MAS275 Probability Modelling

2 Renewal Theory

2.1 Renewal processes in discrete time

Renewal processes form a family of stochastic processes with a fairly simple definition. They also appear in the study of the long-term behaviour of Markov chains, so understanding their theory will be useful when we come back to Markov chains later in the course.

We suppose that on a discrete scale, from time to time there is an occurrence called a **renewal**, and at the points of time between these renewals, nothing happens. For example, we assume that the light bulb in a room is inspected at regular intervals, and if the bulb is found to have failed it is replaced with a new one; the replacement of the light bulb here is a renewal.

We will model the lengths of time between renewals (e.g. the life-lengths of light bulbs) as random variables, and we will assume that these lengths of time are **independent, identically distributed (i.i.d.)** random variables, which, because of the discrete time scale, will be positive integer valued. Denote them by $T_1, T_2, T_3, \ldots$, and assume for the moment that the process behaves as if there was an initial renewal at time zero. Then the "clock times" of successive renewals will be

$$0, T_1, T_1 + T_2, T_1 + T_2 + T_3, \ldots.$$

So the random behaviour of the process will be determined by specifying the (common) distribution of $T_1, T_2, T_3, \ldots$. We use the notation

$$f_n = P(T_i = n) \text{ for } n = 1, 2, 3, \ldots$$

to denote this distribution. (We do not allow two renewals at the same time, so T_i cannot be zero. It will be useful to recognise this by adopting the convention that $f_0 = 0$.)

A process constructed in this way is called a **renewal process** (in discrete time). We may be interested in the number of renewals, N_t say, which have occurred up to and including time t. In the light bulb example this is the number of light bulbs which have been replaced. Considering the process in this way allows us to fit it into the framework of stochastic processes described previously.

Write $f = \sum_{n=1}^{\infty} f_n$. In applications such as the light bulb case, the f_n will form a proper probability distribution in the sense that $f_n \geq 0$ and $f = 1$.

However, in some applications we allow the possibility that

$$f = \sum_{n=1}^{\infty} f_n < 1$$

giving what is known as a **defective** probability distribution; the positive number $1 - f$ is called the **defect** of the distribution. This is interpreted as follows: with probability $1 - f$ the random variable T_i will take the value infinity (in the light bulb example, this means the ith light bulb is "immortal") and the first such infinite T_i represents the event that the ith inter-renewal interval never ends, i.e. there is never any further renewal after the ith renewal.

The process is called **recurrent** if $f = 1$ and **transient** if $f < 1$. If a renewal process is recurrent, then renewals will continue for ever; if it is transient, then a fairly straightforward calculation (see Exercise 19) shows that with probability 1 there will be an infinite T_i for some i, and so renewals will stop.

If a renewal process is recurrent, then the mean of the random variables T_i may or may not be finite. The first case is called the **positive recurrent** case and the second is called the **null recurrent** case. This dichotomy will be important later in the course when we study the long term behaviour of Markov chains.

Example 8. *Bernoulli trials*

2.2 Generating functions

Given a sequence $(a_n) = a_0, a_1, a_2, \ldots$, we can define the **generating function** of the sequence (a_n) to be the power series

$$A(s) = \sum_{k=0}^{\infty} a_k s^k,$$

where s is a so-called dummy variable. In general, the generating function may or may not converge, but in the case where the a_k are probabilities and therefore non-negative and not greater than 1, it can be seen that it converges at least for $|s| < 1$ by comparison with a geometric series. The generating function is a way of encoding the information about the sequence in a function; note that it is possible to recover the sequence (a_n) from the generating function by repeated differentiation:

$$\frac{d^n}{ds^n} A(s) = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} a_k s^{k-n},$$

and evaluating this at zero gives

$$a_n = \frac{\frac{d^n}{ds^n} A(s) \mid_{s=0}}{n!}.$$

Generating functions are particularly useful in probability because if X is a random variable taking non-negative integer values and we write $f_n = P(X = n)$, we can use the definition of expectation to interpret the generating function $F_X(s)$ of the sequence $f_0, f_1, f_2, \ldots$ (which we refer to as the generating function – or **probability generating function** – of the distribution of X) as

$$F_X(s) = \sum_{k=0}^{\infty} f_k s^k = \sum_{k=0}^{\infty} P(X = k) s^k = E(s^X).$$

A simple application of this is the following result, which tells us that the addition of independent random variables can be neatly represented using generating functions:

Lemma 3. *If we have two independent random variables X and Y the generating function of the distribution of their sum $X+Y$ is the product of the generating functions of the distributions of X and Y.*

Proof. By independence,

$$F_{X+Y}(s) = E(s^{X+Y}) = E(s^X s^Y) = E(s^X)E(s^Y) = F_X(s)F_Y(s).$$

□

We can extend this by induction to adding together larger numbers of independent random variables, and we also note that the same result applies to defective probability distributions.

Another useful fact is that if $F(s)$ is the generating function of a non-defective probability distribution on the non-negative integers, then within the radius of convergence

$$F'(s) = \sum_{n=1}^{\infty} f_n n s^{n-1},$$

and so in particular if $F(s)$ is differentiable at $s = 1$ we have

$$F'(1) = \mu = E(X),$$

where μ is the expected value of a random variable X with the given distribution. Hence we can find the mean by differentiating the generating function and setting s equal to 1.

We additionally note that if $F(s)$ is the generating function of the (possibly defective) distribution of a random variable taking non-negative integer values, then because $\sum_{k=0}^{\infty} a_k \leq 1$ we have that for $0 \leq s < 1$

$$F(s) = \sum_{k=0}^{\infty} a_k s^k \leq \sum_{k=0}^{\infty} a_k \leq 1.$$

Example 9. *Examples of generating functions*

Example 10. *Calculation of mean using generating function*

2.3 Generating functions and renewal processes

For each $n = 0, 1, 2, \ldots$ let E_n be the event that a renewal takes place at (clock) time n, and let

$$u_n = P(E_n). \tag{1}$$

Because, after a renewal at time t, the process starts again as if from the beginning, we can also think of this as

$$u_n = P(E_{t+n}|E_t), \tag{2}$$

which by the construction of the process does not depend on t. (Note that this implies $u_0 = 1$, consistent with the idea that we think of there being a renewal at time 0.)

Then (u_n) is a sequence of probabilities, but it is **not** a probability distribution, since occurrences of renewals at different clock times are not mutually exclusive events; in particular there is no need for it to sum to 1 and usually it will not do so.

We can also write the probabilities f_n in terms of the events $E_0, E_1, E_2, \ldots$: we have

$$f_n = P(E_1^c, E_2^c, \ldots, E_{n-1}^c, E_n).$$

Again, because the process starts again as if from the beginning after a renewal, we can also write

$$f_n = P(E_{t+1}^c, E_{t+2}^c, \ldots, E_{t+n-1}^c, E_{t+n}|E_t).$$

(**Summary of the difference between f_n and u_n**: f_n is the conditional probability, given that there was a renewal at time t, that the next renewal is at time $t+n$, whereas u_n is simply the conditional probability that a renewal, not necessarily the next one, occurs at time $t+n$.)

We may calculate

$$\begin{aligned} u_1 &= f_1 \\ u_2 &= f_2 + f_1^2 \\ u_3 &= f_3 + 2f_1f_2 + f_1^3 \end{aligned}$$

and so on: for example, the second equation is found by noting that there are two ways a renewal can occur at clock time 2: either the first renewal (after 0) occurs at that time, or the first renewal after t occurs at time 1 and then a second renewal occurs at time 2.

To generalise these equations, we consider the generating functions of the sequences (u_n) and (f_n)

$$\begin{aligned} F(s) &= \sum_{k=1}^{\infty} f_k s^k \\ U(s) &= \sum_{k=0}^{\infty} u_k s^k. \end{aligned}$$

We obtain the following result:

Theorem 4. *The generating functions $F(s)$ and $U(s)$ satisfy*

$$U(s) = \frac{1}{1 - F(s)} \text{ for } 0 \le s < 1.$$

Proof. We note that for $n \ge 1$

$$u_n = P(E_n) = P(T_1 + T_2 + \ldots + T_k = n \text{ for some } k).$$

Now the event $\{T_1 + T_2 + \ldots + T_k = n\}$ can only happen for at most one value of k, since, by construction, the T_i are positive integer valued random variables and so at most one renewal can occur at any particular time point. It follows that

$$E_n = \bigcup_{k=1}^{\infty} \{T_1 + T_2 + \ldots + T_k = n\}$$

and that this is a disjoint union. We have written this as a union of an infinite sequence of events for convenience, but in fact for each n it is really only a finite union because for $k > n$ the event $\{T_1 + T_2 + \ldots + T_k = n\}$ is empty, again because the T_i are positive integer valued.

It follows that

$$u_n = \sum_{k=1}^{\infty} P(T_1 + T_2 + \ldots + T_k = n)$$

where, similarly, only at most the first n terms of this series are positive, and the rest are zero.

Because adding together independent random variables corresponds to multiplying their generating functions (by Lemma 3), the random variable $T_1 + T_2 + \ldots + T_k$ has the generating function $(F(s))^k$ for each $k = 1, 2, \ldots$. If we now go back to our expression for u_n, multiply both sides by s^n, and then sum these equations over all n (including the trivial $u_0 = 1$) we get

$$\begin{aligned} U(s) &= 1 + \sum_{k=1}^{\infty} \left(\sum_{n=1}^{\infty} P(T_1 + T_2 + \ldots + T_k = n) s^n \right) \\ &= 1 + \sum_{k=1}^{\infty} (F(s))^k. \end{aligned}$$

This is the sum of an infinite geometric series, and the fact, already noted, that $0 \leq F(s) < 1$ ensures that it converges for $0 \leq s < 1$. Summing this series, we get the result:

$$U(s) = \frac{1}{1 - F(s)} \text{ for } 0 \leq s < 1.$$

□

In principle this may be used to obtain $U(s)$ in terms of $F(s)$, or vice versa.

Example 11. *Theorem 4 for Bernoulli trials*

Example 12. *Bernoulli trials with "blocking"*

Theorem 4 is a relationship between functions which we know exist and are finite for $0 \leq s < 1$. Consider what happens when s approaches 1. Note that

$$F(1) = \sum_{k=1}^{\infty} f_k = f$$

where f is as defined previously, and so we conclude that $F(1) = 1$ if the process is recurrent and $F(1) < 1$ if the process is transient. But then, in Theorem 4, if the process is recurrent then as $s \to 1$ the right hand side tends to infinity since the denominator of the fraction tends to zero, whereas if the process is transient then the right hand side tends to a finite limit. We conclude that

- if $U(1) = \sum_{n=0}^{\infty} u_n = \infty$ then the process is **recurrent**;
- if $U(1) = \sum_{n=0}^{\infty} u_n < \infty$ then the process is **transient**.

This gives us an alternative criterion for recurrence which is sometimes useful.

2.4 Simple random walk

A **simple random walk** on the integers is a Markov chain whose state space is the integers $\mathbb{Z}$ and with transition probabilities

$$p_{ij} = \begin{cases} p & j = i+1 \\ 1-p & j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

(So, at each step, we move up one with probability p and down one with probability $1-p$, in a way which is independent of how we got to where we are now. As we often do, we will sometimes write q for $1-p$.)

Assume the walk starts at 0. Let E_n be the event that after n steps we return to zero. Then because whenever this happens occurs the process effectively "starts again from scratch" independently of what went before, we may regard visits to zero as forming a renewal process. (This is in fact true for returns to the starting point in any Markov chain; we will come back to this later in the course.)

We can ask whether this renewal process is recurrent or transient: will the walk keep returning to its starting point, or will it eventually leave and never come back?

Theorem 5. *Returns to zero in the simple random walk, started from zero, are recurrent if* $p = \frac{1}{2}$ *and transient otherwise.*

Proof. Obviously E_n cannot occur if n is odd, because return to zero can only occur after an even number of steps. So $u_n = 0$ for odd n. If n is even, $n = 2m$ say, then return to zero occurs if and only if the first $2m$ steps contain m upward steps and m downward steps, and since the number of upward steps in $2m$ steps has the binomial distribution $Bi(2m, p)$, we may immediately write down, for $m \geq 1$,

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

But note that

$$\begin{aligned}\binom{2m}{m} &= \frac{2m.(2m-1).(2m-2).\dots.2.1}{(m!)^2} \\ &= 2^m \frac{(2m-1).(2m-3).\dots.3.1}{m!},\end{aligned}$$

cancelling each even number in the numerator with a factor in the denominator.

Now, reversing order of factors in the numerator and dividing each by -2 gives

$$\binom{2m}{m} = (-4)^m \frac{(-\frac{1}{2}).(-\frac{3}{2}).\dots.(-\frac{1}{2}-(m-1))}{m!}.$$

Substituting back,

$$u_{2m} = \frac{(-\frac{1}{2}).(-\frac{3}{2}).\dots(-\frac{1}{2}-(m-1))}{m!}(-4pq)^m$$

and so, by the binomial expansion with negative non-integer index and the definition that $u_0 = 1$,

$$\begin{aligned} U(s) &= u_0 + \sum_{m=1}^{\infty} \frac{(-\frac{1}{2}).(-\frac{3}{2}).\dots(-\frac{1}{2}-(m-1))}{m!}(-4pqs^2)^m \\ &= 1 + \sum_{m=1}^{\infty} \frac{(-\frac{1}{2}).(-\frac{3}{2}).\dots(-\frac{1}{2}-(m-1))}{m!}(-4pqs^2)^m \\ &= (1-4pqs^2)^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{1-4pqs^2}}. \end{aligned}$$

Putting $s = 1$, we get

$$U(1) = \frac{1}{\sqrt{1-4pq}}.$$

Noting that

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

we deduce that

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

Hence if $p = \frac{1}{2}$ then $U(1) = \infty$ (so returns are recurrent) whereas if $p \neq \frac{1}{2}$ then $U(1) < \infty$ (so returns are transient). □

We can also ask whether in the recurrent case the expected time to return is finite or infinite.

Theorem 6. *In the simple random walk started from* 0 *with* $p = \frac{1}{2}$, *returns to zero are null recurrent.*

Proof. Putting $p = q = \frac{1}{2}$, we can see from the proof of Theorem 5 that

$$U(s) = \frac{1}{\sqrt{1-s^2}}.$$

It follows easily from Theorem 4 that

$$F(s) = 1 - \sqrt{1-s^2}.$$

Differentiating,

$$F'(s) = -\frac{1}{2}(1-s^2)^{-\frac{1}{2}}.(-2s) = \frac{s}{\sqrt{1-s^2}}$$

which tends to infinity as $s \to 1$, because the denominator tends to zero. Hence the mean inter-renewal time $F'(1)$ is infinite and the process is null recurrent.

□

2.5 Periodicity

The simple random walk of the previous section is an example where the only values of n for which $u_n > 0$ are the even numbers, namely the multiples of 2; we call 2 the period of the renewal process. More generally, we define the **period** of any renewal process as

$$d = \text{h.c.f.}\{n : f_n > 0\},$$

where h.c.f. stands for highest common factor. In other words, d is the largest positive integer such that a renewal can only occur at time n if n is a multiple of d. d is not necessarily the smallest n such that $f_n > 0$; for example, if $f_1 = 0$ but $f_2, f_3, f_4, \ldots > 0$ then

$$d = \text{h.c.f.}\{2, 3, 4, \ldots\} = 1$$

and so in this case $f_d = 0$.

If $d = 1$ (which will often be the case) we call the process **aperiodic** and if $d > 1$ then we call it **periodic with period** d.

Example 13. *Equalisations in rolling a six-sided dice*

2.6 Delayed renewal processes

It is sometimes useful, instead of assuming that a renewal has happened at time zero, to allow a random length of time D, known as a **delay**, to elapse until the first renewal occurs, after which the process carries on as before, independently of the length of the delay. This delay will have its own (possibly defective) distribution, whose generating function we will denote by

$$B(s) = \sum_{n=0}^{\infty} b_n s^n$$

where $b_n = P(D = n)$ is the probability that the delay is of length n, $n \geq 0$. The clock times of renewals are now $D, D + T_1, D + T_1 + T_2, \ldots$.

Previously, we had two definitions, (1) and (2), for u_n. In the non-delayed case these were the same, but with a delay we now need to distinguish them. We introduce the notation v_n for the probability of a renewal happening at time n, $v_n = P(E_n)$ (where again we let E_n be the event that there is a renewal at time n) with corresponding generating function

$$V(s) = \sum_{n=0}^{\infty} v_n s^n.$$

We retain the notation u_n for $P(E_{t+n}|E_t)$, the probability that, given that there is a renewal at time t, there is another renewal n steps later, and we write the corresponding generating function $U(s)$.

In this context, we define

$$f_n = P(T_i = n) = P(E^c_{t+1}, E^c_{t+2}, \ldots, E^c_{t+n-1}, E_{t+n}|E_t),$$

with corresponding generating function $F(s)$. Here f_n is the probability, given that there is a renewal at time t, than the next renewal occurs after another n time steps, i.e. at time $t + n$.

Using a similar technique to the one by which we derived Theorem 4 in Section 2.3, we can obtain the following extension to the delayed case.

Theorem 7. *The generating functions $V(s)$, $B(s)$ and $F(s)$ are related by*

$$V(s) = \frac{B(s)}{1 - F(s)} = B(s)U(s).$$

Proof. We note that for $n \geq 0$

$$v_n = P(E_n) = P(D + T_1 + T_2 + \ldots + T_k = n \text{ for some } k \geq 0).$$

(If $n = 0$, the only possibility is that $D = 0$: there is a renewal at time zero if and only if the delay is zero.)

As in the proof of Theorem 4, the event $\{D + T_1 + T_2 + \ldots + T_k = n\}$ can only happen for at most one value of k, so

$$E_n = \bigcup_{k=0}^{\infty} \{T_1 + T_2 + \ldots + T_k = n\}$$

and that this is a disjoint union.

It follows that

$$v_n = \sum_{k=0}^{\infty} P(D + T_1 + T_2 + \ldots + T_k = n)$$

where, again as in the proof of Theorem 4, only finitely many of the terms in the sum are positive, and the rest are zero.

Because adding together independent random variables corresponds to multiplying their generating functions (by Lemma 3), the random variable $D+T_1+T_2+\ldots+T_k$ has the generating function $B(s)(F(s))^k$ for each $k = 0, 1, 2, \ldots$. If we multiply both sides of the expression for v_n by s^n, and then sum these equations over all n we get

$$\begin{aligned} V(s) &= \sum_{k=0}^{\infty}\left(\sum_{n=0}^{\infty}P(D+T_1+T_2+\ldots+T_k=n)s^n\right) \\ &= \sum_{k=0}^{\infty}B(s)(F(s))^k \\ &= B(s)\sum_{k=0}^{\infty}(F(s))^k. \end{aligned}$$

Summing the geometric series as in the proof of Theorem 4, we get the result:

$$V(s)=\frac{B(s)}{1-F(s)} \text{ for } 0 \le s < 1.$$

(That $U(s)=\frac{1}{1-F(s)}$ can be seen by considering a non-delayed renewal process with the same renewal time distribution and applying Theorem 4.) □

Example 14. *Bernoulli trials with blocking revisited*

2.7 Recurrent patterns in coin-tossing

If a coin is tossed repeatedly, then the outcome of the whole experiment may be written as a string such as

$$HHTTHTTTHHHTHTHT\ldots$$

If we focus attention on a particular short finite sequence such as HHT, then every so often it will appear in this string; in the above example it appears twice:

$$\underline{HHT}THTTTH\underline{HHT}HTHT\ldots$$

Assuming as usual that different tosses are independent with constant probabilities of H and T, as soon as the sequence HHT has appeared, the whole process effectively starts again from scratch, independently of what happened before; hence the index numbers of the tosses on which the sequence is completed form a renewal process. In the above example, renewals occur at the tosses numbered 3 and 12.

However, if we change our sequence to say THT, we have an example of one which can overlap with itself; in the above string we have

$$HHT\underline{THT}TTHHH\underline{THTHT}\ldots$$

where the tosses numbered from 12 to 16 inclusive yield two overlapping occurrences of the sequence THT. We can overcome this complication by noting that if a THT has just occurred, we know that we have just had a T, and so a further occurrence of THT can be achieved either by immediately tossing another HT, or by not doing so and then later tossing another complete THT, disjoint from the previous one. Therefore the numbers of tosses between occurrences will be independent and identically distributed, but the number of tosses until the first occurrence will have a different distribution, because at the start of the sequence we do not already have a T, so to achieve a THT we have to toss a complete THT. So occurrences form a delayed renewal process.

Example 15. *Expected time until sequence completed*