Bhat + Basawa 1992
pp 177-223

10
Inequalities concerning the waiting-time in single-server queues: a survey*

*D. J. Daley, A. Ya. Kreinin, and C. D. Trengove**

ABSTRACT

The range of techniques used to study bounds for the stationary solution W of the equation $W =_d (W + U)_+$, where on the right-hand side the random variables W and U are independent, with $EU < 0$, is reviewed. Some of the paper concerns the random walk with independent identically distributed steps, with generic step U. More useful results are available for the particular queueing context in which $U = S - T$ for generic independent random variables S and T for service and inter-arrival times respectively. There is some study of the question of extreme queueing systems, meaning here, what distributions for S and/or T maximize or minimize the mean waiting-time EW; partial answers are complemented by conjectures. The paper concludes with a discussion of bounds and conjectured bounds and extremal systems for EW in a many-server queueing system $GI/G/k$.

The methods are applicable to somewhat broader questions outside the specific queueing setting.

10.1 INTRODUCTION

The study of approximations and bounds for parameters of interest in queueing theory emerged in a significant way in the mathematical literature through the work of Kingman, in particular Kingman (1962a, 1962b). Before that time, as evidenced for example by Saaty (1961), the emphasis had been on obtaining exact or closed-form expressions. However, alternative approaches were sometimes to be found, especially in the engineering and teletraffic literature, where there was some effort expended in obtaining approximations in specific cases: at least one general principle, the equivalent

* Work done during tenure of Commonwealth Postgraduate Award at the Statistics Department, University of Melbourne.

random method for handling overflow traffic (see e.g. Riordan, 1962, p. 120) had been developed and is still used today.

The present chapter is a revision of an unpublished manuscript (Daley and Trengove, 1977) which in turn drew heavily on an unpublished thesis (Trengove, 1978). The material of the original survey has been supplemented by more recent work, and abbreviated by putting more emphasis on the diverse methodological features of the problem. While it is clear that our discussion is centred on the particular queueing content, the implications are somewhat broader: for example, several of the techniques reviewed here have been successfully applied to another seemingly intractable problem (Daley and Haslett, 1982), and several models are discussed in Stoyan (1983, Chapter 7 in particular).

The specific problem we discuss concerns the stationary waiting time random variable (r.v.) W of a single-server queueing system with arrivals at the successive epochs of a renewal process with generic inter-arrival time r.v. T and independent identically distributed (i.i.d.) service times with generic r.v. S such that

$$\infty > ET > ES \equiv \rho ET > 0, \tag{1.1}$$

i.e. a $GI/G/1$ system. For any real number x, write $x_+ = \max(0, x)$ and $x_- = (-x)_+$. Then it is well known that W and $(W + S + T)_+$ have the same distribution, i.e.

$$W =_d (W + S + T)_+ \tag{1.2}$$

where on the right-hand side, W, S and T are mutually independent r.v.s, and that

$$\text{when (1.1) is satisfied,} \qquad EW < \infty \quad \text{if and only if} \quad ES^2 < \infty \tag{1.3}$$

(Kiefer and Wolfowitz, 1956; see also Lemoine, 1976). In view of (1.3) we assume $ES^2 < \infty$ throughout the discussion concerning EW.

One view of the queueing process embodied in (1.2) is that it is subsumed by the theory of random walks. While some results concerning W can be deduced from that theory, there are other, distinct, characteristics stemming from the representation

$$W =_d (W + U)_+ \qquad \text{where } U \equiv S - T, \tag{1.4}$$

that depend crucially on the r.v. U being the difference of two independent non-negative r.v.s. The behaviour of the positive (negative) tail of U is dominated by the nature of S (resp. T), as is illustrated most graphically when S or T has a negative exponential distribution. Both viewpoints are given emphasis in Chapters 6 and 12 of Feller (1966).

In Section 10.2 we introduce notation and basic relations, and exhibit the general inequalities to be used in subsequent sections. In Sections 10.3–9 we discuss various bounds that have been derived, concentrating on those for the mean EW. In Section 10.10 we consider the possible nature of 'extremal' systems which, subject to the given constraints, may yield the tightest bounds, but the 'conclusions' of this section are conjectures rather than theorems. Section 10.11 contains a discussion of some numerical results and summary comparisons of some of the inequalities.

Stoyan (1977a, 1983) includes several of the bounds given below, but he concentrates on those bounds that are derived via monotonicity and comparison methods.

10.2 FURTHER NOTATION AND BASIC RESULTS

As indicated at (1.4), it is convenient to write $U = S - T$. Let $A(\cdot)$, $B(\cdot)$, $U(\cdot)$ and $W(\cdot)$ denote the distribution functions (d.f.s) of T, S, U and W respectively. We extend Kendall's notation such as $D/G/k$ as follows:

B_α denotes a quasi-Bernoulli r.v. X for which

$$X = \begin{cases} 0 & \text{with probability } 1 - \alpha^{-1}, \\ \alpha EX & \text{with probability } \alpha^{-1} \end{cases},$$

so that $\alpha = EX^2/(EX)^2$; and

$D_{\lim}$ denotes the limit as $n \to \infty$ of a sequence of r.v.s

$$X_n = \begin{cases} EX(1 - \gamma/n) & \text{with probability } n^2/(1 + n^2), \\ EX(1 + n\gamma) & \text{with probability } 1/(1 + n^2), \end{cases}$$

so that for all $n > 0$, $EX_n = EX$ and $EX_n^2 = (EX)^2(1 + \gamma^2)$.

For both B_α and $D_{\lim}$, EX and $\alpha = EX^2/(EX)^2 = 1 + \gamma^2$ are taken as prescribed in the given context. Notation such as $W(T, S)$ or $EW(T, S; k)$ or $W(GI/G/k)$ or $EW(GI/M/1)$ is sometimes useful.

In view of (1.3), we can take expectations in (1.2) and write

$$EW = E(W + U)_+ = EW + EU + E(W + U)_-$$

so

$$E(T - S - W)_+ = E(-U - W)_+ = E(W + U)_- = E(-U) = (1 - \rho)ET. \tag{2.1}$$

If it is assumed that EW^2 and EU^2 are both finite, a similar argument, using $x^2 = x_+^2 + x_-^2$, yields

$$EW^2 = E(W + U)_+^2 = E(W + U)^2 - E(W + U)_-^2,$$

and thus

$$2E(-U)EW = EU^2 - E(-U-W)_+^2 = EU^2 - E(T-S-W)_+^2 \quad (2.2a)$$

$$= \operatorname{var} U - \operatorname{var}(T-S-W)_+. \quad (2.2b)$$

Higher-order analogues of (2.2) can be written down (see Lemoine, 1976): here we note that the relation

$$3E(-U)EW^2 = 3EU^2EW + EU^3 + E(T-S-W)_+^3 \quad (2.3a)$$

can be expressed in the form

$$3E(-U)\operatorname{var} W = 3E(-U)\left(\frac{EU^2}{2EU}\right)^2 + EU^3 + 3E(-U)\operatorname{var} Z \quad (2.3b)$$

where the r.v. Z has $\Pr\{Z > y\} = E((T-S-W)_+ - y)_+/E(-U)$, and thus

$$\operatorname{var} Z = \frac{E(T-S-W)_+^3}{3E(-U)} - \left(\frac{E(T-S-W)_+^2}{2EU}\right)^2$$
$$\geqslant \frac{1}{3}\left(\frac{E(T-S-W)_+^2}{2EU}\right)^2 = \frac{(EZ)^2}{3}.$$

A result of which (1.3) is a special case is that for $r \geqslant 0$,

$$\text{when} \quad ET < \infty, \qquad EW^r < \infty \quad \text{if and only if} \quad ES^{r+1} < \infty \quad (2.4)$$

(Kiefer and Wolfowitz, 1956), so the derivation above of (2.2) holds only for $ES^3 < \infty$. Nevertheless a truncation argument (Kingman, 1970) based on monotonicity properties (Daley and Moran, 1968) shows that (2.2) holds merely when $ES^2 < \infty$: replace the generic r.v. S by $S_K \equiv \min(S, K)$ for some (large) finite K and write W_K for an r.v. satisfying $W_K =_d (W_K + S_K - T)_+$. Then since $S_K \leqslant_d S_{K+1} \leqslant_d S$, $W_K \leqslant_d W_{K+1} \leqslant_d W$, and since as $K \to \infty$, S_K converges weakly to S and $ES < \infty$, W_K converges weakly and monotonically to W. Further, by monotone convergence,

$$\lim_{K\to\infty} EW_K = \lim_{K\to\infty} \int_0^\infty \Pr\{W_K > u\}\, du = \int_0^\infty \Pr\{W > u\}\, du = EW.$$

Similar monotonicity arguments apply to the other terms in (2.2a).

The right-hand side of (2.2) is meaningful as it stands only if $ET^2 < \infty$ but it can be rewritten in a form, useful for another purpose in which it is not necessary that $ET^2 < \infty$ (see Section 10.10 for interpretation of the condition $ET^2 < \infty$). Replace T by $T_K = \min(T, K)$ with K sufficiently large that $ET_K > ES$, and write W_K much as before, so that from (2.2)

$$\begin{aligned}2E(-U_K)EW_K &= ES^2 - 2ESET_K \\ &\quad + E([T_K + (T_K - W_K - S)_+][T_K - (T_K - W_K - S)_+]) \\ &\leqslant ES^2 - 2ESET_K + E(2T_K(W_K + S)) \\ &= ES^2 - 2ESET_K + 2ET_K(EW_K + ES) \\ &= ES^2 + 2ET_KEW_K. \end{aligned} \tag{2.5}$$

Now $T_K \leqslant_d T$ implies $W_K \geqslant_d W$, and monotonicity arguments as before show that the term ET_KEW_K is uniformly bounded in K. Thus, (2.2) may be replaced by

$$\begin{aligned}2E(-U)EW &= ES^2 - ESET \\ &\quad + E(T[W - (W + S - T)_+]) + E((T - W - S)_+(W + S)).\end{aligned} \tag{2.6}$$

We have as consequences of Spitzer's identity (see e.g. Kingman, 1962a; Cohen, 1969)

$$\Pr\{W = 0\} = \exp\left(-\sum_{n=1}^{\infty} \frac{\Pr\{U_1 + \cdots + U_n > 0\}}{n}\right), \tag{2.7a}$$

$$EW = \sum_{n=1}^{\infty} \frac{E(S_1 + \cdots + S_n - T_1 - \cdots - T_n)}{n} = \sum_{n=1}^{\infty} \frac{E(U_1 + \cdots + U_n)_+}{n}, \tag{2.7b}$$

$$\operatorname{var} W = \sum_{n=1}^{\infty} \frac{E(S_1 + \cdots + S_n - T_1 - \cdots - T_n)_+^2}{n}, \tag{2.7c}$$

where $\{S_n\}$ and $\{T_n\}$ are independent sequences of i.i.d. r.v.s with $S_n =_d S$, $T_n =_d T$ and $U_n = S_n - T_n$.

For the sake of completeness we gather here the various inequalities used. Throughout, we assume that the expectations are well defined. The *Cauchy–Schwarz inequality* hardly needs mention, except to recall that it applies equally to conditional expectations: for any r.v.s X, Y,

$$E(X^2 \mid Y) \geqslant_{\text{a.s.}} (E(X \mid Y))^2, \tag{2.8}$$

with equality if and only if $\Pr\{X = EX \mid Y\} =_{\text{a.s.}} 1$ in Y. An equivalent form of (2.8) is

$$\operatorname{var}(X \mid Y) \geqslant_{\text{a.s.}} 0.$$

Jensen's inequality is almost as well known: if $f(\cdot)$ is a convex function, then

$$f(E(X \mid Y)) \leqslant_{\text{a.s.}} E(f(X) \mid Y), \tag{2.9}$$

with equality if and only if $f(\cdot)$ is linear over the support of the r.v. X conditional on Y. In the present context, note that the function

$$f(x) = x_+ \equiv \max(0, x) \tag{2.10}$$

is convex in x as is $x_+^2 \equiv (x_+)^2$.

Tchebychef's monotonic function covariance inequality asserts that if $f(\cdot)$ and $g(\cdot)$ are non-decreasing functions then

$$E(f(X)g(X)) \geqslant Ef(X)Eg(X), \tag{2.11}$$

with equality if and only if $f(X)$ is a.s. constant. Obviously, if $f(\cdot)$ is non-decreasing and $g(\cdot)$ non-increasing then the inequality at (2.11) is reversed.

For *tails of non-negative r.v.s* we use the two inequalities that for $y \geqslant 0$ and $\alpha = EX^2/(EX)^2$,

$$E(X - y)_+^2 \geqslant (\alpha EX - y)_+^2/\alpha, \tag{2.12}$$

equality holding for all $y \geqslant 0$ only if $\Pr\{X = 0 \text{ or } \alpha EX\} = 1$, i.e. the distribution of X is of the B_α type; and

$$E(X - y)_+^2 \geqslant \alpha(E(X - y)_+)^2, \tag{2.13}$$

with equality for all $y \geqslant 0$ only if X is of the B_α type (see Daley, 1977, for proof of a generalization of (2.13)).

10.3 INEQUALITIES DERIVED FROM (2.2)

Kingman (1962a) used the non-negativity of the variance term

$$\operatorname{var}(T - S - W)_+$$

in (2.2b) to deduce

INEQUALITY I. $2(1 - \rho)ETEW \leqslant \operatorname{var} U = \operatorname{var} S + \operatorname{var} T.$ (3.1)

He further showed (Kingman, 1962b) that this inequality is sharp in heavy traffic, i.e. for a suitable sequence of r.v.s $(W^{(\nu)}, S^{(\nu)}, T^{(\nu)})$ satisfying (1.1)–(1.3), and such that $\rho^{(\nu)} \uparrow 1$, $\operatorname{var} S^{(\nu)} \to \operatorname{var} S$, $\operatorname{var} T^{(\nu)} \to \operatorname{var} T$ and $ET^{(\nu)} \to ET$, it is true that

$$2ET(1 - \rho^{(\nu)})EW^{(\nu)} \to \operatorname{var} S + \operatorname{var} T \qquad (\nu \to \infty). \tag{3.2}$$

The inequality (3.1) has been sharpened without using any additional parameters in two ways, by Stoyan (1972) using a convexity argument (see Section 10.4 below), and by Daley (1977) as follows:

$$\begin{aligned}
E(T-S-W)_+^2 &= E[E((T-S-W)_+^2 \mid S, W)] \\
&\geqslant E[\{ET^2/(ET)^2\}\{E((T-S-W)_+ \mid S, W)\}^2], \quad \text{by (2.13)}, \\
&\geqslant \{ET^2/(ET)^2\}\{E[E((T-S-W)_+ \mid S, W)]\}^2, \\
&\qquad\qquad \text{by the Cauchy–Schwarz inequality}, \\
&= (1-\rho)^2 ET^2, \qquad \text{by (2.1)}. \qquad (3.3)
\end{aligned}$$

Combining (3.3) with (2.2a) yields

INEQUALITY II. $2(1-\rho)ETEW \leqslant \operatorname{var} S + [1-(1-\rho)^2] \operatorname{var} T.$ (3.4)

Observing that $(T-S-W)_+ = (U_- - W)_+$, careful perusal of the argument leading to (3.3) shows that it may be replaced by

$$E(T-S-W)_+^2 = E(U_- - W)_+^2 \geqslant \{EU_-^2/(EU_-)^2\}(EU)^2. \qquad (3.5)$$

Thus, a tighter bound than (3.4) is the following (Daley, 1977):

INEQUALITY III.

$$\begin{aligned}
2(1-\rho)ETEW &\leqslant \operatorname{var} S + \operatorname{var} T - (1 - EU_+/EU_-)^2 \operatorname{var} U_- \\
&= \operatorname{var} U - (EU/EU_-)^2 \operatorname{var} U_-. \qquad (3.6)
\end{aligned}$$

Of course, this tighter bound is obtained at the cost of requiring information on the moments of $U_- = (T-S)_+$ rather than of the separate components T and S. Since the distribution of W depends directly on U and only indirectly on S and T, inequalities like (3.6) can be sufficiently close as to be useful as approximations (see Section 10.11 below, and Daley and Haslett, 1982).

Now write

$$\begin{aligned}
E(W+U)_-^2 &= E((W+U)^2 \mid W+U<0) \Pr\{W+U<0\} \\
&= (\operatorname{var}(W+U \mid W+U<0) + [E(W+U \mid W+U<0)]^2) \\
&\qquad \times \Pr\{W+U<0\} \\
&= \operatorname{var}(W+U \mid W+U<0) \Pr\{W+U<0\} + \frac{[E(W+U)_-]^2}{\Pr\{W+U<0\}}. \qquad (3.7)
\end{aligned}$$

By neglecting the variance term and using (2.1), we have

$$E(W+U)_-^2 \geqslant \frac{(EU)^2}{\Pr\{W+U<0\}} = \frac{(EU)^2}{\Pr\{W=0\} - \Pr\{W+U=0\}}, \qquad (3.8)$$

and therefore, using (2.2),

INEQUALITY IVa. $2(1-\rho)ETEW \leqslant \operatorname{var} U - (p_{00}^{-1} - 1)(EU)^2$ (3.9)

where
$$p_{00} = \Pr\{W = 0\} - \Pr\{W + U = 0\} \leqslant p_0 \equiv \Pr\{W = 0\}.$$

This then gives

INEQUALITY IVb. $2(1-\rho)ETEW \leqslant \operatorname{var} U - (p_0^{-1} - 1)(EU)^2.$ (3.10)

Finally, since $p_{00} = \Pr\{W + U < 0\} \leqslant p_U \equiv \Pr(U < 0\}$, we also have

INEQUALITY IVc. $2(1-\rho)ETEW \leqslant \operatorname{var} U - (p_U^{-1} - 1)(EU)^2.$ (3.11)

Equation (3.9) comes from Daley (1976), (3.10) from Lemoine (1976), and (3.11) from Stoyan and Stoyan (1976). Combining (3.8) with $E(U+W)^2 \geqslant EU_-^2 \Pr\{W = 0\}$ leads to

$$E(U + W)_-^2 \geqslant E(-U)\sqrt{(EU_-^2)},$$

and thus

INEQUALITY V.
$$2(1-\rho)ETEW \leqslant \operatorname{var} U - E(-U)(\sqrt{(EU_-^2)} - E(-U)) \quad (3.12a)$$
$$\leqslant \operatorname{var} U - E(-U)EU_+. \quad (3.12b)$$

The inequalities at (3.12) and (3.6) use the same information concerning U and U_-; computation shows that the latter is tighter if and only if $(EU_-)^4 \geqslant (EU)^2 EU_-^2$. Rossberg (1968) used characteristic function techniques to deduce the left-hand side of the summary below of our intermediate inequalities.

INEQUALITY VI.
$$\frac{(EU)^2}{EU^2 - 2E(-U)EW} = \frac{(EU)^2}{E(U+W)_-^2} \leqslant \Pr\{W = 0\} \leqslant \frac{E(U+W)_-^2}{EU_-^2}. \quad (3.13)$$

Except for (3.13), our bounds to date have been upper bounds on EW. Referring to the alternative form (2.6) of (2.2), observe that since $EW = E(W + S - T)_+$),

$$E[T(W - (W + S - T)_+)] = \operatorname{cov}(T, W - (W + S - T)_+),$$

and because $E(W - (W + S - T)_+ | T = t)$ is an increasing function of t, this covariance expression is non-negative. The last term in (2.6) is manifestly non-negative, and since $EW > 0$, we have

INEQUALITY VII. $2(1-\rho)ETEW \geqslant (ES^2 - ESET)_+$. (3.14)

Stoyan and Stoyan (1974) deduced (3.14) by an extremal argument; Mori (1975) deduced it independently using ergodic arguments based partly on Brumelle (1971b). We now use probabilistic reasoning based on the above derivation of (3.14) to establish, as Ott (1987) showed by more analytic argument, that (3.14) is tight.

Denote by $\mathscr{Q}_2$ the class of $GI/G/1$ systems with specified moments ES, ET, ES^2, ET^2.

Theorem 3.1: In the class $\mathscr{Q}_2$, either
(i) $ES^2 < ESET$ and $\inf_{Q_2} EW = 0$; or
(ii) $ES^2 \geqslant ESET$ and $2(1-\rho)ET \inf_{\mathscr{Q}_2} EW = ES^2 - ESET$. Further, this infimum is attained only in $D/G/1$ systems that also satisfy

$$\sum_{j=0}^{\infty} \Pr\{S = jET\} = 1. \tag{3.15}$$

Proof. An intermediate step in the proof of (ii) is the utilization of the property that in $D/G/1$ queue satisfying (3.15),

$$2(1-\rho)ETEW = ES^2 - ESET \tag{3.16}$$

(this is the discrete queue analogue of the Pollaczek–Khintchine equation for EW in $M/G/1$). Consider a sequence $\{W^{(n)}\} \equiv \{W(T^{(n)}, S)\}$ of stationary waiting-time r.v.s derived from the class $\mathscr{Q}_2$ with S having a distribution as at (3.15) and with the members of the sequence $\{T^{(n)}\}$ having the distribution of X_n describing $D_{\lim}$ at the beginning of Section 10.2. Observe that the typical term in (2.7b) for $EW^{(n)}$ is bounded as in

$$\begin{aligned} E(S_1 + \cdots + S_r - rET(1-\gamma/n))_+ &\geqslant E(S_1 + \cdots + S_r - T_1^{(n)} - \cdots - T_r^{(n)})_+ \\ &\geqslant E(S_1 + \cdots + S_r - rET)_+ \end{aligned} \tag{3.17}$$

and consequently, by appealing to (2.7b), it follows that

$$EW^{(n)} \equiv EW(T^{(n)}, S) \to EW(ET, S) \qquad (n \to \infty), \tag{3.18}$$

i.e. the bound in (3.14) is tight as asserted in case (ii).

Referring to case (i), observe that r.v.s S, T exist for the class $\mathscr{Q}_2$ such that S is a two-point distribution on 0 and ES^2/ES, and since $ES^2/ES < ET$, T is such that $\Pr\{T \geqslant ES^2/ES\} = 1$. Then $\Pr\{U \leqslant 0\} = \Pr\{S \leqslant T\} = 1$, and $EW = 0$.

In the borderline case $ES^2/ET = ET$ it follows that we cannot have

$\Pr\{S \leqslant T\} = 1$ unless var $T = 0$ and then we are in case (ii), with $ES^2 = ESET$. In case (ii) otherwise, in order to have equality at (3.14) we must have both of

$$\operatorname{cov}(T, W - (W + S - T)_+) = 0 \quad \text{and} \quad E[(T - W - S)_+(W + S)] = 0. \tag{3.19}$$

Now (cf. below (2.11) the covariance term is zero only if either $T =_{\text{a.s.}} \text{const.}$, i.e. $T =_{\text{a.s.}} ET$, or else $E(W - (W + S - T)_+ | T)$ as a function of the r.v. T is constant a.s., which can only be the case if $T \geqslant_{\text{a.s.}} W + S$, which is impossible when $ES^2/ES > ET$ and $W >_{\text{a.s.}} 0$. Further, when $T =_{\text{a.s.}} ET$, the second term at (3.19) equals zero if and only if

$$\Pr\{W + S = 0 \text{ or } W + S \geqslant ET\} = 1. \tag{3.20}$$

Since $\Pr\{W = 0\} > 0$, and $ES < ET$, we must therefore have both $\Pr\{S=0\} > 0$ and $\Pr\{S = 0\} + \Pr\{S \geqslant ET\} = 1$. From the recurrence relation (1.2) and these facts concerning W and S, if s is in the support of S, then the fractional part of ET given by $\{s/ET\}ET$ is also in the support of W and thus of $W + S$ which conflict with (3.20) unless $\{s/ET\} = 0$, i.e. s is an integer multiple of ET. The theorem is proved. ∎

Some of these manipulations can be applied to the expressions at (2.3) for var W. For example, applying (2.13) to $ES = E(T - S - W)_+^2/2E(-U)$ leads via

$$\operatorname{var} Z \geqslant \frac{(EZ)^2}{3} > \frac{1}{3}\left(\frac{(1-\rho)ET^2}{2ET}\right)^2$$

to

$$\operatorname{var} W \geqslant \left(\frac{EU^2}{2EU}\right)^2 - \frac{E(-U)^3}{3E(-U)} + \left(\frac{E(T - S)_+^2}{(E(T - S)_+)^2}\right)^2 \frac{(EU)^2}{12}, \tag{3.21}$$

and the last term is bounded below by

$$\left(\frac{ET^2}{2(ET)^2}\right)^2 \frac{(EU)^2}{3} = \frac{1}{3}\left(\frac{(1 - \rho)ET^2}{2ET}\right)^2.$$

For an upper bound, write (2.3a) as

$$3E(-U) \operatorname{var} W = 3EW[EU^2 - E(-U)EW] + EU^3 + E(T - S - W)_+^3.$$

Using $x(a - x) \leqslant a^2/4$ and $E(T - S - W)_+^3 E(T - S)_+^3 \leqslant E(S - T)_+^3 - EU^3$ it follows that

$$\operatorname{var} W \leqslant \left(\frac{EU^2}{2EU}\right)^2 + \frac{EU_+^3}{3E(-U)}. \tag{3.22}$$

Other expressions are given in Mori (1975) and Fainberg (1979) while still further manipulations analogous to those above for EW are possible.

10.4 BOUNDS VIA CONVEXITY ARGUMENTS

The first bounds in this section come from Jensen's inequality (2.8) and the closure property of mixtures of convex functions, namely that

$Ef(x, Y)$ is a well-defined function convex in x if $f(x, y)$ is convex in x for every y and $E\max\{0, -f(x, Y)\} < \infty$ for all x. (4.1)

Application of (4.1) shows that the function

$$f(w) \equiv E((w + U)_+ - w) = EU + E(w + U)_- \tag{4.2}$$

is convex in w. By inspection, $f(0) > 0$ while $\lim_{w\to\infty} f(w) < 0$ since $E(w + U)_- \to 0$ $(w \to \infty)$ and $EU < 0$. Because of (1.2), $Ef(W) = 0$, so by applying Jensen's inequality to $f(W)$ we get

$$0 \geqslant E(EW + U)_+ - EW. \tag{4.3}$$

The convexity of $f(\cdot)$ and $f(0) > 0 > f(\infty)$ now gives Marshall's (1968) result:

INEQUALITY VIII. $$EW \geqslant \xi \tag{4.4}$$

where ξ is the root of

$$\xi = E(\xi + U)_+; \tag{4.5a}$$

an equivalent form of (4.5a) is that

$$E(-U) = ET(1 - \rho) = E(T - S - \xi)_+. \tag{4.5b}$$

The function

$$g(w, u) = w^2 - (w + u)_+^2 = -2wu - u^2 + (w + u)_-^2 \tag{4.6}$$

is convex in w for each u so by (4.1),

$$g(w) \equiv E(w^2 - (w + U)_+^2) \tag{4.7}$$

is convex. Using (1.2) and Jensen's inequality much as at (4.3),

$$0 = EW^2 - E(W + U)_+^2 = Eg(W) \geqslant (EW)^2 - E(EW + U)_+^2. \tag{4.8}$$

Now $g(0) < 0$ and $g(w) \to \infty$ for $w \to \infty$, hence

INEQUALITY IX. $$EW \leqslant \eta \tag{4.9}$$

where η is the positive root of

$$\eta^2 = E(\eta + U)_+^2 = E(\eta + S - T)_+^2 \tag{4.10a}$$

which, like (4.5a), has an equivalent form,

$$2E(-U)\eta = EU^2 - E(\eta + U)_+^2. \tag{4.10b}$$

The inequality (4.9) is due to Calo and Schwarz (1977); the derivation above comes from Daley (1976). Note that this upper bound, unlike the bounds in Section 10.3, is finite irrespective of ET^2 being finite or infinite.

Stoyan (1972) used a convexity argument of a rather different kind to refine Kingman's bound (3.1). Writing $W = W(T, S)$, consider, for certain $\gamma > 0$ including $0 < \gamma \leqslant 1$,

$$W_\gamma \equiv W(T, \gamma S).$$

Then by (2.7a),

$$w(\gamma) \equiv EW_\gamma = \sum_{n=1}^{\infty} \frac{E(\gamma(S_1 + \cdots + S_n) - (T_1 + \cdots + T_n))_+}{n} \tag{4.11}$$

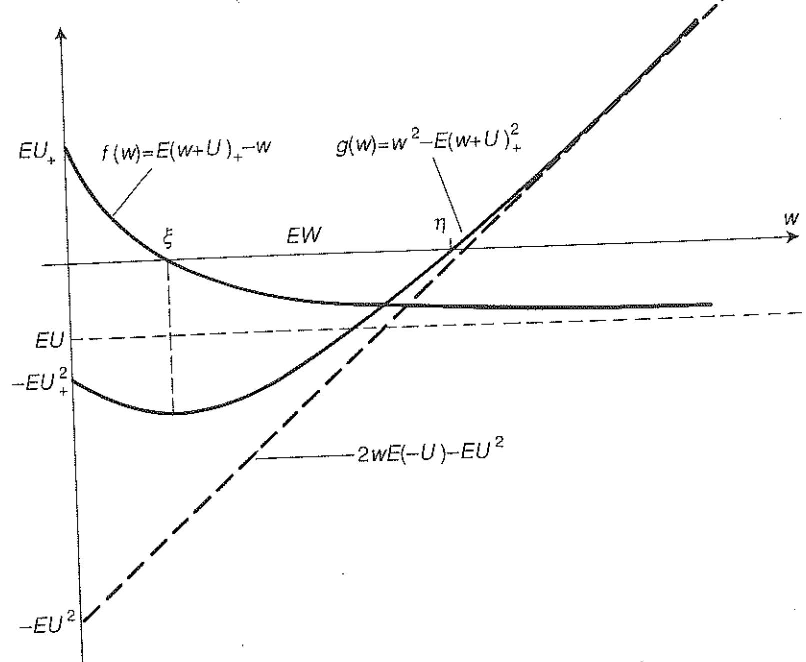

Fig. 10.1 Lower and upper bounds via convexity properties of

$$f(w) = E(w + U)_+ - w, \qquad \text{and} \qquad g(w) = w^2 - E(w + U)_+^2.$$

Note that $g'(w) = -2f(w)$.

is a convex function of γ, and $w(\gamma) \to 0$ as $\gamma \to 0$. By (3.1), provided $\gamma < \rho^{-1}$,

$$w(\gamma) \leqslant \frac{\gamma^2 \operatorname{var} S + \operatorname{var} T}{2(1-\gamma\rho)ET} \equiv w_1(\gamma), \quad \text{say,} \tag{4.12}$$

which has the limit var $T/2ET$ as $\gamma \to 0$. Consequently, the largest convex function $w_0(\gamma)$ with $w_0(0) = 0$ and $w_0(\gamma) \leqslant w_1(\gamma)$ $(0 < \gamma \leqslant \rho^{-1})$ is an upper bound to the convex function $w(\gamma) = EW_\gamma$. In particular, if $w_0(1) < w_1(1)$, then (3.1) is refined. Elementary calculus and algebra shows that $w_0(1) < w_1(1)$ if and only if var $S < (1 - 2\rho)$ var T, in which case

$$EW < w_0(1) = \frac{\operatorname{var} T}{ET}(\rho + \sqrt{(\rho^2 + \operatorname{var} S/\operatorname{var} T)}). \tag{4.13}$$

Figure 10.2 illustrates the argument which is potentially applicable to refining any upper bound on EW that is convex in ρ and positive for $\rho = 0$. Thus, the argument does not apply usefully to the refinement at (3.4). Indeed, (3.4) is always smaller than $w_0(1)$ at (4.13).

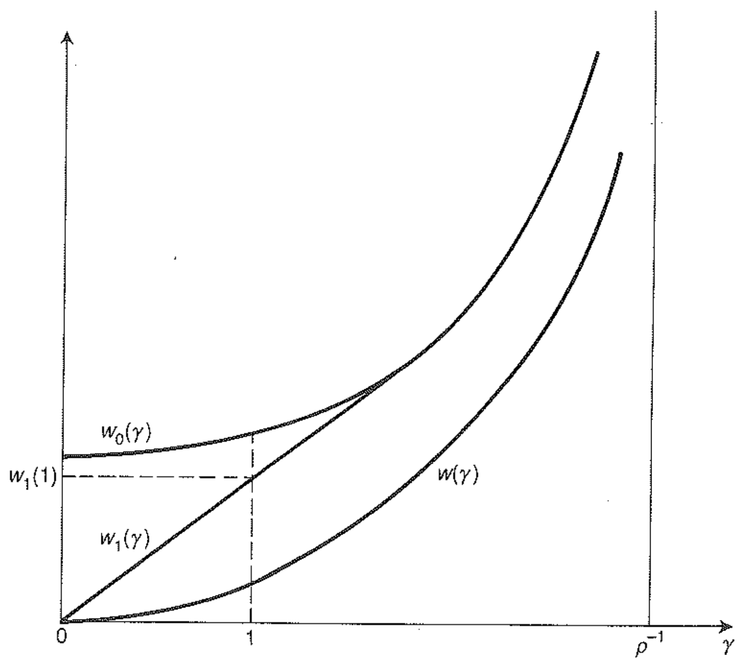

Fig. 10.2 Stoyan's curves for improving bounds via convexity.

10.5 RANDOM WALK METHODS

Given a sequence $\{X_n: n = 1, 2, \ldots\}$ of i.i.d. r.v.s with $EX < 0$, define

$$Y_0 = 0, \quad Y_n = X_1 + \cdots + X_n, \quad \text{and} \quad M = \sup_{n \geq 0}\{Y_n\}. \tag{5.1}$$

Then $M =_d (M + X)_+$. Consequently, as is emphasized in Feller's (1966) comments on queueing theory, much of the study of the stationary waiting-time r.v. W is just part of the study of the r.v. M, as for example the identities at (2.7).

Applying Jensen's inequality to the terms in (2.7a) yields

INEQUALITY Xa.

$$EW \equiv EW(T, S) \geq \sum_{n=1}^{\infty} \frac{E(nES - T_1 - \cdots - T_n)_+}{n} = EW(T, ES), \tag{5.2a}$$

INEQUALITY Xb.

$$EW(T, S) \geq \sum_{n=1}^{\infty} \frac{E(S_1 + \cdots + S_n - nET)_+}{n} = EW(ET, S), \tag{5.2b}$$

with equality holding if and only if either $W =_{\text{a.s.}} 0$, which is equivalent to $S \leq_{\text{a.s.}} T$, or else at (5.2a) $S =_{\text{a.s.}} ES$, or else at (5.2b) $T =_{\text{a.s.}} ET$. This proves Theorem 5.1, more briefly than in Rogozin's (1966) original derivation.

Theorem 5.1 (Rogozin): Within the class of $GI/G/1$ queueing systems prescribed by given ES and ET, EW is least for any given distribution of S when $T =_{\text{a.s.}} ET$, and EW is least for any given distribution of T when $S =_{\text{a.s.}} ES$.

An extension of Rogozin's result to single-server queues with stationary ergodic (not necessarily renewal) arrival process is given in Daley and Rolski (1992a). The proof exploits the representation for M at (5.1), the independence of the sequences $\{S_i\}$ and $\{T_i\}$, and Jensen's inequality. Indeed, the availability of the representation (5.1) is basic to this study of the finiteness of moments of W in such queueing systems, because it can be coupled on the one hand with Jensen's inequality as in the proof of Rogozin's theorem to establish necessary conditions for finiteness, and on the other, in showing the sufficiency of conditions, via the following decomposition inequality (cf. e.g. Sacks, 1960; Wolfson, 1984), valid for any η in $ES < \eta < ET$:

$$\begin{aligned} W =_d \sup_{n \geq 0}\left\{\sum_{i=1}^{n} (S_i - T_i)\right\} &= \sup_{n \geq 0}\left\{\sum_{i=1}^{n} (S_i - \eta + \eta - T_i)\right\} \\ &\leq \sup_{n \geq 0}\left\{\sum_{i=1}^{n} (S_i - \eta)\right\} + \sup_{n \geq 0}\left\{\sum_{i=1}^{n} (\eta - T_i)\right\}. \end{aligned}$$

Identifying these terms as waiting-time r.v.s, taking expectations and introducing a suffix to recall the traffic intensity, we have

$$EW_\rho(G/G/1) \leqslant EW_{\rho_1}(G/D/1) + EW_{\rho_2}(D/G/1), \qquad \text{where} \quad \rho_1\rho_2 = \rho.$$

Returning to the $GI/G/1$ context, it is a direct result of Rogozin's theorem and (2.2) that

$$\begin{aligned} 2(1-\rho)ETEW(T,S) \geqslant 2(1-\rho)&ETEW(ET,S) \\ &= EU^2 - E(ET - S - W)_+^2 \\ &\geqslant EU^2 - E(ET(ET - S - W)_+) \\ &= EU^2 - ETE(-U) = ES^2 - ESET, \end{aligned}$$

i.e. (3.14) is recovered, though with no hint as to the conditions for equality as given in Theorem 3.1.

Set $p = \Pr\{U \geqslant 0\}$, and write V for a r.v. whose distribution has the tail $\Pr\{V > x\} = \Pr\{U > x \mid U \geqslant 0\}$. Then the relation at (2.7b) yields

$$E(U_1 + \cdots + U_n)_+ \geqslant p^n nEV,$$

and hence

INEQUALITY XI. $$EW \geqslant \sum_{n=1}^{\infty} p^n EV = \frac{pEV}{1-p} = \frac{EU_+}{\Pr\{U < 0\}}. \tag{5.3}$$

This inequality can be tightened by introducing, for example, $p_2 = \Pr\{U_1 + U_2 \geqslant 0\}$ and V_2 for a r.v. for which

$$\Pr\{V_2 > x\} = \Pr\{U_1 + U_2 > x \mid U_1 + U_2 \geqslant 0\}.$$

We omit details.

Ladder variable arguments lead to bounds on higher moments of W as follows. Define

$$N = \begin{cases} 0 & \text{if } U_1 + \cdots + U_n \leqslant 0 \quad (\text{all } n < \infty), \\ \inf\{n\colon U_1 + \cdots + U_n > 0\} & \text{otherwise.} \end{cases}$$

On $\{N > 0\}$, introduce

$$L = U_1 + \cdots + U_N. \tag{5.4}$$

Let $L, L_1, L_2, \ldots$ be an i.i.d. sequence,

$$\pi = \Pr\{N > 0\}, \tag{5.5}$$

and let the r.v. $\nu(\pi)$ be independent of $\{L_i\}$ and have the geometric distribution $\{(1-\pi)\pi^j\}$ on $\{0, 1, \ldots\}$. Then W has the representation

$$W = \sum_{i=1}^{\nu(\pi)} L_i, \tag{5.6}$$

so from (5.5),

$$\Pr\{W > 0\} = \pi. \tag{5.7}$$

Since

$$E\nu(\pi) = \sum_{n=1}^{\infty} j(1-\pi)\pi^j = \frac{\pi}{1-\pi} \qquad \text{and} \qquad E[\nu(\pi)(\nu(\pi)-1)] = \frac{2\pi^2}{(1-\pi)^2},$$

$$EW = ELE\nu(\pi) = \frac{\pi EL}{1-\pi},$$

$$EW^2 = EL^2 E\nu(\pi) + (E[\nu(\pi)(\nu(\pi)-1)] = 2(EW)^2 + \frac{\pi EL^2}{1-\pi}, \tag{5.8}$$

so using $EL^2\pi/(1-\pi) \geqslant (EL)^2\pi/(1-\pi) = (EW)^2[(1-\pi)/\pi]$, we have

INEQUALITY XII. $\quad \operatorname{var} W \geqslant (EW)^2/\Pr\{W > 0\}.$ (5.9)

Indeed, by writing

$$\operatorname{var} W = \operatorname{var}(W \mid W > 0)\Pr\{W > 0\} + (EW)^2\left(\frac{1}{\Pr\{W > 0\}} - 1\right),$$

(5.9) implies

$$\operatorname{var}(W \mid W > 0) \geqslant \frac{(EW)^2}{\Pr\{W > 0\}}. \tag{5.10}$$

Also, since

$$W = \begin{cases} 0 & \text{with probability } 1-\pi, \\ W' + L & \text{with probability } \pi, \end{cases}$$

with W' and L independent, $\operatorname{var}(W \mid W > 0) = \operatorname{var} W + \operatorname{var} L$.

Return to the random walk framework of the beginning of this section, and define

$$N(x) = \begin{cases} 0 & \text{if } Y_n < x \text{ for all } n, \\ \inf\{n\colon Y_n \geqslant x\} & \text{otherwise.} \end{cases} \tag{5.11}$$

Thus, $N =_d N(0+)$. On $\{N(x) > 0\}$, define

$$M(x) = Y_{N(x)}. \tag{5.12}$$

Noting that $N(x)$ is a stopping time, it follows that on $\{N(x) > 0\}$,

$\{X_n\colon n \geqslant N(x)+1\}$ is an i.i.d. sequence with the same distribution as the original sequence $\{X^n\}$, and thus, on $\{N(x) > 0\}$,

$$M =_d M' + M(x) \tag{5.13}$$

where M' has the same distribution as M and is independent of $M(x)$. Consequently,

$$\begin{aligned}
\Pr\{M \geqslant x+y \mid M \geqslant x\} &= \Pr\{M \geqslant x+y \mid N(x) > 0\} \\
&= \Pr\{M' + M(x) \geqslant x+y \mid N(x) > 0\} \\
&= \Pr\{M' \geqslant y - (M(x) - x) \mid N(x) > 0\} \\
&\geqslant \Pr\{M' \geqslant y\} = \Pr\{M \geqslant y\}.
\end{aligned}$$

Since W is identified as a r.v. distributed like M, we have

INEQUALITY XIII. $\Pr\{W \geqslant x+y\} \geqslant \Pr\{W \geqslant x\}\Pr\{W \geqslant y\}$. (5.14)

This relationship shows that the distribution of W has the NWU (new worse than used) property studied in reliability theory. It implies the property, weaker than (5.9), that $EW^2/2 \geqslant (EW)^2$. As Stoyan (1977b) observed, it also implies (cf. Korzeniowski and Opawski, 1976) the distributional inequality

$$\Pr\{W > x\} \leqslant \frac{1}{1 + x/EW}, \tag{5.15}$$

which is tighter than the bound on $\Pr\{W > 0\}$ at (5.18) below when

$$x > \frac{(EU)^2 EW}{\operatorname{var} U - 2E(-U)EW}.$$

Köllerström (1976) used the super-multiplicative property (5.14) to deduce a lower bound for EW that decreases as the third moment (essentially, of T) $E(T-S)_+^3$ increases. This behaviour conflicts with Sahin and Perrakis' (1976) numerically based observation involving $K_m/H_n/1$ systems (K_m denotes rational Laplace transform for T of order m) that with increasing ET^3 (ET, ET^2 and S being fixed), lower bounds worsen and upper bounds improve as approximations for EW. The conflict can be explained in part by using the inequality $E\,\mathrm{e}^{-\theta U_-} \geqslant \mathrm{e}^{-\theta EU_-}$ (θ real) in place of $\mathrm{e}^{-x} \geqslant 1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3$ as used by Köllerström (1976) below his equation (15). Another explanation emerges from the identification in Daley and Rolski (1992a) of large values of EW being associated with increased clustering of the arrival process: from this point of view, values of the third moments of T are interpreted as surrogate measures of clustering.

Since $\lim_{\delta \downarrow 0} \Pr\{W > x - \delta\} = \Pr\{W \geqslant x\}$, (5.14) is equivalent to the original form

$$\Pr\{W > x + y\} \geqslant \Pr\{W > x\}\Pr\{W > y\} \tag{5.14'}$$

established by Köllerström using analytical arguments based on (1.2). Trengove (1978) established (5.14) for any r.v. with the random geometric sum representation at (5.6) noting that for i.i.d. geometric r.v.s like $\nu(\pi)$,

$$\Pr\{\nu(\pi) \geqslant i + j\} = \pi^{i+j} = \Pr\{\nu(\pi) \geqslant i\}\Pr\{\nu(\pi) \geqslant j\},$$

and suggesting that the super-multiplicative property of the tails of the d.f. of W may in part be attributed to the multiplicative property of the tails of the geometrically distributed r.v.s $\nu(\pi)$. The derivation above of (5.14) is essentially due to Asmussen (private communication); a similar argument is also given in Köllerström (1978) (see also Daley, 1984d).

Another use of the fundamental random walk relationship $M = \sup_{n \geqslant 0}\{Y_n\}$ at (5.1) was introduced by Karpelevich and Kreinin (1980) (see also Kreinin, 1979c) as being applicable to those queueing systems (e.g. $E_j/D/1$ or $D/E_j/1$) for which the r.v.s T and S are finitely divisible, i.e. are expressible for some positive integer $k \geqslant 2$ in terms of i.i.d. sequences $\{T'_i\}$, $\{S'_i\}$ as

$$T =_d T'_1 + \cdots + T'_k, \qquad S =_d S_1 + \cdots + S'_k.$$

It is then an immediate consequence of (5.1) that

$$W(T, S) \leqslant_d W(T', S').$$

A similar argument applies to certain $GI/D/k$ queues (see Mori, 1975; Daley and Rolski, 1984b).

Kreinin (1981a) bounded $\Pr\{W > 0\}$ by using the inequality

$$\Pr\{X > x\} \leqslant \frac{1}{1 + x^2/\operatorname{var} X} = 1 - \frac{x^2}{\operatorname{var} X + x^2}, \tag{5.16}$$

valid for any r.v. X with $EX = 0$ and $EX^2 < \infty$ on each term $n = 1, 2, \ldots$ in (2.7a) via

$$n^{-1}\Pr\{U_1 + \cdots + U_n > 0\} = n^{-1}\Pr\{U_1 + \cdots + U_n - nEU > nE(-U)\}$$

$$\leqslant \frac{\operatorname{var} U}{n \operatorname{var} U + n^2(EU)^2}.$$

The resulting bound is not as tight as the next inequality which follows from (3.13) and (3.14) (cf. also Daley, 1984b, c), that

$$\Pr\{W > 0\} \leqslant 1 - \frac{(EU)^2}{EU^2 - 2E(-U)EW} \leqslant 1 - \frac{(EU)^2}{EU^2 - (ES^2 - ESET)_+}. \tag{5.17}$$

INEQUALITY XIV. $\Pr\{W > 0\} \leqslant 1 - \dfrac{(1-\rho)^2(ET)^2}{ET^2 - ESET - (ESET - ES^2)_+}$. (5.18)

In the case of the maximum M of the random walk described at the beginning of this section, we have the tight inequality

$$\Pr\{M > 0\} \leqslant \frac{\operatorname{var} X}{EX^2}. \tag{5.19}$$

However, in not being symmetric in ES^2 and ET^2, (5.18) already reflects the different influence that ES^2 and ET^2 may have on $\Pr\{W > 0\}$. Daley (1984c) showed that (5.18) is tight if either $\operatorname{var} T = 0$ or $\operatorname{var} S = 0$ and conjectured that

$$\sup_{\mathscr{D}_2} \Pr\{W > 0\} = \sup_{\mathscr{D}_2} \Pr\{U \geqslant 0\} = \sup_{\mathscr{D}_2} \Pr\{U = 0\} \tag{5.20}$$

$$= \sup_{ES < x < ET} \left\{ \frac{\operatorname{var} T}{\operatorname{var} T + (ET - x)^2} \min\left(\frac{ES}{x}, \frac{\operatorname{var} S}{\operatorname{var} S + (xES)^2} \right) \right\} \tag{5.21}$$

Observe that (5.17) comes from omitting the variance term in the equality at (3.7). This term can be zero, but essentially this and equality at (3.14) require either $\operatorname{var} S = 0$ or $\operatorname{var} T = 0$ (see Daley (1984c) for details).

The Hájek–Rényi inequality can be used with the bound (5.16) to deduce that

$$\Pr\{W > x\} \leqslant \sum_{k=1}^{\infty} \frac{\operatorname{var} U}{(x + kE(-U))^2} \leqslant \frac{\operatorname{var}(U/EU)}{\frac{1}{2} + x/E(-U)}. \tag{5.22}$$

However, given knowledge of EW which is known to be less than $(\operatorname{var} U)/2E(-U)$, this bound is better than (5.15) only for sufficiently small x (and, indeed, for no $x > 0$ if $\operatorname{var} U > \frac{1}{2}(EU)^2$), namely, for

$$x < \frac{\frac{1}{2}(EU)^2 - \operatorname{var} U}{(\operatorname{var} U)/EW - E(-U)}.$$

Bounds on the kth order cumulants γ_k of W for $k \geqslant 2$ were established in Bergmann *et al.* (1979) via the representation, analogous to (2.7c),

$$\gamma_k = \sum_{n=1}^{\infty} \frac{E(S_1 + \cdots + S_n - T_1 - \cdots - T_n)_+^k}{n}.$$

In particular, they are least in $D_{\lim}/G/1$ and $GI/B_\alpha/1$, and largest in $B_\alpha/G/1$ and $GI/D_{\lim}/1$ (but, they are infinite in $GI/D_{\lim}/1$).

Assuming the necessary moments of S and T are finite, Mihoc (1980) used a Gram–Charlier expansion to derive series expressions for $\Pr\{W = 0\}$, a density for the d.f. $W(\cdot)$ (assuming such a density exists), and approximations for the cumulant of W.

A somewhat different inequality involving random walk arguments and EW occurs in a renewal theory problem in Daley (1980): writing $H(x) = \sum_{n=1}^{\infty} F^{n*}(x)$ where $F(\cdot)$ is the d.f. of $-U$, it is shown that with $\lambda = 1/E(-U)$,

$$|H(x) - \lambda x_+ - \tfrac{1}{2}(\lambda^2 \operatorname{var} U - 1)| \leqslant \tfrac{1}{2}\lambda^2 EU^2 - \lambda EW \leqslant \tfrac{1}{2}\lambda^2 EU_-^2.$$

10.6 MONOTONICITY AND COMPARISON METHODS

An expository account of monotonicity and comparison methods, with particular reference to their queueing applications, is given in Stoyan (1977a, 1983) so here we only outline the results. More recently, Shaked, Shanthikumar and Yao (see e.g. Shanthikumar and Yao (1991) for references and some applications) have considered what they term strong stochastic convexity.

Referring to (1.2), it is sought to establish a result of the form $W' \prec W''$ from a condition like $S' \prec S''$ or $-T' \prec -T''$ where $\prec$ denotes a partial ordering. When the ordering is either the stochastic (or, distributional) ordering $\leqslant_d$ or the convex ordering $\leqslant_c$ defined for r.v.s X, Y by

$$X \leqslant_d Y \quad \text{if} \quad \Pr\{X > t\} \leqslant \Pr\{Y > t\} \quad (\text{all } t),$$
$$X \leqslant_c Y \quad \text{if} \quad E(X - t)_+ \leqslant E(Y - t)_+ \quad (\text{all } t),$$

such implications hold because these orderings are preserved under addition of independent r.v.s and truncation, and these are precisely the two operations involved in the right-hand side of the recurrence relation underlying (1.2). Convex ordering is appropriate for establishing inequalities on the mean EW, such as Rogozin's results in Theorem 5.1 (Stoyan and Stoyan, 1969; Borovkov, 1970 (reproduced in Borovkov, 1976, §24)), and Marshall's inequality at (4.4) (see Stoyan and Stoyan, 1974). It does not seem possible to establish (4.9) using $\leqslant_c$ because the square $E(\eta + U)_+^2$ is involved rather than $E(\xi + U)_+$ as at (4.5).

Stochastic ordering has been used by Kingman (1970) in the truncation arguments as around (2.3) and also in developing distributional inequalities of the form

INEQUALITY XVa. $a\,\mathrm{e}^{-\tilde{\theta}x} \leqslant \Pr\{W > x\} \leqslant b\,\mathrm{e}^{-\tilde{\theta}x} \quad (x > 0)$ (6.1)

when $\tilde{\theta} \equiv \sup\{\theta \geqslant 0: E\,\mathrm{e}^{\theta U} \leqslant 1\} < 0$, the constants a and b being identified via martingale methods (Kingman, 1970; Ross, 1974; see also §7 below) as

$$a = \inf_{x>0}\{1/E(e^{\theta(U-x)}|U > x)\}, \tag{6.2a}$$

$$b = \sup_{x>0}\{1/E(e^{\theta(U-x)}|U > x)\}. \tag{6.2b}$$

Bergmann and Stoyan (1976) used the recurrence relation (1.2) and a comparison argument to give another derivation of (6.2b). From (6.1) we have

INEQUALITY XVb. $$a/\tilde{\theta} \leqslant EW \leqslant b/\tilde{\theta}, \tag{6.3a}$$

INEQUALITY XVc. $$a \leqslant \Pr\{W > 0\} \leqslant b. \tag{6.3b}$$

Perhaps the earliest use of stochastic ordering was in Lindley (1952) who, generating a sequence of r.v.s $\{W_n: n = 0, 1, \ldots\}$ via

$$W_{n+1} = (W_n + U_n)_+, \qquad W_0 = 0,$$

observed that

$$W_n \leqslant_d W_{n+1} \qquad (\text{all } n = 0, 1, \ldots).$$

In particular, $W_1 \leqslant_d U_+$, so since $W \geqslant_d W_1$,

INEQUALITY XVI. $$\begin{aligned} 2(1-\rho)ETEW &\geqslant EU^2 - E(-U' - U'')_+^2 \\ &= EU^2 - E(U'_- - U'')_+^2 \end{aligned} \tag{6.4}$$

where U', U'' are i.i.d. like U (Calo and Schwarz, 1977; Daley, 1976). Kingman (1970) gave the weaker result, stemming from $W \geqslant_{\text{a.s.}} 0$,

INEQUALITY XVII. $$2(1-\rho)ETEW \geqslant EU^2 - EU_-^2 = EU_+^2. \tag{6.5}$$

It is worth noting that in systems for which $\Pr\{W = 0\}$ is large (typically, when the traffic intensity ρ is small), (6.4) can be expected to give a reasonable approximation to $2E(-U)EW$, because

$$\begin{aligned} E(W+U)_-^2 = E(U_- - W)_+^2 &= E(U_- - (W + U')_+)_+^2 \\ &\geqslant E((U_- - U'_+)_+^2;\, W = 0) \\ &= E(U_- - U'_+)_+^2 \Pr\{W = 0\}, \end{aligned}$$

whence the 'sandwich' for $2E(-U)EW$:

$$EU^2 - E(U_- - U'_+)_+^2 \Pr\{W = 0\} \geqslant 2E(-U)EW \geqslant EU^2 - E(U_- - U'_+)_+^2.$$

Moreover, the left-hand side here is tightened in

$$2E(-U)EW \leqslant EU^2 - EU_-^2 \Pr\{W=0\} = EU_+^2 + EU_-^2 \Pr\{W=0\}.$$

Bloomfield and Cox (1972) used a comparison argument of a different kind, involving a modification to the queueing discipline. Numbering customers in reverse order of arrival, suppose that customer 0 arrives at time $T_1 + \cdots + T_n$ after customer n whose service time is S_n and that up until the arrival of customer 0 all customers are served immediately on arrival. Then the residual service time at the arrival of customer 0 equals $\sum_{n=1}^{\infty} (S_n - T_1 - \cdots - T_n)_+$, and this represents a lower bound on the waiting time of customer 0 in a first-come–first-served system. Consequently

INEQUALITY XVIIIa. $$W \geqslant_d \sum_{n=1}^{\infty} (S_n - T_1 - \cdots - T_n)_+, \tag{6.6}$$

INEQUALITY XVIIIb.

$$EW \geqslant \sum_{n=1}^{\infty} E(S_n - T_1 - \cdots - T_n)_+ = \int_0^{\infty} (1 - B(u))H(u)\,\mathrm{d}u \tag{6.7}$$

where $H(\cdot) = \sum_{n=1}^{\infty} A^{n*}(\cdot)$ is the renewal function of the arrival process. This work of Bloomfield and Cox is an early example of a light traffic argument (cf. Daley and Rolski, 1991).

10.7 MARTINGALE METHODS

Write $\phi(\theta) = E\,\mathrm{e}^{\theta U}$ for real θ for which the right-hand side is finite, and assume as for (6.1) that

$$\tilde{\theta} \equiv \sup\{\theta \geqslant 0 \colon \phi(\theta) \leqslant 1\} > 0, \tag{7.1}$$

so $\phi(\tilde{\theta}) \leqslant 1$. Then the sequence of r.v.s $\{V_n \colon n = 0, 1, \ldots\}$ defined by

$$V_n = \exp(\tilde{\theta}(U_1 + \cdots + U_n)\}, \tag{7.2}$$

where the r.v.s U_i are i.i.d. like U, is a non-negative super-martingale or indeed a martingale if $\phi(\tilde{\theta}) = 1$. By using Kolmogorov's inequality for (super)-martingales, Kingman (1964) deduced that for all $x \geqslant 0$,

$$\Pr\{V_n \geqslant \exp(\tilde{\theta}x) \text{ for some } n = 1, 2, \ldots\} \leqslant 1/\exp(\tilde{\theta}x),$$

i.e.

$$\Pr\{W \geqslant x\} \leqslant \exp(-\tilde{\theta}x) \qquad (x \geqslant 0). \tag{7.3}$$

Ross (1974) used an appropriately chosen stopping time to study $\{V_n\}$

much as in using martingale methods to deduce Wald's identity, and thereby deduced the inequalities noted previously at (6.2). Tan (1979) refined Kingman and Ross's martingale approach and showed that

$$\Pr\{W > x\} \leqslant \begin{cases} \exp(\kappa(\bar\theta) - \bar\theta\kappa'(\bar\theta)) & (0 \leqslant x \leqslant \kappa'(\tilde\theta)), \\ \exp(\kappa(\tilde\theta) - \tilde\theta x) & (x > \kappa'(\tilde\theta)), \end{cases}$$

where $\kappa(\theta) = \log(E\,e^{\theta U}) = \log\phi(\theta)$, $\tilde\theta$ is as at (7.1), and $\bar\theta \equiv \bar\theta(x)$ satisfies

$$\kappa'(\bar\theta) = x \qquad \text{for} \quad 0 \leqslant x \leqslant \kappa'(\tilde\theta).$$

Neither Tan's bound nor the bound at (6.1) is uniformly better than the other.

10.8 MOMENT GENERATING FUNCTION METHODS

Kingman (1962a) used the easily proved inequality

$$\text{for } \theta > 0, \qquad (\theta x)_+ \leqslant e^{\theta x - 1} \quad (\text{all real } x), \tag{8.1}$$

to show that if θ is such that $\phi(\theta) < 1$ (see the beginning of Section 10.7), then

$$EW = \sum_{n=1}^{\infty} \frac{(U_1 + \cdots + U_n)_+}{n} \leqslant \frac{1}{e\theta} \sum_{n=1}^{\infty} \frac{(\phi(\theta))^n}{n} = \frac{-\log(1 - \phi(\theta))}{e\theta}. \tag{8.2}$$

Consequently, when $\tilde\theta > 0$ with $\tilde\theta$ as at (7.1),

INEQUALITY XIX. $$EW \leqslant \inf_{0 < \theta < \tilde\theta} \left(-\frac{\log(1 - \phi(\theta))}{e\theta} \right), \tag{8.3}$$

and the infimum on the right-hand side is attained at some point in $0 < \theta \leqslant \tilde\theta$. Kingman noted that a weaker bound follows from using the point

$$\theta_0 = \sup\{\theta' \colon \phi(\theta) \geqslant \phi(\theta')\} \qquad (0 < \theta' \leqslant \tilde\theta),$$

and he also used (2.7b) and (2.7c) and the logarithmic series to bound var W and the rate of convergence to EW of

$$EW_n \equiv E\left(\sup_{0 < k < n} \{U_1 + \cdots + U_k\} \right).$$

10.9 INEQUALITIES INVOLVING CLASSES OF DISTRIBUTIONS FOR T

In much of the discussion above, especially in Section 10.3, attention has inevitably focused on the last term in (2.2a), $E(T - S - W)_+^2$. In considering the separate contributions of S and T to the properties of W there is more to be gained from restricting the class of distributions for T. Since

$$E(T-x)_+^2 = 2\int_x^\infty E(T-y)_+ \,\mathrm{d}y = 2\int_x^\infty \mathrm{d}y \int_y^\infty \Pr\{T>z\}\,\mathrm{d}z, \quad (9.1)$$

an inequality on the behaviour of T such as $E(T-x)_+ \leqslant \mu \Pr\{T>x\}$ (all $x \geqslant 0$) leads by integration to $E(T-x)_+^2 \leqslant 2\mu E(T-x)_+$ and thence by a conditioning argument to

$$E(T-S-W)_+^2 \leqslant 2\mu E(T-S-W)_+ = 2\mu E(T-S).$$

Marshall (1968) used such arguments in conjunction with the following pairs of concepts which are more familiar in the reliability theory literature (see e.g. Barlow and Proschan, 1975).

1. The r.v. X has *increasing (resp., decreasing) failure rate*, abbreviated to IFR (DFR), when for every fixed $y \geqslant 0$, $\Pr\{X > x + y \mid X > x\}$ is non-increasing (non-decreasing) in $x > 0$.
2. The r.v. X has a *new better (worse) than used distribution*, abbreviated to NBU (NWU), when for all $x, y > 0$,
$$\Pr\{X > x + y \mid X > x\} \leqslant (\geqslant) \Pr\{X > y\}.$$
3. The r.v. X has *decreasing (increasing) mean residual life*, abbreviated to DMRL (IMRL), when $E(X - x \mid X > x)$ decreases (increases) with increasing $x > 0$.
4. The r.v. X is *new better (worse) than used in expectation*, abbreviated to NBUE (NWUE), when $E(X - x \mid X > x) \leqslant (\geqslant) EX$ (all $x > 0$).

When an r.v. has IFR, it necessarily has DMRL and is NBU, and either of these latter properties implies the NBUE property. Some of the results in this section were originally established with T satisfying more stringent conditions than are indicated in the statements below. Also, reversing the inequality in the assumed property reverses the inequality in the conclusion. Finally, when T has the exponential distribution, so that the system is $M/G/1$, equality holds throughout the defining constraints.

When T is NBUE, $E(X-x)_+ \leqslant EX \Pr\{X > x\}$, and the argument below (9.1) leads to the first part of (9.2).

INEQUALITY XX. When T is NBUE,

$$ES^2 - [(ET)^2 - \operatorname{var} T] \leqslant 2(1-\rho)ETEW \leqslant ES^2, \quad (9.2)$$

$$\Pr\{W > 0\} \leqslant \rho. \quad (9.3)$$

The right-hand side of (9.2) comes from integrating the defining NBUE inequality on $(0, x)$ instead of (x, ∞), leading first to

$$ET^2 - E(T-x)_+^2 \leqslant 2ET[ET - E(T-x)_+],$$

and then by the conditioning argument

$$E(T-S-W)_+^2 = E(E[(T-S-W)_+^2 \mid S, W])$$

and (2.2) to the inequality. Inequality (9.3) comes from the conditioning argument and the defining relation:

$$\begin{aligned} ET - ES = E(T-S-W)_+ &= E(E[(T-S-W)_+ \mid S, W]) \\ &\leqslant E(ET \Pr\{T > S + W \mid S, W\}) \\ &= ET \Pr\{T > S + W\} = ET[\Pr\{W = 0\} - \Pr\{T = S - W\}] \\ &\leqslant ET \Pr\{W = 0\}. \end{aligned}$$

The DMRL property is slightly stronger than NBUE and can be expressed in the form

$$E(T-x-y)_+ \leqslant E(T-x \mid T > x) \Pr\{T > x + y\}.$$

Integrating on $0 \leqslant y < \infty$ leads to

$$E(T-x)_+^2 \leqslant 2E(T-x \mid T > x)E(T-x)_+.$$

Differentiation gives

$$\frac{d}{dx}\left(\frac{E(T-x)_+^2}{E(T-x)_+}\right) = -2 + \frac{E(T-x)_+^2 \Pr\{T > x\}}{(E(T-x)_+)^2},$$

so T being DMRL implies that $E(T-x)_+^2/E(T-x)_+$ is monotonic also, and

$$E(T-x)_+^2 \leqslant \frac{ET^2}{ET} E(T-x)_+ \qquad (\text{all } x > 0). \tag{9.4}$$

Using the same sort of conditioning argument as before yields

INEQUALITY XXI. When T is DMRL,

$$ES^2 - \rho[2(ET)^2 - ET^2]_+ \leqslant 2(1-\rho)ETEW \leqslant ES^2. \tag{9.5}$$

Mori (1975) and Fainberg (1979) give bounds on var W (cf. (2.3), (3.21) and (3.22) above) when T is NBUE, and also when T is IFR (DFR) in which case it is enough to assume that T has DMRL (IMRL). For example, when T is DMRL (IMRL),

$$\operatorname{var} W - \left(\frac{EU^2}{2EU}\right)^2 + \frac{E(-U)^3}{3E(-U)} \geqslant (\leqslant) \left(\frac{ET^2}{2ET}\right)^2.$$

Karpelevitch and Kreinin (1976) initiated another approach which is particularly useful when T is DMRL or IMRL. Consider the parametrically specific curve

$$\Gamma \equiv \{(E(T-x)_+, E(T-x)_+^2): 0 \leqslant x < \infty\}$$

which traces a downward path from (ET, ET^2) to the origin (see Fig. 10.3). The slope of this curve at the parametric point x equals

$$2E(T-x)_+/\Pr\{T > x\} = 2E(T-x \mid T > x),$$

which is monotone when T is DMRL or IMRL. Consequently, Γ is convex (concave) when T is DMRL (IMRL), so when ξ is chosen so that

$$E(T-S-W)_+ = ET - ES = E(T-\xi)_+,$$

we have $E(T-S-W)_+^2 \geqslant (\leqslant) E(T-\xi)_+^2$, i.e.

INEQUALITY XXII. When T is DMRL and ξ satisfies $E(T-\xi)_+ = (1-\rho)ET$,

$$2(1-\rho)ETEW \leqslant EU^2 - E(T-\xi)_+^2 = \operatorname{var} U - \operatorname{var}(T-\xi)_+. \quad (9.6)$$

Another aspect of considering the curve Γ is that it demonstrates immediately why the IMRL bounds analogous to (9.5), namely,

$$\begin{aligned} ES^2 &\leqslant 2(1-\rho)ETEW \\ &\leqslant ES^2 + \rho(ET^2 - 2(ET)^2) = \operatorname{var} S + \rho \operatorname{var} T - \rho(1-\rho)(ET)^2 \quad (9.7) \end{aligned}$$

for IMRL T are tight (Whitt, 1982). This is because for IMRL T, Γ is concave, and is bounded above and below by the straight lines joining (ET, ET^2) back to $(0, ET^2 - 2(ET)^2)$ and the origin respectively. Thus, on these bounding lines, we have IMRL r.v.s with $E(T-x \mid T > x)$ constant in $x > 0$.

The corresponding bounding curves for convex Γ, equivalently, for DMRL T, are not linear, so the lower bound at (9.5) need not be tight. Indeed, for the extreme case that $T =_{\text{a.s.}} ET$, this lower bound equals $ES^2 - 2ESET$ which is smaller by at least $ESET$ than the tight lower bound $(ES^2 - ESET)_+$.

The domain of definition of the curves Γ is bounded by parabolas as in

$$\frac{ET^2}{(ET)^2}[E(T-x)_+]^2 \leqslant E(T-x)_+^2 \leqslant \operatorname{var} T + [E(T-x)_+]^2. \quad (9.8)$$

The left-hand side comes from (2.13). The right-hand side comes from using the Cauchy–Schwarz inequality on the last term in

$$\begin{aligned} E(T-x)_+^2 &= E(T-x)^2 - E(x-T)_+^2 \\ &\leqslant \operatorname{var}(T-x) + [E(T-x)_+ - E(x-T)_+]^2 - [E(x-T)_+]^2 \\ &= \operatorname{var} T + [E(T-x)_+]^2 - 2E(T-x)_+E(x-T)_+; \end{aligned}$$

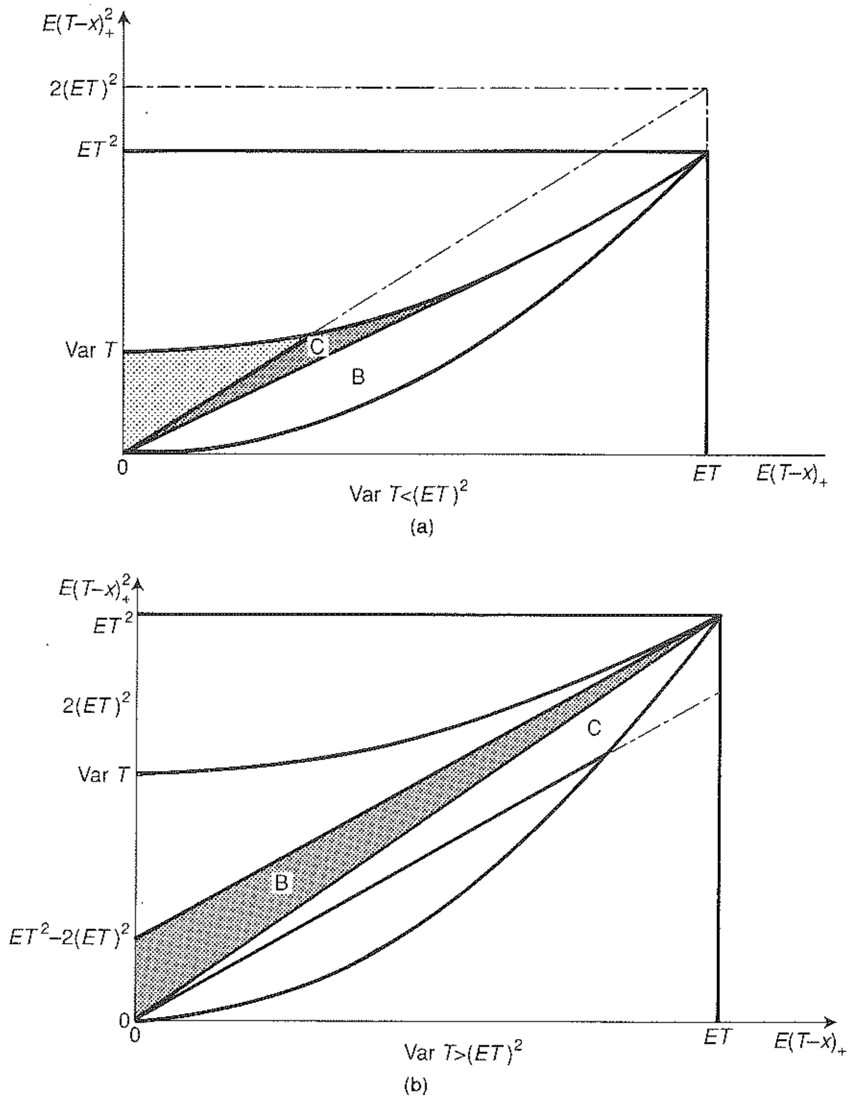

Fig. 10.3 Domain for Karpelevitch and Kreinin curves.

$$\Gamma = \{(E(T-x)_+, E(T-x)_+^2) \colon \infty > x \geqslant 0\}$$

is a continuous increasing curve from $(0, 0)$ to (ET, ET^2) confined to the region bounded by parabolas as at (9.8). When T is NBUE (NWUE), Figure 10.3(a) (Figure 10.3(b)) applies. When T has DMRL (IMRL), Γ is convex (concave) and lies in region B. When T is NBUE (NWUE), Γ has slope $\leqslant$ $(\geqslant)2ET$ and lies in $B \cup C$.

the inequality is tight because it is the supremum of a sequence of r.v.s occurring in the limit $D_{\lim}$ (see Section 10.2 above).

When T is DMRL (IMRL), T is convex (concave) and restricted to the region B in Figures 3(a) and 3(b). When T is NBUE (NWUE), Γ has its slope bounded by $2ET$ and is confined to the union of the regions B and C.

A common feature of these relationships is the use of the functions

$$A_i(x) \equiv E(T-x)_+^i \qquad (i = 0, 1, 2), \tag{9.9}$$

interpreting $E(T-x)_+^0 = EI_{\{T>x\}} = \Pr\{T > x\}$. Suppose (Kreinin, 1981b) that for all $x \geqslant 0$,

$$(ET)^2 A_2(x) \geqslant [(ET)^2 - \text{var } T][A_1(x)]^2 + 2(\text{var } T)ETA_1(x). \tag{9.10}$$

Then when $(ET)^2 > \text{var } T$,

$$(ET)^2 EA_2(S+W) \geqslant [(ET)^2 - \text{var } T][EA_1(S+W)]^2 + 2(\text{var } T)ETEA_1(S+W)$$

and from (2.2), writing $c_T^2 = (\text{var } T)/(ET)^2$, we arrive after some algebra at

INEQUALITY XXIIIa. When T satisfies (9.10) and $c_T^2 \leqslant 1$,

$$2(1-\rho)ETEW \leqslant \text{var } S + \rho^2 \text{ var } T = ES^2 - (1 - c_T^2)(ES)^2. \tag{9.11a}$$

Reversing the inequality at (9.10) leads to

INEQUALITY XXIIIb. When T satisfies the reverse inequality to (9.10) and $c_T^2 \geqslant 1$,

$$2(1-\rho)ETEW \geqslant \text{var } S + \rho^2 \text{ var } T = ES^2 + (c_T^2 - 1)(ES)^2. \tag{9.11b}$$

The exponential distribution satisfies (9.10) with equality for all $x \geqslant 0$. Indeed, any gamma r.v. satisfies either (9.10) or the reverse inequality, and thus (9.11a) or (9.11b) holds for $E_\gamma/G/1$ according as $\gamma \geqslant$ or $\leqslant 1$. To see this for $\gamma \geqslant 1$, let T_γ denote a r.v. with the distribution

$$\Pr\{T_\gamma > x\} = \int_x^\infty e^{-u} u^{\gamma-1} [\Gamma(\gamma)]^{-1} \, du \qquad (\gamma > 0, x \geqslant 0). \tag{9.12}$$

Then on integrating by parts,

$$E(T_\gamma - x)_+^2 = E(T_{\gamma+1} - x)_+^2 - 2E(T_{\gamma+1} - x)_+, \tag{9.13}$$

and also

$$
\begin{aligned}
E(T_\gamma - x)_+^2 &= \int_x^\infty (u-x)^2 \, e^{-u} u^{\gamma-1} [\Gamma(\gamma)]^{-1} \, du \\
&= \gamma(\gamma+1) E(1 - x/T_{\gamma+2})_+^2 \\
&\geqslant \gamma(\gamma+1)[E(1 - x/T_{\gamma+2})_+]^2 \\
&= \gamma(\gamma+1)^{-1}[E(T_{\gamma+1} - x)_+]^2. \qquad (9.14)
\end{aligned}
$$

Since $(\text{var } T_{\gamma+1})/(ET_{\gamma+1})^2 = 1/(\gamma+1) = 1/ET_{\gamma+1}$, (9.10) can now be verified. See Berenshtein *et al.* (1986) for another approach via the queue length.

It may be asked whether the class of r.v.s satisfying (9.10) is related to (e.g.) the IFR or NBU class. Vainshtein and Kreinin (1981) showed by various examples that the class of r.v.s satisfying (9.10) includes some IFR r.v.s, excludes others, and excludes some NBUE r.v.s.

The explicit solution of (1.2) for $E_k/G/1$ systems with $k \geqslant 2$ leads to an equation with $k-1$ roots, and (for example) EW is then described via these roots (see e.g. Cohen, 1969), namely

$$EW = \frac{ES^2 - (1-k^{-1})(ET)^2}{2ET(1-\rho)} + \sum_{j=1}^{k-1} \frac{1}{\theta_j}, \qquad (9.15)$$

where $\theta_1, \ldots, \theta_{k-1}$ are the $k-1$ roots in $\text{Re}(\theta) > 0$ of

$$(1 - \theta ET/k)^k = E(e^{-\theta S}) \equiv \beta(\theta).$$

In various papers Kreinin (see e.g. 1979a, 1980, 1981b) considered the question of bounding certain symmetric functions of these roots (see also Vainshtein, 1983; Fainberg, 1976; Vainshtein and Kreinin, 1981). Daley (1986) showed that the roots in (9.15) lie in a circle $|\theta ET/k - \frac{1}{2}r_k| \leqslant \frac{1}{2}r_k$ where r_k is the root in $1 < r_k < 2$ of $r_k - 1 = [\beta(r_k/(ET/k))]^{1/k}$, and hence that

$$EW \geqslant \frac{ES^2 - (1-k^{-1})(ET)^2}{2ET(1-\rho)} + \frac{ET(1-k^{-1})}{r_k}. \qquad (9.16)$$

This bound is exact for $k = 2$ and improves (9.11b) for $k \geqslant 2$. For most k and ρ, the bound (9.16) is also better than (3.14).

10.10 TIGHT BOUNDS AND EXTREMAL SYSTEMS

All the bounds given in this paper may be viewed as attempting to delimit characteristics of a queueing system in terms of one or two parameters or other constraints on S or T or both. Ideally, we should aim at delimiting precisely the behaviour of (for example) EW in terms of, say, the first two moments of S and T; either additionally or as an alternative, the distributions of S and T yielding these tight bounds should be identified (cf. also Whitt, 1984c).

This last step may be regarded as being of at least the same importance as the bound to which it relates, because it indicates the circumstances under which the bound may be useful as an approximation, and it may also provide a route to the calculation of the bound. We shall call such a queueing system *extremal* (with respect to a given set of constraints). For example, Theorem 3.1 describes extremal systems for $\inf_{\mathcal{Q}_2} EW$.

Recent work has emphasized the nature of factors affecting delays in a queueing system under conditions of light traffic ($\rho \approx 0$) (Asmussen, 1992; Daley and Rolski, 1991 and 1992b) and with a general ergodic arrival process (i.e. not necessarily a renewal process) (Wolff, 1991; Daley and Rolski, 1992a,b). In both these contents, delays occur as a result of any tendency of customers to arrive in clusters. In particular, for $\rho \approx 0$,

$$\Pr\{\text{delay occurs}\} \approx \Pr\{\text{arrivals occur in a cluster}\},$$

$$EW = E(\text{delay}) \approx E[(\text{no. arrivals in cluster} - 1)_+ | \text{cluster}] \times \Pr\{\text{delay occurs}\} ES.$$

Also, it is noted in Daley and Rolski (1992a) (see also (10.7) and (10.11) below) that the term involving $\rho \operatorname{var} T$ in the upper bound on EW at (3.4) can be interpreted as coming from a uniform bound on the extent of clustering. This represents the influence on W of the negative tail of the increments $U = S - T$: because occasional large values of T result in the dissipation of any customer backlog, large values of $\operatorname{var} T$ can increase W only to the extent that it permits smaller non-negative values of T. On the other hand, as emerges in the argument leading to (10.2) below, the positive tail, essentially S, and in particular, occasional large values of S, result in a backlog which can only be dissipated by steady attrition. Consequently, EW is affected directly by the extent of any inherent fluctuations in S, irrespective of the size of ρ.

What heavy traffic results show is that central limit properties of large numbers of the random walk increments dominate W for $\rho \approx 1$. Under these conditions, fluctuations as measured by $\operatorname{var} U$ are appropriate, irrespective of their arising from the positive or negative tail.

These contrasting influences of clustering and fluctuations on delays indicate that identification of extremal systems can be expected to be difficult. Here we mainly concentrate on the class $\mathcal{Q}_2$ or subsets thereof. We attempt to identify within this class the functions $a(\rho)$ and $b(\rho)$ such that

$$2(1-\rho)ETEW \leqslant a(\rho) \operatorname{var} T + b(\rho) \operatorname{var} S, \tag{10.1}$$

noting that we already know from (3.4) that in $\mathscr{Q}_2$, $a(\rho) \leqslant \rho(2-\rho)$ and $b(\rho) \leqslant 1$.

The nature of $b(\rho)$ is easily established. Weak convergence methods as in Daley (1984a) show that

$$\begin{aligned} EW(GI/D_{\lim}/1) &= EW(GI/D/1) + EW(D/D_{\lim}/1) \\ &= EW(GI/D/1) + \frac{\operatorname{var} S}{2(1-\rho)ET}. \end{aligned} \tag{10.2}$$

Consequently,

$$\sup_{\mathscr{Q}_2} b(\rho) = 1. \tag{10.3}$$

Further, by writing the right-hand side of the lower bound at (3.14) as $(\operatorname{var} S - \rho(1-\rho)(ET)^2)_+$, it follows that (10.3) holds in any subset of $\mathscr{Q}_2$.

The nature of $a(\rho)$ is harder to determine. From (10.2), if we can find the function $a_D(\rho)$ and the distribution (or, sequence of distributions) yielding

$$a_D(\rho) \equiv \sup_{0<\operatorname{var} T<\infty} \left\{ \frac{2(1-\rho)ET}{\operatorname{var} T} \sup_{\mathscr{Q}_2} EW(GI/D/1) \right\}, \tag{10.4}$$

then certainly we should have

$$a(\rho) \geqslant a_D(\rho). \tag{10.5}$$

Observe that use of the inequality (2.12) in the last term of (2.2) yields

$$\begin{aligned} 2(1-\rho)ETEW &= EU^2 - E(E((T-S-W)_+^2 \mid S, W)) \\ &\leqslant EU^2 - E(ET^2/ET - S - W)_+^2 (ET)^2/ET^2, \end{aligned} \tag{10.6}$$

and equality holds if $\Pr\{T = 0 \text{ or } ET^2/ET\} = 1$, i.e. in a $B_\alpha/G/1$ system with $\alpha = ET^2/(ET)^2$. Now in such a $B_\alpha/G/1$ system,

$$ET - ES = E(T-S-W)_+ = \alpha^{-1}E(\alpha ET - S - W)_+$$

and

$$\begin{aligned} E(T-S-W)_+^2 &= \alpha^{-1}E(\alpha ET - S - W)_+^2 \geqslant \alpha^{-1}[E(\alpha ET - S - W)_+]^2 \\ &= \alpha(ET-ES)^2, \end{aligned}$$

so

$$\begin{aligned} 2(1-\rho)ETEW &= EU^2 - \alpha^{-1}E(\alpha ET - S - W)_+^2 \\ &\leqslant EU^2 - \alpha(ET-ES)^2 = \operatorname{var} S + \rho(2-\rho)\operatorname{var} T, \end{aligned}$$

which is just (3.4).

Still in such a $B_\alpha/G/1$ system, we have from (4.4) and (4.5) that $EW \geqslant \xi$, where

$$ET - ES = E(T - S - \xi)_+ = \alpha^{-1}E(\alpha ET - S - \xi)_+$$
$$\geqslant \alpha^{-1}(\alpha ET - ES - \xi)_+ \geqslant ET - \alpha^{-1}(ES - \xi),$$

so

$$EW \geqslant \xi \geqslant (\alpha - 1)ES = \frac{\rho \operatorname{var} T}{ET}.$$

Combining this result with (3.4), it follows that

$$\rho(2 - 2\rho) \leqslant a_D(\rho) \leqslant a(\rho) \leqslant \rho(2 - \rho). \tag{10.7}$$

A slightly more delicate argument in Daley and Rolski (1992a), involving instead of B_α the two-point distributions with $ET - x$ for some $0 < x < ES$ as the lower point of support, shows that

$$\tfrac{2}{3}\rho(2 - \rho) \leqslant a_D(\rho) \leqslant a(\rho) \leqslant \rho(2 - \rho). \tag{10.7a}$$

From the identity (2.7c) for var W, and an application of (2.12) again to each T_i in each term, it follows as in Bergmann *et al.* (1979) that for given S,

$$\sup_{\mathscr{D}_2} \operatorname{var} W \qquad \text{occurs in} \qquad B_\alpha/G/1. \tag{10.8}$$

Now var $W \geqslant (EW)^2$, so it is plausible that in $\mathscr{D}_2$ the same extremal system, namely $B_\alpha/G/1$, may yield both $\sup_{\mathscr{D}_2} EW$ and $\sup_{\mathscr{D}_2}$ var W.

The fact that equality holds at (10.6) in $B_\alpha/G/1$ whereas inequality holds for any other distribution for T, also suggests that the largest mean EW over $\mathscr{D}_2$ may occur in $B_\alpha/G/1$. Certainly, such a statement is true in $GI/M/1$ systems and in their discrete analogues $GI/D/1$ (for these analogues T is restricted to distributions for which $\Pr\{T = jES$ for some $j = 0, 1, \ldots\} = 1$), as may be verified from the exponential or geometric form of the distributions of W known for these cases (see Whitt, 1984a, b for a review and discussion of $GI/M/1$).

It is also known, by direct computation and consideration of the roots of equations involving Laplace–Stieltjes transforms, that when S is a mixture of exponential r.v.s (and hence, $E\,e^{\theta S}$ is of the form $\sum_{i=1}^{\infty} p_i\mu_i/(\mu_i + \theta)$ for constants $\mu_i > 0$, $0 \leqslant p_i < 1$, $\sum p_i = 1$, and $ES = \sum p_i/\mu_i$), $\sup_{\mathscr{D}_2} EW$ occurs for $B_\alpha/G/1$. Such a distribution may be called a hyperexponential distribution, so using the symbol H_m in Kendall's notation, the conclusion can be restated as

$$\sup_{\mathscr{D}_2} EW(GI/H_m/1) = EW(B_\alpha/H_m/1).$$

Within the IMRL class for T, we referred at (9.7) above to Whitt's (1982) observation that the upper bound of Marshall (1968) there is tight. The linear relation between $E(T-x)_+^2$ and $E(T-x)_+$ required for this bound implies that the extremal IMRL system for $\sup_{\mathcal{D}_2 \cap \{T \text{ has IMRL}\}} EW$ has the distribution

$$T =_d \begin{cases} 0 & \text{with probability } (\operatorname{var} T - (ET)^2)/ET^2, \\ \operatorname{Exp}(ET^2/2ET) & \text{with probability } 2(ET)^2/ET^2. \end{cases}$$

All this leads to

CONJECTURE I.

$$\sup_{\mathcal{D}_2} EW(GI/G/1) = \sup_{\mathcal{D}_2} EW(GI/D/1) + \frac{\operatorname{var} S}{2(1-\rho)ET}; \tag{10.9a}$$

CONJECTURE II.

$$\sup_{\mathcal{D}_2} EW(GI/D/1) = EW(B_\alpha/D/1). \tag{10.9b}$$

It would follow from (10.9) that we should study W in $B_\alpha/G/1$ and, in particular, $B_\alpha/D/1$ systems. A numerical study via discrete queues is given in Daley (1990), and is consistent with (10.9b).

Another way of doing so is to consider modified systems $D^*/G^*/1$ in which the generic r.v.s T^* and S^* are related to those of the $B_\alpha/G/1$ system by

$$T^* =_{\text{a.s.}} ET^2/ET = \alpha ET \qquad \text{and} \qquad S^* = S_1 + \cdots + S_J,$$

where the r.v. J is independent of the i.i.d. sequence $\{S_n\}$ and $\Pr\{J=j\} = \alpha^{-1}(1-\alpha^{-1})^{j-1}$ $(j = 1, 2, \ldots)$. W is related to the stationary waiting time r.v. W^* of the $D^*/G^*/1$ system by

$$W =_d W^* + S_1 + \cdots + S_{J-1} \tag{10.10}$$

where on the right-hand side all of the r.v.s $\{S_n\}$ and J are mutually independent. It follows from (10.10) that

$$EW = EW^* + E(J-1)ES = EW^* + (\alpha - 1)ES = EW^* + \frac{\rho \operatorname{var} T}{ET}. \tag{10.11}$$

Since $EW^* \geqslant 0$, (10.11) provides an alternative route to establishing the left-hand side of (10.7).

Suppose given a $B_\alpha/G/1$ system for which $S \leqslant_c \operatorname{Exp}(ES)$ where $\operatorname{Exp}(ES) \equiv S'$ denotes an exponentially distributed r.v. with mean ES. Then (cf. e.g. Stoyan, 1983, Chapter 1) since for i.i.d. sequences $\{S_n\}$ and $\{S'_n\}$ with $S_n =_d S$ and $S'_n =_d S'$ we have

$$S_1 + \cdots + S_j \leqslant_c S'_1 + \cdots + S'_j,$$

we also have

$$S_1 + \cdots + S_J \leqslant_c S_1' + \cdots + S_J' =_d \operatorname{Exp}(EJES),$$

because J is geometrically distributed on $\{1, 2, \ldots\}$. Since $EJ = \alpha$, rescaling yields

$$(S_1 + \cdots + S_J)/\alpha \leqslant_c \operatorname{Exp}(ES). \tag{10.12}$$

This then implies (cf. §6 above; Stoyan, 1983, Chapter 5) that

$$\begin{aligned} \alpha^{-1} W^* &=_d \alpha^{-1} W(\alpha ET, S_1 + \cdots + S_J) \\ &=_d W(ET, (S_1 + \cdots + S_J)/\alpha) \\ &\leqslant_c W(ET, \operatorname{Exp}(ES)) =_d W(D/M/1). \end{aligned} \tag{10.13}$$

Therefore,

$$\alpha^{-1} EW^* \leqslant EW(D/M/1) = \frac{\delta ES}{1-\delta} \tag{10.14}$$

where δ is the root in $0 < \delta < 1$ of

$$\delta = E(e^{-(1-\delta)ET/ES}) = e^{-(1-\delta)/\rho}. \tag{10.15}$$

Referring to (10.11), we conclude that when $S \leqslant_c \operatorname{Exp}(ES)$ in $B_\alpha/G/1$, and this covers the case that $S =_{\text{a.s.}} ES$, (for any r.v. X, $EX \leqslant_c X$),

$$\begin{aligned} \frac{2(1-\rho)ETEW}{\operatorname{var} T} &\leqslant \rho(2-2\rho) + \frac{2(1-\rho)ET\alpha\delta ES/(1-\delta)}{\operatorname{var} T} \\ &= \rho(2-2\rho) + \frac{\rho(2-2\rho)\delta/(1-\delta)}{1-\alpha^{-1}} \\ &\to \frac{\rho(2-2\rho)}{1-\delta} \qquad (\alpha \to \infty, \text{ i.e. } \operatorname{var} T \to \infty). \end{aligned} \tag{10.16}$$

This limit $< \rho(2-\rho)$ if and only if $\rho > 2\delta/(1+\delta)$, a condition which is satisfied because, from (10.15),

$$\begin{aligned} \rho = \frac{1-\delta}{-\log \delta} &= \frac{1-\delta}{1-\delta+(1-\delta)^2/2+(1-\delta)^3/3+\cdots} \\ &> \frac{1}{1+\frac{1}{2}(1-\delta)[1+(1-\delta)+(1-\delta)^2+\cdots]} \\ &= \frac{2\delta}{1+\delta}. \end{aligned}$$

Conjecture II can now be combined with (10.5) to give

CONJECTURE III. $\sup_{\mathscr{D}_2} EW(GI/D/1)/\mathrm{var}\, T$ occurs in the limit var $T \to \infty$ of a sequence of $B_\alpha/D/1$ systems, for which

$$a(\rho) = a_D(\rho) = \frac{2(1-\rho)\rho}{1-\delta} \qquad \text{where} \qquad 0 < \delta < 1 \quad \text{and} \quad \delta = e^{-(1-\delta)/\rho}. \tag{10.17}$$

Further support for Conjecture II comes from studying light traffic approximations. Daley and Rolski (1984a) used the Spitzer identity relation (2.7b) to observe that

$$\lim_{\gamma\to\infty} EW(\gamma T, S) = \frac{ES \Pr\{T=0\}}{\Pr\{T>0\}}, \tag{10.18}$$

and this is largest when $\Pr\{T=0\}$ is largest (see e.g. Feller, 1966, §V (7.5)), i.e. when the system is $B_\alpha/G/1$ (see also Daley and Rolski, 1991).

10.11 NUMERICAL COMPARISONS OF SOME BOUNDS

As has been mentioned, part of the motivation for studying bounds is to find an alternative to the classical approach to determining exact solutions for queueing problems. From this point of view, we should naturally be interested in how well these bounds perform as approximations to the actual value of, say, EW. Even if their overall performance is poor, it is still of interest to compare the sharpness and simplicity of the various bounds against each other.

These attributes of sharpness and simplicity are clearly of prime importance in examining the performance of any approximation, though they will tend to be incompatible. We hope to strike a balance between a sharp inequality which remains as difficult to evaluate as EW, and one which, while being simple, tells us nothing about EW because of its inaccuracy.

For the remainder of this section the dimensionless quantity EW/ES will be considered in place of EW so that the mean stationary waiting time is measured in terms of mean service times. In this context it is also worthwhile defining the random variables S_* and T_* by

$$S_* = S/ES, \qquad T_* = T/ET. \tag{11.1}$$

EW/ES and bounds for this quantity are therefore expressible in terms of S_*, T_* and ρ.

It is natural to regard S_* and T_* as specifying a family of queues, with each family parameterized by the traffic intensity, ρ. The families chosen for examination in this section are the $M/M/1$, $M/D/1$ and $D/M/1$ systems. Apart from being relatively easy to handle algebraically, they afford some idea of the range of behaviour to be expected. A distribution of the DFR type, such as hyperexponential distribution, might also be worth considering as a substitute for M or D (cf. Miyazawa, 1976; Whitt, 1984a).

Some values of the more interesting inequalities are given in Tables 10.1 and 10.2 and are displayed graphically in Figs. 10.4(a)–(c). The upper bounds considered in some detail are found at (3.1), (3.4), (3.6), (4.9), (6.3) and (8.3), while the lower bounds are found at (4.4), (6.3)–(6.5) and (6.7).

Table 10.1 Some upper bounds on mean waiting time

				(a) $M/M/1$				
ρ	0.00	0.20	0.40	0.60	0.80	0.90	0.95	$^*(1-\rho)w(\rho)$
EW/ES	0.00	0.25	0.67	1.50	4.00	9.00	19.00	1.0
(3.1)	∞	3.25	2.42	2.83	5.13	10.06	20.03	1.0
(3.4)	1.00	1.25	1.67	2.50	5.00	10.00	20.00	1.0
(3.5)	0.00	0.45	1.07	2.10	4.80	9.90	19.95	1.0
(4.9)	0.00	0.90	1.59	2.63	5.20	10.11	20.05	1.0
(6.3)			[Exact: exponential service time]					
(8.3)	0.00	0.60	1.58	3.67	10.95	28.21	68.24	∞
				(b) $M/D/1$				
ρ	0	0.20	0.40	0.60	0.80	0.90	0.95	$^*(1-\rho)w(\rho)$
EW/ES	0.00	0.13	0.33	0.75	2.00	4.50	9.50	0.5
(3.1)	∞	3.13	2.08	2.08	3.13	5.56	10.53	0.5
(3.4)	1.00	1.13	1.33	1.75	3.00	5.50	10.50	
(3.5)	0.00	0.24	0.60	1.20	2.69	5.34	10.42	0.5
(4.9)	0.00	0.42	0.83	1.47	2.99	5.58	10.55	0.5
(6.3)	0.00	0.38	0.62	1.06	2.32	4.83	9.83	0.5
(8.3)	0.00	0.18	0.54	1.39	4.56	12.40	30.93	∞
				(c) $D/M/1$				
ρ	0.00	0.20	0.40	0.60	0.80	0.90	0.95	$^*(1-\rho)w(\rho)$
EW/ES	0.00	0.01	0.12	0.48	1.69	4.18	9.17	0.5
(3.1)	0.00	0.13	0.33	0.75	2.00	4.50	9.50	0.5
(3.4)			[Same as (3.1): deterministic arrival process]					
(3.5)	0.00	0.01	0.16	0.60	1.91	4.45	9.48	0.5
(4.9)	0.00	0.12	0.53	1.03	2.13	4.56	9.53	0.5
(6.3)			[Exact: exponential service time]					
(8.3)	0.00	0.04	0.45	1.47	4.88	12.92	31.64	∞

* That is, $\lim_{\rho \uparrow 1}(1-\rho)w(\rho)$.

Table 10.2 Some lower bounds on mean waiting time

				(a) $M/M/1$				
ρ	0.00	0.20	0.40	0.60	0.80	0.90	0.95	$^*(1-\rho)w(\rho)$
EW/ES	0.00	0.25	0.67	1.50	4.00	9.00	19.00	1.00
(4.4)	0.00	0.20	0.44	0.74	1.28	1.85	2.45	0.00
(6.3)			[Exact: exponential service time]					
(6.4)	0.00	0.19	0.43	0.93	2.46	5.58	11.82	0.63
(6.5)	0.00	0.04	0.19	0.56	1.78	4.26	9.26	0.50
(6.7)	0.00	0.20	0.40	0.60	0.80	0.90	0.95	0.00
				(b) $M/D/1$				
EW/ES	0.00	0.13	0.33	0.75	2.00	4.50	9.50	0.50
(4.4)	0.00	0.12	0.28	0.53	1.01	1.56	2.15	0.00
(6.3)	0.00	0.07	0.23	0.61	1.81	4.28	9.27	0.50
(6.4)	0.00	0.10	0.21	0.41	0.98	2.12	4.41	0.23
(6.5)	0.00	0.01	0.04	0.13	0.44	1.09	2.41	0.13
(6.7)	0.00	0.20	0.40	0.60	0.80	0.90	0.95	0.00
				(c) $D/M/1$				
EW/ES	0.00	0.01	0.12	0.48	1.69	4.18	9.17	0.50
(4.4)	0.00	0.01	0.09	0.24	0.45	0.60	0.71	0.00
(6.3)			[Exact: exponential service time]					
(6.4)	0.00	0.01	0.10	0.39	1.35	3.36	7.41	0.41
(6.5)	0.00	0.002	0.05	0.28	1.15	2.96	6.63	0.37
(6.7)	0.00	0.01	0.09	0.23	0.40	0.49	0.54	0.00

* That is, $\lim_{\rho\uparrow 1}(1-\rho)w(\rho)$.

For all the bounds $w(\rho)$ say that we consider for given S_* and T_*, we also tabulate $\lim_{\rho\to 1}(1-\rho)w(\rho)$ because, as follows from Kingman's (1962b) limit result,

$$\lim_{\rho\to 1}(1-\rho)EW_*/EW_* = \tfrac{1}{2}(\operatorname{var} S_* + \operatorname{var} T_*), \tag{11.2}$$

which equals 1 for $M/M/1$ and $\frac{1}{2}$ for $M/D/1$ and $D/M/1$.

10.11.1 Notes concerning upper bounds

(1) With the exception of the moment generating function bound (8.3), all the bounds displayed in Table 10.1 are asymptotically sharp for $\rho\uparrow 1$ (i.e. they satisfy (11.2)), while as $\rho\downarrow 0$ all bounds approach 0, except (3.1) and (3.4) which do so if and only if T_* is degenerate. This

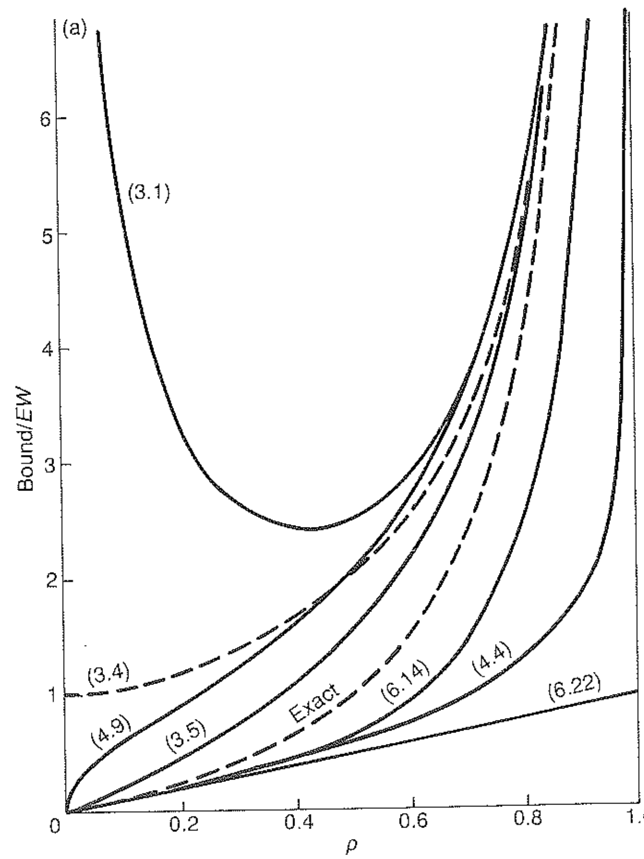

Fig. 10.4 Waiting-time bounds for (a) $M/M/1$, (b) $M/D/1$, (c) $D/M/1$. Waiting time or the bound is given with ES as the unit. Exact is shown as ----. (3.4) is Daley's refinement of Kingman's bound (3.1), and (3.5) a further refinement. (4.4) is Marshall's bound, and (4.9) an analogous upper bound from another convexity argument. See text for (6.14) and (6.22). *(continued)*

asymptotic behaviour is in fact true for all $GI/G/1$ systems with finite second moments for S_* and T_* (cf. Trengove, 1978; Daley and Rolski, 1984a).

(2) Rough comparisons can be made between the upper bounds by reference to Fig. 10.4(a)–(c). Clearly (3.4) supersedes (3.1) since they are almost equally simple to calculate. We may regard as the next simplest those with terms such as EU_+, EU_+^2 and $P\{U > 0\}$ and within this class (3.5) offers marked improvement on (3.4). Slightly more difficult to calculate is (4.9), since the bound is the root of an equation and cannot in general be given explicitly; further, the extra effort does not appear worthwhile, for (4.9) does not improve on (3.5) in the cases considered. The same remarks apply to (8.3) which is even more trouble to calculate.

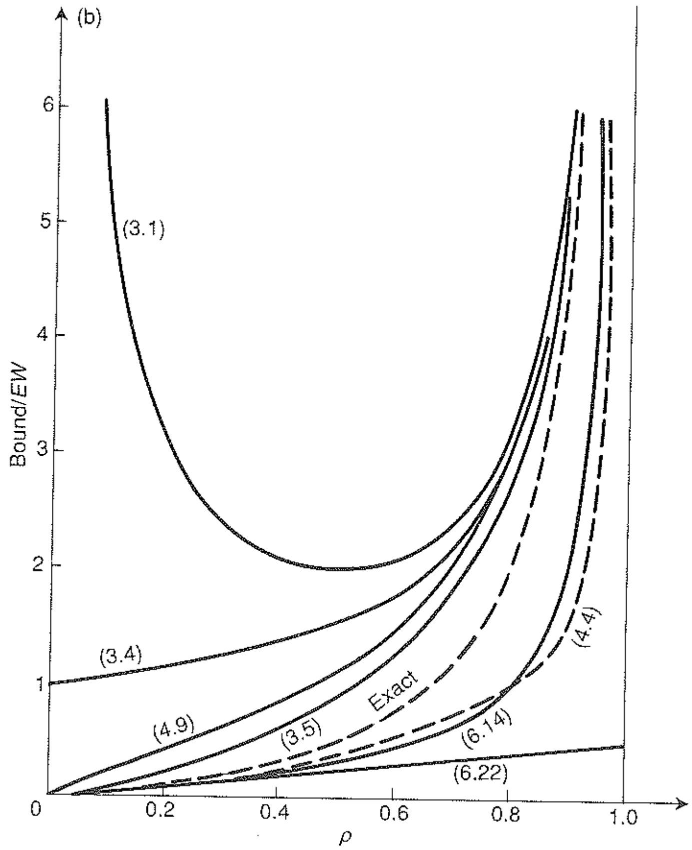

Fig. 10.4 (*continued*).

There is a case for the martingale bound (6.3): when the service time is exponential the bound is exact, and in the case $M/D/1$ it tracks the value of EW_*/ES_* fairly closely except for smaller values of ρ. However, the calculation of the bound in most cases is quite daunting.

(3) We discuss briefly the other upper bounds, using in part unpublished work from Trengove (1978). (3.9) improves on (3.1) but appears to be inferior to (3.4) while being more trouble to calculate; (3.11) is weaker but computationally simpler. (3.12a) is inferior to (3.4) for small ρ while it improves on (3.6) for the cases examined for ρ larger than about 0.8. (3.12b) is necessarily weaker than (3.12a) but still improves on (3.6) for the cases examined for ρ larger than about 0.9.

Stoyan's improvement to (3.1) has been illustrated in principle in Fig. 10.2; in Fig. 10.4(a,b) a horizontal tangent is required to (3.1). As noted in the text, (3.4) is always better.

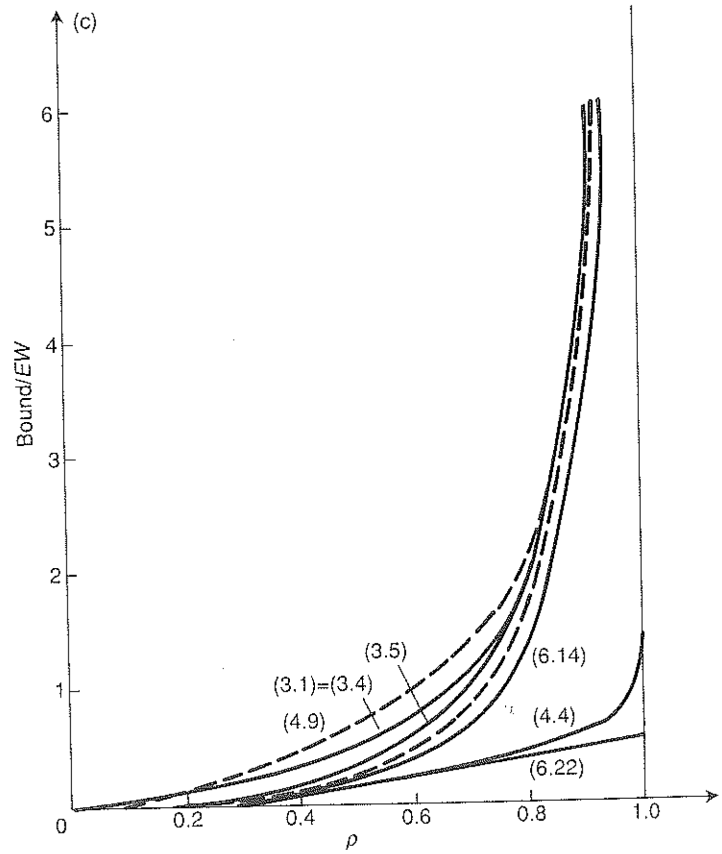

Fig. 10.4 (*continued*).

10.11.2 Notes concerning lower bounds

(1) Unless a lower bound $w(\rho)$ is trivially negative, we generally have $w(\rho) \to 0$ as $\rho \downarrow 0$ so the only limit of interest is $\lim_{\rho \uparrow 1}(1-\rho)w(\rho)$. For (4.4) and (6.7) this limit is zero, being attributable in the former case to a rate of growth slower than $(1-\rho)^{-1}$ as $\rho \uparrow 1$, and in the latter to the finiteness of the bound at $\rho = 1$. The bounds (6.4) and (6.5), while of $O((1-\rho)^{-1})$ for $\rho \uparrow 1$, are not asymptotically sharp and so their error as an approximation to EW_*/ES_* diverges. (6.3) is the only tabulated lower bound which is asymptotically sharp.

(2) Unlike the upper bounds, there are no useful bounds requiring knowledge only of the first two moments of S and T (see Note (3) in Section 10.11.2). The bound at (6.5) requires knowledge of the moments of U_+ and while (6.4) is somewhat tedious to calculate, in principle it requires only those computational techniques used for example in (3.6). The convexity bound

(4.4) is similar in complexity to (4.9) while the comments on the lower bound at (6.3) are similar to those concerning the upper bound.

With the exception of (6.3), all the lower bounds perform poorly as approximations for all but small values of ρ. Thus, it is from the upper bounds that we must in general select a bound as an approximation, unless our interest is confined to small values of ρ.

(3) We discuss briefly the remaining lower bounds. While the bound (3.14) is expressed in terms of the moments of S and T and is known to be the best possible for such, it is zero and trivial for

$$\rho < (ES)^2/ES^2 = 1/(1 + \text{var } S_*).$$

Its performance as an approximation is generally very poor. (5.3) behaves much like (6.7), and remains bounded for $\rho \uparrow 1$.

If T is NBUE, (9.2) gives a lower bound for EW/ES that is asymptotically sharp as $\rho \uparrow 1$. But, like (3.14), the bound is trivially zero for $\rho^2 < (2 - ET_*^2)/ES_*^2$. Similar remarks apply to (9.5) when T is DMRL (or IFR). In the cases we have studied, these bounds appear to improve on (6.4) only for ρ larger than about 0.9.

Since we have not included hyperexponential distributions in the present study, and these are examples of T being NWUE or IMRL or DFR when (9.2) become upper bounds, we cannot compare their performance with (3.6) or (6.3) for example. But some indication of their behaviour is available from Whitt's (1982) identification of conditions for the bounds to be tight.

10.12 CONCLUDING REMARKS

We claimed at the outset that the methods used to derive bounds indicated in this paper have found application elsewhere—in more complex queueing models and in other models of applied probability.

As examples of variants of the basic $GI/G/1$ system to which (some of) the methods above have been applied, we cite systems with autonomous service (i.e. idle time) (see Powell and Avi-Itzhak, 1967; Delbrouck, 1969; Kreinin, 1979a), and systems with warming-up time (see Kreinin, 1979b; and the survey paper of Rossberg and Siegel, 1974).

For systems with finite waiting room the relations that exist between stationary time and (stationary) customer state probabilities have been exploited by Schmidt (1978) (see also Franken *et al.*, 1981; Rolski, 1981).

One method that has not (yet) found its way into the literature on queueing inequalities is that concerned with 'robustness' of the system with respect to perturbations of the governing parameters of the system (cf. Kalashnikov and Anichkin, 1981; Stoyan, 1983, Chapter 8).

Trengove (1978) explored some possibilities using the techniques of this chapter to study bounds on EW in many-server systems of the $GI/G/k$ type. The identity (2.2) exploited so much above is of limited use because its analogue involves the work loads of all k servers at an arrival epoch. Daley (1984a) made some progress with (1.2) via analogues of both Ott's result (cf. Theorem 3.1) for Stoyan's lower bound (3.14) and the limit result at (10.2). In particular, he showed that

$$\inf_{\mathscr{Q}_2} EW(GI/G/k) = 0 \qquad \text{when} \qquad \rho < 1 - k^{-1},$$

a result with no counterpart in the single-server case.

There is a technique available for bounding $EW(GI/G/k)$ that has no counterpart in the single-server case. It involves comparing queues with different disciplines. Wolff (1977, 1987) and Gittins (1978) showed that amongst systems where the allocation rule of each customer to a server is independent of that customer's service time, the mean waiting time is least under the first-come–first-served discipline. Consequently, a $GI/G/k$ queue with the cyclic allocation discipline, which can be viewed as k interwoven single-server queues, gives an upper bound on the mean waiting time. A convenient calculus for studying allocation rules was given by Foss (1980, 1981) (see also Daley, 1987). It is a weak majorization inequality that is preserved under the first-come–first-served discipline, and because of this and the fact that a customer's waiting time is not a convex function of inter-arrival and service times to earlier arrivals, special arguments as in Wolff are needed to study the r.v. W. In particular, while some convex ($\leqslant_c$) inequalities may hold, neither distributional ($\leqslant_d$) nor almost sure ($\leqslant_{a.s.}$) inequalities can be expected to hold in any generality.

REFERENCES

Al-Khayyal, F. A. and Gross, D. (1976). On approximating and bounding $GI/M/c$ queues—a summary. *Tech. Report*, Inst. Management Sci. and Engineering, George Washington University.

Asmussen, S. (1992). Light traffic equivalence in single server queues. *Ann. Appl. Probab.*, **2** (to appear).

Barlow, R. E. and Proschan, F. (1975). *Statistical theory of reliability and life testing.* Holt, Rinehart, and Winston, New York.

Berenshtein, A., Kreinin, A. and Vainshtein, A. (1986). Convex properties of Poisson distribution and their applications. (In Russian). In *Stability Problems for Stochastic Models*, 17–22 (cf. *Math. Reviews*, **89**h: 60027).

Bergmann, R. (1974). Qualitätsunterscheide bei Wartemodellen vom Typ $G/G/1$. *Math. Operat.-forschung Statist.*, **5**, 709–24.

Bergmann, R., Daley, D. J., Rolski, T., and Stoyan, D. (1979). Bounds for cumulants of waiting times in $GI/G/1$ queues. *Math. Operat.-forschung Statist., Ser. Optim.*, **10**, 257–63.

Bergmann, R. and Stoyan, D. (1976). On exponential bounds for the waiting-time distribution function in $GI/G/1$. *J. Appl. Probab.*, **13**, 411–17.

Bloomfield, P. and Cox, D. R. (1972). A low traffic approximation for queues. *J. Appl. Probab.*, **9**, 832–40.

Borovkov, A. A. (1970). On factorization identities and the properties of the distribution of the supremum of sequential sums. *Teor. Verojatnost. i Primenen*, **14**, 377–418. (Transl. *Theor. Probab., Appl.*, **15**, 359–402).

Borovkov, A. A. (1976). *Stochastic processes in queueing theory.* Springer-Verlag, New York. (Transl. of Russian original, publ. 1972, Nauka, Moscow.)

Brumelle, S. L. (1971a). On the relation between customer and time averages in queues. *J. Appl. Probab.*, **8**, 508–20.

Brumelle, S. L. (1971b). Some inequalities for parallel-server queues. *Operat. Res.*, **19**, 402–13.

Calo, S. B. and Schwartz, S. C. (1977). An upper bound for the equilibrium mean wait in a stationary $GI/G/1$ queue. *Math. Operat. Res.*, **2**, 240–43.

Cohen, J. W. (1969). *The single-server queue.* North-Holland, Amsterdam.

Daley, D. J. (1976). Bounds on mean waiting times in $GI/G/1$ queues. *Research Report*, School of Mathematical Sciences, University of Melbourne.

Daley, D. J. (1977). Inequalities for moments of tails of random variables, with a queueing application. *Z. Wahrs.*, **41**, 139–43.

Daley, D. J. (1980). Tight bounds for the renewal function of a random walk. *Ann. Probab.*, **8**, 615–21.

Daley, D. J. (1984a). Some results for the mean waiting-time and work-load in $GI/G/k$ queues. (Manuscript).

Daley, D. J. (1984b). One- and two-sided Kolmogorov type inequalities for a random walk. (Manuscript).

Daley, D. J. (1984c). Bounds for the delay probability of a $GI/G/1$ queue. (Manuscript).

Daley, D. J. (1984d). Inequalities for geometric random sums. (Manuscript).

Daley, D. J. (1986). A lower bound for mean characteristics in $E_k/G/1$ and $GI/E_k/1$ queues. *Math. Operat.-forschung Statist., Ser. Optim.*, **17**, 117–24.

Daley, D. J. (1987). Certain optimality properties of the first-come first-served discipline for $GI/G/s$ queues. *Stoch. Proc. Appl.*, **25**, 301–8.

Daley, D. J. (1990). Discrete-time queues with explicitly computable solutions. *Techn. Report*, Statistics Research Section, Australian National University.

Daley, D. J. and Haslett, J. (1982). A thermal-energy storage process with controlled input. *Adv. Appl. Prob.*, **14**, 257–71.

Daley, D. J. and Moran, P. A. P. (1968). Two-sided inequalities for waiting time and queue size distributions in $GI/G/1$. *Teor. Verojatnost. i Primenen.* **13**, 356–9 (Transl. *Theor. Probab. Appl.*, **13**, 338–41).

Daley, D. J. and Rolski, T. (1984a). A light traffic approximation for a single-server queue. *Math. Operat. Res.*, **9**, 624–8.

Daley, D. J. and Rolski, T. (1984b). Some comparability results for waiting times in single- and many-server queues. *J. Appl. Probab.*, **21**, 887–900.

Daley, D. J. and Rolski, T. (1991). Light traffic approximations in queues. *Math. Operat. Res.*, **16**, 57–71.

Daley, D. J. and Rolski, T. (1992a). Finiteness of waiting time moments in a single-server queue with general arrival process. *Ann. Appl. Probab.*, **2**.

Daley, D. J. and Rolski, T. (1992b). Light traffic approximations in general stationary single-server queues. *Stoch. Proc. Appl.*

Daley, D. J. and Trengove, C. D. (1977). Bounds for mean waiting times in single-server queues: a survey. Unpublished preprint, Statistics Dept. (IAS), Australian National University.

Delbrouck, L. E. N. (1969). A multiserver queue with enforced idle time. *Operat. Res.*, **17**, 506–19.

Delbrouck, L. E. N. (1975). Convexity properties, moment inequalities and asymptotic exponential approximations for delay distributions in $GI/G/1$ systems. *Stoch. Proc., Appl.*, **3**, 193–207.

Fainberg, M. A. (1976). Estimates of the parameters of a queue in multiline queueing systems of the type $E_k/G/s$. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1976** (5), 71–8. (Transl. *Engng. Cybernetics*, **14**(5), 66–73.)

Fainberg, M. A. (1979). Estimates of the waiting time in single-channel queueing systems. *Izv. Akad. Nauk SSSR Techn. Kibernet.*, **1979**(2), 217–19. (Transl. *Engng. Cybernetics*, **17**(2), 144–6.)

Feller, W. (1966). *An introduction to probability theory and its applications*, Vol. II. Wiley, New York.

Foss, S. G. (1980). Approximation of multichannel queueing systems (in Russian). *Sibirski Mat. Zh.*, **21**(6), 132–40. (Transl. *Sib. Math. J.*, **21**, 851–7.)

Foss, S. G. (1981). Comparison of servicing strategies in multichannel queueing systems (in Russian). *Sibirski Mat. Zh.*, **22**(1), 190–7. (Transl. *Sib. Math J.*, **22**, 142–7.)

Franken, P. and Stoyan, D. (1975). Stability, monotonic, and extremal solutions of problems in queueing and reliability theory. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1975**(3), 89–102.

Franken, P., König, D., Arndt, U., and Schmidt, V. (1981). *Queues and point processes.* Akademie-Verlag, Berlin.

Gittins, J. C. (1978). A comparison of service disciplines for $GI/G/m$ queues. *Math. Operat.-forschung Statist., Ser. Optim.*, **9**, 255–60.

Kalashnikov, V. V. and Anichkin, S. A. (1981). Continuity of random sequences and approximation of Markov chains. *Adv. Appl. Probab.*, **13**, 402–14.

Karpelevitch, F. I. and Kreinin, A. Ya. (1976). Estimates of the mean waiting time in a single-queue system ($GI/G/1$). *Izv. Akad. Nauk SSSR Techn. Kibernet.*, **1976**(3), 105–7. (Transl. *Engng. Cybernetics*, **14**(3), 81–3.)

Karpelevitch F. I. and Kreinin, A. Ya. (1980). Some bounds for the mean waiting time in $E_k/GI/1$ queues. *Math. Operat.-forschung Statist., Ser. Statist.*, **11**, 85–8.

Kiefer, J. and Wolfowitz, J. (1956). On the characteristics of the general queueing process with applications to random walk. *Ann. Math. Statist.*, **27**, 147–61.

Kingman, J. F. C. (1962a). Some inequalities for the queue $GI/G/1$. *Biometrika*, **49**, 315–24.

Kingman, J. F. C. (1962b). On queues in heavy traffic. *J. Roy. Statist. Soc., Ser. B*, **24**, 383–92.

Kingman, J. F. C. (1964). A martingale inequality in the theory of queues. *Proc. Camb. Phil. Soc.*, **59**, 359–61.

Kingman, J. F. C. (1970). Inequalities in the theory of queues. *J. Roy. Statist. Soc., Ser. B*, **32**, 102–10.

Köllerström, J. (1976). Stochastic bounds for the single-server queue. *Math. Proc. Camb. Phil. Soc.*, **80**, 521–5.

Köllerström, J. (1978). Stochastic bounds for the queue $GI/G/1$ in heavy traffic. *Math. Proc. Camb. Phil. Soc.*, **84**, 361–75.

Korzeniowski, A. and Opawski, A. (1976). Bounds for reliability in the NBU, NWU and NBUE, NWUE classes. *Zast. Mat.*, **15**, 1–5.

Kreinin, A. Ya. (1979a). A method of obtaining estimates of the mean characteristics of queueing systems. *Izv. Akad. Nauk SSR Tehn. Kibernet.*, **1979**(5), 68–70. (Transl. *Engng. Cybernetics*, **17**(5), 53–6.)

Kreinin, A. Ya. (1979b). Estimates of the mean waiting time in a single-queue system with readjustment. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1979**(3), 207–10. (Transl. *Engng. Cybernetics*, **17**(3), 143–7.)

Kreinin, A. Ya. (1979c). On some queueing systems. *Applied Mathematics and Railway Transport Problems*. MIIT, Moscow, **640**, 84–7.

Kreinin, A. Ya. (1980). Bounds on mean characteristics of $E_k/GI/1/\infty$ queues. *Math. Operat.-forschung Statist., Ser. Statist.*, **11**, 515–19.

Kreinin, A. Ya. (1981a). Inequality for the probability of idleness in the system $GI/G/1/\infty$. (In Russian) *Kibernetika*, **1981**(5), 139–40. *Cybernetics.*

Kreinin, A. Ya. (1981b). Bounds on waiting time characteristics of a system with Erlang arrivals and general service distribution. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1981**(3), 187–91. (Not transl. in *Engng. Cybernet.*)

Lemoine, A. J. (1976). On random walks and stable $GI/G/1$ queues. *Math. Operat. Res.*, **1**, 159–64.

Lindley, D. V. (1952). The theory of queues with a single server. *Proc. Camb. Phil. Soc.*, **58**, 277–89.

Marshall, K. T. (1988). Some inequalities in queueing. *Operat. Res.*, **16**, 651–65.

Mihoc, A. I. (1980). Determination of the characteristics of waiting in a single-line system from the moments of the initial distributions. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1980**(3), 57–63. (Transl. *Engng. Cybernetics*, **18**(3), 43–9.)

Miyazawa, M. (1975). Stochastic order relations among $GI/G/1$ queues with a common traffic intensity. *J. Operat. Res. Soc. Japan*, **19**, 193–208.

Mori, M. (1976). Some bounds for queues. *J. Operat. Res. Soc. Japan*, **18**, 152–81.

Ott, T. J. (1987). Simple inequalities for the $D/G/1$ queue. *Operat. Res.*, **35**, 589–97.

Powell, B. A. and Avi-Itzhak, B. (1967). Queueing systems with enforced idle time. *Operat. Res.*, **15**, 1145–56.

Riordan, J. (1962). *Stochastic service systems.* Wiley, New York.

Rogozin, B. A. (1966). Some extremal problems in queueing theory. *Teor. Verojatnost. i Primenen.*, **11**, 161–9. (Transl. *Theor. Probab. Appl.*, **11**, 144–51.)

Rolski, T. (1972). Some inequalities for $GI/M/n$ queues. *Zast. Mat.*, **13**, 43–7.

Rolski, T. (1976). A note on orders in the set of distribution functions. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.*, **24**, 513–16.

Rolski, T. (1981). *Stationary random processes associated with point processes.* (Springer Lecture Notes Statistics 5), Springer-Verlag, New York.

Rolski, T. and Stoyan, D. (1974). Two classes of semi-orderings and their application in the queueing theory. *Z. Angew. Math. Mech.*, **54**, 127–8.

Rolski, T. and Stoyan, D. (1976). On the comparison of waiting times in $GI/G/1$ queues. *Operat. Res.*, **24**, 197–200.

Ross, S. M. (1974). Bounds on the delay distribution in $GI/G/1$ queues. *J. Appl. Probab.*, **11**, 417–21.

Rossberg, H.-J. (1968). Optimale Eigenschaften einiger Wartesysteme bei regelmäßigem Eingang bzw. konstanten Bedienungszeiten. *Z. Angew. Math. Mech.*, **48**, 395–403.

Rossberg, H.-J. (1970). Bemerkungen zu einer Arbeit von K. T. Marshall über das Wartemodell $G/G/1$. *Z. Angew. Math. Mech.*, **50**, 423.

Rossberg, H.-J. and Siegel, G. (1974). The $GI/G/1$ model with warming-up time. *Zast. Mat.*, **14**, 17–26.

Saaty, T. L. (1961). *Elements of queueing theory with applications.* McGraw-Hill, New York.

Sacks, J. (1960). Ergodicity of queues in series. *Ann. Math. Statist.*, **31**, 579–88.

Sahin, I. and Perrakis, S. (1976). Moment inequalities for a class of single server queues. *INFOR-Canad. J. Operat. Res. Inf. Proc.*, **14**, 144–52.

Schmidt, V. (1978). On some relations between stationary time and customer state probabilities for queueing systems $G/GI/s/r$. *Math. Operat.-forschung Statist., Ser. Optim.*, **9**, 261–72.

Shanthikumar, J. G. and Yao, D. D. (1991). Strong stochastic convexity: closure properties and applications. *J. Appl. Probab.*, **28**, 131–45.

Siegel, G. (1974). Abschätzungen für die Wartezeitverteilungen und ihrer Momente beim Wartemodell $GI/G/1$ mit und ohne 'Erwärmung'. *Z. Angew. Math. Mech.*, **54**, 609–19.

Stoyan, D. (1972). Abschätzungen für die Bedienungs- und Zuverlässigkeitstheorie. *Zast. Mat.*, **13**, 173–85.

Stoyan, D. (1976). Bemerkungen zu meiner Arbeit 'Abschätzungen für die Konvergenz gegen die stationäre Verteilung monotoner Markowscher Prozesse'. *Math. Nachr.*, **72**, 321–2.

Stoyan, D. (1977a). *Qualitativ Eigenschaften Abschätzungen und stochastischer Modelle.* Akademie-Verlag, Berlin.

Stoyan, D. (1977b). Further stochastic order relations among $GI/GI/1$ queues with a common traffic intensity. *Math. Operat.-forschung Statist., Ser. Optim.*, **8**, 541–8.

Stoyan, D. (1983). *Comparison methods for queues and other stochastic models.* Wiley, Chichester. (Translation with revision of Stoyan (1977a), ed. D. J. Daley.)

Stoyan, D. and Stoyan, H. (1969). Monotonieeigenschaften der Kundenwartezeiten im Modell $GI/G/1$. *Z. Angew. Math. Mech.*, **49**, 729–34.

Stoyan, D. and Stoyan, H. (1974). Inequalities for the mean waiting time in single-line queueing systems. *Izv. Akad. Nauk SSSR Tehn. Kibernet.*, **1974**(6), 104–6. (Transl. *Engng. Cybernetics*, **12**(6), 79–81.)

Stoyan, D. and Stoyan, H. (1976). Some qualitative properties of single server queues. *Math. Nachr.*, **70**, 29–35.

Tan, H. H. (1979). Another martingale bound on the waiting-time distribution in $GI/G/1$ queues. *J. Appl. Probab.*, **16**, 454–7.

Trengove, C. D. (1978). *Bounds for the mean waiting time in queues.* Unpublished M.Sc. thesis, University of Melbourne.

Vainshtein, A. D. (1983). Improved lower bounds for mean characteristics of $E_k/GI/m$ queues. *Math. Operat.-forschung Statist., Ser. Optim.*, **14**, 601–5.

Vainshtein, A. D. and Kreinin, A. Ya. (1981). One estimate of the mean length of a queue in a single-line service system. *Automat. Telemekh.*, **1981**(11), 60–5. (Transl. *Automat. Remote Control*, **42**, 1478–82.)

Whitt, W. (1982). The Marshall and Stoyan bounds for IMRL/$G/1$ queues are tight. *Operat. Res. Letters*, **1**, 209–13.

Whitt, W. (1984a). On approximations for queues, I: Extremal distributions. *AT&T Bell. Lab. Tech. J.*, **63**, 115–38.

Whitt, W. (1984b). On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Lab. Tech. J.*, **63**, 163–75.

Whitt, W. (1984c). Minimizing delays in the *GI/G/1* queue, *Operat. Res.*, **32**, 41–51.

Wolff, R. W. (1977). An upper bound for multi-channel queues. *J. Appl. Probab.*, **14**, 884–8.

Wolff, R. W. (1987). Upper bounds on work for multi-channel queues. *J. Appl. Probab.*, **24**, 547–51.

Wolff, R. W. (1991). On finite delay-moment conditions in queues. *Operat. Res.*, **40**, 771–5.

Wolfson, B. (1984). Some moment results for certain tandem and multiple-server queues. *J. Appl. Probab.*, **21**, 901–10.

Queueing and Related Models

Edited by

U. NARAYAN BHAT

Department of Statistical Science,
Southern Methodist University,
Dallas, Texas, USA

and

ISHWAR V. BASAWA

Department of Statistics,
University of Georgia,
Athens, Georgia, USA

CLARENDON PRESS · OXFORD

1992