3 Probability Theory and Simple Markov Processes

In this note we will review some elementary concepts from probability theory which will be used in the rest of the course. We will also use some of these ideas to define a Markov process and show how such processes may be used for performance modelling.

3.1 Probability Theory

Probability theory deals with the analysis of events which are unpredictable when viewed individually, but which exhibit some regularity and predictability when viewed in large numbers. We are familiar with such ideas applied to tossing a coin or rolling a die. Here the unpredictability of an individual outcome is due to the non-deterministic nature of the situation. When we are considering a computer system the unpredictability of an individual outcome is likely to arise because of our inability to measure or specify the exact conditions under which the event occurs. This distinction, however, is largely philosophical; probability theory is equally applicable to both these situations.

There are three axioms of probability theory:

1. There is a *sample space*, Ω, which encompasses all possible observations or outcomes.
2. There is a collection of subsets of Ω, denoted E, and termed *events*[1]; these subsets are usually identified as sample points which satisfy some condition.
3. There is a *probability mapping*, Pr, from E to $\mathbb{R}$. Pr must satisfy three simple conditions:
   (a) For any event A, $A \in E$, the mapping Pr is defined and satisfies $0 \leq \Pr(A) \leq 1$.
   (b) $\Pr(\Omega) = 1$.
   (c) If A and B are mutually exclusive, that is, they contain no sample points in common, then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

If we consider the example given in lecture note 1, of observing the department's dial-in lines the sample space is the complete set of observations. Example events might be that more than three lines are in use or that a particular user is making use of the facility. These events corresponds to subsets of the sample points.

Various properties of probabilities can be derived from the axioms and simple set theory. For example, the probability of the union of two events A and B which are **not** mutually exclusive is $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Similarly, the probability of the complement of event A, denoted by $\neg A$, is $\Pr(\neg A) = 1 - \Pr(A)$.

3.1.1 Conditional Probabilities

If we know that some event has occurred then this may affect the probability that other events have occurred. The *conditional probability* of A occurring, given that B has occurred, is

[1] Note that this is a different use of the term "event" from discrete *event* system.

$$\boxed{\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}}$$

Example: In a communication network, there are two routes A and B, between a particular pair of nodes which are chosen with equal probability. Message transmission time along route A is 1 second and along route B is 3 seconds. However, the probability of a transmission error is 0.2 on route A and 0.1 on route B. We wish to find the probability of a message being transmitted correctly in 1 second.

Let us denote the event of a successful message transmission by C. We wish to calculate the probability that a transmission takes route A and is successful, i.e. $\Pr(A \cap C)$. We know the conditional probability that a transmission is successful, given that it took route A, is $\Pr(C \mid A) = 0.8$ and the probability that route A is taken is $\Pr(A) = 0.5$. So it follows that $\Pr(A \cap C) = \Pr(C \mid A) \times \Pr(A) = 0.8 \times 0.5 = 0.4$.

If A and B are mutually exclusive $\Pr(A \mid B) = 0$. If B is a precondition for A, then $\Pr(A \cap B) = \Pr(A)$. Two events are *independent* if knowledge of the occurrence of one of them tells us nothing about the probability of the other, i.e. $\Pr(A \mid B) = \Pr(A)$. Thus, when two events are independent, the probability of them both occurring is

$$\boxed{\Pr(A \cap B) = \Pr(A) \times \Pr(B).}$$

3.1.2 Random Variables and Probability Distributions

As discussed in lecture note 1, a random variable is a mapping from the sample space to the real or natural numbers. We are often interested in calculating, for each possible value, with what probability the random variable takes that value. Conversely, knowing these probabilities gives us a lot of information about the sample space modelled by the random variable.

When the random variable takes discrete values such as the natural numbers we associate probabilities with the random variable in an intuitive way. That is, the probability that a random variable X has the value v is equal to the probability of the union of all events that lead to $X = v$.

Suppose that random variable X has values $\{x_i \mid i \in I\}$ and that for all $i \in I$ the probability that $X = x_i$ is equal to p_i. Then $\{p_i \mid i \in I\}$ is the *probability distribution* of X. This mapping from x_i to p_i is sometimes called the *mass function*. In the simple example of rolling a fair die, the values of the random variable are clearly $\{1, 2, 3, 4, 5, 6\}$ and the probability distribution is $\{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$.

The *expectation* of X is defined to be

$$\boxed{E[X] = \sum_{i \in I} x_i \times p_i.}$$

Informally, we can think of this as the average value that X will have. For rolling a fair die this will be

$$(1 \times 1/6) + (2 \times 1/6) + (3 \times 1/6) + (4 \times 1/6) + (5 \times 1/6) + (6 \times 1/6) = 21/6 = 3.5$$

For some sample spaces, discrete values of random variables are not appropriate. For example, if we are interested in the number of dial-in lines which are used each hour then a discrete random variable is ideal. However, if we are interested in the duration of each connection, the variable can take almost any non-negative value, so a continuous random variable is appropriate.

For continuous random variables we can no longer assign a probability to each sample point in a meaningful way. Instead we define the probability that the random variable takes a value which is less than a given value. This is done by a *distribution function*[2], F. For a continuous random variable X the distribution function is defined as

$$\boxed{F(x) = \Pr(X \leq x)}$$

Probability distribution functions satisfy four conditions:

1. $F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$;
2. $F(x) \longrightarrow 0$ as $x \longrightarrow -\infty$;
3. $F(x) \longrightarrow 1$ as $x \longrightarrow \infty$;
4. $F(x)$ is continuous[3].

The derivative of the distribution is known as the *density function* $f(x)$,

$$\boxed{f(x) = \frac{dF(x)}{dx}.}$$

It is convenient to regard the infinitesimal quantity $f(x)dx$ as the probability that the value of X lies in the interval $(x, x + dx)$. The *expectation* of the continuous random variable X is defined as

$$\boxed{E[X] = \int_{-\infty}^{\infty} x f(x) dx.}$$

The expectation of the sum of a number of random variables is the sum of their individual expectations. This is true even if the variables are dependent in some way. In other words, for continuous random variables, X and Y, $E[X + Y] = E[X] + E[Y]$.

Two continuous random variables X and Y with distribution functions $F_X(x)$ and $F_Y(y)$, are independent if their *joint* distribution function, $F(x, y)$ satisfies

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y) = F_X(x) \times F_y(y).$$

[2]Some authors call this the *cumulative distribution function*.

[3]Strictly, $F(x)$ only needs to be *right continuous*, but we will assume that $F(x)$ is continuous for all the continuous random variables which we will encounter in this course.

3.1.3 Two important distributions

The most important continuous probability distribution for performance modelling purposes is the *exponential distribution*[4] which has the distribution function

$$F(x) = 1 - e^{-\mu x} \quad x \geq 0 \qquad \text{and density function} \qquad f(x) = \mu e^{-\mu x} \quad x \geq 0$$

The expectation of an exponential random variable with parameter μ is $1/\mu$ and its variance is $1/\mu^2$. This random variable is often used to denote the time which elapses until some event occurs, such as the arrival of a task at a computer system, the arrival of a bus at a bus-stop or the outbreak of war.

Example: Suppose that the time for which a computer system is up is an exponential random variable with mean 20 days. What is the probability that the machine will stay up for more than 40 days?

Let X denote the random variable representing the up-time of the machine. Then we are interested in $\Pr(X > 40) = 1 - \Pr(X \leq 40)$ (i.e. the distribution function evaluated when $x = 40$ and $\mu = 1/20$):

$$\Pr(X > 40) = 1 - (1 - e^{-40/20}) = e^{-2} = 0.135 \qquad \text{to 3 significant figures}$$

The exponential distribution function is closely related to a discrete random variable, the *Poisson* distribution. This random variable takes values in the set $\{0, 1, 2, \ldots\}$ and has the mass function

$$p_i = e^{-\mu} \frac{\mu^i}{i!} \qquad i \geq 0.$$

The expectation of a Poisson random variable with parameter μ is also μ. The Poisson random variable arises frequently in computer and communication system modelling. It is typically used as a counting variable, recording the number of events that occur in a given period of time; for example, the number of messages transmitted in a communication network node in one minute or the number of tasks arriving at a computer in one hour. If we observe a Poisson process with parameter μ for some time period of length h then:

- the probability that one event occurs is $\mu h + o(h)$. [5].
- the probability that two or more events occur is $o(h)$.
- the probability that no events occur is $1 - \mu h + o(h)$.

Thus if we observe a Poisson process for a infinitesimal time period dt the probability that an event occur is μdt. If the occurrence of events is governed by a Poisson distribution then the inter-event times are governed by an exponential distribution with the same parameter, and vice versa. Therefore, if we know that the delay until an event is

[4]Some authors call this the *negative exponential distribution*.

[5]The notation $o(h)$ denotes terms of small relative size; formally a function f is $o(h)$ if $\lim_{h \to 0} \frac{f(h)}{h} = 0$ In practice, what this means is that these are terms which we can ignore when h is small.

exponentially distributed then the probability that it will occur in the infinitesimal time interval of length dt, is μdt.

The reason that the exponential distribution is used extensively in performance modelling is because it has several attractive mathematical features. Here we will state those properties without proof, but the proofs can be found in any textbook on stochastic modelling, Markov processes or performance modelling.

- The exponential distribution is *memoryless*: if X is an exponentially distributed random variable, then

  $$\Pr(X > t + s | X > t) = \Pr(X > s) \qquad t > 0, s > 0$$

  If X represents the waiting time until an event, as is often the case in a performance model, this implies that knowing that t time units have already elapsed without the event occuring does not give us any additional information about when the event will occur—the distribution of the further waiting time is the same as if no waiting time had already passed.

- If X_1 and X_2 are two exponentially distributed random variables, with parameters λ_1 and λ_2 respectively, and Y is defined to be their minimum, i.e. $Y = \min(X_1, X_2)$, then Y is also exponentially distributed, with parameter $\lambda_1 + \lambda_2$. This is sometimes called the *superposition* property.

- Let X be an exponentially distributed random variable, with parameter λ, representing time until an event. Suppose that the events are probabilistically divided into two categories, with events belonging to stream A with probability p_A and belonging to stream B with probability $p_B = 1 - p_A$. Then stream A and stream B are Poisson streams with parameters $\lambda \times p_A$ and $\lambda \times p_B$ respectively, and the waiting times between events of category A and between events of category B are exponentially distributed with parameters $\lambda \times p_A$ and $\lambda \times p_B$ respectively. This is sometimes called the *decomposition property.*

We will encounter more distributions later in course, when we come to study simulation. However, for Markov processes, we make the assumption that all inter-event times are exponentially distributed.

3.2 Stochastic Processes

As we discussed in lecture note 1, stochastic processes provide models for evolving systems by describing the sequences of states that they pass through. These states are characterised by one or more random variables. Recall that formally a stochastic process is defined to be family of random variables $\{X(t), t \in T\}$. Each $X(t)$ takes its values from some set S which is called the *state space* of the process. Any set of instances of $\{X(t), t \in T\}$ can be regarded as a path of a particle moving randomly in the state space, S, its position at time t being $X(t)$. These paths are called *sample paths* or *realisations* of the stochastic process.

The stochastic processes we will be considering during the first half of the course will all have the following properties:

$\{X(t)\}$ is a Markov process. This implies that $\{X(t)\}$ has the *Markov* or *memoryless property*: given the value of $X(t)$ at some time $t \in T$, the future path $X(s)$ for $s > t$ does not depend on knowledge of the past history $X(u)$ for $u < t$, i.e. for $t_1 < \cdots < t_n < t_{n+1}$,

$$\Pr(X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n, \ldots, X(t_1) = x_1) = \Pr(X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n)$$

$\{X(t)\}$ is stationary. This means that for any $t_1, \ldots t_n \in T$ and $t_1 + \tau, \ldots, t_n + \tau \in T$ $(n \geq 1)$, then the process's joint distributions are unaffected by the change in the time axis and so,

$$F_{X(t_1+\tau)\ldots X(t_n+\tau)} = F_{X(t_1)\ldots X(t_n)}$$

$\{X(t)\}$ is time homogeneous. This implies that the behaviour of the system does not depend on when it is observed. In particular, the transition rates between states are independent of the time at which the transitions occur. Thus, for all t and s, it follows that $\Pr(X(t + \tau) = x_k \mid X(t) = x_j) = \Pr(X(s + \tau) = x_k \mid X(s) = x_j)$.

$\{X(t)\}$ is irreducible. This implies that all states in S can be reached from all other states, by following the transitions of the process.

In this course, our primary objective with respect to a Markovian model will be to calculate the probability distribution of the random variable $X(t)$ over the state space S, as $t \longrightarrow \infty$. This is termed the *steady state probability distribution*. From this probability distribution we will derive performance measures based on subsets of states where some condition holds. Solution of Markov processes and derivation of performance measures will be discussed in detail in lecture note 4. In the rest of this note we will be concerned with the definition of a Markov process and how such models can be expressed.

3.3 Markov Processes

A Markov process is a stochastic process which satisfies the Markov property. Thus, like any stochastic process, a Markov process is characterised by a random variable $X(t)$, indexed over time t, and a state space S such that $X(t) \in S$ for all t. The dynamic behaviour of the system being modelled is represented by the transitions between these states, and times spent in states, *sojourn times*. In general, these holding times represent the periods while processing is occurring in the system, the transitions represent events in the system.

If a state $i \in S$ is entered at time t and the next state transition takes place at time $t + T$, T is the sojourn time in state i. By the Markov property, at any time point τ, the distribution of the time to the next change of state is independent of the time of the previous change of state. In other words, sojourn times are memoryless. Since the only probability distribution function which has this property is the exponential distribution this tells us that the sojourn time in any state of a Markov process is an exponentially distributed random variable. Hence at time τ, the probability that there is a state transition in the interval $(\tau, \tau + dt)$ is $q_i dt + o(dt)$, where q_i is the parameter of the exponential distribution of the sojourn time in state i.

Suppose that when a transition out of state i occurs, the new state is j with probability p_{ij}. By the Markov property, this must only depend on i and j. Thus we have, for $i \neq j$, $i, j \in S$,

$$\Pr(X(t + dt) = j \mid X(t) = i) = q_{ij} dt + o(dt)$$

where the $q_{ij} = q_i p_{ij}$, by the decomposition property. The q_{ij} are called the *instantaneous transition rates*, and if the average time delay before a transition from i to j is exponentially distributed with parameter μ, then $q_{ij} = \mu$.

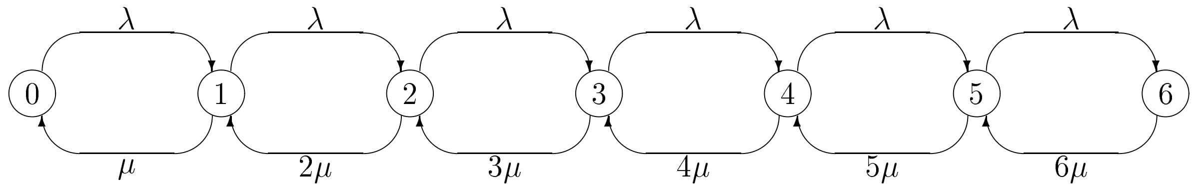
In practice, we are usually given the q_{ij} because we will know the parameter of the distribution governing the time delay associated with the event which is represented by the state transition. For example, we might be told that the average duration of a connection to the department's dial-in facility is $1/\mu$ seconds. If we are using Markovian analysis, we then *choose to assume* that the duration is exponentially distributed with parameter μ.

From the q_{ij} we can derive the *exit rate*, q_i, and the transition probabilities, p_{ij}, as follows. The exit rate is the rate at which the system leaves state i, i.e. it is the parameter of the exponential distribution governing the sojourn time. This will be the minimum of the delays until any of the possible transitions occurs. Thus, by the superposition property it is the sum of the individual transition rates, i.e. $q_i = \sum_{j \in S} q_{ij}$. The transition probability p_{ij} is the probability, given that a transition out of state i occurs, that it is the transition to state j. By the definition of conditional probability, this is $p_{ij} = q_{ij}/q_i$.

3.3.1 State transition diagrams

For small Markov processes the simplest way to represent the process is often in terms of its state transition diagram. In such a diagram we represent each state of the process as a node in a graph. The arcs in the graph represent possible transitions between states of the process. The arcs are labelled by the *transition rates* between states. Since every transition is assumed to be governed by an exponential distribution, the rate of the transition will be the parameter of the distribution. Irreducibility implies that the state transition diagram must be strongly connected.

Example: Consider again the department's dial-in lines and assume we take the simplest representation of the states of the system, i.e. the number of lines in use. Then, if there are 6 lines, the state space is $\{0, 1, 2, 3, 4, 5, 6\}$. The only possible transition from state 0 is to state 1 and this will be caused by a new connection. Similarly, from state 6 the only possible transition is to state 5, caused by a line being released. For any other state, j, $1 \leq j \leq 5$ there are two possible transitions, to states $j + 1$ or $j - 1$, caused by a new connection or release of an old one, respectively. Let us assume that the arrival of dial-in requests form a Poisson stream with parameter λ. This implies that the rate at which $j \to j + 1$ transitions occur is λ. As above, we assume that the mean duration of a connection is $1/\mu$, so that the rate at which release transitions occur on each line is μ. Then the state transition diagram for this process is:



3.3.2 Generator matrices

A Markov process with n states can also be characterised by its $n \times n$ *infinitesimal generator matrix*, $\mathbf{Q}$. The entry in the jth column of the ith row of the matrix ($j \neq i$) will be q_{ij}, the instantaneous transition rate. In other words, it will be the sum of the parameters labelling arcs connecting nodes i and j in the state transition diagram. The diagonal elements are chosen to ensure that the sum of the elements in every row is zero, i.e. $q_{ii} = -\sum_{j \in S, j \neq i} q_{ij}$.

Example: For the dial-in lines example, whose state transition diagram is shown in the previous section, the infinitesimal generator matrix is:

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & 0 \\ \mu & -(\mu+\lambda) & \lambda & 0 & 0 & 0 & 0 \\ 0 & 2\mu & -(2\mu+\lambda) & \lambda & 0 & 0 & 0 \\ 0 & 0 & 3\mu & -(3\mu+\lambda) & \lambda & 0 & 0 \\ 0 & 0 & 0 & 4\mu & -(4\mu+\lambda) & \lambda & 0 \\ 0 & 0 & 0 & 0 & 5\mu & -(5\mu+\lambda) & \lambda \\ 0 & 0 & 0 & 0 & 0 & 6\mu & -6\mu \end{pmatrix}$$

As we will see in the next lecture note, the generator matrix plays a crucial role in finding the probability distribution of the Markov process over its state space.

Jane Hillston ⟨jeh@dcs.ed.ac.uk⟩. January 16, 1996.