X

Steady-State Properties of $GI/G/1$

1 Notation. The Actual Waiting Time

We consider the (FIFO) $GI/G/1$ queue in the notation of III.1b. That is, the customers are numbered $n = 0, 1, 2, \ldots$, U_n is the service time of n, T_n the time between the arrivals of n and $n+1$ and $A(x) = \mathbb{P}(T_n \le x)$ is the interarrival distribution, $B(x) = \mathbb{P}(U_n \le x)$ the service-time distribution (we assume $A(0) = \mathbb{P}(T_n = 0) = 0$, $B(0) = \mathbb{P}(U_n = 0) = 0$). We let $\mu_A = \mathbb{E}T_n$ denote the interarrival mean and $\mu_B = \mathbb{E}U_n$ the mean service time (μ_A, μ_B are assumed finite throughout). Then $\rho = \mu_B/\mu_A$ is the traffic intensity. Unless otherwise stated, *it is assumed that customer* 0 *has just arrived at time* $t = 0$ *to an empty queue*.

Some basic tools in the analysis of the system are: random walks that yield information on the waiting–time distribution; regenerative processes that permit conclusions to be made on the existence of limits of other functionals such as queue lengths; and rate conservation that will provide relations between the limits and in particular express the distributions of workload and queue size in terms of the waiting-time distribution.

Some of the basic facts on the waiting times have already been touched upon, but will now be put together. Define $X_n = U_n - T_n$, $\mu = \mathbb{E}X_n = \mu_B - \mu_A$, $S_0 = 0$, $S_n = X_0 + \cdots + X_{n-1}$, $M_n = \max_{0 \le k \le n} S_k$, $M = \max_{0 \le k < \infty} S_k$. Then the cases $\mu < 0$, $\mu = 0$ and $\mu > 0$ correspond to $\rho < 1$, $\rho = 1$, resp. $\rho > 1$, and III.6 yields:

Proposition 1.1 *The* (actual) *waiting time process* $\{W_n\}$ *is a Lindley process generated by* $\{S_n\}$, *i.e.* $W_{n+1} = (W_n + X_n)^+$. *In particular,*

$$W_n = \max\bigl(S_n, S_n - S_1, \ldots, S_n - S_{n-1}, 0\bigr) \tag{1.1}$$

$$\stackrel{\mathscr{D}}{=} M_n \tag{1.2}$$

and if $\rho < 1$, *then a limiting steady–state distribution exists and is given by* $\mathbb{P}_e(W_n \le x) = \mathbb{P}(M \le x)$.
[The formulas (1.1) and (1.2) require slight variants for $W_0 \neq 0$, cf. III.6. However, the limit result still holds true.]

Our interest in the following is centered around the so–called *stable case* $\rho < 1$ and we shall only briefly as a digression indicate the typical behaviour for $\rho \ge 1$.

Proposition 1.2 (i) *If* $\rho = 1$, $\sigma^2 = \mathbb{V}arX_n < \infty$, *then the limiting distribution of* $W_n/\sqrt{n}$ *exists and is that of the absolute value of a normal r.v. with mean zero and variance* σ^2; (ii) *if* $\rho > 1$, *then* $W_n/n \stackrel{\text{a.s.}}{\to} \mu = \mu_A(\rho - 1)$.

Proof. In case (i), it is well known that $M_n/\sqrt{n}$ has the asserted limit properties (the easiest proof is presumably by Donsker's theorem, Billingsley, 1968, Ch. 2; for a direct proof, see Chung, 1974, pp. 217–222). In case (ii), we have $S_n/n \stackrel{\text{a.s.}}{\to} \mu > 0$. Hence by (1.1), $W_n > 0$ eventually. Hence if η is the last n with $W_n = 0$, we have $W_n = S_n - S_\eta$, $n \ge \eta$, from which we get $W_n/n \sim S_n/n \sim \mu$. □

Now define $\sigma(0) = 0$, $\sigma = \inf\{n \ge 1 : W_n = 0\}$, $\sigma(1) = \sigma$, $\sigma(k+1) = \inf\{n > \sigma(k) : W_n = 0\}$. Since $W_0 = 0$, we may interpret σ as the number of customers served in the first busy period and $\sigma(k)$ as the index of the customer initiating the kth busy cycle.

Proposition 1.3 *The* $\sigma(k)$ *are regeneration points for the waiting–time process. We have* $\mathbb{P}(\sigma < \infty) = 1$ *if and only if* $\rho \le 1$. *Hence for* $\rho \le 1$, $\{W_n\}$ *is aperiodic regenerative with imbedded renewal sequence* $\{\sigma(k)\}$. *Furthermore,* $\sigma = \sigma(1)$ *coincides with the weak descending ladder epoch,* $\sigma = \tau_- = \inf\{n \ge 1 : S_n \le 0\}$. *We have*

$$W_n = S_n = U_0 + \cdots + U_{n-1} - T_0 - \cdots - T_{n-1}, \quad n = 0, \ldots, \sigma - 1, \tag{1.3}$$

$$-S_\sigma = -S_{\tau_-} = I \tag{1.4}$$

where I *is the idle period corresponding to the first busy cycle, and furthermore* $\mathbb{E}\sigma < \infty$ *if and only if* $\rho < 1$.

Proof. By the Lindley process property, we have $W_n = S_n$, $n = 0, \ldots, \sigma - 1$, and this makes it clear that $\sigma = \tau_-$. Also I is the amount by which the last interarrival time exceeds the residual work at the time of the last arrival in the cycle,

$$I = T_{\sigma-1} - (W_{\sigma-1} + U_{\sigma-1}) = -S_\sigma = -S_{\tau_-}.$$

It is clear that the $\sigma(k)$ are regeneration points, and by general random walk results we have finally $\sigma = \tau_- < \infty$ a.s. if and only if $\mu = \mathbb{E}X_n \le 0$, i.e. $\rho \le 1$, and $\mathbb{E}\tau_- = \mathbb{E}S_{\tau_-}/\mu < \infty$ if and only if $\mu < 0$, i.e. $\rho < 1$. Finally aperiodicity follows from $\mathbb{P}(\sigma = 1) = \mathbb{P}(U \le T) > 0$. □

For the sake of easy reference, some of the main r.v.'s occurring in the rest of the chapter will now be introduced.

Definition 1.4 *Suppose $\rho < 1$. Then throughout this chapter:*
(i) *W will denote a random variable having the steady–state distribution H, say, of W_n, $H(x) = \mathbb{P}(W \le x) = \mathbb{P}_e(W_n \le x)$; similarly,*
(ii) *V, Q have the steady state distributions of the workload V_t, resp. the queue length Q_t* (which will be shown to exist if the interarrival distribution A is nonlattice)*;*
(iii) *Q_n^A, Q_n^D denote the queue length just prior to the nth arrival and just after the nth departure, and Q^A, Q^D the corresponding steady–state quantities;*
(iv) *U, T, X, $\{T^{(k)}\}_0^\infty$ have the distributions of U_n, T_n, $X_n = U_n - T_n$, $\{T_0 + \cdots + T_{k-1}\}_0^\infty$, respectively, and are mutually independent and independent of W, V, Q, Q^A, etc.; similar conventions apply for*
(v) *U^*, T^* having densities $\mathrm{d}B_0(x)/\mathrm{d}x = \overline{B}(x)/\mu_B$ and $\mathrm{d}A_0(x)/\mathrm{d}x = \overline{A}(x)/\mu_A$.*

The distributions B_0, A_0 are familiar from renewal theory, V.3. Also, from the independence of W_n and U_n, it is seen that we may identify $W + U$ by the sojourn time in the steady state.

A main problem for the study of the actual waiting time is obviously to study the distribution of W. Various expressions are available for $H(x) = \mathbb{P}(W \le x)$. From Proposition 1.1 and VIII.2.2 we have

$$H(x) \;=\; \big(1 - \|G_+\|\big)U_+(x) \;=\; \big(1 - \|G_+\|\big)\sum_{n=0}^{\infty} G_+^{*n}(x), \tag{1.5}$$

whereas Proposition 1.3 and VI.(1.5) yield

$$H(x) \;=\; \frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1} I(W_n \le x) \;=\; \frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x)\,. \tag{1.6}$$

These formulas are, however, not intrinsically different in view of VIII.2.3(b). A somewhat different characterization of H is as the unique solution to Lindley's integral equation III.(6.6) with

$$F(x) \;=\; \mathbb{P}(X_n \le x) \;=\; \mathbb{P}(U_n - T_n \le x) \;=\; \int_0^\infty B(x+y)\,A(\mathrm{d}y), \quad x \in \mathbb{R}.$$

Also, the characteristic function has been found in VIII.4 but is obviously quite complicated.

The representation (1.5) will turn out to be particularly useful when combined with VIII.3.1(b) stating that

$$G_+(A) \;=\; U_- * F(A) \;=\; \int_{-\infty}^{0} F(A-x)\, U_-(\mathrm{d}x), \quad A \subseteq (0,\infty). \tag{1.7}$$

Proposition 1.5 *$W \stackrel{\mathscr{D}}{=} (W+X)^+$, whereas the conditional distribution of $(W+X)^-$ given $W+X \le 0$ coincides with the common distribution of $-S_{\tau_-}$ and I. In particular, for $f : [0,\infty) \to [0,\infty)$*

$$\mathbb{E}f\big((W+X)^-\big) \;=\; \frac{\mathbb{E}f(-S_{\tau_-})}{\mathbb{E}\tau_-} \;=\; -\mathbb{E}X\frac{\mathbb{E}f(I)}{\mathbb{E}I}, \tag{1.8}$$

$$\mathbb{E}(W+X)^- \;=\; -\mathbb{E}X. \tag{1.9}$$

Proof. The first statement was noted previously in III.6.6 and yields in particular

$$\mathbb{P}(W+X \le 0) \;=\; \mathbb{P}\big((W+X)^+ = 0\big) = \; \mathbb{P}(W=0) \;=\; 1 - \|G_+\| \;=\; 1/\mathbb{E}\tau_-,$$

cf. VIII.2.3(c). Also, by VIII.3.2(b),

$$\begin{aligned}
&\mathbb{E}f\big((W+X)^-\big) \\
&\quad= \int_{-\infty}^{0} f(-x)\, H * F(\mathrm{d}x) \;=\; \big(1 - \|G_+\|\big) \int_{-\infty}^{0} f(-x)\, U_+ * F(\mathrm{d}x) \\
&\quad= \big(1 - \|G_+\|\big) \int_{-\infty}^{0} f(-x)\, G_-(\mathrm{d}x) \;=\; \frac{1}{\mathbb{E}\tau_-}\mathbb{E}f(-S_{\tau_-}) \\
&\quad= \mathbb{P}(W+X \le 0)\mathbb{E}f(-S_{\tau_-}).
\end{aligned}$$

Recalling $\mathbb{E}S_{\tau_-} = \mathbb{E}\tau_-\mathbb{E}X$ and (1.4), the proof is complete. □

Problems

1.1 Give a direct proof of (1.9) by using $W + X = (W+X)^+ - (W+X)^-$.

2 The Moments of the Waiting Time

The problem is to study conditions for the existence of $\mathbb{E}W^p$, $p > 0$, and, as far as possible, to derive an explicit expression. In view of $W \stackrel{\mathscr{D}}{=} M$, this is really a random walk problem (as in the case for many other aspects of the behaviour of the waiting time, cf. e.g. Sections 6 and 7) and can therefore be formulated in that setting alone. The queueing interpretation may, however, require some slight reformulations: for example, in the following existence result, $\mathbb{E}(X^+)^{p+1} = \mathbb{E}\big((U-T)^+\big)^{p+1} < \infty$ is readily seen to be equivalent to $\mathbb{E}U^{p+1}$, whereas $\mathbb{E}X^- < \infty$ is automatic in view of $\mathbb{E}T = \mu_A < \infty$.

Theorem 2.1 *Consider a random walk with $\mu = \mathbb{E}X < 0$ and let $p > 0$. Then $\mathbb{E}M^p < \infty$ provided that $\mathbb{E}(X^+)^{p+1} < \infty$. Conversely, if $\mathbb{E}M^p < \infty$ and $\mathbb{E}\,X^- < \infty$, then $\mathbb{E}(X^+)^{p+1} < \infty$.*

Proof. We first note that the pth moment ν_n of a sum $Y_1 + \cdots + Y_n$ of nonnegative i.i.d. summands with $\mathbb{E}Y_1^p < \infty$ is $\mathrm{O}(n^p)$. Indeed, if $p \le 1$ Jensen's inequality gives $\nu_n \le (n\mathbb{E}Y)^p$, whereas for $p \ge 1$ we have

$$\nu_n^{1/p} \;=\; \big[\mathbb{E}(Y_1 + \cdots + Y_n)^p\big]^{1/p} \;=\; \|Y_1 + \cdots + Y_n\|_p \;\le\; n\|Y\|_p.$$

Hence if $\alpha = \mathbb{E}\big[S_{\tau_+}^p;\, \tau_+ < \infty\big] < \infty$,

$$\mathbb{E}M^p \;=\; \big(1 - \|G_+\|\big) \sum_{n=0}^{\infty} \int_0^{\infty} x^p G_+^{*n}(\mathrm{d}x) \;=\; \big(1 - \|G_+\|\big) \sum_{n=0}^{\infty} \|G_+\|^n \mathrm{O}(n^p)$$

will be finite in view of $\|G_+\| < 1$, whereas if $\alpha = \infty$ then the term corresponding to $n = 1$ in the sum is infinite and hence $\mathbb{E}M^p = \infty$.

Write $U(y) = U_-[-y, 0]$. Then by VIII.3.1(b)

$$\begin{aligned} \frac{\alpha}{p} \;&=\; \int_0^{\infty} x^{p-1} \mathbb{P}\big(S_{\tau_+} > x,\, \tau_+ < \infty\big)\, \mathrm{d}x \;=\; \int_0^{\infty} x^{p-1} U_- * F(x, \infty)\, \mathrm{d}x \\ &=\; \int_0^{\infty} F(\mathrm{d}y) \int_0^{y} x^{p-1} U(y - x)\, \mathrm{d}x\,. \end{aligned} \tag{2.1}$$

By the elementary renewal theorem (the proof is valid also if the interarrival distribution has an atom at 0 as G_-) we have for suitable c_1, c_2 that $U(z) \le c_1 + c_2 z$, and since for large y

$$\begin{aligned} &\int_0^{y} x^{p-1} \big[c_1 + c_2(y - x)\big]\, \mathrm{d}x \\ &\quad=\; \frac{1}{p} y^p c_1 + \frac{1}{p} y^{p+1} c_2 - \frac{1}{p+1} y^{p+1} c_2 \;\sim\; \frac{c_2}{p(p+1)} y^{p+1} \end{aligned} \tag{2.2}$$

it follows that $\mathbb{E}(X^+)^{p+1} < \infty$ implies $\alpha < \infty$ and hence $\mathbb{E}M^p < \infty$. Conversely, if $\mathbb{E}X^- < \infty$, then $\mathbb{E}S_{\tau_-} = \mathbb{E}\tau_- \mathbb{E}X > -\infty$ and hence $U(z) \ge d_1 + d_2 z$ with $d_2 > 0$. If $\mathbb{E}M^p < \infty$, then $\alpha < \infty$ and combining (2.1) and (2.2) yields $\mathbb{E}(X^+)^{p+1} < \infty$. □

Not even the moments of M (if they exist) can be found very explicitly. For example, VIII.4.5 and (1.5) yield the expressions

$$\mathbb{E}M \;=\; \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{E}S_n^+ \;=\; \frac{\mathbb{E}[S_{\tau_+};\, \tau_+ < \infty]}{1 - \|G_+\|}\,. \tag{2.3}$$

A further important relation is the following:

Theorem 2.2 *If $\mathbb{E}|X|^{p+1} < \infty$ for some $p = 1, 2, \ldots$, then*

$$\sum_{q=0}^{p} \binom{p+1}{q} \mathbb{E}M^q \mathbb{E}X^{p+1-q} \;=\; \mathbb{E}\big[-(M + X)^-\big]^{p+1} \;=\; \frac{\mathbb{E}S_{\tau_-}^{p+1}}{\mathbb{E}\tau_-}\,. \tag{2.4}$$

(Note that in the queueing setting, we may rewrite the r.h.s. of (2.4) as $(-1)^p \mathbb{E}X\mathbb{E}I^{p+1}/\mathbb{E}I$; cf. (1.8).)

Proof. The last identity in (2.4) follows from (1.8). To show the remaining part of the theorem, first suppose $\mathbb{E}(X^+)^{p+2} < \infty$. Then

$$\mathbb{E}M^{p+1} < \infty, \quad \mathbb{E}\big[(M+X)^-\big]^{p+1} \le \mathbb{E}|X|^{p+1} < \infty,$$

and since $(M+X)^+(M+X)^- = 0$ we get

$$\begin{aligned}
(M+X)^{p+1} &= \big[(M+X)^+ - (M+X)^-\big]^{p+1} \\
&= \big[(M+X)^+\big]^{p+1} + \big[-(M+X)^-\big]^{p+1}, \\
\mathbb{E}(M+X)^{p+1} &= \sum_{q=0}^{p+1} \binom{p+1}{q} \mathbb{E}M^q \mathbb{E}X^{p+1-q} \\
&= \mathbb{E}\big[(M+X)^+\big]^{p+1} + \mathbb{E}\big[-(M+X)^-\big]^{p+1} \\
&= \mathbb{E}M^{p+1} + \mathbb{E}[-(M+X)^-]^{p+1}
\end{aligned}$$

and cancelling $\mathbb{E}M^{p+1}$, (2.4) follows. In the general case, replace X_n by $X_n^{(k)} = X_n \wedge k$ and let $M^{(k)}$ be defined in terms of the $X_n^{(k)}$ rather than the X_n. Then $\mathbb{E}({X^{(k)}}^+)^{p+2} < \infty$, hence

$$\sum_{q=0}^{p} \binom{p+1}{q} \mathbb{E}{M^{(k)}}^q \mathbb{E}{X^{(k)}}^{p+1-q} \;=\; \mathbb{E}[-(M^{(k)} + X^{(k)})^-]^{p+1} \;<\; \infty\,.$$

But clearly, $M^{(k)} \le M$ and $M^{(k)} \uparrow M$ as $k \to \infty$. Hence the desired conclusion follows by monotone convergence as $k \to \infty$. □

Rewriting in queueing notation, we get in particular for the mean waiting time ($p = 1$) that

$$\begin{aligned}
2\mathbb{E}(-X)\mathbb{E}W &= \mathbb{E}X^2 - \mathbb{E}\big[(W+X)^-\big]^2 \;=\; \mathbb{V}arX - \mathbb{V}ar(W+X)^- \\
&= \mathbb{E}X^2 - \frac{\mathbb{E}S_{\tau_-}^2}{\mathbb{E}\tau_-} \;=\; \mathbb{E}X^2 - \frac{\mathbb{E}(-X)\mathbb{E}I^2}{\mathbb{E}I} \qquad (2.5)
\end{aligned}$$

(here the second equality follows from (1.9)). Considerable effort has been put into converting these expressions into bounds or approximations that are more explicit in the sense that only the distribution of X (or U, T) is invoked, and preferably only even the first few moments. We return to the approximations in Section 7 and XIII.6, and here present only some of the roughest bounds,

$$\mathbb{E}U^2 - \mathbb{E}U\mathbb{E}T \;\le\; 2\mathbb{E}(-X)\mathbb{E}W \;\le\; \mathbb{V}ar\,X \;=\; \mathbb{V}ar\,U + \mathbb{V}ar\,T\,. \qquad (2.6)$$

(The lower bound may be negative and hence trivial. The upper bound is in fact sharp in an asymptotic sense; cf. Section 7.) Here the upper bound is obvious from $\mathbb{V}ar(W+X)^- \ge 0$. For the lower bound, rewrite (2.5) as

$$\mathbb{E}U^2 - 2\mathbb{E}U\mathbb{E}T + \mathbb{E}T^2 - \mathbb{E}[(W+X)^-]^2 \;=\; \mathbb{E}U^2 - 2\mathbb{E}U\mathbb{E}T + \mathbb{E}(CD)$$

where $C = T + (W+X)^-$, $D = T - (W+X)^-$. Here

$$D \;=\; T + W + X - (W+X)^+ \;=\; W + U - (W+X)^+$$

so that

$$\mathbb{E}(CD) \;=\; \mathbb{E}\big[T\big(W - (W+X)^+\big)\big] + \mathbb{E}T\mathbb{E}U + \mathbb{E}\big[(W+X)^-(W+U)\big] \quad (2.7)$$

The two last terms in (2.7) are obviously nonnegative, and thus it is sufficient to show that the first one is so too. But $f(T) = T$ and $g(T) = W - (W+U-T)^+$ are both nondecreasing in T for fixed W, U. Hence by a well–known inequality (Problem 2.2)

$$\begin{aligned} \mathbb{E}\big[f(T)g(T)\,\big|\,W,U\big] &\geq \mathbb{E}\big[f(T)\,\big|\,W,U\big]\cdot\mathbb{E}\big[g(T)\,\big|\,W,U\big] \\ &= \mathbb{E}T\cdot\mathbb{E}\big[W-(W+X)^+\,\big|\,W,U\big], \\ \mathbb{E}\big[T\big(W-(W+X)^+\big)\big] &\geq \mathbb{E}T\cdot\mathbb{E}\big[W-(W+X)^+\big] \;=\; \mathbb{E}T\cdot 0 \;=\; 0. \end{aligned}$$

□

Problems

2.1 Consider a random walk with $\|G_+\| = \|G_-\| = 1$. Show that $\mathbb{E}S_{\tau_+} < \infty$, $\mathbb{E}S_{\tau_-} > -\infty$ if and only if $\mathbb{E}X^2 < \infty$, $\mathbb{E}X = 0$, and that then $\mathbb{E}X^2 = -2\mathbb{E}S_{\tau_+}\mathbb{E}S_{\tau_-}$. [*Hint:* Necessity and the stated identity follows by Wiener–Hopf factorization of the ch.f.]

2.2 (CHEBYCHEFF'S COVARIANCE INEQUALITY) Let X be a r.v. and f, g nondecreasing functions. Show that $\mathbb{E}[f(X)g(X)] \geq \mathbb{E}f(X)\mathbb{E}g(X)$ [*Hint:* Reduce to the case $\mathbb{E}f(X) = 0$ and consider $\mathbb{E}[f(X)(g(X) - g(b))]$ where b is the point at which f changes sign.]

2.3 Carry out the last step in the proof of Corollary IX.3.4.

Notes Theorem 2.1 goes back to Kiefer & Wolfowitz (1956) and there are many proofs around. As one of many applications of (2.5), we mention in particular the observation by Minh and Sorli (1983) that when estimating $\mathbb{E}W$ by simulation, the only unknown quantities are $\mathbb{E}I$ and $\mathbb{E}I^2$, and that simulating these rather than $\mathbb{E}W$ increases precision. Bounds for $\mathbb{E}W$ and related quantities are surveyed in Stoyan (1983) and Daley *et al.* (1994). For Problem 2.2, see also Thorisson (2000), p. 2.

3 The Workload

In continuous time, there is a regenerative structure similar to the one in Proposition 1.3: the instants with a customer entering an empty queue are regeneration points. Letting C be the first such instant after $t = 0$ and recalling that we start with customer 0 having just arrived, it is seen that C is just the length of the first busy cycle. Furthermore, $C < \infty$ a.s. is equivalent to $\sigma < \infty$ a.s., i.e. to $\rho \leq 1$ (cf. Proposition 1.3). In fact, there

is a close relation between σ, C and the first busy period G: since precisely the customers $0, 1, \ldots, \sigma - 1$ are served in the first busy period, we have $G = U_0 + \cdots + U_{\sigma-1}$ and the first busy cycle ends at the arrival time $C = T_0 + \cdots + T_{\sigma-1}$ of customer σ. One checks immediately that $\{\sigma \leq n\}$ is independent of $T_n, T_{n+1}, \ldots, U_n, U_{n+1}$, and hence Wald's identity yields the first part of

Proposition 3.1 *Suppose $\rho \leq 1$. Then the mean busy cycle is $\mathbb{E}C = \mu_B \mathbb{E}\sigma$, the mean busy period is $\mathbb{E}G = \mu_A \mathbb{E}\sigma$ and the mean idle period is $\mathbb{E}I = \mathbb{E}C - \mathbb{E}G = -\mu \mathbb{E}\sigma$. Furthermore the mean busy period is nonlattice if and only if the interarrival distribution A is so, and spread out if and only if A is so.*

The second part is often stated to be obvious, but some care is needed (cf. Problem 3.3), and we give the proof (when $\rho < 1$) in the form of the following more general result:

Proposition 3.2 *Let $T_0 > 0, T_1 > 0, \ldots$ be i.i.d. with common distribution A with $\mu_A < \infty$, and let $\sigma \geq 1$ be a random time such that $\mathbb{E}\sigma < \infty$ and $T_n, T_{n+1}, \ldots$ are independent of $\{\sigma \leq n\}$ for all n. Then then distribution K of $C = T_0 + \cdots + T_{\sigma-1}$ is nonlattice if and only if A is so, and spread out if and only if A is so.*

Proof. By Wald's identity, we have $\mathbb{E}C < \infty$. Also by an obvious iterative procedure we may assume that random times $\sigma(1) = \sigma < \sigma(2) < \cdots$ have been constructed such that $\{T_0 + \cdots + T_{\sigma(k)-1}\}$ is a renewal process governed by K. Then, in the obvious notation, the renewal measures satisfy $U_A \geq U_K$. Suppose K was lattice, say aperiodic on $\mathbb{N}$, but A not. Then by Blackwell's renewal theorem,

$$\frac{h}{\mu_A} = \lim_{n\to\infty} \big[U_A(n) - U_A(n-h)\big] \geq \lim_{n\to\infty} \big[U_K(n) - U_K(n-h)\big] = \frac{1}{\mu_K}$$

for all $h < 1$, which is impossible. Similarly, assume that A is spread out but K not. Then U_K is concentrated on a Lebesgue null set N, and Stone's decomposition shows that the U_A–measure of N is finite, whereas the U_K–measure is infinite, contradicting $U_A \geq U_K$.

If, conversely, A is not spread out, then U_A is concentrated on a Lebesgue null set N. Hence U_K is concentrated on N, and K cannot be spread out. That K is lattice if A is so is even more trivial. □

The remaining part of Proposition 3.1 now follows immediately when $\rho < 1$. When $\rho = 1$, replace B by an equivalent (in the sense of null sets) and stochastically smaller distribution $\widetilde{B}$. Then the busy cycle distributions are equivalent, and since $\widetilde{\rho} < 1$, Proposition 3.2 applies to $\widetilde{C}$.

Corollary 3.3 *Suppose $\rho < 1$ and that A is nonlattice. Then a limiting steady–state distribution of the workload V_t exists and is given by*

$$\mathbb{E}f(V) \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C f(V_s)\,\mathrm{d}s. \tag{3.1}$$

If A is spread out, then $V_t \to V$ in total variation.

Proof. For $\rho < 1$, we have $\mathbb{E}\sigma < \infty$. Hence Proposition 3.1 ensures that the basic limit theorems for regenerative processes in VI.1 and VII.1 are applicable. □

As a first application of (3.1), note that the time spent by $\{V_t\}$ in state 0 in the time interval $[0, C)$ is the just the idle period. Thus combining with Proposition 3.1, we get

$$\begin{aligned}\mathbb{P}(V=0) \;&=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C I(V_s=0)\,\mathrm{d}s\\ &=\; \frac{\mathbb{E}I}{\mathbb{E}C} \;=\; \frac{(\mu_A-\mu_B)\mathbb{E}\sigma}{\mu_A\mathbb{E}\sigma} \;=\; 1-\rho\,.\end{aligned} \tag{3.2}$$

[Note that this is always explicit in contrast to $\mathbb{P}(W=0) = 1/\mathbb{E}\sigma$.]

We next express the distribution of V in terms of the steady–state waiting time distribution (for the meaning of U^*, T^*, see Definition 1.4):

Theorem 3.4 *The conditional distribution of V given $V > 0$ is the same as the distribution $H * B_0$ of $W + U^*$. Equivalently,*

$$\mathbb{P}(V \le x) \;=\; 1-\rho+\rho\mathbb{P}(W+U^* \le x) \;=\; 1-\rho+\rho H * B_0(x), \tag{3.3}$$

cf. (3.2). *An alternative characterization is $V \stackrel{\mathscr{D}}{=} (W+U-T^*)^+$.*

Proof. Let $X_t = (V_t - x)^+$. Then $\{X_t\}$ has derivative -1 when $V_t > x$ and 0 otherwise, whereas the jump at the arrival of customer n is $(W_n + U_n - x)^+ - (W_n - x)^+$. Hence the rate conservation law VII.6.6 applied to a stationary version yields

$$\mathbb{P}(V>x) \;=\; \frac{1}{\mu_A}\big[\mathbb{E}(W+U-x)^+ - \mathbb{E}(W-x)^+\big] \;=\; \frac{\mu_B}{\mu_A}\mathbb{P}\big(U^* > (W-x)^-\big)$$

where the last identity follows from

$$\mathbb{E}\big[(U+a)^+ - a^+\big] \;=\; \int_{a-}^{\infty} \overline{B}(u)\,\mathrm{d}u \;=\; \mu_B\mathbb{P}(U^* > a^-) \tag{3.4}$$

(integration by parts) by conditioning upon $a = W - x$. Hence $\mathbb{P}(V > x) = \rho\mathbb{P}\big(U^* > (x-W)^+\big)$ which (since $U^* > 0$) is the same as $\mathbb{P}(V > x) = \rho\mathbb{P}(U^* > x - W)$, $x \ge 0$, and (3.3) follows.

For $V \stackrel{\mathscr{D}}{=} (W+U-T^*)^+$, consider $X_t = \int_t^{M_t} I(V_s > x)\,\mathrm{d}x$ where M_t is the next arrival instant after t. The process $\{X_t\}$ decreases linearly at unit rate on intervals where $V_s > x$, is 0 at the nth arrival instant and then

jumps to $\int_0^{T_n} I(W_n + U_n - s > x)\,\mathrm{d}s$. It follows by rate conservation that for $x \geq 0$

$$\begin{aligned}\mathbb{P}(V > x) &= \frac{1}{\mu_A}\int_0^\infty \overline{A}(s) I(W + U - s > x)\,\mathrm{d}s \;=\; \mathbb{P}(T^* < W + U - x) \\ &= \mathbb{P}(W + U - T^* > x) \;=\; \mathbb{P}\big((W + U - T^*)^+ > x\big).\end{aligned}$$

□

Note that in the $M/G/1$ case, we have $T^* \stackrel{\mathscr{D}}{=} T$ and hence

$$V \;\stackrel{\mathscr{D}}{=}\; (W + U - T^*)^+ \;\stackrel{\mathscr{D}}{=}\; (W + U - T)^+ \;\stackrel{\mathscr{D}}{=}\; W$$

so that we obtain another proof that $V \stackrel{\mathscr{D}}{=} W$ in $M/G/1$, as found already in III.9 and VII.6.

Since $\mathbb{E}U^* = \mathbb{E}U^2/2\mu_B$, it follows also by combining with (3.2) that:

Corollary 3.5 $\mathbb{E}V \;=\; \rho\left\{\dfrac{\mathbb{E}U^2}{2\mu_B} + \mathbb{E}W\right\}$.

It is instructive to consider the following two direct proofs of Corollary 3.5. The first uses rate conservation applied to $X_t = V_t^2$. Here in steady state, $\mathbb{E}X_t' = \mathbb{E}[2V;\, V > 0] = 2\mathbb{E}V$, so by rate conservation $\mathbb{E}V$ is

$$\frac{1}{2\mu_A}\mathbb{E}\big[(W + U)^2 - W^2\big] \;=\; \frac{1}{2\mu_A}\big[\mathbb{E}U^2 + 2\mu_B\mathbb{E}W\big] \;=\; \rho\left\{\frac{\mathbb{E}U^2}{2\mu_B} + \mathbb{E}W\right\}$$

(if $\mathbb{E}U^3 = \infty$ so that $\mathbb{E}W^2 = \infty$, use a truncation argument as in the proof of Theorem 2.2). The second proof uses a sample part decomposition of a regenerative cycle, cf. the partitioning of the subgraph of $\{V_t\}_{0 \leq t < C}$ into triangles and parallelograms in Fig. 3.1.

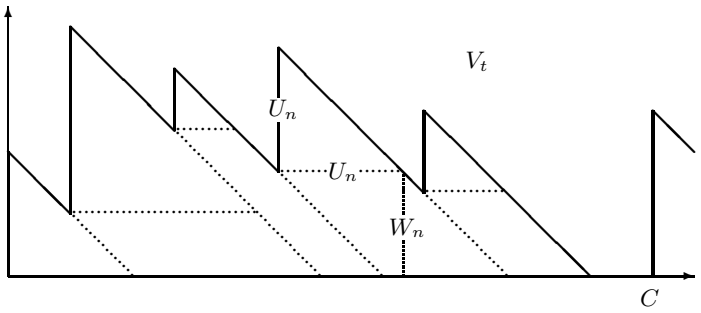

Figure 3.1

The area is $U_n^2/2$ of the nth triangle and W_nU_n of the nth parallelogram, hence

$$\mathbb{E}V \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C V_s\,\mathrm{d}s \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\sigma-1}\big[U_n^2/2 + W_nU_n\big]$$

$$
\begin{aligned}
&= \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\infty}\mathbb{E}\big[U_n^2/2 + W_nU_n;\,\sigma > n \,\big|\, U_k, T_k, k = 0,\dots,n-1\big] \\
&= \frac{1}{\mathbb{E}C}\,\mathbb{E}\sum_{n=0}^{\infty}\mathbb{E}\big[U^2/2 + W_nU;\,\sigma > n\big] \\
&= \frac{1}{\mu_A\mathbb{E}\sigma}\left\{\frac{1}{2}\mathbb{E}U^2\mathbb{E}\sigma + \mu_B\mathbb{E}\sum_{n=0}^{\sigma-1}W_n\right\} \\
&= \rho\left\{\frac{\mathbb{E}U^2}{2\mu_B} + \frac{1}{\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1}W_n\right\} \;=\; \rho\left\{\frac{\mathbb{E}U^2}{2\mu_B} + \mathbb{E}W\right\}.
\end{aligned}
$$

Problems

3.1 Define R_t as the residual service time of the customer being served at time t ($R_t = 0$ if the server is idle). Show that $\mathbb{P}(R \le x) = 1 - \rho + \rho B_0(x)$.
3.2 Let $\{B_t\}$ be the forward recurrence time of the arrival process. Show that $\{(B_t, V_t)\}$ is strong Markov.
3.3 Show that the assumption $\mathbb{E}\sigma < \infty$ of Proposition 3.2 is indispensable. [*Hint:* $T_n = 1 + \theta Z_n$ where $\theta \in (0,1)$ is irrational, $Z_n = \pm 1$ w.p. $1/2$ and $\sigma = \inf\{n \ge 1 :\, Z_0 + \cdots + Z_{n-1} = 0\}$.]

Notes Since rate conservation holds in a more general stationary setting, it is clear that many results of the present section have parallels in such situations too, sometimes at a cost of a slightly more complicated formulation. See e.g. Sigman (1995).

4 Queue Length Processes

In the same way that the (actual) waiting time process is obtained by observing the virtual waiting time (workload) just before arrival times, it is sometimes of interest to look at the queue length (number in system) at certain random times. In particular, seen from the point of view of the arriving customer, the queue length at the time of arrival is a basic quantity and motivates the study of $\{Q_n^A\}_{n\in\mathbb{N}}$; cf. Definition 1.4. To distinguish from $\{Q_n^A\}$, $\{Q_n^D\}$, we use the terminology "at an arbitrary point of time" when considering $\{Q_t\}_{t\ge 0}$ in the steady state, and Q in Definition 1.4 refers to this case (i.e. $\mathbb{P}(Q = k) = \lim_{t\to\infty}\mathbb{P}(Q_t = k)$). We start by an elementary but celebrated result:

Theorem 4.1 (LITTLE'S LAW) *Suppose $\rho < 1$ and that A is nonlattice. Then the arrival rate $\lambda = \mu_A^{-1}$, the mean steady–state queue length $\ell = \mathbb{E}Q$ at an arbitrary point of time and the mean steady–state sojourn time $w = \mathbb{E}(W + U)$ are related by $\ell = \lambda w$.*

Proof. By reference to Proposition 3.1, regenerative processes apply to $\{Q_t\}_{t\geq 0}$ exactly as to the workload to show that a limiting steady–state r.v. Q exists (the convergence is always in t.v. since the state space $\mathbb{N}$ is discrete) and has distribution given by

$$\mathbb{E}f(Q) \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C f(Q_s)\,\mathrm{d}s. \tag{4.1}$$

Letting $f(x)=x$, it is seen that each of the customers $n=0,1,\ldots,\sigma-1$ provide a contribution W_n+U_n to $\int_0^C Q_s\,\mathrm{d}s$. Hence

$$\ell \;=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C Q_s\,\mathrm{d}s \;=\; \frac{1}{\mu_A\mathbb{E}\sigma}\mathbb{E}\sum_{n=0}^{\sigma-1}(W_n+U_n) \;=\; \lambda\mathbb{E}(W+U) \;=\; \lambda w.$$

□

Theorem 4.2 (DISTRIBUTIONAL LITTLE'S LAW) *Let $\{N^*(t)\}$ be a time–stationary version of the renewal arrival process that is independent of W,U, etc. Then $Q \stackrel{\mathscr{D}}{=} N^*(W+U)$, i.e. $\mathbb{P}(Q=0)=\rho$ and*

$$\mathbb{P}(Q\geq k) \;=\; \mathbb{P}(N^*(W+U)\geq k) \;=\; \mathbb{P}\big(W+U>T^*+T^{(k-1)}\big)$$

for $k=1,2,\ldots$. An alternative characterization is

$$\mathbb{P}(Q\geq k) \;=\; \mathbb{P}\big(V>T^{(k-1)}\big) \;=\; \rho\mathbb{P}\big(W+U^*>T^{(k-1)}\big), \tag{4.2}$$

$$\mathbb{P}(Q=k) \;=\; \int_0^\infty \big[A^{*(k-1)}(t)-A^{*k}(t)\big]\,H*B_0(\mathrm{d}t).$$

Proof. Let $\tau_n=T_0+\cdots+T_{n-1}$ be the arrival time of customer n. Then his service interval is $[\tau_n+W_n,\tau_n+W_n+U_n)$, and the time during this interval where $Q_t\geq k$ is the intersection with $[\tau_{n+k-1},\infty)$ which has length

$$r_n \;=\; (\tau_n+W_n+U_n-\tau_{n+k-1})^+-(\tau_n+W_n-\tau_{n+k-1})^+.$$

Since $\tau_{n+k-1}-\tau_n$ is independent of $W_n,U_n,\{\sigma>n\}$ and distributed as $T^{(k-1)}$, it follows that

$$\begin{aligned}
\mathbb{P}(Q\geq k) \;&=\; \frac{1}{\mathbb{E}C}\,\mathbb{E}\int_0^C I(Q_s\geq k)\,\mathrm{d}s \\
&=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\int_0^C I(Q_s\geq k)\,\mathrm{d}s \;=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\sum_{n=0}^{\sigma-1} r_n \\
&=\; \frac{1}{\mu_A\mathbb{E}\sigma}\,\mathbb{E}\sum_{n=0}^{\sigma-1}\Big[(W_n+U_n-T^{(k-1)})^+-(W_n-T^{(k-1)})^+\Big].
\end{aligned}$$

But conditionally upon $T^{(k-1)}=x$, this last expression is of the same form as (3.4) and so becomes

$$\rho\mathbb{P}\big(U^*>T^{(k-1)}-W\big) \;=\; \rho\mathbb{P}\big(V>T^{(k-1)}\,\big|\,V>0\big) \;=\; \mathbb{P}\big(V>T^{(k-1)}\big) \tag{4.3}$$

$$=\; \mathbb{P}\big(W+U-T^*>T^{(k-1)}\big) \;=\; \mathbb{P}\big(N^*(W+U)\geq k\big), \tag{4.4}$$

where we used the first part of Theorem 3.4 for (4.3) and the last for (4.4). The first part of the present theorem now follows from (4.4) and the last from (4.3). □

Theorem 4.3 *Suppose $\rho < 1$. Then Q^A, Q^D are welldefined and have the same distribution given by*

$$\mathbb{P}(Q^A \geq k) \;=\; \mathbb{P}(Q^D \geq k) \;=\; \mathbb{P}\big(W + U \geq T^{(k)}\big). \tag{4.5}$$

If either A or B is continuous, this may be rewritten as

$$\mathbb{P}(Q^A = 0) \;=\; \mathbb{P}(Q^D = 0) \;=\; \mathbb{P}(W = 0) \;=\; H(0), \tag{4.6}$$

$$\mathbb{P}(Q^A \geq k) \;=\; \mathbb{P}(Q^D \geq k) \;=\; \mathbb{P}\big(W > T^{(k-1)}\big), \quad k = 1, 2, \ldots, \tag{4.7}$$

$$\mathbb{P}(Q^A = k) \;=\; \mathbb{P}(Q^D = k) \;=\; \mathbb{P}\big(T^{(k-1)} < W < T^{(k)}\big), \tag{4.8}$$

$$\mathbb{P}(Q^A = k) \;=\; \mathbb{P}(Q^D = k) \;=\; \int_{0+}^{\infty} \big[A^{*(k-1)}(t) - A^{*k}(t)\big]\, H(\mathrm{d}t). \tag{4.9}$$

Proof. Clearly, $\{Q_n^A\}$, $\{Q_n^D\}$ are regenerative w.r.t. the renewal sequence $\{\sigma(k)\}$ and hence the existence of the limiting distribution is immediate. That the distributions are equal follow by rate conservation applied to $X_t = I(Q_t \geq k+1)$, since upward jumps occur at arrival epochs with $Q_n^A = k$ and downward at departure epochs with $Q_n^D = k$.

As in the proof of Theorem 4.2, we have

$$\big\{Q_{n+k}^A \geq k\big\} \;=\; \big\{\tau_{n+k} \leq \tau_n + W_n + U_n\big\}.$$

From this (4.5) follows by taking probabilities and letting $n \to \infty$. If either A or B is continuous, then so is the distribution of $W + U - T$ so that for $k \geq 1$ (4.5) becomes

$$\mathbb{P}\big(W + U > T^{(k)}\big) \;=\; \mathbb{P}\big(W + U - T > T^{(k-1)}\big) \;=\; \mathbb{P}\big(W > T^{(k-1)}\big)$$

(using $W \stackrel{\mathscr{D}}{=} (W + U - T)^+$). From this (4.6)–(4.9) follow by easy manipulations (in (4.9), 0 must be excluded from the domain of integration to deal with the case $k = 1$ where $T^{(k-1)} = 0$). □

Problems

4.1 Consider the set–up of Theorem 4.3. Explain that if $\mathbb{P}(U = T) > 0$, it may happen that $\mathbb{P}(Q^A = 0)$ is effectively smaller than $\mathbb{P}(W = 0)$.
4.2 Derive the distribution of Q^D by a direct argument similar to the one used for Q^A.

Notes It is clear as in Section 3 that much of the analysis carries over beyond the independence assumptions in $GI/G/1$. In particular, this is the case for Little's law which basically does not require anything more than the existence of limits of the Cesaro averages of the number of customers in continuous time and of the sojourn times in discrete time. The literature is extensive; see e.g.

Sigman (1995) and El–Taha and Stidham (1999). A series of papers by Glynn and Whitt, e.g. Glynn and Whitt (1986, 1989), deal with broader interpretations of $\ell = \lambda w$. For an extension $H = \lambda G$ that has been studied extensively, see e.g. Sigman (1995); it is essentially equivalent to the rate conservation law and applies typically to the same kind of problems.

5 $M/G/1$ and $GI/M/1$

Most of the steady–state characteristics of $M/G/1$ and $GI/M/1$ have already been found at various places; see in particular Sections 3, 4 and VIII.5. We collect here some of the main facts, give some complements and sketch some alternative approaches.

Theorem 5.1 *Consider the $GI/M/1$ queue with interarrival distribution A, service intensity δ and $\rho = (\delta\mu_A)^{-1} < 1$. Then in the steady state:*
(a) *The distribution of the waiting time W is a mixture of an atom at 0 and an exponential distribution with intensity η on $(0,\infty)$ with weights $1-\theta$, resp. θ. Here $\theta = \mathbb{E}\mathrm{e}^{-\eta T} = 1 - \eta/\delta$, where η is the solution > 0 of*

$$1 \;=\; \mathbb{E}\mathrm{e}^{\eta(U_n - T_n)} \;=\; \frac{\delta}{\delta-\eta}\int_0^\infty \mathrm{e}^{-\eta x}\, A(\mathrm{d}x)\,. \tag{5.1}$$

(b) *The distribution of the workload V is a mixture of an atom at 0 and an exponential distribution with intensity η on $(0,\infty)$ with weights $1-\rho$, resp. ρ.*
(c) *The distribution of the queue length Q at an arbitrary point of time is modified geometric and given by $\mathbb{P}(Q=0) = 1-\rho$, $\mathbb{P}(Q \ge k) = \rho\theta^{k-1}$, $k = 1, 2, \ldots$.*
(d) *The common distribution of the queue lengths Q^A, Q^D just before arrivals, resp. just after departures, is geometric with parameter θ, i.e. with point probabilities $\pi_n = (1-\theta)\theta^n$.*

Proof. (a) was shown in VIII.5.8. When U is exponential, we have $U^* \stackrel{\mathscr{D}}{=} U$, and we get the Laplace transform of $W + U^*$ as

$$\Big[1 - \theta + \theta\frac{\eta}{\eta+s}\Big]\frac{\delta}{\delta+s} \;=\; \frac{[\eta + (1-\theta)s]\delta}{(\eta+s)(\delta+s)} \;=\; \frac{\eta\delta + \eta s}{(\eta+s)(\delta+s)} \;=\; \frac{\eta}{\eta+s}$$

which proves (b); cf. Theorem 3.4. For (c) and (d), note first that conditioning upon $T^{(k-1)}$ in (4.7) yields

$$\mathbb{P}\big(W > T^{(k-1)}\big) \;=\; \theta\mathbb{E}\mathrm{e}^{-\eta T^{(k-1)}} \;=\; \theta\left[\mathbb{E}\mathrm{e}^{-\eta T}\right]^{k-1} \;=\; \theta^k,$$

and (d) follows. (c) is obtained similarly from (4.2) and $\mathbb{P}\big(V > T^{(k-1)}\big) = \rho\theta^{k-1}$. □

Imbedded Markov chain analysis plays an important historical role in the proof of results like (c) and (d) and is also applicable to a number of further

models. We therefore next present the main steps of this approach, though it is certainly neither the shortest nor the most elegant one for simple queues such as $GI/M/1$ (and $M/G/1$ below). It was found in III.6.2 that the Markov chain $\{Q_n^A\}$ has transition matrix

$$\boldsymbol{P} \;=\; \begin{pmatrix} r_0 & q_0 & 0 & 0 & \dots \\ r_1 & q_1 & q_0 & 0 & \\ r_2 & q_2 & q_1 & q_0 & \\ \vdots & & & & \ddots \end{pmatrix},$$

where $q_k \;=\; \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^k}{k!}\,A(\mathrm{d}t)$ and $r_n \;=\; q_{n+1} + q_{n+2} + \cdots$. By direct insertion it is now seen that $\pi_n = (1-\theta)\theta^n$ solves $\boldsymbol{\pi P} = \boldsymbol{\pi}$, provided that θ satisfies (i) $\sum_0^\infty r_n\theta^n = 1$, (ii) $\sum_0^\infty q_n\theta^n = \theta$. An elementary calculation shows that (i) follows from (ii). If η, θ are connected by $\eta = \delta(1-\theta)$, $\theta = 1 - \eta/\delta$, we may rewrite (ii) as

$$1 - \frac{\eta}{\delta} \;=\; \sum_{n=0}^\infty q_n\theta^n \;=\; \int_0^\infty \mathrm{e}^{-\eta t}\,A(\mathrm{d}t),$$

which is the same as (5.1). Alternatively, $\boldsymbol{\pi}$ can be derived by remarking that $\{Q_n^A\}$ is a Lindley process governed by $f_1 = q_0$, $f_0 = q_1$, $f_{-1} = q_2$, …, hence the stationary distributions is that of the random walk maximum M which was found in VIII.5.5(b).

To proceed from Q^A to Q, we use semi–regeneration; cf. VII.5. The cycle length is an interarrival time T and we let $\mathbb{E}_k$ refer to the case where k customers were present just before the start of the interarrival interval. The imbedded Markov chain in VII.5 is just $\{Q_n^A\}$ with stationary distribution $\boldsymbol{\pi}$, and thus by VII.(5.1), we have

$$\mathbb{P}(Q = j) \;=\; \frac{1}{m}\sum_{k=0}^\infty \pi_k\,\mathbb{E}_k\int_0^T I(Q_t = j)\,\mathrm{d}t,$$

where $m = \sum_0^\infty \pi_k\mathbb{E}_k T = \mu_A$. If $\{N_s\}$ is a Poisson process with intensity δ and $j \le k+1$, integration by parts yields

$$\begin{aligned}\mathbb{E}_k\int_0^T I(Q_t = j)\,\mathrm{d}t \;&=\; \int_0^\infty \mathbb{P}(N_t = k+1-j)\overline{A}(t)\,\mathrm{d}t \\ =\; \int_0^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^{k+1-j}}{(k+1-j)!}\overline{A}(t)\,\mathrm{d}t \;&=\; \int_0^\infty \delta^{-1}\sum_{\ell=k+2-j}^\infty \mathrm{e}^{-\delta t}\frac{(\delta t)^\ell}{\ell!}\,A(\mathrm{d}t)\end{aligned}$$

which equals $\delta^{-1}r_{k+1-j}$. For $j > k+1$, we get 0 and hence

$$\begin{aligned}\mathbb{P}(Q = j) \;&=\; \frac{1}{\mu_A}\sum_{k=j-1}^\infty \pi_k\delta^{-1}r_{k+1-j} \;=\; \rho\sum_{i=0}^\infty (1-\theta)\theta^{j-1}\theta^i r_i \\ &=\; \rho(1-\theta)\theta^{j-1}\end{aligned}$$

(using (i) above) and (c) is shown. For an alternative proof using rate conservation, let $X_t = I(Q_t \geq j+1)$. The rate of upward jumps is $\mu_A^{-1}\pi_j$ and the rate of downward jumps is $\delta\mathbb{P}(Q = j+1)$. Equating these two quantities yields $\mathbb{P}(Q = j+1) = \rho\pi_j$ which shows that (c) follows immediately from (d).

Theorem 5.2 *Consider the $M/G/1$ queue with interarrival intensity β, service time distribution B, and $\rho = \beta\mu_B < 1$. Then in the steady state:*
(a) *The distributions of the waiting time W and the workload V are the same and given as $H = (1-\rho)\sum_0^\infty \rho^n B_0^{*n}$, where $B_0(x) = \mu_B^{-1}\int_0^x \overline{B}(y)\,\mathrm{d}y$ is the stationary excess distribution.*
(b *The distributions of the queue lengths Q, Q^A, Q^D at an arbitrary point, just before arrivals, resp. just after departures, are the same, say $\boldsymbol{\pi}$, which can be expressed in terms of H and the Poisson distribution by $\pi_0 = 1-\rho$,*

$$\pi_k \;=\; \int_{0^+}^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{k-1}}{(k-1)!}\,H(\mathrm{d}t) \;=\; \rho\int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{k-1}}{(k-1)!}\,H*B_0(\mathrm{d}t)\,, \quad (5.2)$$

$k = 1, 2, \ldots$. *In particular,*

$$\mathbb{E}Q \;=\; \rho\big[1+\beta(\mathbb{E}W + \mathbb{E}U^*)\big] \;=\; \rho + \beta\mathbb{E}W \;=\; \rho + \frac{\rho^2\mu_B^{(2)}}{2(1-\rho)\mu_B^2}\,. \quad (5.3)$$

Proof. For $W \stackrel{\mathscr{D}}{=} V$, see III.9.2, VII.6.7 and Section 3. Further, $H = (1-\rho)\sum_0^\infty B_0^{*n}$ is just the Pollaczeck–Khinchine formula VIII.(5.5).

In (b), $Q \stackrel{\mathscr{D}}{=} Q^A \stackrel{\mathscr{D}}{=} Q^D$ follows from $W \stackrel{\mathscr{D}}{=} V$ and Theorems 4.3 and 4.2. Also, Theorem 4.3 yields $\mathbb{P}(Q=0) = 1-\rho$ and

$$\mathbb{P}(Q \geq k) \;=\; \mathbb{P}\big(W > T^{(k-1)}\big) \;=\; \int_{0^+}^\infty \sum_{\ell=k-1}^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^\ell}{\ell!}\,H(\mathrm{d}t)\,,$$

from which the first part of (5.2) follows; the second follows since (a) shows that $\rho\,H*B_0$ coincides with H on $(0,\infty)$. The proof of (5.3) is now easy. □

The alternative approach of imbedded Markov chain analysis for $M/G/1$ starts by noting that

$$Q_{n+1}^D \;=\; (Q_n^D - 1)^+ + K_n \quad (5.4)$$

where K_n is the number of customers arriving while customer n is being served. Clearly, $\{Q_n^D\}$ is a Markov chain with transition matrix

$$\boldsymbol{P} \;=\; \begin{pmatrix} q_0 & q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ 0 & q_0 & q_1 & q_2 & \cdots \\ 0 & 0 & q_0 & q_1 & \cdots \\ 0 & 0 & 0 & q_0 & \cdots \\ \vdots & & & & \ddots \end{pmatrix},$$

where $q_k = \mathbb{P}(K_n = k) = \int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^k}{k!}\,B(\mathrm{d}t)$. Irreducibility is obvious since all $q_k > 0$, and also $\mathbb{E}K_n$ is the expected number $\beta\mu_B = \rho$ of arriving customers in a service interval. Thus $\mathbb{E}_i Q_1^D = \rho + i - 1$ for $i \geq 1$, and it is a matter of routine to check from Foster's criteria in I.5 that we have recurrence when $\rho \leq 1$ and ergodicity when $\rho < 1$ (when $\rho = 1$, there is in fact null recurrence, and when $\rho > 1$ there is transience; cf. Problem 5.4).

Assume in the following that $\rho < 1$. Then the equation $\boldsymbol{\pi P} = \boldsymbol{\pi}$ becomes

$$\begin{aligned} \pi_0 &= \pi_0 q_0 + \pi_1 q_0, \\ \pi_1 &= \pi_0 q_1 + \pi_1 q_1 + \pi_2 q_0, \\ \pi_2 &= \pi_0 q_2 + \pi_1 q_2 + \pi_2 q_1 + \pi_3 q_0 \\ &\vdots \end{aligned} \tag{5.5}$$

Letting $r_n = q_{n+1} + q_{n+2} + \cdots$, it follows by adding equations $0, \ldots, n$ and solving for $\pi_{n+1}q_0$ that

$$\begin{aligned} \pi_1 q_0 &= \pi_0 r_0, \\ \pi_2 q_0 &= \pi_0 r_1 + \pi_1 r_1, \\ \pi_3 q_0 &= \pi_0 r_2 + \pi_1 r_2 + \pi_2 r_1 \\ &\vdots \end{aligned} \tag{5.6}$$

If we sum these equations and note that $\sum_0^\infty r_n = \rho$, we get

$$(1 - \pi_0)q_0 \;=\; \pi_0\rho + (1 - \pi_0)(\rho - r_0),$$

from which it easily follows that $\pi_0 = 1 - \rho$. The remaining π_n are then recursively determined by (5.6), but cannot be found in closed formulas.

However, many properties of $\boldsymbol{\pi}$ can be derived directly from equations (5.4)–(5.6). Let us look at (5.4) which in the limit becomes

$$Q^D \;\stackrel{\mathscr{D}}{=}\; (Q^D - 1)^+ + K \;=\; Q^D - I(Q^D > 0) + K \tag{5.7}$$

(in obvious notation). Taking squared expectations yields

$$\mathbb{E}{Q^D}^2 \;=\; \mathbb{E}{Q^D}^2 + \mathbb{P}(Q^D > 0) + \mathbb{E}K^2 - 2\mathbb{E}Q^D + 2\mathbb{E}Q^D\,\mathbb{E}K - 2\mathbb{P}(Q^D > 0)\mathbb{E}K.$$

Eliminating $\mathbb{E}{Q^D}^2$ and solving for $\mathbb{E}Q^D$ using $\mathbb{E}K = \rho$, $\mathbb{P}(Q^D = 0) = 1 - \pi_0 = \rho$ and

$$\mathbb{E}K^2 \;=\; \int_0^\infty \sum_{k=0}^\infty k^2 \mathrm{e}^{-\beta t}\frac{(\beta t)^k}{k!}\,B(\mathrm{d}t) \;=\; \int_0^\infty [\beta t + (\beta t)^2]\,B(\mathrm{d}t) \;=\; \rho + \beta^2\mu_B^{(2)}$$

then easily yields the same expression as in (5.3) ($Q^D \stackrel{\mathscr{D}}{=} Q$ will be shown in a moment). Also the generating function $\widehat{\pi}[s] = \sum_0^\infty s^n\pi_n = \mathbb{E}s^{Q^D}$ can

be found in the same way. In fact, (5.7) yields

$$\begin{aligned}
\widehat{\pi}[s] &= \mathbb{E}s^{Q^D - I(Q^D>0)}\mathbb{E}s^K = (\pi_0 + \pi_1 + s\pi_2 + s^2\pi_3)\sum_{n=0}^{\infty} s^n q_n, \\
s\widehat{\pi}[s] &= \big[\widehat{\pi}[s] + \pi_0(s-1)\big]\widehat{q}[s] = \big[\widehat{\pi}[s] + (1-\rho)(s-1)\big]\widehat{q}[s], \\
\widehat{\pi}[s] &= \frac{(1-\rho)(1-s)\widehat{q}[s]}{\widehat{q}[s] - s} \qquad (5.8)
\end{aligned}$$

where, letting $\widehat{B}[\cdot]$ denote the Laplace transform of B,

$$\widehat{q}[s] = \int_0^\infty \sum_{k=0}^{\infty} \mathrm{e}^{-\beta t}\frac{(s\beta t)^k}{k!}\, B(\mathrm{d}t) = \int_0^\infty \mathrm{e}^{-\beta t(1-s)}\, B(\mathrm{d}t) = \widehat{B}\big[\beta(1-s)\big].$$

To proceed from Q^D to Q, we use again semi–regeneration. The imbedded Markov chain is $\{Q_n^D\}$ with stationary distribution $\boldsymbol{\pi}$, and a cycle C started by $Q_n^D = k \geq 1$ is just a service interval of length U; for $k = 0$ we have to add the idle period of expected length $1/\beta$. It follows that for $j \geq 1$ we have

$$\mathbb{P}(Q = j) = \frac{1}{m}\sum_{k=0}^{\infty} \pi_k\, \mathbb{E}_k \int_0^C I(Q_t = j)\,\mathrm{d}t,$$

where

$$m = \pi_0\,(1/\beta + \mu_B) + \sum_{i=1}^{\infty}\pi_i\mu_B = \frac{1-\rho}{\beta} + \mu_B = \frac{1}{\beta}.$$

For $j \geq 1$ fixed, write $\alpha_k = \mathbb{E}_k \int_0^C I(Q_t = j)\mathrm{d}t$. Then $\alpha_0 = \alpha_1$. For $k > j$ we have $\alpha_k = 0$, whereas for $1 \leq k \leq j$ we get

$$\begin{aligned}
\alpha_k &= \int_0^\infty \mathrm{e}^{-\beta t}\frac{(\beta t)^{j-k}}{(j-k)!}\overline{B}(t)\,\mathrm{d}t = \int_0^\infty \beta^{-1}\sum_{\ell=j-k+1}^{\infty}\mathrm{e}^{-\beta t}\frac{(\beta t)^\ell}{\ell!}\, B(\mathrm{d}t) \\
&= \beta^{-1} r_{j-k}.
\end{aligned}$$

It follows that

$$\mathbb{P}(Q = j) = \beta\sum_{k=0}^{\infty}\pi_k\alpha_k = \pi_0 r_{j-1} + \sum_{k=1}^{j}\pi_k r_{j-k} = \pi_j,$$

where the last equality follows from (5.6). The truth of this for all $j \geq 1$ implies $Q \stackrel{\mathscr{D}}{=} Q^D$.

For an alternative proof using rate conservation, let $X_t = I(Q_t \geq j+1)$. The rate of upward jumps is $\beta\mathbb{P}(Q = j)$ and the rate of downward jumps is $\beta\mathbb{P}(Q^D = j)$ (interpret β as the departure rate). Equating these two quantities yields $Q \stackrel{\mathscr{D}}{=} Q^D$.

Problems

5.1 Derive the steady–state characteristics of the $GI/G/1$ queue where $U-1$ is exponential with rate say δ and $T \geq 1$.
5.2 Check that the formulas for $\mathbb{E}W$ (see VIII.5.7) and $\mathbb{E}Q$ in $M/G/1$ are in agreement with Little's formula, and that $\mathbb{E}W = \mathbb{E}V$ in agreement with Corollary 3.5.
5.3 Show that $\mathbb{E}W$ and $\mathbb{E}Q$ in $M/G/1$ are minimized by $M/D/1$ subject to the constraints that β and μ_B are fixed. See further XI.5.
5.4 Show by a modification of the derivation of $\pi_0 = 1-\rho$ from (5.6) that the stationary measure is infinite for $\rho = 1$ and that therefore $\{Q_n^D\}$ is null recurrent. Show also that there is transience for $\rho > 1$. [*Hint:* $Y_{n+1} \geq Y_n - 1 + K_n$.]
5.5 Give a direct derivation of (5.8) by multiplying equation n in (5.6) by s^{n+1} and summing over n. Check the formula for the mean by differentiation.
5.6 Let R_t denote the attained service of the customer in service at time t (if any) and define $D_n(x) = \mathbb{P}(R \leq x, Q = n)$, $n = 1, 2, \ldots$ (thus $\|D_n\| = \pi_n$). Show that D_n has density

$$\beta\overline{B}(x)\mathrm{e}^{-\beta t}\left\{(\pi_0+\pi_1)\frac{(\beta t)^{n-1}}{(n-1)!} + \sum_{k=0}^{n-2}\pi_{n-k}\frac{(\beta t)^k}{k!}\right\}.$$

Notes A further classical topic for the $M/G/1$ queue is the connection of the busy period to branching processes. This is most readily understood in the preemptive LCFS setting (where the busy period distribution is the same as for FCFS). Here one defines the children of a particular customer as the customers who arrived while he was in service. A simple example of the connection is then that the number of customers served in the busy period is the same as the total number of progeny of the customer initiating the busy period. A fairly general formulation is in Shalmon (1988) who also gives references to earlier work (to which we add Neuts, 1969). A recent generalization goes from the compound Poisson $M/G/1$ case to Lévy processes, see LeGall and Le Yan (1998).

6 Continuity of the Waiting Time

We consider here and in the next two sections a family of $GI/G/1$ queueing systems indexed by $k = 0, 1, 2, \ldots$ with service time distribution $B^{(k)}$, interarrival distribution $A^{(k)}$ and $U_n^{(k)}$, $T_n^{(k)}$, $X_n^{(k)}$, $S_n^{(k)}$, $W_n^{(k)}$, $W^{(k)}$, etc. defined the obvious way. The problem, stated in a rough form, is to study the limiting behaviour of $W^{(k)}$ as $k \to \infty$ under appropriate conditions, assuming that $\rho_k < 1$ for $k = 1, 2, \ldots$ and that $A^{(k)} \xrightarrow{w} A^{(0)}$, $B^{(k)} \xrightarrow{w} B^{(0)}$ (weak convergence). In Sections 7, 8 we consider the extreme cases where the limit has traffic intensity $\rho_0 = 1$ or $\rho_0 = 0$, whereas the situation here is $0 < \rho_0 < 1$. It is then reasonable to ask for conditions under which $W^{(k)} \xrightarrow{\mathcal{D}} W^{(0)}$. This is denoted as a *continuity* (or *stability* or *robustness*) property of the waiting time, and is of importance for example to justify

the approximation of a queueing system with given $A^{(0)}$, $B^{(0)}$ by systems with $A^{(k)}$, $B^{(k)}$ of phase type (cf. III.4).

To facilitate notation, we suppress from now on indices n and $k = 1, 2, \ldots$ whenever convenient (thus, e.g. $\mathbb{E}U \to \mathbb{E}U^{(0)}$ or $\lim_{k\to\infty} \mathbb{E}U = \mathbb{E}U^{(0)}$ means $\mathbb{E}U_n^{(k)} \to \mathbb{E}U_n^{(0)}$).

We shall first state and prove the main result in random walk terms, and thereafter reformulate in terms more natural for queues.

Theorem 6.1 *Consider random walks* $\{S_n\}_{n\in\mathbb{N}}$, $\{S_n^{(0)}\}_{n\in\mathbb{N}}$ *with* $\mu = \mathbb{E}X < 0$, $F \xrightarrow{w} F^{(0)}$, $k \to \infty$, $\mu_0 = \mathbb{E}X_n^{(0)} < 0$. *Then* $M \xrightarrow{\mathscr{D}} M^{(0)}$ *provided that the* X^+ *are uniformly integrable or equivalently that* $\mathbb{E}X^+ \to \mathbb{E}X^{(0)^+}$.

The key step of the proof is

Lemma 6.2 *Define* $K_n = \max_{r\ge n} S_r$. *Then* $\overline{\lim_{n\to\infty}}\, \overline{\lim_{k\to\infty}}\, \mathbb{P}(K_n > 0) = 0$.

Proof. By general results on weak convergence, $\mathbb{E}X^+ \to \mathbb{E}X^{(0)^+}$ is equivalent to the uniform integrability of the X^+ since $X^+ \xrightarrow{\mathscr{D}} X^{(0)^+}$. Choose $c < 0$ such that $\mathbb{E}\big[X^{(0)^+} \vee c\big] < 0$ and define $\check{X}_n = X_n \vee c$. Then $\check{X}_n \xrightarrow{\mathscr{D}} \check{X}^{(0)}$, $\check{S}_n \ge S_n$, $\check{K}_n \ge K_n$. Hence for the proof it is no restriction to assume that the X are uniformly bounded below, say by c. Then the X themselves are uniformly integrable, hence $\mu = \mathbb{E}X \to \mu_0 < 0$. Now for $\mu < 0$,

$$\begin{aligned}\mathbb{P}(K_n > 0) &= \mathbb{P}\left(\max_{r\ge n} \frac{S_r}{r} > 0\right) = \mathbb{P}\left(\max_{r\ge n}\left\{\frac{S_r}{r} - \mu\right\} > -\mu\right)\\ &\le \frac{1}{|\mu|}\mathbb{E}\left|\frac{S_n}{n} - \mu\right|,\end{aligned}$$

using the fact that $\{S_r/r - \mu\}_{r=n,n+1,\ldots}$ is a backward martingale and Kolmogorov's inequality. Decompose $S_n - n\mu$ as $\widetilde{S}_n + \widetilde{\widetilde{S}}_n$, where

$$\widetilde{X}_n = X_n I(X_n \le d) - \mathbb{E}[X_n; X_n \le d],\ \widetilde{\widetilde{X}}_n = X_n I(X_n > d) - \mathbb{E}[X_n; X_n > d]$$

with d satisfying $\mathbb{P}(X_n^{(0)} = d) = 0$, $\mathbb{E}[X_n; X_n > d] < \epsilon$ for all k. Then $\widetilde{\sigma}^2 = \mathbb{V}ar\widetilde{X}_n \to \widetilde{\sigma}_0^2 = \mathbb{V}ar\widetilde{X}_n^{(0)}$ since the $\widetilde{X}$ are bounded uniformly in k, so that

$$\mathbb{E}|S_n/n - \mu| \le \mathbb{E}|\widetilde{S}_n/n| + \mathbb{E}|\widetilde{\widetilde{S}}_n/n| \le \widetilde{\sigma}/\sqrt{n} + 2\epsilon,$$

using the Cauchy–Schwarz inequality. Hence

$$\overline{\lim_{n\to\infty}}\, \overline{\lim_{k\to\infty}}\, \mathbb{P}(K_n > 0) \le \overline{\lim_{n\to\infty}} \frac{1}{|\mu_0|}\big[\widetilde{\sigma}_0/\sqrt{n} + 2\epsilon\big] = \frac{2\epsilon}{|\mu_0|},$$

and since ϵ is arbitrary, the proof is complete. □

Proof of Theorem 6.1. From $X \xrightarrow{\mathscr{D}} X^{(0)}$ it follows that $\{X_r\}_{r=0}^n \xrightarrow{\mathscr{D}} \{X_r^{(0)}\}_{r=0}^n$ and hence by the continuous mapping theorem $M_n \xrightarrow{\mathscr{D}} M_n^{(0)}$.

Now let $x > 0$ satisfy $\mathbb{P}(M^{(0)} = x) = 0$. Then also $\mathbb{P}(M_n^{(0)} = x) = 0$ for each n and hence

$$\begin{aligned}\limsup_{k\to\infty} \mathbb{P}(M^{(k)} > x) &\leq \limsup_{k\to\infty} \big[\mathbb{P}(M_n^{(k)} > x) + \mathbb{P}(K_n^{(k)} > 0)\big] \\ &= \mathbb{P}(M_n^{(0)} > x) + \limsup_{k\to\infty} \mathbb{P}(K_n^{(k)} > 0), \\ \liminf_{k\to\infty} \mathbb{P}(M^{(k)} > x) &\geq \liminf_{k\to\infty} \mathbb{P}(M_n^{(k)} > x) \;=\; \mathbb{P}(M_n^{(0)} > x).\end{aligned}$$

Letting $n \to \infty$ yields $\mathbb{P}(M > x) \to \mathbb{P}(M^{(0)} > x)$. Hence $M \stackrel{\mathscr{D}}{\to} M^{(0)}$ (note that $x = 0$ is not a continuity point of M). □

Apparently the point mass of M at zero is of particular interest, but $\mathbb{P}(M = 0) \to \mathbb{P}(M^{(0)} = 0)$ does not follow alone from $M \stackrel{\mathscr{D}}{\to} M^{(0)}$. However:

Proposition 6.3 *If in addition to the assumptions of Theorem 6.1 the distribution $F^{(0)}$ of $X^{(0)}$ is continuous, then $\mathbb{P}(M = 0) \to \mathbb{P}(M^{(0)} = 0)$.*

Proof. The assumptions ensure that $\mathbb{P}(S_n^{(0)} = 0) = 0$ for each $n \geq 1$ and hence $\mathbb{P}(M_n = 0) \to \mathbb{P}(M_n^{(0)} = 0)$. Now argue exactly as above. □

Corollary 6.4 *Consider for $k = 0, 1, 2, \ldots$ $GI/G/1$ queues with $A \stackrel{w}{\to} A^{(0)}$, $B \stackrel{w}{\to} B^{(0)}$, $\rho_0 < 1$. Then $W \stackrel{\mathscr{D}}{\to} W^{(0)}$ provided that the U are uniformly integrable, or equivalently, that $\mathbb{E}U \to \mathbb{E}U^{(0)}$. If in addition either $A^{(0)}$ or $B^{(0)}$ is continuous, then also $\mathbb{P}(W = 0) \to \mathbb{P}(W^{(0)} = 0)$.*

Proof. Appealing to the interpretation $X = U - T$, $W \stackrel{\mathscr{D}}{=} M$, it is straightforward to check the assumptions of Theorem 6.1 and Proposition 6.3 (the uniform integrability of the X^+ follows from $X^+ \leq U$ and the uniform integrability of the U). □

Problems

6.1 Let $F^{(k)}$ be concentrated at $-1, k$ with point masses $1 - 1/2k$, $1/2k$ and let $F^{(0)} = \lim F^{(k)}$. Show that

$$\mathbb{P}(M^{(k)} \geq 1) \;\geq\; \mathbb{P}(X_n = k \text{ for some } n = 1, \ldots, k) \;\to\; \mathrm{e}^{-1/2}$$

and deduce that $M^{(k)} \stackrel{\mathscr{D}}{\to} M^{(0)}$ does not hold.

Notes Continuity problems are treated e.g. in Borovkov (1976), Stoyan (1983), Brandt *et al.* (1990) and Kalashnikov (1994). A classical reference for Markov chains is Karr (1975).

7 Heavy Traffic Limit Theorems

If, in the set–up of Section 6, the limiting traffic intensity ρ_0 is 1 rather than < 1, we are in the situation of *heavy traffic* where all queueing systems

are heavily congested. We expect again $W = W^{(k)} \stackrel{\mathscr{D}}{\to} W^{(0)}$, but now $W^{(0)} = \infty$ a.s. It will turn out that a more precise result can be obtained, namely that under weak conditions $|\mu|W$ is approximately exponentially distributed. We start again by formulating this for the random walk setting (σ^2 denotes $\mathbb{V}arX$).

Theorem 7.1 *Consider random walks* $\{S_n\}_{n\in\mathbb{N}}$, $\{S_n^{(0)}\}_{n\in\mathbb{N}}$ *with* $\mu < 0$, $\mu \to 0$, $\underline{\lim}_{k\to\infty} \sigma^2 > 0$, *and the* X^2 *uniformly integrable. Then* $Y = |\mu|M/\sigma^2$ *is approximately exponentially distributed with intensity* 2, *i.e.* $\mathbb{P}(Y > y) \to \mathrm{e}^{-2y}$. *Furthermore,* $\mathbb{E}Y \to 1/2$.

Remark 7.2 The conditions of Theorem 7.1 are not intrinsically different from the apparently stronger

$$X \stackrel{\mathscr{D}}{\to} X^{(0)}, \quad \sigma^2 \to \sigma_0^2 > 0, \quad \mu_0 = 0. \tag{7.1}$$

Indeed, the uniform integrability ensures that $\{F\} = \{F^{(k)}\}$ is tight. Thus every subsequence $\{k'\}$ has a weakly convergent subsequence $\{k''\}$, i.e. $X^{(k'')} \stackrel{\mathscr{D}}{\to} X^{(0)}$ for some $X^{(0)}$. But then by uniform integrability, $\mu_0 = \lim \mu_{k''} = 0$, $\sigma_0^2 = \lim \sigma_{k''}^2 > 0$. Furthermore, a standard analytical argument shows that if we can show the asymptotic exponentiality for $\{k''\}$, then it will hold for $\{k'\}$ as well. Hence *for the proof we can* (and shall) *assume that* (7.1) *holds.* Also, by rescaling, we may take $\sigma^2 = 1$; then $Y = |\mu|M = -\mu M$. □

Two approaches to Theorem 7.1 will be considered, the first being based on characteristic functions $\varphi_Y(y) = \mathbb{E}\mathrm{e}^{\mathrm{i}yY}$. Thus we have to show $\varphi_Y(y) = \varphi_M(-\mu y) \to (1 - \mathrm{i}y/2)^{-1}$. In the proof, we let $\mu_2 = \mathbb{E}X^2$ (thus $\mu_2 \to 1$ since $\mu \to 0$, $\sigma^2 \to 1$).

Lemma 7.3 *For each* y, *it holds as* $k \to \infty$ *that*

$$\varphi_X(-\mu y) \;=\; 1 - \mathrm{i}\mu^2 y - \frac{\mu^2 y^2}{2} + \mathrm{o}(\mu^2). \tag{7.2}$$

Proof. Define $g(z) = \mathrm{e}^{\mathrm{i}yz} - 1 - \mathrm{i}yz + y^2z^2/2$. Then for each $\epsilon > 0$, we can bound $|g(z)|$ by $c_\epsilon|z^3|$ for $|z| \le \epsilon$ and by $d_\epsilon|z^2|$ for $|z| > \epsilon$. Hence

$$\begin{aligned} \big|\mathbb{E}g(-\mu X)\big| &\le c_\epsilon \mathbb{E}\big[|-\mu X\big|^3; \big|-\mu X\big| \le \epsilon] + d_\epsilon \mathbb{E}\big[(\mu X)^2; \big|-\mu X\big| > \epsilon] \\ &\le \mu^2\big\{\epsilon c_\epsilon \mathbb{E}X^2 + d_\epsilon \mathbb{E}\big[(\mu X)^2; |-\mu X| > \epsilon\big]\big\} \end{aligned}$$

and therefore $\limsup_{k\to\infty} \mu^{-2}|\mathbb{E}g(-\mu X)| \le \epsilon c_\epsilon$ by uniform integrability. Since c_ϵ remains bounded as $\epsilon \downarrow 0$, it follows that

$$\varphi_X(-\mu y) - \Big(1 - \mathrm{i}\mu^2 y - \frac{\mu^2 y^2}{2}\mu_2\Big) \;=\; \mathbb{E}g(-\mu X) \;=\; \mathrm{o}(\mu^2),$$

and the lemma follows since $\mu_2 \to 1$. □

Proof of Theorem 7.1. We first note that as in Section 6 we have $M \stackrel{\mathscr{D}}{\to} \infty$. But

$$\mathbb{E}\big[(M+X)^-\big]^2 \;\le\; \mathbb{E}(X^-)^2\mathbb{P}(M \le c) + \mathbb{E}[X^2;\, X < -c].$$

Letting first $k \to \infty$ and next $c \to \infty$ yields

$$\mathbb{E}\big[(M+X)^-\big]^2 \;\to\; 0 \quad (\text{hence } \mathbb{E}(M+X)^- \to 0). \tag{7.3}$$

From this $\mathbb{E}Y \to 1/2$ is clear from (2.5). Now for each z, $\mathrm{e}^{\mathrm{i}yz^+} = \mathrm{e}^{\mathrm{i}yz} + 1 - \mathrm{e}^{-\mathrm{i}yz^-}$. Letting $Z = M + X$ and taking expectations we get

$$\varphi_M(y) \;=\; \varphi_M(y)\varphi_X(y) + 1 - \varphi_{-(M+X)^-}(y) \;=\; \frac{1-\varphi_{-(M+X)^-}(y)}{1-\varphi_X(y)}\,. \tag{7.4}$$

Since $\mathrm{e}^{\mathrm{i}z} - 1 - \mathrm{i}z \;=\; z^2\mathrm{O}(1)$ for z real, we get

$$\begin{aligned}\varphi_{-(M+X)^-}(-\mu y) &= 1 + \mathrm{i}\mu y\mathbb{E}(M+X)^- + \mathrm{O}(1)\mu^2y^2\mathbb{E}\big[(M+X)^-\big]^2\\ &= 1 - \mathrm{i}\mu^2 y + o(\mu^2),\end{aligned}$$

using (1.9) and (7.3). Hence by Lemma 7.3 and (7.4),

$$\varphi_M(-\mu y) \;=\; \frac{\mathrm{i}\mu^2 y + \mathrm{o}(\mu^2)}{\mathrm{i}\mu^2 y + \mu^2y^2/2 + \mathrm{o}(\mu^2)} \;\to\; \frac{1}{1-\mathrm{i}y/2}\,. \qquad \square$$

The second proof of Theorem 7.1 involves more advanced tools (weak convergence in function space) but is perhaps more illuminating and yields additional information, namely asymptotics of the $M_n^{(k)}$. We let $\{B_\xi(t)\}_{t\ge0}$ denote Brownian motion with unit variance and drift ξ. The *inverse Gaussian distribution function* $G(t;\,\xi,c)$ with parameters $\xi \in \mathbb{R}$, $c > 0$ is the c.d.f. of the first passage time $\tau(\xi,c) = \inf\{t > 0 :\, B_\xi(t) \ge c\}$,

$$G(T;\,\xi,c) \;=\; \mathbb{P}(\tau(\xi,c) \le T) \;=\; \mathbb{P}\Big(\max_{0\le t\le T} B_\xi(t) \ge c\Big)\,. \tag{7.5}$$

This distribution (defective for $\xi < 0$) can in fact be found explicitly. We defer the derivation to XIII.4 and here use only the formula

$$\big\|G(\cdot;\,\xi,c)\big\| \;=\; \mathbb{P}\Big(\max_{0\le t<\infty} B_\xi(t) \ge c\Big) \;=\; \mathrm{e}^{2\xi c}, \quad \xi < 0. \tag{7.6}$$

Proposition 7.4 *Under the conditions of Theorem 7.1, it holds for any $T < \infty$ that*

$$\frac{|\mu|}{\sigma^2}M_{\lfloor T\sigma^2/\mu^2\rfloor} \stackrel{\mathscr{D}}{\to} \max_{0\le t\le T} B_{-1}(t), \quad \mathbb{P}\left(\frac{|\mu|}{\sigma^2}M_{\lfloor T\sigma^2/\mu^2\rfloor} > y\right) \to G(T;\,-1,y).$$

Proof. We may again assume that (7.1) holds with $\sigma_0^2 = 1$. Let $\{c\} = \big\{c^{(m)}\big\}$ be any sequence with $c^{(m)} \to \infty$ and define

$$B(t) \;=\; B^{(m)}(t) \;=\; \frac{1}{\sqrt{c}}\left[S_{\lfloor ct\rfloor} - \lfloor ct\rfloor\mu\right].$$

It then follows from the invariance principle (Donsker's theorem) in its standard form (e.g. Billingsley, 1968, Ch. 3) that $B \overset{\mathscr{D}}{\to} B_0$ in D. Taking $c = \mu^{-2}$ we have $\lfloor ct \rfloor \mu / \sqrt{c} \to -t$, i.e.

$$\begin{aligned} \big\{ |\mu| S_{\lfloor t/\mu^2 \rfloor} \big\}_{0 \le t < \infty} &= \big\{ B(t) + \lfloor ct \rfloor \mu / \sqrt{c} \big\}_{0 \le t < \infty} \\ &\overset{\mathscr{D}}{\to} \{ B_0(t) - t \}_{0 \le t < \infty} \overset{\mathscr{D}}{=} B_{-1}. \end{aligned}$$

Hence, since $f \to \sup_{0 \le t \le T} f(t)$ is continuous a.e. on D w.r.t. any probability distribution concentrated on the continuous functions, it follows from the continuity of B_{-1} that

$$|\mu| M_{\lfloor T/\mu^2 \rfloor} = \sup_{0 \le t \le T} |\mu| S_{\lfloor t/\mu^2 \rfloor} \overset{\mathscr{D}}{\to} \max_{0 \le t \le T} B_{-1}(t)$$

which yields the desired conclusion in view of $\sigma^2 \to 1$. □

Proof of Theorem 7.1. We assume again $\sigma^2 \to 1$ and write

$$Y = Y_1 \vee Y_2 = \big(|\mu| M_{\lfloor T/\mu^2 \rfloor} \big) \vee \Big(\sup_{n > T/\mu^2} |\mu| S_n \Big).$$

Here by (7.6) and Proposition 7.4,

$$\lim_{T \to \infty} \lim_{k \to \infty} \mathbb{P}(Y_1 > y) = \lim_{T \to \infty} G(T; -1, y) = \mathrm{e}^{-2y}, \tag{7.7}$$

whereas $\big\{ (S_n - n\mu)^2 \big\}$ is a backward submartingale, hence

$$\begin{aligned} \mathbb{P}(Y_2 > 0) &= \mathbb{P}\Big(\max_{n > T/\mu^2} (S_n/n - \mu) > -\mu \Big) \\ &\le \frac{1}{\mu^2} \mathbb{E}\big[S_{\lfloor T/\mu^2 \rfloor} / \lfloor T/\mu^2 \rfloor - \mu \big]^2 = \frac{\sigma^2}{\mu^2 \lfloor T/\mu^2 \rfloor}, \end{aligned}$$

$$\lim_{T \to \infty} \lim_{k \to \infty} \mathbb{P}(Y_2 > 0) \le \lim_{T \to \infty} \frac{1}{T} = 0. \tag{7.8}$$

Combining (7.7) and (7.8), the desired conclusion is obtained exactly as in the proof of Theorem 6.1. □

Corollary 7.5 *Consider $GI/G/1$ queueing systems with $A \overset{w}{\to} A^{(0)}$, $B \overset{w}{\to} B^{(0)}$, where $A^{(0)}$, $B^{(0)}$ are not both degenerate, $\rho < 1$, $\rho \to \rho_0 = 1$ and the U^2, T^2 uniformly integrable. Then $Y = |\mu| W / \sigma^2$ is approximately exponentially distributed with intensity 2 and $\mathbb{E}Y \to 1/2$. Here $\mu = \mathbb{E}X = \mathbb{E}U - \mathbb{E}T$, $\sigma^2 = \mathbb{V}arX$. Furthermore, for each T*

$$\mathbb{P}\left(\frac{|\mu|}{\sigma^2} W_{\lfloor T\sigma 2/\mu^2 \rfloor} > y \right) \to G(T; -1, y).$$

The proof is a routine application of Theorem 7.1 and is omitted.

Results of the type in Corollary 7.5 are of high potential relevance, since the heavy traffic situation occurs widely in practice (when designing a service facility, one usually avoids for economical reasons to keep the server

idle for a large proportion of the time). Given a queue with ρ smaller than but close to 1, we may imbed the system in the set–up of Corollary 7.5, writing $A = A^{(k)}$, $B = B^{(k)}$ for some large k. It is then suggested that the following approximations may be used:

$$\mathbb{E}W \approx \mathbb{V}arX/2\mathbb{E}(-X), \quad \mathbb{P}(W > y) \approx \exp\bigl\{-2\mathbb{E}(-X)y/\mathbb{V}arX\bigr\}. \quad (7.9)$$

Note that when $\mathbb{E}X \approx 0$, we have $\mathbb{E}X^2 \approx \mathbb{V}arX$ and one may thus replace $\mathbb{V}arX$ by $\mathbb{E}X^2$ in (7.9). However, inspection of (2.5) shows that $\mathbb{E}X^2/2\mathbb{E}(-X)$ and $\mathbb{V}arX/2\mathbb{E}(-X)$ are both upper bounds for $\mathbb{E}W$ and hence $\mathbb{V}arX/2\mathbb{E}(-X)$ is the best approximation.

We return to a special aspect of heavy traffic approximations in XIII.6, but finally we mention that in view of the formulas of Sections 3 and 4, it is straightforward to derive analogues of Corollary 7.5 for workload, queue length and so on (cf. Problem 7.1).

Problems

7.1 Show that under the conditions of Corollary 7.5 the steady–state workload V has the same limiting distribution as W. Show similarly, using the results of Section 4, that $|\mu|Q/\sigma^2$, $|\mu|Q^A/\sigma^2$ have limiting exponential distributions with intensities $2\mu_A$.

7.2 Show (7.6) by optional stopping of the martingale $\{\mathrm{e}^{-2\xi B_\xi(t)}\}$ at $\tau(\xi, c) \wedge T$.

Notes Heavy traffic limit theory was largely initiated by Kingman in the 1960s, with the functional CLT point of view being developed by Iglehart and Whitt. For surveys, see Glynn (1990) and Whitt (2002).

Without second moments, one often gets a stable rather than a Brownian limit. See e.g. Furrer *et al.* (1997) and Heath *et al.* (1999) for recent papers in the area, and Whitt (2002) for a survey and references.

A notable recent development is heavy traffic limit theory for queueing networks, where the limit is reflected Brownian motion in an orthant. See further the Notes to IV.6 and IX.2.

8 Light Traffic

Intuitively, light traffic means that the generic interarrival time T is much larger than the generic service time U, implying that typically the system is idle in the steady state. When considering the $GI/G/1$ queue at an arbitrary point of time, the idleness probability is $1 - \rho = \mathbb{P}(Q = 0) = \mathbb{P}(V = 0)$, so that light traffic certainly requires ρ to be close to 0. A more refined question is to study the behaviour of Q, V given $\{Q > 0\} = \{V > 0\}$. To this end, we consider a sequence of $GI/G/1$ queues in the notation of Sections 6 and 7, assuming throughout that the interarrival time $T = T^{(k)}$ satisfies $T \xrightarrow{\mathscr{D}} \infty$, $k \to \infty$, and that the service time distribution B is fixed, i.e. does not depend on $k = 1, 2, \ldots$ (this certainly implies $\rho \to 0$).

Proposition 8.1 *As $k \to \infty$, it holds without further conditions that* (a) $W \to 0$ *in t.v.,* (b) V *conditionally upon* $V > 0$ *converges to the equilibrium service time* U^* *in t.v.,* (c) Q *conditionally upon* $Q > 0$ *converges to* 1 *in t.v.*
[For basic facts about total variation convergence, see A8.]

Proof. Let Z_ϵ denote a r.v. that is 0 w.p. ϵ and ϵ^{-1} w.p. $1-\epsilon$. Then $T^{(k)}$ is stochastically larger than Z_ϵ for all large k so that $W \le_{\rm so} M_\epsilon$ (stochastical order), the maximum of a random walk with increments distributed as $U - Z_\epsilon$. Since $M_\epsilon \le_{\rm so} M_\delta \stackrel{\mathscr{D}}{=}$ when $\epsilon < \delta$, we get

$$M_\epsilon \stackrel{\mathscr{D}}{=} (M_\epsilon + U - Z_\epsilon)^+ \le_{\rm so} (M_\delta + U - Z_\epsilon)^+$$

which converges in t.v. to 0 as $\epsilon \downarrow 0$. Hence $\mathbb{P}(M_\epsilon > 0) \to 0$ and therefore $P(W > 0) \to 0$, proving (a). It follows by Theorem 3.4 that

$$\mathbb{P}(V \in A \,|\, V > 0) \;=\; \mathbb{P}(W + U^* \in A) \;\sim\; \mathbb{P}(U^* \in A)$$

uniformly in $A \subseteq (0, \infty)$, showing (b). For (c), (4.2) then yields

$$\mathbb{P}(Q \ge 2) \;=\; \rho \mathbb{P}(W + U^* > T) \;\sim\; \rho \mathbb{P}(U^* > T) \;=\; \mathrm{o}(\rho),$$

$$\mathbb{P}(Q = 1 \,|\, Q > 0) \;=\; 1 - \mathbb{P}(Q \ge 2 \,|\, Q > 0) \;=\; 1 - \frac{\mathrm{o}(\rho)}{\rho} \;\to\; 1. \qquad \square$$

The intuitive content of Proposition 8.1(b),(c) is that a busy cycle in light traffic with high probability only contains one customer and that if we observe the system at an arbitrary point of time and see it busy, it is because we sample the single service time in the cycle rather than the following idle period. The situation at arrival instants is different: if a customer has to wait ($W > 0$), we expect him to be customer $n = 1$ in the cycle, not $n = 0$ who does not have to wait, so that W given $W > 0$ should most often be the residual service time $U_0 - T_0$ of the previous customer given it is positive. To rigorously verify this intuition as well as to derive precise asymptotics of $\mathbb{P}(W > 0)$ is, however, more difficult than in the case of V and will occupy the rest of this section.

We start again in a triangular array random walk setting, where we are given random walks $\{S_n\} = \{S_n^{(k)}\}$ with increments $X_0, X_1, \ldots$, increment distributions $F(x) = \mathbb{P}(X \le x)$, maxima $M = \max_{n=0,1,\ldots} S_n$ etc. (indexed by $k = 1, 2, \ldots$). Call two families $\{R\} = \{R^{(k)}\}$, $\{S\} = \{S^{(k)}\}$ of r.v.'s with values in $[0, \infty)$ *light traffic equivalent* if

$$\mathbb{P}(R > 0) \to 0,\ \mathbb{P}(S > 0) \to 0,\ \frac{\mathbb{P}(R > 0)}{\mathbb{P}(S > 0)} \to 1 \tag{8.1}$$

as $k \to \infty$ and the conditional t.v. distance converges to 0,

$$\big\| \mathbb{P}(R \in \cdot \,|R > 0) - \mathbb{P}(S \in \cdot \,|S > 0) \big\| \;\to\; 0. \tag{8.2}$$

Theorem 8.2 *Assume that* $X \overset{\mathscr{D}}{\to} -\infty$ *as* $k \to \infty$, *and that*

$$0 \;=\; \lim_{a\uparrow\infty} \overline{\lim_{k\to\infty}} \; \frac{\mathbb{E}[X; X > a]}{p_+} \;=\; \lim_{a\uparrow\infty} \overline{\lim_{k\to\infty}} \; \frac{\int_a^\infty x\, F(\mathrm{d}x)}{p_+} \tag{8.3}$$

where $p_+ = \mathbb{P}(X > 0) = \int_0^\infty F(\mathrm{d}x)$. *Then* M *and* X^+ *are light traffic equivalent.*

Remark 8.3 Let F_+ denote the conditional distribution of X given $X > 0$. Then (8.3) means that the family $\{F_+\}$ is uniformly integrable. This should be compared with the unconditional uniform integrability conditions for heavy traffic in Section 7. □

The key step in the proof is (take $S_{\tau_+} = 0$ when $\tau_+ = \infty$):

Lemma 8.4 *The ascending ladder heights* S_{τ_+} *and the* X^+ *are light traffic equivalent.*

Proof. By (1.7), we can write $G_+ = L + K$ where L, K are the restriction to $(0,\infty)$ of F, resp. $F * \sum_1^\infty G_-^{*n}$ (L is the contribution from the atom of U_- at zero). For simplicity of notation, let $R = R^{(k)}$ be the measure $R(\mathrm{d}x) = \sum_1^\infty G_-^{*n}\big(\mathrm{d}(-x)\big)$ on $(0,\infty)$. Then for $z \ge 0$,

$$\begin{aligned} \overline{K}(z) \;&=\; \sum_{n=1}^\infty \int_{-\infty}^0 \overline{F}(z-x)\, G_-^{*n}(\mathrm{d}x) \;=\; \int_0^\infty \overline{F}(z+x) R(\mathrm{d}x) \\ &=\; \int_z^\infty R(y-z) F(\mathrm{d}y) \;\le\; \int_z^\infty R(y) F(\mathrm{d}y). \end{aligned} \tag{8.4}$$

To proceed from (8.4), we will need the estimate

$$R(t) \;\le\; \varphi(t)(1+t), \tag{8.5}$$

where $\varphi(t)$ is bounded uniformly in k, nondecreasing and tends to 0 for any fixed t as $k \to \infty$. First $X \overset{\mathscr{D}}{\to} -\infty$ implies $S_{\tau_-} \overset{\mathscr{D}}{\to} -\infty$ (with high probability S_{τ_-} coincides with X_0). In particular, $\overline{G}_-(-t) \to 0$ for all $t > 0$. Since $R(1) \le \overline{G}_-(-1)\big(1 + R(1)\big)$, this implies that $R(1)$ is bounded. Similarly, $R(n-1, n] \le \overline{G}(-n)\big(1 + R(1)\big)$ so that $R(n) \le \overline{G}_-(-n)n\big(1 + R(1)\big)$, and from these estimates (8.5) follows.

Letting $z = 0$ in (8.4), we get

$$\begin{aligned} \limsup_{k\to\infty} \frac{\mathbb{P}\left(S_{\tau_+} > 0, S_{\tau_+} \ne X_0^+\right)}{p_+} \;&=\; \limsup_{k\to\infty} \frac{K(0,\infty)}{p_+} \\ &\le\; \limsup_{k\to\infty} \frac{\int_0^\infty \varphi(y)(1+y) F(\mathrm{d}y)}{p_+} \\ &\le\; \limsup_{k\to\infty} \Big\{ (1+a)\varphi(a) + 2\varphi(\infty) \frac{\int_a^\infty y F(\mathrm{d}y)}{p_+} \Big\} \\ &=\; \limsup_{k\to\infty} 2\varphi(\infty) \frac{\int_a^\infty y F(\mathrm{d}y)}{p_+}. \end{aligned}$$

Letting $a \to \infty$, this converges to 0 according to (8.3), which easily yields the assertion. □

Proof of Theorem 8.2: Just note that $\mathbb{P}(M > 0) = \mathbb{P}(S_{\tau_+} > 0)$,

$$\mathbb{P}(M \neq S_{\tau_+} \,|\, M > 0) = \mathbb{P}\big(\tau_+(2) < \infty \,\big|\, \tau_+ < \infty\big) = \mathbb{P}(S_{\tau_+} > 0) \to 0$$

and appeal to Lemma 8.4. □

We next consider $GI/G/1$ queues. In view of $W \stackrel{\mathscr{D}}{=} M$, Theorem 8.2 states that W and $(U-T)^+$ are light traffic equivalent provided $X = U-T$ satisfies (8.3). It remains to carry out the relevant translation to conditions in terms of A, B, and to give some examples.

The first example is thinning of the arrival process where the results are in terms of $\Gamma(t) = \sum_1^\infty A^{*n}(t)$ (the renewal function except that the $n = 0$ term is not included).

Corollary 8.5 *Given a $GI/G/1$ queueing system specified in terms of U, T, define for each $k = 1, 2, \ldots$ another $GI/G/1$ system by thinning of the arrival process with retention probability $1/k$. That is, $T = T_0 + \cdots + T_{N-1}$ where N is independent of $T_0, T_1, \ldots$ with $\mathbb{P}(N = \ell) = (1 - 1/k)^{\ell-1}/k$, $\ell = 1, 2, \ldots$ Then W and $(U-T)^+$ are light traffic equivalent provided that $\mathbb{E}U^2 < \infty$. Writing $\Gamma = \sum_1^\infty A^{*n}$, one then has*

$$\mathbb{P}(W > 0) \;\sim\; p_+ \;=\; \mathbb{P}(U - T > 0) \;\sim\; \frac{1}{k}\mathbb{E}\Gamma(U). \tag{8.6}$$

Proof. Obviously,

$$\mathbb{P}(U - T > y) \;=\; \int_y^\infty \sum_{\ell=1}^\infty \frac{1}{k}(1 - 1/k)^{\ell-1} A^{*\ell}(u - y) B(\mathrm{d}u)$$

for $y > 0$ so that

$$k\mathbb{P}(U - T > y) \;\uparrow\; \int_y^\infty \Gamma(u - y) B(\mathrm{d}u),\; k \to \infty. \tag{8.7}$$

Taking $y = 0$ gives $p_+ \sim \mathbb{E}\Gamma(U)/k$. Further, by integration by parts we have

$$\int_a^\infty x F(\mathrm{d}x) \;=\; a\mathbb{P}(U - T > a) \,+\, \int_a^\infty \mathbb{P}(U - T > x)\,\mathrm{d}x. \tag{8.8}$$

We can bound $\Gamma(y)$ by $c(1 + y)$, and therefore an upper bound for (8.8) is

$$\begin{aligned} & \frac{c}{k}\Big[a\overline{B}(a) + a\mathbb{E}(U - a)^+ + \int_a^\infty \big\{\overline{B}(x) + \mathbb{E}(U - x)^+\big\}\mathrm{d}x\Big] \\ = \;& \frac{c}{k}\Big[a\overline{B}(a) + (1 + a)\mathbb{E}(U - a)^+ + \frac{1}{2}\mathbb{E}(U - a)^{+2}\Big]. \end{aligned}$$

Here $[\cdots] \to 0$ as $a \to \infty$ because of $\mathbb{E}U^2 < \infty$, and combining with $p_+ \sim \mathbb{E}\Gamma(U)/k$ shows that (8.3) holds. □

Next consider the scaling case $T = kT_*$ with $A_*(t) = \mathbb{P}(T_* \le t)$ independent of k.

Corollary 8.6 *Assume $T = kT_*$ where $\mathbb{P}(T_* \le t) \sim ct^\alpha$, $t \downarrow 0$. Then W and $(U-T)^+$ are light traffic equivalent provided that $\mathbb{E}U^{\alpha+1} < \infty$, and then*

$$\mathbb{P}(W > 0) \;\sim\; p_+ \;=\; \mathbb{P}(U - T > 0) \sim \frac{c}{k^\alpha}\mathbb{E}U^\alpha \tag{8.9}$$

Proof. For $y > 0$,

$$\mathbb{P}(U - kT_* > y) \;=\; \int_y^\infty A_*\Big(\frac{u-y}{k}-\Big)B(\mathrm{d}u).$$

Letting $y = 0$, we get

$$k^\alpha p_+ \;= \int_0^\infty k^\alpha A_*\Big(\frac{u}{k}-\Big)\, B(\mathrm{d}u) \;\to c\int_0^\infty u^\alpha B(\mathrm{d}u)$$

(using dominated convergence and $c_1 = \sup_t A_*(t-)/t^\alpha < \infty$). These estimates show also that an upper bound for (8.8) is

$$\begin{aligned}\frac{c_1}{k^\alpha}\Big[a\mathbb{E}(U-a)^{+\alpha} + \int_a^\infty \mathbb{E}(U-x)^{+\alpha}\mathrm{d}x\Big]&\\ = \;\frac{c}{k}\Big[a\mathbb{E}(U-a)^{+\alpha} + \frac{1}{\alpha+1}\mathbb{E}(U-a)^{+\alpha+1}\Big]&.\end{aligned}$$

Here $[\cdots] \to 0$ as $a \to \infty$ because of $\mathbb{E}U^{\alpha+1} < \infty$, and combining with $p_+ \sim c\mathbb{E}U^\alpha/k^\alpha$ shows that (8.3) holds. □

Let $B^{(x)}$ denote the overshoot distribution, $\overline{B}^{(x)}(y) = \overline{B}(x+y)/\overline{B}(x)$.

Corollary 8.7 *Assume that there exists a distribution G with finite mean such that $B^{(x)}$ is stochastically dominated by G for all x. Then W and $(U-T)^+$ are light traffic equivalent.*

Proof. For $y > 0$,

$$\mathbb{P}(U - T > y) \;\;=\;\; p_+\mathbb{P}(U > T + y \,|\, U > T) \;\le\; p_+\overline{G}(y).$$

Hence an upper bound for (8.8) is

$$p_+\left[a\overline{G}(a) + \int_a^\infty \overline{G}(x)\,\mathrm{d}x\right].$$

Here $[\cdots] \to 0$ as $a \to \infty$ when $\mu_G = \int_0^\infty \overline{G}(x)\,\mathrm{d}x < \infty$, and therefore (8.3) holds. □

Remark 8.8 Intuitively, what are the reasons that delay occurs in light traffic? Two reasons come immediately to mind: short interarrival times (clustering) or long service times. To make such a study more rigorous, one way is to describe the conditional distribution of U, T given $X = U - T > 0$ For example, in the scaling case $T = kT_*$ in Corollary 8.6, one has in

$M/D/1$ that $U \equiv 1$ (being constant) is unchanged in this distribution, whereas the conditional distribution of T_* is that of T_* given $T_* \le 1/k$ (which is asymptotically the uniform distribution on $(0, 1/k)$) so that delay is caused by short interarrival times. If instead one considers $D/M/1$, $T = k$ is unchanged in the conditional distribution, whereas the conditional distribution of U is that of $U + k$ so that delay is caused by long service times. See further the Problems, which also contain an example (Problem 8.3) where it is necessary to have *both* long service times and short interarrival times if delay is to occur in light traffic. □

Problems

8.1 Show that in Corollary 8.6, one has

$$\mathbb{P}\,(\,U \le u, T_* \le t/k \,|\, U - kT_* > 0) \;\to\; \frac{\int_0^u (y \wedge t)^\alpha B(\mathrm{d}y)}{\int_0^\infty y^\alpha B(\mathrm{d}y)}, \quad 0 < t < u.$$

8.2 Show that if the service time U has a nondecreasing failure rate, then (8.3) holds.

8.3 Take $\mathbb{P}(U > u) = \mathrm{e}^{-u^2}$, $T = kT_*$, $\mathbb{P}(T_* \le t) = \mathrm{e}^{-1/\sqrt{t}}$. Show using Problem 8.2 that (8.3) holds, and that conditionally upon $U - kT_* > 0$, $U/k^{1/5} \xrightarrow{\mathbb{P}} K$, $k^{4/5}T_* \xrightarrow{\mathbb{P}} K$, $U - kT_* \xrightarrow{\mathbb{P}} 0$, where $K = 4^{-2/5}$ is the unique point where $\varphi(z) = z^{-1/2} + z^2$ attains it minimum.

Notes The study of light traffic goes back to Bloomfield and Cox (1972), but the first mathematically more substantial results are those of Daley and Rolski (1984, 1991). The present exposition follows Asmussen (1992b), who also gives further examples and conditions for $\mathbb{E}W^{(k)^p} \sim \mathbb{E}(U - T)^p$, $p > 0$, together with the corresponding asymptotics. See also Sigman (1992) for workloads.

Whitt (1989) suggests approximations in the whole range $\rho \in (0, 1)$ using interpolating between heavy traffic ($\rho \uparrow 1$) and light traffic ($\rho \downarrow 0$); the details involve the explicit solution of $M/M/1$. Further frequently studied topics in light traffic limit theory are Taylor expansions such as $\mathbb{E}W \approx a_1\rho + \cdots + a_n\rho^n$ and, of course, models beyond $GI/G/1$ such as networks. See e.g. Kovalenko (1995) and Baccelli and Schmidt (1996) for these and further subjects.

9 Heavy–Tailed Asymptotics

We now assume that the service time distribution B is heavy–tailed, more precisely that B is long–tailed (for all y, $\overline{B}(x - y)/\overline{B}(x) \to 1$ as $x \to \infty$) and that its stationary excess (integrated tail) distribution $B_0(x) = \int_0^x \overline{B}(y)\,\mathrm{d}y \,/\, \mu_B$ is in the class $\mathscr{S}$ of subexponential distributions (see A5 for these concepts). We will derive tail asymptotics first for the steady–state waiting time W and later, under the added regularity condition $B \in \mathscr{S}^*$ (see (A.5.3)), for the maximal waiting time in a busy cycle (the parallel results for light tails are given in XIII.5 and state that both tails decay

with the same exponential rate). We assume throughout $\rho < 1$. The result on W is as follows:

Theorem 9.1 (a) *Consider a random walk such that $\mu = \mathbb{E}X < 0$ and that $\overline{F}(x) \sim B(x)$, $x \to \infty$, for some distribution B on $(0,\infty)$ which is long–tailed and satisfies $B_0 \in \mathscr{S}$. Then, writing $\overline{F}_I(x) = \int_x^\infty \overline{F}(y)\,\mathrm{d}y$, it holds that*

$$\mathbb{P}(M > x) \;\sim\; \frac{1}{|\mu|}\overline{F}_I(x), \quad x \to \infty; \tag{9.1}$$

(b) *for a $GI/G/1$ queue with $\rho < 1$ and the service time distribution B satisfying the assumptions of* (a)*,*

$$\mathbb{P}(W > x) \;\sim\; \frac{\rho}{1-\rho}\overline{B}_0(x), \quad x \to \infty. \tag{9.2}$$

The proof uses the following lemma:

Lemma 9.2 *Let $Y_1, Y_2, \ldots$ be i.i.d. with common distribution $G \in \mathscr{S}$ and let N be an independent integer–valued r.v. with $\mathbb{E}z^N < \infty$ for some $z > 1$. Then $\mathbb{P}(Y_1 + \cdots + Y_N > u) \sim \mathbb{E}N\,\overline{G}(u)$.*

Proof. Recall from A5 that $\overline{G^{*n}}(u) \sim n\overline{G}(u)$, $u \to \infty$, and that for each $z > 1$ there is a $D < \infty$ such that $\overline{G^{*n}}(u) \le \overline{G}(u) D z^n$ for all u. Therefore we can use dominated convergence with $\sum \mathbb{P}(N = n)\, Dz^n$ as majorant to obtain

$$\frac{\mathbb{P}(Y_1 + \cdots + Y_N > u)}{\overline{G}(u)} \;=\; \sum_{n=0}^{\infty} \mathbb{P}(N = n) \frac{\overline{G^{*n}}(u)}{\overline{G}(u)} \;\to\; \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot n \;=\; \mathbb{E}N.$$

□

For the proof of Theorem 9.1, it is instructive to first consider the $M/G/1$ case where A is exponential with rate β. The Pollaczeck–Khinchine formula states that $W \stackrel{\mathscr{D}}{=} Y_1 + \cdots + Y_K$ where the Y_i have distribution B_0 and K is geometric with parameter ρ, $\mathbb{P}(K = k) = (1-\rho)\rho^k$. Since $\mathbb{E}K = \rho/(1-\rho)$ and $\mathbb{E}z^K < \infty$ whenever $\rho z < 1$, the result follows immediately from Lemma 9.2. The argument for the general random walk or $GI/G/1$ case is similar. In fact, we have a similar representation $M = Y_1 + \cdots + Y_K$ where K is the number of ladder steps and $Y_1, Y_2, \ldots$ are i.i.d. with common distribution $G = G_+/\|G_+\|$. The difficulty is that whereas K is still geometric, then the parameter $\theta = \|G_+\|$ is not explicit as for $M/G/1$, and also it is not a priori clear that the tail behaviour of G is the same as that of B_0.

Write $\overline{G}_+(x) = G_+(x,\infty) = \mathbb{P}(S_{\tau_+} > x,\, \tau_+ < \infty)$ and let μ_{G_-} be the mean of G_-, $U_- = \sum_0^\infty G_-^{*n}$.

Lemma 9.3 $\overline{G}_+(x) \sim \overline{F}_I(x)/|\mu_{G_-}|$, $x \to \infty$.

Proof. By (1.7),

$$\overline{G}_+(x) \;=\; \int_{-\infty}^{0} \overline{F}(x-y)\, U_-(dy).$$

The heuristics is now that the contribution from the interval $(-N, 0]$ to the integral is $\mathrm{O}(\overline{F}(x))$ which by long–tailedness is $\mathrm{o}(\overline{F}_I(x))$, whereas for large y, $U_-(\mathrm{d}y)$ is close to Lebesgue measure on $(-\infty, 0]$ normalized by $|\mu_{G_-}|$ so that we should have

$$\overline{G}_+(x) \;\sim\; \frac{1}{|\mu_{G_-}|}\int_{-\infty}^{0} \overline{F}(x-y)\,\mathrm{d}y \;=\; \frac{1}{|\mu_{G_-}|}\overline{F}_I(x)\,.$$

We now make this precise. If G_- is nonlattice, then by Blackwell's renewal theorem $U_-(-n-1,-n] \to 1/|\mu_{G_-}|$. In the lattice case, we can assume that the span is 1 and then the same conclusion holds since then $U_-(-n-1,-n]$ is just the probability of a renewal at $-n$.

Given ϵ, choose N such that $\overline{F}(n-1)/\overline{F}(n) \le 1+\epsilon$ for $n \ge N$ (this is possible since B long–tailed, cf. A5.1(a)), and that $U_-(-n-1,-n] \le (1+\epsilon)/|\mu_{G_-}|$ for $n \ge N$. We then get

$$\begin{aligned}
&\varlimsup_{x\to\infty} \frac{\overline{G}_+(x)}{\overline{F}_I(x)} \\
&\quad\le\; \varlimsup_{x\to\infty} \int_{-N}^{0} \frac{\overline{F}(x-y)}{\overline{F}_I(x)}\, U_-(\mathrm{d}y) \;+\; \varlimsup_{x\to\infty} \int_{-\infty}^{-N} \frac{\overline{F}(x-y)}{\overline{F}_I(x)}\, U_-(\mathrm{d}y) \\
&\quad\le\; \varlimsup_{x\to\infty} \frac{\overline{F}(x)}{\overline{F}_I(x)}\, U_-(-N,0] \;+\; \varlimsup_{x\to\infty} \frac{1}{\overline{F}_I(x)} \sum_{n=N}^{\infty} \overline{F}(x+n) U_-(-n-1,-n] \\
&\quad\le\; 0 \;+\; \varlimsup_{x\to\infty} \frac{1}{\overline{F}_I(x)}\, \frac{1+\epsilon}{|\mu_{G_-}|} \sum_{n=N}^{\infty} \overline{F}(x+n) \\
&\quad\le\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\, \varlimsup_{x\to\infty} \frac{1}{\overline{F}_I(x)} \int_{N}^{\infty} \overline{F}(x+y)\,\mathrm{d}y \\
&\quad=\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\, \varlimsup_{x\to\infty} \frac{\overline{F}_I(x+N)}{\overline{F}_I(x)} \;=\; \frac{(1+\epsilon)^2}{|\mu_{G_-}|}\,.
\end{aligned}$$

Here in the third step we used that $\overline{B}(x)/\overline{B}_0(x) \to 0$ (since B is long–tailed) and hence $\overline{F}(x)/\overline{F}_I(x) \to 0$, and in the last that $\overline{F}_I$ is asymptotically proportional to $B_0 \in \mathscr{S}$. Similarly,

$$\varliminf_{x\to\infty} \frac{\overline{G}_+(x)}{\overline{F}_I(x)} \;\ge\; \frac{(1-\epsilon)^2}{|\mu_{G_-}|}\,.$$

Letting $\epsilon \downarrow 0$, the proof is complete. □

Proof of Theorem 9.1. We first show part (a). By Lemma 9.3, $\mathbb{P}(Y_i > x) \sim \overline{F}_I(x)/(\theta|\mu_{G_-}|)$. Hence using dominated convergence precisely as for

$M/G/1$, $M = Y_1 + \cdots + Y_K$ yields

$$\mathbb{P}(M > u) \;\sim\; \sum_{k=1}^{\infty}(1-\theta)\theta^k\, k\,\frac{\overline{F}_I(u)}{\theta|\mu_{G_-}|} \;=\; \frac{\overline{F}_I(u)}{(1-\theta)|\mu_{G_-}|}\,.$$

Now just observe that $(1-\theta)|\mu_{G_-}| = (1-\|G_+\|)|\mu_{G_-}| = |\mu|$ by VIII.(2.1).

To get (b) from (a), just observe that

$$\frac{\overline{F}(x)}{\overline{B}(x)} \;=\; \int_0^\infty \frac{\overline{B}(x+y)}{\overline{B}(x)}\,A(\mathrm{d}y) \;\to\; \int_0^\infty 1\cdot A(\mathrm{d}y) \;=\; 1$$

by dominated convergence. This implies $\overline{F}_I(x) \sim \mu_B\overline{B}_0(x)$ and, using $|\mu| = \mu_A - \mu_B$, that

$$\mathbb{P}(W > x) \;=\; \mathbb{P}(M > x) \;\sim\; \frac{\mu_B}{|\mu|}\overline{B}_0(x) \;=\; \frac{\rho}{1-\rho}\overline{B}_0(x)\,. \qquad \square$$

Now consider the cycle maximum. In the random walk case, we consider a reflected version (Lindley process) $\{W_n\}$ starting from $W_0 = 0$ and define the cycle σ as for $GI/G/1$,

$$\sigma \;=\; \inf\{n \ge 1 :\, W_n = 0\} \;=\; \tau_- \;=\; \inf\{n \ge 1 :\, S_n \le 0\}\,.$$

The cycle maximum is

$$M_\sigma \;=\; \max_{0\le n<\sigma} S_n \;=\; \max_{0\le n<\sigma} W_n.$$

Its relevance for extreme value theory has been explained in VI.4, and in fact, VI.4.10 and the following result immediately show that $\max_{0\le k\le n} W_n$ after a suitable normalization has a Fréchet limit distribution as $n\to\infty$ when B is regularly varying (analogously Problem VI.4.1 gives a Gumbel limit when B is heavy–tailed Weibull; it is straighforward to adapt the argument to see that the same is the case for the log–normal distribution).

Theorem 9.4 *Consider a reflected random walk (Lindley process) $\{W_n\}$ such that $\mu = \mathbb{E}X < 0$ and that $\overline{F}(x) \sim B(x)$, $x\to\infty$, for some $B \in \mathscr{S}^*$. Then*

$$\mathbb{P}(M_\sigma > x) \;\sim\; \mathbb{E}\sigma\overline{F}(x), \quad x\to\infty. \tag{9.3}$$

The same conclusion holds for the $GI/G/1$ waiting time when the service time distribution B satisfies $B \in \mathscr{S}^*$.

For the proof, we first introduce some notation. Define

$$\begin{aligned} N_1(x,x_0) &= \#\big\{n < \sigma :\, S_n \le x_0, S_{n+1} > x\big\},\\ p_1(x,x_0) &= \mathbb{P}\big(S_{n+1} > x \text{ for some } n < \sigma \text{ with } S_n \le x_0\big),\\ p_2(x,x_0) &= \mathbb{P}\big(\tau(x) < \sigma, x_0 \le S_{\tau(x)-1} \le x\big)\,. \end{aligned}$$

where $\tau(x) = \inf\{n \ge 1 :\, S_n > x\}$ (note that the definitions of $p_1(x,x_0)$ and $p_2(x,x_0)$ are not symmetric in the sets $[0,x_0]$ and (x_0,∞)). Then

$$p_1(x,x_0) \;\le\; \mathbb{P}(M_\sigma > x) \;\le\; p_1(x,x_0) + p_2(x,x_0)\,. \tag{9.4}$$

Lemma 9.5 $\mathbb{E}N_1(x,x_0) \sim \mathbb{E}\sigma\mathbb{P}(M \le x_0)\overline{F}(x)$.

Proof. Define $C(A) = \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \in A) = \mathbb{E}\sigma\mathbb{P}(W \in A)$. We get

$$\begin{aligned}\mathbb{E}N_1(x,x_0) &= \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x_0, S_{n+1} > x) = \mathbb{E}\sum_{n=0}^{\sigma-1} I(S_n \le x_0)\overline{F}(x-S_n)\\ &= \int_0^{x_0} \overline{F}(x-y)C(\mathrm{d}y) = \mathbb{E}\sigma\int_0^{x_0} \overline{F}(x-y)\mathbb{P}(W \in \mathrm{d}y)\ .\end{aligned}$$

Now just divide by $\overline{F}(x)$ and use $\overline{F}(x-y)/\overline{F}(x) \to 1$ uniformly in $0 \le y \le x_0$, as follows from $1 \le \overline{F}(x-y)/\overline{F}(x) \le \overline{F}(x-x_0)/\overline{F}(x) \to 1$. □

Lemma 9.6 $p_1(x,x_0) \sim \mathbb{E}\sigma\mathbb{P}(W \le x_0)\overline{F}(x)$.

Proof. After $\tau(x)$, the expected time $\{S_n\}$ spends in $(0,x_0)$ before hitting $(-\infty,0]$ is bounded by $a_1 + a_2x_0$. Hence with $\alpha(x,x_0) = (a_1+a_2x_0)\overline{F}(x-x_0)$, we have

$$\begin{aligned}\mathbb{P}\big(N_1(x,x_0) \ge k+1 \,\big|\, N_1(x,x_0) \ge k\big) &\le \alpha(x,x_0),\\ \mathbb{P}\big(N_1(x,x_0) \ge k+1\big) &\le p_1(x,x_0)\alpha(x,x_0)^k,\\ \mathbb{E}\big[N_1(x,x_0);\, N_1(x,x_0) \ge 2\big] &\le \frac{p_1(x,x_0)\alpha(x,x_0)}{1-\alpha(x,x_0)}\ ,\\ p_1(x,x_0) \le \mathbb{E}N_1(x,x_0) &\le p_1(x,x_0) + \frac{p_1(x,x_0)\alpha(x,x_0)}{1-\alpha(x,x_0)}.\end{aligned}$$

Now just note that $\alpha(x,x_0) \to 0$ and use Lemma 9.5. □

Letting first $x \to \infty$ and next $x_0 \to \infty$ in (9.4), the following estimate will complete the proof of Theorem 9.4:

Lemma 9.7 $\lim\limits_{x_0\to\infty} \limsup\limits_{x\to\infty} \dfrac{p_2(x,x_0)}{\overline{F}(x)} = 0.$

The proof is based upon a downcrossing argument. Define $m_+ = \mathbb{E}X_+$, $m_- = \mathbb{E}X_-$ (thus $m = -\mu = m_- - m_+$) and

$$\begin{aligned}D_\sigma(x) &= \mathbb{E}\sum_{n=0}^{\sigma-1} I\big(S_n > x, S_{n+1} \le x\big),\\ D(x) &= \mathbb{E}\sum_{n=0}^{\infty} I\big(S_n > -x, S_{n+1} \le -x\big).\end{aligned}$$

Lemma 9.8 $\lim\limits_{x\to\infty} D(x) = \dfrac{m_-}{m}$.

Proof. Let U denote the occupation (renewal) measure of the random walk, $U(A) = \sum_0^\infty \mathbb{P}(S_n \in A)$. Then (Problem VIII.3.4) $U[x,x+z] \le a_1 + a_2z$ for all x,z and has limit z/m as $z \to -\infty$ in the nonlattice case (that is, $U(\mathrm{d}z - x)$ converges vaguely to Lebesgue measure normalized by m).

Similar estimates as in the proof of the key renewal theorem then yield

$$\begin{aligned} D(x) &= \int_{-x}^{\infty} U(\mathrm{d}y)F(-x-y) \;=\; \int_0^{\infty} U(\mathrm{d}z-x)F(-z) \\ &\to \frac{1}{m}\int_0^{\infty} F(-z)\mathrm{d}z \;=\; \frac{m_-}{m}\;. \end{aligned}$$

The lattice case is similar though easier. □

Proof of Lemma 9.7. By regenerative process theory,

$$\frac{D_\sigma(x)}{\mathbb{E}\sigma} \;=\; \lim_{n\to\infty}\mathbb{P}(W_n > x, W_{n+1}\le x) \;=\; \int_x^{\infty}\mathbb{P}(W\in\mathrm{d}y)F(x-y).$$

Theorem 9.1 makes it plausible that we can replace $\mathbb{P}(W \in \mathrm{d}y)$ by $m^{-1}\overline{F}(y)\,\mathrm{d}y$; for the rigorous proof which indeed uses $B \in \mathscr{S}^*$ in an essential way, see Asmussen *et al.* (2002). We then get

$$\begin{aligned} \frac{D_\sigma(x)}{\mathbb{E}\sigma} &\sim \frac{1}{m}\int_x^{\infty}\overline{F}(y)\mathrm{d}y\int_{-\infty}^{x-y}F(\mathrm{d}z) \\ &= \frac{\overline{F}(x)}{m}\int_{-\infty}^{0}F(\mathrm{d}z)\int_x^{x-z}\frac{\overline{F}(y)}{\overline{F}(x)}\mathrm{d}y \\ &\sim \frac{\overline{F}(x)}{m}\int_{-\infty}^{0}|z|F(\mathrm{d}z) \;=\; \overline{F}(x)\frac{m_-}{m}\;, \end{aligned}$$

where the third step is an easy consequence of long–tailedness.

On the other hand, the overshoot over x after an upcrossing from a level $\le x_0$ converges in distribution to ∞ by long–tailedness, so that the expected subsequent number of downcrossings of level x before $[0,x_0]$ is hit is approximately m_-/m by Lemma 9.8. Hence we get

$$\begin{aligned} \mathbb{E}\sigma\overline{F}(x)\frac{m_-}{m} &\sim D_\sigma(x) \;\ge\; \mathbb{E}N_1(x,x_0)\frac{m_-}{m} + p_2(x,x_0) \\ &\sim \mathbb{E}\sigma\overline{F}(x)\mathbb{P}(M\le x_0)\frac{m_-}{m} + p_2(x,x_0)\;, \\ \limsup_{x\to\infty}\frac{p_2(x,x_0)}{\overline{F}(x)} &\le \mathbb{E}\sigma\mathbb{P}(M>x_0)\frac{m_-}{m}\;. \end{aligned}$$

Let $x_0\uparrow\infty$. □

Notes Theorem 9.1 has a long history associated with the names of (in alphabetical order) von Bahr, Borovkov, Cohen, Pakes and Veraverbeke. These contributions are given a final form in Embrechts and Veraverbeke (1982). There are numerous recent analogues for more general models, e.g. Whitt (2001) and Boxma *et al.* (2002) for many–server queues, Heath *et al.* (1999), Jelenkovic and Momcilovic (2001) and Zwart *et al.* (2003) for fluid queues, and Baccelli *et al.* (1999) and Baccelli and Foss (2003) for (feed-forward) networks. Also tail asymptotics for the busy period has received considerable attention, see Baltrunas *et*

al. (2002) and references therein. For other queue disciplines than FIFO, see the Notes to III.9.

For some remarkable explicit waiting–time distributions in $M/G/1$ with heavy tails, see Abate and Whitt (1999).

Theorem 9.4 was given independently by Samorodnitsky *et al.* (1997), assuming regular variation, and Asmussen (1998a); the latter paper used the "plausible" step in the proof of Lemma 9.7, which was only recently justified by Asmussen *et al.* (2002; in connection with the results of that paper, see also Bertoin and Doney, 1994b, and Asmussen *et al.*, 2003).

A current trend in the literature related to stressing the importance of heavy tails is the study of *long–range dependence* (LRD). In a stationary process setting, this means that the dependence between X_0 and X_t decays slowly; a common precise definition is that $|\mathbb{C}ov(X_0, X_t)|$ is not integrable (note that this is a necessary condition for a CLT for $\int_0^T X_t \, \mathrm{d}t$ with variance constant proportional to $\sqrt{T}$; cf. the Notes to VI.3). Again, statistical studies are taken as the main motivation, but they are far from uncontroversial; see Mikosch and Stărică (2003). LRD is related to *self–similarity*, i.e. the existence of a constant H (the *Hurst parameter*) such that $\{c^{-H} X_{tc}\}_{t \geq 0} \stackrel{\mathscr{D}}{=} \{X_t\}_{t \geq 0}$. The volumes edited by Park and Willinger (2000) and Taqqu *et al.* (2002) may be taken as a starting point for the area. A main example is *fractional Brownian motion* (FBM), a certain Gaussian process with stationary long–range dependent increments; see e.g. Massoulie and Simonian (1999), Norros (2000) and Piterbarg (2001).

The simplest result pointing to the connection between heavy tails and LRD is covariance asymptotics for renewal processes (Daley, 1999). Another simple case is alternating renewal processes where in the notation of VI.2b one of F_0, F_1 is heavy–tailed; this is in turn relevant for fluid models involving on–off sources with heavy–tailed on periods. See Heath *et al.* (1998, 1999).

http://www.springer.com/978-0-387-00211-8

Applied Probability and Queues
Asmussen, S.
2003, XII, 438 p., Hardcover
ISBN: 978-0-387-00211-8