Lecture 1, 6.10.2022

MATH70031/M4P70 MARKOV PROCESSES

Professor N. H. BINGHAM
Mathematical Finance Section, 702 Weeks Hall;
n.bingham@imperial.ac.uk; 7594-2085

Place: HUXLEY 139
Times:
Thursday 10 - 11 am, Weeks 1 - 10 (6 Oct - 8 Dec)
Friday 9 - 10 am, Weeks 1 - 10 (7 Oct - 9 Dec)
Monday, 2-3 pm, Weeks 2 - 10 (10 Oct - 12 Dec)
Office hour: TBA
Class representative: To be selected. First task, to arrange possible times for the office hour. Suggestion:
(a) Get a class list, and a timetable grid (day of week x time of day). Ask in the Departmental Office to have these printed for you, and tell them why (or, DIY).
(b) Delete Tuesdays and Wednesdays. Pass it round the class, asking each person to put a small cross in any hour that would clash for them.
(c) Delete 9 am (I'm doing it once already, but that's enough!) Take your pick from any uncrossed squares. If there are none, take your pick from the squares with the smallest number of crosses.
(d) Tell me, and I'll announce it and stick it on the course website.

Main texts
[Nor] NORRIS, J. R., *Markov chains*, CUP, 1997.
[Wil] WILLIAMS, David, *Probability with martingales*, CUP, 1991.

Also useful
[HaiL] HAIRER, Martin and LI, Xue-Mei, *Markov processes*, 204 p. See the course website. [This is designed for private study, not for use in the lecture room. I am teaching the course in Xue-Mei's absence.]
[GriS] GRIMMETT, Geoffrey R. and STIRZAKER, David R., *Probability and random processes*, 3rd ed., 2001/4th ed., 2020, OUP.

The classics

[Chu] CHUNG, Kai-Lai, *Markov chains with stationary transition probabilities*, 2nd ed., Grundlehren math. Wiss. 104, Springer, 1967 (1st ed. 1960).
[Dyn] DYNKIN, E. B., *Theory of Markov processes*, Dover, p/b, 2006 (repr. Eng. tr. 1961, Russ. 1959).
[Fel] FELLER, W., *An introduction to probability theory and its applications*, Volume I, 3rd ed., Wiley, 1968.
[MeyT] MEYN, S. and TWEEDIE, R. L., *Markov chains and stochastic stability*, 2nd ed., p/b, CUP, 2009 (1st ed. 1993).

Cited in the text

Ch. 0

[Bil] P. BILLINGSLEY, *Probability and measure*. Wiley, 1986;
[Kal] O. KALLENBERG, *Foundations of modern probability*, 2nd ed., Springer, 2002 (1st ed. 1997).
[KinT] J. F. C. KINGMAN and S. J. TAYLOR, *Introduction to measure and probability*, CUP, 1966;
[Rud] Walter RUDIN, *Real and complex analysis*, McGraw-Hill, 2nd ed. (1974)/3rd ed. (1987).

Ch. 1.

[BinK] N. H. BINGHAM and R. KIESEL, *Risk-neutral valuation: Pricing and hedging of financial derivatives*, 2nd ed., Springer, 2004 (1st ed. 1998).
[Rev] D. REVUZ, *Markov chains*, North-Holland, 1975

Ch. 2.

[Bil61] P. BILLINGSLEY, *Statistical inference for Markov processes*. U. Chicago Press, 1961 (75p).
[CoxM] D. R. COX and H. D. MILLER, *The theory of stochastic processes*. Chapman & Hall, 1965 (p/b, 1977).
[DoyS] P. G. DOYLE and J. L. SNELL, *Random walks and electric networks*. Carus Math. Monographs 22, Math. Assoc. America, 1984.
[Ewe] W. J. EWENS, *Mathematical population genetics*, Springer, 1979.
[Kel] F. P. KELLY, *Reversibility and stochastic networks*, Wiley, 1979.
[KemS] J. G. KEMENY and J. L. SNELL, *Finite Markov chains*. Van Nostrand, 1960.

Ch. 3. [Nor], above.

Ch. 4.
[Asm] S. ASMUSSEN, *Applied prob. and queues*, 2nd ed., Springer, 2003.
[Bre] L. BREIMAN, *Probability*. Addison-Wesley, 1968.
[Øks] B. ØKSENDAL, *Stochastic differential equations: An introduction with applications*, 6th ed., Universitext, Springer, 2003 [5th ed. 1998].
[RevY] D. REVUZ and M. YOR, *Continuous martingales and Brownian motion*, 3rd ed. Grundl. math. Wiss. 293, Springer, 1999
[RogW] L. C. G. ROGERS and D. WILLIAMS, *Diffusions, Markov processes and martingales. Volume 1: Foundations* 2nd ed., Wiley, 1994; *Volume 2: Itô calculus*, Wiley, 1987.
[Ste] J. M. STEELE, *Stochastic calculus and financial applications*, Springer, 2001 (BM, Ch. 3).
[StrV] D. W. STROOCK and S. R. S. VARADHAN, *Multidimensional diffusions*. Grundl. math. Wiss. 233, Springer, 1979.

Table of contents

Time-line: Who did what when

Bernoulli, Daniel (1700 - 1782), Bernoulli-Laplace model of diffusion, 1769

Bernoulli, Jacob (Jacques) (1654-1705), *Ars conjectandi*, 1713 (posth.)

Boltzmann, Ludwig (1844 - 1906), H-theorem, thermodynamics, 1896

Brown, Robert (1773 - 1858), Brownian motion, 1828

Chung, Kai-Lai (1917 - 2009), the classic *Markov chains with stationary transition probabilities*, 1960, 1967

Clausius, Rudolf (1822 - 1888), entropy, First and Second Laws of Thermodynamics, 1865

Cramér, Harald (1893 - 1985), Cramér-Lundberg collective-risk model, 1920s; Cramér estimate of ruin in insurance, 1930

Daniell, Percy John (1889 - 1946), Daniell-Kolmogorov theorem (1918)

Ehrenfest, Paul (1880 - 1933) and Tatiana, Ehrenfest urn, 1907

Fisher, R. A. (Sir Ronald) (1890 - 1962), Wright-Fisher model in genetics, 1930

Frobenius, Georg (1849 - 1917), Perron-Frobeius theorem, 1908

Haar, Alfred (1885 - 1910), Haar functions, 1910; Haar measure, 1933

Kolmogorov, Andrei Nikolaevich (1903 - 1987), Daniell-Kolmogorov theorem; the classic *Grundbegriffe der Wahrscheinlichkeitsrechnung*, (1933)

Laplace, P.-S. de (1749 - 1827), Bernoulli-Laplace model of diffusion, 1812

Lundberg, Filip (1876 - 1965), Cramér-Lundberg collective risk model, 1920s

Markov, A. A. (1856 - 1922), Markov chains, 1908 [book, Calculus of probabilities (Russian), 2nd ed., 1908 (1st ed., 1900); German transl. *Wahrscheinlichkeitsrechnung*, 1912]

Meyer, P.-A. (1934 - 2003), filtrations, 1970s

Neumann, John von (1903-57), construction of the natural numbers $\mathbb{N}$, $\mathbb{N}_0$ in 1923

Paley, Raymond (R. E. A. C.) (1907-1933), Paley-Wiener-Zygmund (PWZ) theorem, 1933

Perron, Oskar (1880 - 1975), Perron-Frobenius theorem, 1907

Poisson, Siméon-Denis (1781 - 1840), Poisson distribution, 1837

Pólya, George (1897 - 1985), Pólya's theorem (§2.9)

Schauder, Juliusz (1899 - 1943), Schauder functions, 1927

Stirling, James (1692 - 1770), Stirling's formula, 1730.

Wiener, Norbert (1894 - 1964), Wiener process, 1923; Paley-Wiener-Zygmund theorem, 1933

Wright, Sewall (1889 - 1988), Wright-Fisher model in genetics, 1931

Zygmund, Antoni (1900 - 1992), Paley-Wiener-Zygmund theorem, 1933

Background and revision

Theorem (Conditional Mean Formula. For $\mathcal{B}$ any σ-field,

$$E[E[X|\mathcal{B}]] = E[X].$$

Proof. In the tower property, take $\mathcal{C}$ the trivial σ-field $\{\emptyset, \Omega\}$. This contains no information, so an expectation conditioning on it is the same as an unconditional expectation. The the tower property now gives

$$E[E[X|\mathcal{B}] \ |\{\emptyset, \Omega\}] = E[X|\{\emptyset, \Omega\}] = E[X]. \qquad //$$

Theorem (Conditional Variance Formula).

$$var(Y) = E[var(Y|X)] + var(E[Y|X]).$$

Proof. Recall $varX := E[(X - EX)^2]$. Expanding the square,

$$varX = E[X^2 - 2X.(EX) + (EX)^2] = E(X^2) - 2(EX)(EX) + (EX)^2 = E(X^2) - (EX)^2.$$

Conditional variances can be defined in the same way. Recall that $E(Y|X)$ is constant when X is known ($= x$, say), so can be taken outside an expectation over X, E_X say. Then

$$var(Y|X) := E(Y^2|X) - [E(Y|X)]^2.$$

Take expectations of both sides over X:

$$E_X var(Y|X) = E_X[E(Y^2|X)] - E_X[E(Y|X)]^2.$$

Now $E_X[E(Y^2|X)] = E(Y^2)$, by the Conditional Mean Formula, so the right is, adding and subtracting $(EY)^2$,

$$\{E(Y^2) - (EY)^2\} - \{E_X[E(Y|X)]^2 - (EY)^2\}.$$

The first term is $varY$, by above. Since $E(Y|X)$ has E_X-mean EY, the second term is $var_X E(Y|X)$, the variance (over X) of the random variable $E[Y|X]$ (random because X is). Combining, the result follows. //

Interpretation.

$varY$ = total variability in Y,

$E_X var(Y|X)$ = variability in Y not accounted for by knowledge of X,

$var_X E(Y|X)$ = variability in Y accounted for by knowledge of X:

variance = mean of conditional variance + variance of conditional mean. This is extremely useful in Statistics, in breaking down uncertainty, or variability, into its contributing components. There is a whole area of Statistics devoted to such Components of Variance.

Measure Theory

We assume that the audience has taken MATH50006 Lebesgue measure and integration, currently taught by Dr P.-F. Rodriguez.

We list below the results from Measure Theory that we shall need:

The framework of Lebesgue measure theory; σ-additivity, σ-algebras
Borel sets, non-measurable sets; properties (regularity, invariance)
Convergence theorems:
Lebesgue monotone convergence theorem (monotone convergence, MCT);
Lebesgue dominated convergence theorem (dominated convergence, DCT);
Fatou's lemma
General measure theory (similar to the canonical case of Lebesgue; often based on Carathéodory's extension theorem)
Uniform integrability
Absolute continuity
Lebesgue decomposition; Lebesgue differentiation theorem
Radon-Nikodym theorem

Suggested texts:
[KinT] Kingman and Taylor (above), Ch. 1 - 9;
[Wil] Williams (above), Ch. 1 - 8 and their Appendices;
[Rud] Rudin (above), Ch. 1-8.

Ch. 0. Probability and Measure Theory: Prerequisites and revision

What is probability? (This is the title of Ch. 10 of the book Kingman & Taylor [KinT] cited above. I'm a probabilist, by the way.)

This is (in some sense) a silly question. Use short words rather than long words when one can. Let's ask instead: What is chance? First, we have to have a common language in which to communicate. Ours is (happens to be) English. So, one could (should?) reply: I'm speaking to you in plain English. If you don't understand plain English, that's your problem, not mine.

We deliberately do not '*define probability*' (or chance, if you prefer short words) – we are not dictionary compilers. In non-mathematical terms, we have as good an idea of what chance is/means as of any other ordinary word. In mathematical terms: we have in Measure Theory just the mathematics we need to do the job: *a* probability (measure) is just *a measure of mass* 1; a probability (of an event – a measurable set) is the measure that probability measure gives to the set. That's it.

This passage was suggested by a statement (H&L, p.13) 'The 'randomness' describes the lack of complete information about the system' (we could have used 'randomness' instead of probability or chance above). This statement is worth considering here, before we engage with the main mathematical content of the course. Several comments:

1. *Tossing a coin.* A coin is a rigid body. We learn in (Newtonian) Dynamics how the motion of rigid body is determined by its initial conditions (add air resistance, if you like). We have no way to predict or describe the fine detail of how a coin is tossed – which is why the toss of a coin is routinely used to break symmetry, on occasions such as the start of a football match. As both captains and the referee would agree, arguing about Newtonian Dynamics is beside the point here. Life's too short: stick to the point, toss the coin, and get the match started.

2. *Chance in your life.* A young person has their life ahead of them, and most of the big choices still to make: of career and partner; what job to take, where to live, how many kids to have, etc. The 'information' concerning your life is always going to be incomplete until your life is over. You/we/I experience the uncertainties of life as chance, or randomness. I doubt whether most people would happily describe this as 'not randomness, but incomplete information as to what will happen before we die'.

3. *Mortality* (leading on from this). 'Call no man happy till he take his

happiness down to a quiet grave' (Aeschylus, *Oedipus Rex*, last line). My paraphrases of this: 'While you're alive, your vulnerable' (know anyone who's been in a life-changing accident, or any old person suffering from dementia?) Or (football again): 'You don't know the score until the game is over'.

Returning to 2 above: the only viewpoint from which I regard 'not randomness but incomplete information' as natural when applied to individual lives (yours, mine etc.) is that of an obituarist (I'm sensitive to this: I happen to have written lot of obituaries!)

Moral: that's enough of 'philosophising' for this course. From now on, we're going to use good, proper Mathematics: that of *measures of mass* 1, which we will call *probability measures*, and call their values *probabilities*. One can avoid these terms and speak instead of 'measures of mass 1' and 'values of this measure of mass 1' instead (not recommended).

Probability and Statistics

Statistics is an eminently practical and very useful subject (and the area of one of the four Sections in this Department). One way to think about Probability is as the theoretical underpinning of Statistics (we have probabilists in all four Sections). Another is bringing Measure Theory – abstract Pure Mathematics – 'to life', by applying it in the real world all around us. Random variables (measurable functions, below) belong to Probability.

Sampling. Is data random?

When we sample the value of a random variable drawn from a distribution (usually to study the distribution, as this is the only way we can get at it), what we get is a number (datum), and if we do it lots of times (the more the better), what we get is data. Data are still random in some sense: they are realised values of random variables, and if you (or someone else) re-sampled, you (or they) would get different values. In another sense, they are not random – you have them written down, in front of you. They are *numbers*, preferably lots of them; you have them, in some format (list, array etc.), stored in some way (computer, paper etc.) The statistician's job is to extract as much information from it as possible about where it came from – the distribution from which it was sampled. Statistics depends on Probability; Statistics is eminently useful and widely applied. Its new frontiers with 'Big Data' – Markov chain Monte Carlo (MCMC), data handling, Computer Science, Machine Learning etc. – are constantly growing.

Lecture 2. 7.10.2022

Measure and Probability

We will freely abbreviate 'probability' [11 letters] to pr [or prob, if you prefer], and (probability) distribution function to (probability) d/n, d/n fn or *law*.

We have a well-known correspondence between sets of technical terms here. With Measure Theory of the left and Probability Theory on the right:
measure [as a set-function] $\leftrightarrow$ probability [when the measure has mass 1]
measure [of a measurable set] $\leftrightarrow$ probability [of an event]
measure space $\leftrightarrow$ probability space
measurable set $\leftrightarrow$ event
measurable function $\leftrightarrow$ random variable
integral [wrt a measure] $\leftrightarrow$ expectation [wrt a pr measure]
convergence in measure $\leftrightarrow$ convergence in probability
almost everywhere, a.e. [except on a set of measure 0] $\leftrightarrow$ almost surely, a.s. [except on a set of pr 0]
a.e. convergence $\leftrightarrow$ a.s. convergence
L_2-convergence $\leftrightarrow$ mean-square convergence
weak convergence $\leftrightarrow$ convergence in distribution

Stochastic processes

The term *stochastic* (coined in the 1920s by A. Ya. Khinchin/Ȟinčin (depending on transliteration from Cyrillic to Roman), or Khintchine (when he was writing in French), can generally be used interchangeably with'probabilistic', or even 'random' (though note, as we shall see, a *stochastic matrix* is one whose row-sums are 1, while a *random matrix* is one whose elements are random). The term *process* denotes a phenomenon evolving, or unfolding, with time (if with space, we speak of a spatial process, if both of a space-time process). ['Stochastic' comes from the Greek word meaning to aim at or guess; compare the title of the first classic book in probability,
Jacob Bernoulli, *Ars conjectandi*, 1713 ('The art of guessing').
Perhaps *random process* as in [GriS] would be better.]

To begin, we need a space to model where the randomness takes place: a *probability space*. This is a *triple* $(\Omega, \{\mathcal{F}\}, \mathbb{P})$, consisting of:
a *sample space*, Ω – the set whose elements are the individual random outcomes, the *sample points*, ω;
a σ-field $\mathcal{F}$ of *events* – the subsets A of Ω whose probability $\mathbb{P}(A)$ is defined;
a probability measure $\mathbb{P}$.

General note on notation.

Our first duty is to avoid ambiguity – a sin, in mathematics.

Our second duty is to use the best notation for the job in hand, which is usually the minimal one (all mathematicians are minimalists at heart!) This is *context-dependent*, just as ordinary language is. The same word can mean quite different things in different contexts.

Not changing notation to suit a changed context compares with not changing one's clothes to suit a changed temperature or changed degree of formality etc. There is no virtue in 'uniform consistency', in either context. Remember: *notation is our servant, not our master.*

You have probably been exposed by now to enough different lecturers using different notation to be used to this sort of thing by now. I hope so!

Notation and sampling

Suppose we have a random variable, X say (below we shall use X for a stochastic process, a whole collection of random variables indexed by an index set I, but the two are consistent: one random variable corresponds to I being a singleton). The randomness here is in the sample point ω, so in full this is $X(\omega)$. It is customary, and very convenient, to omit the ω unless we need to mention it specifically, and we shall usually do this. To begin with, for your own use: if in doubt, put them in; if not in doubt, leave them out.

Note.

Chung used 'ωs everywhere' in the 1960 edition of his classic book [Chu] *Markov chains*. In the Preface to the 1967 edition (p. X), he writes:
'Personal taste and habit not being stationary in time, I should have liked to make more radical departures such as deleting hundreds of the ω's in the cumbersome notation, but have generally decided to leave well alone'.
I too have moved in the same way (perhaps under his influence).

A *stochastic process* (in brief, stoch proc, SP, or just *process*) X is a mathematical model of a phenomenon evolving in time, the model taking the form of a *family* of random variables $\{X_i : i \in I\}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by some index set I.

With a stochastic process indexed by time (below), the value of the process X at time t will be written $X_t(\omega)$, or just X_t. But, we frequently have multiple time-points t_i, say, and rather than have 'suffices within suffices', it is convenient to write $X_t(\omega) = X_t$ as $X(t, \omega) = X(t)$. I use them interchangeably, just as I do L_p and L^p, and hardly notice the difference.

Lecture 3. 10.10.2022

Usually, I will be infinite. If I is finite, of dimension $d < \infty$ say, the X_i are the elements of a d-dimensional random vector, which can be handled using the familiar machinery of a first course in probability, and/or statistics, particularly the vast and important area of statistics called Multivariate Analysis. The probabilistic structure here is described by the corresponding d-dimensional pr law:

$$F(\mathbf{x}),\ F_{\mathbf{X}}(\mathbf{x}) := \mathbb{P}(X_1 \leq x_1, \cdots, X_d \leq x_d), \qquad (*)$$

labelling I here as $\{1, 2, \cdots, d\}$, as we may.

When I is infinite, we have no familiar machinery to hand, and it is far from obvious how to find an infinite-dimensional analogue of the above, or even whether there is one. There is definitely something to prove!

Finite-dimensional distributions and the Daniell-Kolmogorov theorem

To proceed: note first that I, being infinite, has (infinitely many) finite-dimensional subsets, $(i_1, \cdots, i_d)$ say, to each of which the above applies. These are called the *finite-dimensional distributions* (fi-di d/ns, fdds) of the infinite collection. Write the corresponding d-dimensional probability law as $F(i_1, \cdots, i_d)$. These satisfy two *consistency conditions*:
(C1) Each d-dimensional distribution is invariant under permutation of the d indices. For, this just permutes the conditions $X_i \leq x_i,\ i = 1, \cdots d$ in $(*)$, so their intersection is unchanged.
(C2) Recall that if $x_i \uparrow \infty$ in $(*)$, the effect is to delete the condition on x_i, so taking a d-dimensional law into a $(d-1)$-dimensional one.

It turns out that these two *Daniell-Kolmogorov consistency conditions* (C1), (C2) are not only *necessary* for the existence of a stochastic process defined in this infinite-dimensional setting, but also *sufficient*. This is the Daniell-Kolmogorov theorem (P. J. Daniell (1889 - 1946) in 1918, A. N. Kolmogorov (1903 - 1987) in 1933; also called the Daniell extension theorem and the Kolmogorov extension theorem):

Daniell-Kolmogorov Theorem, D-K. For I any infinite index set, and any collection of finite-dimensional distributions satisfying the consistency conditions (C1), (C2) above, there exists a unique measure μ defined on the

Borel subsets $\mathcal{B}(\mathbb{R}^I)$ of $\mathbb{R}^I$ such that I restricted to each $(i_1, \cdots, i_d)$ gives $F(i_1, \cdots, i_d)$.

This classic result is the fundamental *existence theorem for stochastic processes*. We shall assume it here. We cite three textbook proofs (see References, Lecture 1):
[KinT] (measure, 159 - 161, by a compactness argument; probability, 381),
[Bil] (§36; two proofs, 513 - 515 and 515 - 517),
[Kal] (projective limits, 114 - 115).

In words: whenever speaking of a stochastic process $X = (X(i) : i \in I)$ on an infinite index set I makes sense (i.e., whenever its finite-dimensional distributions satisfy the consistency conditions (C1), (C2)), it can be constructed as above, and we can use it. So: in view of D-K, existence is no problem for us: *if it could exist, it does exist*. This will suffice for us.

Path properties

The Daniell-Kolmogorov theorem clearly uses all the information in the finite-dimensional d/ns, and *only* that. But one needs to go beyond this.

The map $t \mapsto X(t, \omega)$ (or to $X_t(\omega)$, $X(t)$ or X_t) is called the *path* (or *sample path*) of the process. Path properties only arise in continuous time. They involve information going *beyond* the finite-dimensional distributions! One wants to work with as well-behaved a *version* of the process as can be *realised* (constructed), consistent with the finite-dimensional distributions. The nicest property here is (path-)*continuity* (example: *Brownian motion*, Ch. 4). The next nicest is continuity from one side and limits from the other (this usually suffices, and is all we need in this course). Of the two possibilities here, the commoner is '*continuous on the right, limits on the left*' (corlol in English) (example: the *Poisson process*, Ch. 3). But as the area was developed by Paul-André Meyer and his colleagues in the Strasbourg school, it is usual to use the French, 'continu à droite, limite à gauche' (càdlàg, or cadlag). With right and left interchanged, one has càglàd, or caglad ('collor' isn't used). One even needs both together (in stochastic integration, one needs the random integrator cadlag and the random integrand caglad).

Lecture 4, 13.10.2022

Independence and independent copies

Recall the definition of independence:
1. Events $(A_i : i \in I)$, or $(A(i) : i \in I)$ to avoid suffices within suffices as before, are *independent* if and only if [iff], for every finite subset $(i_1, \cdots i_d)$ of I, the events $A(i_j)$, $j = 1, \cdots, d$, are independent, that is,

$$\mathbb{P}(A(i_1) \cap \cdots A(i_d)) = \prod_{j=1}^{d} \mathbb{P}(A(i_j)).$$

2. Random variables $(X_i : i \in I)$ are independent iff all events of the form $A_i := (X_i \in B_i)$, for B_i in the σ-field of the pr space on which X_i is defined, are independent.

The Daniell-Kolmogorov Theorem allows us to construct all these random variables on the same probability space (the *product* probability space), as the relevant consistency conditions are clearly satisfied.

In particular, we may have all the X_i defined on the same probability space, and so with the same distribution. The resulting set of X_i $(i \in I)$ are then *independent and identically distributed*, or *iid* for short. They are often called *independent copies* of each other – or just *copies*.
Note.

Recall von Neumann's construction of the natural numbers with zero: $\mathbb{N} := \{1, 2, 3, ...\}$, the set of *natural numbers*, and $\mathbb{N}_0 := \{0, 1, 2, ...\} = \mathbb{N} \cup \{0\}$. We can take these for granted, or proceed as follows:

$$0 \leftrightarrow \emptyset, \quad 1 \leftrightarrow \{\emptyset\}, \quad 2 \leftrightarrow \{0, 1\}, \quad 3 \leftrightarrow \{0, 1, 2\}, \cdots.$$

etc. (John von Neumann (1903-57) in 1923).

Example: Lebesgue measure and infinite coin-tossing: binary expansion

We note the simplest example of this situation. Let the x_i $(i \in \mathbb{N})$ be independent copies of a *binary* random variable (coin toss):

$$\mathbb{P}(x_i = 0) = \mathbb{P}(x_i = 1) = \frac{1}{2}, \qquad (i \in \mathbb{N}).$$

The relevant probability space for a single toss is $(\{0, 1\}, \mathcal{P}((\{0, 1\}), \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1)$, where $\mathcal{P}(A)$ is the *power set* of A, the set of all its subsets, and δ_n is

the Dirac mass (unit point mass) at n, which by above we could write as $(2, \mathcal{P}(2), \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1)$. For infinitely many tosses, we use the infinite product of this, written $(2, \mathcal{P}(2), \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1)^{\mathbb{N}}$. On this, the consistency conditions are clearly satisfied (automatic with independence). So the Daniell-Kolmogorov theorem gives us infinitely many independent copies of a coin toss.

This is a familiar result in a new setting! Recall the decimal expansion of a real number $x \in \mathbb{R}$: with $[x]$ for its integer part,

$$x = [x] \,.\, x_1 x_2 \cdots x_n \cdots, \qquad [x] \in \mathbb{Z}, \quad x_n \in \{0, 1, 2, \cdots, 9\}.$$

In just the same way, we can take the *binary expansion* instead: with the point now binary, not decimal,

$$x = [x] \,.\, x_1 x_2 \cdots x_n \cdots, \qquad [x] \in \mathbb{Z}, \quad x_n \in \{0, 1\},$$

Consider now the simplest continuous probability space, the *Lebesgue space*: the unit interval $[0, 1]$ endowed with its Borel σ-field and Lebesgue measure (a probability measure here). Drawing X from $[0, 1]$ under Lebesgue measure and taking its binary expansion as above, one can check inductively that the expansion coefficients X_n are independent coin-tosses.

The converse also holds. To see this, pick any $c \in \mathbb{R}$, and translate x by c, $x \mapsto x + c$. Discarding the integer part, the fractional part (binary expansion) of $x + c$ is also a sequence of independent coin-tosses, by the previous argument. So, the measure on x is translation-invariant. So it is Lebesgue measure (defined uniquely to within a positive multiplicative constant – needed to adjust for the unit of length, cm or inches say).

Thus *the Lebesgue space is identified with the countable independent product of the coin-tossing space.* So we have here two remarkably different ways of looking at the same thing.

Note: Haar measure

Translation-invariance of Lebesgue measure is invariance under the group action of the additive group of reals. This can be generalised, up to *locally compact topological groups* [locally compact: points have compact neighbourhoods; topological group: a group with a topology under which the group operations are continuous] – Alfred Haar (1885 - 1933), in 1933. See MacTutor for Haar (and the poem on Haar and von Neumann).

Lecture 5, 14.10.2022

Ch. 1: Stochastic Processes: Foundations

Filtrations

Much of Ch. 0 is *static*. We turn now to its dynamic counterpart, typically where randomness unfolds with time (as in life!)

Stochastic processes occur naturally even in finite situations. An example is in *mathematical finance*, when time is measured discretely (every day, say, or every hour – even every second or micro-second, depending on context), and space – here price (the value of a stock, say, to the nearest cent/penny, or (for interest rates and other percentages) basis point (bp – a hundredth of 1%). If the stock price at time n ($n = 0, 1, \cdots, N$) is S_n, we may be required to price *options* on the stock over time. Here we regard $S := (S_n)_{n=0}^N$as a stochastic process. Although everything here is finite (including the sample space Ω in the probability space needed to describe the model), a stochastic-process view point is needed here. The mathematics is fairly recent (1970s – e.g., the *Black-Scholes formula* of 1973), non-trivial, and of obvious practical importance. See e.g. [BinK] (above), Ch. 4.

Usually, however, the time-set will be infinite, e.g. $\mathbb{N}_0 := 0 \cup \mathbb{N}$, $\mathbb{R}_+ := [0, \infty)$, or some interval $I := [t_0, t_1]$. The 'space variable' (set of values taken) is also usually infinite, e.g. $\mathbb{Z}, \mathbb{N}, \mathbb{R}$.

Note.

1. Turning to the typical case where both time-set and value-set are infinite: we have an immediate split for each, between countable and uncountable. If both are countable, everyone calls a process with the Markov property (below) a *Markov chain*. Chung's classic (cited above) reserves the term *chain* for when the *state* variable is discrete, and divides his book between Part I: Discrete Parameter (recall the parameter $i \in I$ corresponds to time in Ch. 0) and Part II: Continuous Parameter. By contrast, for Revuz ([Rev], p. 13) a chain is in discrete time, whether state is discrete or continuous. It is more usual to use the term *Markov process* for the continuous-state case, and we shall follow this.

2. Recall that countability is built into Measure Theory because of the *countable additivity* property of measures. So we must expect extra difficulties when we pass from discrete to continuous, in either space or time, as the setting there is *uncountable*.

We begin in discrete time, with a random phenomenon producing, at each time $n = 0, 1, 2, \cdots$ (say) a random variable X_n. The information available to us at time n is the set of values $\{X_0, X_1, \cdots, X_n\}$. We can consider the *conditional* distribution of X_{n+1}, or X_m for $m > n$, *given* this information. Note that as n increases, the information available to us increases also.

Consider all events involving $\{X_0, X_1, \cdots, X_n\}$ – that is, all events of the form $\{(X_0, X_1, \cdots, X_n) \in A\}$, for A a (measurable) subset of $\mathbb{R}^{n+1}$ ('measurable' just means that the probability $P((X_0, X_1, \cdots, X_n) \in A)$ is defined). This class of events is called the σ-field *generated* by $(X_0, X_1, \cdots, X_n)$, written $\sigma(X_0, X_1, \cdots, X_n)$. It should be thought of as 'the information contained in $X_0, X_1, \cdots, X_n$', or 'what we know when we know $X_0, X_1, \cdots, X_n$'. For, knowing $X_0, X_1, \cdots, X_n$ we know exactly *which* of these events

$$\{(X_0, X_1, \cdots, X_n) \in A)\} \qquad (A \in \mathcal{B}(\mathbb{R}^{n+1}))$$

occur. We write

$$\mathcal{F}_n := \sigma(X_0, X_1, \cdots, X_n)$$

('$\mathcal{F}$ for field': recall we use script capitals for classes of sets). Then as the class of sets $\{(X_0, X_1, \cdots, X_n) \in A)\}$ increases with n,

$$\mathcal{F}_n \subset \mathcal{F}_{n+1} :$$

this just says that we learn more as time progresses and new information becomes available.

Definition (P.-A. Meyer, 1970s). A *filtration* is an increasing family $\{\mathcal{F}_n\}$ of σ-fields.

Here $\mathcal{F}_\infty := \cup_{n=0}^{\infty} \mathcal{F}_n$ is also a σ-field (check). If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{F}_n$, $\mathcal{F}_\infty \subset \mathcal{F}$, then $(\Omega, \{\mathcal{F}_n\}, \mathcal{F}, \mathbb{P})$ is called a *filtered probability space*.

If $X = \{X\}_n$ is a stochastic process, its *natural filtration* is $\{\mathcal{F}_n\}$ with $\mathcal{F}_n := \sigma(X_0, X_1, \cdots, X_n)$. We will usually be concerned with natural filtrations in this course.

Background and revision

If X_n is $\mathcal{F}_n$-measurable for each n, the stochastic process $X = \{X\}_n$ is said to be *adapted* to this filtration. We will *always* deal with adapted filtrations in this course.

Just as a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a mathematical model for a *static* random experiment, so a filtered probability space $(\Omega, \{\mathcal{F}_n\}, \mathcal{F}, \mathbb{P})$ is a mathematical model for a *dynamic* random experiment. It can serve as a basis for the mathematical description of any stochastic process $\{X_n\}$ adapted to it – that is, to $\{\mathcal{F}_n\}$: X_n must be $\mathcal{F}_n$-measurable for each n. The filtered probability space may – should – be regarded as part of the definition of the stochastic process $\{X_n\}$. Think of it as the space on which the stochastic process 'lives'.

Now that we are reassured by the Daniell-Kolmogorov theorem that all the stochastic processes we need *exist*, and have filtered probability spaces on which they live, we will often (usually; whenever possible) omit explicit reference to the filtered probability space. We need to know that it's there, but don't need (or want) to see it normally (the analogy with underwear comes to my mind here; if you find that helpful, fine; if not, also fine – *but then think of your own analogy*, or the subject will seem too abstract, and it needn't, and shouldn't).

Now that we know what a stochastic process is and 'where one lives', we will usually drop the 'stochastic', and leave it to be understood from context. From now on, 'process' *means* 'stochastic process' (again: if this bothers you, fine – go on writing 'stochastic' in every time – unless/until you realise you no longer need it).

Dependence and independence

We have lots of experience of dealing with independence – in your first course(s) on Probability and (implicitly) on Statistics (the different readings in a sample being typically assumed to be independent copies from the same distribution).

Note on errors in Statistics

Independent errors tend to cancel. This is the essence of the Laws of Large Numbers (Week and Strong); see MATH70028/M4P6 PROBABILITY (Dr I. Krasowsky, Term 2).

It is also what makes Statistics work. This is basically why large samples (though more expensive to gather and analyse) are better than small ones – there is more cancellation). This is *not true* for dependent errors. *Correlated errors in Statistics are very dangerous.*

Example.

Imagine a Physics lesson at school, in which an experiment is to be performed. The teacher divides the class into 10 pairs, and then goes into his back room to catch up on exam marking. The two best experimenters are in the same pair; the others gang up on them and force them to do the experiment for them, while they revise for their exam. A physical constant is to be measured, correct to several significant figures. Unfortunately the instrument the 'good' pair are using reads way too high (it was dropped that morning; the culprit did not own up). When the 'good' pair have their result, the others copy the first three significant figures, but attempt to disguise their cheating by each of the 9 pairs inventing their own 'nonsense figures' after that. These, being independent, *will* tend to cancel. But the first three significant figures will not: they will be way too high for all 10 pairs. There is *no* cancellation there, only *replication* of a wrong result.

Independence is much easier to handle than dependence. It is here that the three classical limit theorems, LLN (Law of Large Number, Weak and Strong), CLT (Central Limit Theorem) and LIL (Law of the Iterated Logarithm) find their simplest forms. Beyond that, the two main areas in which *dependence* can be handled are *Markov* chains and processes (this course) and *martingales* (M4P6 next term again). The other main areas where much can be said are *weak dependence* (mainly a hierarchy of *mixing* conditions), *stationarity*, and *Gaussianity*.

Lecture 6, 17.10.2022

Ch. 2. MARKOV CHAINS: DISCRETE TIME

1. Notation and Examples

The Markov property

A *Markov process* in discrete time is a stochastic process $X = (X_n)$ with

$$\mathbb{P}(X_n \in A \mid X_m, B) = \mathbb{P}(X_n \in A \mid X_m) \qquad \forall\ B \in \mathcal{F}_{m-1},\ m < n :$$

in words, the conditional probability of the future (time n) given the present (time $m < n$) and past (times up to $m-1$) is the same as that of the future given the present. So: where you are is all that counts, not how you got there.

Conditional independence of past and future given the present

We can express the Markov property informally and schematically by writing it (with F denoting future) as

$$\mathbb{P}(F \mid Past, Present) = \mathbb{P}(F \mid Present).$$

Conditioning on a present event of positive probability, this says

$$\frac{\mathbb{P}(F \cap Past \cap Present)}{\mathbb{P}(Past \cap Present)} = \frac{\mathbb{P}(F \cap Present)}{\mathbb{P}(Present)}.$$

Multiply by $\mathbb{P}(Past \cap Present)/\mathbb{P}(Present)$:

$$\frac{\mathbb{P}(F \cap Past \cap Present)}{\mathbb{P}(Present)} = \frac{\mathbb{P}(F \cap Present)}{\mathbb{P}(Present)} . \frac{\mathbb{P}(Past \cap Present)}{\mathbb{P}(Present)} :$$

$$\mathbb{P}(F \cap Past \mid Present) = \mathbb{P}(F \mid Present) . \mathbb{P}(Past \mid Present) :$$

past and future are conditionally independent given the present. This statement is symmetrical between past and future. This suggests that one may be able to *reverse time*. This is true in that reversing time does give *a* Markov property, but in general under a different transition probability matrix (below; [Nor, §1.9]; think of the Second Law of Thermodynamics – entropy increases with time), so the chain 'looks different when run backwards'. All that one can usefully say in general is that the present *splits the past and the*

future, as above. But some Markov chains 'look the same when run backwards'; these *reversible* Markov chains are very useful.

The values taken by the process may be discrete or continuous. The discrete case is easier, so we begin with it. The X-values form a countable set, $\{x_n\}$ say. It is usually possible to disregard the precise values x_n and replace them by *labels*, n. Usually this label set will be $\mathbb{N}$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $\mathbb{N}_n := \{1, 2, \cdots, n\}$,$\mathbb{N}_n^0 := \mathbb{N}_n \cup \{0\}$ or $\mathbb{Z}$, depending on context. In general, write E_k for state k.

Example.

Simple random walk (below) on $\mathbb{Z}$: label set and value set both $\mathbb{Z}$.

It is convenient to refer to a Markov process with both time and state discrete as a *Markov chain*. The theory originated with the Russian probabilist A. A. MARKOV (1856 - 1922), in the second (1908) Russian edition (1st, 1900) of his book *Calculus of Probabilities* (German translation, *Wahrscheinlichkeitsrechnung*, 1912).

The classic text for Markov chains is Feller's book [Fel], which has influenced both our style of treatment and choice of material, examples etc.

To describe such a Markov chain, we need the *transition probabilities*,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i).$$

We confine ourselves here, for simplicity, to the most important special case, when these transition probabilities are *stationary* (do not depend on time n):

$$p_{ij} := \mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(i \to j).$$

We assemble these transition probabilities p_{ij} into a *transition (probability) matrix*

$$P := (p_{ij}).$$

Similarly, we define the *n-step* transition probabilities

$$p_{ij}^{(n)} := \mathbb{P}(X_{m+n} = j \mid X_m = i) = \mathbb{P}(i \to j \text{ in } n \text{ steps})$$

(by stationarity, this does not depend on m), and form the n-step transition probability matrix

$$P^{(n)} := (p_{ij}^n).$$

Theorem. $P^{(n)} = P^n$: the n-step transition probability matrix is the nth matrix power of the (1-step) transition probability matrix P.

Lecture 7. 20.10.2022

Proof. First, the case $n = 2$:

$$\begin{aligned} p_{ij}^{(2)} &= \mathbb{P}(i \to j \text{ in 2 steps}) = \sum_k \mathbb{P}(i \to k \to j) \\ &= \sum_k \mathbb{P}(k \to j \text{ on 2nd step} \mid i \to k \; on \; 1st)\mathbb{P}(i \to k \; on \; 1st). \end{aligned}$$

Using the Markov property on the last RHS above and reversing factors:

$$p_{ij}^{(2)} = \sum_k \mathbb{P}(k \to j)P(i \to k).$$

This says that

$$p_{ij}^{(2)} = \sum_k p_{kj}p_{ik} = \sum_k p_{ik}p_{kj} = (P^2)_{ij},$$

for all i, j. Combining,

$$P^{(2)} = P^2.$$

The general case is similar (or, use induction on n). □

This result shows one of the great advantages of Markov chain theory: it is perfectly adapted to the theory of matrices and linear algebra, which is very well developed (and familiar!)

Initial distribution

Suppose that the position at time $n = 0$ is random, with

$$p_i := \mathbb{P}(X_0 = i).$$

form the (in general infinite) *row-vector*

$$p := (p_0, p_1, \cdots, p_i, \cdots).$$

Then

$$\begin{aligned} \mathbb{P}(X_n = j) &= \sum_i \mathbb{P}(X_n = j \; \cap \; X_0 = i) = \sum_i \mathbb{P}(X_0 = i)\mathbb{P}(X_n = j | X_0 = i) \\ &= \sum_i p_i p_{ij}^{(n)} = (pP^n)_j : \end{aligned}$$

$$\mathbb{P}(X_n = .) = pP^n :$$

the row-vector pP^n gives the distribution of the position at time n.

Examples.

1. *Two states.*

$$P = \begin{pmatrix} 1-p & p \\ \alpha & 1-\alpha \end{pmatrix}.$$

Think of motion on $\mathbb{R}$ with constant speed,

$$p = \mathbb{P}(\text{change direction to left} \mid \text{going right}),$$

$$\alpha = \mathbb{P}(\text{change direction to right} \mid \text{going left}).$$

Cox & Miller [CoxM] (§3.2, Example 3.2) use this model to discuss rainfall data (with the two states being wet and dry days) in Tel Aviv. For a general treatment of statistical issues here, see e.g. Billingsley [Bil].

2. *Random walk with absorbing barriers*

'Gambler's ruin' with finite total capital a is modelled by a Markov chain with $a+1$ states, and (with p, q the probabilities of winning or losing on each play, $p+q=1$)

$$P = \begin{pmatrix} 1 & 0 & & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & \ddots & \ddots & \ddots & \\ & & & q & 0 & p \\ & & & & 0 & 1 \end{pmatrix}.$$

3. *Random walk with reflecting barriers*

Suppose that in the gambler's ruin case above, the players wished to continue playing. Then a player who runs out of money needs to have his last stake returned, leading to

$$P = \begin{pmatrix} q & p & 0 & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & \ddots & \ddots & \ddots & \\ & & & q & 0 & p \\ & & & & q & p \end{pmatrix}.$$

Lecture 8. 21.10.2022

4. *Cyclic random walks*

Suppose the states represent positions on a circle:

$$P = \begin{pmatrix} q_0 & q_1 & & & q_{a-1} \\ q_{a-1} & q_0 & q_1 & & q_{a-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_1 & q_2 & & q_{a-1} & q_0 \end{pmatrix}.$$

5. *The Ehrenfest model of diffusion*

Suppose that a balls are distributed between two 'urns' (two parts of a contained, A and B). At each stage, a ball is chosen at random (each with prob $1/a$) and changed to the *other* urn. The state is the number of balls in A (say). Then

$$p_{i,i+1} = 1 - i/a, \qquad p_{i,i-1} = i/a, \qquad p_{ij} = 0 \text{ else.}$$

$$P = \begin{pmatrix} 0 & 1 & & & & \\ a^{-1} & 0 & 1-a^{-1} & & & \\ & 2a^{-1} & 0 & 1-2a^{-1} & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1-a^{-1} & 0 & a^{-1} \\ & & & & 0 & 1 \end{pmatrix}.$$

This model was introduced by Paul and Tatiana Ehrenfest in 1907, motivated by statistical mechanics. The balls represent molecules of a gas (so will be present in astronomical numbers – recall Avogadro's number, c. 6.02×10^{23}, is the number of gas molecules per standard volume under standard conditions). If A has an excess (more than half) of molecules, it is more likely to lose molecules to B than gain them, and vice versa. The model exhibits a *central force*, or *restoring force*, towards equilibrium, but allows departures from equilibrium by spontaneous fluctuations. We return to its (very interesting) physical interpretation later.

6. *Bernoulli-Laplace model of diffusion*

This is an earlier variant on the Ehrenfest model (Daniel Bernoulli in 1769, P.-S. de Laplace in 1812). Here there are $2a$ balls, of two colours, a black and a white, say, and two containers, 1 and 2, each of which contains a balls. At each stage, a ball is chosen at random from each container and

they are interchanged. The state is the number of white balls in 1 (say):

$$p_{i,i-1} = (i/a)^2, \quad p_{i,i} = 2i(a-i)/a^2, \quad p_{i,i+1} = (a-i)^2/a^2, \quad p_{ij} = 0 \text{ else.}$$

Again: this is motivated by statistical mechanics; see later.

7. *Wright-Fisher model in mathematical genetics*

Here (Sewall Wright in 1931, R. A. Fisher in 1930) the $2N$ genes in each generation are obtained by sampling with replacement from the genes in the previous one, leading to

$$p_{ij} = \binom{2N}{j} (i/2N)^j (1-(i/2N))^{2N-j} \qquad (i,j = 0,1,\cdots,2N).$$

Note. Mathematical genetics makes extensive use of Markov chain methods. For background, see e.g. Ewens [Ewe].

2. Classification of States

Call E_k is *accessible* from E_j if $p_{jk}^{(n)} > 0$ for some n, and write $j \to k$. If both $j \to k$ and $k \to j$, j and k *intercommunicate*, $j \leftrightarrow k$.

Call a set C of states *closed* if no state outside C is accessible from any state in C (that is, once the process enters C, it stays there). The *closure* $\overline{C}$ of any set of states C is the smallest closed set containing it (cf. Topology).

A singleton closed set is an *absorbing state*, or *trap* (example: the extreme states 0 and a in the gambler's ruin problem (§1, Example 2).

A Markov chain is called *irreducible* if there is no closed set other than the set of all states.

Stochastic matrices and subchains

Call a matrix Q *stochastic* if its elements are non-negative and its row-sums are 1 (example: a Markov chain transition pr matrix $P = (p_{ij})$, §1).

If we have a closed set C of states, let us for convenience label sets in C first, then those outside it. Then P is partitioned as

$$P = \begin{pmatrix} Q & 0 \\ U & V \end{pmatrix},$$

where Q governs transitions within C, the 0 shows we can't leave C, U governs transitions from outside C to C and V governs those outside C.

We can now imagine *deleting* all states outside C from the state space. We are left with a Markov chain, with state space C and transition pr matrix Q, called the Markov chain *restricted* to C, or the *subchain* on C. Note that

$$P^n = \begin{pmatrix} Q^n & 0 \\ U_n & V^n \end{pmatrix},$$

where P^n, Q^n, V^n are matrix powers, but U_n is not in general.

Lecture 9, 24.10.2022

Periodicity

A state E_j has *period* $t > 1$ if $P_{jj}^{(n)} = 0$ unless n is a multiple of t, and t is the largest such integer. Otherwise E_j is called *aperiodic*.

Example.

Simple random walk: all states are periodic with period 2. So is the Ehrenfest urn model.

Stopping times and the strong Markov property

Call a random variable $X : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ a *stopping time* (for the process X) if the events $\{T = n\}$ depend only on $X_0, X_1, \cdots, X_n$, i.e.

$$\{T = n\} \in \mathcal{F}_n := \sigma(X_0, \cdots, X_n) \qquad \forall n.$$

The name makes clear how to think about this: we know when to stop without having to look into the future (think of a gambler, who would like to stop 'at the end of a winning streak', but has no way to do so).

The *strong Markov property* (SMP) is that the Markov property extends from fixed times to stopping times T, when $T < \infty$. We shall assume this (see e.g. [Nor, Th. 1.4.2] for the proof, which is quite short). But note that the strong Markov property, though not restrictive in the discrete-time setting here, becomes so in continuous time (Ch. 3, 4).

Transience and persistence (recurrence)

Call state E_j *persistent* (or *recurrent* – usage varies, even here!) if

$$\mathbb{P}(\text{return to } E_j \mid \text{start at } E_j) = 1,$$

(return to E_j is certain), *transient* otherwise. Then return to E_j n times is certain for each n. Let $n \to \infty$: as the decreasing limit of certain sets is certain (countable additivity; equivalently, the union of countably many null sets is null, which just says $\sum 0 = 0$): writing 'i.o.' for 'infinitely often',

$$\mathbb{P}(\text{return to } E_j \text{ i.o.} \mid \text{start at } E_j) = 1.$$

For E_j *transient*, $\mathbb{P}(\text{return to } E_j \mid \text{start at } E_j) < 1$. Return to E_j is a stopping time. Use the strong Markov property to re-start at every return. Then

$$0 \leq \mathbb{P}(E_j \text{ i.o.} \mid \text{start at } E_j) \leq \mathbb{P}(E_j \ n \text{ times} \mid \cdots) \leq \mathbb{P}...|...)^n \to 0 :$$

$\mathbb{P}(E_j \text{ i.o.}|...) = 0$. There is thus a clean split between persistence and transience here – an example of a *zero-one law* (0-1 law; cf. Problems 3).

3. The Feller relation

Consider a situation Φ such as return to a state E_j of a Markov chain (X_n), which *regenerates* – 'starts afresh','forgets its past' – each time it occurs. To describe this, we need two related sequences. Write 'fft' as short for 'for the first time'.

$$u_n := \mathbb{P}(\text{it happens at time } n) \qquad (u_0 := 1),$$

$$f_n := \mathbb{P}(\text{it happens at time } n \text{ fft}) \qquad (f_0 := 0).$$

Form the generating functions (GFs)

$$U(s) := \sum_0^\infty u_n s^n, \qquad F(s) := \sum_0^\infty f_n s^n.$$

The link between them, the *Feller relation*, is Th. 1, §13.3 in his book [Fel] (§12.3 in the 1st (1950) edition).

Theorem (Feller relation) $\qquad U(s) = 1/(1 - F(s))$.

Proof.

$$u_n = \mathbb{P}(\text{happens at } n) = \sum_0^n \mathbb{P}(\text{at } n \text{ and fft at } k) = \sum_0^n \mathbb{P}(\text{at } n \mid \text{fft at } k).\mathbb{P}(\text{fft at } k)$$

By regeneration at k, the first factor is u_{n-k}, while the second is f_k. So

$$u_n = \sum_1^n u_{n-k} f_k \qquad (n \geq 1), \qquad u_0 = 1.$$

(restricting the sum to start at 1 as $f_0 = 0$). Multiply by s^n and sum:

$$U(s) = 1 + \sum_1^\infty u_{n-k} s^{n-k}.f_k s^k.$$

Write $j := n - k$. The limits of summation are now $j \geq 0$, $k \geq 0$ or 1, giving

$$U(s) = 1 + \sum_0^\infty u_j s^j . \sum_0^\infty f_k s^k = 1 + U(s).F(s) :$$

$$U(s)(1 - F(s)) = 1 : \qquad U(s) = 1/(1 - F(s)). \qquad \square$$

Lecture 10, 27.10.2022

Cor. Write f for the probability that regeneration ever occurs. Then if $u := \sum_0^\infty u_n$,

$$f = 1 - 1/u.$$

Thus regeneration Φ is certain iff $u = \sum u_n$ diverges.

Proof. As $u_n \geq 0$, $U(s)$ is increasing. so for each N,

$$\sum_0^N u_n \leq \lim_{s\uparrow 1} U(s) \leq \sum_0^\infty u_n = u.$$

So $U(s) \uparrow u$ as $s \uparrow 1$. Likewise $F(s) \uparrow f$ as $s \uparrow 1$. So letting $s \uparrow 1$,

$$u = 1/(1-f), \qquad f = 1 - 1/u. \qquad \square$$

If Φ is persistent (certain to occur eventually) and T is the time to first regeneration,

$$\mathbb{E}[s^T] = \sum f_k s^k = F(s): \qquad \mathbb{E}[Ts^{T-1}] = \sum k f_k s^{k-1} = F'(s):$$

$$\mu := \mathbb{E}[T] = \sum k f_k = F'(1).$$

If $f < 1$, call Φ *defective*, with *defect* $1 - f = 1/u$. Thus Φ is defective or non-defective according as $u = \sum u_n$ converges or diverges.

Theorem (Erdős-Feller Pollard theorem, EFP, 1949). If Φ is persistent and $\mathbb{E}[T] = \mu$:
(i) in the aperiodic case, $u_n \to 1/\mu$ $(n \to \infty)$.
(ii) with period t, $u_{nt} \to t/\mu$ (while $u_k = 0$ for k not divisible by t).

Proof. See Feller [Fel, XIII.11] (analysis, not probability; not examinable). $\square$

4. Transience and persistence (recurrence)

Call a persistent E_j *null* if its mean recurrence time $\mu_j = \infty$, *positive* if $\mu_j < \infty$.

By EFP, as $p_{jj}(n) \to 1/\mu_j$ $(n \to \infty)$ (aperiodic case; similarly for the periodic case). So for persistent E_j,

$$E_j \text{ is null iff} \qquad p_{jj}(n) \to 0 \qquad (n \to \infty).$$

Adapting the notation above, write $U_j(s) = \sum u_{jn} s^n$, $F_j(s) = \sum f_{jn} s^n$ for the functions $U(s), F(s)$ corresponding to state E_j. So E_j is persistent iff $\sum u_{jn} = \infty$, i.e. $\sum p_{jj}(n) = \infty$, i.e. iff the corresponding Φ_j is non-defective.

Call i *ergodic* if i is aperiodic and positive persistent.

Theorem (Solidarity Theorem). If states i and j intercommunicate, they have the same properties:
(i) Both are persistent or both are transient; if persistent, both are null or both are positive.
(ii) Both have the same period.
(iii) Both or neither are ergodic.

Proof. (i) If $i \leftrightarrow j$, then $p_{ij}^{(m)} > 0$ for some m and $p_{ji}^{(n)} > 0$ for some n. Then

$$p_{ii}(m+n+r) = \sum_k p_{ik}(m) p_{ki}(n+r) = \sum_{k,\ell} p_{ik}(m) p_{k\ell}(r) p_{\ell i}(n)$$

$$\geq p_{ij}(m) p_{jj}(r) p_{ji}(n)$$

(all terms, being products of probabilities, are non-negative). So if $\alpha := p_{ij}(m) p_{ji}(n)$, $\alpha > 0$ and

$$p_{ii}(m+n+r) \geq \alpha \, p_{jj}(r).$$

So if j is persistent, i.e. $\sum u_{jn} = \infty$, the above gives $p_{ii}(n) = \infty$, so i is persistent. By symmetry in i and j: i persistent iff j persistent. As transience/persistence is a dichotomy, this gives i transient iff j transient (or, the two series converge together).

If i is null persistent, $p_{ii}(n) \to 0$ $(n \to \infty)$, so the above gives $p_{jj}(n) \to 0$ $(n \to \infty)$, so j is null-persistent, etc.
(ii) If i has period t, $r = 0$ above gives $m + n$ a multiple of t. Then the LHS above $= 0$ unless r is a multiple of t. So the RHS $= \alpha p_{jj}(r) = 0$ unless r is a multiple of t, so j too has period t.
(iii) From (i) and (ii). $\square$

Lecture 11, 28.10.2022

Generating functions. As above, write

$$P_{ij}(s) := \sum_0^\infty p_{ij}(n)s^n, \qquad F_{ij}(s) := \sum_0^\infty f_{ij}(n)s^n.$$

As above, one finds

$$P_{ii}(s) = 1 + F_{ii}(s)P_{ii}(s) : \qquad P_{ii}(s) = 1/(1 - F_{ii}(s)),$$

$$P_{ij}(s) = F_{ij}(s)P_{jj}(s) \qquad (i \neq j)$$

(read: LHS, from i to j; RHS, from i to j for the first time, then back to j).

Equivalence classes

The relation $i \leftrightarrow j$ is reflexive, symmetric and transitive, so is an *equivalence relation.* It thus decomposes to set of all states into *equivalence classes/* By above, each equivalence class consists of states with the same properties, and is irreducible.

Let $C_1, C_2, \cdots$ be the persistent equivalence classes. There may be many. E.g. Gambler's run: $C_1 := \{0\}$ (gambler is ruined) and $C_2 := \{a\}$ (gambler wins) are distinct persistent equivalence classes.

Proposition. Each C_r is closed.

Proof. If $i \in C_r$ and $i \to j$ for $j \notin C_r$, j does not lead back to i (or j would be in C_r). So

$$\mathbb{P}(X_n \neq i \ \forall n \geq 1 \mid X_0 = i) \geq \mathbb{P}(X_1 = j \mid X_0 = i) > 0,$$

which means that return to i is uncertain, contradicting the persistence of i. So no such j can exist, so C_r is closed. □

The transient states are less important than the persistent ones, and there is no need to distinguish between them. So with T the set of transient states:

Decomposition Theorem. The state space S can be partitioned uniquely as

$$S = T \cup C_1 \cup \cdots \cup C_r \cup \cdots,$$

where T is the set of transient states and the C_r are the irreducible closed sets of persistent states.

As a corollary: we could form $C := \bigcup_i C_i$, the set of all persistent states, which is *closed* as each C_i is, and partition the state space as $S = C \cup T$. This partitions the transition probability matrix P as before as

$$P = \begin{pmatrix} Q & 0 \\ U & V \end{pmatrix}.$$

Then Q is again a Markov chain transition probability matrix, that of the chain *restricted* to C (subchain on C), the class of persistent sets.

Finite chains
Theorem. For a finite Markov chain, it is impossible for all states to be transient: a finite chain must contain some persistent states.

Proof. If the state space is $\{1, \cdots, N\}$, for each i and each n

$$1 = \sum_{j=1}^{N} p_{ij}(n).$$

Let $n \to \infty$: if j were transient, $\sum p_{ij}(n) < \infty$, so $p_{ij}(n) \to 0$ as $n \to \infty$. If *all* states j were transient, letting $n \to \infty$ here would give the contradiction

$$1 = \lim_{n\to\infty} \sum_{1}^{N} p_{ij}(n) = \sum_{1}^{N} \lim p_{ij}(n) = \sum_{1}^{n} 0 = 0.$$

So not all states can be transient. □

Note. 1. An infinite chain can easily be transient (consist only of transient states). A trivial example is moving to the right on $\mathbb{Z}$: $p_{i,i+1} = 1$: all states are transient. A highly non-trivial example is simple symmetric RW on $\mathbb{Z}^d$ for $d \geq 3$, by Pólya's Theorem (§2.9 below; see e.g. [Fel, XIV.7], [DoyS]).
2. Writing $\mathbb{E}_i$, $\mathbb{P}_i$ for expectation and probability starting at i,

$$p_{ij}(n) = \mathbb{P}_i(\text{in } j \text{ at time } n) = \mathbb{E}_i I(\text{in } j \text{ at time } n,$$

$$\sum_n p_{ij}(n) = \mathbb{E}_i \sum_n I(\text{in } j \text{ at time } n) = \mathbb{E}_i(\text{total time spent in } j).$$

This is finite if j is transient, infinite if j is persistent.

The Theorem above is now obvious. There is infinite time altogether. Only finite expected time can be spent in any transient state, so also in any finite union of transient states.

Lecture 12. 31.10.2022

Theorem. A persistent state j in a finite chain is positive (non-null).

Proof. If the finite chain has state-space $\{1, \cdots, N\}$, assume there is a null state. Let C_j be the equivalence class containing it. Since C_j is closed, we can consider the subchain induced on C_j. Then

$$1 = \sum_{k \in C_j} p_{ik}(n) \qquad \text{(finite sum)}.$$

Let $n \to \infty$: each $p_{ik}(n) \to 0$, so the sum on RHS $\to 0$, giving $1 = 0$. This contradiction gives the non-existence of null states in a finite chain. □

We quote the following important classical result, which despite its probabilistic relevance is an algebraic result about finite non-negative matrices (i.e., matrices with non-negative elements) with row-sums 1 (i.e. stochastic matrices). It is due to O. Perron (1907) and G. Frobenius (1908).

Theorem (Perron-Frobenius Theorem). Let P be the transition probability matrix of a finite irreducible Markov chain with period d.
(i) $\lambda_1 = 1$ is always an eigenvalue of P; if $d > 1$, so too are the other dth roots of unity, $\lambda_2 = \omega, \cdots, \lambda_d = \omega^{d-1}$, where $\omega := \exp(2\pi i/d)$.
(ii) All other eigenvalues λ_j have modulus $|\lambda_j| < 1$.

5. Limit distributions and invariant distributions

Recall: a state is ergodic if it is persistent, positive and aperiodic; a chain is ergodic if it is irreducible (all states the same type), with all states ergodic.

Theorem. In an ergodic chain (not necessarily finite):
(i) $\exists\ \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$, independent of i.
(ii) $\pi_j > 0$, and $\sum \pi_j = 1$.
(iii) $\pi_j = \sum_i \pi_i p_{ij}$, or in matrix notation, if $\pi := (\pi_0, \pi_1, \cdots)$,

$$\pi = \pi P.$$

Conversely, if (ii), (iii) hold for an irreducible aperiodic chain, (i) holds, $\pi_k = 1/\mu_k$ (μ_k the mean recurrence time of k), and the chain is ergodic.

Proof. By the Erdő-Feller-Pollard theorem,

$$p_{jj}^{(n)} \to \pi_j = 1/\mu_j \qquad (n \to \infty),$$

and $\pi_j > 0$ as the states are positive. As

$$p_{ij}^{(n)} = \sum_k \mathbb{P}(i \to j \text{ for the first time at } k)\, p_{jj}^{(n-k)},$$

letting $n \to \infty$, dominated convergence gives

$$p_{ij}^{(n)} \to \sum_k \mathbb{P}(i \to j \text{ for the first time at } k)\, \pi_j = f_{ij}\pi_j = \pi_j, \qquad \forall\ i,$$

as the chain is ergodic, proving (i). Now

$$1 = \sum_{j=1}^{\infty} p_{ij}^{(n)} \geq \sum_{j=1}^{N} p_{ij}^{(n)}$$

for each N. Let $N \to \infty$:

$$s := \sum_1^{\infty} \pi_j \leq 1.$$

$$p_{ij}^{(n+1)} = \sum_k p_{ik}^{(n)} p_{kj} \geq \sum_{k=1}^{N} p_{ik}^{(n)} p_{kj} \qquad \forall\ N.$$

Let $n \to \infty$:

$$\pi_j \geq \sum_{k=1}^{N} \pi_k p_{kj}.$$

Let $N \to \infty$:

$$\pi_j \geq \sum_k \pi_k p_{kj}. \qquad (*)$$

Sum over j:

$$s := \sum_j \pi_j \geq \sum_j \sum_k \pi_k p_{kj} = \sum_k \pi_k \sum_j p_{kj} = \sum_k \pi_k = s.$$

This chain of inequalities with equality at each end forces *equality* in each *inequality* in $(*)$, proving (iii).

For (ii): that $\sum_j \pi_j = 1$ follows formally from $\sum_j p_{ij}^{(n)} = 1$ and (i) on interchanging limit and sum. A full proof requires more care; see [Nor, §1.7, p.35 - 38] (cf. [GriS, §6.4, p.207-217]). □

Lecture 13, 3.11.2022

The d/n π in the Theorem is called the *limit distribution* of the chain, because $(p_{ij}^{(n)})_j$ is the d/n at time of starting from i and $p_{ij}^{(n)} \to \pi_j$ as $n \to \infty$. It is also an *invariant (stationary, equilibrium)* d/n, in the sense that if π is the *starting* d/n, the d/n after one step is πP, which is still π as $\pi = \pi P$. Similarly after any number of steps. So:

Cor. If an ergodic chain is started in its invariant (or limit) distribution π, it stays in it for all time.

Coupling

We mention in passing an important idea in proofs of the limit theorem for Markov chains: *coupling*. One constructs *two* independent copies of the chain on the *same* probability space; call them $X = (X_n)$ and $Y = (Y_n)$. Start X from state i; start Y in π, where π is the d/n we would like to prove is the limit d/n. If we can prove that eventually X and Y *meet* – with probability 1, there exists n with $X_n = Y_n$ – then we can imagine X and Y 'stuck together' on first meeting: say $Y_m := X_m$ for $m \geq n$. Using the Markov (lack of memory) property of Y at time n, this does not change the d/n of Y, which is $\phi = \pi P$. So Y stays forever in d/n π, so in particular has limit d/n π, so X has limit d/n $/pi$. For a nice short treatment, see [Nor, §1.8].

Examples.

1. *Gambler's ruin.*

There is no limit d/n. The chain is not irreducible: 0 and a are persistent (absorbing, traps), $1, /cdots, a-1$ are transient. There are, trivially, two different invariant d/ns, 'start at 0' and 'start at a'.

2. *Ehrenfest urn.*

Again, there is no limit d/n: the chain is periodic with period 2. But apart from this, the chain comes as close to having a limit d/n as possible: it has an *invariant* distribution, the *binomial* d/n:

$$\pi = (\pi_j), \qquad \pi_j = 2^{-d}\binom{d}{j} :$$

$$p_{i,i-1} = i/d, \qquad p_{i,i+1} = (d-i)/d, \qquad \pi_i = 2^{-d}\binom{d}{i},$$

$$(\pi P)_j = \pi_{j-1}p_{j-1,j} + \pi_{j+1}p_{j+1,j}$$
$$= 2^{-d}\binom{d}{j-1}\frac{(d-j+1)}{d} + 2^{-d}\binom{d}{j+1}\frac{j+1}{d}$$
$$= \frac{2^{-d}}{d}\Big(\frac{d!}{(j-1)!(d-j+1)!}(d-j+1) + \frac{d!}{(j+1)!(d-j-1)!}(j+1)\Big).$$

Now $(d-j+1)/(d-j+1)! = 1/(d-j)!$, $(j-1)! = j/j!$ in the first term, $(j+1)/(j+1)! = 1/j!$, $1/(d-j-1)! = (d-j)/(d-j)!$ in the second. So the RHS is

$$\frac{2^{-d}}{d}\binom{d}{j}\{j + (d-j)\} = 2^{-d}\binom{d}{j} = \pi_j.$$

This shows that $\pi = (\pi_i) = (2^{-d}\binom{d}{i})$ is invariant, as required. □

6. Eigenvalue decompositions

Recall that the rows of P sum to 1:

$$\sum_j p_{ij} = 1 \qquad \forall\ i$$

(from i, we must go somewhere). So with $\mathbf{1}$ for a column vector of 1s,

$$P\mathbf{1} = \mathbf{1}.$$

This says that $\lambda = 1$ is an *eigenvalue* of P, with eigenvector $\mathbf{1}$.

Recall the Perron-Frobeniuis theorem of §4: if the chain is finite and irreducible with period d,
(i) 1 is an eigenvalue, and if $d > 1$, so are the other $d-1$ dth roots of unity.
(ii) All other eigenvalues λ_j have modulus $|\lambda_j| < 1$.
The other eigenvalues are the roots λ of the characteristic equation

$$|\ P - \lambda I\ | = 0.$$

We assume for convenience that the eigenvalues (e-values, for short) are all distinct. This assumption can be weakened, but at the cost of more complicated algebra, and it is satisfied in the range of examples we shall consider. Then we can order the λ_i in order of decreasing modulus:

$$\lambda_1 = 1, \qquad 1 = |\lambda_1| \geq \cdots \geq |\lambda_d|.$$

Lecture 14, 4.11.2022

Let v_i be a right (column) eigenvector (e-vector) of λ_i, with corresponding left (row) e-vector u_i:

$$Pv_i = \lambda_i v_i, \qquad u_i P = \lambda_i u_i.$$

Form the matrices

$$U := \begin{pmatrix} , u_1 \\ \vdots \\ u_d \end{pmatrix}, \qquad V = (v_1, \cdots, v_d), \qquad \Lambda = diag(\lambda_i).$$

Then

$$UP = \Lambda P, \qquad PV = V\Lambda.$$

Because the λ_i are *distinct*, the u_i are *linearly independent*, and similarly for the v_i (we quote this from Linear Algebra). So U, V are non-singular, and

$$P = U^{-1}\Lambda U = V\Lambda V^{-1}.$$

This is called the *spectral decomposition* of P; the set of e-values is called the *spectrum* of P (cf. optics, and resonant frequencies in mechanics).

For $i \neq j$,

$$u_i P v_j = u_i \lambda_j v_j = \lambda_j u_i v_j,$$

and symmetrically

$$u_i P v_j = \lambda_i u_i v_j.$$

Subtract: as the e-values were assumed distinct,

$$(\lambda_j - \lambda_i) u_i v_j = 0 : \qquad u_i v_j = 0 \qquad (i \neq j).$$

Recall that e-vectors are determined only to within a scale factor (if x is an e-vector, so is cx for c a non-zero constant). We can choose the scale-factors to make

$$u_i v_i = 1 \qquad \forall\ i : \qquad u_i v_j = \delta_{ij}$$

(this assumes $u_i v_i \neq 0$; we quote this from Linear Algebra). Then

$$UV = I : \qquad U = V^{-1}.$$

Then

$$P = U^{-1}\Lambda U = V\Lambda U = \sum_i \lambda_i v_i u_i = \sum_i \lambda_i A_i,$$

where $A_i := v_i u_i$ is a $d \times d$ matrix (v is a $d \times 1$ column, u is a $1 \times d$ row) satisfying

$$A_i A_j = v_i u_i v_j u_j = 0 \qquad (i \neq j), \qquad A_i A_i = v_i u_i v_i u_i = v_i u_i = A_i \qquad (i = j):$$

$$A_i A_j = \delta_{ij} A_i.$$

$$\sum_i A_i = \sum_i v_i u_i = \begin{pmatrix} v_i & \cdots & v_d \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix} = VU = I.$$

Theorem (Spectral Decomposition). For each n, $P^n = \sum_{i=1}^d \lambda_i^n A_i$.

Proof. By induction, assume this holds for n (it holds for $n = 1$ by above). Then

$$P.P^n = (\sum_i \lambda_i A_i)(\sum_j \lambda_j^n A_j) = \sum_{ij} \lambda_i \lambda_j^n A_i A_j:$$

$$P.P^n = \sum_{ij} \lambda_i \lambda_j^n \delta_{ij} A_i = \sum_i {\lambda_i}^{n+1} A_i,$$

completing the induction. □

With $a = (a_i)$ the initial d/n (row-vector), as the u_i are linearly independent we can decompose a w.r.t. (u_i):

$$a = \sum_i \alpha_i u_i,$$

say. Then the d/n at time n is aP^n, with a and P as above.

Now $u_i v_j = \delta_{ij}$, so $av_j = \sum_i \alpha_i u_i v_j = \sum_i \alpha_i \delta_{ij} = \alpha_j$:

$$\alpha_i = av_i, \qquad a = \sum_i (av_i) u_i,$$

$$aP^n = \sum_i (av_i) u_i . \sum_j \lambda_j^n A_j = \sum_{ij} av_i \lambda_j^n . u_i A_j.$$

Lecture 15, 7.11.2022

But $u_i A_j = u_i v_j u_j = \delta_{ij} u_j$, so $\quad aP^n = \sum_i a v_i \lambda_i^n u_i$.
In the aperiodic case, the λ_1^n term is $1^n = 1$, while all other λ_i have modulus $|\lambda_i| < 1$, so $\lambda_i^n \to 0$ as $n \to \infty$. Then $\quad aP^n \to (av_1)u_1 \qquad (n \to \infty)$.
Now $a = (a_1, \cdots, a_d)$ is the initial d/n, v_1 is a column-vector of 1s, so $av_1 = \sum a_i = 1$. So

$$aP^n \to u_1 \qquad (n \to \infty),$$

for all initial d/ns a. Similarly,

$$P^n \to A_i = v_1 u_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} u_1 = \begin{pmatrix} u_1 \\ \vdots \\ u_1 \end{pmatrix},$$

a matrix (of rank 1) with *all rows identical*. In particular,

$$p_{ij}^n \to u_{1j} \qquad (n \to \infty) \qquad \forall\ i.$$

This is the matrix equivalent of the limit theorem

$$p_{ij}^n \to \pi_j \qquad (n \to \infty) \qquad \forall\ i,$$

and recovers the above result, identifying π with u_1:

Theorem. In the irreducible aperiodic case, the limit distribution π is u_1, the left (row) eigenvector corresponding to the Perron-Frobenius eigenvalue 1.

Rate of convergence.

All e-values λ_i with $|\lambda_i| < 1$ have $\lambda_i^n \to 0$ geometrically fast as $n \to \infty$. So the rates of convergence above are governed by the largest e-value modulus $|\lambda_i| < 1$. Then $1 - |\lambda_i|$ is called the *spectral gap*. For background, see e.g. P. DIACONIS and D. STROOCK, Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.* **1** (1991), 36-61.
Note.

The spectral decomposition $P^n = \sum \lambda_i^n A_i$, $A_i = v_i u_i$ is important because it splits the dependence of p_{ij}^n in n from that in i, j, enabling the evolution as n increases to be handled easily. So a finite Markov chain is usually regarded as 'solved' when the e-values λ_i and e-vectors u_i, v_i are known.

Example: The Ehrenfest urn (P. and T. Ehrenfest, 1907, 1912).

$P = (p_{ij})$, where $p_{i.i+1} = (d-i)/d$, $p_{i,i-1} = i/d$, $p_{ij} = 0$ else, $i, j = 0, \cdots, d$:

Theorem (Kac, 1947).
(i) $\lambda_j = 1 - 2j/d$ $(j = 0, 1, \cdots, d)$ (the period is 2; $\lambda_0 = 1$, $\lambda_d = -1$).
(ii) $p_{ik}^{(n)} = \sum_{j=0}^{d} b_{ij} b_{jk} \lambda_j^n$, where $B = (b_{ij})$, $B^2 = I$, $BP = \Lambda B$, $P^n = B\Lambda^n B$, where

$$2^{-\frac{1}{2}d}(1-z)^i(1+z)^{d-i} = \sum_{j=0}^{d} b_{ij} z^j.$$

Note. 1. The b_{ij} are related to the *Krawtchouk polynomials* (discrete orthogonal polynomials from special-function theory).

The corresponding result for the – closely related – Bernoulli-Laplace chain (§1, Ex. 6) is due to Karlin and McGregor (1961, 1962). The eigenvectors are again discrete orthogonal polynomials, the *Hahn polynomials*.
2. The motivation for the Ehrenfest urn is how to reconcile the observed *irreversibility* of physical phenomena such as heat flow in thermodynamics and statistical mechanics at *macroscopic* level with the *reversibility* (in time) of the laws of dynamics giving rise to them at *microscopic* level. Recall that the invariant d/n of the Ehrenfest urn is binomial: $\pi = (\pi_i)$, where $\pi_i = 2^{-d}\binom{d}{i}$. The 'extremal' states π_0 and π_d have $\pi_0 = \pi_d = 2^{-d}$. So the mean recurrence time of these states is $\mu_0 = 1/\pi_0 = 2^d$. Now d is the number of molecules present, and at macroscopic level this will be of the order of magnitude of Avogadro's number (c.6.02×10^{23} – enormous). Then 2^d is astronomically vast – effectively, infinite. So irreversibility corresponds to 'impossible' states being possible, but having so astronomically vast mean recurrence times that they are never seen in practice. This is the Ehrenfests' answer to objections by Mach, Loschmidt and others to Boltzmann's work in statistical mechanics (Paul Ehrenfest was a pupil of Boltzmann's). Tragically, Ludwig Boltzmann (1844 - 1906), with J. C. Maxwell and Willard Gibbs one of the three founding fathers of statistical mechanics, had been driven to suicide in 1906 by the criticism to which his work had been subjected to – only a year before the Ehrenfests' work, which could have saved him.
3. Recall that the Bernoulli-Laplace model (an earlier and more complicated variant on the Ehrenfest model) came much earlier (1769 and 1812). It seems to have been forgotten; perhaps it came 'before its time'. It could have saved Boltzmann! The relevant field of statistical mechanics 'took off' with the work of Maxwell (1872), Boltzmann (1872) and Gibbs (1870s; book of 1902).

Lecture 16, 10.11.2022

7. Reversibility

Definition. A stochastic process $X = (X_t)$ is *reversible* if, for times $t_1, \cdots, t_n$ and τ, $(X(t_1), \cdots, X(t_n))$ and $(X(\tau - t_1), \cdots, X(\tau - t_n))$ have the same distribution.

Thus reversibility allows one to *run time backwards*: if $t_1 < \cdots < t_n$, $\tau - t_1 > \cdots > \tau - t_n$.

Lemma. A reversible process is stationary.

Proof. Both $(X(t_1), \cdots, X(t_n))$ and $(X(t_1 + \tau), \cdots, X(t_n + \tau))$ have the same d/n as $(X(-t_1), \cdots, X(-t_n))$. So the finite-dimensional d/ns of X are invariant under translation through time τ, for any τ, and this is stationarity. □

Definition. A Markov chain (p_{ij}) satisfies the *detailed balance condition* if there exist $\pi_j > 0$, $\sum \pi_j = 1$ so that

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad \forall i, j. \tag{DB}$$

Theorem. A stationary Markov chain is reversible iff it satisfies the detailed balance condition (DB). Then $\pi = (\pi_i)$ is the invariant distribution.

Proof. If the process is reversible: by stationarity, $\mathbb{P}(X_t = j)$ is independent of t. Then writing $\pi_j := \mathbb{P}(X_t = j)$, $\pi_j > 0$ (if not, $\mathbb{P}(X_t = j) = 0$ for all t, and then we could exclude state j from the sample space), and $\sum \pi_j = 1 (= \sum \mathbb{P}(X_t) = j)$. By reversibility,

$$\mathbb{P}(X_t = j, X_{t+1} = k) = \mathbb{P}(X_t = k, X_{t+1} = j).$$

The LHS is $\mathbb{P}(X_t = j)\mathbb{P}(X_{t+1} = k | X_t = j) = \pi_j p_{jk}$, and similarly for the RHS, giving detailed balance,

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

Conversely, assume (DB). Sum over j: as $\sum_j p_{ij} = 1$ for each i,

$$\pi_i = \sum_j \pi_j p_{ji}, \qquad \forall\ i: \qquad \pi = (\pi_i) = \pi P.$$

So π is the invariant d/n.

Choose any sequence of states $j_0, \cdots, j_m$. Then

$$\mathbb{P}(X_t = j_0, \cdots, X_{t+m} = j_m) = \pi(j_0)p(j_0j_1)\cdots p(j_{m-1}j_m),$$

and

$$\mathbb{P}(X_u = j_m, \cdots, X_{u+m} = j_0) = \pi(j_m)p(j_mj_{m-1})\cdots p(j_1j_0).$$

On the first RHS, $\pi(j_0)p(j_0j_1) = p(j_1j_0)\pi(j_1)$, and then $\pi(j_1)p(j_1j_2) = p(j_12j_1)\pi(j_2)$, etc., giving the first RHS as $p(j_1j_0)p(j_2j_1)\cdots p(j_mj_{m-1})\pi(j_m)$, which is the second RHS on reversing the order of multiplication. So the RHSs are equal, so the LHSs are equal. Write $\tau := t + u + m$. We get that $(X_t, X_{t+1}, \cdots, X_{t+m})$ has the same d/n as $(X_{\tau-t}, X_{\tau-t-1}, \cdots, X_{\tau-t-m}$. From this it follows that $(X(t_1), \cdots, X(t_n))$ has the same d/n as $(X(\tau - t_1), \cdots, X(\tau - t_n))$ for all τ, n and integers t_i. For, we can choose $m \geq n$ so large that of the two sets of integer time-points $(t, t+1, \cdots, t+m)$, $(t_1, t_2, \cdots, t_n)$, the first includes the second. Then the joint d/n corresponding to the first contains that for the second, by omitting (or summing out) unwanted arguments, giving the required equality of d/ns. $\square$

Reversibility and Entropy

The concept of *entropy* originates in physics in 1865, in the work of Rudolf CLAUSIUS (1822 - 1888) as a measure of *disorder*. Clausius' 1865 paper in *Annalen der Physik* ends with two famous sentences:

"Die Energie der Welt ist constant.

Die Entropie der Welt strebt einem Maximum zu."

[The energy of the world (he means, universe) is constant. The entropy of the world (universe) strives towards a maximum.] These two statements have become known as the First Law of Thermodynamics (Principle of Conservation of Energy) and the Second Law of Thermodynamics. The Second Law was later reformulated by Ludwig BOLTZMANN (1844 - 1906) in his 1896 book *Vorlesungen über Gastheorie* [Lectures on Gas Theory] as his 'H-theorem': he defined the entropy H, and gave arguments to show that it increases with time. Now increase of entropy with time is obviously incompatible with time-reversibility. Hence the intense interest among physicists in the early 20th C on reversibility arguments, and the importance of the Ehrenfest and Bernoulli-Laplace models above.

Lecture 17, 11.11.2022

Reversibility and Electric Networks

Imagine a network of nodes $x, y, \cdots$ linked by wires $xy, \cdots$ of resistance R_{xy} ($R_{xy} = R_{yx}$ as current flow is reversible). In mathematical terms, we have a *graph*, with *vertices* X, y etc. and *edges* xy. Write

$$C_{xy} := 1/R_{xy},$$

the *conductance* of the edge xy,

$$C_x := \sum_y C_{xy}, \qquad P_{xy} := C_{xy}/Cx.$$

Then $P = (P_{xy})$ defines a Markov chain transition probability matrix (as $P_{xy} \geq 0$, $\sum_y P_{xy} = 1$), and this chain is *reversible*, as

$$C_x P_{xy} = C_{xy} = C_{yx} = C_y P_{yx}.$$

So detailed balance holds, and $C = (C_x)$ is the equilibrium d/n). Conversely, reversible chains can arise in this way from networks:
Reversibility characterises those ergodic chains which can arise from electric networks.
For an excellent treatment of the area, see Doyle & Snell [DoyS].

The electric analogy can be exploited by *cutting* (removing wires) and *shorting* (short-circuiting nodes, to bring them to the same potential).
Thomson's Principle: electric flow minimises energy dissipation; [DoyS] §3.5;
Rayleigh's Monotonicity Law: increasing resistance anywhere can only increase effective resistance everywhere; [DoyS] Ch. 4.
The motivation of [DoyS] is an electric poof of Pólya's Theorem (§2.9 below).
Reversibility in general.

For a monograph treatment, see Kelly [Kel].

Note: Traffic management

Experienced drivers (or passengers) will recognise that Rayleigh's Monotonicity Law is not applicable as it stands to the management of traffic on our roads. In general, increasing road space reduces congestion and time wasted in traffic jams (at the cost of money, damage to the environment, delays during construction and extra maintenance, etc.) But, new roads generate their own new traffic, etc. This is an important and specialised area!

8. Random walks; gambler's ruin

We turn to one of the simplest non-trivial stochastic processes *random walk* on $\mathbb{Z}$. Here each step X_n is ± 1:

$$\mathbb{P}(X_n = +1) = p, \qquad \mathbb{P}(X_n = -1) = q, \qquad p, q \in (0,1),\ p+q = 1.$$

When $p = q = \frac{1}{2}$, we have *symmetric*, or *simple*, random walk. The name (due to Pólya, 1921: Irrfahrt) suggests the motion of a drunkard wandering aimlessly. It also gives an idealised model of, say, the motion of a particle (in gas or fluid) subject to collisions with surrounding molecules under thermal agitation (Robert Brown (1773 - 1858) in 1828). For background, see e.g. [GriS, §3.9, 3.10, 5.3, 5.10, 13.7].

Consider $S_n := \sum_1^n X_k$ as the net gain of a gambler betting on heads (probability p of success at each toss). What is the probability a_n of being ahead (by 1) for the first time at time n? – or, writing T for the waiting time till we are first ahead ($+\infty$ if we never are!), what is the d/n of T? In particular, when is T a.s. finite, when is $\mathbb{E}[T]$ finite, and how does this depend on p, q? Similarly for $a_n^{(k)}$, the pr of being k units ahead for the first time at time n.

Theorem. Let T be the waiting time to first being ahead (by 1) in a gambling game, gaining 1 with probability p and losing 1 with probability$q = 1-p$ at each trial.
(i) In the unfavourable case ($p < q$): T has positive probability $1 - (p/q)$ of being $+\infty$ (and so $\mathbb{E}[T] = +\infty$).
(ii) in the fair case ($p = q = \frac{1}{2}$), $T < \infty$ a.s,, but $\mathbb{E}[T] = +\infty$.
(iii) In the favourable case ($p > q$),

$$\mathbb{E}[T] = \frac{1}{p-q} < +\infty.$$

Proof. Abbreviating 'for the first time' to 'fft',

$$\begin{aligned} a_n^{(2)} &:= \mathbb{P}(\text{net gain 2 for 1st time at nth trial}) \\ &= \sum_1^{n-1} \mathbb{P}(\text{net gain 1 fft at } k;\ \text{net gain 1 fft in next } n-k \text{ trials at } n) \\ &= \sum_1^{n-1} \mathbb{P}(\cdots)\mathbb{P}(\cdots) \qquad (\text{tosses independent}) : \end{aligned}$$

Lecture 18, 14.11.2022
Proof (continued)

$$a_n^{(2)} = \sum_1^{n-1} a_k a_{n-k},$$

if we consider the sequence of tosses as 'starting afresh' at time k.
So as $a_0 = 0$,

$$a_n^{(2)} = \sum_1^{n-1} a_k a_{n-k} = \sum_1^{n-1} a_k a_{n-k}.$$

The RHS is the *convolution* of (a_n) with itself. Form the generating function (GF), of the random variable T and the sequence (a_n) giving its distribution:

$$A(s) := \sum_0^\infty a_n s^n = \sum_0^\infty \mathbb{P}(T = n) s^n = \mathbb{E}[\ s^T\ ].$$

Now GFs *multiply* under convolution (just as characteristic functions – CFs – do; this comes from the independence, as probabilities of independent events multiply: $\mathbb{E}[s^{T+U}] = \mathbb{E}[s^T.s^U] = \mathbb{E}[s^T].\mathbb{E}[s^U]$ if T, U are independent). So

$$A^{(2)}(s) := \sum_0^\infty a_n^{(2)} s^n = A(s)^2.$$

Similarly, or by induction,

$$A^{(k)}(s) = A(s)^k.$$

This has a simple probabilistic interpretation: the time to first get k ahead is the sum of k independent times to 'get one ahead'.

Now $a_0 = 0$, $a_1 = p$, and for $n > 1$,

$$a_n = q.a_{n-1}^{(2)} :$$

for, to get ahead for the first time at time $n > 1$, we have to *lose* the first trial (pr q), and then get *two* ahead from there in the next $n - 1$ trials.

Multiply by s^n and sum over $n = 2, 3, \cdots$:

$$LHS = \sum_2^\infty a_n s^n = A(s) - ps;$$

$$RHS = q\sum_{n=2}^{\infty} a_{n-1}^{(2)} s^n = q.s\sum_{k=1}^{\infty} a_k^{(2)} s^k = qsA^{(2)}(s) = qsA(s)^2.$$

So $A(s)$ satisfies the *quadratic equation*

$$qsA(s)^2 - A(s) + ps = 0.$$

Solving,

$$A(s) = [1 \pm \sqrt{1 - 4pqs^2}]/(2qs).$$

Now $A(s)$, being a PGF, is bounded for $s \to 0$, but the + sign here gives $A(s)$ unbounded near 0. So:

$$A(s) = [1 - \sqrt{1 - 4pqs^2}]/(2qs).$$

Expanding $(1 - 4pqs^2)^{\frac{1}{2}}$ by the Binomial Theorem,

$$a_{2k} = 0, \qquad a_{2k-1} = \frac{(-)^k (4pq)^k}{2q}\binom{\frac{1}{2}}{k} \qquad (k = 1, 2, \cdots).$$

Now

$$\begin{aligned} A(1) &= \sum_0^{\infty} a_n \\ &= \sum_0^{\infty} \mathbb{P}(\text{net gain for 1st time after } n \text{ trials}) \\ &= \mathbb{P}(\text{net gain ever 1}). \end{aligned}$$

Since $1 - (p-q)^2 = (p+q)^2 - (p-q)^2 = 4pq$,

$$\sqrt{1 - 4pq} = \sqrt{(p-q)^2} = |p - q|.$$

So

$$A(1) = \sum_n \mathbb{P}(T = n) = \mathbb{P}(T < \infty) = \frac{1 - |p-q|}{2q}.$$

If $p < q$, this is

$$\frac{1 - (q-p)}{2q} = \frac{p}{q} \qquad (\text{as } p + q = 1).$$

If $p \geq q$, this is

$$\frac{1 - (p-q)}{2q} = \frac{q}{q} = 1.$$

Lecture 19, 17.11.2022
Proof (continued)
So

$$\mathbb{P}(T<\infty)=\mathbb{P}(\text{net gain ever 1})=\begin{cases} p/q<1, & (p<q);\\ 1, & (p\geq q).\end{cases}$$

To find $\mathbb{E}[T]$, we have to differentiate its GF $A(s)=\sum_0^\infty s^n\mathbb{P}(T=n)$:

$$A'(s)=\sum ns^{n-1}\mathbb{P}(T=n); \qquad A'(1)=\sum n\mathbb{P}(T=n)=\mathbb{E}[T]:$$

$$2qA'(s)=-\frac{1}{s^2}+\frac{\sqrt{1-4pqs^2}}{s^2}-\frac{1}{s}.\frac{\frac{1}{2}.2s(-4pq)}{\sqrt{1-4pqs^2}}$$

$$=\frac{4pq}{\sqrt{\cdots}}+\frac{\sqrt{\cdots}}{s^2}-\frac{1}{s^2}.$$

If $p=q$, $\sqrt{1-4pqs^2}=0$ when $s=1$; otherwise $4pq<1$, and $\sqrt{\cdots}>0$ when $s=1$. So $\mathbb{E}[T]=A'(1)=+\infty$ if $p=q=\frac{1}{2}$.
As before, when $p>q$, $\sqrt{\cdots}=p-q$ at $s=1$. So

$$2qA'(1)=\frac{4pq}{p-q}+p-q-1=\frac{4pq}{p-q}-2q:$$

$$\mathbb{E}[T]=A'(1)=\frac{2p}{p-q}-1=\frac{1}{p-q} \qquad (p+q=1).$$

This completes the proof. □

Gambling interpretation

Suppose we gamble in the fair case, with the strategy 'quit when first ahead'. Then our eventual net profit of 1 is *certain*. More: we can make our eventual certain profit as large as we like – N, say (repeat N times, or play in units of N).

Beware! Despite this, the strategy above is suicidal (even for $N=1$), because the waiting-time T till we quit has *infinite mean*. This is because we are quite likely to go far into net loss *before* realising our certain profit. In practice, our capital is finite, and so is our lifetime, and we have positive probability of going bankrupt, or dying, before quitting when ahead. No wonder the name *gambler's ruin* is used here.

The early history of probability theory was motivated by gambling questions (and 'martingale' once meant a gambling game of the type above).

Return to the origin

Let $u_n := \mathbb{P}(\text{at 0 at } n\text{th trial})$ (not necessarily for the first time). This can only happen for n *even*; it happens at time $2n$ iff there have been n successes and n failures. So

$$u_{2n} = \binom{2n}{n} p^n q^n.$$

The sequence (u_n) has GF

$$U(s) = \sum_0^\infty u_{2n} s^{2n} = (1 - 4pqs^2)^{-\frac{1}{2}}.$$

For,

$$\binom{-\frac{1}{2}}{n} = (-\frac{1}{2})(-\frac{3}{2}) \cdots (-\frac{1}{2} - n + 1)/n! = (-)^n \frac{(2n-1)}{2} \frac{(2n-3)}{2} \cdots . \frac{3}{2} . \frac{1}{2} / n!$$

$$= (-)^n (2n-1)(2n-3) \cdots 3.1/(2^n n!).$$

Insert factors $2.4. \cdots .2n$ top and bottom:

$$\binom{-\frac{1}{2}}{n} = (-)^n \frac{(2n)!}{2^n n! . 2^n n!} = \frac{(-)^n}{4^n} \binom{2n}{n}.$$

So

$$u_{2n} = (-)^n 4^n p^n q^n \binom{-\frac{1}{2}}{n} = (-4pq)^n \binom{-\frac{1}{2}}{n},$$

giving the GF

$$U(s) = \sum_0^\infty u_{2n} s^{2n} = \sum_0^\infty (-4pqs^2)^n \binom{-\frac{1}{2}}{n} = (1 - 4pqs^2)^{-\frac{1}{2}},$$

by the Binomial Theorem, as required.

By this and the Feller relation,

$$U(s) = (1 - 4pqs^2)^{-\frac{1}{2}} = 1/(1 - F(s)),$$

$$F(s) = 1 - \sqrt{1 - 4pqs^2} = 1 - \sum_0^\infty (-4pqs^2)^n \binom{\frac{1}{2}}{n}.$$

Lecture 20, 18.11.2022

Here

$$F(s) = \sum_0^\infty f_n s^n, \qquad f_n = \mathbb{P}(\text{1st return to the origin at time } n).$$

As above,

$$\binom{\frac{1}{2}}{n} = (\frac{1}{2})(-\frac{1}{2})\cdots(\frac{1}{2}-n+1)/n! = (-)^{n-1}(2n-3)(2n-5)\cdots 3.1.(-1)/(2^n n!)$$

Insert a factor $(2n-1)$ top and bottom. As above, we obtain

$$\binom{\frac{1}{2}}{n} = \frac{(-)^{n-1}}{(2n-1)}.\frac{1}{4^n}\binom{2n}{n}, \qquad f_{2n} = u_{2n}/(2n-1).$$

In particular, by above,

$$f := F(1) = 1 - \sqrt{1-4pq} = 1 - |p-q|.$$

So: (i) If $p = q = \frac{1}{2}$ (symmetric random walk), $f = 1$: return to the origin is *certain* (non-defective).
(ii) If $p \neq q$ (asymmetric random walk), $f = \mathbb{P}$(return to the origin in finite time) $= 1 - |p-q| < 1$, $\mathbb{P}$(no return to the origin) $= |p-q| > 0$, and return to the origin is *uncertain* (defective).

9. Random walk in higher dimensions; Pólya's Theorem

We can construct a two-dimensional analogue of symmetric random walk on the line, on $\mathbb{Z}^2$. Here each point (m,n) has four neighbours, $(m\pm 1, n\pm 1)$, and moves to each with probability $1/4$. In $\mathbb{Z}^3$, each point (i,j,k) has six neighbours $(i\pm 1, j\pm 1, k\pm 1)$, and moves to each with equal prob $1/6$. In $\mathbb{Z}^d$, we move to each of the $2d$ neighbours with equal pr $1/(2d)$.

What are the probabilistic properties of simple symmetric random walk in d dimensions? The answer depends on the dimension d (which gives the results geometric as well as probabilistic interest). The result below was proved by George Pólya in 1921. See e.g.
K. L. Chung, Pólya's work in probability, BLMS 19 (1987), 570-576.

Pascal's triangle.

First, we recall some facts about Pascal's triangle. From the recurrence relation for the binomial coefficients

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}, \qquad (bin)$$

we can build up the triangle starting from a central 1 at the top (row 0) with two 1s below left and right (row 1), filling in 1s at each end and other entries in each row as the sum of its upper left- and upper right-hand neighbours. Then (*bin*) translates into:
(i) the kth entry in row n is $\binom{n}{k}$,
(ii) each entry is the *number of routes* from the vertex to it, each step being 'down left' or 'down right'.
Proof of (bin)

The RHS $\binom{n+1}{k}$ is the number of ways of choosing k from $n+1$. Number them, and split according to whether the last one is chosen. If it isn't, all k are chosen from the first n ($\binom{n}{k}$ ways); if it is, the other $k-1$ are chosen from the first n ($\binom{n}{k-1}$ ways). □

Note. This is vastly preferable to using fractions of factorials and bringing to a common denominator!

Corollary.

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}.$$

Short proof. Equate coefficients of x^n left and right in

$$(1+x)^{2n} = (1+x)^n.(1+x)^n.$$

Nice proof. Draw Pascal's triangle, as far as row $2n$, highlighting its central entry $\binom{2n}{n}$ and row n. By (ii), each entry in row n is the number of routes from the vertex down to it. But by symmetry, 'reflecting in the nth row', this is also the number of routes from it down to $\binom{2n}{n}$. So there are $\binom{n}{k}^2$ routes from the vertex down to $\binom{2n}{n}$ crossing row n at $\binom{n}{k}$. Now sum over k. □

Lecture 21, 21.11.2022

Theorem (Pólya, 1921). Simple symmtric random walk is persistent (recurrent) in dimensions $d = 1, 2$, transient in dimensions $d \geq 3$.

Proof.
$d = 1$.

We proved above (§8) that return to the origin is certain (non-defective), so the random walk is persistent.
$d = 2$.

If $u_{2n} = \mathbb{P}(\text{return to the origin at time } 2n)$,

$$u_{2n} = \frac{1}{4^{2n}} \sum_{k=0}^{n} \binom{2n}{k, k, n-k, n-k} := \frac{1}{4^{2n}} \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!}.$$

For, the multinomial coefficient on RHS counts the number of ways of choosing k steps right and k left along the x-axis, $n-k$ steps up and $n-k$ down along the y-axis; each such choice has pr $1/4$ at each of its $2n$ steps, so pr $1/4^{2n}$ altogether.

Now

$$u_{2n} = \frac{1}{4^{2n}} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k}^2.$$

So by the Corollary,

$$u_{2n} = \frac{1}{4^{2n}} \binom{2n}{n}^2.$$

By Stirling's formula,

$$n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} \qquad (n \to \infty).$$

Substituting and cancelling,

$$u_{2n} \sim \frac{1}{\pi n} \qquad (n \to \infty).$$

So $U(1) = \sum u_n = \infty$, as $\sum 1/n$ diverges. So return to the origin is again certain, and the walk is persistent.

$d = 3$.

Similarly,

$$\begin{aligned} u_{2n} &= \frac{1}{6^{2n}} \sum_{j,k} \frac{(2n)!}{j!j!k!k!(n-j-k)!(n-j-k)!} \\ &= \frac{1}{2^{2n}} \binom{2n}{n} \sum_{j,k} \Big(\frac{1}{3^n} . \frac{n!}{j!k!(n-j-k)!} \Big)^2 . \end{aligned}$$

Now the terms in $\Big(\cdots \Big)$ are the terms of a probability d/n (a *trinomial* d/n). So the double sum is at most its maximum term (as the terms of a pr d/n sum to 1). This maximum term is attained when $j, k, n-j-k$ are all approximately equal – about $n/3$ each. Stirling's formula shows that this maximum term is $O(1/n)$ (check). So the double sum is $O(1/n)$ also. Stirling's formula also shows that

$$\frac{1}{2^{2n}} \binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}} = O(1/\sqrt{n})$$

(the calculation above). Combining,

$$u_{2n} = O(1/n^{3/2}).$$

So $U(1) = \sum u_n$ converges, like $\sum 1/n^{3/2}$. Since $U(1) < \infty$, return to the origin is uncertain (defective), and the random walk is transient.

$d \geq 3$.

The same argument applies as with $d = 3$: one obtains a multinomial (d-nomial) d/n, attaining its maximum when each suffix $j_1, \cdots, j_d$ is about n/d. Stirling's formula gives the sum as $O(1/n^{\frac{1}{2}(d-1)})$, and u_{2n} as $O(1/n^{\frac{1}{2}d})$ as above, so for $d \geq 3$, $U(1) = \sum u_n < \infty$, giving transience as above. $\square$

Note. Pólya's Theorem inspired the fascinating book [DoyS] by Doyle & Snell – highly recommended.

2. Geometrically, transience results in higher dimensions ($d \geq 3$) because there is 'more room' – more ways to *avoid* returning to the origin – in higher dimensions.

3. In 3 dimensions, the probability of returning to the origin is about 0.35, and the expected number of returns about $0.65 \sum k(0.35)^k \sim 0.53$ (Whipple & McCrea, 1940; see [Fel, p.360]).

Lecture 22, 24.11.2022

Ch. 3. MARKOV CHAINS: CONTINUOUS TIME; POISSON PROCESSES

1. The exponential distribution: Lack-of-memory property

Recall the *exponential distribution* $E(\lambda)$ with *parameter* $\lambda > 0$. This is the law on $\mathbb{R}_+$ with density

$$f(t) = \lambda e^{-\lambda t} I(t \geq 0).$$

Think of a random variable $T \sim E(\lambda)$ is representing the *lifetime* of some component. Given that it has been in use already for time $s > 0$, what information does that give us about its 'residual lifetime'? Plausible answers:
(a). Less than with new – things 'wear out' (human lifespan, etc!)
(b). More than with new: it's stood the test of time, got over initial 'teething troubles' etc. (RAF pilots in WWII, etc.); (c). No information.
It is (c) that we need to pursue here.

Theorem (Lack-of-memory property, memoryless property). A random variable $T : \Omega \to \mathbb{R}_+$ has the *lack-of-memory* property,

$$\mathbb{P}(T > s + t |\ T > s) = \mathbb{P}(T > t) \qquad \forall\ s, t \geq 0$$

iff it has an exponential distribution: $T \sim E(\lambda)$ for some $\lambda > 0$.

Proof. If $T \sim E(\lambda)$,

$$\mathbb{P}(T > s+t |\ T > s) = \mathbb{P}(T > s+t)/\mathbb{P}(T > s) = e^{-\lambda(s+t)}/e^{-\lambda s} = e^{-\lambda t} = \mathbb{P}(T > t).$$

Conversely, if T has the memoryless property, $g(t) := \mathbb{P}(T > t)$ satisfies

$$g(s + t) = g(s)g(t) \qquad \forall\ s, t \geq 0.$$

As $T > 0$, $g(1/n) > 0$ for some n. Then by induction

$$g(1) = g(1/n + \cdots + 1/n) = g(1/n)^n > 0 : \qquad g(1) = e^{-\lambda}$$

for some λ, $0 \leq \lambda < \infty$. Similarly, for integers $m, n \geq 1$,

$$g(m/n) = g(1/n)^m = g(1)^{m/n} = e^{-\lambda m/n} : \qquad g(r) = e^{-\lambda r} \quad (r \in \mathbb{Q}_+).$$

As g is decreasing: approximate each real t by rationals $r \le t \le s$:

$$e^{-\lambda r} = g(r) \ge g(t) \ge g(s) = e^{-\lambda s}.$$

Letting $s \downarrow t$, $r \uparrow t$ forces $g(t) = e^{-\lambda t}$, so $T \sim E(\lambda)$. □

2. The Poisson process

Renewal theory. Suppose that a new lightbulb (say) is installed at time $t = 0$, and used continuously until it fails. It is then replaced by a new one, used till it fails, then replaced, etc. We assume the lifetimes of the lightbulbs are $X_1, \cdots, X_n, \cdots$, iid with law F.

Write $S_n := \sum_1^n X_k$ ($S_0 = 0$) for the partial sums of (X_n)) ((S_n) is a *random walk* with *step-length* d/n F), and write

$$N_t, \ N(t) := \max\{k : S_k \le t\} \qquad (t \ge 0).$$

In the exponential case $F = E(\lambda)$, $N = (N_t)$ is called a *Poisson process* with *rate* (or *intensity*) λ, $Pp(\lambda)$. The name is from the link (below) with the *Poisson distribution* with *parameter* $\lambda > 0$, $P(\lambda)$,

$$\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!, \qquad (k = 0, 1, 2, \cdots).$$

Lemma. If $X_n \sim E(\lambda)$ are independent, $S_n := \sum_1^n S_k$ has density

$$g_n(x) := \lambda(\lambda x)^{n-1}e^{-\lambda x}/(n-1)!/qq(x > 0).$$

Proof. Each X_i has moment-generating function (MGF)

$$M(t) := \int_0^\infty \lambda e^{-\lambda x}e^{tx}dx = \int_0^\infty \lambda e^{-(\lambda - t)x}dx = \lambda/(\lambda - t) \ \ (t < \lambda).$$

Since adding independent random variables multiplies MGFs, S_n has MGF $\lambda^n/(\lambda - t)^n$. Differentiate $\int_0^\infty e^{-sx}dx = 1/s$ n times:

$$\int_0^\infty x^{n-1}e^{-sx}dx/(n-1)! = 1/s^n.$$

Replace s by $\lambda - t$, multiply through by λ^n and compare: g_n and the density of S_n both have MGF $\lambda^n/(\lambda - t)^n$. So by uniqueness of MGFs, S_n has density g_n. □

We can now prove the link with the Poisson d/n (so justifying the name Poisson process).

Lecture 23, 25.11.2022

Theorem. If $N = (N_t)$ is a Poisson process with rate λ, $N \sim Pp(\lambda)$, each N_t has the Poisson d/n with parameter λt, $N_t \sim P(\lambda t)$:

$$\mathbb{P}(N_t = n) = e^{-\lambda t}(\lambda t)^n/n! \qquad (n = 0, 1, 2, \cdots,\ t \geq t).$$

Proof. $\{N_t = n\} = \{S_n \leq t < S_{n+1}\}$ (there are exactly n failures by time t iff $S_n \leq t$ but $S_{n+1} > t$). This and $\{S_{n+1} \leq t\} \subset \{S_n \leq t\}$ give

$$\mathbb{P}(N_t = n) = \mathbb{P}(S_n \leq t < S_{n+1}) = \mathbb{P}(S_n \leq t) - \mathbb{P}(S_{n+1} \leq t).$$

By the Lemma, S_n, S_{n+1} have densities g_n, g_{n+1}, so

$$\begin{aligned} \mathbb{P}(N_t = n) &= \int_0^t \frac{e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}dx - \int_0^t \frac{e^{-\lambda x}\lambda^{n+1}x^n}{n!}dx \\ &= \frac{\lambda^n}{n!}\int_0^t (nx^{n-1}e^{-\lambda x} - \lambda x^n e^{-\lambda x})dx \\ &= \lambda^n t^n e^{-\lambda t}/n! \end{aligned}$$

(as $t^n e^{-\lambda t}$ has derivative $nt^{n-1}e^{-\lambda t} - \lambda t^n e^{-\lambda t}$). So $N_t \sim P(\lambda t)$, as required. □

The result extends. If the time-interval $[0, t]$ is subdivided into $[0, u]$ and $(u, t]$, by the lack-of-memory property of the exponential d/n, the lightbulb in use at time u behaves like a new one. So the number of Poisson points N_u in $[0, u]$, $N_t - N_u$ in $(u, t]$ are independent, Poisson $P(\lambda u), P(\lambda(t-u))$ respectively. The argument extends to any number of *disjoint* intervals (which need not be contiguous), and beyond intervals to (Lebesgue-)measurable sets:

Theorem. If $N \sim Pp(\lambda)$ and $A_1, \cdots A_n$ are disjoint measurable sets with measures $|A_k|$, the Poisson counts $N(A_n), \cdots, N(A_n)$ are independent, Poisson $P(\lambda|A_1|), \cdots, P(\lambda|A_n|)$.

3. Markov chains in continuous time: rates and jump chains

This lack-of-memory (or memoryless) property of the exponential d/n is clearly closely linked to the Markov property, in which the process 'forgets its past'. As in Ch. 2, we again have discrete states, which we label E_i, $i \in \mathbb{N}$ or $\mathbb{N}_N$, but now time is continuous. In particular, the time-set is now

uncountable. We will be brief; we follow Norris [Nor, Ch. 2.3]

Each state E_i, or just i, will have its own exponential law $E(q_i)$ with parameter $q_i > 0$ – the 'rate of leaving i'. When it leaves i, it will go to some other state j, with probability q_{ij}/q_i. Thus

$$\sum_{j \neq i} q_{ij}/q_i = 1 : \qquad \sum_{j \neq i} q_{ij} = q_i < \infty.$$

Writing

$$q_{ii} := -q_i,$$

and forming the matrix $Q := (q_{ij} : i, j \in I)$, this *Q-matrix* satisfies
(i) $0 \leq q_i = -q_{ii} < \infty$ for all i;
(ii) $q_{ij} \geq 0$ for all $i \neq j$;
(iii) the row-sums are zero: $\sum_{j \in I} q_{ij} = 0$ for all i.

Call the successive times spent in the successive states S_n ('S for stay, or sojourn'). The transitions to new states are the *jump times*, J_n ($J_0 = 0$: there is no jump at the initial time $t = 0$). So

$$S_n = J_n - J_{n-1}, \qquad J_n = \sum_1^n S_k.$$

The states jumped to at these jump times form a discrete-time Markov chain, the *jump chain*, $Y = (Y_n)$ say.

The new phenomenon here is the possibility of *infinitely many jumps in finite time*, if

$$\zeta := \sup_n J_n = \sum_1^\infty S_n < \infty.$$

This is called *explosion*. When it happens, it is convenient to send the chain to a 'graveyard state', call it ∞, and keep it there: $X_t = \infty$ for $t \geq \zeta$. Such a process is called *minimal*.

The simplest example of such a chain is that when all the q_i are equal, to λ, say, and all jumps are up by one, $i \mapsto i + 1$. The resulting process is then (§2 above) the *Poisson process* with *rate* λ, $\mathrm{Pp}(\lambda)$. There are three alternative descriptions, below (see e.g. [Nor, Th. 2.4.3]).

Lecture 24, 28.11.2022

Theorem (Poisson process definitions). For $X = (X_t : t \geq 0)$ an increasing, right-continuous, integer-valued process starting from 0, $\lambda \in (0, \infty)$, the following are equivalent:
(a) (jump chain/holding times): the holding times S_n are independent exponential $E(\lambda)$ and the jump chain is $Y_n = n$;
(b) (infinitesimal): X has independent increments and

$$\mathbb{P}(X_{t+h} - X_t = 0) = 1 - \lambda h + o(h), \qquad \mathbb{P}(X_{t+h} - X_t = 1) = \lambda h + o(h),$$

as $h \downarrow 0$, uniformly in t;
(c) (increments/Poisson): X has stationary increments and for each t, $X_t \sim P(\lambda t)$.

One can re-write (b) above as a differential equation. It is then the *forward equation* for the transition probability matrix $P(t)$ of X at time t, $P'(t) = P(t)Q$, showing the link with the Q-matrix above [Nor, Th. 2.1.1].

One thinks of a Poisson process of rate λ as one counting 'accidents' occurring 'completely at random' at rate λ. If one were to add two independent processes of this kind, one would expect to get another of the same type (Poisson) and for the rates to add. This is indeed what happens [Nor, Th. 2.4.4]:

Theorem. If $X \sim Pp(\lambda)$ and $Y \sim Pp(\mu)$ are independent, $X + Y \sim Pp(\lambda + \mu)$.

Continuing to think in this way of the points of a Poisson process as occurring 'completely at random', one would expect that knowing that an interval I contained one Poisson point, it would be distributed uniformly within I, and that if n Poisson points occurred, these n points would be independently and uniformly distributed in I. Again, this is what happens [Nor, Th. 2.4.5, 2.4.6]. (With n points, as the time-axis is ordered they are automatically ordered by size. They thus form the *order statistics* of the sample of n points.)

Most of the properties of states for the discrete-time chains of Ch. 2 go over to the *jump chain* here in discrete time. One has [Nor, Th. 3.4.1]: (i) if i is persistent/transient for the jump chain (Y_n), it is persistent/transient

for the continuous-time chain (X_t);
(ii) every state is transient or persistent;
(iii) transience and persistence are class properties (so the Solidarity Theorem extends to continuous time).
So we may call Q transient, etc., when the chain X it comes from is.

The results of Ch. 2 on invariant d/ns also go over [Nor, Th. 3.5.5]:

Theorem. For Q irreducible and persistent and μ a measure, the following are equivalent:
(i) $\mu Q = 0$;
(ii) $\mu P(t) = \mu$ for some $t > 0$;
(iii) $\mu P(t) = \mu$ for all $t > 0$.

As (ii), (iii) correspond to μ being invariant in discrete time, this justifies calling μ *invariant* if $\mu Q = 0$.

For irreducible and non-explosive chains: if the chain has an invariant d/n, π, then [Nor, Th. 3.6.2]

$$p_{ij}(t) \to \pi_j \qquad (t \to \infty) \qquad \forall\ i.$$

Time-reversibility also extends to continuous time [Not, Th. 3.7.1]. The chain (or its Q-matrix) and the measure π are in *detailed balance* if

$$\pi_i q_{ij} = \pi_j q_{ji} \qquad \forall\ i, j. \tag{DB}$$

Then for irreducible and non-explosive chains, reversibility and (DB) are equivalent, as before.

Ch. 4. MARKOV PROCESSES IN CONTINUOUS TIME; BROWNIAN MOTION

1. Markov Processes.

X is *Markov* if for each t, each $A \in \sigma(X_s : s > t)$ (the 'future') and $B \in \sigma(X_s : s < t)$ (the 'past'),

$$\mathbb{P}(A|X_t, B) = \mathbb{P}(A|X_t).$$

That is, if you know where you are (at time t), how you got there doesn't matter so far as predicting the future is concerned – equivalently, past and future are conditionally independent given the present.

The same definition applied to Markov processes in discrete time. If both time and state are discrete, the term *Markov chain* is usually used. These, and their transition probability matrices $P = (p_{ij})$, are considered in detail in Ch. 2.

Markov processes (and chains) have been much studied. They have an extensive and interesting theory, and they provide models for many of the standard situations studied in Applied Probability. See e.g. Norris [Nor], Asmussen [Asm], Meyn & Tweedie [MeyT].

A situation is Markov if knowing the present is all that is needed to study the future. Roughly speaking, non-Markovian situations, in which one needs to know not only the present but also how one got there, are much harder, and are usually intractable. Again roughly speaking, the two main kinds of dependence where one can get useful results are martingales [mgs] and Markov processes.

Lecture 25, 1.12.2022

X is said to be *strong Markov* if the Markov property holds with the *fixed* time t replaced by a *stopping time* T (a random variable). This is a real restriction of the Markov property in the continuous-time case (though not in discrete time).
Example. If we take T an exponentially distributed random variable, and define a stochastic process X by

$$X(t) = \begin{cases} 0, & \text{if } t \leq T, \\ t - T, & \text{if } t \geq T, \end{cases}$$

(unit speed, to the right at 0 till the exponential time T, then a left turn through $\pi/4$); the Markov property holds at any fixed time t, but not at T.

Another standard example of a process which is Markov but not strong Markov is provided by the *left-continuous* Poisson process, i.e., a (right-continuous) Poisson process (Ch. 3) made left-continuous at its jumps.

1a. Diffusions.

A diffusion is a path-continuous strong-Markov process such that for each time t and state x the following limits exist:

$$\mu(t,x) := \lim_{h \downarrow 0} \frac{1}{h} E[(X_{t+h} - X_t)|X_t = x],$$

$$\sigma^2(t,x) := \lim_{h \downarrow 0} \frac{1}{h} E[(X_{t+h} - X_t)^2|X_t = x].$$

Then $\mu(t,x)$ is called the *drift*, $\sigma^2(t,x)$ the *diffusion coefficient*.

Diffusions are closely linked to Brownian motion $B = (B_t)$ (below), and to martingales. We quote: the *Itô integral* allows one to integrate a suitable random integrand $Y = (Y_t)$ with respect to Brownian motion, thus defining a *stochastic integral* $\int_0^t Y(u)dB(u)$, or $\int_0^t Y dB$. One may then study *stochastic differential equations* (SDEs), such as

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t.$$

Under suitable conditions, such an SDE has a solution $X = (X_t)$, which is a diffusion with drift μ and diffusion coefficient σ. All this extends to the multidimensional case. In $\mathbb{R}^d$, X_t, μ are d-vectors, σ a $d \times d$ matrix.
Note. As with ODEs and PDEs, one needs to have existence theorems and uniqueness theorems – and one has more than one sense in which 'solution'

can be taken. With SDEs, one needs to discriminate between *weak* and *strong* solutions. For background, see e.g. Øksendal [Øks].

Generators.

Write $D = d/dx$ for the differentiation operator in one dimension, $D_i = \partial/\partial x_i$ in $\mathbb{R}^d$; thus $D^2 = d^2/dx^2$, $D_{ij} = \partial^2/\partial x_i \partial x_j$. Write

$$L_t := \frac{1}{2}\sigma(t,.)D^2 + \mu(t,.)D, \qquad \text{or} \qquad \frac{1}{2}\sum_{i,j=1}^{d} \sigma_{ij}(t,.)D_{ij} + \sum_{i=1}^{d} \mu_i(t,.)D_i;$$

then L is an elliptic differential operator (linear, second-order, partial if $d > 1$). Under suitable conditions, the parabolic PDE

$$L_t f + \partial f/\partial t = 0 \qquad (PPDE)$$

has as solutions the transition prob. density function for the diffusion X.

Example: Brownian motion. The prototype here is Brownian motion (below), where $\mu = 0$, $\sigma = 1$ (or I in higher dimensions), $L = \frac{1}{2}D^2$ (or $\frac{1}{2}\Delta$ in higher dimensions, with Δ the Laplacian) and $(PPDE)$ is the *heat equation*.

In one dimension, the usual treatment of diffusions uses the *scale function* and *speed measure*; see e.g. Breiman [Bre], Ch. 16, Rogers & Williams [R-W2], V.46, 47. Here one uses the total ordering of the real line (so this is specific to one dimension). In higher dimensions, one uses the Stroock-Varadhan approach via *martingale problems*; see [SV].

2. Gaussian Processes.

Recall the multivariate normal distribution $N(\mu, \mathbf{\Sigma})$ in n dimensions. If $\mu \in \mathbb{R}^n$, $\mathbf{\Sigma}$ is a non-negative definite $n \times n$ matrix, $\mathbf{X}$ has distribution $N\mu, \mathbf{\Sigma})$ if it has characteristic function

$$\phi_{\mathbf{X}}(\mathbf{t}) := E\exp\{i\mathbf{t}^T.\mathbf{X}\} = \exp\{i\mathbf{t}^T.\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\} \qquad (\mathbf{t} \in \mathbb{R}^n).$$

If further $\mathbf{\Sigma}$ is positive definite (so non-singular), $\mathbf{X}$ has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}$$

(*Edgeworth's Theorem*, 1893).

A process $X = (X_t)_{t\geq 0}$ is *Gaussian* if all its finite-dimensional distributions are Gaussian. Such a process can be specified by:
(i) a measurable function $\mu = \mu(t)$ with $EX_t = \mu(t)$,
(ii) a non-negative definite function $\sigma(s,t)$ with $\sigma(s,t) = cov(X_s, X_t)$.

Lecture 26, 2.12.2022

3. Brownian Motion (BM).

The Scottish botanist Robert Brown (1773 - 1858) observed pollen particles in suspension under a microscope in 1828 and 1829 (though Leeuwenhoek had observed the phenomenon before him – indeed, so had Lucretius in antiquity, in *De rerum naturae*), and observed that they were in constant irregular motion.

In 1900 L. Bachelier considered BM a possible model for stock-market prices – the first time BM had been used to model financial or economic phenomena, and before a mathematical theory had been developed.

In 1905 Albert Einstein considered BM as a model of particles in suspension, and used it to estimate *Avogadro's number* ($N \sim 6 \times 10^{23}$), based on the diffusion coefficient D in the *Einstein relation*

$$varX_t = Dt \qquad (t > 0).$$

Definition. **Brownian motion** (BM) on $\mathbb{R}$ is the process $B = (B_t : t \geq 0)$ such that:
(i) $B_0 = 0$;
(ii) B has stationary independent increments (so B is a Lévy process);
(iii) B has Gaussian increments: for $s, t \geq 0$, $B_{t+s} - B_s \sim N(0, t)$;
(iv) B has continuous paths: $t \mapsto B_t$ is continuous ($t \mapsto B(t, \omega)$ is continuous for all $\omega \in \Omega$).
[The path-continuity in (iv) can be relaxed by assuming it only a.s.; we can then get continuity by excluding a suitable null-set from our probability space.]

The fact that BM so defined *exists* is quite deep, and was first proved by Norbert Wiener (1894-1964) in 1923. In honour of this, BM is also known as the *Wiener process*, and the probability measure generating it – the measure W on $C[0, 1]$ (one can extend to $C[0, \infty)$) by

$$W(A) = P(B_\cdot \in A) = P(\{t \mapsto B_t(\omega)\} \in A)$$

for all Borel sets $A \in C[0, 1]$ is called *Wiener measure*.

So what we prove below (by constructing it!) is the *existence theorem*: BM *exists*. However, some authors omit the requirement (iv) of path-continuity, and then they prove the *continuity theorem*: BM has *continuous paths* (more

precisely: there exists a version of BM with continuous paths).

Covariance.

Before addressing existence, we first find the covariance function. For $s \leq t$, $B_t = B_s + (B_t - B_s)$, so as $EB_t = 0$,

$$cov(B_s, B_t) = E(B_s B_t) = E(B_s^2) + E[B_s(B_t - B_s)].$$

The last term is $E(B_s)E(B_t - B_s)$ by independent increments, $= 0$, so

$$cov(B_s, B_t) = E(B_s^2) = s \qquad (s \leq t): \qquad cov(B_s, B_t) = \min(s, t).$$

A Gaussian process (one whose finite-dimensional distributions are Gaussian) is specified by its mean function and its covariance function, so among centred (zero-mean) Gaussian processes, the covariance function $\min(s, t)$ serves as the signature of BM.

Finite-Dimensional Distributions.

For $0 \leq t_1 < \ldots < t_n$, the joint law of $X(t_1), X(t_2), \ldots, X(t_n)$ can be obtained from that of $X(t_1), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$. These are jointly Gaussian, hence so are $X(t_1), \ldots, X(t_n)$: the finite-dimensional distributions are *multivariate normal*. Recall that the multivariate normal law in n dimensions, $N_n(\mu, \Sigma)$ is specified by the mean vector μ and the covariance matrix Σ (non-negative definite) by its CF:

$$E \exp\{i\mathbf{u}^T\mathbf{X}\} = \exp\{i\mathbf{u}^T\mathbf{X} - \frac{1}{2}\mathbf{u}^T\Sigma\mathbf{u}\},$$

and when Σ is positive definite (so non-singular), the joint density is given by Edgeworth's theorem. So to check the finite-dimensional distributions of BM – stationary independent increments with $B_t \sim N(0, t)$ – it suffices to show that they are multivariate normal with mean zero and covariance $cov(B_s, B_t) = \min(s, t)$ as above.

Construction of BM.

It suffices to construct BM for $t \in [0, 1]$). This gives $t \in [0, n]$ by dilation, and $t \in [0, \infty)$ by letting $n \to \infty$.

First, take $L^2[0, 1]$, and any complete orthonormal system (cons) (ϕ_n) on it. Now L^2 is a Hilbert space, under the inner product and norm

$$\langle f, g \rangle = \int_0^1 f(x)g(x)dx \quad (\text{or } \int fg), \qquad \|f\| := (\int f^2)^{1/2}).$$

Lecture 27. 5.12.2022

By Parseval's identity,

$$\int_0^1 fg = \sum_{n=0}^{\infty} \langle f, \phi_n \rangle \langle g, \phi_n \rangle$$

(where convergence of the series on the right is in L^2, or in mean square: $\|f - \sum_0^n \langle f, \phi_k \rangle \phi_k\| \to 0$ as $n \to \infty$). Now take, for $s, t \in [0,1]$,

$$f(x) = I_{[0,s]}(x), \qquad g(x) = I_{[0,t]}(x).$$

Parseval's identity becomes

$$\min(s,t) = \sum_{n=0}^{\infty} \int_0^s \phi_n dx \int_0^t \phi_n(x) dx.$$

Now take (Z_n) independent and identically distributed $N(0,1)$, and write

$$B_t = \sum_{n=0}^{\infty} Z_n \int_0^t \phi_n(x) dx.$$

This is a sum of independent random variables. Kolmogorov's theorem on random series ('three-series theorem' – see e.g. [Brei] §3.4, [GriS], 7.11.35) says that it converges a.s. if the sum of the variances converges. This is $\sum_{n=0}^{\infty} (\int_0^t \phi_n(x) dx)^2$, $= t$ by above. So the series above converges a.s., and by excluding the exceptional null set from our prob. space (as we may), everywhere.

The Haar System (1910).

Define

$$H(t) = \begin{cases} 1, & \text{on } [0, \frac{1}{2}), \\ -1, & \text{on } [\frac{1}{2}, 1], \\ 0, & \text{else.} \end{cases}$$

Write $H_0(t) \equiv 1$, and for $n \geq 1$, express n in dyadic form as $n = 2^j + k$ for a unique $j = 0, 1, \ldots$ and $k = 0, 1, \ldots, 2^j - 1$. Using this notation for n, j, k throughout, write

$$H_n(t) := 2^{j/2} H(2^j t - k) \qquad (n = 2^j + k,\ j \geq 0,\ 0 \leq k < 2^j)$$

(so H_n has support $[k/2^j, (k+1)/2^j]$). So if $m \neq n$ have the same j, $H_m H_n \equiv 0$, while if m, n have different js, one can check that $H_m H_n$ is $2^{(j_1+j_2)/2}$ on half its support, $-2^{(j_1+j_2)/2}$ on the other half, so $\int H_m H_n = 0$. Also H_n^2 is 2^j on $[k/2^j, (k+1)/2^j]$, so $\int H_n^2 = 1$. Combining:

$$\int H_m H_n = \delta_{mn},$$

and (H_n) form an orthonormal system, called the *Haar system*. For completeness: the indicator of any dyadic interval $[k/2^j, (k+1)/2^j]$ is in the linear span of the H_n (difference two consecutive H_ns and scale). Linear combinations of such indicators are dense in $L^2[0,1]$. Combining: the Haar system (H_n) is a cons in $L^2[0,1]$.

The Schauder System (1927).

We obtain the *Schauder system* by integrating the Haar system. Consider the triangular function (or 'tent function')

$$\mathbb{D}(t) := \begin{cases} 2t, & \text{if } 0 \leq t \leq \frac{1}{2}, \\ 2(1-t), & \text{if } \frac{1}{2} \leq t \leq 1, \\ 0 & \text{else.} \end{cases}$$

Write $\mathbb{D}_0(t) := t$, $\mathbb{D}_1(t) := \mathbb{D}(t)$, and define the nth *Schauder function* $\mathbb{D}_n$ by

$$\mathbb{D}_n(t) := \mathbb{D}(2^j t - k) \qquad (n = 2^j + k \geq 1, \quad j \geq 0, \ 0 \leq k < 2^j).$$

Note that $\mathbb{D}_n$ has support $[k/2^j, (k+1)/2^j]$ (so is 'localized' on this dyadic interval, which is small for n, j large). We see that

$$\int_0^t H(u)du = \frac{1}{2}\mathbb{D}(t), \qquad \int_0^t H_n(u)du = l_n \mathbb{D}_n(t),$$

where $l_0 = 1$ and for $n \geq 1$,

$$l_n = \frac{1}{2}.2^{-j/2} \qquad (n = 2^j + k \geq 1).$$

The Schauder system $(\mathbb{D}_n)$ is again a cons on $L^2[0,1]$.

Lecture 28, 8.12.2022

THEOREM (Paley-Wiener-Zygmund, PWZ, 1933). For $(Z_n)_0^\infty$ independent $N(0,1)$ random variables, l_n, $\mathbb{D}_n$ as above,

$$B_t := \sum_{n=0}^{\infty} l_n Z_n \mathbb{D}_n(t)$$

converges uniformly on $[0,1]$, a.s. The process $B = (B_t : t \in [0,1])$ is Brownian motion.

Lemma. For Z_n independent $N(0,1)$,

$$|Z_n| \le C\sqrt{\log n} \qquad \forall n \ge 2,$$

for some random variable $C < \infty$ a.s.

Proof. For $x > 1$,

$$P(|Z_n| \ge x) = \frac{2}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}u^2} du \le \sqrt{2/\pi} \int_x^\infty u e^{-\frac{1}{2}u^2} du = \sqrt{2/\pi} e^{-\frac{1}{2}x^2}.$$

So for any $a > 1$,

$$P(|Z_n| > \sqrt{2a \log n}) \le \sqrt{2/\pi} \exp(-a \log n) = \sqrt{2/\pi}.n^{-a}.$$

Since $\sum n^{-a} < \infty$ for $a > 1$, the Borel-Cantelli lemma (Problems/Solutions 1; see e.g. [Brei] §3.3, or [G-S] §7.3 Th. 10) gives

$$P(|Z_n| > \sqrt{2a \log n} \text{ for infinitely many } n) = 0 : \qquad C := \sup_{n \ge 2} \frac{|Z_n|}{\sqrt{\log n}} < \infty \qquad a.s.$$

Proof of the Theorem.

1. Convergence.

Choose J and $M \ge 2^J$; then

$$\sum_{n=M}^{\infty} l_n |Z_n| \mathbb{D}_n(t) \le C \sum_{M}^{\infty} l_n \sqrt{\log n} \mathbb{D}_n(t).$$

The right is majorized by

$$C. \sum_{J}^{\infty} \sum_{k=0}^{2^j - 1} \frac{1}{2}.2^{-j/2} \sqrt{j+1} \mathbb{D}_{2^j + k}(t),$$

using $n = 2^j + k < 2^{j+1}$, $\log n \leq (j+1)\log 2$, and $\mathbb{D}_n(.) \geq 0$. So the series is absolutely convergent. In the inner sum, only one term is non-zero (t can belong to only one dyadic interval $[k/2^j, (k+1)/2^j)$), and each $\mathbb{D}_n(t) \in [0,1]$. So

$$LHS \leq C \sum_{j=J}^{\infty} \frac{1}{2}.2^{-j/2}\sqrt{j+1} \qquad \forall t \in [0,1],$$

and this tends to 0 as $J \to \infty$, so as $M \to \infty$. So the series $\sum l_n Z_n \mathbb{D}_n(t)$ is absolutely and uniformly convergent, a.s. Since continuity is preserved under uniform convergence and each $\mathbb{D}_n(t)$ (so each partial sum) is continuous, B_t is continuous in t.

2. Covariance.

By absolute convergence, we can interchange integral and expectation (Fubini's theorem):

$$\mathbb{E}[B_t] = \mathbb{E}[\sum_0^{\infty} l_n Z_n \mathbb{D}_n(t)] = \sum l_n \mathbb{D}_n(t).\mathbb{E}[Z_n] = \sum 0 = 0.$$

So the covariance is

$$\mathbb{E}[B_s B_t] = \mathbb{E}[\sum_m Z_m \int_0^s \phi_m . \sum_n Z_n \int_0^t \phi_n] = \sum_{m,n} \mathbb{E}[Z_m Z_n] \int_0^s \phi_m \int_0^t \phi_n,$$

$$= \sum_{m,n} \delta_{m,n} \int_0^s \phi_m \int_0^t \phi_n = \sum_n \int_0^s \phi_m \int_0^t \phi_n = \min(s,t),$$

by the Parseval calculation above.

3. Joint Distributions.

Take $t_1, \ldots, t_m \in [0,1]$, we have to show that $(B(t_1), \ldots, B(t_n))$ is multivariate normal, with mean vector 0 and covariance matrix $(\min(t_i, t_j))$. The multivariate CF is

$$\mathbb{E}\ [\exp\{i \sum_{j=1}^{m} u_j B(t_j)\}] = \mathbb{E}\ [\exp\{i \sum_{j=1}^{m} u_j \sum_{n=0}^{\infty} l_n Z_n \mathbb{D}_n(t)\}].$$

As the Z_n are independent, this is

$$\prod_{n=0}^{\infty} \mathbb{E}\ [\exp\{i l_n Z_n \sum_{j=1}^{m} u_j \mathbb{D}_n(t_j)\}].$$

Lecture 29, 9.12.2022

Proof (ctd)
As the $Z_n \sim SN(0,1)$, this is

$$\prod_{n=0}^{\infty} \exp\{-\frac{1}{2} l_n^2 (\sum_{j=1}^{m} u_j \mathbb{D}_n(t_j))^2\} = \exp\{-\frac{1}{2} \sum_{n=0}^{\infty} l_n^2 (\sum_{j=1}^{m} u_j \mathbb{D}_n(t))^2\}.$$

The sum in the exponent on the right is

$$\sum_{n=0}^{\infty} l_n^2 \sum_{j=1}^{m} \sum_{k=1}^{m} u_j u_k \mathbb{D}_n(t_j) \mathbb{D}_n(t_k) = \sum_{j=1}^{m} \sum_{k=1}^{m} u_j u_k \sum_{n=0}^{\infty} \int_0^{t_j} H_n(u)du. \int_0^{t_k} H_n(u)du.$$

This is

$$\sum_{j=1}^{m} \sum_{k=1}^{m} u_j u_k \min(t_j, t_k),$$

by the Parseval calculation, as (H_n) are cons. Combining,

$$\mathbb{E}\left[\exp\{i \sum_{j=1}^{m} u_j B(t_j)\}\right] = \exp\{-\frac{1}{2} \sum_{j=1}^{m} \sum_{k=1}^{m} u_j u_k \min(t_j, t_k)\}.$$

This says that $(B(t_1), \ldots, B(t_n))$ is multinormal with mean 0 and covariance function $\min(t_j, t_k)$ as required. This completes the construction of Brownian motion. □

Wavelets.

The Haar system (H_n) and the Schauder system $(\mathbb{D}_n)$ are examples of *wavelet systems*. The original function, H or $\mathbb{D}$, is a *mother wavelet*, and the 'daughter wavelets' are obtained from it by dilation and translation. The PWZ expansion is the *wavelet expansion of BM* with respect to the Schauder system $(\mathbb{D}_n)$. For any $f \in C[0,1]$, we can form its wavelet expansion

$$f(t) = \sum_{n=0}^{\infty} c_n \mathbb{D}_n(t); \qquad c_n = f(\frac{k+\frac{1}{2}}{2^j}) - \frac{1}{2}[f(\frac{k}{2^j}) + f(\frac{k+1}{2^j})].$$

are the *wavelet coefficients*. This is the form that gives the $\mathbb{D}_n(.)$ term its correct triangular influence, localized on the dyadic interval $[k/2^j, (k+1)/2^j]$.

Thus for f BM, $c_n = l_n Z_n$, with l_n, Z_n as above. The wavelet construction of BM above is, in modern language, the classical 'broken-line' construction of BM due to Lévy in his book of 1948. This account is from Steele [Ste].
Note.
1. The expansion in the PWZ Theorem can be generalised (*Karhunen-Loève expansions*).
2. We shall see that Brownian motion is a *fractal*, and wavelets are a useful tool for the analysis of fractals more generally.
3. Wavelets are very useful in *data compression*. This is because many signals with lots of 'local discontinuities' may be accurately summarized by a *sparse* wavelet expansion (one with only a few non-zero coefficients). For example, the FBI digitized its finger-print data bank using wavelets, 1993-94 (without this, the US criminal justice system would have collapsed long ago).
4. Backgound on wavelets: see e.g. the work of Professor Guy Nason here,
C. M. BRISLAWN: Fingerprints go digital, *Notices AMS* 42 (1995), 1278 - 1283.

We mention some of the many application areas:
(i) digitization and reproduction of images (near-perfect reconstruction is possible using only the 1% of wavelet coefficients largest in magnitude);
(ii) automatic recognition of: fingerprints (above); irises (electronic passports); vehicle number plates; bank cards for cashless transactions, etc.

Zeros.

It can be shown that Brownian motion *oscillates*:

$$\limsup_{t\to\infty} X_t = +\infty, \qquad \liminf_{t\to\infty} X_t = -\infty \qquad a.s.$$

Hence, for every n there are zeros (times t with $X_t = 0$) of X with $t \geq n$ (indeed, infinitely many such zeros). So, denoting the zero-set of $BM(\mathbb{R})$ by

$$Z := \{t \geq 0 : X_t = 0\} :$$

1. Z is an *infinite* set. We quote also:
2. Z is a (Lebesgue) *null* set: Z has Lebesgue measure zero.
3. Z is a *closed* set (contains its limit points – from path-continuity).
Less obvious are the next two properties:
4. Z is a *perfect* set: every point $t \in Z$ is a limit point of points in Z.
So there are *infinitely many* zeros in *every* neighbourhood of *every* zero (so the paths must oscillate amazingly fast!). This shows that *it is impossible to draw a realistic picture of a Brownian path.*

Lecture 30, 12.3.2022

5. **Brownian Scaling.** For each $c \in (0,\infty)$, $X(c^2t)$ is $N(0,c^2t)$, so $X_c(t) := c^{-1}X(c^2t)$ is $N(0,t)$. Thus X_c has all the defining properties of a BM (check). So, X_c **IS** a BM:

Theorem. If X is $BM(\mathbb{R})$ and $c > 0$, $X_c(t) := c^{-1}X(c^2t)$, then X_c is again a $BM(\mathbb{R})$.

Corollary. X is *self-similar* (reproduces itself under scaling), so a Brownian path $X(.)$ is a *fractal*. So too is the zero-set Z.

As a result of the incredibly fast oscillation of the paths and their fractal nature, it is not surprising that Brownian paths are a.s. *nowhere differentiable*. We quote this.

BM owes part of its importance to belonging to *all* the important classes of stochastic processes: it is (strong) Markov, a (continuous) martingale, Gaussian, a diffusion, a Lévy process (process with stationary independent increments), etc.

4. The Ornstein-Uhlenbeck process

Because Brownian paths are nowhere differentiable, using BM to model the movement of particles etc. is unrealistic, as *velocity* cannot be defined for them. Consequently, for a more realistic model, one may start with a velocity process, subject to frictional drag as it moves through a medium, and subject to random bombardment, which *can* be modelled by a BM.

The *Ornstein-Uhlenbeck process* (O-U process; G. E. Uhlenbeck and L. S. Ornstein, 1930) models a velocity process $V = (V_t)$, given by the following stochastic differential equation (SDE): for some $\beta > 0$,

$$dV_t = -\beta V_t + dB_t \qquad (V_0 = 0), \tag{OU}$$

where $B = (B_t)$ is Brownian motion. Using the integrating factor $e^{\beta t}$ gives $d(e^{\beta t}v_t) = e^{\beta t}dB_t$. Integrating from 0 to t,

$$V_t e^{\beta t} = \int_0^t e^{\beta u} dB_u : \qquad V_t = \int_0^t e^{-\beta(t-u)} dB_u.$$

So V_t is Gaussian, has mean 0, and is continuous. For the covariance:

$$cov(V_t, V_{t+s}) = \mathbb{E}\,[(\int_0^t e^{-\beta(t-u)} dB_u)(\int_0^{t+s} e^{-\beta(t+s-v)} dB_v] \qquad (s > 0).$$

Split the $\int_0^{t+s}$ term into $\int_0^t + \int_t^{t+s}$. By independence of the Brownian increments in $[0, t]$ and $[t, t+s]$, the 2nd term here gives no contribution, so

$$\begin{aligned} cov(V_t, V_{t+s}) &= \mathbb{E}\,[(\int_0^t e^{-\beta(t-u)} dB_u)(\int_0^t e^{-\beta(t+s-v)} dB_v] \\ &= e^{-2\beta t} e^{-\beta s} \int_0^t \int_0^t e^{\beta u} e^{\beta v} \mathbb{E}[dB_u dB_v]. \end{aligned}$$

By independence of Brownian increments, $\mathbb{E}[dB_u dB_v] = 0$ for $u \neq v$, while for $u = v$, $\mathbb{E}[(dB_u)^2] = du$ (Lévy's result on quadratic variation of Brownian motion; we quote this). So

$$cov(V_t, V_{t+s}) = e^{-2\beta t} e^{-\beta s} \int_0^t e^{2\beta u} du = e^{-2\beta t} e^{-\beta s} [e^{2\beta t} - 1]/(2\beta)$$

$$= e^{-\beta s} [1 - e^{-2\beta t}]/(2\beta).$$

So as $t \to \infty$,

$$cov(V_t, V_{t+s}) \to e^{-\beta s}/(2\beta) \qquad (s > 0),$$

with a similar calculation for $s < 0$. Combining, we obtain

$$cov(V_t, V_{t+s}) \to e^{-\beta|s|}/(2\beta) \qquad (t \to \infty).$$

So to within a scale factor, the *equilibrium* process has covariance $e^{-\beta|s|}$. This result is obtained via stationarity by Revuz and Yor [RevY, p.37]. See also Rogers and Williams [RogW, p.115].

The O-U *velocity* process V is Markov: the past before time t has played no role, and we have obtained its exact distribution (Gaussian with mean 0 and the above covariance) with no reference to it. Integrating it to obtain the O-U *displacement* process U is *not* Markov: knowing its position does not give full information about the future, as the direction of travel is needed to know e.g. whether it will move first to the left or the right. However, the bivariate process (U, V) is Markov.