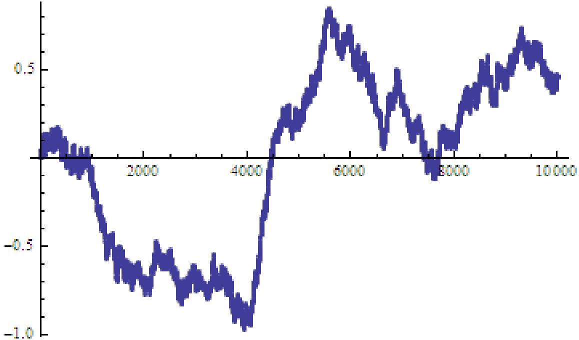
$$\frac{1}{s_n^2}\sum_{m=0}^{n-1}\mathbb{E}\left[|X_{m+1}-X_m|^2|\mathcal{F}_m\right] \;\to\; 1\,,$$

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

for each $\varepsilon > 0$.

Convergence to Brownian motion

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



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

No mean and/or variance (I)

In all of the cases we have discussed today, our random variables have had finite variances.

Let $X_1, X_2, \ldots$ be independent Cauchy random variables i.e. having density

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

The mean of this distribution is undefined because

$$\int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \infty.$$

(So it certainly doesn't have a variance!)

No mean and/or variance (II)

The characteristic function of X_1 is $\mathbb{E}\left[e^{itX_1}\right] = e^{-|t|}$ and so it's easy to see that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \stackrel{d}{=} X_1.$$

Compare to the standard normal distribution for which

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \stackrel{d}{=} X_1.$$

Stable laws

The standard normal and Cauchy distributions are examples of **stable laws**, that is distributions for which if we take i.i.d. copies $X_1, X_2, \ldots, X_n$ then there exist constants a_n and b_n such that

$$\frac{\sum_{i=1}^{n} X_i - a_n}{b_n} \stackrel{d}{=} X_1.$$

Stable laws are special because they are the possible limits in distribution for (recentred and normalised) sums of i.i.d. random variables. It turns out that necessarily $b_n = n^{1/\alpha}$ for some $\alpha \in (0, 2]$. The stable laws form a four-parameter family of distributions: $\alpha = 2$ corresponds to the normal distributions, which are the only ones to have finite variance; the mean exists iff $\alpha \in (1, 2]$.

Generalised CLT

Here's an example of the sort of result we can get.

Theorem (Gnedenko-Kolmogorov)

Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables such that

$$\frac{\mathbb{P}\left[X_1 < -x\right]}{x^{-\alpha}} \to c_1 \text{ and } \frac{\mathbb{P}\left[X_1 > x\right]}{x^{-\alpha}} \to c_2 \text{ as } x \to \infty$$

for some constants c_1, c_2 and some $\alpha \in (0,2)$. Then, for some constants m_n,

$$\frac{1}{n^{1/\alpha}} \left(\sum_{i=1}^{n} X_i - m_n \right) \xrightarrow{d} S_\alpha \text{ as } n \to \infty,$$

where S_α has a stable distribution (of index α).

Recurrence

"A bad penny always turns up"
Old English proverb.

Motivation from MCMC

Given a probability density $p(x)$ of interest, for example a Bayesian posterior, we could address the question of drawing from $p(x)$ by using, for example, Gaussian random-walk Metropolis-Hastings:

- 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.

Recurrence

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?

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: $\mathcal{L}(X_t|X_0 = x)$ (the distribution of X_t) 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 $\mathcal{L}(X_t|X_0 = x)(A) = \mathbb{P}[X_t \in A|X_0 = x]$ and $\pi(A)$:

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

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

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

(There are *many* other possible measures of distance . . .)

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

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

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.

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

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

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

ϕ-irreducibility (I)

We make two observations about Markov chain irreducibility:

1. The discrete theory fails to apply directly even to well-behaved chains on non-discrete state-spaces.
2. Suppose ϕ is a measure on the state-space: then we could ask for the chain to be irreducible *on sets of positive ϕ-measure*.

Definition

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

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

ϕ-irreducibility (II)

1. We call ϕ an **irreducibility measure**. It is possible to modify ϕ to construct a **maximal irreducibility measure** ψ; one such that any set B of positive measure under some irreducibility measure for X is of positive measure for ψ.
2. Irreducible chains on countable state-space are c-irreducible where c is counting measure ($c(A) = |A|$).
3. If a chain has unique equilibrium measure π then π will serve as a maximal irreducibility measure.

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.

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

Regeneration and small sets (II)

Definition

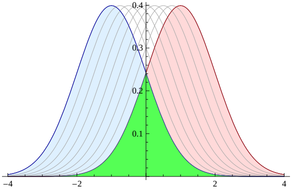
A set E of positive ϕ-measure is a **small set of lag** k for X if there is $\alpha \in (0,1)$ and a probability measure ν such that **for all** $x \in E$ the following **minorisation condition** is satisfied

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

We can interpret this as follows: if we sub-sample X every k time-steps 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 (III)

Consider the Gaussian random walk described above. **Any** bounded set is small of lag 1. For example, consider the set $E = [-2, 2]$.

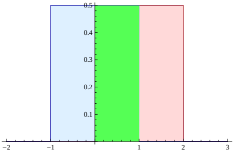


The green region represents the overlap of all the Gaussian densities centred at all points in E. Let α be the area of the green region and let f be its upper boundary. Then $f(x)/\alpha$ is a probability density (concentrated on E) and, for any $x \in E$,

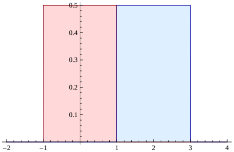
$$\mathbb{P}\left[X_1 \in A | X_0 = x\right] \geq \alpha \int_A \frac{f(x)}{\alpha} dx = \alpha\nu(A).$$

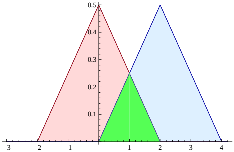
Regeneration and small sets (IV)

Let X be a RW with transition density $p(x, \mathrm{d}\, y) = \frac{1}{2}\, \mathbb{1}_{\{|x-y|<1\}}$. Consider the set $[0, 1]$: this is small of lag 1, with $\alpha = 1/2$ and ν the uniform distribution on $[0, 1]$.



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





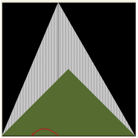
Regeneration and small sets (V)

Small sets would not be very interesting except that:

1. All ϕ-irreducible Markov chains X possess small sets;
2. Consider chains X with continuous transition density kernels. They possess *many* small sets of lag 1;
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.

Animated example: a random walk on $[0,1]$



▸ ANIMATION

Transition density $p(x,y) = 2\min\{\frac{y}{x}, \frac{1-y}{1-x}\}$.
Detailed balance equations (in terms of densities):

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

Spot an invariant probability density: $\pi(x) = 6x(1-x)$.
For any $A \subset [0,1]$ and all $x \in [0,1]$,

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

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

Regeneration and small sets (VI)

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 of lag 1 i.e. there exists a probability measure ν and $\alpha \in (0,1)$ such that

$$\mathbb{P}\left[X_1 \in A | X_0 = x\right] \geq \alpha \nu(A) \textit{ for all } x \in \mathcal{X}.$$

Then

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

and so X is uniformly ergodic.

Harris-recurrence

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

Definition

X is **(ϕ-)recurrent** if, for ϕ-almost all starting points x and any subset B with $\phi(B) > 0$, when started at x the chain X hits B eventually with probability 1.

Definition

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

Examples of ϕ-recurrence

- Random walks with continuous jump densities. (In fact, measurable jump densities suffice.)
- Chains with continuous or even measurable transition densities with the exception that the chain may stay put.
- Vervaat perpetuities:

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

  where U_1, U_2, … are independent Uniform$(0,1)$.
- Volatility models:

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

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

Foster-Lyapunov criteria

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

Renewal and regeneration

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

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

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

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

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

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

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

Then $\mathbb{E}\left[T_A | X_0 = x\right] < \infty$ 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, X has an equilibrium distribution.

Sketch of proof

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

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

   Hence, $Y_{\min\{n,T\}}$ converges.
2. So $\mathbb{P}\left[T < \infty\right] = 1$ (otherwise $\Lambda(X_n) > c$, $Y_n > c + an$ and so $Y_n \to \infty$). Moreover, $\mathbb{E}\left[Y_T | X_0\right] \leq \Lambda(X_0)$ (Fatou argument) so $a\,\mathbb{E}\left[T | X_0\right] \leq \Lambda(X_0)$.
3. Now use the finiteness of b to show that $\mathbb{E}\left[T^* | X_0\right] < \infty$, where T^* is the time of the first regeneration in C.
4. ϕ-irreducibility: X has a positive chance of hitting A before first regeneration in C. Hence, $\mathbb{E}\left[T_A | X_0\right] < \infty$.

A converse

Suppose, on the other hand, that $\mathbb{E}\left[T|X_0 = x\right] < \infty$ for all starting points x, where C is some small set and T is the first time for X to return to C. The Foster-Lyapunov criterion for positive recurrence follows for $\Lambda(x) = \mathbb{E}\left[T|X_0 = x\right]$ as long as $\mathbb{E}\left[T|X_0 = x\right]$ is bounded for $x \in C$.

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

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

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.

Sketch of proof

1. $Y_n = \Lambda(X_n)/\gamma^n$ defines non-negative supermartingale up to time $T = \inf\{m \geq 0 : X_m \in C\}$: if $T > n$ then

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

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

2. So $\mathbb{P}\left[T < \infty\right] = 1$ (otherwise $\Lambda(X) > c$ and so $Y_n > c/\gamma^n$ does not converge). Moreover, $\mathbb{E}\left[\gamma^{-T}|X_0\right] \leq \Lambda(X_0)$.
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 before regeneration in C. Hence, $\mathbb{E}\left[\gamma^{-T_A}|X_0\right] < \infty$.

Two converses

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

Uniform ergodicity follows if the function Λ is bounded above.

But more is true! Strikingly, for Harris-recurrent Markov chains the existence of a geometric Foster-Lyapunov condition is equivalent to the property of geometric ergodicity.

Examples

1. General reflected random walk: $X_{n+1} = \max\{X_n + Z_{n+1}, 0\}$, 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) Foster-Lyapunov criterion for positive recurrence applies.

   Similar considerations often apply to Metropolis-Hastings Markov chains based on random walks.
2. Reflected simple asymmetric random walk: $X_{n+1} = \max\{X_n + Z_{n+1}, 0\}$ for $Z_1, Z_2, \ldots$ i.i.d. such that $\mathbb{P}[Z_1 = -1] = q = 1 - p = 1 - \mathbb{P}[Z_1 = +1] > \frac{1}{2}$.
   (a) X is counting-measure-irreducible on non-negative integers;
   (b) Foster-Lyapunov criterion for geometric ergodicity applies.

Reflected simple asymmetric random walk: more details

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

$$\mathbb{E}\left[\Lambda(X_1) | X_0 = x_0\right] = \begin{cases} \Lambda(x_0) - (q - p) & \text{if } x_0 \not\in C\,, \\ 0 + p & \text{if } x_0 \in C\,. \end{cases}$$

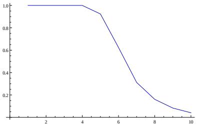
- Geometric ergodicity criterion: check for $\Lambda = e^{ax}$, $C = \{0\} = \Lambda^{-1}(\{1\})$:

$$\mathbb{E}\left[\Lambda(X_1) | X_0 = x_0\right] = \begin{cases} \Lambda(x_0) \times (pe^{a} + qe^{-a}) & \text{if } x_0 \not\in C\,, \\ 1 \times (pe^{a} + q) & \text{if } x_0 \in C\,. \end{cases}$$

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

Cutoff

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



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

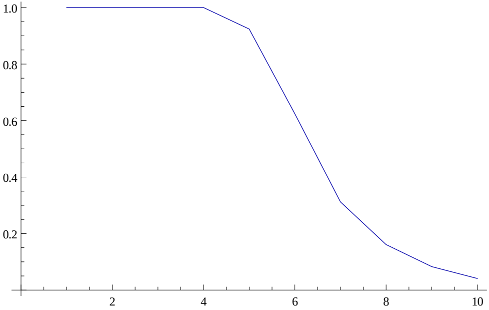
Consider repeatedly shuffling a pack of n cards using a **riffle shuffle**.

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

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:



Riffle shuffle: sharp result (Bayer and Diaconis 1992)

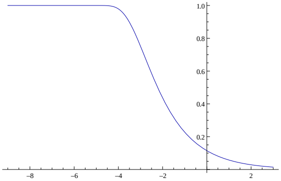
Let

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

Then

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

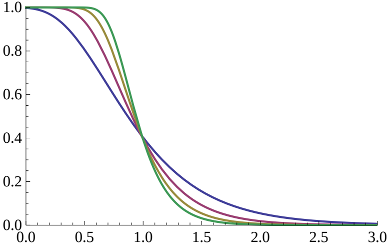
As a function of θ this looks something like:



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

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.

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 and group representation theory ...

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

Strong uniform times

Recall from lecture 3 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 $\sigma \in G$,

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

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

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 $k \geq 0$,

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

Proof.

For any set $A \subseteq G$,

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

So $|P^k(A) - \pi(A)| \leq \mathbb{P}\left[T > k\right]$. $\square$

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 $Q\heartsuit$;
- ...
- 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;
- ...

... 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

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

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

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

$$\begin{aligned}\mathbb{P}\left[W_n > k\right] = \mathbb{P}\left[\cup_d A_d\right] &\leq \sum_d \mathbb{P}\left[A_d\right] \\ &= n\left(1 - \frac{1}{n}\right)^k \leq n e^{-k/n}.\end{aligned}$$

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

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

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.

The lower bound

To prove lower bounds of cutoffs, a frequent trick is to find a set B such that

$$|P_n^{\tau_n(\theta)}(B) - \pi_n(B)| \quad \text{is large.}$$

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,

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

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	
S_n	Overhand shuffle	

Aldous, D. and P. Diaconis (1986).
Shuffling cards and stopping times.
The American Mathematical Monthly 93(5), 333–348.

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

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

Athreya, K. B. and P. Ney (1978).
A new approach to the limit theory of recurrent Markov chains.
Trans. Amer. Math. Soc. 245, 493–501.

Bayer, D. and P. Diaconis (1992).
Trailing the dovetail shuffle to its lair.
Ann. Appl. Probab. 2(2), 294–313.

Breiman, L. (1992).
Probability, Volume 7 of *Classics in Applied Mathematics*.
Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
Corrected reprint of the 1968 original.

Brown, B. M. (1971).
Martingale central limit theorems.
Ann. Math. Statist. 42, 59–66.

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

Durrett, R. (2010).
Probability: theory and examples (Fourth ed.), Volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*.
Cambridge: Cambridge University Press.

Fleming, I. (1953).
Casino Royale.
Jonathan Cape.

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

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

Jerrum, M. (2003).
Counting, sampling and integrating: algorithms and complexity.
Lectures in Mathematics ETH Zürich. Basel: Birkhäuser Verlag.

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

Kendall, W. S. (2004).
Geometric ergodicity and perfect simulation.
Electron. Comm. Probab. 9, 140–151 (electronic).

Kendall, W. S., F. Liang, and J.-S. Wang (Eds.) (2005).
Markov chain Monte Carlo: Innovations and Applications.
Number 7 in IMS Lecture Notes. Singapore: World Scientific.

Kendall, W. S. and G. Montana (2002).
Small sets and Markov transition densities.
Stochastic Process. Appl. 99(2), 177–194.

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

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

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

Murdoch, D. J. and P. J. Green (1998).
Exact sampling from a continuous state space.
Scand. J. Statist. 25(3), 483–502.

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

Nummelin, E. (1978).
A splitting technique for Harris recurrent Markov chains.
Z. Wahrsch. Verw. Gebiete 43(4), 309–318.

Nummelin, E. (1984).
General irreducible Markov chains and nonnegative operators, Volume 83 of *Cambridge Tracts in Mathematics*.
Cambridge: Cambridge University Press.

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

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

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

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

Photographs used in text

- Police phone box
  en.wikipedia.org/wiki/Image:Earls_Court_Police_Box.jpg
- The standing martingale en.wikipedia.org/wiki/Image:Hunterhorse.jpg
- The cardplayers en.wikipedia.org/wiki/Image:
  Paul_C%C3%A9zanne%2C_Les_joueurs_de_carte_%281892-95%29.jpg
- Chinese abacus en.wikipedia.org/wiki/Image:Boulier1.JPG
- Error function en.wikipedia.org/wiki/Image:Error_Function.svg
- Boomerang en.wikipedia.org/wiki/Image:Boomerang.jpg
- Alexander Lyapunov
  en.wikipedia.org/wiki/Image:Alexander_Ljapunow_jung.jpg
- Riffle shuffle (photo by Johnny Blood)
  en.wikipedia.org/wiki/Image:Riffle_shuffle.jpg