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 second year. Other students from various M.Sc.'s are welcome as long as they are aware of 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, 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 that were treated at the very end of the Part A syllabus. The parts most relevant to us in today's lecture are again included here, and some more material is on the first assignment sheet.

This is a mathematics course. "Applied probability" means that we apply probability, but not so much Part A Probability but further probability building on Part A and not covered there, so effectively, we will be spending a lot of our time developing theory as required for certain examples and applications.

For the rest of the course, let $\mathbb{N} = \{0, 1, 2, \ldots\}$ 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 thereof). However, most results in the theory of Markov chains will be treated for any at most *countable* state space $\mathbb{S}$, which does not pose any complications as compared with $\mathbb{N}$, since one can always enumerate all states in $\mathbb{S}$ and hence give them labels in $\mathbb{N}$. For uncountable state spaces, however, several technicalities arise that are beyond the scope of this course, at least in any reasonable 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 choose an elementary definition that happens to be the most illustrative since it allows to draw pictures straight away.

Definition 1 Let $(Z_n)_{n\geq 1}$ be a sequence of independent exponential random variables $Z_n \sim Exp(\lambda)$ for a parameter (inverse mean) $\lambda \in (0,\infty)$, $T_0 = 0$, $T_n = \sum_{k=1}^n Z_k$, $n \geq 1$. Then the process $X = (X_t)_{t\geq 0}$ defined by

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

is called *Poisson process with rate* λ.

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 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 $(Z_n)_{n\geq 1}$ we have defined a more complex object $(X_t)_{t\geq 0}$, that we call Poisson process.

Let us collect some properties that, apart from some technical details, can serve as an alternative definition of the Poisson process.

Remark 2 A Poisson process X with rate λ has the following properties

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

ii) X has independent increments, i.e. for all $t_0 \leq \ldots \leq t_n$, $X_{t_j} - X_{t_{j-1}}$, $j = 1, \ldots, n$, are independent.

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

To justify i), just calculate

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

and it is easily checked that 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.

ii)-iii) can be derived from the following Proposition 3.

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 $\mathbb{S}$. For every $s_0 \in \mathbb{S}$ this specifies the distribution of a Markov chain $(M_n)_{n\geq 0}$ starting from s_0 (under $\mathbb{P}_{s_0}$, say), by

$$\mathbb{P}_{s_0}(M_1 = s_1, \ldots, M_n = s_n) = \mathbb{P}(M_1 = s_1, \ldots, M_n = s_n | M_0 = s_0) = \prod_{j=1}^{n} \pi_{s_{j-1},s_j}$$

We say that $(M_n)_{n\geq 0}$ is a Markov chain with transition matrix Π starting from s_0. There are several formulations of the Markov property:

- For all paths $s_0, \ldots, s_{n+1} \in \mathbb{S}$ of positive probability, we have

$$\mathbb{P}(M_{n+1} = s_{n+1} | M_0 = s_0, \ldots, M_n = s_n) = \mathbb{P}(M_{n+1} = s_{n+1} | M_n = s_n) = \pi_{s_n,s_{n+1}}$$

- For all $s \in \mathbb{S}$ and events $\{(M_0, \ldots, M_n) \in A\}$ and $\{(M_n, M_{n+1}, \ldots) \in B\}$, we have: if $\mathbb{P}(M_n = s, (M_j)_{0\leq j\leq n} \in A) > 0$, then

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

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

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.

In analogy, we will here study Poisson processes X starting from initial states $X_0 = k \in \mathbb{N}$ (under $\mathbb{P}_k$), by which we just mean that we consider $X_t = k + \widetilde{X}_t$, $t \geq 0$, where $\widetilde{X}$ is a Poisson process starting from 0 as in the above definition.

Proposition 3 (Markov property) *Let X be a Poisson process with rate λ and $t \geq 0$ a fixed time. Then the following hold.*

i) 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) Given $X_t = k$, $(X_r)_{r\le t}$ and $(X_{t+s})_{s\ge 0}$ are independent, and the conditional distribution of $(X_{t+s})_{s\ge 0}$ is the same as the distribution of X given $X_0 = k$.

iii) $(X_{t+s} - X_t)_{s\ge 0}$ is also a Poisson process with rate λ, independent of $(X_r)_{r\le t}$.

We will prove a more general Proposition 17 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 (models that have the Markov property) 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 3ii) 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 for the arrival of claims. 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 keep their independence and identical distribution. This class of counting processes is 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 book will be 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. If 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 times by any other distribution. There are also systems with $k = 2, 3, \ldots, \infty$ servers. The abstract queueing systems can be more concretely applied in telecommunication, computing networks, etc.

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

Lecture 2

Conditioning and stochastic modelling

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

This lecture consolidates the ideas of conditioning and modelling preparing a more varied range of applications and a less mechanical use of probability than what was the focus of Part A. Along the way, we explain the meaning of statements such as the Markov properties in Lecture 1.

2.1 Modelling of events

Much of probability theory is about events and probabilities of events. Informally, this is an easy concept. Events like $A_1 =$"the die shows an even number" and $A_2 =$"the first customer arrives before 10am" make perfect sense in real situations. When it comes to assigning probabilities, things are less clear. We seem to be able to write down some ($\mathbb{P}(A_1) = 0.5$?) probabilities directly without much sophistication (still making implicit assumptions about the fairness of the die and the conduct of the experiment). Others ($\mathbb{P}(A_2)$) definitely require a mathematical model.

Hardly any real situations involve genuine randomness. It is rather our incomplete perception/information that makes us think there was randomness. In fact, assuming a specific random model in 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. Quite commonly, we have the following phenomenon: 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 a non-mathematical task. Analysing a model is purely mathematical. Models have to be consistent, i.e. not contain contradictions. This statement may seem superfluous, but 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 nonexponential distribution. Well, such a model does not exist (we won't prove this statements, it's hard). 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 say, that a customer arrives exactly every six minutes. Another model could be to say, customers arrive according to a Poisson process at rate $\lambda = 10$ (time unit=1 hour). Whichever model you prefer, fact is, you can "calculate" $\mathbb{P}(A_2)$, and it is not the same in the two models, so we should reflect this in our notation. We don't want it to be A_2 that changes, so it must be $\mathbb{P}$, and we may wish to write $\tilde{\mathbb{P}}$ for the second model. $\mathbb{P}$ 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 you should wonder whether $\mathbb{P}(X_0 = k) > 0$).

Aside: Formally, there is a way to define random variables as functions $X_t : \Omega \to \mathbb{N}$, $Z_n : \Omega \to [0, \infty)$ etc. $\mathbb{P}_k$ can then be defined as a *measure* on Ω for all k, and this measure is compatible with our distributional assumptions which claim that

$$\text{the probability that } X_t = k + j \text{ is } \frac{(\lambda t)^j}{j!} e^{-\lambda t}$$

in that

$$\mathbb{P}_k(X_t = j + k) = \mathbb{P}_k(\{\omega \in \Omega : X_t(\omega) = j + k\}) = \frac{(\lambda t)^j}{j!} e^{-\lambda t}.$$

In the mathematical sense, the set $A_{j+k} := \{X_t = j+k\} := \{\omega \in \Omega : X_t(\omega) = j+k\} \subset \Omega$ is called an *event*. Technically[1], we cannot call all subsets of Ω events if Ω is uncountable, but we will not worry about this, since it is very hard to find examples of *non-measurable sets*. ω should be thought of as a scenario, a realisation of all the randomness. Ω collects the possibilities and $\mathbb{P}$ tells us how likely each event is to occur. If we denote the set of all events by $\mathcal{A}$, then $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *stochastic basis*. Its form is usually irrelevant. It is important that it 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 Ω. 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 first should be apparent: countable families of independent variables are conceptually easier than uncountable families of dependent variables. The second is that a result in measure theory says that there exists a stochastic basis on which we can define countable families of independent variables whereas

[1]The remainder of this paragraph is in a smaller font. This means (now and whenever something is in small font), that it can be skipped on first reading, and the reader may or may not want to get back to it at a later stage.

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. When learning probability, it is usual to first apply analytic *rules* to calculate distributions of 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). 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. Every effort will be made to point out technicalities, but without drawing attention away from the probabilistic arguments that constitute this course and that are useful for applications.

B10a Martingales Through Measure Theory provides as pleasant an introduction to measure theory as can be given. That course nicely complements this course in providing a formal basis for probability theory in general and this course in particular.

2.2 Conditional probabilities, densities and expectations

Conditional probabilities were introduced in Part A (or even Mods) 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 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 $t > s$,

$$f_{T_2|T_1=s}(t) = \frac{f_{T_1,T_2}(s,t)}{f_{T_1}(s)} = \frac{f_{Z_1,Z_2}(s,t-s)}{f_{Z_1}(s)} = f_{Z_2}(t-s),$$

by the transformation formula for bivariate densities to relate f_{T_1,T_2} to f_{Z_1,Z_2}, and independence of Z_1 and Z_2.

Conditioning has to do with available information. Many models are stochastic only because the detailed deterministic structure is too complex. E.g. the counting process of natural catastrophies (hurricanes, earthquakes, volcanic eruptions, spring tide) etc. is genuinely deterministic. Since we cannot observe and model precisely weather, tectonic movements etc. it is much more fruitful to write down a stochastic model, e.g. a Poisson process, as a first approximation. We observe this process over time, and we can update the stochastic process by its realisation. Suppose we know the value of the intensity parameter $\lambda \in (0, \infty)$. (If we don't, any update will lead to a new estimate of λ, but we do not worry about this here). If the first arrival takes a long time to happen, this gives us information about the second arrival time T_2, simply since $T_2 = T_1 + Z_2 > T_1$. 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 $T_1 = s$, $T_2 = T_1 + Z_2$ ought to have the distribution of Z_2 shifted by s.

Example 6 Conditional probabilities and conditional densities are compatible in that

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

provided only that f_T is right-continuous. To see this, write the right hand side as

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

and under the continuity condition (and by dominated convergence), 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

$$\mathbb{P}(X = k | T = t) = \lim_{\varepsilon \downarrow 0} \mathbb{P}(X = k | t \leq T \leq t + \varepsilon)$$

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*

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

If b) 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

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

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

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

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

You have defined independence of two random variables 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

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

You may also know (or can easily check) that this is equivalent, in both cases, to

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

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 [2] Condional independence is in many ways like ordinary (unconditional) independence. E.g., if X is discrete, it suffices to check the condition for $A = \{x\}$, $x \in \mathbb{X}$. If Y is real-valued, it suffices to consider $B = (-\infty, y]$, $y \in \mathbb{R}$. If Y is bivariate, it suffices to consider all B of the form $B = B_1 \times B_2$.

If $\mathbb{X} = \{f : [0, t] \to \mathbb{N} \text{ right-continuous}\}$, it suffices to consider $A = \{f \in \mathbb{X} : f(r_1) = n_1, \ldots, f(r_m) = n_m\}$ for all $0 \leq r_1 < \ldots < r_m \leq t$ and $n_1, \ldots, n_m \in \mathbb{N}$. This is the basis for the meaning of Proposition 3ii).

We conclude by a fact that may seem obvious, but does not follow immediately from the definition. 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.

[2] Facts are results that are true, but that we cannot prove in this course. Note also that there is a grey zone between facts and propositions, since proofs or partial proofs sometimes appear on assignment sheets, in the main or optional parts.

Lecture 3

Birth processes and explosion

Reading: Grimmett-Stirzaker 6.8 (11),(18)-(20), Norris 2.3

In this lecture we introduce birth processes in analogy with our definition of Poisson processes. The common description (also for Markov chains) is always as follows: given the current state is m, what is the next part of the evolution, and (for the purpose of an inductive description) how does it depend on the past? (Answer to the last bit: conditionally independent given the current state, but this can here be expressed in terms of genuine independence).

3.1 Definition and an example

If we consider the Poisson process as a model for a growing population, it is not always sensible to assume that new members are born at the same rate regardless what the size of the population. You would rather expect this rate to increase with size (more births in larger populations), although some saturation effects may make sense as well.

Also, if the Poisson process is used as a counting process of alpha particle emissions of a decaying radioactive substance, it makes sense to assume that the rate is decreasing with the number of emissions, particularly if the half-life time is short.

Definition 11 A stochastic process $X = (X_t)_{t\geq 0}$ of the form

$$X_t = k + \#\left\{ n \geq 1 : \sum_{j=1}^{n} Z_j \leq t \right\}$$

is called a *simple birth process of rates* $(\lambda_n)_{n\geq 0}$ starting from $X_0 = k \in \mathbb{N}$, if the *inter-arrival times* Z_j, $j \geq 1$, are independent exponential random variables with parameters λ_{k+j-1}, $j \geq 1$. X is called a $(k, (\lambda_n)_{n\geq 0})$-birth process.

Note that the parameter λ_n is attached with height n. The so-called holding time at height n has an $Exp(\lambda_n)$ distribution.

"Simple" refers to the fact that no two births occur at the same time, which one would call "multiple" births. Multiple birth processes can be studied as well, and, given

certain additional assumptions, form examples of continuous-time Markov chains, like simple birth processes do, as we will see soon.

We will not use the second part of this definition yet. However, it is vital to point out that, in general, the distribution of a Markov process (e.g. a Markov chain) is determined by an intial distribution and the transition mechanism. Here, we chose to describe the transition mechanism by the exponential holding times Z_n, $n \geq 1$, whose distributions depend on the initial state $k \in \mathbb{N}$.

Example 12 Consider a population in which each individual gives birth after an exponential time of parameter λ, all independently and repeatedly. If n individuals are present, then the first birth will occur after an exponential time of parameter $n\lambda$. Then we have $n+1$ individuals and, by the lack of memory property (see Assignment 1), the process begins afresh. Thus the size of the population performs a birth process with rates $\lambda_n = n\lambda$. Let X_t denote the number of individuals at time t and suppose $X_0 = 1$. Write T for the time of the first birth. Then by Proposition 7c)

$$\mathbb{E}(X_t) = \int_0^\infty \lambda e^{-\lambda s}\mathbb{E}(X_t|T=s)ds = \int_0^t \lambda e^{-\lambda s}\mathbb{E}(X_t|T=s)ds + e^{-\lambda t}$$

since $X_t = X_0 = 1$ if $T > t$.

Put $\mu(t) = \mathbb{E}(X_t)$, then for $0 < s < t$, intuitively $\mathbb{E}(X_t|T=s) = 2\mu(t-s)$ since from time s, the two individuals perform independent birth processes of the same type and we are interested in their population sizes $t-s$ time units later. We will investigate a more formal argument later, when we have a Markov property at our disposal.

Now

$$\mu(t) = \int_0^t 2\lambda e^{-\lambda s}\mu(t-s)ds + e^{-\lambda t}$$

and setting $r = t - s$

$$e^{\lambda t}\mu(t) = 1 + 2\lambda \int_0^t e^{\lambda r}\mu(r)dr.$$

Differentiating we obtain

$$\mu'(t) = \lambda\mu(t)$$

so the mean population size grows exponentially, and $X_0 = 1$ implies

$$\mathbb{E}(X_t) = \mu(t) = e^{\lambda t}.$$

3.2 Tonelli's Theorem, monotone and dominated convergence

The following is a result from measure theory that we cannot prove or dwell on here.

Fact 13 (Tonelli) *You may interchange order of integration, countable summation and expectation whenever the integrand/summands/random variables are* nonnegative*, e.g.*

$$\mathbb{E}\left(\sum_{n\geq 0} X_n\right) = \sum_{n\geq 0}\mathbb{E}(X_n), \qquad \int_0^\infty \sum_{n\geq 0} f_n(x)dx = \sum_{n\geq 0}\int_0^\infty f_n(x)dx$$
$$\int_0^\infty \int_0^x f(x,y)dydx \;=\; \int_0^\infty \int_y^\infty f(x,y)dxdy.$$

There were already two applications of this, one each in Lectures 1 and 2. The focus was on other parts of the argument, but you may wish to consider the justification of Remark 2 and Proposition 7 again now.

Interchanging limits is more delicate, and there are monotone and dominated convergence for this purpose. In this course we will only interchange limits when this is justified by monotone or dominated convergence, but we do not have the time to work out the details. Here are indicative statements.

Fact 14 (Monotone convergence) *Integrals (expectations) of an increasing sequence of nonnegative functions (random variables Y_n) converge (in the sense that $\lim_{n\to\infty}\mathbb{E}(Y_n) = \mathbb{E}(\lim_{n\to\infty} Y_n)$.*

Fact 15 (Dominated convergence) *Integrals (expectations) of a pointwise convergent sequence of functions $f_n \to f$ (applied to a random variable) converge, if the sequence $|f_n| \leq g$ is dominated by an integrable function g (function which when applied to the random variable has finite expectation), i.e.*

$$\int g(x)dx < \infty \quad \Rightarrow \quad \lim_{n\to\infty}\int f_n(x)dx = \int \lim_{n\to\infty} f_n(x)dx$$
$$\mathbb{E}(g(X)) < \infty \quad \Rightarrow \quad \lim_{n\to\infty}\mathbb{E}(f_n(X)) = \mathbb{E}(\lim_{n\to\infty} f_n(X)).$$

We refer to Part B Martingales through measure theory for those of you who follow that course. In any case, our working hypothesis is that, in practice, we may interchange all limits that come up in this course.

3.3 Explosion

If the rates increase too quickly, it may happen that infinitely many individuals are born in finite time (as with deterministic inter-birth times). We call this phenomenon *explosion*. Formally, we can express the possibility of explosion by $\mathbb{P}(T_\infty < \infty) > 0$ where $T_\infty = \lim_{n\to\infty} T_n$. For simple birth processes, we have the following necessary and sufficient condition.

Proposition 16 *Let X be a birth process with rates $(\lambda_n)_{n\geq 0}$. Then*

$$\mathbb{P}(T_\infty < \infty) > 0 \quad \textit{if and only if} \quad \sum_{n=0}^{\infty} \frac{1}{\lambda_n} < \infty.$$

Furthermore, in this case $\mathbb{P}(T_\infty < \infty) = 1$.

Proof: Note that

$$\begin{aligned} \mathbb{E}(T_\infty) &= \mathbb{E}\left(\sum_{n=1}^{\infty} Z_n\right) \\ &= \sum_{n=1}^{\infty} \mathbb{E}(Z_n) \\ &= \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \end{aligned}$$

where Tonelli's Theorem allows us to interchange summation and expectation. Therefore, if the series is finite, then $\mathbb{E}(T_\infty) < \infty$ implies $\mathbb{P}(T_\infty < \infty) = 1$. The other implication does not hold, in general. However, (using monotone convergence), and also the independence of the Z_n, we can calculate

$$\begin{aligned} -\log \mathbb{E}\left(e^{-T_\infty}\right) &= -\sum_{n=0}^{\infty} \log \mathbb{E}\left(e^{-Z_n}\right) \\ &= \sum_{n=1}^{\infty} \log\left(1 + \frac{1}{\lambda_n}\right). \end{aligned}$$

Either this latter sum is greater than $\log(2)\sum_{n=n_0}^{\infty} \frac{1}{\lambda_n}$ if $\lambda_n \geq 1$ for $n \geq n_0$, by linear interpolation and concavity of the logarithm. Or otherwise a restriction to any subsequence $\lambda_{n_k} \leq 1$ shows that the sum is infinite as each of these summands contributes at least $\log(2)$.

Therefore, if the series diverges, then $\mathbb{E}(e^{-T_\infty}) = 0$, i.e. $\mathbb{P}(T_\infty = \infty) = 1$. □

Note that we have not explicitly specified what happens after T_∞ if $T_\infty < \infty$. With a population size model in mind, $X_t = \infty$ for all $t \geq T_\infty$ is a reasonable convention. Formally, this means that X is a process in $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$. This process is often called the *minimal process*, since it is "active" on a minimal time interval. We will show the Markov property for minimal processes. It can also be shown that there are other ways to specify X after explosion that preserve the Markov property. The next natural thing to do is to start afresh after explosion. Such a process is then called *non-minimal*.