PM background material

Introduction

This course teaches various aspects of computer-aided modelling for the performance evaluation of computer systems and communication networks. The performance of a system influences its design, procurement and subsequent use. Performance analysis may be carried out at any stage of that life-cycle. In general, the problem is to achieve the highest possible performance given the constraints on the system. Typical constraints might be the number of users, the bandwidth or the cost. Modelling and simulation are methods which are commonly used by performance analysts to represent constraints and optimise performance.

Many of the ideas of the course will be presented by examples; as well as computer systems and communication networks, other dynamic systems such as manufacturing and transport systems will sometimes be considered. With all the examples it will be as important to consider the underlying concepts as the details of the example itself.

Discrete event systems

A common feature of all the systems which we will consider is that they are *discrete event* systems. What this means is that our view of the system (its *state*) is characterised by variables that take distinct values and which change at distinct times or *events* in the system. For example, we might be interested in the number of nodes in a wireless network which are currently waiting to send a message (N say). This number will change when one of two events occur: either a node, which was not previously waiting, generates a message and is now waiting to send ($N \to N + 1$), or a node, which was previously waiting, successfully transmits its message ($N \to N - 1$).

We will not consider *continuous* systems. These are systems in which the state variables characterising the system are constantly changing. An example of such a model would be the speed of a falling object or the drag experienced by a plane. Nevertheless we will sometimes find it useful to *approximate* the behaviour of a discrete system by a continuous one.

In all our modelling studies one of our major tasks will be to choose the parameters which *characterise* the system under study.

Discrete time discrete event models only consider the system at *predetermined moments in time*, which are typically evenly spaced, eg. at each clock "tick". State changes are observed only at these moments, which are not necessarily the moment at which the events themselves occur. Such models are often used to approximate systems with continuously changing states. *Continuous time* discrete event models, consider the system at the *time of each event*. Hence the time parameter in such models is conceptually continuous. At levels of abstraction above the hardware clock these models are generally appropriate for computer and communication systems.

Probability Theory

Although computer systems by their very nature are deterministic systems the models we will be considering will be *stochastic*. Loosely speaking, a stochastic model is one in which behaviour is random. We use such models because the world that computers interact with is by no means deterministic, but instead full of random behaviour. Also, the stochastic nature of the models will allow us to consider typical or average behaviour rather than particular behaviour. We will rely heavily on basic probability theory when constructing and solving our models.

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 a mobile phone antenna using frequency division multiplexing the sample space is the complete set of observations. Example events might be that more than three frequencies 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 $\Pr(\neg A) = 1 - \Pr(A)$. $P(A, B)$ is the *joint probability* that both A and B occur.

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

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

Random Variables and Probability Distributions

Often when we conduct an experiment or investigation we are interested in some number associated with the experiment rather than the actual outcome. For example, if we consider the mobile phone antenna again, we are likely to be interested in the number of frequencies in use rather than the details of each connection. Thus we naturally construct a mapping from the observations we make of the actual interactions, the *sample space*, to the value we are interested in. Such a function, from the sample space to the real or natural numbers is a continuous or discrete *random variable*, respectively. 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 $\{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$. This is a *uniform* distribution.

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 frequencies 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

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

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, $E[X + Y] = E[X] + E[Y]$.

The expectation only gives us an idea of the average value assumed by a random variable, not how much individual values may differ from this average. The *variance*, $Var(X)$, gives us an indication of the "spread" of values:

$$\boxed{Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2.}$$

[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.

Two important distributions

The most important continuous probability distribution for performance modelling purposes is the *(negative) exponential distribution* 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 sms messages transmitted from a mobile phone in one hour or the number of database queries arriving at a server in one minute. If we observe a Poisson process with parameter μ for some time period of length h then:

- the probability that one event occurs is $\mu h + o(h)$.[4]
- 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 occurs 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 exponentially distributed then the probability that it will occur in the infinitesimal time interval of length dt, is μdt.

[4]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.

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, 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*.

Linear Algebra

We will use *vectors* (columns of numbers) to represent (discrete) probability distributions over the discrete states of the systems we consider. Typically such a vector will be denoted $\pi(t)$ if it is the probability at a particular time t, or simply π when we consider the *equilibrium* or *steady state distribution*. We use the notation e to denote the vector in which all entries are 1, and e_i the vector in which all entries are 0 except for the ith which has value 1.

A *matrix* is a two-dimensional grid of numbers and we will use matrices to represent the relationship between states. For example, in a matrix Q entry $q_{i,j}$ will capture the rate of moving from state i to state j. We use the notation I to denote the *identity matrix*, i.e. the matrix for which

$$I_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Given a matrix Q, Q^T is the *transpose* of Q

$$\left(Q^T\right)_{i,j} = Q_{j,i}$$

Jane Hillston ⟨Jane.Hillston@ed.ac.uk⟩. January 8, 2013.