MATH3013: Simulation and Queues

Athanassios (Thanos) N. Avramidis
School of Mathematics
University of Southampton
Highfield, Southampton, United Kingdom

April 2012

Contents

.

Chapter 1

Probability

This chapter builds foundations in probability, the main mathematical tool behind queues and stochastic simulation. Because it is more foundational than an end in itself, the examination under-emphasizes it. Guidance for the examination is given in Chapter 5.

1.1 Preliminaries

The expression "x := y" defines x as being equal to y, while "x =: y" defines y as being equal to x.

1.2 Probability Spaces*

A *random experiment* involves an outcome that cannot be determined in advance. The set of all possible outcomes is called the *certain event* and denoted Ω.

An *event* is a subset of the certain event. An event A is said to occur if and only if the observed outcome ω of the experiment is an element of the set A.

Example 1.1 Consider the experiment of flipping a coin once. The two possible outcomes are "Heads" and "Tails", and a natural certain event is the set $\{H, T\}$.

Given a certain event Ω and an event A, the event that occurs if and only if A does *not* occur is called the *complement* of A and is denoted A^c. That is,

$$A^c = \{\omega \in \Omega : \omega \notin A\}.$$

Given two events A and B, their *union* is the event "A or B", also written

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$

The *intersection* of A and B is the event "A and B", also written

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$

The operations of complement, union, and intersection arise frequently and give new events, for which we observe the following identities:

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c. \tag{1.1}$$

The first says "not (A or B)" equals "(not A) and (not B)". The second says "not (A and B)" equals "(not A) or (not B)".

The set containing no elements is called the *empty event* and is denoted $\emptyset$. Note that $\Omega^c = \emptyset$ and $\emptyset^c = \Omega$.

Event A is said to *imply* event B, written $A \subset B$, if every element of A belongs to B; in other words, the occurrence of A makes the occurrence of B certain. This is also written as $B \supset A$.

The union of several (i.e., more than two) events means the occurrence of *any* one of the events, and the intersection of several events means the occurrence of *all* the events.

Two events A and B are called *disjoint* if they cannot happen simultaneously; equivalently, they have no elements in common, i.e.,

$$A \cap B = \emptyset,$$

A family of events is called disjoint if every pair of them are disjoint.

The following is our definition of probability.

Definition 1.2 *Let Ω be a certain event and let $\mathbb{P}$ be a function that assigns a number to each event. Then $\mathbb{P}$ is called a* probability *provided that*

1. *For any event A, $0 \leq \mathbb{P}(A) \leq 1$;*
2. *$\mathbb{P}(\Omega) = 1$;*
3. *for any sequence $A_1, A_2, \ldots$ of disjoint events,*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{1.2}$$

The following summarises properties of a probability $\mathbb{P}$.

Proposition 1.3 *(a)* $\mathbb{P}(\emptyset) = 0$.

(b) For any event A, $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.

(c) If events A and B satisfy $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(d) If events $A_1, A_2, \ldots A_n$ are disjoint, then $\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$.

Proof. Left as exercise. □

1.3 Random Variables and Distributions

Definition 1.4 *A* random variable *(rv) X with values in the set E is a function that assigns a value $X(\omega)$ in E to each outcome ω in Ω.*

We will refer to the set E as the *support* of X. In all cases of interest to us, E is one of the following: (a) the set of integer numbers; (b) the set of real numbers; (c) a subset of these sets.

Any set containing a support is also a support, so it is most useful to prescribe the smallest possible support. An example of a discrete random variable is the number, X, of successes during n independent trials each having success probability p, i.e., X has the Binomial(n, p) distribution. Here, the smallest possible support is the set $\{0, 1, \ldots, n\}$.

Example 1.1 (continued) Define X by putting $X(H) = 1$, $X(T) = -1$. Then X is a random variable with support $\{-1, 1\}$.

Notationally, we generally abbreviate $\mathbb{P}(\{\omega : X(\omega) \leq b\})$ as $\mathbb{P}(X \leq b)$.

The function F defined by

$$F(b) = \mathbb{P}(X \leq b), \quad -\infty < b < \infty$$

is called the *(Cumulative) Distribution Function* of the random variable X. We call F in short the *cdf of X*.

Example 1.1 (continued) Suppose the probability of "Heads" is 0.6. We have

$$\mathbb{P}(X = -1) = \mathbb{P}(\{T\}) = 1 - \mathbb{P}(\{H\}) = 0.4, \quad \mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 0.6,$$

and thus the cdf of X is

$$F(b) = \mathbb{P}(X \leq b) = \begin{cases} 0 & b < -1, \\ 0.4 & -1 \leq b < 1, \\ 1 & 1 \leq b. \end{cases} \tag{1.3}$$

A cdf has the following properties.

Proposition 1.5 *If F is a cdf of a finite-valued rv, then:*

(a) F is nondecreasing.

(b) F is right-continuous .

(c) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$.

(d) $F(\infty) = \lim_{x\to\infty} F(x) = 1$.

Proof. We only prove (a). If $a \le b$, then

$$\{X \le a\} \subset \{X \le b\} \tag{1.4}$$

and Proposition 1.3(c) gives

$$\mathbb{P}(X \le a) \le \mathbb{P}(X \le b). \tag{1.5}$$

□

1.3.1 Discrete distributions

In Example 1.1 the support had two points. Consider now the general case of a discrete rv X with support the *ordered* points

$$x_1 < x_2 < x_3 < \dots. \tag{1.6}$$

Then, for each $k = 1, 2, 3, \dots$, the cdf at x_k is the sum of the probabilities associated to the values less than or equal to x_k:

$$F(x_k) = \sum_{j:x_j \le x_k} \mathbb{P}(X = x_j) = \sum_{j:j\le k} \mathbb{P}(X = x_j).$$

Because each term in the sum is non-negative, we have $F(x_1) \le F(x_2) \le \dots$. Thus, the cdf is the non-decreasing step function

$$F(x) = \begin{cases} 0 & & x & < x_1 \\ F(x_1) & x_1 \le & x & < x_2 \\ F(x_2) & x_2 \le & x & < x_3 \\ \vdots & & & \\ F(x_k) & x_k \le & x & < x_{k+1} \\ \vdots & & & \end{cases} \tag{1.7}$$

In the opposite direction, the individual probabilities follow from the cdf as $\mathbb{P}(X = x_k) = F(x_k) - F(x_{k-1})$ for all k.

1.3.2 Continuous distributions

A major class of non-discrete distributions has a support that is not *countable*, meaning it cannot be enumerated, i.e., cannot be represented as in (1.6). A typical such support is any subinterval of the real numbers, i.e, $[a, b]$, with $a < b$. (Enumerating

this set is impossible because for any real number $x_1 > a$, there exists a real number x_2 such that $a < x_2 < x_1$.)

Suppose further that F is differentiable at x, i.e., the left- and right-derivatives of F at x are equal, i.e., we can define

$$F'(x) = \lim_{\epsilon \to 0+} \frac{F(x+\epsilon) - F(x)}{\epsilon} = \lim_{\epsilon \to 0+} \frac{F(x) - F(x-\epsilon)}{\epsilon} = \lim_{\epsilon \to 0+} \frac{F(x+\epsilon) - F(x-\epsilon)}{2\epsilon}. \tag{1.8}$$

The last enumerator is $\mathbb{P}(x - \epsilon < X \leq x + \epsilon)$, so $F'(x)$ describes the "probabilistic intensity" of falling "arbitrarily close to x". For this reason, it is called the *probability density function* (pdf) (of F and of the associated rv).

The cdf F and its pdf F' are mirrors of each other: they are linked by a differentiation step when going from F to F' (by definition); and they are linked by an integration step when going from F' to F (this is said by the Fundamental Theorem of Calculus (FTC)). The integration link is

$$F(b) = F(b) - F(-\infty) = \int_{-\infty}^{b} F'(t)\mathrm{d}t, \quad -\infty < b < \infty. \tag{1.9}$$

Alternatively, we could calculate the function $1 - F(b) = \mathbb{P}(X > b)$ (called the *tail probability* or *complementary cdf*) as

$$1 - F(b) = F(\infty) - F(b) = \int_{b}^{\infty} F'(t)\mathrm{d}t, \quad -\infty < b < \infty.$$

The integral in (1.9) is the area under the graph of F' between $-\infty$ and b. If F' changes form anywhere there, then the integral (area) is typically calculated piecewise. The following is a minimal example illustrating this.

Example 1.6 Equiprobable outcomes on $[2, 4] \cup [5, 6]$ (i.e., the union of these real intervals) are described by the pdf

$$f(x) = \begin{cases} \frac{1}{3}, & 2 \leq x \leq 4 \\ \frac{1}{3}, & 5 \leq x \leq 6 \end{cases}$$

$F(x)$ is calculated as the area piecewise. The answer is

$$F(x) = \begin{cases} \frac{x-2}{3}, & 2 \leq x \leq 4 \\ \frac{2}{3}, & 4 < x \leq 5 \\ \frac{2}{3} + \frac{x-5}{3}, & 5 < x \leq 6 \end{cases}$$

The Uniform Distribution Let $a < b$. The Unif(a, b) distribution has support $[a, b]$ and constant pdf. Thus, the cdf F satisfies $F(a) = 0$ and $F(b) = 1$. The

constant value of the pdf, call it c, is determined from

$$1 = F(b) = \int_a^b c\mathrm{d}t = c(b-a),$$

i.e., $c = 1/(b-a)$. Thus,

$$F(x) = \int_a^x \frac{1}{b-a}\mathrm{d}t = \frac{x-a}{b-a}, \quad a \le x \le b. \tag{1.10}$$

In particular, the Unif$(0,1)$ distribution has cdf $F(x) = x$ and pdf $f(x) = 1$ with support $[0,1]$.

The Exponential Distribution Let $\lambda > 0$. We write $X \sim \text{Expon}(\lambda)$ and read "X has the exponential distribution with rate λ" to mean that X has support $[0,\infty)$ and has *tail probability*

$$\mathbb{P}(X > x) = e^{-\lambda x}, \quad x > 0. \tag{1.11}$$

Equivalently, the cdf is

$$\mathbb{P}(X \le x) = 1 - e^{-\lambda x}, \quad x \ge 0,$$

and the pdf is

$$\frac{d}{dx}(1 - e^{-\lambda x}) = \lambda e^{-\lambda x}, \quad x > 0.$$

1.4 Conditional Probability

Let Ω be a certain event and let $\mathbb{P}$ be a probability on it.

Definition 1.7 *Let B be an event such that $\mathbb{P}(B) > 0$. For any event A, the* conditional probability *of A given B, written $\mathbb{P}(A|B)$, is a number satisfying*

(a) $0 \leq \mathbb{P}(A|B) \leq 1$*;*

(b) $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$*.*

Remark 1.8 Fix B such that $\mathbb{P}(B) > 0$. Then, the $\mathbb{P}(A|B)$, viewed as a function of A, satisfies all the conditions in Definition 1.2. That is:

1. $0 \leq \mathbb{P}(A|B) \leq 1$;
2. $\mathbb{P}(\Omega|B) = 1$;
3. For any sequence $A_1, A_2, \ldots$ of disjoint events, $\mathbb{P}(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$.

That is, the probability $\mathbb{P}(\cdot|B)$ satisfies all the usual properties as the unconditional probability, $\mathbb{P}(\cdot)$.

The intuition is: knowledge that the event B has occurred *forces* us to revise the probability of all events. A key exception to this happens when A and B are independent, as will be seen shortly.

We now give a fundamental idea for probabilistic calculations.

Definition 1.9 *(Partition.) The set of events $\{B_1, B_2, \ldots\}$ is called a* partition *if it satisfies:*

1. *(Disjointness.) $B_i \cap B_j = \emptyset$ for all $i \neq j$ (B_i and B_j cannot happen simultaneously).*
2. *(Exhaustiveness.) $\cup_{i=1}^{\infty} B_i = \Omega$. (One of the B_i must happen.)*

Theorem 1.10 *(Law of Total Probability (LTP)). Let B_1, B_2, ... be a partition, i.e., the B_i are disjoint and exhaustive. Then, for any event A,*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i). \tag{1.12}$$

Proof. Write

$$A = \cup_{i=1}^{\infty}(A \cap B_i) \tag{1.13}$$

and take probabilities, noting that the $A \cap B_i$ are disjoint because the B_i are disjoint, to obtain

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i).$$

Finally, replace $\mathbb{P}(A \cap B_i)$ by $\mathbb{P}(A|B_i)\mathbb{P}(B_i)$. $\square$

1.5 Independence

Definition 1.11 *Events A and B are said to be* independent *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.14}$$

If (1.14) holds, then Definition 1.7 gives $\mathbb{P}(A|B) = \mathbb{P}(A)$ (and also $\mathbb{P}(B|A) = \mathbb{P}(B)$), i.e., the conditional probabilities equal the unconditional ones. The intuition behind independence is that the occurrence of one event does not contain information about the occurrence of the other.

Definition 1.12 *The random variables X_1 and X_2 are said to be* independent *if*

$$\textit{the events } \{X_1 \leq b_1\} \textit{ and } \{X_2 \leq b_2\} \textit{ are independent for all } b_1, b_2. \tag{1.15}$$

It turns out that (1.15) implies that all "interesting" events about X_1 and X_2 are independent (e.g., events where "$\leq$" is replaced by "$\geq$" or by "$=$" are also independent).

1.6 Mean and Variance, and Sums

Definition 1.13 *Suppose X is a discrete random variable with support $\{1, 2, 3, \ldots\}$ and its distribution is $\mathbb{P}(X = i) = p_i$ for all i. The expected value (mean) of X is*

$$\mathbb{E}[X] = 1p_1 + 2p_2 + 3p_3 + \cdots = \sum_{i=1}^{\infty} ip_i. \tag{1.16}$$

$\mathbb{E}[X]$ can be seen to be the area on the (x, y) plane, bounded below by the cdf of X, bounded above by the line $y = 1$, and bounded to the left by the line $x = 0$.

Definition 1.14 *Suppose X is a random variable with mean μ. The* variance *of X is the mean square deviation of X from its mean μ:*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2]. \tag{1.17}$$

Here is how summation of random variables affects the mean and variance.

Fact 1.15 *1. For any random variable X and constants a, b,*

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X]. \tag{1.18}$$

2. For any random variables X_1 and X_2,

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]. \tag{1.19}$$

Fact 1.16 *If X_1 and X_2 are independent random variables, then*

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2). \tag{1.20}$$

Properties (1.19) and (1.20) extend to any sum of finitely many random variables, as can be seen by mathematical induction; that is,

$$\begin{aligned} \mathbb{E}[X_1 + X_2 + \ldots + X_n] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_n], \\ \mathrm{Var}(X_1 + X_2 + \ldots + X_n) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) \quad \text{for independent } X\text{'s.} \end{aligned}$$

1.7 Probability Generating Functions

Definition 1.17 *Let X be a discrete random variable with support $\{0, 1, \ldots\}$ and probabilities $p_n := \mathbb{P}(X = n)$, where $\sum_{n=0}^{\infty} p_n = 1$. The* probability generating function *(pgf) of X is the function*

$$G(z) = \sum_{n=0}^{\infty} p_n z^n, \quad 0 \le z \le 1. \tag{1.21}$$

The pgf contains the distribution of X, as stated below.

Fact 1.18 *The function G and its derivatives of all orders are continuous functions on $[0, 1]$.*

Write the n-order derivative of G as $G^{(n)}$, and put $G^{(0)} = G$. We have:

$$G^{(1)}(z) = \frac{d}{dz} G(z) = \sum_{n=0}^{\infty} \frac{d}{dz} p_n z^n = \sum_{n=1}^{\infty} p_n n z^{n-1}, \tag{1.22}$$

$$G^{(2)}(z) = \frac{d}{dz} G^{(1)}(z) = \sum_{n=1}^{\infty} \frac{d}{dz} p_n n z^{n-1} = \sum_{n=2}^{\infty} p_n n(n-1) z^{n-2}. \tag{1.23}$$

The derivatives of higher order work analogously. At $z = 0$, powers 0^k appear; they equal zero unless $k = 0$, where $0^0 = 1$. That is, we get

$$G(0) = p_0, \quad G^{(1)}(0) = p_1, \quad G^{(2)}(0) = 2p_2,$$

and, in the same way, $G^{(n)}(0) = n! p_n$ for all $n > 0$. We now see all of the following:

Fact 1.19 *The pgf gives the distribution as*

$$p_0 = G(0), \quad p_n = \frac{G^{(n)}(0)}{n!}, \quad n = 1, 2, \ldots. \tag{1.24}$$

Fact 1.20 $G(1) = \sum_{n=0}^{\infty} p_n = 1$, *directly from (1.21).*

Fact 1.21 *The pgf determines the mean of X as*

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} p_n n = G^{(1)}(1) \quad \textit{from (1.22).}$$

1.8 Two Major Limit Theorems

For independent and identically distributed (iid) random variables X_1, X_2, ..., there are strong theorems about their average and their sum, as the number n being averaged or summed goes to infinity. The first theorem is about the average.

Theorem 1.22 (Strong Law of Large Numbers (SLLN)) *If X_1, X_2, ... are independent and identically distributed (iid) random variables with finite mean μ, then their average, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, converges to μ in a probabilistic sense:*

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{1.25}$$

In words, the average converges to the mean *with probability one*, or *w.p. 1.*

The second theorem is about the sum, $S_n = X_1 + X_2 + \cdots + X_n$, as n goes to infinity. Here is an example where the theorem is useful.

Example 1.23 A person, JD, arrives at a single-server system and finds $n = 100$ customers ahead, one of them already in service. Suppose the individual customer service (processing) times are iid random variables X_1, X_2, ..., with mean $\mu = 5$ and variance $\sigma^2 = 4$. Can we approximate JD's waiting time in queue?

In general, the in-service customer has a *remaining service time*, S^{rem}, whose distribution depends on the elapsed time in service (unless service times are exponentially distributed, as we will see). But provided n is not small, the sum of n service times should not be affected too much by any one. Let us then pretend the in-service customer's remaining service time is like a "fresh" one, so JD's waiting time is the sum $S_{100} = X_1 + X_2 + \ldots + X_{100}$, where the X_i are iid with mean 5 and variance 4.

One approximation could be based, roughly, on the SLLN:

$$S_{100} = 100 \underbrace{\frac{1}{100}(X_1 + \ldots + X_{100})}_{\approx \mu = 5} \approx 500. \tag{1.26}$$

Using only the mean service time, we got a deterministic approximation of S_{100}.

The Central Limit Theorem, (1.27) below, says that the distribution of S_n is approximately Normal with mean

$$\mathbb{E}[S_n] = n\mu \quad \text{by (1.19)}$$

and variance

$$\text{Var}(S_n) = n\sigma^2 \quad \text{by (1.20)}.$$

Thus, $S_{100} \overset{\text{approx}}{\sim} \text{N}(500, 400)$, where $\overset{\text{approx}}{\sim}$ means "is distributed approximately as", and $\text{N}(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2.

The CLT is recorded below.

Theorem 1.24 (Central Limit Theorem (CLT)) *Let X_1, X_2, ... be independent and identically distributed (iid) random variables with mean μ and variance $\sigma^2 < \infty$, and put $S_n = X_1 + X_2 + \cdots + X_n$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{as } n \to \infty,$$

where $\xrightarrow{\mathcal{D}}$ means convergence in distribution. More practically,

$$S_n \overset{\text{approx}}{\sim} \mathrm{N}(n\mu, n\sigma^2) \overset{\mathcal{D}}{=} \underbrace{n\mu}_{\textit{deterministic}} + \underbrace{\sigma\sqrt{n}\mathrm{N}(0,1)}_{\textit{stochastic}} \quad \text{as } n \to \infty, \tag{1.27}$$

where $\overset{\mathcal{D}}{=}$ means equality in distribution [1].

Remarks:

1. No assumption is made about the distribution of the X_i.

2. Assuming $\mu \neq 0$, we can see that as $n \to \infty$, the stochastic effect becomes (in the limit) negligible relative to the deterministic effect. To see this, divide the multiplier of the stochastic term $\mathrm{N}(0,1)$ over the deterministic term:

$$\frac{\sqrt{n}\sigma}{n\mu}$$

   and note this goes to zero.

[1] Recall that $X \sim \mathrm{N}(0,1)$ is equivalent to $aX + b \sim \mathrm{N}(b, a^2)$, for any a, b.

.

1.9 Questions / Exercises

Indications: "P": proof; "S": short answer that refers appropriately to definitions.

1. Two standard dice are thrown, and the outcomes X_1, X_2 (a number in $\{1, 2, \ldots 6\}$ for each) are recorded. Assume X_1 and X_2 are independent. Are X_1 and X_1+X_2 independent random variables? Explain carefully.

2. For a random variable with pdf $f()$, the expected value can be defined as

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x. \tag{1.28}$$

   Its variance can then be defined as

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) \mathrm{d}x. \tag{1.29}$$

   Compute the mean and variance of the exponential distribution seen in Section 1.3.2. *Hint:* Note the identity $\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2$.

3. Out of n servers in a system (e.g., persons answering calls at a call center), seen at a certain time point, say at noon, some are temporarily absent (e.g., they are at the bathroom). Assume individual-server absences occur independently of each other with probability p. Let N be the number of absent servers. Express the mean and variance of N in terms of n and p and use the CLT to approximate the distribution of N.

4. In Example 1.23, let the number of customers ahead (number of rv's in the sum) be n. Express the probability that the approximated (i.e., normally-distributed) waiting time is at least 10% above its mean, writing it as a function of the $\mathrm{N}(0,1)$ cdf and n, then compute it explicitly for $n = 4$, 16, 64, and 256. Determine the limit of this as $n \to \infty$.

5. (S) Suppose a customer arrives at a system where multiple servers each serve their own queue and chooses to join the shortest queue. Suggest assumptions, focusing on the customer service times seen as random variables, that would explain this customer behaviour. Strive for the weakest assumption(s) possible. Then, assuming shorter waiting is preferred, suggest a situation where join-the-shortest queue would not necessarily make sense.

6. (P)

   (a) Let X and Y be independent random variables, each with support the integer numbers. Show that

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{1.30}$$

   Hint: Put $A_i := \{X = i\}$ and $B_j := \{Y = j\}$. Use the fact that A_i and B_j are independent events for all i and j.

   (b) Suppose X_1 and X_2 are rvs, and let μ_1, μ_2 be their respective means. Observe

$$\mathrm{Var}(X_1 + X_2) = \mathbb{E}[(X_1 - \mu_1 + X_2 - \mu_2)^2].$$

   Show that when X_1 and X_2 are discrete and independent, the above equals $\mathrm{Var}(X_1) + \mathrm{Var}(X_2)$. *Hint:* expand the square appropriately, and use properties of $\mathbb{E}[]$, including (1.30).

7. (P) Throughout, X is a real-valued random variable with cdf F, and b is a real (number). We give an idea why F is right-continuous at b (the result in Proposition 1.5(b)). If the events $A_1, A_2, \ldots$ "decrease to" A, meaning that $A_1 \supset A_2 \supset A_3 \supset \ldots$ and $A = \cap_{n=1}^{\infty} A_n$, then it can be shown that

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{1.31}$$

Now, note that the events $A_n = \{X \leq b + 1/n\}$, $n = 1, 2, 3, \ldots$, decrease to the event $\{X \leq b\}$. Applying (1.31),

$$F(b) = \mathbb{P}(X \leq b) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} F(b + \frac{1}{n})$$

i.e., F is right-continuous at b.

(a) If the events $A_1, A_2, \ldots$ "increase to" A, meaning that $A_1 \subset A_2 \subset A_3 \subset \ldots$ and $A = \cup_{n=1}^{\infty} A_n$, then again (1.31) holds. Using this property, show that

$$\mathbb{P}(X < b) = F(b-) \tag{1.32}$$

where $F(b-) = \lim_{x\to b-} F(x)$, the left-limit of F at b.

(b) Show that

$$\mathbb{P}(X = b) = F(b) - F(b-). \tag{1.33}$$

(c) Define X to be a *continuous* random variable if

$$\mathbb{P}(X = b) = 0, \quad \text{all } b. \tag{1.34}$$

Restate this definition via a property of F, the cdf of X.

1.10 Solutions to Exercises

1. Short argument: as $X_1 + X_2$ contains X_1, we do not expect X_1 and $X_1 + X_2$ to be independent. As a more careful argument, we show one conditional probability about $X_1 + X_2$, given an event about X_1, that is different from the unconditional one: $\mathbb{P}(X_1 + X_2 = 8 | X_1 = 1) = 0$, versus the (easily seen) $\mathbb{P}(X_1 + X_2 = 8) > 0$. The inequality of the two shows that independence fails.

2. We need the integration-by-parts rule [2]

$$\int_a^b f(x)g'(x)\mathrm{d}x = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)\mathrm{d}x. \tag{1.35}$$

where $f'(x) = \frac{d}{dx}f(x)$, and likewise for g. The calculation is:

$$\mathbb{E}[X] = \int_0^\infty y\lambda e^{-\lambda y}\mathrm{d}y = \frac{1}{\lambda}\int_0^\infty \underbrace{x}_{f(x)}\underbrace{e^{-x}}_{g'(x)}\mathrm{d}x$$

[2] This is obtained from $f(x)g(x)\big|_a^b = \int_a^b (fg)'(x)\mathrm{d}x = \int_a^b f'(x)g(x)\mathrm{d}x + \int_a^b f(x)g'(x)\mathrm{d}x$.

(change of variable $x = \lambda y$), where the last integral is

$$I := \int_0^\infty x e^{-x} \mathrm{d}x = x(-e^{-x})\big|_0^\infty - \int_0^\infty 1 \cdot (-e^{-x}) \mathrm{d}x = 0 + e^{-x}\big|_0^\infty = 1,$$

(by (1.35) with $f(x) = x$, $g(x) = -e^{-x}$, $a = 0$, $b = \infty$). Thus, $\mathbb{E}[X] = 1/\lambda$.

Now, calculate $\mathbb{E}[X^2]$ using similar steps:

$$\mathbb{E}[X^2] = \int_0^\infty y^2 \lambda e^{-\lambda y} \mathrm{d}y = \frac{1}{\lambda^2} \int_0^\infty \underbrace{x^2}_{f(x)} \underbrace{e^{-x}}_{g'(x)} \mathrm{d}x,$$

(again $x = \lambda y$), and the last integral is, using (1.35) with $f(x) = x^2$ and g, a, b as above,

$$\int_0^\infty x^2 e^{-x} \mathrm{d}x = x^2(-e^{-x})\big|_0^\infty - \int_0^\infty (2x)(-e^{-x}) \mathrm{d}x = 0 + 2I = 2.$$

Thus, $\mathbb{E}[X^2] = 2/\lambda^2$, and now $\mathrm{Var}(X) = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$.

3. Let I_j be random, taking value 1 if server j is absent, and 0 otherwise. The number of absent servers is $N = I_1 + I_2 + \ldots + I_n$. By assumption, the I_j are independent and identically distributed, where $\mathbb{P}(I_1 = 1) = p$, so $\mathbb{P}(I_1 = 0) = 1 - p$. The CLT gives $N \overset{\text{approx}}{\sim} \mathrm{Normal}(\mathbb{E}[N], \mathrm{Var}(N))$, where

$$\begin{aligned} \mathbb{E}[N] &= n\mathbb{E}[I_1] \\ \mathrm{Var}(N) &= n\mathrm{Var}(I_1) \\ \mathbb{E}[I_1] &= 1 \cdot p + 0 \cdot (1-p) = p \\ \mathrm{Var}(I_1) &= (1-p)^2 \cdot p + (0-p)^2(1-p) = p(1-p). \end{aligned}$$

4. By the CLT, the approximated waiting time when there are n customers ahead is $X \sim \mathrm{N}(n\mu, n\sigma^2)$, or equivalently $\frac{X - n\mu}{\sqrt{n}\sigma} \sim \mathrm{N}(0,1)$.

   The exercise inadvertently asked about the event "X is 10% above its mean". This event has probability zero, as will be seen. Now we calculate the probability that "$X \geq 10\%$ above its mean", i.e., $\{X \geq 1.1n\mu\}$.

$$\mathbb{P}(X \geq 1.1n\mu) = \mathbb{P}(X > 1.1n\mu) = \mathbb{P}\left(\frac{X - n\mu}{\sqrt{n}\sigma} > \frac{0.1n\mu}{\sqrt{n}\sigma}\right) = 1 - \Phi\left(\frac{0.1n\mu}{\sqrt{n}\sigma}\right)$$

   where $\Phi()$ is the $\mathrm{N}(0,1)$ cdf. In the first step we used $\mathbb{P}(X = 1.1n\mu) = 0$. [3] As $n \to \infty$, the argument inside Φ goes to ∞, so the $\mathbb{P}()$ goes to $1 - \Phi(\infty) = 0$. For calculations, I used `matlab` and the code "`n=[4 16 64 256]; proba= 1 - normcdf(0.1*5*n./(sqrt(4*n)))`". The probabilities are, in order, 0.3085, 0.1587, 0.0228, and 0.00003167.

5. Assume first-come first-serve discipline and let n_i be the number of people at server i upon arrival of a test customer X. Assume X joins one server immediately, and does not switch to another server. If service times are identically distributed, then,

[3] this follows from (1.34) and the continuity of the normal distribution.

by (1.19), the mean waiting time of X at server i is $n_i b$, where b is the mean service time. Minimising the mean waiting time is then equivalent to minimising n_i across i, i.e., joining the shortest queue. Independence of service times is not necessary in this argument. If a server tends to be "faster" than others, then the assumption "identically distributed service times across servers" does not seem sensible. *Note:* The argument generalises: if the mean service time (per customer) at server i is b_i, then the mean waiting time at i is $n_i b_i$, and we could minimise this.

6(a)

$$\mathbb{E}[XY] = \sum_i \sum_j ij\mathbb{P}(X = i, Y = j). \tag{1.36}$$

Using the independence, this equals $\sum_i \sum_j ij\mathbb{P}(X = i)\mathbb{P}(Y = j) = \sum_i i\mathbb{P}(X = i) \sum_j j\mathbb{P}(Y = j) = \mathbb{E}[X]\mathbb{E}[Y]$.

Note: it might be more natural to only sum the distinct outcomes of XY, unlike (1.36) above, but this again results in (1.36), as we now explain. For any k, write $\{XY = k\}$ as the union (logical "or") of disjoint events $\cup_{(i,j):ij=k}\{X = i, Y = j\}$ (for example, $\{XY = 3\} = \{X = 1, Y = 3\} \cup \{X = 3, Y = 1\}$). Thus $\mathbb{P}(XY = k) = \sum_{(i,j):ij=k} \mathbb{P}(X = i, Y = j)$, and (1.36) follows from

$$\begin{aligned} \mathbb{E}[XY] &= \sum_k k\mathbb{P}(XY = k) \\ &= \sum_k \sum_{(i,j):ij=k} ij\mathbb{P}(X = i, Y = j) = \sum_i \sum_j ij\mathbb{P}(X = i, Y = j). \end{aligned}$$

(b) $(X_1 - \mu_1 + X_2 - \mu_2)^2 = (X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2)$. The mean of this equals

$$\begin{aligned} &\mathbb{E}[(X_1 - \mu_1)^2] + \mathbb{E}[(X_2 - \mu_2)^2] + 2\mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \quad \text{by (1.19) and (1.18)} \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathbb{E}[X_1 - \mu_1]\mathbb{E}[X_2 - \mu_2] \quad \text{by part (a) and Var() definition} \end{aligned}$$

and the rightmost term is zero ($\mathbb{E}[X_1 - \mu_1] = \mathbb{E}[X_1] - \mu_1 = 0$).

7(a) The events $A_n = \{X \le b - 1/n\}$, $n = 1, 2, \ldots$ increase to the event $\{X < b\}$, so (1.31) gives

$$\mathbb{P}(X < b) = \lim_{n\to\infty} \mathbb{P}(A_n),$$

which is $\lim_n F\left(b - \frac{1}{n}\right) = F(b-)$.

(b) Now, $\{X \le b\} = \{X < b\} \cup \{X = b\}$, and the events $\{X < b\}$ and $\{X = b\}$ are disjoint, so, by Proposition 1.3(d), $\mathbb{P}(X \le b) = \mathbb{P}(X < b) + \mathbb{P}(X = b)$, i.e., $F(b) = F(b-) + \mathbb{P}(X = b)$, which is (1.33).

(c) From (1.33) and (1.34) for a fixed b, we have

$$F(b) - F(b-) = \mathbb{P}(X = b) = 0,$$

i.e., F is left-continuous at b, which is equivalent to F being continuous at b because F is always right-continuous. The re-stated definition reads: a continuous random variable is one whose cdf is an everywhere-continuous function.

Chapter 2

Poisson and Related Processes

2.1 Preliminaries

Function Order If a function f satisfies

$$\lim_{h\to 0} \frac{f(h)}{h} = 0,$$

then we write $f(h) = o(h)$. Examples:

- h^2: $\frac{h^2}{h} = h \to 0$, so h^2 *is* $o(h)$
- $\sqrt{h}$: $\frac{\sqrt{h}}{h} = \frac{1}{\sqrt{h}} \to \infty$, so $\sqrt{h}$ is *not* $o(h)$.

It is easy to check that:

1. $f(h) = o(h) \Rightarrow cf(h) = o(h)$ for any constant c.
2. $f(h) = o(h), g(h) = o(h) \Rightarrow f(h) + g(h) = o(h)$.

Later on, we encounter expressions such as $\lambda h + o(h)$, where $\lambda > 0$. These mean to say that the $o(h)$ term becomes negligible compared to λh in the limit as $h \to 0$ ($o(h)/h$ tends to zero, whereas $\lambda h/h = \lambda > 0$). In particular, the sign in front of $o(h)$ is irrelevant, as

$$\lim_{h\to 0} \frac{\lambda h + o(h)}{h} = \lim_{h} \frac{\lambda h - o(h)}{h} = \lim_{h} \frac{\lambda h}{h} = \lambda.$$

(Notation: the limit detail is dropped after first use.)

Expansion of the Exponential Function Expanding the exponential function around zero via Taylor's Theorem with remainder of order 2,

$$e^x = 1 + x + o(x). \tag{2.1}$$

The Poisson Distribution The discrete random variable N with support $\{0, 1, 2, \ldots\}$ is said to have the Poisson distribution with mean λ if

$$\mathbb{P}(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad n = 0, 1, 2, \ldots. \tag{2.2}$$

We write this in short $N \sim \text{Poisson}(\lambda)$. A calculation gives $\mathbb{E}[N] = \lambda$ and $\text{Var}(N) = \lambda$.

2.2 Residual Time and Memoryless Property

We are interested in the time X until a specified event happens, having in mind events such as customer arrivals and customer service completions. X is modelled as a positive-real-valued random variable with known distribution.

Suppose an observer knows that, as of s time units ago, the event has not occurred. Then, the time until the event occurs, called the *residual time*, is $X - s$. The observer is then interested in the conditional distribution of $X - s$ given $X > s$:

$$\mathbb{P}(X - s > t | X > s) \quad t \geq 0. \tag{2.3}$$

For $s = 5$, for example, this is the function

$$\mathbb{P}(X - 5 > t | X > 5), \quad t \geq 0.$$

In general the distribution changes with s, reflecting a "memory" mechanism. A key exception is described below.

Definition 2.1 *A positive-valued random variable X is said to be* memoryless *if it satisfies*

$$\mathbb{P}(X - s > t | X > s) = \mathbb{P}(X > t), \quad s, t \geq 0. \tag{2.4}$$

This says that the conditional distribution of $X - s$ given $X > s$ equals the unconditional distribution of X; we write this in short

$$(X - s | X > s) \stackrel{\mathcal{D}}{=} X. \tag{2.5}$$

Proposition 2.2 *X satisfies (2.4) $\Leftrightarrow$ X has an exponential distribution.*

Proof of the "$\Leftarrow$": Observe that

$$\mathbb{P}(X - s > t | X > s) = \mathbb{P}(X > s + t | X > s) = \frac{\mathbb{P}(X > s + t, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)};$$

now for $X \sim \text{Expon}(\lambda)$, i.e., satisfying (1.11), this equals

$$\frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t),$$

i.e. (2.4) holds. The proof of "$\Rightarrow$" is omitted.

Example 2.3 Suppose we are putting out (servicing) fires, and X is the time required to service a fire, in minutes. Suppose we try to predict when the service will finish given that service started s minutes ago, and our predictor is the mean remaining service time,

$$g(s) = \mathbb{E}[X - s | X > s].$$

First, we give a simulation method for estimating $g(s)$.

1. Fix s. Obtain random samples of $(X - s|X > s)$, as in (2.6) below, independently of others.

2. Calculate $\hat{g}(s)$ as the sample average. This is an estimate of $g(s)$.

How do we obtain random samples of $(X - s|X > s)$? The simplest method is:

Sample (randomly) X; if $X > s$, record $X - s$; otherwise, reject (no record is kept). (2.6)

The distribution of X can affect $g(s)$ a lot. Suppose $X \sim \text{N}(40, 100)$, a Normal distribution with mean 40 and standard deviation 10. 1 million trials of the form (2.6) gave the following estimates (which are accurate for our purpose):

s	$\hat{g}(s)$
20	20.5
30	12.9
40	8.0
50	5.2
60	3.7

In contrast, suppose $X \sim \text{Expon}(1/40)$, the Exponential distribution with the same mean as the Normal, 40. Then, from (2.5), $g(s) = \mathbb{E}[X] = 40$ for all s. This is very different: for example, for a service that began $s = 40$ minutes ago, the Normal gives a mean remaining service time of 8.0, much smaller than the Exponential's 40.

2.3 Counting Processes

How do we model events occurring "randomly" over time? Let t denote time, and let it range from 0 to infinity.

One approach focuses on the (cumulative) count of events over time:

$$N_t = \text{number of events that occur after time 0 and up to time } t, \quad t \geq 0.$$

Another approach focuses on the times of events:

$$S_n = \text{time of occurrence of } n\text{-th event}, \quad n = 1, 2, \ldots, \tag{2.7}$$

or, equivalently, the *inter-event times* (times between successive events)

$$\begin{aligned} X_1 &= S_1 = \text{time of 1st event} \\ X_2 &= S_2 - S_1 = \text{time between 1st and 2nd event} \\ \ldots &= \ldots \\ X_n &= S_n - S_{n-1} = \text{time between the } (n-1)\text{-st and } n\text{-th event}, \quad n = 1, 2, 3, \ldots \end{aligned}$$

(Put $S_0 = 0$ so that $X_n = S_n - S_{n-1}$ for $n = 0$ as well.) We make some natural assumptions:

1. $N_0 = 0$; that is, the counting starts at time zero.
2. N_t is integer-valued, and it is a nondecreasing function of t.
3. $X_n > 0$ for all n; that is, events occur one at a time–two events cannot happen at the same time.

As the counts $(N_t : t \geq 0)$ and the event times $(S_n,\ n = 0, 1, 2, \ldots)$ are alternative descriptions of the same (random) experiment, they are connected by

$$\{\text{"time of the } n\text{-th event"} > t\} \Leftrightarrow \{\text{"\# events in } (0, t]\text{"} < n\} \quad \text{for all } n \text{ and } t$$

i.e., we have the equality of events

$$\{S_n > t\} = \{N_t < n\} \tag{2.8}$$

and

$$\{N_t = n\} = \{S_n \leq t < S_{n+1}\} \tag{2.9}$$

for any $n = 0, 1, 2, \ldots$ and $t \geq 0$. These event equalities are fundamental to calculations later on.

2.4 The Poisson Process

We study a special counting process, the Poisson, that is analytically simple and commonly used.

2.4.1 Distribution of counts (the N_t process)

For $s, t \geq 0$, we have by definition

$$N_{s+t} - N_s = \#\text{ (number of) events that occur after } s \text{ and up to } s+t.$$

This is the *count* or *increment* on (the time interval) $(s, s+t]$.

The description is via the (probabilistic) behaviour of such counts, particularly on disjoint (non-overlapping) intervals.

Definition 2.4 *The counting process* $(N_t : t \geq 0)$ *is called a* Poisson process of rate λ *if:*

1. *(Independence) The counts on disjoint intervals are independent rv's.*
2. *(Identical distribution, or stationarity) The distribution of* $N_{s+t} - N_s$ *depends on the interval's length,* t*, and not the startpoint,* s*.*
3. *(Small-interval behaviour)*

$$\mathbb{P}(N_h = 1) = \lambda h + o(h), \quad \mathbb{P}(N_h \geq 2) = o(h), \quad \text{as } h \to 0. \tag{2.10}$$

Here is another definition:

Definition 2.5 *The counting process* $(N_t : t \geq 0)$ *is called a* Poisson process of rate λ *if:*

1. *(Independence) The counts on disjoint intervals are independent rv's.*
2. *(Poisson distribution) The count in any interval of length* t $(t \geq 0)$ *has the Poisson distribution with mean* λt*. That is, for any* $s, t \geq 0$*,*

$$\mathbb{P}(N_{s+t} - N_s = n) = \mathbb{P}(N_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \ldots \tag{2.11}$$

For $s, t \geq 0$, the intervals $(0, s]$ and $(s, s+t]$ are disjoint, so N_s and $N_{s+t} - N_s$ are independent.

Theorem 2.6 *Definition 2.4* $\Leftrightarrow$ *Definition 2.5.*

Proof. Proof of "⇒". It suffices to show that the functions

$$p_n(t) = \mathbb{P}(N_t = n), \quad t \geq 0, \quad n = 0, 1, 2, \ldots,$$

are as specified on the right of (2.11). We will show this by writing and solving differential equations linking the $p_n(t)$ and their derivatives, $(\mathrm{d}/\mathrm{d}t)p_n(t)$, across n. As $N_0 = 0$, these functions at time 0 are:

$$p_0(0) = 1, \quad p_n(0) = 0, \quad n = 1, 2, \ldots.$$

We put $p_{-1}(t) = 0$, all t; this function has no physical meaning and serves to simplify the notation.

Fix $h > 0$ and put $D_h = N_{t+h} - N_t$. Note $D_h \stackrel{\mathcal{D}}{=} N_h$, by the identical-distribution assumption. Fix n. Then

$$\{N_{t+h} = n\} = \{N_t = n, D_h = 0\} \text{ or } \{N_t = n-1, D_h = 1\} \text{ or } \{D_h \geq 2, N_t = n - D_h\} \tag{2.12}$$

where the "or"'s are between disjoint events (as the value of D_h is different across). Now note

$$\mathbb{P}(D_h \geq 2, N_t = n - D_h) \leq \mathbb{P}(D_h \geq 2) = o(h) \quad \text{by (2.10)}.$$

Taking probabilities in (2.12) and using the above gives

$$p_n(t+h) = \mathbb{P}(N_t = n, D_h = 0) + \mathbb{P}(N_t = n-1, D_h = 1) + o(h), \tag{2.13}$$

and the probabilities on the right can be written

$$\begin{aligned} \mathbb{P}(N_t = n, D_h = 0) &= \mathbb{P}(N_t = n)\mathbb{P}(D_h = 0) \quad \text{by independence} \\ &= p_n(t)[1 - \lambda h + o(h)] \quad \text{by (2.10)}, \end{aligned}$$

and, similarly,

$$\begin{aligned} \mathbb{P}(N_t = n-1, D_h = 1) &= \mathbb{P}(N_t = n-1)\mathbb{P}(D_h = 1) \quad \text{by independence} \\ &= p_{n-1}(t)[\lambda h + o(h)] \quad \text{by (2.10)}. \end{aligned}$$

Thus (2.13) gives (substitute the above, re-arrange, and divide by h):

$$\frac{p_n(t+h) - p_n(t)}{h} = -\left(\lambda + \frac{o(h)}{h}\right) p_n(t) + \left(\lambda + \frac{o(h)}{h}\right) p_{n-1}(t) + \frac{o(h)}{h}.$$

Now take limits as $h \downarrow 0$ (i.e., h decreases to zero); the $o(h)/h$ terms tend to zero, so

$$\frac{\mathrm{d}}{\mathrm{d}t} p_n(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \quad n = 0, 1, 2, \ldots. \tag{2.14}$$

To solve this set of differential equations, multiply both sides by $e^{\lambda t}$ and re-arrange to obtain:

$$\lambda e^{\lambda t} p_{n-1}(t) = e^{\lambda t} \frac{d}{\mathrm{d}t} p_n(t) + \lambda e^{\lambda t} p_n(t) = \frac{\mathrm{d}}{\mathrm{d}t} \left(e^{\lambda t} p_n(t) \right). \tag{2.15}$$

The above can be solved for $p_n(t)$ in the order $n = 0, 1, 2, \ldots$, as follows. For $n = 0$: $p_{-1}(t) = 0$, and analytical integration of (2.15) gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} \left(e^{\lambda t} p_0(t) \right) \Rightarrow e^{\lambda t} p_0(t) = c \Rightarrow p_0(t) = c e^{-\lambda t}, \quad t \geq 0$$

for some constant c, and the requirement $p_0(0) = 1$ gives $c = 1$. Now we use induction on n. Assume

$$p_{n-1}(t) = e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}.$$

Putting this into the left of (2.15),

$$\frac{\lambda^n t^{n-1}}{(n-1)!} = \frac{\mathrm{d}}{\mathrm{d}t} \left(e^{\lambda t} p_n(t) \right). \tag{2.16}$$

Integrating analytically,

$$e^{\lambda t} p_n(t) = \frac{\lambda^n t^n}{n!} + c$$

for some constant c, and the condition $p_n(0) = 0$ gives $c = 0$. This concludes the induction step and the proof.

Proof of "$\Leftarrow$": Using (2.11) and expressing the exponentials as in (2.1), we find

$$\begin{aligned}
\mathbb{P}(N_h = 0) &= e^{-\lambda h}, \\
\mathbb{P}(N_h = 1) &= e^{-\lambda h} \lambda h = \lambda h + o(h) \\
\mathbb{P}(N_h \geq 2) &= 1 - \mathbb{P}(N_h = 0) - \mathbb{P}(N_h = 1) = 1 - e^{-\lambda h} - e^{-\lambda h} \lambda h = o(h).
\end{aligned}$$

□

Example 2.7 (Calculating joint probabilities.) Let $(N_t : t \geq 0)$ be a Poisson process of rate λ. For times $0 = t_0 < t_1 < t_2 < \ldots < t_k$ and natural numbers $n_1 \leq n_2 \leq \ldots \leq n_k$, probabilities of the form

$$\mathbb{P}(N_{t_1} = n_1, N_{t_2} = n_2, \ldots, N_{t_k} = n_k)$$

are calculable via the independence and Poisson-distribution properties:

$$\begin{aligned}
&\mathbb{P}(N_{t_1} = n_1, N_{t_2} = n_2, \ldots, N_{t_k} = n_k) \\
&= \mathbb{P}(N_{t_1} - N_{t_0} = n_1, N_{t_2} - N_{t_1} = n_2 - n_1, \ldots, N_{t_k} - N_{t_{k-1}} = n_k - n_{k-1}) \\
&= \mathbb{P}(N_{t_1} - N_{t_0} = n_1) \mathbb{P}(N_{t_2} - N_{t_1} = n_2 - n_1) \cdots \mathbb{P}(N_{t_k} - N_{t_{k-1}} = n_k - n_{k-1}) \\
&\qquad \text{by independence} \\
&= \prod_{i=1}^{k} e^{-\lambda(t_i - t_{i-1})} \frac{[\lambda(t_i - t_{i-1})]^{n_i - n_{i-1}}}{(n_i - n_{i-1})!} \quad \text{by (2.11).}
\end{aligned}$$

2.4.2 Distribution of Times (the X_n and S_n)

Theorem 2.8 *The counting process, $N = (N_t : t \geq 0)$ is a Poisson process of rate $\lambda \Leftrightarrow$ The associated inter-event times, $X_1 = S_1, X_2 = S_2 - S_1, X_3 = S_3 - S_2, \ldots$ are independent exponentially distributed rv's of rate λ.*

Proof. (Partial proof.) Apply (2.8) for $n = 1$:

$$\{S_1 > t\} = \{N_t = 0\}. \tag{2.17}$$

Taking probabilities,

$$\mathbb{P}(S_1 > t) = \mathbb{P}(N_t = 0) = e^{-\lambda t}, \quad t \geq 0$$

where the last step holds by (2.11). That is, S_1 has the Expon(λ) distribution, as claimed. The remainder is a sketch of the remaining proof. Using the independence property of the N_t process, it can be shown that $S_2 - S_1$ is independent of S_1; moreover, using the stationarity property, it can be shown that $S_2 - S_1$ has the same distribution as S_1. Similarly, one can show that for any n, $S_n - S_{n-1}$ is independent of the corresponding differences for any smaller n, and moreover it has the same distribution as S_1. $\square$

Now we give the distribution of S_n by taking probabilities in (2.8):

$$\mathbb{P}(S_n > t) = \mathbb{P}(N_t < n) = \mathbb{P}(\cup_{k=0}^{n-1}\{N_t = k\}) = \sum_{k=0}^{n-1} \mathbb{P}(N_t = k) = \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad t \geq 0. \tag{2.18}$$

(Note in step 3 that the events $N_t = k$ are disjoint across k.) In the last step, we used the Poisson formula, (2.11). The above distribution is known as Gamma(n, λ) and also as Erlang-n.

2.4.3 Simulating a Poisson Process

To simulate a Poisson process, the result from Theorem 2.8 is applied. Specifically, we simulate the inter-event times by using the fact that for any $\lambda > 0$, a random sample of the Expon(λ) distribution can be obtained as $(-1/\lambda)\log(U)$, where log denotes natural logarithm and $U \sim \text{Unif}(0, 1)$ denotes a pseudo-random number. Note that $\log(U)$ is always negative, and the result is always positive, as it should.

Method 2.9 (How to sample a Poisson process on a finite interval.) Given $T > 0$ and $\lambda > 0$, the event times of a Poisson process of rate λ on the interval $[0, T]$ can be simulated based on the inter-event times, as follows. Set $S_0 = 0$, and for $n = 1, 2, \ldots$, set $S_n = S_{n-1} + X_n$, where X_n, the n-th inter-event time is sampled as $(-1/\lambda)\log(U_n)$, where $U_n \sim \text{Unif}(0, 1)$ is independent of all other U's; stop as soon as $S_n > T$.

2.4.4 Merging Poisson Processes

Suppose we merge the events of two independent Poisson processes of rates λ_1 and λ_2 (i.e., all the N_t are independent across the two; or, equivalently, all the X_n, or all the S_n, are independent across the two). Put

$$\begin{aligned} N_t^i &= \text{number of events of process } i \text{ that occur up to time } t,\ i = 1, 2 \\ N_t &= N_t^1 + N_t^2 = \text{number of events of the merged process that occur up to time } t \end{aligned}$$

Then the *merged* process $(N_t : t \geq 0)$ is Poisson of rate $\lambda_1 + \lambda_2$. This property generalises to merging any finite number of processes, by induction. That is:

Proposition 2.10 *The merging of the events of independent Poisson processes gives a Poisson process of rate equal to the sum of the rates of the merged processes.*

Sketch of proof (merge two processes for simplicity):

1. The merged process has independent counts on disjoint intervals; this comes from the independence of counts across the processes being merged and across disjoint time intervals for the same process.

2. Check that the merged process satisfies (2.10) with the rate claimed.

Application 2.11 Customer JD arrives to find q customers waiting in queue, at a system of k non-idling servers (servers do not idle if work is available), where customer service times are independent Expon(μ) rvs. Assuming first-come first-serve discipline, how long is JD's waiting time?

1. JD joins the queue in position $q + 1$ (due to first-come first-serve discipline). Thus, with time 0 being the time he joins, his waiting time equals the time of the $(q + 1)$-st (customer) departure (service completion).

2. While JD waits, each server is working (due to no-server-idling). The independent Expon(μ) service times mean that the departures (i.e., the counting process of departure events) at a particular server is a Poisson processes of rate μ (Theorem 2.8), and independent of the departures at other servers.

3. Departures from the system are simply the merging of departures at individual servers, so they form a Poisson process of rate k *times* μ (Proposition 2.10).

4. The distribution of the time of the $(q + 1)$-st event of a Poisson process was derived earlier; see (2.18). So JD's waiting time has the Gamma$(q + 1, k\mu)$ distribution, that is, tail probability function (2.18) with $n = q+1$ and $\lambda = k\mu$.

2.5 Poisson Process of General Rate Function

So far, the constant λ was the probabilistic rate of an event occurrence in any small time interval. The forthcoming model replaces this constant by a general (nonnegative) function of time, $\lambda(u)$, $u \geq 0$.

Definition 2.12 *The counting process $(N_t : t \geq 0)$ is called a Poisson process of (with) rate function $\lambda(t)$ if:*

1. *(Independence) The counts on disjoint intervals are independent rv's.*
2. *(Small-interval behaviour)*

$$\mathbb{P}(N_{s+h} - N_s = 1) = \lambda(s)h + o(h), \quad \mathbb{P}(N_{s+h} - N_s \geq 2) = o(h), \quad \text{as } h \to 0. \tag{2.19}$$

The term *Non-Homogeneous* (or *time-inhomogeneous*) Poisson Process (NHPP) indicates the rate function is non-constant.

In this case the distribution of N_t is again Poisson, but its mean is now the *mean function*

$$m(t) = \int_0^t \lambda(u)\mathrm{d}u, \tag{2.20}$$

that is, the area under the graph of the rate function from 0 to t. By solving differential equations like those in the proof of Theorem 2.6, we obtain the summary result below:

Proposition 2.13 *In a Poisson process of rate function $\lambda(\cdot)$, the count $N_e - N_s$ has the Poisson distribution with mean $\int_s^e \lambda(u)\mathrm{d}u = m(e) - m(s)$. That is,*

$$\mathbb{P}(N_e - N_s = n) = e^{-[m(e)-m(s)]}\frac{[m(e) - m(s)]^n}{n!}, \quad n = 0, 1, \ldots \tag{2.21}$$

Proof. Left as Exercise 3. □

Example: Exam E03 3. Recognise it is an NHPP and identify the rate function; the solution is then standard, exactly the main result above for $s = 0$.

2.5.1 Thinning a Poisson Process

Suppose:

1. $N = (N_t : t \geq 0)$ is a Poisson process of rate K.
2. Conditional on any event of N occurring at time t, accept the event with probability $p(t)$ and reject it otherwise, doing so independently of anything else. (The acceptance step is simulated via pseudo-random numbers later.)

Let $\tilde{N}_t$ be the number of accepted events occurring up to time t inclusive. Call $(\tilde{N}_t : t \geq 0)$ the *thinned process*.

Proposition 2.14 *(Thinning a Poisson process.) The process* $(\tilde{N}_t : t \geq 0)$ *is a Poisson process of rate function* $Kp(t)$.

Sketch of proof (assume the function $p(\cdot)$ is continuous, for simplicity):

1. The thinned process has independent counts on disjoint intervals; this comes from the independence of counts of the original process together with the independence of the acceptance/rejection random variables from everything else.

2. Check that the thinned process satisfies (2.19) with the rate claimed.

2.5.2 Simulating an NHPP via Thinning

Problem 2.15 Simulate (or sample) on a given time interval $[0, T]$ a non-homogeneous Poisson process of given rate function $\lambda(t)$, $0 \leq t \leq T$.

If the rate function is piece-wise constant, a simple solution is: simulate a constant-rate process on each interval where the rate is constant, via Method 2.9.

If the rate function is not piecewise constant (examples include: linear with a nonzero slope, non-constant polynomial, trigonometric, exponential), then a simple solution is as follows.

Method 2.16 (The Thinning Method.) To simulate the event times of a Poisson process of rate function $\lambda(t)$ on the interval $[0, T]$, do:

1. Calculate the rate
$$K = \max_{0 \leq t \leq T} \lambda(t) < \infty. \tag{2.22}$$

2. Using Method 2.9, simulate the event times of a Poisson process of constant rate K on the given interval. That is, set $S_0 = 0$, and for $n = 1, 2, \ldots$, set $S_n = S_{n-1} + (-1/K)\log(U_n)$, where $U_n \sim \text{Unif}(0, 1)$ is independent of other U's, stopping as soon as $S_n > T$. The set $S_1, S_2, \ldots, S_{n-1}$ is a random sample of the event times of a Poisson process of rate K on the given interval.

3. (Thinning step.) For $i = 1, 2, \ldots, n-1$, accept S_i with probability $\lambda(S_i)/K$ (reject with the remaining probability; note that $\lambda(S_i)/K$ is always between 0 and 1, by the choice of K). The acceptance/rejection is done by randomly sampling $V_i \sim \text{Unif}(0, 1)$ independently of everything else and accepting if $V_i \leq$

$\lambda(S_i)/K$. The set of accepted S's is a random sample of the event times of a Poisson process of rate function $\lambda(t)$ on the given interval.

2.6 Estimating a Poisson Process: Brief View

Suppose we want to estimate (construct) the functions $\lambda(t)$ and $m(t)$ from given event times $s_1 < s_2 < \ldots < s_n$. Note:

- $m()$ can always be found by integrating $\lambda()$ as in (2.20).
- Obtaining $\lambda(t)$ from $m()$ is possible, assuming m is differentiable at t:

$$\lambda(t) = \frac{\mathrm{d}}{\mathrm{d}t} m(t).$$

- A constant rate λ is equivalent to a linear mean function, $m(t) = \lambda t$.

As a start, consider:

Step-function estimate of $m(t)$.

$$\hat{m}(t) = \text{observed count up to } t = \text{number of } s\text{'s that are} \leq t. \tag{2.23}$$

That is,

$$\hat{m}(s_k) = k, \quad k = 1, 2, \ldots n,$$

and it is constant between s_{k-1} and s_k. Unattractively, it is impossible to infer $\lambda()$ by differentiation [1].

One way to get a reasonable estimate of $\lambda()$ is by a slight revision of the above:

Piece-wise linear estimate of $m(t)$ (linear in-between event times). Define the function $\tilde{m}(t)$ to be as above at the event times, i.e.,

$$\tilde{m}(s_k) = k, \quad k = 1, 2, \ldots n,$$

and estimate $\tilde{\lambda}$ as the slope of $\tilde{m}$:

$$\tilde{\lambda}(t) = \frac{\tilde{m}(s_k) - \tilde{m}(s_{k-1})}{s_k - s_{k-1}} = \frac{1}{s_k - s_{k-1}} \quad \text{for } s_{k-1} \leq t < s_k. \tag{2.24}$$

If the $\tilde{\lambda}(t)$ are close between adjacent $(s_{k-1}, s_k]$ intervals, we could average them, loosing little information in the averaging, to simplify the estimate.

[1] $\hat{m}(t)$ has derivative zero at all points other than the s's; at the s's it is not differentiable.

2.7 Large-t $\frac{N_t}{t}$ via Strong Law of Large Numbers

Let N_t be the count of events up to time t. Assume

A1. Inter-event times are independent, identically distributed, with mean $\mu > 0$.

A2. $\lim_{t\to\infty} N_t = \infty$.

What is the average number of events per unit time, N_t/t, for t large ?
For any t,

$$S_{N_t} \le t < S_{N_t+1}.$$

Dividing by N_t,

$$\frac{S_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{S_{N_t+1}}{N_t+1}\frac{N_t+1}{N_t}. \tag{2.25}$$

By the SLLN (Theorem 1.22), both the right and left side of (2.25) converge to μ as $t \to \infty$, with probability one. Hence the middle must converge to the same, with probability one (the event of convergence of the left and right, call it A, implies the event of convergence of the middle, call it B; thus $1 = \mathbb{P}(A) \le \mathbb{P}(B)$, proving $\mathbb{P}(B) = 1$). We conclude

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu} \quad \text{w.p. 1.} \tag{2.26}$$

2.8 Exercises

1. Let $N = (N_t : t \geq 0)$ be a Poisson process of rate $\lambda = 0.4$.

   i Calculate: (a) $\mathbb{P}(N_5 = 3)$; (b) $\mathbb{P}(N_5 = 3, N_{15} = 7)$; (c) $\mathbb{P}(N_{15} = 7|N_5 = 3)$; (d) $\mathbb{P}(N_5 = 3|N_{15} = 7)$.

   ii Let X be the time between two successive events of this process. Specify the distribution of X fully, including its mean. Write $\mathbb{P}(X \leq 0.5)$ and $\mathbb{P}(X \leq 2)$ explicitly.

   iii State, briefly, how the results in (i) change if the process is Poisson of mean function $m()$ with $m(0) = 0$ as usual.

2. In Proposition 2.10, two independent Poisson processes of rates λ_1 and λ_2 are merged; N_t^i is the count of process i up to time t; and $N_t = N_t^1 + N_t^2$ is the count of the merged process. Show that as $h \to 0$, we have $\mathbb{P}(N_h = 2) = o(h)$ and $\mathbb{P}(N_h = 1) = (\lambda_1+\lambda_2)h+o(h)$. *Note: similar, slightly more complex calculations are seen later in (3.1) to (3.3).*

3. (NHPP)

   (a) (Distribution of N_t.) For the (NHPP) N_t in Definition 2.12, write $p_n(t) = \mathbb{P}(N_t = n)$, $n = 0, 1, 2, \ldots$ (put $p_{-1}(t) = 0$ for later convenience), and show that these functions must satisfy the set of differential equations

   $$\frac{\mathrm{d}}{\mathrm{d}t}p_n(t) = -\lambda(t)p_n(t) + \lambda(t)p_{n-1}(t), \quad n = 0, 1, 2, \ldots. \tag{2.27}$$

   Then verify that the functions $p_n(t) = e^{-m(t)}[m(t)]^n/n!$, $n = 0, 1, 2, \ldots$ satisfy (2.27). Note that $(\mathrm{d}/\mathrm{d}t)m(t) = \lambda(t)$. Note $p_0(0) = 1$ and $p_n(0) = 0$ for $n > 0$, effectively saying that $N_0 = 0$.

   (b) (Distribution of S_n.) Working as for the constant-rate process, show that the distribution of S_n for the general case above is

   $$\mathbb{P}(S_n > t) = \sum_{k=0}^{n-1} e^{-m(t)} \frac{[m(t)]^k}{k!}, \quad t \geq 0$$

   so the distribution of the time of the 1st event, S_1, is

   $$\mathbb{P}(S_1 > t) = e^{-m(t)}, \quad t \geq 0. \tag{2.28}$$

4. (Simulating Poisson processes.)

   (a) Simulate the event times of a Poisson process of rate $\lambda = 2$ during $[0, T = 2]$ based on assumed pseudo-random numbers $\{.8187, .2466, .5488, .3679, .4066\}$.

   (b) Simulate the event times of the NHPP of rate function $\lambda(t) = 1 + \sin(\pi t)$, $0 \leq t \leq 2$, where $\pi \doteq 3.14159$, by thinning the process sampled previously. For the acceptance test, assume pseudo-random numbers 0.7, 0.4, 0.2, 0.6, adding your own if needed.

   (c) It can be shown that a random variable with cdf $F(x)$ can be simulated by solving for x the equation $F(x) = U$, where $U \sim \mathrm{Unif}(0, 1)$ (a pseudo-random number). Using this property together with (2.28), show how the time of the 1st event of the NHPP in (b) may be simulated. You may use that $\int_0^t \sin(\pi u)\mathrm{d}u = (1/\pi)[1 - \cos(\pi t)]$.

.

2.9 Solutions to Exercises

1. For $s \le e$, $N_e - N_s$ has the Poisson distribution with mean $\lambda(e-s)$ the rate λ times the length of the interval, $e-s$. Recall the Poisson distribution is given in (2.2).

   i (a) Recall the assumption $N_0 = 0$, so $N_5 = N_5 - N_0$; here $e - s = 5$, so $N_5 \sim$ Poisson(2) and thus $\mathbb{P}(N_5 = 3) = e^{-2}2^3/3!$.

   (b) In calculating this joint probability, note that N_5 and N_{15} refer to overlapping time intervals, so they are *not* independent. The "trick" is to write $N_{15} = N_5 + N_{15} - N_5$ and use that $N_{15} - N_5$ is independent of N_5 and Poisson(4)-distributed ($e - s = 15 - 5 = 10$, times rate, 0.4).Thus

$$\begin{aligned}\mathbb{P}(N_5 = 3, N_{15} = 7) &= \mathbb{P}(N_5 = 3, N_{15} - N_5 = 7 - 3) \\ &= \mathbb{P}(N_5 = 3)\mathbb{P}(N_{15} - N_5 = 4) \\ &= e^{-2}\frac{2^3}{3!}e^{-4}\frac{4^4}{4!}.\end{aligned}$$

   (c)

$$\begin{aligned}\mathbb{P}(N_{15} = 7 | N_5 = 3) &= \frac{\mathbb{P}(N_{15} = 7, N_5 = 3)}{\mathbb{P}(N_5 = 3)} \\ &= \mathbb{P}(N_{15} - N_5 = 4) \quad \text{by (b), the } \mathbb{P}(N_5 = 3) \text{ cancels out} \\ &= e^{-4}\frac{4^4}{4!}.\end{aligned}$$

   (d)

$$\mathbb{P}(N_5 = 3 | N_{15} = 7) = \frac{\mathbb{P}(N_5 = 3, N_{15} = 7)}{\mathbb{P}(N_{15} = 7)} = \frac{e^{-2}\frac{2^3}{3!}e^{-4}\frac{4^4}{4!}}{e^{-6}\frac{6^7}{7!}}.$$

   ii We know $X \sim$ Expon(0.4), whose mean is $1/0.4$ (seen elsewhere). Then

$$\begin{aligned}\mathbb{P}(X \le \frac{1}{2}) &= 1 - e^{-0.4 \cdot \frac{1}{2}} = 1 - e^{-0.2}. \\ \mathbb{P}(X \le 2) &= 1 - e^{-0.4 \cdot 2} = 1 - e^{-0.8}.\end{aligned}$$

   iii In the Poisson process with mean function $m()$, $N_e - N_s$ has the Poisson distribution with mean $m(e) - m(s)$, and such counts on disjoint intervals are independent, as with the constant-rate process. Thus, we only modify the Poisson means; the mean of N_5 is $m(5) - m(0) = m(5)$ and the mean of $N_{15} - N_5$ is $m(15) - m(5)$.

2. The key idea is to express the events of interest as (note superscripts are not powers):

$$\begin{aligned}\{N_h = 2\} &= \{N_h^1 = 2, N_h^2 = 0\} \cup \{N_h^1 = 1, N_h^2 = 1\} \cup \{N_h^1 = 0, N_h^2 = 2\} \\ \{N_h = 1\} &= \{N_h^1 = 1, N_h^2 = 0\} \cup \{N_h^1 = 0, N_h^2 = 1\} \qquad (2.29)\end{aligned}$$

where the events on the right are disjoint. Taking probabilities,

$$\begin{aligned}\mathbb{P}(N_h = 2) &= \mathbb{P}(N_h^1 = 2, N_h^2 = 0) + \mathbb{P}(N_h^1 = 1, N_h^2 = 1) + \mathbb{P}(N_h^1 = 0, N_h^2 = 2) \\ &= \mathbb{P}(N_h^1 = 2)\mathbb{P}(N_h^2 = 0) + \mathbb{P}(N_h^2 = 1)\mathbb{P}(N_h^2 = 1) + \mathbb{P}(N_h^1 = 0)\mathbb{P}(N_h^2 = 2) \\ &\qquad \text{by independence.}\end{aligned}$$

The first term is $\leq \mathbb{P}(N_h^1 = 2) = o(h)$; likewise, the third term is $\leq \mathbb{P}(N_h^2 = 2) = o(h)$. The middle term is

$$[\lambda_1 h + o(h)][\lambda_2 h + o(h)] = o(h)$$

(terms $\lambda_i h o(h)$ and $\lambda_1\lambda_2 h^2$ are $o(h)$), as required. Similarly,

$$\begin{aligned}\mathbb{P}(N_h = 1) &= \mathbb{P}(N_h^1 = 1, N_h^2 = 0) + \mathbb{P}(N_h^1 = 0, N_h^2 = 1) \\ &= \mathbb{P}(N_h^1 = 1)\mathbb{P}(N_h^2 = 0) + \mathbb{P}(N_h^1 = 0)\mathbb{P}(N_h^2 = 1) \quad \text{by independence} \\ &= [\lambda_1 h + o(h)][1 - \lambda_2 h + o(h)] + [1 - \lambda_1 h + o(h)][\lambda_2 h + o(h)] \\ &= \lambda_1 h + \lambda_2 h + o(h)\end{aligned}$$

where, again, various original terms are combined into the $o(h)$.

3(a) The count from t to $t + h$, denoted $D_{t,h} = N_{t+h} - N_t$, is independent of N_t by assumption. The essential difference relative to the homogeneous case is that the distribution of $D_{t,h}$ depends on both t and h, whereas that of D_h there depended only on h. The derivation mimics closely the steps from (2.12) to (2.14), "correcting" for the difference above.

$\{N_{t+h} = n\} = \{N_t = n, D_{t,h} = 0\}$ or $\{N_t = n-1, D_{t,h} = 1\}$ or $\{D_{t,h} \geq 2, N_t = n - D_{t,h}\}$

where the events on the right are disjoint, so the left probability is the sum of the events' probabilities on the right. Work these out:

$$\mathbb{P}(D_{t,h} \geq 2, N_t = n - D_{t,h}) \leq \mathbb{P}(D_{t,h} \geq 2) = o(h) \quad \text{by (2.19)},$$

$$\mathbb{P}(N_t = n, D_{t,h} = 0) = p_n(t)[1 - \lambda(t)h + o(h)] \quad \text{by independence and (2.19)},$$

and

$$\mathbb{P}(N_t = n - 1, D_{t,h} = 1) = p_{n-1}(t)[\lambda(t)h + o(h)] \quad \text{by independence and (2.19)}.$$

Putting these together,

$$p_n(t + h) = p_n(t)[1 - \lambda(t)h + o(h)] + p_{n-1}(t)[\lambda(t)h + o(h)] + o(h).$$

Exactly as in the constant-rate case (move $p_n(t)$ from right to left, divide by h, take limits as $h \to 0$), we get (2.27). To verify, differentiate the given functions $p_n(t) = e^{-m(t)}[m(t)]^n/n!$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-m(t)}\frac{[m(t)]^n}{n!}\right) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}e^{-m(t)}\right)\frac{[m(t)]^n}{n!} + e^{-m(t)}\frac{\mathrm{d}}{\mathrm{d}t}\frac{[m(t)]^n}{n!} \\ &= -e^{-m(t)}\left(\frac{\mathrm{d}}{\mathrm{d}t}m(t)\right)\frac{[m(t)]^n}{n!} + e^{-m(t)}\frac{n[m(t)]^{n-1}}{n!}\frac{\mathrm{d}}{\mathrm{d}t}m(t) \\ &= -\lambda(t)p_n(t) + \lambda(t)p_{n-1}(t),\end{aligned}$$

as required.

(b)

$$\mathbb{P}(S_n > t) = \mathbb{P}(N_t < n) = \sum_{k=0}^{n-1} \mathbb{P}(N_t = k) = \sum_{k=0}^{n-1} e^{-m(t)} \frac{[m(t)]^k}{k!}.$$

4(a) The event times are simulated iteratively as $S_0 = 0$, $S_i = S_{i-1} + X_i$ for $i = 1, 2, \ldots$, stopping as soon as $S_i > T = 2$, where the inter-event times X_i are simulated by the formula $X_i = (-1/\lambda)\log(U_i)$, the U_i being Unif$(0, 1)$ random numbers (Method 2.16). We obtain:

i	U_i	$X_i = -(1/\lambda)\log(U_i)$	$S_i = S_{i-1} + X_i$
1	.8187	0.1	0.1
2	.2466	0.7	0.8
3	.5488	0.3	1.1
4	.3679	0.5	1.6
5	.4066	0.45	2.05

(b) Following Method 2.16, compute the maximum: $K = \max_{0 \le t \le 2}[1 + \sin(\pi t)] = 2$. The event times in (a) may be used as the times to be thinned because they were sampled with the appropriate rate, 2. Denoting by V_i the given pseudo-random numbers, calculate:

S_i	$\lambda(S_i)/K$	V_i	Is $V_i \le \lambda(S_i)/K$?
0.1	0.65	0.7	No
0.8	0.79	0.4	Yes
1.1	0.34	0.2	Yes
1.6	0.02	0.6	No

Thus, the simulated process has events at times 0.8 and 1.1 only.

(c) The mean function is

$$m(t) = \int_0^t (1 + \sin(\pi u))\mathrm{d}u = t + (1/\pi)[1 - \cos(\pi t)].$$

Aim for the point at which the cdf of S_1 equals U:

$$F(t) = 1 - e^{-m(t)} = U \Leftrightarrow m(t) = -\log(1 - U).$$

The explicit point t is not pursued here.

.

Chapter 3

Queues

3.1 Preliminaries

Vectors are column vectors. The transpose of a vector $\mathbf{p}$ is denoted $\mathbf{p}^T$.

We often take limits as the length h of a time interval goes to zero; we write this as "$\lim_{h\to 0}$", or even as "lim" when the condition is obvious.

A probability distribution with support a set E is called in short a *distribution on* E.

3.1.1 Queues: Terminology and Global Assumptions

Jobs (customers) arrive at a system at which a number of servers are stationed. A job may have to wait in a *waiting area* (queue) until a server becomes available. After being served in a *service area*, jobs leave. The *system* includes both these areas.

1. The times between arrivals are independent identically distributed (iid) random variables.

2. The service times are random variables and independent of the inter-arrival times.

3. *No idling.* Server(s) will not idle if jobs are waiting.

4. $M/M/c/k$ means **M**emoryless inter-arrival times, **M**emoryless service times, c Servers, and a maximum of k jobs waiting in queue, where $k = \infty$ if not specified. "Memoryless" means that these times are exponentially distributed.

5. The inter-arrival times and the service times have means that are positive and finite.

3.2 The Birth-Death Process

Denote X_t the number of jobs in the system at time t, also called the *state*. We are interested in the (stochastic) process $X = (X_t : t \geq 0)$. If there is a maximum allowed number in the system, say k, then X_t takes values in the finite set $E = \{0, 1, 2, \ldots, k\}$, 0 being the lowest state, reflecting an empty system. If there is no such maximum, then X_t takes values in the infinite set $E = \{0, 1, 2, \ldots\}$. A birth-death process arises as follows:

1. Whenever $X_t = i$, events whose effect is to increase X by one ("births") occur according to a Poisson process $B = (B_t)$ of (birth) rate λ_i (the time T_B until the next birth is an Expon(λ_i) random variable).

2. In addition, again given $X_t = i$, events whose effect is to decrease X by one ("deaths") occur according to a Poisson process $D = (D_t)$ of (death) rate μ_i (the time T_D until the next death is an Expon(μ_i) random variable).

3. These Poisson processes are independent.

The process X moves as follows. The next state of X, as well as the timing of the state change, are the result of the competition between the birth and death processes. This means that: (a) Starting from state i, the time when the process X first changes state, $T = \min(T_B, T_D)$, has an Expon($\lambda_i + \mu_i$) distribution (as seen in merging); and (b) The next state is $i + 1$ with probability $\lambda_i/(\lambda_i + \mu_i)$ (a birth happens before a death), or $i - 1$ with the remaining probability (a death happens before a birth).

For each i, we now derive conditional (transition) probabilities of the future state of the process given $X_t = i$, h time units in the future. Put:

- B = number of births (occurring) in $(t, t + h]$, and D = number of deaths.
- $N := B + D$ is the number of events (births plus deaths) in $(t, t + h]$.

Then,

$$\begin{aligned}
\mathbb{P}(X_{t+h} = i + 1 | X_t = i) &= \mathbb{P}(B - D = 1 | X_t = i) \\
&= \mathbb{P}(B - D = 1, N = 1 | X_t = i) + \underbrace{\mathbb{P}(B - D = 1, N \geq 2 | X_t = i)}_{o(h)} \\
&= \mathbb{P}(B = 1, D = 0 | X_t = i) + o(h) \\
&= \mathbb{P}(B = 1 | X_t = i)\mathbb{P}(D = 0 | X_t = i) + o(h) \\
&= [\lambda_i h + o(h)][1 - \mu_i h + o(h)] + o(h) = \lambda_i h + o(h), \qquad (3.1)
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{P}(X_{t+h} = i-1 | X_t = i) \\
&\quad = \mathbb{P}(B - D = -1 | X_t = i) \\
&\quad = \mathbb{P}(B - D = -1, N = 1 | X_t = i) + \underbrace{\mathbb{P}(B - D = -1, N \geq 2 | X_t = i)}_{o(h)} \\
&\quad = \mathbb{P}(B = 0, D = 1 | X_t = i) + o(h) \\
&\quad = \mathbb{P}(B = 0 | X_t = i)\mathbb{P}(D = 1 | X_t = i) + o(h) \\
&\quad = [1 - \lambda_i h + o(h)][\mu_i h + o(h)] + o(h) = \mu_i h + o(h),
\end{aligned} \tag{3.2}
$$

and X changes by 2 or more with probability that is negligible as $h \to 0$:

$$
\mathbb{P}(X_{t+h} = j | X_t = i) \leq \mathbb{P}(N \geq 2 | X_t = i) = o(h) \quad \text{for } |j - i| \geq 2. \tag{3.3}
$$

Consequently, X remains at the same state with the probability that remains, i.e.,

$$
\mathbb{P}(X_{t+h} = i | X_t = i) = 1 - (\lambda_i + \mu_i)h + o(h).
$$

The distribution of X_t can be expressed via differential equations obtained similarly to those for the Poisson process, (2.14). In these equations, to be seen later for a more general process, the following limits appear:

$$
q_{i,j} := \lim_{h \to 0} \frac{\mathbb{P}(X_{t+h} = j | X_t = i)}{h} = \left\{ \begin{array}{ll} \lambda_i & j = i+1 \\ \mu_i & j = i-1 \\ 0 & \text{otherwise } (j \notin \{i-1, i+1\}) \end{array} \right\}, j \neq i. \tag{3.4}
$$

The $q_{i,j}$, sometimes called *transition rates* or *probability rates*, are probabilities of transitions (state changes) *relative to time*. Note that if a $i \to j$ transition requires at least 2 events in order to happen, then $q_{i,j} = 0$. On the other hand, if $i \to j$ is caused by a single event, then $q_{i,j}$ is the rate of the underlying process.

The transition rates are sometimes summarised in a *transition diagram*, as in Figure 3.1.

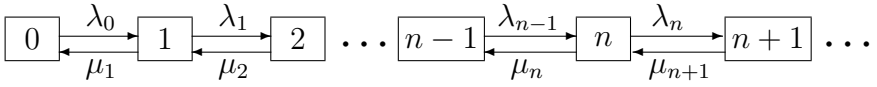

Figure 3.1: Transition diagram of a birth-death process.

3.3 Continuous-Time Markov Chains (CTMCs)

Definition 3.1 *The stochastic process* $(X_t : t \geq 0)$ *is called a continuous-time Markov chain (CTMC) with state space (set of possible states)* E *if for all* $i, j, i_1, \ldots, i_k \in E$*, all* $t \geq 0$*, and all* $p_1, p_2, \ldots p_k \leq s$*,*

$$\mathbb{P}(X_{s+t} = j | X_{p_1} = i_1, \ldots, X_{p_k} = i_k, X_s = i) = \mathbb{P}(X_{s+t} = j | X_s = i). \tag{3.5}$$

If the right side of (3.5) does not depend on s*, then the CTMC is called* homogeneous*.*

In words, given the process' history up to time s, the conditional distribution of X at future time points is the same as the conditional distribution given only its present value, X_s. Yet in other words, we can say that the X process is *memoryless*, because its future behaviour depends on its history only through the present; the strict past is irrelevant. We are mainly interested in homogeneous CTMCs, where

$$p_{i,j}(t) := \mathbb{P}(X_{s+t} = j | X_s = i)$$

is a function of i, j and t.

3.3.1 Generator

We assume there exist $q_{i,j}$ such that

$$q_{i,j} := \lim_{h \to 0} \frac{p_{i,j}(h)}{h}, \quad i, j \in E, i \neq j.$$

That is, $q_{i,j}$ is a "probability rate" of going from i to j. The birth-death process seen earlier is a special case of a homogeneous CTMC with $q_{i,j}$ being the λ_i for $j = i + 1$, the μ_i for $j = i - 1$, and 0 otherwise.

We also define

$$\begin{aligned} q_i &= \lim_{h \to 0} \frac{\mathbb{P}(X_{t+h} \neq X_t | X_t = i)}{h} = \lim_{h \to 0} \frac{1 - p_{i,i}(h)}{h} \\ &= \lim_{h \to 0} \frac{\sum_{j \in E, j \neq i} p_{i,j}(h)}{h} = \sum_{j \in E, j \neq i} \left(\lim_{h \to 0} \frac{p_{i,j}(h)}{h} \right) = \sum_{j \in E, j \neq i} q_{i,j} \end{aligned}$$

where we will always assume the interchange between limit and summation is valid (it could fail, but only if E is infinite and under further unusual conditions). (We used that $\sum_{j \in E} p_{i,j}(t) = 1$ for all i and t, a consequence of the fact that X_t must take a value in E.) That is, q_i is a "probability rate" of leaving i (to go anywhere else). We summarise these in the matrix

$$\mathbf{Q} = [q_{i,j}]_{i,j,\in E}$$

where we put $q_{i,i} = -q_i$. $\mathbf{Q}$ is called the *(CTMC) generator* and is central to later calculations. It can be calculated without taking limits, after a little experience.

Example 3.2 A repair shop has two workers named fast (F) and slow (S). Jobs arrive according to a Poisson process of rate 4. Service times are exponentially distributed with mean 1/3 and 1/2 at F and S respectively. Whenever both F and S are available to process a job, preference is given to F. Arrivals that find a job waiting are lost.

Focus on the system state over time, with the following possible states:

Numeric Code	State Description
1	empty
2	F working, S idle
3	S working, F idle
4	both servers working, 0 jobs in queue
5	both servers working, 1 job in queue

A detailed derivation of the generator would resemble that done for the birth-death process in (3.1) to (3.3). Here, a $5 \to 4$ transition is caused by a departure at F or S, i.e., an event of the process that merges these two event types, whose rate is $3 + 2 = 5$ (merging result). Thus, similar to (3.2), $p_{5,4}(h) = 5h + o(h)$ as $h \to 0$, so $q_{5,4} = \lim_{h\to 0} p_{5,4}(h)/h = 5$. Consider now $q_{5,3}$. Similar to (3.3), $p_{5,3}(h) = o(h)$ because $5 \to 3$ requires at least two departures, so $q_{5,3} = \lim_{h\to 0} p_{5,3}(h)/h = 0$. And so on for other i, j.

Generally, a systematic calculation of the generator can go as follows. For each type of event (Poisson process), write the corresponding rate, and list all transitions caused by a single such event in the form "from i to j", for specified i and j. Here this gives:

Causing Event	Rate	List of State Changes
Arrival	4	from 1 to 2, from 2 to 4, from 3 to 4, from 4 to 5
Departure at F	3	from 2 to 1, from 4 to 3, from 5 to 4
Departure at S	2	from 3 to 1, from 4 to 2, from 5 to 4

Then, for each i and $j \neq i$, locate all $i \to j$ transitions listed and find $q_{i,j}$ as the sum of the corresponding rates. Here this gives (empty matrix entries indicate zeros):

$$\mathbf{Q} = \begin{bmatrix} -4 & 4 & & & \\ 3 & -7 & & 4 & \\ 2 & & -6 & 4 & \\ & 2 & 3 & -9 & 4 \\ & & & 3+2 & -5 \end{bmatrix}$$

3.3.2 Time-dependent Distribution and Kolmogorov's Differential System

Assume the process state is known at time 0, e.g., $X_0 = 5$. Put $p_i(t) = \mathbb{P}(X_t = i)$. We focus on the vector $\mathbf{p}(t) = (p_i(t))_{i \in E}$, which is the distribution at time t of X_t. The whole distribution is related to that at previous times via

$$\begin{aligned} p_i(t+h) &= \mathbb{P}(X_{t+h} = i) \\ &= \sum_{k \in E} \mathbb{P}(X_{t+h} = i, X_t = k) \\ &= \sum_{k \in E} \mathbb{P}(X_t = k)\mathbb{P}(X_{t+h} = i | X_t = k) \\ &= \sum_{k \in E} p_k(t) p_{k,i}(h) \\ &= p_i(t)p_{i,i}(h) + \sum_{k \in E, k \neq i} p_k(t)p_{k,i}(h) \quad t, h \geq 0 \end{aligned}$$

(Theorem 1.10, Law of Total Probability). Subtract $p_i(t)$ from both left and right, and divide by h:

$$\frac{p_i(t+h) - p_i(t)}{h} = -p_i(t)\frac{1 - p_{i,i}(h)}{h} + \sum_{k \in E, k \neq i} p_k(t)\frac{p_{k,i}(h)}{h}.$$

Take limits as $h \to 0$. In the rightmost term, we assume the limit can pass inside the sum (which is the case if E is finite), i.e.,

$$\lim_{h \to 0} \sum_{k \in E, k \neq i} p_k(t)\frac{p_{k,i}(h)}{h} = \sum_{k \in E, k \neq i} p_k(t) \left(\lim_{h \to 0} \frac{p_{k,i}(h)}{h} \right) = \sum_{k \in E, k \neq i} p_k(t) q_{k,i}$$

and arrive at *Kolmogorov's differential system*:

$$p_i'(t) = -p_i(t)q_i + \sum_{k \in E, k \neq i} p_k(t)q_{k,i}, \quad i \in E. \tag{3.6}$$

where $p_i'(t) = \frac{\mathrm{d}}{\mathrm{d}t} p_i(t)$.

Supposing for example $X_0 = 5$, we would have the initial condition $p_5(0) = 1$ and $p_j(0) = 0$ for $j \neq 5$, and $\mathbf{p}(t)$ should satisfy (3.6) together with the initial condition.

3.4 Long-Run Behaviour

With X_t being a count of jobs in the system at time t, or something similar (repair-shop example), and all events causing changes to X "being" independent Poisson processes (inter-event times having exponential distributions), $(X_t : t \geq 0)$ will be a CTMC.

The theory of (Homogeneous) CTMCs is rich, especially for time going to ∞. It is built upon a corresponding theory for discrete-time Markov chains (t integer). The theory is technical and will not be seen in depth. A key summary result, Fact 3.3 below, is our basic supporting theory.

$1_{\{A\}}$ is the indicator function, valued 1 on the set (event) A and 0 elsewhere. Thus, for example, $1_{\{X_u=i\}}$ indicates if the process X at time u is at state i. Put

$$T_t^i = \text{time from 0 to } t \text{ that } X \text{ is in state } i = \int_0^t 1_{\{X_u=i\}} \mathrm{d}u. \tag{3.7}$$

Fact 3.3 *The CTMC $(X_t : t \geq 0)$ corresponding to "interesting" and "stable" queues converges, regardless of the initial state X_0, as follows.*

(a)

$$\lim_{t\to\infty} p_i(t) = \pi_i, \quad \lim_{t\to\infty} p_i'(t) = 0, \quad i \in E. \tag{3.8}$$

(b) Putting (3.8) into (3.6), we have the balance equations

$$\pi_i q_i = \sum_{k\in E: k\neq i} \pi_k q_{k,i}, \quad i \in E \tag{3.9}$$

($\pi^T \mathbf{Q} = 0$). If the equation set (3.9) together with the normalising equation

$$\sum_{i\in E} \pi_i = 1 \tag{3.10}$$

has a solution $(\pi_i)_{i\in E}$, then the solution is unique and $\pi_i > 0$ for all i. This is called the stationary *(or* steady-state *or* limit*)* distribution *of the CTMC.*

(c) If a stationary distribution $(\pi_i)_{i\in E}$ exists, then

$$\lim_{t\to\infty} \frac{T_t^i}{t} = \pi_i \quad \text{w.p. 1} \quad \textit{for all } i \in E \tag{3.11}$$

and as a consequence of $\pi_i > 0$ we have

$$\lim_{t\to\infty} T_t^i = \infty, \quad i \in E. \tag{3.12}$$

Result (3.11) is basic to calculations below. For this reason, we want at least some idea why it holds. We argue why the corresponding mean converges to π_i:

$$\begin{aligned}\mathbb{E}T_t^i = \mathbb{E}\left[\int_0^t 1_{\{X_u=i\}}\mathrm{d}u\right] &= \int_0^t \mathbb{E}[1_{\{X_u=i\}}]\mathrm{d}u \quad \text{(interchange of } \mathbb{E}[] \text{ and } \textstyle\int \text{ assumed valid)} \\ &= \int_0^t p_i(u)\mathrm{d}u.\end{aligned}$$

Then

$$\lim_{t\to\infty} \frac{\mathbb{E}\left[\int_0^t 1_{\{X_u=i\}}\mathrm{d}u\right]}{t} = \lim_{t\to\infty} \frac{\int_0^t p_i(u)\mathrm{d}u}{t} = \pi_i,$$

the last step following from $\lim_{u\to\infty} p_i(u) = \pi_i$, which is said in (3.8).

As π_i is the (long-run) fraction of time spent at state i, the balance equation (3.9) for state i says

$$(\text{fraction of time at } i) \times q_i = \sum_{k\in E, k\neq i} (\text{fraction of time at } k) \times q_{k,i}$$

the left side being a "rate out of i" and the right side being a "rate into i".

3.4.1 Long-Run Average Cost

Proofs here are non-examinable. (2.26) will be used heavily, for arrival events. [1]

Cost a Deterministic Function of System State

Suppose that whenever $X_u = i$, we incur a cost $f(i)$ per unit time, where f is a given function (while thinking "cost" may help, f could represent anything, e.g., a gain). Then the cost per unit time, up to time t, is

$$\frac{\int_0^t f(X_u)\mathrm{d}u}{t} = \sum_{i\in E} f(i)\frac{\int_0^t 1_{\{X_u=i\}}\mathrm{d}u}{t},$$

Thus the long-run cost per unit time is

$$\lim_{t\to\infty}\sum_{i\in E} f(i)\frac{\int_0^t 1_{\{X_u=i\}}\mathrm{d}u}{t} = \sum_{i\in E} f(i)\left(\lim_{t\to\infty}\frac{\int_0^t 1_{\{X_u=i\}}\mathrm{d}u}{t}\right) = \sum_{i\in E} f(i)\pi_i \quad \text{w.p. 1} \tag{3.13}$$

by (3.11) in the last step, and by assuming that the limit can pass inside the sum, which is the case if E is finite. This is abbreviated as "average of $f(X)$". Note this is $\mathbb{E}[f(X)]$ for the random variable X whose distribution is $(\pi_i)_{i\in E}$.

We now apply this.

Example 3.2 (continued) Find the following long-run quantities: (a) the average number of jobs in the system; (b) the average number of busy servers; and (c) the fraction of time that both servers are busy.

To identify the stationary distribution, solve $\sum_{i=1}^5 \pi_i = 1$ together with (3.9), i.e. (the generator was given previously):

State	Rate Out	=	Rate In
1	$4\pi_1$	=	$2\pi_3 + 3\pi_2$
2	$7\pi_2$	=	$4\pi_1 + 2\pi_4$
3	$6\pi_3$	=	$3\pi_4$
4	$9\pi_4$	=	$4\pi_2 + 4\pi_3 + 5\pi_5$
5	$5\pi_5$	=	$4\pi_4$

Obtaining the solution is then straightforward. [2]

(a) Following the cost result (3.13), the average number of jobs in the system is $1\cdot(\pi_2+\pi_3) + 2\pi_4 + 3\pi_5$ ($f(1)=0$; $f(2)=f(3)=1$; $f(4)=2$; $f(5)=3$).

(b) $1\cdot(\pi_2+\pi_3) + 2(\pi_4+\pi_5)$ ($f(1)=0$; $f(2)=f(3)=1$; $f(4)=f(5)=2$).

(c) $\pi_4+\pi_5$ (f is an indicator function, valued 1 at states 4 and 5, and 0 elsewhere).

[1] Assume all inter-arrival times are finite; then the arrival times S_n are finite for all n; then the number of events up to t, $N_t = \sum_{n=1}^\infty 1_{\{S_n\le t\}}$, satisfies $\lim_{t\to\infty} N_t = \infty$, verifying the A2 there.

[2] The solution is the vector $(65, 60, 40, 80, 64)/309$.

Cost Events

We consider here costs that arise as counts of certain *cost events*. In the first model, while the system state is i, cost events are a Poisson process with rate c_i. Put

$$C_t^i = \text{number of cost events that occur while the state is } i, \text{ up to time } t \tag{3.14}$$

so $C_t = \sum_{i \in E} C_t^i$ is the number of cost events up to time t. Write

$$\frac{C_t}{t} = \sum_{i \in E} \frac{C_t^i}{T_t^i} \frac{T_t^i}{t}$$

and recall T_t^i is the time spent at i, see (3.7). The fractions on the right converge:

1. T_t^i/t converges (w.p. 1) to the state-i stationary probability, π_i, by (3.11).
2. C_t^i/T_t^i converges (w.p. 1) to the underlying rate, c_i, by (2.26) (the assumption there that time goes to ∞ is checked by (3.12)).

Thus

$$\lim_{t \to \infty} \frac{C_t}{t} = \sum_{i \in E} \lim_{t \to \infty} \frac{C_t^i}{T_t^i} \lim_{t \to \infty} \frac{T_t^i}{t} = \sum_{i \in E} c_i \pi_i \tag{3.15}$$

where the "w.p. 1" will be dropped for convenience.

In the second model, assume a cost event is triggered for each arrival if and only if it finds the system (X_t) in one of the states in a set A (for example, A could have a single "system is full" state). Let C_t be the number of cost events up to t; then the average number of cost events per unit time is

$$\frac{C_t}{t} = \sum_{i \in A} \frac{N_t^i}{N_t} \frac{N_t}{t},$$

where

$$N_t^i = \text{number of arrivals that find the system in state } i, \text{ up to time } t \tag{3.16}$$

and N_t is the number of arrivals up to time t. As $t \to \infty$:

1. N_t/t converges (w.p. 1) to the arrival rate, call it λ, again by (2.26).
2. We will later see a Theorem called PASTA (Poisson Arrivals see Time Averages) that ensures that $\frac{N_t^i}{N_t}$ converges (w.p. 1) to the state-i stationary probability, π_i.

Thus

$$\lim_{t \to \infty} \frac{C_t}{t} = \sum_{i \in A} \lim_{t \to \infty} \frac{N_t^i}{N_t} \lim_{t \to \infty} \frac{N_t}{t} = \sum_{i \in A} \pi_i \lambda. \tag{3.17}$$

Example 3.4 Jobs (customers) arrive to a system according to a Poisson process of rate $\lambda = 5$. There are two servers, and service times at either server are exponentially distributed with mean $1/\mu = 1/3$ (rate $\mu = 3$), and independent of everything else. At most two jobs may wait for service, so a job that arrives to find 2 jobs waiting *balks*, meaning it does not join the queue. Moreover, each job i *abandons*, meaning it leaves without being served, as soon as its waiting time in queue is equal to Y_i, where the Y_i are exponentially distributed with mean $1/\eta = 1/2$ (rate $\eta = 2$), independently of everything else.

(The system state, X_t, defined as the number of jobs in the system, thus takes values in $E = \{0, 1, 2, 3, 4\}$ and the independent exponential distributions imply that $(X_t : t \geq 0)$ is a CTMC.) We want to calculate:

(a) The long-run average number of abandons per unit time.

(b) The long-run average number of balks per unit time.

First, calculate the generator. Single events that cause a state change are the arrivals, departures of served customers, and (what is new here) abandons, which act the same way as departures, decreasing the state by one. The balk events do not change the state.

The rate of abandons depends on the state: when in a state with k jobs in queue (importantly the state determines k: $k = 1$ in state 3, $k = 2$ in state 4, $k = 0$ in other states), and since η is given as the rate at which an individual job abandons, the abandon rate is $k\eta$ (merging property). Summing the rates of the events causing the same state transition (as usual), we find the generator is

$$\mathbf{Q} = \left[\begin{array}{c|c|c|c|c} & \lambda & & & \\ \hline \mu & & \lambda & & \\ \hline & 2\mu & & \lambda & \\ \hline & & 2\mu+\eta & & \lambda \\ \hline & & & 2\mu+2\eta & \end{array}\right]$$

(the main diagonal is implicit, and other entries are zero). This is a birth-death process (i.e., we have zeros everywhere except on the diagonals above and below the main diagonal). In Section 3.4.2 below, we derive the stationary distribution of a general birth-death process (with infinite state space); this essentially also shows how the finite-state space solution can be obtained. So we do not pursue a solution here. Denote the stationary distribution $(\pi_i)_{i=0}^4$.

(a) Following the first cost result, (3.15), the long-run average number of abandons per unit time is $\eta\pi_3 + 2\eta\pi_4$ (cost rates $c_3 = \eta$, $c_4 = 2\eta$, and 0 at other states).

(b) Following the second cost result, (3.17), the long-run average number of balks per unit time is $\pi_4\lambda$ ($A = \{4\}$).

♣ Study the following. E03 2, Two Stations in Tandem. Involves thinning and thereby a relatively tricky calculation of the generator.

3.4.2 Explicit Stationary Distributions for Specific Models

Birth-Death Process with Infinite State Space

The generator is in (3.4), so the balance equations become

$$\left.\begin{array}{rl} \pi_0\lambda_0 &= \pi_1\mu_1 \\ \pi_i(\lambda_i+\mu_i) &= \pi_{i-1}\lambda_{i-1}+\pi_{i+1}\mu_{i+1}, \quad i=1,2,\ldots. \end{array}\right\} \tag{3.18}$$

Solving iteratively gives

$$\pi_i\lambda_i = \pi_{i+1}\mu_{i+1}, \quad i=0,1,\ldots \tag{3.19}$$

i.e., $\pi_{i+1} = \pi_i\lambda_i/\mu_{i+1}$, from which we obtain (proceeding down to state zero)

$$\pi_k = \pi_{k-1}\frac{\lambda_{k-1}}{\mu_k} = \pi_{k-2}\frac{\lambda_{k-2}}{\mu_{k-1}}\frac{\lambda_{k-1}}{\mu_k} = \ldots = \pi_0\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu_1\mu_2\ldots\mu_k}, \quad k=1,2,\ldots. \tag{3.20}$$

Normalise, i.e., require (3.10) (substitute the π's in terms of π_0):

$$\sum_{i=0}^{\infty}\pi_i = \pi_0\left(1+\frac{\lambda_0}{\mu_1}+\frac{\lambda_0\lambda_1}{\mu_1\mu_2}+\ldots\right) = \pi_0\left(1+\sum_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu_1\mu_2\ldots\mu_k}\right) = 1. \tag{3.21}$$

We see that a unique solution (stationary distribution) exists if and only if

$$1+\sum_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu_1\mu_2\ldots\mu_k} < \infty. \tag{3.22}$$

It is then given by (3.20), with π_0 determined from (3.21).

Application 3.5 Our standard notation for the M/M/c model will be an arrival rate denoted λ (i.e., inter-arrival times exponentially distributed with mean $1/\lambda$) and an individual-server service rate denoted μ (i.e., service times exponentially distributed with mean $1/\mu$). With X_t the number of jobs in the system at time t, $X = (X_t : t \geq 0)$ is a birth-death process with state space $\{0, 1, 2, \ldots\}$, as there is infinite waiting space. As servers are non-idling, μ_n, the death rate when $X_t = n$, is the rate of departures aggregated over all busy servers, i.e., the individual-server departure rate, μ, times the number of busy servers; thus

$$\mu_n = \begin{cases} n\mu & n < c \\ c\mu & n \geq c \end{cases}$$

The birth (arrival) rate is λ, regardless of X_t; that is, $\lambda_n = \lambda$ for all n. Putting

$$\rho := \frac{\lambda}{c\mu},$$

the birth-death solution (3.20) becomes

$$\pi_n = \begin{cases} \pi_0 \frac{\lambda^n}{\mu \cdot 2\mu \cdot 3\mu \cdots n\mu} = \pi_0 \frac{c^n \rho^n}{n!} & n \leq c \\ \pi_c \rho^{n-c} = \pi_0 \frac{c^c \rho^n}{c!} & n \geq c \end{cases} \tag{3.23}$$

(the two branches agree at $n = c$) and π_0 must satisfy

$$1 = \sum_{n=0}^{c-1} \pi_n + \sum_{n=c}^{\infty} \pi_n = \pi_0 \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \sum_{n=c}^{\infty} \rho^{n-c} \right].$$

Thus, a stationary distribution exists if and only if $\sum_{j=0}^{\infty} \rho^j < \infty$, i.e., if and only if $\rho < 1$. In this case, $\sum_{j=0}^{\infty} \rho^j = \frac{1}{1-\rho}$, and

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!(1-\rho)} \right]^{-1}. \tag{3.24}$$

For future reference, for $c = 1$ (M/M/1 model), we get $\pi_0 = 1 - \rho$ (the sum is 1, the second term is $\rho/(1-\rho)$), and we find $\pi_n = (1-\rho)\rho^n$ for all n.

With X_t being the number of jobs in the system in the finite-waiting space model M/M/c/k and in extensions that allow balks and abandons (with exponentially distributed patience times), a stationary distribution exists for any ρ (as there are only $c + k + 1$ π's); it is not difficult to compute by using the general birth-death solution (3.20) and (3.21).

♣ Study the following M/M/c-type problems. • E04 1, except for the question on waiting times (last paragraph) : M/M/1 versus an M/M/2 with half-as-fast servers, i.e., same overall speed. • E06 1: compare three separate M/M/1 systems to a centralised M/M/2 that handles all work.

Transitions Beyond Infinite-Case Birth-Death, Solvable via PGFs

♣ Study the following.

- Exercise 6. Then recognise that E06 3, E07 2, E10 3 are all similar to this. The modelling novelty is to allow batch arrivals. An infinite-state-space CTMC arises, with structure more complex than the birth-death case. Solvable via the pgf of the stationary distribution, as shown in the exercise.

- E04 2. This is the M/E_k/1 Model, where E_k refers to the Erlang-k distribution, meaning that the service time is the sum of k independent exponentially distributed service *stages* (of known rate each). A reduction to an infinite-state-space CTMC is done, the state variable now counting stages (in the system) rather than jobs. The balance equations are exactly as in Exercise 6 (but the state variable has different meaning across the two problems), and thus so is the pgf of the stationary distribution. In answering questions involving the number of jobs, we must translate from a job count to (a set of) stage counts: for example, if $k = 2$, then the state "1 job in the system" is equivalent to "1 stage in the system" or "2 stages in the system".

3.5 Arrivals That See Time Averages

We now state carefully a key theorem that states, under conditions, the equality of long-run customer averages to associated stationary probabilities.

Theorem 3.6 (Arrivals See Time Averages (ASTA)) *Let $X = (X_t : t \geq 0)$ be the number of jobs in the system, assumed to be a CTMC with state space E and stationary distribution $(\pi_i)_{i \in E}$. Write T_j for the j-th arrival time. Assume that for all t, the arrival-counting process (essentially the T_j's) forward from t is independent of the history of X up to t. Then*

$$\lim_{n \to \infty} \frac{\sum_{j=1}^{n} 1_{\{X_{T_j} = i\}}}{n} = \pi_i \quad \text{w.p. 1}, \quad i \in E. \tag{3.25}$$

Taking expected value of the left side results in the indicator random variables being replaced by probabilities, so the above gives

$$\lim_{n \to \infty} \frac{\sum_{j=1}^{n} \mathbb{P}(X_{T_j} = i)}{n} = \pi_i. \tag{3.26}$$

If the arrivals are a Poisson process, then the theorem's assumptions are true, and this special case is called PASTA, the added "P" standing for "Poisson". In Exercises and Exams from Year 2007/08 onwards, PASTA is mentioned when used. In pre-2007/08 exams, PASTA is implicitly assumed without being mentioned.

With deterministic arrival and service times, the ASTA conclusions tend to fail: although there exists a long-run fraction of time in state i, it tends to differ from the long-run fraction of arrivals that find the system in this state.

3.5.1 A Steady-State Delay Distribution

Problem 3.7 Customers arrive to an M/M/1 queue at rate $\lambda = 1/2$ per hour. Service discipline is first come first served (FCFS). Two rates of service are possible. The faster rate of service is $\mu = 1$ per hour and costs £40 per hour; the slower service rate is $\mu = 0.75$ per hour and costs £30 per hour. A cost of £200 is incurred for each customer that waits in queue (i.e., excluding service) more than one hour. Find which service rate gives the lowest average cost per hour.

We will analyse a slightly more general problem, where we replace the constant "1" (the "1 hour") by any $s \geq 0$, and assume the M/M/c model for c general. Let W_j be the delay (time spent in queue) of the j-th job (customer). Write

$$C_t = \text{number of arrivals up to } t \text{ that wait more than } s = \sum_{j=1}^{N_t} 1_{\{W_j > s\}}.$$

Similar to the cost model that gave (3.17), write

$$\frac{C_t}{t} = \frac{C_t}{N_t}\frac{N_t}{t}$$

where N_t is the number of arrivals up to t. Now, as $t \to \infty$, N_t/t converges (w.p. 1) to the arrival rate, as seen in (2.26), and $\lim_{t\to\infty} N_t = \infty$. We attempt to guess $\lim_{t\to\infty} C_t/N_t$ from the corresponding (limit of) expected value:

$$F(s) := \lim_{t\to\infty} \mathbb{E}\left[\frac{C_t}{N_t}\right] = \lim_{t\to\infty} \mathbb{E}\left[\frac{1}{N_t}\sum_{j=1}^{N_t} 1_{\{W_j > s\}}\right] = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(W_j > s), \quad s \geq 0. \tag{3.27}$$

We call $F()$ a *steady-state delay distribution* (it is usually a proper distribution).

To find F, let T_j be the arrival time of the j-th job, and condition on the number of jobs it finds on the system, denoted X_{T_j}:

$$\mathbb{P}(W_j > s) = \sum_{k=0}^{\infty} \mathbb{P}(W_j > s | X_{T_j} = k)\mathbb{P}(X_{T_j} = k) \tag{3.28}$$

for any $s \geq 0$ (this is the Law of Total Probability (LTP), Theorem 1.10, with partition events $\{X_{T_j} = k\}$, $k = 0, 1, 2, \ldots$).

Supposing the system has c servers, it suffices to sum over $k \geq c$, since otherwise the waiting time is zero; thus

$$\begin{aligned} F(s) &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n}\sum_{k=0}^{\infty} \mathbb{P}(W_j > s | X_{T_j} = c+k)\mathbb{P}(X_{T_j} = c+k) \quad \text{by (3.28)} \\ &= \sum_{k=0}^{\infty} \mathbb{P}(W > s | X = c+k) \cdot \left(\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(X_{T_j} = c+k)\right) \end{aligned} \tag{3.29}$$

(limit passed inside sum), where we abbreviate W_j as W and X_{T_j} as X, as j does not affect the probability. Now note that:

- The term in parenthesis, $\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{P}(X_{T_j} = c + k)$, equals the stationary probability that the process X is at state $c + k$, by PASTA, (3.26).

- The conditional probability has been considered, under the *First-Come, First-Serve* (FCFS) discipline, in Application 2.11: given $X = c + k$ with $k \geq 0$, W has a Gamma distribution. Specifically, assuming Expon(μ) service times, plugging the Gamma result from there and the π's from (3.23) into (3.29) and simplifying gives

$$F(s) = \frac{\pi_c}{1-\rho} e^{-(c\mu-\lambda)s}, \quad s \geq 0, \tag{3.30}$$

  where $F(0) = \pi_c/(1-\rho) = \sum_{n=c}^{\infty} \pi_n$ is the stationary probability that all servers are busy. The formula describes the steady-state delay as a *mixed distribution* with two components: with probability $1 - F(0)$, it is zero; with the remaining probability, $F(0)$, it is distributed as Expon($c\mu - \lambda$).

The answers to Problem 3.7 are now direct from (3.30) and the M/M/1 stationary distribution, given following (3.24). In the slow system, $\pi_c/(1-\rho) = \rho = 2/3$,

$$F^{\text{slow}}(1) = \frac{2}{3} e^{-\frac{1}{4}\cdot 1} \doteq 0.519,$$

and the long-run average cost per hour is $30 + 200\lambda F^{\text{slow}}(1) = £81.9$. Exactly in the same way, in the fast system, $\pi_c/(1-\rho) = \rho = 1/2$,

$$F^{\text{fast}}(1) = \frac{1}{2} e^{-\frac{1}{2}\cdot 1} \doteq 0.303$$

and the long-run average cost per hour is $40 + 200\lambda F^{\text{fast}}(1) = £70.3$. Thus the faster system has lower cost.

♣ Study the following. Exams E03 1, E07 3 (last 10 points), E11 4(b) are all essentially the above analysis.

3.6 Little's Law

Little's Law states the existence of three long-run limits and a link between them. We need the following.

N_t	number of arrivals up to time t
X_t	number of jobs present in the system at time t
W_j	time spent in the system by job j

Very roughly speaking, the central idea is that there exist (random) times such that all the processes "probabilistically restart" themselves, independently of the past, and moreover, the processes go through infinitely many such "independent identically distributed" cycles. Taking these times as

$\tau_i = i$-th time the system becomes empty after having been non-empty, $\quad i = 1, 2, \ldots$

where we put $\tau_0 = 0$, and putting

$A_i = \int_{\tau_{i-1}}^{\tau_i} X_u \mathrm{d}u$	= "area" of X process during the i-th cycle
$\tilde{N}_i = N_{\tau_i} - N_{\tau_{i-1}}$	= number of arrivals during the i-th cycle

the following is a set of precise supporting assumptions:

A1. The cycle lengths $(\tau_i - \tau_{i-1})_{i=1}^{\infty}$ are iid. The areas $(A_i)_{i=1}^{\infty}$ are iid. The arrival counts $(\tilde{N}_i)_{i=1}^{\infty}$ are iid.

A2. $\mathbb{E}[\tau_1] < \infty$ (finite mean cycle length) and $\mathbb{E}[\tilde{N}_1] < \infty$ (finite mean number of arrivals per cycle).

A3. The cycle containing t, $R_t := \min\{i : \tau_i \geq t\}$, and the cycle during which job k arrives, $C_k := \min\{i : N_{\tau_i} \geq k\}$, satisfy

$$\lim_{t\to\infty} R_t = \infty, \quad \lim_{k\to\infty} C_k = \infty, \quad \text{w.p. 1.}$$

Theorem 3.8 (Little's Law) *Assume A1 to A3. Then,*

(i) Long-run average number in system:

$$\lim_{t\to\infty} \frac{\int_0^t X_u \mathrm{d}u}{t} = \frac{\mathbb{E}[A_1]}{\mathbb{E}[\tau_1]} =: L \quad \text{w.p. 1.} \tag{3.31}$$

(ii) Long-run average arrival rate:

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{\mathbb{E}[\tilde{N}_1]}{\mathbb{E}[\tau_1]} =: \lambda \quad \text{w.p. 1.} \tag{3.32}$$

(iii) Long-run average waiting time:

$$\lim_{k\to\infty} \frac{\sum_{i=1}^{k} W_i}{k} = \frac{\mathbb{E}[A_1]}{\mathbb{E}[\tilde{N}_1]} =: W \quad \text{w.p. 1.} \tag{3.33}$$

Thus, in particular, $L = \lambda W$.

Proof. The proof is non-examinable and given for completeness.

(i) For any t and for $n = R_t$ being the cycle containing t, we have $\tau_{n-1} \le t \le \tau_n$ and

$$A_1 + \ldots + A_{n-1} \le \int_0^t X_u \mathrm{d}u \le A_1 + \ldots + A_n.$$

Dividing the latter inequality by the (reversed) former inequality,

$$\frac{A_1 + \ldots + A_{n-1}}{\tau_n} \le \frac{\int_0^t X_u \mathrm{d}u}{t} \le \frac{A_1 + \ldots + A_n}{\tau_{n-1}}. \tag{3.34}$$

We will take limits of the bounds as $t \to \infty$ and thus $n = R_t \to \infty$. Consider the lower bound (left side of (3.34)) carefully. The enumerator is a sum of a large number of A's, $n-1$ of them, and the denominator $\tau_n = \sum_{i=1}^{n}(\tau_i - \tau_{i-1})$ is the sum of n cycle lengths, suggesting a "Strong Law of Large Numbers" effect for both. Indeed, as $n \to \infty$,

$$\frac{A_1 + \ldots + A_{n-1}}{\tau_n} = \underbrace{\frac{A_1 + \ldots + A_{n-1}}{n-1}}_{\to \mathbb{E}[A_1] \text{ w.p. } 1} \underbrace{\frac{n-1}{n}}_{\to 1} \underbrace{\frac{1}{\frac{\sum_{i=1}^{n}(\tau_i - \tau_{i-1})}{n}}}_{\to 1/\mathbb{E}[\tau_1 - \tau_0] = 1/\mathbb{E}[\tau_1] \text{ w.p. } 1} \tag{3.35}$$

by using the SLLN for the areas (A's) and for the cycle lengths ($\tau_i - \tau_{i-1}$'s). That is, the above converges to $\mathbb{E}[A_1]/\mathbb{E}[\tau_1]$ w.p. 1. Checking that the upper bound (right side of (3.34)) has the same limit, the proof is complete.

(ii) The ideas are similar to (i). For any t and for $n = R_t$, we have

$$\frac{N_{\tau_{n-1}}}{\tau_n} \le \frac{N_t}{t} \le \frac{N_{\tau_n}}{\tau_{n-1}} \tag{3.36}$$

and the limit of the lower bound as $t \to \infty$ (thus $n = R_t \to \infty$) is

$$\lim_{n\to\infty} \frac{N_{\tau_{n-1}}}{\tau_n} = \lim_{n\to\infty} \frac{\tilde{N}_1 + \ldots + \tilde{N}_{n-1}}{n-1} \frac{n-1}{n} \frac{n}{\tau_n} = \mathbb{E}[\tilde{N}_1] \frac{1}{\mathbb{E}[\tau_1]} \quad \text{w.p. 1} \tag{3.37}$$

by using the SLLN for the $\tilde{N}$'s and for the cycle lengths. Checking that the upper bound (right side of (3.36)) has the same limit, the proof is complete.

(iii). Since a job always departs by the end of the cycle during which it arrives, we have

$$\sum_{i=1}^{N_{\tau_n}} W_i = A_1 + \ldots + A_n \quad \text{for all } n.$$

Consider the k-th arrival and let $n = C_k$ be the cycle during which it occurs. Then, $N_{\tau_{n-1}} \le k \le N_{\tau_n}$, and

$$\frac{A_1 + \ldots + A_{n-1}}{N_{\tau_n}} = \frac{\sum_{i=1}^{N_{\tau_{n-1}}} W_i}{N_{\tau_n}} \le \frac{\sum_{i=1}^{k} W_i}{k} \le \frac{\sum_{i=1}^{N_{\tau_n}} W_i}{N_{\tau_{n-1}}} = \frac{A_1 + \ldots + A_n}{N_{\tau_{n-1}}}. \quad (3.38)$$

The limit of the lower bound as $k \to \infty$ (thus $n = C_k \to \infty$) is

$$\lim_{n\to\infty} \frac{A_1 + \ldots + A_{n-1}}{N_{\tau_n}} = \lim_{n\to\infty} \frac{A_1 + \ldots + A_{n-1}}{\tau_n} \lim_{n\to\infty} \frac{\tau_n}{N_{\tau_n}} = \frac{\mathbb{E}[A_1]}{\mathbb{E}[\tau_1]} \frac{\mathbb{E}[\tau_1]}{\mathbb{E}[\tilde{N}_1]} = \frac{\mathbb{E}[A_1]}{\mathbb{E}[\tilde{N}_1]} \text{ w.p. 1}$$

(the first limit is proved in (3.35) and the second limit is essentially proved in (3.37)). Checking that the upper bound (right side of (3.38)) has the same limit, the proof is complete. □

Application 3.9 Based on the π's in (3.23) for the M/M/c model, we give formulas for certain (long-run) averages for the system and for the queue only (i.e., excluding service), using Little's Law to go from average numbers to average waiting times (for which we previously had no theory). The long-run average number of jobs in the M/M/c queue, call it L_q, is (apply (3.13) with cost function "# in queue", $f(n) = n - c$ for $n \ge c$ and 0 otherwise):

$$L_q = \text{average \# in queue} = \sum_{n=c}^{\infty} (n-c)\pi_n = \pi_c \sum_{j=1}^{\infty} j\rho^j = \pi_c \frac{\rho}{(1-\rho)^2}.^3 \quad (3.39)$$

Let W and W_q be the (long-run) average waiting times in the system and in queue, respectively. These are linked as follows: for any job j, the time in the system is the sum of $W_{j,q}$ = time in queue plus S_j = time in service. Then

$$W := \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} W_j = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} W_{j,q} + \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} S_j = W_q + \frac{1}{\mu} \quad \text{w.p. 1}, \quad (3.40)$$

the W_q arising by definition, and $1/\mu$ arising by the SLLN for the S_j. Then, letting L = average # in system $= \sum_{n=1}^{\infty} n\pi_n$ we have the links

$$\begin{aligned} L &= \lambda W \qquad \text{(Little's Law for the system)} \\ &= \lambda \left(W_q + \frac{1}{\mu} \right) \\ &= L_q + \frac{\lambda}{\mu} \qquad \text{(via Little's Law for the queue).} \end{aligned} \quad (3.41)$$

[3] For $\rho < 1$, we have $\sum_{i=1}^{k} i\rho^i = \rho \sum_{i=1}^{k} \frac{\mathrm{d}}{\mathrm{d}\rho} \rho^i = \rho \frac{d}{d\rho} \sum_{i=0}^{k} \rho^i = \rho \frac{\mathrm{d}}{\mathrm{d}\rho} \frac{1-\rho^{k+1}}{1-\rho} = \rho \frac{1-(k+1)\rho^k + k\rho^{k+1}}{(1-\rho)^2}$; taking limits as $k \to \infty$, we have $\sum_{i=1}^{\infty} i\rho^i = \frac{\rho}{(1-\rho)^2}$.

♣ Study the following problems, which involve average waiting times and Little's Law. • E04 1 remainder on waiting times. • Exercise 3. Comparison of M/M/1 versus M/M/2 (each server has the same speed) in terms of a (long-run average) cost that has waiting and staffing components.

3.7 Exercises

Recall: (i) $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. (ii) For $0 \le \rho < 1$, $\sum_{i=0}^{\infty} \rho^i = 1/(1-\rho)$.

1. (Random Balking.) Consider an M/M/1 queue with arrival rate λ and service rate μ. An arrival that finds i jobs in the system acts randomly, independently of everything else, as follows: with probability p_i, it joins the system; with probability $1 - p_i$, it balks, i.e., does not join. Let X_t be the number of jobs in the system at time t. (i) Determine $\lim_{h\to 0} \mathbb{P}(X_{t+h} = i+1 | X_t = i)/h$ carefully, showing all your work. (ii) Identify $q_{ij} := \lim_{h\to 0} \mathbb{P}(X_{t+h} = j | X_t = i)$ for all $i \neq j$; no derivation is required. Given that (X_t) is a birth-death process, identify the birth and death rates. (iii) For the case $p_i = 1/(i+1)$ for all $i = 0, 1, \ldots$, find the stationary distribution of (X_t) in terms of λ and μ.

2. (Number of servers adapts to the number of customers in the system.) Customers arrive at a system according to a Poisson process of rate λ. With X the number of customers in the system, assume that the number of servers is one whenever $X \le m$ and two whenever $X > m$. Assume service times are Expon(μ). Show that the stationary distribution of X has the form

$$\pi_k = \begin{cases} \pi_0 (2\rho)^k & k = 1, 2, \ldots, m \\ \pi_0 (2\rho)^m \rho^{k-m} & k = m+1, m+2, \ldots \end{cases} \tag{3.42}$$

   where $\rho = \lambda/(2\mu)$, identifying results you use without proof. How can π_0 be determined?

3. Customers arrive at a system at rate 5 per hour, and the average time to service them is 1/6 hours. It costs $8x$ per hour to have x customers in the system (waiting cost) plus $5c$ per hour to have c servers (server cost). Find the c in $\{1, 2\}$ that minimises the long-run average cost. State any assumptions made. *Hint:* Use (3.41) and the M/M/c stationary distribution, (3.23).

4. For the M/M/c model with arrivals of rate λ and with Expon(μ) service times, verify (3.30) by combining Application 2.11 (waiting time when encountering k customers in queue has the Gamma($k+1, c\mu$) distribution, see (2.18)) and the fact (from Example 3.5) $\pi_{c+k} = \pi_c \rho^k$ for $k = 0, 1, \ldots$, where $\rho = \lambda/(c\mu)$. *Hint*: In the resulting doubly-indexed sum, change the order of summation.

5. (Failure of ASTA Conclusions in a deterministic queue.) Let X_t be the number of jobs in a single-server system at time t, where $X_0 = 0$, arrivals occur at times 0, 10, 20, 30, ..., and service times are 9 for each job. Here, the function $(X_t : t \ge 0)$ is deterministic. State it and determine the following limits:

   (a) The long-run fraction of arrivals that find the server busy.

   (b) The long-run fraction of time the server is busy.

   Does the ASTA Theorem apply in this case?

6. (Solving certain balance equations via probability generating functions.) Jobs arrive in a system in batches of size $b > 1$, with batch-arrival events occurring according to a Poisson process of rate λ. There is one server at which service times

are Expon(μ) rv's, independent of everything else. There is infinite waiting space. Let X_t be the number of jobs in the system at time t.

(i) Given that (X_t) is a CTMC with values in $\{0, 1, 2, \ldots\}$, briefly argue that its stationary distribution $\{\pi_i\}_{i=0}^{\infty}$, when it exists, satisfies the following.

State	Rate Out	=	Rate In
0	$\pi_0\lambda$	=	$\pi_1\mu$
i $(1 \le i < b)$	$\pi_i(\lambda+\mu)$	=	$\pi_{i+1}\mu$
i $(b \le i)$	$\pi_i(\lambda+\mu)$	=	$\pi_{i+1}\mu + \pi_{i-b}\lambda$

(ii) Let $G(z) = \sum_{i=0}^{\infty} \pi_i z^i$ be the probability generating function (pgf) of the distribution $\{\pi_i\}_{i=0}^{\infty}$. Show that $G(z) = \frac{N(z)}{D(z)}$, where $N(z) = \mu\pi_0(1-z)$ and $D(z) = \lambda z^{b+1} - (\lambda+\mu)z + \mu$. *Hint:* Multiply the state-i equation above by z^i, for each i, then sum over all i (from 0 to ∞), then make appropriate "corrections".

(iii) To find π_0 we require $1 = \sum_{i=0}^{\infty} \pi_i = G(1)$ (normalising equation). It is given that $G(1) = \lim_{z\uparrow 1} G(z)$ (z increases to 1; continuity of $G()$, Fact 1.18). Using L'Hopital's rule, work out the limit as a function of π_0, and thus show that

$$\pi_0 = 1 - \frac{b\lambda}{\mu}.$$

Note that the remaining π_i's are then determined by (1.24).

3.8 Solutions to Exercises

1. Short answer to (i). Arrivals happen with rate λ. When $X_t = i$, an arrival joins with probability p_i, so "customer-joins-the-system" events happen with rate λp_i.

   Full answer to (i). Let $A = \#$ of arrivals in $(t, t+h]$ and $D = \#$ of departures in the same interval. I indicates if a given arrival joins (1-yes, 0-no), I is independent of everything else, and $\mathbb{P}(I = 1|X_t = i) = p_i$. Then, intending $h \to 0$,

   $$\begin{aligned}
   &\mathbb{P}(X_{t+h} = i+1 | X_t = i) \\
   &= \mathbb{P}(A = 1, D = 0, I = 1 | X_t = i) + \underbrace{\mathbb{P}(A + D \ge 2 \text{ and other conditions} | X_t = i)}_{o(h)} \\
   &= \mathbb{P}(A = 1|X_t = i)\mathbb{P}(D = 0|X_t = i)\mathbb{P}(I = 1|X_t = i) + o(h) \quad \text{by independence} \\
   &= [\lambda h + o(h)][1 - \mu h + o(h)]p_i + o(h) \\
   &= \lambda p_i h + o(h)
   \end{aligned}$$

   where we need not specify the "other conditions". Thus

   $$\lim_{h\to 0} \frac{\mathbb{P}(X_{t+h} = i+1|X_t = i)}{h} = \lim_{h\to 0} \left(\lambda p_i + \frac{o(h)}{h}\right) = \lambda p_i.$$

   Note: The thinning Proposition 2.14 is a similar idea.

(ii) The set of values that X_t may take is $\{0, 1, 2, \ldots\}$. In (i) we showed $q_{ij} = \lambda p_i$ for $j = i+1$ and all i. Similar to the derivations (3.1) to (3.3), we see that $q_{ij} = \mu$ for all $i > 0$ and $j = i - 1$; and $q_{ij} = 0$ for all other $i \neq j$. That is, (X_t) is a birth-death process with birth rates $\lambda_i = \lambda p_i$ for all i and death rates $\mu_i = \mu$ for all $i > 0$.

(iii) From the general birth-death solution (3.20) and $p_i = 1/(i+1)$ we have

$$\pi_k = \pi_0 \frac{\lambda \cdot (\lambda/2) \cdot (\lambda/3) \cdots (\lambda/k)}{\mu^k} = \pi_0 \frac{\rho^k}{k!}, \quad k = 1, 2, \ldots.$$

where $\rho = \lambda/\mu$. Normalising gives $\pi_0 = \left(1 + \sum_{k=1}^{\infty} \frac{\rho^k}{k!}\right)^{-1} = (e^{\rho})^{-1} = e^{-\rho}$.

2. With X_t the number of customers in the system at time t, (X_t) is a birth-death process taking values in $\{0, 1, 2, \ldots\}$, with birth rates $\lambda_n = \lambda$ for all n, and death rates $\mu_n = \mu$ for $n \leq m$ and $\mu_n = 2\mu$ for $n > m$. Then, using the standard birth-death result (3.20), we obtain the stated equations. To determine π_0, insert the π_i from (3.42) into the normalising equation $\sum_{i=0}^{\infty} \pi_i = 1$ and solve for π_0.

3. We assume that inter-arrival and service times are exponentially distributed and independent. Then, the one-server and two-server systems are the M/M/1 and M/M/2 model, respectively. In our standard notation, $\lambda = 5$ and $\mu = 6$.

   By "cost" we mean long-run average cost per hour, in £. The waiting cost is $8L$, where L, the long-run average number of customers in the system, is a known function of $\rho = \lambda/(c\mu)$, via L_q, from (3.39) and (3.41). The cost is $8L + 5c$.

   M/M/1 calculation. Using (3.23), that is, the stationary distribution for the M/M/c model, for $c = 1$, we have $\pi_1 = (1-\rho)\rho$. Then, by (3.39), $L_q = \pi_1\rho/(1-\rho)^2 = \rho^2/(1-\rho)$; and (3.41) becomes $L = \rho^2/(1-\rho) + \rho = \rho/(1-\rho)$. Then, $\rho = \lambda/\mu = 5/6$, $L = 5$, and the cost is $8L + 5 \cdot 1 = 45$.

   M/M/2 calculation. Again by (3.23), this time for $c = 2$, we have $\pi_2 = \pi_0 2\rho^2$, where, from (3.24),

$$\pi_0 = \left[1 + 2\rho + \frac{2\rho^2}{1-\rho}\right]^{-1} = \frac{1-\rho}{1+\rho}.$$

   Then (3.39) becomes

$$L_q = \pi_2 \frac{\rho}{(1-\rho)^2} = (\pi_0 2\rho^2) \frac{\rho}{(1-\rho)^2} = \frac{2\rho^3}{1-\rho^2}$$

   and (3.41) becomes

$$L = \frac{2\rho^3}{1-\rho^2} + 2\rho = \frac{2\rho}{1-\rho^2}.$$

   We have $\rho = \lambda/(2\mu) = 5/12$, $L = 120/119 \doteq 1.0084$, and the total cost is $8L + 5 \cdot 2 \doteq 18.067$. Thus, the 2-server system has lower cost.

4. First, the conditional probability in (3.29) is

$$\mathbb{P}(W > s | X = c + k) = \sum_{i=0}^{k} e^{-c\mu s} \frac{(c\mu s)^i}{i!}, \quad s \geq 0, \ k = 0, 1, \ldots.$$

(Application 2.11; Gamma$(k+1, c\mu)$ tail probability, (2.18)). Recall that we have used ASTA to claim that $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(X_{T_j} = c+k) = \pi_{c+k}$. Now, by (3.23) we have $\pi_{c+k} = \pi_c \rho^k$ for $k = 0, 1, \ldots$, where $\rho = \lambda/(c\mu)$. Thus (3.29) becomes

$$\begin{aligned} F(s) &= \sum_{k=0}^{\infty}\sum_{i=0}^{k} e^{-c\mu s}\frac{(c\mu s)^i}{i!}\cdot \pi_c \rho^k \\ &= \pi_c e^{-c\mu s}\sum_{i=0}^{\infty}\frac{(c\mu s)^i}{i!}\sum_{k=i}^{\infty}\rho^k \quad \text{(reversed the summation order)} \\ &= \pi_c e^{-c\mu s}\sum_{i=0}^{\infty}\frac{(c\mu s)^i}{i!}\frac{\rho^i}{1-\rho} = \frac{\pi_c}{1-\rho}e^{-(c\mu-\lambda)s}. \quad (e^x = \textstyle\sum_{i=0}^{\infty}\frac{x^i}{i!} \text{ for } x = c\mu s\rho = \lambda s) \end{aligned}$$

5. For t from 0 to 9, we have $X_t = 1$, then for t from 9 to 10, we have $X_t = 0$; and this pattern of "cycles", of length 10 each, repeats forever. (We choose to not specify if X_t is 0 or 1 at times when it jumps from one to the other, as this does not matter.) Thus:

   (a) All arriving jobs find the server is idle ($X_t = 0$ just before each arrival time).

   (b) The long-run fraction of time that the server is busy (equivalently that $X_t = 1$) is 9/10.

   The (a) and (b) above are, respectively, the left and right side in the ASTA Theorem's conclusion, (3.25). The theorem does not apply here.

6. (i) State changes occur as follows: when at state i, a batch-arrival event changes the state to $i+b$, while a service-completion event changes the state to $i-1$; thus the generator entries are: $q_{i,i+b} = \lambda$ for all i; $q_{i,i-1} = \mu$ for all $i > 0$; and $q_{ij} = 0$ for other $i \neq j$. The stated equations are the usual balance equations (3.9) associated to this generator.

   (ii) Multiplying and summing as suggested gives

$$\begin{aligned} \lambda\pi_0 + (\lambda+\mu)\underbrace{\sum_{i=1}^{\infty}\pi_i z^i}_{=G(z)-\pi_0} &= \mu\sum_{i=0}^{\infty}\pi_{i+1}z^i + \lambda\sum_{i=b}^{\infty}\pi_{i-b}z^i \\ &= \mu\frac{1}{z}\underbrace{\sum_{i=0}^{\infty}\pi_{i+1}z^{i+1}}_{=G(z)-\pi_0} + \lambda z^b\underbrace{\sum_{i=b}^{\infty}\pi_{i-b}z^{i-b}}_{=G(z)}. \end{aligned}$$

   The key idea above is that each of the infinite sums in the first equation gives $G(z)$ after the "correction" steps seen in the second equation. Now, re-arrange to solve for G:

$$G(z)\left(\lambda+\mu-\frac{\mu}{z}-\lambda z^b\right) = \mu\pi_0\left(1-\frac{1}{z}\right) \Rightarrow G(z) = \frac{\mu\pi_0(1-z)}{\lambda z^{b+1}-(\lambda+\mu)z+\mu}.$$

   (iii) The derivatives of $N()$ and $D()$ are $N'(z) = -\mu\pi_0$ and $D'(z) = (b+1)\lambda z^b - \lambda - \mu$. Using L'Hopital's rule, $1 = G(1) = \lim_{z\uparrow 1} G(z) = \frac{N'(1)}{D'(1)} = \frac{-\mu\pi_0}{b\lambda-\mu} \Rightarrow \pi_0 = 1 - \frac{b\lambda}{\mu}$.

Chapter 4

Sampling from Distributions

We study some general methods for simulating (sampling) a value from a given univariate distribution, with support contained in the real numbers. The source of randomness is a pseudo-random number generator, assumed to return independent samples from the Unif$(0, 1)$ distribution, defined after (1.10).

4.1 Preliminaries

F will generally denote a cdf. Given a cdf F, we write "$X \sim F$" to mean "X is a sample of the rv whose cdf is F". Given a pdf f, we write "$X \sim f$" to mean "X is a sample of the rv whose pdf is f". The notions of "inf" (infimum) and "min" are taken to coincide. Likewise, "sup" (supremum) and "max" will coincide.

4.2 Inversion

Definition 4.1 *The* inverse *of the cdf F is the function*

$$F^{-1}(u) = \inf\{x : F(x) \geq u\} = \min\{x : F(x) \geq u\}, \quad 0 < u < 1.$$

To give a more explicit definition of the inverse, note first that any cdf F has a left limit everywhere (as it is non-decreasing), i.e., for any real a we can put

$$F(a-) = \lim_{x \to a-} F(x).$$

Definition 4.2 *Let F be a cdf.*

(a) If F is continuous and strictly increasing on an interval (a, b), then, for any u in $(F(a), F(b))$, define $F^{-1}(u)$ as the unique x in (a, b) that solves $F(x) = u$.

(b) If F has a jump at a, i.e., $F(a-) < F(a)$, then, for any u in $(F(a-), F(a)]$, define $F^{-1}(u) = a$.

Remark 4.3 In case (a), the existence of x follows from the Intermediate Value Theorem, and the uniqueness results from the strictly-increasing assumption. In this case, $F(F^{-1}(u)) = u$ for all u in (0,1). In case (b), where F has a jump at a, note that $F(F^{-1}(u)) = F(a) > u$ for u in $(F(a-), F(a))$, so F^{-1} is not the standard inverse function.

The inversion method returns $F^{-1}(U)$, where $U \sim \text{Unif}(0,1)$, and determined by a pseudo-random-number generator. We now show the method's correctness.

Proposition 4.4 *Let F be a cdf and let F^{-1} be its inverse. If $U \sim \text{Unif}(0,1)$, then $F^{-1}(U)$ has cdf F.*

Proof. The key fact is the equivalence

$$F(x) \geq u \Leftrightarrow x \geq F^{-1}(u) \tag{4.1}$$

which is a consequence of the fact that F is a nondecreasing, right-continuous function. We omit a detailed proof of (4.1). Now

$$\begin{aligned} \mathbb{P}(F^{-1}(U) \leq x) &= \mathbb{P}(U \leq F(x)) \qquad \text{by (4.1)} \\ &= F(x) \qquad \text{since } U \sim \text{Unif}(0,1) \end{aligned}$$

i.e., $F^{-1}(U)$ has cdf F. $\square$

Remark 4.5 On a computer, provided only that we can evaluate $F(x)$ at any x, $F^{-1}(u)$ can easily be computed for any u, via a *bracketing/bisection method*, for example. This method is simple, but beyond our scope.

4.2.1 Calculating the Inverse Explicitly: Examples

Inversion of a discrete cdf. Section 1.3.1 explained that a discrete distribution can always be reduced to *ordered* support points $x_1 < x_2 < x_3 < \ldots$, each x occurring with positive probability, and the corresponding cdf F satisfies $F(x_1) < F(x_2) < F(x_3) < \ldots$. Here, the inverse of F is (Definition 4.2(b)):

$$F^{-1}(u) = \begin{cases} x_1 & 0 < u \leq F(x_1) \\ x_2 & F(x_1) < u \leq F(x_2) \\ \vdots & \\ x_k & F(x_{k-1}) < u \leq F(x_k) \\ \vdots & \end{cases}$$

The inverse of some simple continuous cdf's is calculated explicitly below.

Example 4.6 Consider the Unif$(5, 8)$ distribution. By (1.10), the cdf is

$$F(x) = \frac{x-5}{3}, \quad 5 \le x \le 8.$$

Definition 4.2(a) applies on the entire support, so solve

$$\frac{x-5}{3} = u \Leftrightarrow x = 5 + 3u = F^{-1}(u).$$

Example 4.7 In Example 1.6 we found that $X =$"equiprobable outcome on $[2, 4] \cup [5, 6]$" has cdf

$$F(x) = \begin{cases} \frac{x-2}{3}, & 2 \le x \le 4 \\ \frac{2}{3}, & 4 < x \le 5 \\ \frac{2}{3} + \frac{x-5}{3}, & 5 < x \le 6 \end{cases}$$

This cdf is continuous, so its inverse is found by solving $F(x) = u$ (Definition 4.2(a)). For u in $(F(2), F(4))$, x must be between 2 and 4, so solve

$$\frac{x-2}{3} = u \Leftrightarrow x = 2 + 3u = F^{-1}(u).$$

For u in $(F(5), F(6))$, x must be between 5 and 6, so solve

$$\frac{2}{3} + \frac{x-5}{3} = u \Leftrightarrow x = 3 + 3u = F^{-1}(u).$$

Finally, $F^{-1}(2/3) = 4$. In summary,

$$F^{-1}(u) = \begin{cases} 2 + 3u, & u \le \frac{2}{3} \\ 3 + 3u, & u > \frac{2}{3} \end{cases}$$

Example 4.8 Consider the Expon(λ) distribution. Recall that the cdf is

$$F(x) = 1 - e^{-\lambda x}, \quad x \ge 0.$$

Definition 4.2(a) applies on the entire support, so solve

$$1 - e^{-\lambda x} = u \Leftrightarrow x = -\frac{1}{\lambda} \log(1 - u) = F^{-1}(u).$$

♣ Study the following. • Exercise 1 (part (b) = inversion of Geometric distribution = E11 3(a)). • E04 4(i). Triangular distribution with support parametrised by some α. Helps emphasise that in solving $F(x) = u$, only x in the support are relevant. Failing to apply this restriction, as $F(x)$ is a quadratic, there are two solutions, and the one outside the support is irrelevant. • E03 4 (inversion part, first 5 lines). • E10 2(c) (mixture of exponentials with disjoint support).

4.3 Acceptance-Rejection

4.3.1 Method and Theory

The problem is to sample from a given distribution with pdf f and support $\mathcal{S}$. That is, the output X should satisfy, for all $x \in \mathcal{S}$,

$$\lim_{\epsilon \to 0+} \frac{\mathbb{P}(x - \epsilon < X \leq x + \epsilon)}{2\epsilon} = f(x), \quad \text{or, less precisely, } \mathbb{P}(X \in \mathrm{d}x) = f(x)\mathrm{d}x \quad (4.2)$$

where $\mathrm{d}x$ means an infinitesimal (i.e., arbitrarily small) interval containing x.

The acceptance-rejection (A/R) method samples from another pdf, g (it is assumed known how to sample from g) and rejects samples in a way so that accepted samples have the desired pdf, f. There is the key requirement that there exists a *finite* constant K such that

$$a(x) := \frac{f(x)}{Kg(x)} \leq 1 \quad \text{for all } x \in \mathcal{S} \quad (4.3)$$

The need for this is explained in Remark 4.9 below. The sampling works as follows:

1. Sample $Y \sim g$ and $U \sim \mathrm{Unif}(0, 1)$, independently of any other samples.

2. If $U \leq a(Y)$, set $X \leftarrow Y$ (accept) and exit. Otherwise (reject), return to step 1.

Note that the output X is the Y conditioned by the *acceptance* event

$$A = \{U \leq a(Y)\}.$$

g is called the *instrumental (trial, candidate, proposal)* pdf, and Kg the *envelope*.

That X has pdf f can be seen roughly as follows:

$$\begin{aligned} \mathbb{P}(X \in \mathrm{d}x) &= \mathbb{P}(Y \in \mathrm{d}x | A) = \frac{\mathbb{P}((Y \in \mathrm{d}x) \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A | Y \in \mathrm{d}x)\, \mathbb{P}(Y \in \mathrm{d}x)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}\left(U \leq a(Y) \middle| Y \in \mathrm{d}x\right) g(x)\mathrm{d}x}{\mathbb{P}(A)} = \frac{a(x)g(x)\mathrm{d}x}{\mathbb{P}(A)} = \frac{f(x)}{K\mathbb{P}(A)}\mathrm{d}x. \end{aligned}$$

For simplicity, assume the above is a pdf; then it must integrate to one, so

$$1 = \int_{\mathcal{S}} \frac{f(x)}{K\mathbb{P}(A)}\mathrm{d}x = \frac{1}{K\mathbb{P}(A)} \underbrace{\int_{\mathcal{S}} f(x)\mathrm{d}x}_{=1} = \frac{1}{K\mathbb{P}(A)}$$

(as the pdf f has support $\mathcal{S}$). That is, the "less precisely" form in the right of (4.2) is true. Moreover, we see that

$$\mathbb{P}(A) = \frac{1}{K} \quad (4.4)$$

and that $K \geq 1$ always.

Remark 4.9 The step $\mathbb{P}(U \leq a(Y)|Y \in \mathrm{d}x) = a(x)$ in the proof holds only if $a(x) \leq 1$ for all x (since $a(x) > 1$ cannot be a probability). This is why (4.3) is required.

A more careful proof of correctness of A/R shows the left of (4.2), as follows.

Lemma 4.10 *Assume that $a()$ satisfies*

$$|Y - x| \leq \epsilon \quad \Rightarrow \quad a(x) - M\epsilon \leq a(Y) \leq a(x) + M\epsilon$$

where $M < \infty$ can be taken as the maximum of the absolute derivative (slope) of $a()$. Then

$$\lim_{\epsilon \to 0^+} \frac{\mathbb{P}(|Y - x| \leq \epsilon | A)}{2\epsilon} = \frac{f(x)}{K\mathbb{P}(A)} \tag{4.5}$$

provided that $f(x) > 0$ and $g(x) < \infty$, hence $a(x) > 0$.

Proof. Write the conditional probability in (4.5) as $\mathbb{P}(B)/\mathbb{P}(A)$, where

$$B = \{|Y - x| \leq \epsilon, U \leq a(Y)\}.$$

The main idea is to bound the event B by a subset and a superset whose probabilities tend to the correct limit. The bounds are

$$\{|Y - x| \leq \epsilon, U \leq a(x) - M\epsilon\} \subset B \subset \{|Y - x| \leq \epsilon, U \leq a(x) + M\epsilon\}.$$

Taking probabilities,

$$\begin{aligned}
\mathbb{P}(|Y - x| \leq \epsilon, U \leq a(x) - M\epsilon) \quad &\leq \mathbb{P}(B) \leq \quad \mathbb{P}(|Y - x| \leq \epsilon, U \leq a(x) + M\epsilon) \\
\Rightarrow \mathbb{P}(|Y - x| \leq \epsilon)\mathbb{P}(U \leq a(x) - M\epsilon) \quad &\leq \mathbb{P}(B) \leq \quad \mathbb{P}(|Y - x| \leq \epsilon)\mathbb{P}(U \leq a(x) + M\epsilon) \\
\Rightarrow \mathbb{P}(|Y - x| \leq \epsilon)[a(x) - M\epsilon] \quad &\leq \mathbb{P}(B) \leq \quad \mathbb{P}(|Y - x| \leq \epsilon)[a(x) + M\epsilon]
\end{aligned}$$

by the independence of Y and U, and by then using that $U \sim \mathrm{Unif}(0, 1)$. Dividing by $2\epsilon\mathbb{P}(A)$ throughout, we have lower and upper bounds for our target:

$$\frac{\mathbb{P}(|Y - x| \leq \epsilon)[a(x) - M\epsilon]}{2\epsilon\mathbb{P}(A)} \leq \frac{\mathbb{P}(B)}{2\epsilon\mathbb{P}(A)} \leq \frac{\mathbb{P}(|Y - x| \leq \epsilon)[a(x) + M\epsilon]}{2\epsilon\mathbb{P}(A)}. \tag{4.6}$$

Now, take limits as $\epsilon \to 0^+$ and observe:

- $\lim_{\epsilon \to 0^+} \mathbb{P}(|Y - x| \leq \epsilon)/2\epsilon = g(x)$, as Y has pdf g;
- $\lim_{\epsilon \to 0+}[a(x) - M\epsilon] = \lim_{\epsilon \to 0+}[a(x) + M\epsilon] = a(x) > 0$.

Thus, both the lower and upper bound in (4.6) converge to $g(x)a(x)/\mathbb{P}(A) = f(x)/[K\mathbb{P}(A)]$, and thus so does the middle. $\square$

4.3.2 Feasibility and Efficiency

Each trial is accepted with probability $\mathbb{P}(A) = \frac{1}{K}$, independently of other trials. Thus, the number of trials until acceptance has the Geometric($\mathbb{P}(A)$) distribution, whose mean is $1/\mathbb{P}(A) = K$. Let us think "maximum sampling efficiency" equals "minimum mean number of trials until acceptance", i.e., "maximum $\mathbb{P}(A)$", i.e., "minimum K". The constraint (4.3) on K can be rewritten as $K \geq \sup_{x\in\mathcal{S}} \frac{f(x)}{g(x)} = \max_{x\in\mathcal{S}} \frac{f(x)}{g(x)}$, so the minimum K that satisfies the constraint is

$$K = \sup_{x\in\mathcal{S}} \frac{f(x)}{g(x)} = \max_{x\in\mathcal{S}} \frac{f(x)}{g(x)}. \tag{4.7}$$

The A/R method always requires *setup*, meaning choosing g and then determining K by solving this maximisation problem. A considerable limitation is that if the K in (4.7) is infinite, then A/R is impossible for this f and g; see Example 4.13 below.

♣ Study the following. • E07 4. Develops an A/R sampler for the target density proportional to $x^{\alpha-1}e^{-x}$, $x > 0$ (the Gamma$(\alpha, 1)$ distribution), for the case $\alpha < 1$. Note the density goes to ∞ as $x \to 0$. The problem statement gives the envelope, and the problem is then straightforward. Choosing the envelope was a more subtle problem, which was not asked here, because of the requirement of the existence of a finite K; this was achieved by the envelope choice (up to a proportionality constant) $x^{\alpha-1}$ for $x < 1$, and e^{-x} (the Expon(1) pdf) for $x > 1$. The envelope is sampled by inversion (that is, inverting the associated cdf).

Example 4.11 The pdf to sample from is

$$f(x) = \begin{cases} \frac{2}{25}(6-x) & \text{for } 1 \leq x \leq 6 \\ 0 & \text{otherwise.} \end{cases}$$

This is the Triangular(1,1,6) distribution, the three parameters being minimum, mode, and maximum. Let g be the pdf of the Unif(1,6) distribution), i.e., the constant $1/5$ on $[1, 6]$. As setup, we must find (4.7). Focus on $\frac{f(x)}{g(x)} = \frac{2(6-x)}{5}$ in the support of f, i.e., $[1, 6]$, and see that it attains the maximum value of 2 at $x = 1$. Thus, $K = 2$, $a(x) = (6 - x)/5$, and $\mathbb{P}(A) = 1/K = 1/2$. To see the sampling in action, suppose the sampled (Y, U) pairs are (4.9, 0.3) and (1.7, 0.8). Then calculate:

Y	U	$a(Y)$	Is $U \leq a(Y)$?
4.9	0.3	0.22	No
1.7	0.8	0.86	Yes

Here, two trials were needed until acceptance. The number 1.7 is a random sample from the pdf f.

When f and g involve the exponential function, it is typically easier to optimise $\log(f/g)$ ($\log$ = "natural logarithm", the inverse function to the exponential one) rather than f/g, as the former has simpler derivatives. A maximiser of f/g is also a maximiser of $\log(f/g)$, and vice versa. The following is a simple illustration of this.

Example 4.12 Suppose we want to sample from the pdf $f(x) = (2/\pi)^{1/2} e^{-x^2/2}$ with support $(0, \infty)$. (This is the pdf of $|Z|$, where $|\cdot|$ denotes absolute value and $Z \sim \mathrm{N}(0, 1)$, the standard normal distribution (mean 0, variance 1).) Consider acceptance-rejection with g the pdf of the Expon(1) distribution, i.e., $g(x) = e^{-x}$ on $(0, \infty)$. To find the optimal K, maximise $\log[f(x)/g(x)] = \log(2/\pi)/2 + x - x^2/2$ on $(0, \infty)$. The maximiser is $x = 1$, giving $K = (2\mathrm{e}/\pi)^{1/2}$ and acceptance probability $1/K \approx 0.76$.

We write $f(x) \propto h(x)$ on a given support $\mathcal{S}$ to mean that $f(x) = h(x)/c$, where $c = \int_{\mathcal{S}} h(t)\mathrm{d}t$ does not matter in the argument.

Example 4.13 The Gamma(α, λ) distribution (shape $\alpha > 0$, scale $\lambda > 0$), seen earlier for α integer, has pdf $f(x) \propto x^{\alpha-1} e^{-\lambda x}$ with support $x > 0$. Consider A/R with f and g being Gamma with respective shapes a, b and common scale. Then $f(x)/g(x) \propto x^{a-b}$. If we choose $b < a$, then $\max_{x>0} x^{a-b} = \infty$, since $x^{a-b} \to \infty$ as $x \to \infty$, and A/R is impossible. If we choose $b > a$, then $\max_{x>0} x^{a-b} = \infty$, now because $x^{a-b} \to \infty$ as $x \to 0$, so again A/R is impossible.

A general case where A/R is possible is when the pdf f has support (a, b) with a and b both finite, and moreover f is bounded, i.e., $\max_{a<x<b} f(x) < \infty$. An example of such f is the Beta(α, β) distribution for $\alpha, \beta > 1$, where $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$ with support $(0, 1)$. For any f as above, we can choose $g(x) = 1/(b-a)$ on $[a, b]$, that is, the Unif(a, b) pdf (which is easy to sample from). Then, $f(x)/g(x) = (b-a)f(x)$, and (4.7) gives $K = (b-a)\max_{a<x<b} f(x)$, which is finite because both terms in the product are finite by assumption.

♣ Study the following. • E03 4. Both the support and the density f are bounded, so feasibility of A/R is ensured. • E08 4. A triangular distribution, which is sampled by inversion, forms the envelope of an A/R sampler for a Beta distribution (the cdf of that beta is a cubic polynomial, so inversion would require solving a cubic equation). • Exercise 3. Illustrates an involved A/R setup. We consider a whole family of g's; for each of them we find the usual K as in (4.7); with K now being a function of g, we use the (g, K) pair resulting in the smallest K.

4.4 Exercises

1. Describe in full the inversion method of simulating each of the distributions below.

   (a) Uniform on the interval $(-5, 5)$.

   (b) Suppose we do independent trials, where each trial succeeds with probability p, and let X be the number of trials up to and including the first success. Then X is said to have the Geometric(p) distribution. Derive the cdf of X and then describe how X can be sampled by the inversion method.

2. Based on the definition of the Geometric(p) distribution given above, give a method for simulating it based only on an unlimited supply of Unif$(0, 1)$ pseudo-random numbers. *Hint:* Simulate the success indicators; neither inversion nor acceptance-rejection is needed.

3. We want to sample from the Gamma$(a, 1)$ distribution whose pdf is

$$f_a(x) = \frac{1}{\Gamma(a)} x^{a-1} e^{-x}, \quad x > 0$$

   where $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \mathrm{d}x$. Consider acceptance-rejection methods based on the family of trial pdf's

$$g_\lambda(x) = \lambda e^{-\lambda x},$$

   where λ may depend on a. For given $a > 1$, show that across envelopes of the form Kg_λ, $0 < \lambda < 1$, the choice $\lambda = 1/a$ maximises the acceptance probability. Then state the resulting acceptance probability. *Hint:* By maximising $\frac{f_a(x)}{g_\lambda(x)}$ across x, determine the corresponding best A/R constant, $K = K^*(a, \lambda)$. Then, with a being fixed, minimise $K^*(a, \lambda)$ across λ. It will be easier to work with logarithms of functions when seeking the min or max.

4.5 Solutions to Exercises

1. In each case, we first write down the cdf F explicitly, and then find the inverse F^{-1} explicitly. Having done that, the random sample from F is $F^{-1}(U)$, where U is a Unif$(0, 1)$ (pseudo)-random number.

   (a) Let f be the pdf. The uniform distribution on $[-5, 5]$ has pdf that is a constant $c > 0$ on this interval and 0 elsewhere. Find c:

$$1 = \int_{-\infty}^{\infty} c \, dt = \int_{-5}^{5} c \, dt = 10c \Rightarrow c = \frac{1}{10}.$$

   That is,

$$f(t) = \begin{cases} \frac{1}{10}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

   Thus, the cdf is

$$F(x) = \int_{-\infty}^{x} f(t) \, dt = \int_{-5}^{x} \frac{1}{10} \, dt = \frac{x+5}{10} \quad \text{for } -5 \le x \le 5.$$

Find the inverse by solving for x:

$$\frac{x+5}{10} = u \Leftrightarrow x = -5 + 10u = F^{-1}(u).$$

(b) (i) Find the cdf. To this end, the easiest way is:

$$\mathbb{P}(X > k) = \mathbb{P}(\text{first } k \text{ trials failed}) = (1-p)^k, \quad k = 0, 1, 2, \ldots.$$

The cdf, F, is

$$F(k) = \mathbb{P}(X \leq k) = 1 - \mathbb{P}(X > k) = 1 - (1-p)^k, \quad k = 0, 1, 2, \ldots.$$

(ii) To state the inverse explicitly, we need to find the smallest integer k satisfying

$$1 - (1-p)^k \geq u \Leftrightarrow (1-p)^k \leq 1 - u \Leftrightarrow k \log(1-p) \leq \log(1-u) \Leftrightarrow k \geq \frac{\log(1-u)}{\log(1-p)}.$$

(In the last step, the division by $\log(1-p) < 0$ changed the direction of inequality.) Thus, $F^{-1}(u) = \left\lceil \frac{\log(1-u)}{\log(1-p)} \right\rceil$, where $\lceil x \rceil$ is the standard ceiling function (smallest integer that is $\geq x$). Inversion returns $F^{-1}(U)$, where $U \sim \text{Unif}(0,1)$.

2. Let $U_i \sim \text{Unif}(0,1)$, $i = 1, 2, \ldots$, and independent (determined by a pseudo-random-number generator). Return the first i such that $U_i \leq p$.

3. For fixed a and λ, the smallest possible K is $K^*(a,\lambda) = \max_{x:x>0} r_{a,\lambda}(x)$, where

$$r_{a,\lambda}(x) := \frac{f_a(x)}{g_\lambda(x)} = \frac{x^{a-1}e^{(\lambda-1)x}}{\lambda\Gamma(a)}, \quad x > 0$$

Maximise $\log r_{a,\lambda}(x)$ over the support of f_a, i.e., $x > 0$:

$$\begin{aligned}
\log r_{a,\lambda}(x) &= (a-1)\log x + (\lambda-1)x - \log[\lambda\Gamma(a)] \\
\frac{\mathrm{d}\log r_{a,\lambda}}{\mathrm{d}x} &= \frac{a-1}{x} + (\lambda-1) = 0 \Rightarrow x^*(a,\lambda) = \frac{a-1}{1-\lambda} \\
\frac{\mathrm{d}^2}{\mathrm{d}x^2}\log r_{a,\lambda} &= -(a-1)\frac{1}{x^2} < 0.
\end{aligned}$$

Thus the $x^*(a,\lambda)$ above is the maximiser. (Note that the given constraints on a and λ imply $x^* > 0$, which is in the support. If it was not in the support, it would be irrelevant; it is the maximum over the support that matters.)

For fixed a, maximising the acceptance probability is equivalent to minimising the A/R constant $K^*(a,\lambda)$ over λ; first, form the logarithm:

$$\begin{aligned}
&\log K^*(a,\lambda) = \log r_{a,\lambda}(x^*) \\
&= (a-1)\log(a-1) - (a-1)\log(1-\lambda) + (1-a) - \log\lambda - \log\Gamma(a).
\end{aligned}$$

Now minimise this over $0 < \lambda < 1$:

$$\begin{aligned}
\frac{\partial \log K^*}{\partial\lambda} &= \frac{a-1}{1-\lambda} - \frac{1}{\lambda} = 0 \Rightarrow \lambda^*(a) = \frac{1}{a} \\
\frac{\partial^2 \log K^*}{\partial^2\lambda} &= \frac{a-1}{(1-\lambda)^2} + \frac{1}{\lambda^2} > 0.
\end{aligned}$$

Thus $\lambda^*(a)$ above is the minimiser. Putting $K^*(a) = K^*(a, \lambda^*)$, i.e., the constant associated to the best trial pdf, $g_{\lambda^*(a)}$, we have

$$\begin{aligned}
&\log K^*(a) := \log K^*(a, \lambda^*(a)) \\
&= (a-1)\log(a-1) - (a-1)\log\left(\frac{a-1}{a}\right) + (1-a) + \log a - \log\Gamma(a) \\
&= a\log a + (1-a) - \log\Gamma(a) \quad \text{after cancellations}
\end{aligned}$$

i.e., $K^*(a) = a^a e^{1-a}/\Gamma(a)$. The acceptance probability is, as always, $1/K^*(a)$.

Chapter 5

Guidance for Exam Preparation

Questions 1 and 2 are compulsory. Of Questions 3 and 4, only the best answer will count.
Generally, it is good to state clearly any assumptions made or missing information so as to show what you know, even if incomplete.
More detail is given below for each question. Listed is the material (including past exam papers) that can be given higher priority. "E04 4(i)" means "Exam of June 2004, Question 4(i)", and so on.

Question 1 (a): 9 points. The Exponential Distribution and Section 2.2, particularly the memoryless property (2.4), its derivation and significance. Exam E11 4(a).

Question 1 (b): Modelling, 31 points. A modelling question, typically involving the following.

1. Identify an appropriate CTMC, its generator, and the equations that determine the stationary distribution, assuming one exists. A classic model is a birth-death process with finite or infinite state space.

2. Given the stationary distribution of this CTMC, calculate performance by applying some or all of (3.13), (3.15), (3.17), and Little's Law in the form $L = \lambda W$.

Examples 3.2 and 3.4 and Application 3.9. Chapter-3 Exercise 2. WORKING
E03 2: Two Stations in Tandem. E08 3: Finite-state-space CTMC; the policy of preferring A over B together with the question "find the long-run fraction of time A is busy" suggest that the state should indicate the idle/busy state of each server separately. E04 1: M/M/1 versus an M/M/2 with half-as-fast servers, i.e., same overall speed. E06 1: Centralised versus decentralised M/M/ systems. E08 1: M/M/1. E11 1: M/M/2. Chapter-3 Exercise 3: A comparison of M/M/1 versus M/M/2 involving waiting and server costs.
WORKING - END

The following problems should be studied only with regard to obtaining the balance equations. The explicit stationary distribution (solution to these equations) via probability generating functions (pgf's) is not examinable. E04 2, M/E_k/1 Model. Exercise 6, E06 3, E07 2, E10 3: batch arrivals.

Question 2: Chapter 4, 26 points.

1. Section 4.2, especially Definitions 4.1 and 4.2 and Proposition 4.4. Exams E10 2(a)-(b), E11 3(b).

2. Applying the inversion method. Exercise 1. Exams E03 4, E04 4(i), E07 4, E08 4, E10 2(c), E11 2(a), E11 3(a).

3. Section 4.3. Exam E11 3(c).

4. Applying the acceptance-rejection method. Exercise 3. Exams E03 4, E06 4(b) excluding the "discuss" part, E07 4, E08 4, E09 2(b), E10 4(a), E11 2(b).

Question 3: Chapter 2, 34 points. Emphasis on definitions and on analysis similar or identical to the notes. Study in priority order:

1. The Poisson-process basics: Definitions 2.4 and 2.5, calculations as in Example 2.7 and Exercise 1. Proof of "⇒" part in Theorem 2.6, which essentially is the proof of (2.11). Section 2.4.2. Exams E04 3, E07 3 (first 15 points).

2. Merging of Poisson processes, Section 2.4.4, especially Proposition 2.10 and Exercise 2. Poisson process of general rate function, Section 2.5. Exercises 3, 4. To solve E03 3, recognise it is an NHPP and identify the rate function; the solution is then standard, obtained exactly as in Exercise 3. Exam E10 4(b).

3. Convergence as in Section 2.7.

Typical partial question:

> Consider a certain randomly occurring event, and let N_t be the number of events that occur up to time t. State appropriate conditions about the N_t such that $(N_t : t \geq 0)$ is a Poisson process. The conditions are about: (i) independence; (ii) stationarity, meaning that certain distributions do not depend on time; and (iii) the probabilities that one and two events occur, respectively, in an interval of length h, as $h \to 0$. (11 points)

Question 4: Chapter 3, 34 points. Emphasis on definitions and on analysis similar or identical to the notes. Study in priority order:

1. Continuous-Time Markov chains, Section 3.3. The birth-death process, Section 3.2. Derivation of Kolmogorov-type differential equations (3.6), or something similar. Exam E06 2.

2. Long-run behaviour, Sections 3.4 and 3.5, especially Section 3.5.1. Exams E03 1, E07 3 (last 10 points), E11 4(b).

3. Careful statement of Little's Law, as in Theorem 3.8.

Typical partial question:

Put X_t for the number of jobs in a system at time t, and assume that for $h \geq 0$ we have $X_{t+h} = X_t + B - D$, where, conditional on $X_t = i$, the random variables B and D (counts of new "births" and "deaths" during $(t, t+h]$, respectively) are independent with

$$\mathbb{P}(B = 1|X_t = i) = \lambda_i h + o(h), \quad \mathbb{P}(B \geq 2|X_t = i) = o(h)$$

$$\mathbb{P}(D = 1|X_t = i) = \mu_i h + o(h), \quad \mathbb{P}(D \geq 2|X_t = i) = o(h), \quad i > 0.$$

1. Determine, showing all your work,

$$\lim_{h \to 0} \frac{\mathbb{P}(X_{t+h} = i + 1|X_t = i)}{h}$$

for any i. (8 points)

2. Determine, showing all your work,

$$\lim_{h \to 0} \frac{\mathbb{P}(X_{t+h} = j|X_t = i)}{h}$$

for $|j - i| \geq 2$, i.e., the target state j differs from the starting state i by 2 or more. (9 points)