Lecture 1

Introduction: Poisson processes, generalisations and applications

Reading: Part A Probability; Grimmett-Stirzaker 6.1, 6.8 up to (10)
Further reading: Ross 4.1, 5.3; Norris Introduction, 1.1, 2.4

This course is, in the first place, a course for 3rd year undergraduates who did Part A Probability in their 2nd year. Other students such as MSc students are welcome, but should note the prerequisites of the course. These are essentially an introductory course in probability *not* based on measure theory. It will be an advantage if this included the central aspects of discrete-time Markov chains. This will be relevant by the time we get to Lecture 5 in week 3.

The aim of Lecture 1 is to give a brief overview of the course. To do this at an appropriate level, we begin with a review of Poisson processes, which were treated at the end of the Part A course. The material most relevant to us is again included here, and some more is on the first assignment sheet.

This is a mathematics course. The name "Applied probability" suggests that we apply probability. However, there is more to it. This course is about "probability" and "applications" or application-driven probability theory. In particular, it is not just Part A Probability that we apply, but also further probability building on Part A. Effectively, we will be spending a fair share of our time developing theory so that we can analyse certain examples and applications.

For the rest of the course, let $\mathbb{N} = \{0, 1, 2, \dots\}$ denote the natural numbers including zero. Apart from very few exceptions, all stochastic processes that we consider in this course will have state space $\mathbb{N}$ (or a subset of $\mathbb{N}$). Specifically, we have in mind that we are counting, and studying the evolution of, numbers of people in a population, affected by a disease, of a certain genetic type, in a queue, etc. or just balls in an urn, bacteria in a dish, numbers of claim-free years for a motor insurance, the wealth of a gambler, numbers of defective items in a production line.

However, most results in the theory of Markov chains will be treated for any *countable*, i.e. finite or countably infinite, state space $\mathbb{S}$. This does not pose any complications as compared with $\mathbb{N}$, since we can always enumerate all states in $\mathbb{S}$ and hence give them labels in $\mathbb{N}$. Important examples are $\mathbb{Z}$, $\mathbb{N}^2$ and finite sets such as $\{\text{bachelor}, \text{married}, \text{divorced}, \text{widowed}\}$ or a set of colours, car makes, universities, shops, or indeed sets like $\{0, 1\}^n$. For uncountable state spaces, however, several technicalities arise that are beyond the scope of this course, at least in any generality – we will naturally come across a few examples of Markov processes in $\mathbb{R}$ towards the end of the course.

1.1 Poisson processes

There are many ways to define Poisson processes. We will use the following definition. We write $Z \sim \text{Exp}(\lambda)$ to say "Z is an exponentially distributed random variable with probability density function $\lambda e^{-\lambda t}$, $t \geq 0$", for some $\lambda > 0$.

Definition 1 Let $Z_n \sim \text{Exp}(\lambda)$, $n \geq 0$, independent, for some $\lambda > 0$. Let $T_n = Z_0 + \cdots + Z_{n-1}$, $n \geq 1$. Then the process $X = (X_t, t \geq 0)$ defined by

$$X_t = \#\{n \geq 1 : T_n \leq t\}, \qquad t \geq 0,$$

is called *Poisson process with rate* λ, abbreviated $\text{PP}(\lambda)$.

Note that $(X_t, t \geq 0)$ is not just a family of (dependent!) random variables but indeed $t \mapsto X_t$ is a random right-continuous function. This point of view is very useful since it is the formal justification for pictures of "typical realisations" of X.

Think of T_n as arrival times of customers (arranged in increasing order). Then X_t is counting the numbers of arrivals up to time t for all $t \geq 0$ and we study the evolution of this counting process. Instead of customers, one might be counting particles detected by a Geiger counter or cars driving through St. Giles, etc. Something more on the link and the important distinction between real observations (cars in St. Giles) and mathematical models (Poisson process) will be included in Lecture 2. For the moment we have a mathematical model, well specified in the language of probability theory. Starting from a simple sequence of independent random variables Z_n, $n \geq 0$, we have defined a more complex object $(X_t, t \geq 0)$, that we call the Poisson process.

Let us collect some properties that, apart from some technical details (to do with handling uncountably many random variables), can serve as an alternative definition of the Poisson process.

Remark 2 A process $X \sim \mathrm{PP}(\lambda)$ has the following properties:

(i) $X_t \sim \mathrm{Poi}(\lambda t)$ for all $t \geq 0$, where $\mathrm{Poi}(\lambda t)$ refers to the Poisson distribution with mean λt.

(ii) X has independent increments, i.e. for all $t_0 \leq \cdots \leq t_n$, the random variables $X_{t_j} - X_{t_{j-1}}$, $1 \leq j \leq n$, are independent.

(iii) X has stationary increments, i.e. $X_{t+s} - X_t \sim X_s$ for all $t \geq 0$, $s \geq 0$, where $\sim$ means "has the same distribution as".

To justify (i), calculate

$$\begin{aligned}
\mathbb{E}\left(q^{X_t}\right) &= \sum_{n=0}^{\infty} q^n \mathbb{P}(X_t = n) \\
&= \sum_{n=0}^{\infty} q^n \mathbb{P}(T_n \leq t, T_{n+1} > t) \\
&= \sum_{n=0}^{\infty} q^n \left(\mathbb{P}(T_n \leq t) - \mathbb{P}(T_{n+1} \leq t)\right) \\
&= 1 - \sum_{j=1}^{\infty} q^{j-1}(1-q)\mathbb{P}(T_j \leq t) \\
&= 1 - \int_0^t \sum_{j=1}^{\infty} q^{j-1}(1-q)\frac{\lambda^j}{(j-1)!} z^{j-1} e^{-\lambda z} dz \\
&= 1 - \int_0^t (1-q)\lambda e^{-\lambda z + \lambda q z} dz = e^{-\lambda t(1-q)},
\end{aligned}$$

where we used the well-known result that T_n as a sum of independent $\mathrm{Exp}(\lambda)$-variables has a $\mathrm{Gamma}(n, \lambda)$ distribution. This is the probability generating function of the Poisson distribution with parameter λt. We conclude by the Uniqueness Theorem for probability generating functions. Note that we interchanged summation and integration. This may be justified by an expansion of $e^{-\lambda z}$ into a power series and using uniform convergence of power series twice. We will see another justification in Lecture 3.

(ii)-(iii) can be derived from the following Proposition 3, see also Lecture 4.

1.2 The Markov property

Let $\mathbb{S}$ be a countable state space, typically $\mathbb{S} = \mathbb{N}$. Let $\Pi = (\pi_{rs})_{r,s\in\mathbb{S}}$ be a Markov transition matrix on $\mathbb{S}$. An $\mathbb{S}$-valued stochastic process $M = (M_n, n \geq 0)$ is called a *discrete time Markov chain with transition matrix* Π *starting from* $i_0 \in \mathbb{S}$ if for all $n \geq 1$ and $i_1, \ldots, i_n \in \mathbb{S}$

$$\mathbb{P}(M_1 = i_1, \ldots, M_n = i_n) = \prod_{j=1}^{n} \pi_{i_{j-1},i_j}.$$

It is often convenient to capture the initial state by writing $\mathbb{P}_{i_0}$ instead of $\mathbb{P}$. We then say that M is starting from i_0 *under* $\mathbb{P}_{i_0}$. We will use notation such as M, $(M_n, n \geq 0)$ and $(M_n)_{n\geq 0}$ interchangeably. Markov chains have the Markov property, which can be stated in several useful ways:

- For all paths $i_0, \ldots, i_{n+1} \in \mathbb{S}$ of positive probability $\mathbb{P}(M_0 = i_0, \ldots, M_n = i_n) > 0$, we have

$$\mathbb{P}(M_{n+1} = i_{n+1} | M_0 = i_0, \ldots, M_n = i_n) = \mathbb{P}(M_{n+1} = i_{n+1} | M_n = i_n) = \pi_{i_n,i_{n+1}}.$$

- For all $k \in \mathbb{S}$ and events $\{(M_j)_{0\leq j\leq n} \in A\}$ and $\{(M_{n+m})_{m\geq 0} \in B\}$, we have: if $\mathbb{P}(M_n = k, (M_j)_{0\leq j\leq n} \in A) > 0$, then

$$\mathbb{P}((M_{n+m})_{m\geq 0} \in B | M_n = k, (M_j)_{0\leq j\leq n} \in A) = \mathbb{P}((M_{n+m})_{m\geq 0} \in B | M_n = k) = \mathbb{P}_k(M \in B).$$

- The processes $(M_j)_{0\leq j\leq n}$ and $(M_{n+m})_{m\geq 0}$ are conditionally independent given $M_n = k$, for all $k \in \mathbb{S}$. Furthermore, given $M_n = k$, the process $(M_{n+m})_{m\geq 0}$ is a Markov chain with transition matrix Π starting from k.

Informally: no matter how we got to a state, the future behaviour of the chain is as if we were starting a new chain from that state. This is one reason why it is vital to study Markov chains not starting from one initial state but from any state in the state space $\mathbb{S}$.

Similarly, we will here study Poisson processes X starting from any initial state $X_0 = k \in \mathbb{N}$ (*under* $\mathbb{P}_k$), by which we just mean that we consider $X_t = k + \#\{n \geq 1 : T_n \leq t\}$, $t \geq 0$, where $T_n = Z_0 + \cdots + Z_{n-1}$ and Z_n, $n \geq 0$, are as in Definition 1.

Proposition 3 (Markov property) *Let $X \sim \mathrm{PP}(\lambda)$, i.e. X is a Poisson process starting from $X_0 = 0$, and consider a fixed time $t \geq 0$. Then the following hold.*

(i) *For all $s \geq 0$, $0 \leq r_1 < \cdots < r_n \leq t$ and $0 \leq i_1 \leq \cdots \leq i_n \leq k \leq \ell$,*

$$\mathbb{P}(X_{t+s} = \ell | X_t = k, X_{r_1} = i_1, \ldots, X_{r_n} = i_n) = \mathbb{P}(X_{t+s} = \ell | X_t = k) = \mathbb{P}_k(X_s = \ell).$$

More generally, for all $k \in \mathbb{N}$ and events $\{(X_r)_{r\leq t} \in A\}$ and $\{(X_{t+s})_{s\geq 0} \in B\}$, we have: if $\mathbb{P}(X_t = k, (X_r)_{r\leq t} \in A) > 0$, then

$$\mathbb{P}((X_{t+s})_{s\geq 0} \in B | X_t = k, (X_r)_{r\leq t} \in A) = \mathbb{P}((X_{t+s})_{s\geq 0} \in B | X_t = k) = \mathbb{P}_k((X_s)_{s\geq 0} \in B).$$

(ii) *The processes $(X_r)_{r\leq t}$ and $(X_{t+s})_{s\geq 0}$ are conditionally independent given $X_t = k$, for all $k \in \mathbb{N}$. Furthermore, given $X_t = k$, the process $(X_{t+s})_{s\geq 0}$ is a Poisson process with rate λ starting from k.*

(iii) *$(X_{t+s} - X_t)_{s\geq 0}$ is a Poisson process with rate λ starting from 0, (unconditionally) independent of $(X_r)_{r\leq t}$.*

We will prove a more general Proposition 18 in Lecture 4. Also, in Lecture 2, we will revise and push further the notion of conditioning. For this lecture we content ourselves with the formulation of the Markov property and proceed to the overview of the course.

Markov models (stochastic processes that have the Markov property and that model real-life evolutions, natural or man-made) are useful in a wide range of applications, e.g. price processes in Mathematical Finance, evolution of genetic material in Mathematical Biology, evolutions of particles in space in Mathematical Physics. The Markov property is a property that makes the model somewhat simple (not easy, but it could be much less tractable). We will develop tools that support this statement.

1.3 Brief summary of the course

Two generalisations of the Poisson process and several applications make up this course.

- The Markov property of Proposition 3(i)-(ii) can be used as a starting point to a bigger class of processes, so-called *continuous-time Markov chains*. They are analogues of discrete-time Markov chains, and they are often better adapted to applications. On the other hand, new aspects arise that did not arise in discrete time, and connections between the two will be studied. Roughly, the first half of this course is concerned with continuous-time Markov chains. Our main reference book will be Norris's book on Markov Chains.

- The Poisson process is the prototype of a counting process. For the Poisson process, "everything" can be calculated explicitly. In practice, though, this is often only helpful as a first approximation. E.g. in insurance applications, the Poisson process is used as a model to count claim arrivals. However, there is empirical evidence that inter-arrival times are neither exponentially distributed nor independent nor identically distributed. The second approximation is to relax exponentiality of inter-arrival times but to keep their independence and identical distribution. These counting processes are called *renewal processes*. Since exact calculations are often impossible or not helpful, the most important results of renewal theory are limiting results. Our main reference will be Chapter 10 of Grimmett and Stirzaker's book on Probability and Random Processes.

- Many applications that we discuss are in queueing theory. The easiest, so-called $M/M/1$ queue consists of a server and customers arriving according to a Poisson process. Independently of the arrival times, each customer has an exponential service time for which he will occupy the server, when it is his turn. When the server is busy, customers queue until being served. Everything has been designed so that the queue length is a continuous-time Markov chain, and various quantities can be studied or calculated (equilibrium distribution, lengths of idle periods, waiting time distributions etc.). More complicated queues arise if the Poisson process is replaced by a renewal process or the exponential service time distribution by any other distribution. There are also systems with $k = 2, 3, \ldots, \infty$ servers. Abstract queueing systems can be applied in telecommunications, computing networks, etc.

- Some other applications include insurance ruin models and the propagation of diseases.

Lecture 2

Conditioning and stochastic modelling

Reading: Grimmett-Stirzaker 3.7, 4.6
Further reading: Grimmett-Stirzaker 4.7; CT4 Unit 1

This lecture consolidates the ideas of conditioning and modelling. Along the way, we explain the full meaning of statements such as the Markov properties of Lecture 1.

2.1 Modelling of events

As in the Prelims and Part A courses, random variables are defined as functions on a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the *sample space*, $\mathcal{F}$ is a collection of subsets of Ω called *events*, the *probability measure* $\mathbb{P}$ assigns a probability in $[0,1]$ to each event. A probability space satisfies

- $\Omega \in \mathcal{F}$;
- $A \in \mathcal{F} \Rightarrow A^c = \Omega \setminus A \in \mathcal{F}$;
- $A_n \in \mathcal{F}, n \geq 1, \Rightarrow \bigcup_{n\geq 1} A_n \in \mathcal{F}$.
- $\mathbb{P}(\Omega) = 1$;
- $A_n \in \mathcal{F}, n \geq 1$, disjoint $\Rightarrow \mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$.

Random variables $X\colon \Omega \to \mathbb{X}$, where $\mathbb{X}$ is typically either $\mathbb{R}$ or $\mathbb{S}$, but can also be a space of functions such as $\mathbb{X} = \{f\colon [0,\infty) \to \mathbb{S} \text{ right-continuous}\}$, are such that

- $\{X \in B\} := X^{-1}(B) = \{\omega \in \Omega\colon X(\omega) \in B\} \in \mathcal{F}$ for all (measurable) $B \subseteq \mathbb{X}$.

This course is not based on measure theory, and in fact we will only occasionally have to refer back to these properties for clarity of argument.

Modelling means specifying a mathematical model for a real-world phenomenon. *Stochastic modelling* include some randomness, i.e. some real-world events are assigned probabilities or some real-world observables are assigned distributions. At first, real-world events can be named e.g. $A_1 =$"the die shows an even number" and $A_2 =$"the first customer arrives before 10am". A stochastic model identifies such an event as a subset of a sample space Ω and assigns probabilities. We seem to be able to write down some probabilities directly without much sophistication ($\mathbb{P}(A_1) = 0.5$? still making implicit assumptions about the fairness of the die and the conduct of the experiment). Others require less obvious specification of a stochastic model ($\mathbb{P}(A_2) =$?).

Hardly any real situations involve genuine randomness. It is rather our incomplete perception/information that makes us think there was randomness. Nevertheless, assuming a specific random model to inform our decision-making can be very helpful and lead to decisions that are sensible/good/beneficial in some sense.

Mathematical models always make assumptions and reflect reality only partially. The following situation is quite common: the better a model represents reality, the more complicated it is to analyse. There is a trade-off here. In any case, we must base all our calculations on the model specification, the model assumptions. Translating reality into models is at least partly a non-mathematical task. Analysing a model is purely mathematical.

Models have to be consistent, i.e. they must not contain contradictions. This statement may seem obvious, but the point is that not all contradictions are immediately apparent. There are models that have undesirable features that cannot be easily removed, least by postulating the contrary. E.g., you may wish to specify a model for customer arrival where arrival counts

over disjoint time intervals are independent, arrival counts over time intervals of equal lengths have the same distribution (cf. Remark 2 (ii)-(iii)), and times between two arrivals have a non-exponential distribution. Well, such a model does not exist (we won't prove this statement now, it's a bit too hard at this stage). On the other hand, within a consistent model, all properties that were not specified in the model assumptions have to be derived from these. Otherwise it must be assumed that the model may not have the property.

Suppose we are told that a shop opens at 9.30am, and on average, there are 10 customers per hour. One model could be to assume that a customer arrives exactly every six minutes. Another model could be to assume customers arrive according to a Poisson process at rate $\lambda = 10$ (time unit=1 hour). Whichever model we use, we can "calculate" $\mathbb{P}(A_2)$, and it is not the same in the two models, so we should reflect this in our notation. Since A_2 does not really change from one model to the other, it had better be $\mathbb{P}$ that changes, and we may wish to write $\widetilde{\mathbb{P}}$ for the second model. The probability measure $\mathbb{P}$ should be thought of as defining the randomness. Similarly, we can express dependence on a parameter by $\mathbb{P}^{(\lambda)}$, dependence on an initial value by $\mathbb{P}_k$. Informally, for a Poisson process model, we set $\mathbb{P}_k(A) := \mathbb{P}(A|X_0 = k)$ for all events A (formally, this should make us wonder whether $\mathbb{P}(X_0 = k) > 0$, and in fact, we first define $\mathbb{P}_k$ and could then write $\mathbb{P}(A|X_0 = k) := \mathbb{P}_k(A)$ as a long-hand notation).

Aside: Technically, we cannot in general call all subsets of Ω events if Ω is uncountable, but we will not worry about this, since it is hard to find examples of *non-measurable sets*. ω should be thought of as a scenario, a realisation of all the randomness, which we typically express in terms of random variables $X(\omega)$. What matters are (joint!) distributions of random variables, not usually the precise form of $(\Omega, \mathcal{F}, \mathbb{P})$. It is important, though, that $(\Omega, \mathcal{F}, \mathbb{P})$ exists for all our purposes to make sure that the random objects we study exist. We will assume that all our random variables can be defined as (measurable) functions on some $(\Omega, \mathcal{F}, \mathbb{P})$. This existence can be proved for all our purposes, using measure theory.

In fact, when we express complicated families of random variables such as a Poisson process $(X_t)_{t\geq 0}$ in terms of a countable family $(Z_n)_{n\geq 1}$ of independent random variables, we do this for two reasons. The intuitive reason may be apparent: countable families of independent variables are conceptually easier than uncountable families of dependent variables. The formal reason is that a result in measure theory says that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we can define countable families of independent variables whereas any more general result for uncountable families or dependent variables requires additional assumptions or other caveats.

It is very useful to think about random variables Z_n as functions $Z_n(\omega)$, because it immediately makes sense to define a Poisson process $X_t(\omega)$ as in Definition 1, by defining new functions in terms of old functions. A certain class of probability problems can be solved by applying analytic *rules* to calculations involving functions of random variables (transformation formula for densities, expectation of a function of a random variable in terms of its density or probability mass function, etc.). Here we are dealing more explicitly with random variables and events themselves, operating on them directly.

This course is not based on measure theory, but you should be aware that some of the proofs are only mathematically complete if based on measure theory. Ideally, this only means that we apply a result from measure theory that is intuitive enough to believe without proof. In a few cases, however, the gap is more serious. We will identify technicalities, but without drawing attention away from the probabilistic arguments that we develop in this course and that are useful for applications.

B8.1 Martingales Through Measure Theory provides as pleasant an introduction to measure theory as can be given. That course nicely complements this course in providing the formal basis for probability theory in general and hence for this course in particular. However, it is by no means a co-requisite, and when we do refer to this course, it is likely to be to material that has not yet been covered there. Williams' Probability with Martingales is the recommended book reference.

2.2 Conditional probabilities, densities and expectations

Conditional probabilities were introduced in Prelims as

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)},$$

where we require $\mathbb{P}(A) > 0$.

Example 4 Let X be a Poisson process. Then

$$\mathbb{P}(X_t = k + j | X_s = k) = \frac{\mathbb{P}(X_t - X_s = j, X_s = k)}{\mathbb{P}(X_s = k)} = \mathbb{P}(X_t - X_s = j) = \mathbb{P}(X_{t-s} = j),$$

by the independence and stationarity of increments, Remark 2 (ii)-(iii).

Conditional densities were introduced in Part A as

$$f_{S|T}(s|t) = f_{S|T=t}(s) = \frac{f_{S,T}(s,t)}{f_T(t)}.$$

Example 5 Let X be a Poisson process. Then, for $t > s$,

$$f_{T_2|T_1=s}(t) = \frac{f_{T_1,T_2}(s,t)}{f_{T_1}(s)} = \frac{f_{Z_0,Z_1}(s,t-s)}{f_{Z_0}(s)} = \frac{f_{Z_0}(s) f_{Z_1}(t-s)}{f_{Z_0}(s)} = f_{Z_1}(t-s) = \lambda e^{-\lambda(t-s)},$$

by the transformation formula for bivariate densities to relate f_{T_1,T_2} to f_{Z_0,Z_1}, and independence of Z_0 and Z_1.

Conditioning has to do with available information. In the real world, we often observe a process with time. When we have set up a stochastic model, e.g. a Poisson process witha known parameter $\lambda > 0$. (If we don't know λ, we should estimate λ and update estimates as we observe the real-world process, but we do not worry about this in this course.) It is instructive to think of updating the stochastic process by its realisation in the real world as time evolves. If the first arrival takes a long time to happen, this gives us information about the second arrival time T_2, simply since $T_2 = T_1 + Z_1 > T_1$. When we eventually observe $T_1 = s$, the conditional density of T_2 given $T_1 = s$ takes into account this observation and captures the remaining stochastic properties of T_2. The result of the formal calculation to derive the conditional density is in agreement with the intuition that if $T_1 = s$, $T_2 = T_1 + Z_1$ ought to have the distribution of Z_1 shifted by s.

Example 6 Conditional probabilities and conditional densities are compatible in that

$$\mathbb{P}(S \in B | T = t) = \int_B f_{S|T=t}(s) ds = \lim_{\varepsilon \downarrow 0} \mathbb{P}(S \in B | t \le T \le t + \varepsilon),$$

provided only that the distribution of (S,T) is sufficiently smooth. To see this, when $f_{S,T}$ is sufficiently smooth, write for all intervals $B = (a,b)$

$$\mathbb{P}(S \in B | t \le T \le t + \varepsilon) = \frac{\mathbb{P}(S \in B, t \le T \le t + \varepsilon)}{\mathbb{P}(t \le T \le t + \varepsilon)} = \frac{\frac{1}{\varepsilon} \int_t^{t+\varepsilon} \int_B f_{S,T}(s,u) ds du}{\frac{1}{\varepsilon} \mathbb{P}(t \le T \le t + \varepsilon)}$$

and under the smoothness condition (by dominated convergence, Fubini-Tonelli etc.), this tends to

$$\frac{\int_B f_{S,T}(s,t) ds}{f_T(t)} = \int_B f_{S|T=t}(s) ds = \mathbb{P}(S \in B | T = t).$$

Similarly, we can also define

$$\mathbb{P}(X=k|T=t)=\lim_{\varepsilon\downarrow 0}\mathbb{P}(X=k|t\le T\le t+\varepsilon)\quad\text{and}\quad f_{T|X=k}(t)=\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\mathbb{P}(t\le T\le t+\varepsilon,X=k).$$

One can define conditional expectations in analogy with unconditional expections, e.g. in the latter case by

$$\mathbb{E}(X|T=t)=\sum_{j=0}^{\infty}j\mathbb{P}(X=j|T=t).$$

Proposition 7 (a) *If X and Y are (dependent) discrete random variables in $\mathbb{N}$, then*

$$\mathbb{E}(X)=\sum_{n=0}^{\infty}\mathbb{E}(X|Y=n)\mathbb{P}(Y=n).$$

(b) *If X and T are jointly continuous random variables in $(0,\infty)$ or*

(c) *if X is discrete and T is continuous, and if T has a right-continuous density, then*

$$\mathbb{E}(X)=\int_0^{\infty}\mathbb{E}(X|T=t)f_T(t)dt.$$

Proof: (c) We start at the right-hand side

$$\int_0^{\infty}\mathbb{E}(X|T=t)f_T(t)dt=\int_0^{\infty}\sum_{j=0}^{\infty}j\mathbb{P}(X=j|T=t)f_T(t)dt$$

and calculate

$$\begin{aligned}\mathbb{P}(X=j|T=t) &= \lim_{\varepsilon\downarrow 0}\frac{\mathbb{P}(X=j,t\le T\le t+\varepsilon)}{\mathbb{P}(t\le T\le t+\varepsilon)}\\ &= \lim_{\varepsilon\downarrow 0}\frac{\frac{1}{\varepsilon}\mathbb{P}(t\le T\le t+\varepsilon|X=j)\mathbb{P}(X=j)}{\frac{1}{\varepsilon}\mathbb{P}(t\le T\le t+\varepsilon)}\\ &= \frac{f_{T|X=j}(t)\mathbb{P}(X=j)}{f_T(t)}\end{aligned}$$

so that we get on the right-hand side

$$\int_0^{\infty}\sum_{j=0}^{\infty}j\mathbb{P}(X=j|T=t)f_T(t)dt=\sum_{j=0}^{\infty}j\mathbb{P}(X=j)\int_0^{\infty}f_{T|X=j}(t)dt=\mathbb{E}(X)$$

after interchanging summation and integration. This is justified by Tonnelli's theorem that we state in Lecture 3.

(b) is similar to (c).

(a) is more elementary and left to the reader. □

Statement and argument hold for left-continuous densities and approximations from the left, as well. For continuous densities, one can also approximate $\{T=t\}$ by $\{t-\varepsilon\le T\le t+\varepsilon\}$ (for $\varepsilon<t$, and normalisation by 2ε, as adequate).

Recall that we formulated the Markov property of the Poisson process as

$$\mathbb{P}((X_{t+s})_{s\ge 0}\in B|X_t=k,(X_r)_{r\le t}\in A)=\mathbb{P}_k((X_{t+s})_{s\ge 0}\in B)$$

for all events $\{(X_r)_{r\le t}\in A\}$ such that $\mathbb{P}(X_t=k,(X_r)_{r\le t}\in A)>0$, and $\{(X_{t+u})_{u\ge 0}\in B\}$. For certain sets A with zero probability, this can still be established by approximation.

2.3 Independence and conditional independence

Recall that independence of two random variables is defined as follows. Two discrete random variables X and Y are independent if

$$\mathbb{P}(X = j, Y = k) = \mathbb{P}(X = j)\mathbb{P}(Y = k) \qquad \text{for all } j, k \in \mathbb{S}.$$

Two jointly continuous random variables S and T are independent if their joint density factorises, i.e. if

$$f_{S,T}(s,t) = f_S(s)f_T(t) \qquad \text{for all } s,t \in \mathbb{R}, \text{ where } f_S(s) = \int_{\mathbb{R}} f_{S,T}(s,t)dt.$$

Recall also (or check) that this is equivalent, in both cases, to

$$\mathbb{P}(S \le s, T \le t) = \mathbb{P}(S \le s)\mathbb{P}(T \le t) \qquad \text{for all } s,t \in \mathbb{R}.$$

In fact, it is also equivalent to

$$\mathbb{P}(S \in A, T \in B) = \mathbb{P}(S \in B)\mathbb{P}(T \in B) \qquad \text{for all (measurable) } A, B \subset \mathbb{R},$$

and we define more generally:

Definition 8 Let X and Y be two random variables with values in any, possibly different spaces $\mathbb{X}$ and $\mathbb{Y}$. Then we call X and Y independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad \text{for all (measurable) } A \subset \mathbb{X} \text{ and } B \subset \mathbb{Y}.$$

We call X and Y conditionally independent given a third random variable Z if for all $z \in \mathbb{S}$ (if Z has values in $\mathbb{S}$) or $z \in [0,\infty)$ (if Z has values in $[0,\infty)$),

$$\mathbb{P}(X \in A, Y \in B | Z = z) = \mathbb{P}(X \in A | Z = z)\mathbb{P}(Y \in B | Z = z).$$

Remark and Fact 9 [1] Conditional independence is in many ways like ordinary (unconditional) independence. E.g., if X is discrete, it suffices to consider $A = \{x\}$, $x \in \mathbb{X}$. If X is real-valued, it suffices to consider $A = (-\infty, x]$, $x \in \mathbb{R}$. If X is bivariate, it suffices to consider all A of the form $A = A_1 \times A_2$.

If $X = (X_r)_{r \le t}$, it suffices to consider $A = \{X_{r_1} = x_1, \ldots, X_{r_n} = x_n\}$ for all $0 \le r_1 < \cdots < r_n \le t$, $x_1, \ldots, x_n \in \mathbb{N}$, $n \ge 1$. This is how Proposition 3(ii) can be interpreted, applied and proved.

We conclude by a fact that may seem obvious, but does not follow immediately from the definitions. Also the approximation argument only gives some special cases.

Fact 10 *Let X be any random variable, and T a real-valued random variable with right-continuous density. Then, for all (measurable) $f : \mathbb{X} \times [0,\infty) \to [0,\infty)$, we have*

$$\mathbb{E}(f(X,T)|T = t) = \mathbb{E}(f(X,t)|T = t).$$

Furthermore, if X and T are independent and $g : \mathbb{X} \to [0,\infty)$ (measurable) we have

$$\mathbb{E}(g(X)|T = t) = \mathbb{E}(g(X)).$$

If X takes values in $[0,\infty)$ also, example for f are e.g. $f(x,t) = 1_{\{x+t>s\}}$, where $1_{\{x+t>s\}} := 1$ if $x + t > s$ and $1_{\{x+t>s\}} := 0$ otherwise; or $f(x,t) = e^{\lambda(x+t)}$ in which case the statements are

$$\mathbb{P}(X + T > s | T = t) = \mathbb{P}(X + t > s | T = t) \text{ and } \mathbb{E}(e^{\lambda(X+T)}|T = t) = e^{\lambda t}\mathbb{E}(e^{\lambda X}|T = t),$$

and the condition $\{T = t\}$ can be removed on the right-hand sides if X and T are independent. This can be shown by the approximation argument.

The analogue of Fact 10 for discrete T is elementary.

[1] Facts are theorems that we cannot fully prove in this course. Note also that there is a grey zone between theorems/propositions and facts, since partial proofs of facts or full proofs of theorems/propositions sometimes appear on assignment sheets, in the main or optional parts.

2.4 Method: One-step analysis, conditioning on T_1

The main conditioning method in this course is to condition on the first event. In the case of a discrete-time Markov chain this is the value after the first step. In the case of a Poisson process (or simple birth process or renewal process, as studied later), this is the time of the first arrival. In the case of a continuous-time Markov chain, it will be a combination of the two.

Example 11 Let $X \sim \mathrm{PP}(\lambda)$ and $m(u) = \mathbb{E}(X_u)$, $u \geq 0$. Then by Proposition 7(c),

$$m(u) = \mathbb{E}(X_u) = \int_0^\infty \mathbb{E}(X_u | T_1 = t) f_{T_1}(t) dt = \int_0^u \mathbb{E}(1 + \widetilde{X}_{u-T_1} | T_1 = t) \lambda e^{-\lambda t} dt,$$

where $\widetilde{X}_s = X_{T_1+s} - 1$ is a $\mathrm{PP}(\lambda)$ independent of $T_1 = Z_0$, since $\widetilde{X}_s = \#\{n \geq 1 : \widetilde{T}_n \leq s\}$, where $\widetilde{Z}_n = Z_{n+1} \sim \mathrm{Exp}(\lambda)$, $n \geq 0$, are independent, and independent of Z_0. By Fact 10, this yields

$$m(u) = \int_0^u (1 + \mathbb{E}(X_{u-t})) \lambda e^{-\lambda t} dt = 1 - e^{-\lambda u} + \int_0^u m(r) \lambda e^{-\lambda(u-r)} dr.$$

If we multiply this by $e^{\lambda u}$, differentiate and cancel $e^{\lambda u}$ again, we find $m'(u) = \lambda$. Since also $m(0) = \mathbb{E}(X_0) = 0$, we obtain $m(t) = \lambda t$ for all $t \geq 0$, using a quite different argument from Remark 2. The real power of this argument will be revealed when applied for processes other than the Poisson process, for which many stronger tools yield stronger results.