Markov Processes Relat. Fields 5, 451–476 (1999)

Markov Processes and Related Fields

Exact Asymptotics for a Large Deviations Problem for the GI/G/1 Queue

S. Asmussen and J.F. Collamore

Centre for Mathematical Sciences, Mathematical Statistics, Lund University, Box 118, S-221 00 Lund, Sweden
E-mail: asmus@maths.lth.se, jfc@maths.lth.se

Received March 17, 1999, revised July 22, 1999

Abstract. Let V be the steady-state workload and Q the steady-state queue length in the GI/G/1 queue. We obtain the exact asymptotics for probabilities of the form $\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}$ as $t \to \infty$. In the light-tailed case, there are three regimes according to the limiting value of $a(t)/b(t)$. Our analysis here extends and simplifies recent work of Aspandiiarov and Pechersky [8]. In the heavy-tailed subexponential case, a lower asymptotic bound is derived and shown to be the exact asymptotics in a regime where $a(t)$, $b(t)$ vary in a certain way determined by the service time distribution.

KEYWORDS: change of measure, conditioned limit theorems, saddlepoint approximations, subexponential distribution, workload

AMS SUBJECT CLASSIFICATION: 60K25

1. Introduction

Consider the GI/G/1 queue in the notation of [3, Chapter VIII], with service times $U_0, U_1, \ldots$ of customers $n = 0, 1, \ldots$ and interarrival times $T_0, T_1, \ldots$, where T_n is the time between the arrival of customer n and $n+1$. The sequences $\{U_n\}$ and $\{T_n\}$ are i.i.d. and mutually independent, with U_n having distribution B (say) and T_n having distribution A. We assume stability $\rho = \mu_B/\mu_A < 1$, where μ_A, μ_B are the means of A, B, and that both A and B are non-lattice. Let V denote the workload in steady state and Q the queue length (the number of customers in the system, including the customer currently being served). A considerable amount of work then deals with tail asymptotics for $\mathsf{P}\{V \geq x\}$, $x \to \infty$, or for $\mathsf{P}\{Q \geq k\}$, $k \to \infty$. The results, which carry far beyond the GI/G/1 setting, show that with light tails one has an exponential decay

$$\mathsf{P}\{V \geq x\} \quad \sim \quad C_V e^{-\gamma x}, \tag{1.1}$$
$$\mathsf{P}\{Q \geq k\} \quad \sim \quad C_Q z^k \tag{1.2}$$

for suitable constants $\gamma > 0$, $z < 1$ and C_V, C_Q (we will return to the exact expressions later); here and in the following, $\sim$ means that the ratio is one in the limit ($x \to \infty$ in (1.1) and $k \to \infty$ in (1.2)). When going beyond the GI/G/1 queue, the exact asymptotics are more difficult to obtain, and as is customary in large deviations theory one considers logarithmic asymptotics. That is, instead of (1.1) one would be satisfied with showing that

$$\log \mathsf{P}\{V \geq x\} \quad \sim \quad -\gamma x$$

and similarly for (1.2) (for examples of such asymptotics in a general setting, see e.g. [19]).

In the heavy-tailed case where the service time distribution B is subexponential (as described in [16] and Section 4), the asymptotics for the GI/G/1 steady-state workload V have long been known:

$$\mathsf{P}\{V \geq t\} \quad \sim \quad \frac{1}{\mu_A - \mu_B} \int_t^\infty \overline{B}(x)\, dx; \tag{1.3}$$

see e.g. [17] and references therein. The asymptotics of the steady-state queue length Q were more recently obtained by Asmussen, Klüppelberg and Sigman [7], who showed that

$$\mathsf{P}\{Q \geq t\} \ \sim \ \mathsf{P}\{V \geq t\mu_A\} \tag{1.4}$$

subject to a certain condition on B, saying basically that the tail decreases slower than $\exp\{-x^{1/2}\}$. For less heavy-tailed subexponential distributions (say Weibull with $\beta \geq 1/2$), the tail of Q is effectively larger than the right-hand side of (1.4), and exact asymptotics were obtained in [7] only for the M/G/1 case.

The present paper is concerned with joint asymptotics for V and Q, that is, we study probabilities of the form $\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}$ as $t \to \infty$. In a recent paper, Aspandiiarov and Pechersky [8] obtained the logarithmic asymptotics for the light-tailed M/G/1 case, where T is exponential, and $a(t) = at$, $b(t) = bt$. Their arguments are based upon a large deviation principle for multidimensional compound Poisson processes. We improve upon the result by Aspandiiarov and Pechersky [8] by

1) replacing the logarithmic asymptotics by exact asymptotics;

2) generalizing from M/G/1 to GI/G/1;

3) presenting a shorter and more straightforward approach; basically, we just use the representation

$$V = U_r + \sum_{i=1}^{Q} U_i', \tag{1.5}$$

where U_r has the residual service distribution B_r and $U_1', U_2', \ldots$ are independent of (Q, U_r) and i.i.d. with distribution B, in combination with sharp estimates (saddlepoint approximations) for sums of i.i.d. random variables and convexity arguments.

In addition, we present various results for the case of heavy-tailed service times, where asymptotic regimes other than $a(t) = at$, $b(t) = bt$ turn out to be of interest. Roughly, the results state that $\mathsf{P}\{V \geq a(t) + b(t)(\mu_A - \mu_B)\}$ is always an asymptotic lower bound for $\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}$ and in some main cases the correct asymptotics.

2. Preliminaries

Notation. For any real-valued distribution F, let

$$\lambda_F(\alpha) = \int_0^\infty e^{\alpha x} F(dx), \qquad \Lambda_F(\alpha) = \log \lambda_F(\alpha)$$

denote the moment generating function and cumulant generating function of F, respectively.

Our starting point is the classical exact asymptotics for the steady-state waiting time W (Feller [18, Chapter XII] or Asmussen [3, Chapter XII])

$$\mathsf{P}\{W > x\} \sim C_W\, e^{-\gamma x}; \tag{2.1}$$

for a strict upper bound it is also known that

$$\mathsf{P}\{W > x\} \leq e^{-\gamma x}. \tag{2.2}$$

The constant γ is the non-zero solution of the equation

$$\lambda_B(\gamma)\lambda_A(-\gamma) = 1, \tag{2.3}$$

and it is assumed (for (2.1)) that $\Lambda_B'(\gamma) < \infty$. In the discussion of light-tailed distributions below, it will always be assumed that this condition holds. (For the corresponding asymptotics when this condition does not hold, see Remark 3.1.)

While the constant γ in (2.1) is readily computed from (2.3), the constant C_W is somewhat more difficult to obtain. Basically it is determined from the solution of a Wiener-Hopf problem. In particular, C_W is explicit if either A or B is exponential (M/G/1 or GI/M/1; Feller [18], Asmussen [3]) and algorithmically computable if either A or B is phase-type (PH/G/1 or GI/PH/1; Asmussen [4]). In view of these references, we will henceforth consider C_W as an explicitly available constant.

The distributions of the steady-state workload V and the steady-state queue length Q can easily be obtained in terms of the distribution of W. Let U_e be a random variable having the stationary excess distribution B_e of U, $B_e(x) = \int_0^x \mathsf{P}\{U > y\}\,dy/\mu_B$, and let $T^{(k)}$ have the distribution A^{*k} of the sum of k independent interarrival times. Then, assuming $W, U_e, T^{(0)}, T^{(1)}, \ldots$ to be independent, one has (see e.g. the books by Cohen [13], Asmussen [3, Chapter VIII.3–4])

$$\mathsf{P}\{V > x\} = \rho\,\mathsf{P}(W + U_e > x), \tag{2.4}$$
$$\mathsf{P}\{Q \geq k\} = \rho\,\mathsf{P}\{W + U_e > T^{(k-1)}\}. \tag{2.5}$$

Combining (2.4)–(2.5) with (2.1), one then obtains in a straightforward way that (1.1), (1.2) hold with

$$C_V = \rho C_W \lambda_{B_e}(\gamma), \quad C_Q = C_V \lambda_B(\gamma), \quad \text{and} \quad z = \lambda_B(\gamma)^{-1}; \tag{2.6}$$

see e.g. Section 2 of [1] or the proof of Corollary 2.1 below for (1.2). Here

$$\lambda_B(\gamma) = \lambda_A(-\gamma)^{-1},$$
$$\lambda_{B_e}(\alpha) = \frac{1}{\mu_B}\left(\frac{\lambda_B(\alpha) - 1}{\alpha}\right), \quad \alpha \neq 0. \tag{2.7}$$

We will need the following extension of (2.5).

Proposition 2.1. $\mathsf{P}\{Q \geq k,\, U_r > y\} = \rho\,\mathsf{P}\{U_e > (T^{(k-1)} - W)^+ + y\}$.

The proof is just as the proof of (2.5) in [3] (in the last display on p. 193, the condition that the residual service time exceeds y leads to replacing U_n by $U_n - y$).

Corollary 2.1. *Let $G_k(y) = \mathsf{P}\{U_r \leq y \mid Q = k\}$, $k \in \mathbf{Z}_+$ and $y \in \mathbf{R}$; and let*

$$G(y) = 1 - \frac{1}{\lambda_{B_e}(\gamma)} \int_y^\infty e^{\gamma(u-y)} B_e(du). \tag{2.8}$$

Then as $k \to \infty$,

(i) *the functions $\{G_k\}$ converge pointwise to G;*

(ii) *the moment generating functions* $\{\lambda_{G_k}\}$ *converge pointwise to* λ_G *on* $\mathfrak{D} = \{\alpha : \lambda_G(\alpha) < \infty\}$. *Also, this convergence is uniform on compact subsets of* int $\mathfrak{D}$.

Proof. (i) The result is given for M/G/1 in [1]. For GI/G/1, we get

$$\mathsf{P}\{Q = k, U_r > y\} = \rho \int_y^\infty B_e(du) \int_0^\infty D(dv)\, \mathsf{P}\{W > v + y - u\},$$

where $D(\cdot) = A^{*(k-1)}(\cdot) - A^{*(k-2)}(\cdot)$. Since $A^{*(k-1)}(a)$ goes to zero faster than any exponential for a fixed a (see [3, p. 113]), the right-hand side behaves like

$$\begin{aligned}
&\rho \int_y^\infty B_e(du) \int_0^\infty D(dv) C_W e^{-\gamma(v+y-u)} \\
&= \rho C_W e^{-\gamma y} \int_y^\infty e^{\gamma u} B_e(du) \int_0^\infty e^{-\gamma v} D(dv) \\
&= \rho C_W e^{-\gamma y} (z^{k-1} - z^{k-2}) \int_y^\infty e^{\gamma u} B_e(du). \qquad (2.9)
\end{aligned}$$

(ii) By a slight variant of (i) (using (2.2) in place of (2.1)),

$$\mathsf{P}\{U_r > y \mid Q = k\} \le \frac{1}{C_W \lambda_{B_e}(\gamma)} \int_y^\infty e^{\gamma(u-y)} B_e(du). \qquad (2.10)$$

Thus the sequence $F_k(\cdot) = 1 - G_k(\cdot)$ is dominated by $\bar{F}(\cdot) = \{1 - G(\cdot)\}/C_W$, all $k \in \mathbf{Z}_+$. Integrate by parts to obtain

$$\lambda_{G_k}(\alpha) = 1 + \alpha \int_0^\infty F_k(u) e^{\alpha u} du, \quad \text{for all } \alpha \in \mathfrak{D}, \qquad (2.11)$$

and similarly for G. Applying the dominated convergence theorem to the right-hand side of (2.11) yields $\lambda_{G_k} \to \lambda_G$ on $\mathfrak{D}$.

Since F_k is dominated by $\bar{F}$, (2.11) also implies $0 \le \lambda_{G_k} \le \lambda_G / C_W$. By the convexity of moment generating functions, it follows that on any compact subset $K \subset \text{int}\, \mathfrak{D}$

$$\sup \left\{ |\lambda'_{G_k}(\alpha)| : \alpha \in K,\ k \in \mathbf{Z}_+ \right\} < \infty.$$

Then the required uniformity follows as a consequence of the mean value theorem, which implies that the oscillation outside a *finite* subset of K is appropriately small. □

Remark 2.1. By a straightforward calculation,

$$\lambda_G(\alpha) = \frac{\alpha\lambda_{B_e}(\alpha) - \gamma\lambda_{B_e}(\gamma)}{(\alpha-\gamma)\lambda_{B_e}(\gamma)}, \quad \alpha \neq \gamma. \tag{2.12}$$

Since moment generating functions are lower semi-continuous, (2.12) also determines Λ_G at the point $\alpha = \gamma$; in particular, $\lambda_G(\gamma) = 1 + \gamma\Lambda'_{B_e}(\gamma)$.

3. Results for the light-tailed case

For light-tailed distributions one obtains distinctively different behavior depending on whether $a/b \le \Lambda'_B(0)$, when the asymptotics are dominated by $\mathsf{P}\{Q \ge bt\}$; or $a/b \ge \Lambda'_B(\gamma)$, when the asymptotics are dominated by $\mathsf{P}\{V \ge at\}$; or $\Lambda'_B(0) < a/b < \Lambda'_B(\gamma)$, when the asymptotics are influenced by *both* V and Q.

We begin by analyzing the intermediate and more interesting case $\Lambda'_B(0) < a/b < \Lambda'_B(\gamma)$. An essential part of the analysis in this case will involve the "saddlepoint approximation" (see e.g. [9, 14, 15, 22–25, 27]).

Definition 3.1. Let F be a distribution function, and let x be any point in the range of the map Λ'_F. The unique solution $s(x)$ to the equation $\Lambda'_F(s(x)) = x$ is called the ***saddlepoint*** of x.

If S_n is the nth sum of an i.i.d. sequence of random variables with law F and one is interested in obtaining large deviation asymptotics of the form $\mathsf{P}\{S_n/n \ge x\}$, then the "saddlepoint" method considers the *conjugate distribution determined by* $s(x)$, namely the distribution

$$\tilde{F}(du) = \frac{e^{s(x)u}}{\lambda_F(s(x))}F(du),$$

which has the important property that its mean $m_F(s(x)) = x$. We also denote the variance of this conjugate distribution by $\sigma_F^2(s(x))$.

Theorem 3.1. *Assume* $\Lambda'_B(0) < a/b < \Lambda'_B(\gamma)$. *Then as* $t \to \infty$

$$\mathsf{P}\{V \ge at,\, Q \ge bt\} \sim \frac{C\exp\{-(at)s + \lceil bt\rceil(\Lambda_B(s) - \Lambda_B(\gamma))\}}{s\sigma_B(s)\sqrt{2\pi bt}}, \tag{3.1}$$

where $C = C_Q\lambda_G(s)(\lambda_B(\gamma)-1)/(\lambda_B(\gamma)-\lambda_B(s))$ *and* $s = s(a/b)$ *is the saddlepoint of* a/b. ($\lceil x\rceil$ *denotes the smallest integer greater than or equal to* x).

Recall that the constants C_Q and $\lambda_G(s)$ were given already in (2.6) and (2.12) and the constant γ in (2.3).

Proof of Theorem 3.1. We begin by establishing the following lemma.

Lemma 3.1. *Let $N \in \mathbb{Z}_+$ and $\Delta > 0$, and define $\mathfrak{B} = \{\beta : b \le \beta \le a/(\Lambda_B'(0) + \Delta)\}$. Then, uniformly for $\beta \in \mathfrak{B}$*

$$\mathsf{P}\{V \ge at \mid Q = \lceil \beta t + N \rceil\} \sim \frac{\lambda_G(s_\beta)\exp\{-(at)s_\beta + \lceil \beta t \rceil \Lambda_B(s_\beta)\}}{s_\beta \sigma_B(s_\beta)\sqrt{2\pi\beta t}} \lambda_B(s_\beta)^N \tag{3.2}$$

as $t \to \infty$, where $s_\beta = s(a/\beta)$ is the saddlepoint of a/β.

Proof of Lemma 3.1. Let $\{U_i'\}$ be an i.i.d. sequence with distribution B, $S_n = U_1' + \cdots + U_n'$, and $G_{\beta t}(u) = \mathsf{P}\{U_r \le u \mid Q = \lceil \beta t + N \rceil\}$. Then $V = S_Q + U_r$ and so

$$\mathsf{P}\{V \ge at \mid Q = \lceil \beta t + N \rceil\} = \int_{u \ge 0} \mathsf{P}\{S_{\lceil \beta t + N \rceil} + u \ge at\} G_{\beta t}(du). \tag{3.3}$$

To determine the limiting form of the quantity inside the integrand, apply the saddlepoint approximation. Namely observe by Theorem 1 of [27] that

$$\mathsf{P}\left\{\frac{S_{\lceil \beta t + N \rceil}}{\lceil \beta t + N \rceil} \ge \frac{(at - u)}{\lceil \beta t + N \rceil}\right\} \sim \frac{\exp(-t\,\mathcal{F}(\alpha_t, \beta_t))}{s(\frac{\alpha_t}{\beta_t})\sigma_B(s(\frac{\alpha_t}{\beta_t}))\sqrt{2\pi\beta_t t}}, \quad t \to \infty, \tag{3.4}$$

where

$$\alpha_t = \frac{at - u}{t}, \quad \beta_t = \frac{\lceil \beta t + N \rceil}{t}, \quad \mathcal{F}(\alpha, \beta) = \alpha s\Big(\frac{\alpha}{\beta}\Big) - \beta \Lambda_B\Big(s\Big(\frac{\alpha}{\beta}\Big)\Big). \tag{3.5}$$

Moreover, this convergence is uniform for $\beta \in \mathfrak{B}$ and $0 \le u \le c_\beta t$, where

$$c_\beta = a - \beta(\Lambda_B'(0) + \delta)$$

and $\delta > 0$ is a constant which will be fixed later in the discussion following (3.16). (To establish uniformity from Petrov's result, note that $u \ge 0$, $b \ge \beta$ and $a/b < \Lambda_B'(\gamma)$ implies $\alpha_t/\beta_t \le \Lambda_B'(\gamma) - \varepsilon$, for some $\varepsilon > 0$. On the other hand, $u \le c_\beta t$ implies $\alpha_t/\beta_t \ge (\Lambda_B'(0) + \delta/2)$, for suitably large t.) Next observe by Theorem 23.5 of [28] that

$$\mathcal{F}(\alpha, \beta) = \beta \Lambda_B^*\Big(\frac{\alpha}{\beta}\Big), \tag{3.6}$$

where Λ_B^* is the convex conjugate of Λ_B. Since $\Lambda_B^{*'}(\cdot) = s(\cdot)$, [28, Theorem 23.5], and Λ_B^* is differentiable on the interior of its domain, [28, Theorem 26.3], it follows by the mean value theorem that

$$\mathcal{F}(\alpha_t, \beta_t) = \mathcal{F}(a, \beta) - \frac{u}{t}s(\theta_t) - \frac{N_t}{t}\Lambda_B(s(\theta_t)), \tag{3.7}$$

where $s(\theta_t)$ is the saddlepoint of some $\theta_t \in [\alpha_t/\beta_t, a/\beta]$ and

$$N_t = N + (\lceil \beta t \rceil - \beta t).$$

Then by (3.4), (3.5) and (3.7)

$$\begin{aligned} &\mathsf{P}\{S_{\lceil \beta t + N \rceil} + u \geq at\} \\ &\quad \sim \frac{\exp\{-(at)s_\beta + (\beta t)\Lambda_B(s_\beta)\}}{s\left(\frac{\alpha_t}{\beta_t}\right)\sigma_B\left(s\left(\frac{\alpha_t}{\beta_t}\right)\right)\sqrt{2\pi\beta t}} \lambda_B(s(\theta_t))^{N_t} e^{s(\theta_t)u}, \quad t \to \infty, \end{aligned} \tag{3.8}$$

uniformly for $\beta \in \mathfrak{B}$ and $0 \leq u \leq c_\beta t$, for some $\theta_t \in [\alpha_t/\beta_t, a/\beta]$.

Roughly speaking, substitution of this last equation into (3.3) will establish the lemma, provided it can be shown that the relevant limit can be brought outside the integral. The remainder of the proof is devoted to formally dealing with this problem. For this purpose we introduce some further notation, as follows. Let

$\mathfrak{g}_\beta(t)$ = the right-hand side of (3.2);

$\mathfrak{f}_\beta(u,t)$ = (the right-hand side of (3.8)) $/\mathfrak{g}_\beta(t)$;

$\mathcal{G}_{\beta,t}(u) = \mathsf{P}\left\{S_{\lceil \beta t + N \rceil} + u \geq at\right\}/\mathfrak{g}_\beta(t)$;

$\tilde{U}_{\beta t}$ = a random variable with distribution
$G_{\beta t}(u) = \mathsf{P}\{U_r \leq u \mid Q = \lceil \beta t + N \rceil\}$
(independent of the service times $\{U_i'\}$).

In this notation, (3.8) becomes

$$\mathcal{G}_{\beta,t}(u) \sim \mathfrak{f}_\beta(u,t) \quad \text{as } t \to \infty, \tag{3.9}$$

uniformly for $\beta \in \mathfrak{B}$ and $0 \leq u \leq c_\beta t$. It needs to be shown that

$$\lim_{t\to\infty} \mathsf{E}\left[\mathcal{G}_{\beta,t}(\tilde{U}_{\beta t})\right] = 1, \quad \text{uniformly for } \beta \in \mathfrak{B} \tag{3.10}$$

(which is a natural consequence of (3.9), since $s(\alpha_t/\beta_t)$, $s(\theta_t) \to s_\beta$ and hence $\mathfrak{f}_\beta(u,t) \cong \exp(s_\beta u)/\lambda_G(s_\beta)$).

Proceeding more formally, we begin by analyzing the function $\mathfrak{f}_\beta(u,t)$. Note that the saddlepoints s_β decrease as β increases, and

$$\beta_\Delta \stackrel{\text{def}}{=} \frac{a}{\Lambda_B'(0) + \Delta} \implies \Lambda_B'(s_{\beta_\Delta}) > \Lambda_B'(0) \implies s_{\beta_\Delta} > 0. \tag{3.11}$$

Thus $\{s_\beta : \beta \in \mathfrak{B} = [b, \beta_\Delta]\}$ forms a bounded interval disjoint from $\{0\}$ (and likewise for $\sigma_B(s_\beta)$); and with minor adjustments (as in the discussion following (3.5)), these same properties are obtained for $s(\alpha_t/\beta_t)$, $\sigma_B(\alpha_t/\beta_t)$ on the restricted range $0 \leq u \leq c_\beta t$ and $\beta \in \mathfrak{B}$. Hence, the definition of $\mathfrak{f}_\beta$ implies

$$\mathfrak{f}_\beta(u,t) \leq \text{const} \cdot \exp(s(\theta_t)u), \qquad 0 \leq u \leq c_\beta t \quad \text{and} \quad \beta \in \mathfrak{B}. \tag{3.12}$$

Since $s(\theta_t) \le s_b < \gamma$ and $\lambda_G(\gamma) < \infty$, we may then apply Corollary 2.1 (ii) to obtain

$$\mathsf{E}\left[\mathfrak{f}_\beta(\tilde{U}_{\beta t}, t);\ 0 \le \tilde{U}_{\beta t} \le c_\beta t\right] \le \text{const} < \infty, \tag{3.13}$$

for all $t \ge T_0$ for some T_0 and $\beta \in \mathfrak{B}$. It then follows by (3.9) that, uniformly for $\beta \in \mathfrak{B}$,

$$\begin{aligned}\lim_{t\to\infty} \Bigg\{ &\mathsf{E}\left[\mathcal{G}_{\beta,t}(\tilde{U}_{\beta t}) - \frac{\exp(s_\beta \tilde{U}_{\beta t})}{\lambda_G(s_\beta)};\ 0 \le \tilde{U}_{\beta t} \le c_\beta t\right] \\ + \ &\mathsf{E}\left[\frac{\exp(s_\beta \tilde{U}_{\beta t})}{\lambda_G(s_\beta)} - \mathfrak{f}_\beta(\tilde{U}_{\beta t}, t);\ 0 \le \tilde{U}_{\beta t} \le c_\beta t\right]\Bigg\} = 0. \end{aligned}\tag{3.14}$$

We now show that the restriction to $\{0 \le \tilde{U}_{\beta t} \le c_\beta t\}$ can be removed from the first term of (3.14), and that the second term of (3.14) can, in an appropriate sense, be neglected.

We start by analyzing the first term. By Chebychev's inequality,

$$\mathsf{P}\{\tilde{U}_{\beta t} \ge c_\beta t\} \le \exp(-\gamma c_\beta t)\lambda_{G_{\beta t}}(\gamma). \tag{3.15}$$

The exponential term on the right decays in t like

$$\begin{aligned}\gamma(a - \beta\Lambda_B'(0) - \beta\delta) &= \frac{\gamma}{s_\beta}(s_\beta a - \beta s_\beta \Lambda_B'(0)) - \gamma\beta\delta \\ &\ge (s_\beta a - \beta\Lambda_B(s_\beta)) + \varepsilon \end{aligned}\tag{3.16}$$

for δ sufficiently small and $\varepsilon > 0$ independent of $\beta \in \mathfrak{B}$. (We have used the hypothesis $a/b < \Lambda_B'(\gamma)$, which implies $s_\beta \le s_b < \gamma$. Also, we have used the assumption $\beta \le a/(\Lambda_B'(0) + \Delta)$, which by (3.6) implies $(s_\beta a - \beta\Lambda_B(s_\beta))$ is bounded away from 0.) We now conclude that the exponential term in (3.15) decays at a faster rate than the exponential term in the definition of $\mathfrak{g}_\beta(t)$. Hence

$$\lim_{t\to\infty} \mathsf{P}\{\tilde{U}_{\beta t} \ge c_\beta t\} \cdot \frac{1}{\mathfrak{g}_\beta(t)} = 0, \tag{3.17}$$

uniformly for $\beta \in \mathfrak{B}$, implying

$$\lim_{t\to\infty} \mathsf{E}[\mathcal{G}_{\beta,t}(\tilde{U}_{\beta t});\ \tilde{U}_{\beta t} \ge c_\beta t] = 0, \tag{3.18}$$

uniformly for $\beta \in \mathfrak{B}$. Also, by a slight variant of Corollary 2.1 (ii),

$$\mathsf{E}\left[\exp(s_\beta \tilde{U}_{\beta t});\ \tilde{U}_{\beta t} > L\right] \to \mathsf{E}\left[\exp(s_\beta U);\ U > L\right] \quad \text{as } t \to \infty, \tag{3.19}$$

where U is a random variable having the distribution function G given in Corollary 2.1 (and this convergence is uniform on $\operatorname{int}\mathfrak{D}$, where $\mathfrak{D}$ is the domain of the right-hand side). The quantity on the right can be made arbitrarily small, for $\beta \in \mathfrak{B}$, for a sufficiently large choice of L. Equations (3.18) and (3.19) imply that the restriction to $\{0 \le \tilde{U}_{\beta t} \le c_\beta t\}$ in the first term on the left of (3.14) can be dropped.

Next consider the second term on the left of (3.14). From the definition of $\mathfrak{f}_\beta$ and the convergence $s(\alpha_t/\beta_t)$, $s(\theta_t) \to s_\beta$, we have

$$\lim_{t\to\infty} \mathfrak{f}_\beta(u,t) = \frac{\exp(s_\beta u)}{\lambda_G(s_\beta)} \tag{3.20}$$

uniformly for $\beta \in \mathfrak{B}$ and $0 \le u \le L < \infty$. Hence

$$\lim_{t\to\infty} \mathsf{E}\left[\frac{\exp(s_\beta \tilde{U}_{\beta t})}{\lambda_G(s_\beta)} - \mathfrak{f}_\beta(\tilde{U}_{\beta t}, t);\ 0 \le \tilde{U}_{\beta t} \le L\right] = 0 \tag{3.21}$$

uniformly for $\beta \in \mathfrak{B}$. It remains to consider this integral over the range $\{L < \tilde{U}_{\beta t} \le c_\beta t\}$. But by (3.12) and another application of (3.19), we see that this part may also be neglected for L suitably large.

By (3.14) we now conclude

$$\lim_{t\to\infty} \mathsf{E}\left[\mathcal{G}_{\beta,t}(\tilde{U}_{\beta t}) - \frac{\exp\left(s_\beta \tilde{U}_{\beta t}\right)}{\lambda_G(s_\beta)}\right] = 0, \tag{3.22}$$

uniformly for $\beta \in \mathfrak{B}$. Since $\lambda_{G_{\beta t}}$ converges uniformly to λ_G for $\beta \in \mathfrak{B}$, by Corollary 2.1 (ii), the desired result follows from (3.22). □

Let $r = \lambda_B(s_\beta)/\lambda_B(\gamma)$. Then by Lemma 3.1 and (1.2), (2.6),

$$\mathsf{P}\left\{V \ge at,\ Q = \lceil \beta t + N \rceil\right\} \sim \frac{C_Q(1-z)\lambda_G(s_\beta)\exp\left\{-(at)s_\beta + \lceil \beta t\rceil\left(\Lambda_B(s_\beta) - \Lambda_B(\gamma)\right)\right\}}{s_\beta \sigma_B(s_\beta)\sqrt{2\pi\beta t}}\, r^N \tag{3.23}$$

as $t \to \infty$, uniformly for $\beta \in \mathfrak{B}$. Hence

$$\mathsf{P}\left\{V \ge at, Q \in \left[\lceil bt\rceil, \lceil bt + M\rceil\right)\right\} \sim \mathfrak{h}(t)(1-r)\sum_{N=0}^{M-1} r^N, \tag{3.24}$$

where $\mathfrak{h}(t)$ is the function on the right-hand side of (3.1) and M is any positive integer. To complete the proof we show

Lemma 3.2. $\lim_{M\to\infty}\left\{\lim_{t\to\infty} \mathsf{P}\{V \ge at,\ Q \ge bt + M\}/\mathfrak{h}(t)\right\} = 0.$

Proof. Let $\Delta > 0$. First consider $\mathsf{P}\left\{V \ge at,\ Q \in [bt + M, \beta_\Delta t)\right\}$, where $\beta_\Delta = a/(\Lambda_B'(0) + \Delta)$. By (3.23), this probability decays asymptotically as $t \to \infty$ like

$$\int_{b+M/t}^{\beta_\Delta} \frac{C_Q(1-z)\lambda_G(s_\beta)\exp\left(-t\,\mathcal{H}_a(\beta)\right)}{s_\beta \sigma_B(s_\beta)\sqrt{2\pi\beta t}}\, t\, d\beta, \tag{3.25}$$

where

$$\mathcal{H}_a(\beta) = \beta\Big(\frac{a}{\beta}s_\beta - \Lambda_B(s_\beta) + \Lambda_B(\gamma)\Big). \tag{3.26}$$

(We ignore the effect of the $\lceil\cdot\rceil$ operation, which will prove negligible here.)

Now consider the behavior of $\mathcal{H}_a(\beta)$ as a function of β. By Theorem 23.5 of [28],

$$\mathcal{H}_a(\beta) \geq \beta\Big(\frac{a}{\beta}\gamma - \Lambda_B(\gamma) + \Lambda_B(\gamma)\Big) = a\gamma \tag{3.27}$$

with equality iff $\Lambda_B'(\gamma) = a/\beta \;\big(= \Lambda_B'(s_\beta)\big) \iff \gamma = s_\beta$. By assumption

$$\Lambda_B'(\gamma) > \frac{a}{b} \geq \frac{a}{\beta} \qquad \text{for } \beta \geq b, \tag{3.28}$$

implying $\gamma \notin \big\{s_\beta : \beta \in [b,\infty)\big\}$. Hence we conclude that $\mathcal{H}_a(\beta)$ does *not* attain its minimum as a function of β on the half-line $[b,\infty)$. Since $\mathcal{H}_a(\beta)$ is convex ((3.6) and [28, p. 35]), it follows that

$$\mathcal{H}_a(\beta) \geq \mathcal{H}_a(b) + K(\beta - b), \quad \beta \in [b,\infty), \tag{3.29}$$

for some $K > 0$. Since the set of saddlepoints $\{s_\beta : \beta \in [b,\, \beta_\Delta]\}$ form a bounded interval disjoint from $\{0\}$ (as in (3.11) above), and similarly for $\{\sigma_B(s_\beta)\}$ and $\{\lambda_G(s_\beta)\}$, it follows by (3.29) that (3.25) decays at least as fast as

$$\text{const}\cdot \frac{C_Q(1-z)\lambda_G(s_\beta)\exp(-t\mathcal{H}_a(\beta))}{s_\beta\sigma_B(s_\beta)\sqrt{2\pi\beta t}}\bigg|_{\beta = b + \frac{M}{t}} = \text{const}\cdot \mathfrak{h}(t)(1-r)r^M. \tag{3.30}$$

This establishes the desired result over the restricted range $Q \in [bt + M, \beta_\Delta t)$.

Finally consider $\mathsf{P}\{V \geq at,\, Q \geq \beta_\Delta t\}$. Note by (3.6) and (3.29) that

$$\mathcal{H}_a(b) < \mathcal{H}_a(\beta_0) = \beta_0\Lambda_B(\gamma), \tag{3.31}$$

where

$$\beta_0 = \lim_{\Delta\to 0}\beta_\Delta = \frac{a}{\Lambda_B'(0)}.$$

Thus $\beta_\Delta\Lambda_B(\gamma) > \mathcal{H}_a(b)$, for some $\Delta > 0$. For this Δ, (1.2), (2.6) then imply that $\mathsf{P}\{Q \geq \beta_\Delta t\} \geq \mathsf{P}\{V \geq at,\, Q \geq \beta_\Delta t\}$ decays faster than $\mathfrak{h}(t)$ as $t \to \infty$. □

We conclude this section by considering the remaining two cases.

Theorem 3.2. (i) *Assume* $0 \leq a/b < \Lambda_B'(0)$. *Then as* $t \to \infty$

$$\mathsf{P}\{V \geq at,\, Q \geq bt\} \sim \kappa\,\mathsf{P}\{Q \geq bt\} \;\sim \kappa C_Q z^{\lceil bt\rceil} \tag{3.32}$$

with $\kappa = 1$. *If* $\Lambda_B'(0) = a/b$ *and* $\Lambda_B''(\gamma) < \infty$, *then this equation holds with* $\kappa = 1/2$.

(ii) *Assume* $\Lambda_B'(\gamma) < a/b \le \infty$. *Then as* $t \to \infty$

$$\mathsf{P}\{V \ge at,\, Q \ge bt\} \sim \tilde{\kappa}\,\mathsf{P}\{V \ge at\} \sim \tilde{\kappa} C_V e^{-\gamma bt} \tag{3.33}$$

with $\tilde{\kappa} = 1$. *If* $\Lambda_B'(\gamma) = a/b$ *and* $\Lambda_B''(\gamma) < \infty$, *then this equation holds with* $\tilde{\kappa} = 1/2$.

Proof. (i) If $a/b < \Lambda_B'(0)$, then (3.32) is immediate from the law of large numbers and (1.5), since the conditional distribution of $\sum_1^Q U_i'/Q$ given $Q \ge k$ converges to the degenerate distribution at μ_B as $k \to \infty$.

If $a/b = \Lambda_B'(0)$ and $\Lambda_B''(\gamma) < \infty$, then (3.32) follows from (1.5) and a slight variant of the central limit theorem as given in Theorem 7.2 of [10]. For further details, see Proposition A.1 below, where a general conditional limit theorem is established.

(ii) First suppose $a/b > \Lambda_B'(\gamma)$. Consider a doubly infinite stationary version $\{(V_s^*, Q_s^*)\}_{-\infty<s<\infty}$ of $\{(V_s, Q_s)\}_{s\ge 0}$. If $V_0^* \ge at$, then it is well-known that this occurred as a consequence of an exponential tilted buildup where $\Lambda_A(\alpha)$, $\Lambda_B(\alpha)$, changed to $\Lambda_A(\alpha-\gamma)-\Lambda_A(-\gamma)$, respectively $\Lambda_B(\alpha+\gamma)-\Lambda_B(\gamma)$ (see [2]). Denote the corresponding exponentially tilted probability measure by P_γ.

More precisely, let $\mathsf{P}^t(\cdot) = \mathsf{P}\{\cdot \mid V_0^* \ge t\}$; let $-\tau \stackrel{\text{def}}{=} \sup\{s \le 0 : V_s^* = 0\}$ denote the start of the busy period straddling 0; and let $N_A^*(s)$, $N_B^*(s)$ denote the number of arrivals, respectively service events, in $[-s, 0]$, $s \ge 0$. Then, by an easy variant of [2, Section 7], one has $\mathsf{P}^t\{F_{t,\varepsilon}\} \to 1$ for any ε, where $F_{t,\varepsilon}$ denotes the event

$$\left\{\tau \in \frac{t(1\pm\varepsilon)}{\overline{\mu}_A/\overline{\mu}_B - 1},\quad N_A^*(\tau) \in \overline{\mu}_A\tau(1\pm\varepsilon),\quad N_B^*(\tau) \in \overline{\mu}_B\tau(1\pm\varepsilon)\right\} \tag{3.34}$$

and $\overline{\mu}_A = 1/\Lambda_A'(-\gamma)$, $\overline{\mu}_B = 1/\Lambda_B'(\gamma)$ are the means in the P_γ distributions of $N_A(1)$, $N_B(1)$. Next observe that $a/b > 1/\overline{\mu}_B$ implies that for some $\varepsilon > 0$

$$\{\overline{\mu}_A a(1-\varepsilon) - \overline{\mu}_B a(1+\varepsilon)\}(1-\varepsilon)\Big(\frac{1}{\overline{\mu}_A/\overline{\mu}_B - 1}\Big) \ge b.$$

Then $Q_0^* = N_A^*(\tau) - N_B^*(\tau) \ge bt$ on $F_{at,\varepsilon}$ and hence

$$\mathsf{P}^{at}\{Q_0^* \ge bt\} \ge \mathsf{P}^{at}\{F_{at,\varepsilon}\} \sim 1.$$

Finally, if $a/b = \Lambda_B'(\gamma)$, then a refinement of (3.34) is needed, namely a conditional central limit theorem in place of a conditional law of large numbers. This is given below in Proposition A.1. If $a/b = \Lambda_B'(\gamma)$ and $\Lambda_B''(\gamma) < \infty$, then it follows as a consequence of Proposition A.1 that $\mathsf{P}^{at}\{Q_0^* \ge bt\} \sim 1/2$. □

Remark 3.1. In the above discussion, it has been assumed that a solution γ to (2.3) exists and $\Lambda_B'(\gamma) < \infty$. We conclude this section with some comments on the asymptotics for light-tailed distributions which do not satisfy this condition.

In that case, (2.1) does not hold; instead one has

$$\mathsf{P}\{W > x\} = o(e^{-\gamma x}), \tag{3.35}$$

where $\gamma = \sup\{\alpha : \lambda_B(\alpha)\lambda_A(-\gamma) \leq 1\}$.

(Under stronger conditions, somewhat more can be said; see [11, Theorem 22.12], or [17, Theorem 5.1].)

Using (3.35) in place of (2.1), one obtains natural analogs to the asymptotics stated in Theorems 3.1 and 3.2, namely,

$$\mathsf{P}\{V \geq at,\, Q \geq bt\} = o(f(t)), \tag{3.36}$$

where $f(t)$ is the right-hand side of either (3.1), (3.32) or (3.33) (depending on whether $\Lambda_B'(0) < a/b < \lim_{\alpha\to\gamma} \Lambda_B'(\gamma)$; $0 \leq a/b \leq \Lambda_B'(0)$; or $a/b \geq \lim_{\alpha\to\gamma} \Lambda_B'(\alpha)$, respectively).

4. Results for the heavy-tailed case

We now assume that the service time distribution B is subexponential (see [16]). As an additional regularity condition, we assume

$$\lim_{x\to\infty} \mathsf{P}\left\{\frac{U-x}{h(x)} > y \,\middle|\, U > x\right\} \;=\; \mathsf{P}\{Z > y\} \tag{4.1}$$

for some random variable Z, where

$$h(x) \;=\; \mathsf{E}\left[U - x \mid U > x\right] \;=\; \frac{1}{\overline{B}(x)} \int_x^\infty \overline{B}(y)\, dy$$

is the so-called auxiliary function. See [20] for a discussion of this condition and [6, 7, 26] for applications where it occurs.

In particular, this set-up covers the regularly varying case, where $\overline{B}(x) = L(x)/x^\alpha$ for some slowly varying function $L(x)$ and $\alpha > 1$ (to ensure finite mean); here one can take

$$h(t) = \frac{x}{\alpha}, \quad \mathsf{P}\{Z > y\} = \frac{1}{(1 + y/(\alpha-1))^\alpha} \tag{4.2}$$

(i.e. Z is Pareto distributed with mean one). In other examples, B is necessarily less heavy-tailed ($\overline{B}(x)/x^\alpha \to 0$ for all $\alpha < \infty$); one has $h(x)/x \to 0$ and Z is standard exponential. Further, one has the self-neglecting property

$$\lim_{x\to\infty} \frac{h(x + yh(x))}{h(x)} \;=\; 1 \tag{4.3}$$

uniformly in $|y| \le y_0$.

For example, B could be lognormal-like, $\overline{B}(x) \sim cx^\beta e^{-\lambda \log^\gamma x}$ with $\gamma > 1$, in which case $h(x) \sim x \log^{1-\gamma} x/(\lambda\gamma)$, or Weibull-like, $\overline{B}(x) \sim cx^\alpha e^{-\lambda x^\beta}$ with $0 < \beta < 1$, in which case $h(x) \sim x^{1-\beta}/(\alpha\beta)$.

We will often use the following standard property of subexponential distributions: for any fixed a,

$$\lim_{t\to\infty} \frac{\overline{B}(t+a)}{\overline{B}(t)} = \lim_{t\to\infty} \frac{\int_{t+a}^\infty \overline{B}(y)\,dy}{\int_t^\infty \overline{B}(y)\,dy}. \tag{4.4}$$

We also assume that the interarrival time T is not too heavy-tailed in the sense that $\mathsf{E}\,T^2 < \infty$.

Note for the following that we can rewrite (1.3) in either of the ways

$$\mathsf{P}\{V \ge t\} \sim \frac{h(t)\overline{B}(t)}{\mu_A - \mu_B} = \frac{1}{\mu_A - \mu_B} \int_t^\infty (y-t)^+ \, B(dy)\,. \tag{4.5}$$

It is immediate as in Theorem 3.2 that

$$\limsup_{t\to\infty} \frac{a(t)}{b(t)} < \mu_B \implies \mathsf{P}\{V \ge a(t), Q \ge b(t)\} \sim \mathsf{P}\{Q \ge b(t)\} \tag{4.6}$$

and thus we assume

$$\liminf_{t\to\infty} \frac{a(t)}{b(t)} > \mu_B \tag{4.7}$$

(the liminf may be infinite). We will present first (Section 4.1) a heuristic argument indicating that one would then expect

$$\mathsf{P}\{V \ge a(t),\, Q \ge b(t)\} \sim \mathsf{P}\{V > c(t)\}, \tag{4.8}$$

where $c(t) = a(t) + b(t)(\mu_A - \mu_B)$.

In Section 4.2, we show that the right-hand side of (4.8) is always a lower bound but in fact not always the correct asymptotics. In Section 4.3, we then supplement this with upper bounds leading to criteria for (4.8) to be true. Note that in heavy-tailed asymptotic problems, it is usually easier to get lower than upper bounds, cf. e.g. [12] and [29].

4.1. Heuristics

In the heavy-tailed case, there are two obstacles for using the representation $V = U_r + \sum_1^Q U_i'$. First, the tail behavior of U_r, given that Q is large, needs to be derived. Second, when both Q and the U_i' are heavy-tailed, the best result on the tail behavior of sums like $\sum_1^Q U_i'$ that we know of is Theorem 3.6

of [7], but the technical condition (c) there creates difficulties. For example, in the regularly varying case (c) would require $n\overline{B^{*n}}(n(\mu_B+\varepsilon)) = o(n^{-(\alpha-1)})$, but according to [26] the correct order is $O(n^{-(\alpha-1)})$. (There are also difficulties in applying [26] say in the Weibull case.)

Our arguments instead use the regenerative representation

$$\mathsf{P}\{V \geq a(t),\, Q \geq b(t)\} \;=\; \frac{1}{\mathsf{E}C}\,\mathsf{E}\int_0^C I\{V_s \geq a(t),\, Q_s \geq b(t)\}\,ds, \tag{4.9}$$

where C is the busy cycle (i.e. the sum of the busy and idle periods), combined with the description of a cycle with V_s exceeding a large level $a(t)$ given in [5]: roughly, the exceedance happens as consequence of one early big service time, say of size $y > a(t)$, and apart from this, everything in the cycle develops normally. Using the law of large numbers and ignoring random fluctuations, this leads to the simplified picture of the cycle in Figure 1.

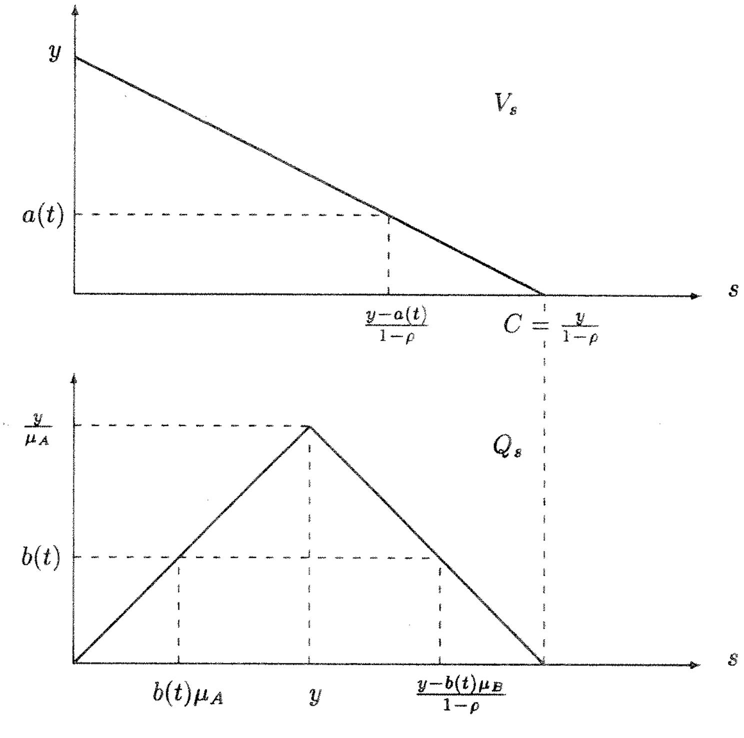

Figure 1

To understand the figure, note that V_s decreases linearly at rate $1 - \mu_B/\mu_A = 1 - \rho$; in particular, level $a(t)$ is downcrossed at time $(y - a(t))/(1-\rho)$, and the busy cycle ends at time $y/(1-\rho)$. The queue length builds up linearly at rate $1/\mu_A$ during the long service time of duration y; in particular, level $b(t)$ is upcrossed at time $b(t)\mu_A$. At time y, we thus have $Q_y = y/\mu_A$ and after that Q_s decreases linearly at rate $1/\mu_B - 1/\mu_A$, implying that level $b(t)$ is downcrossed at time

$$y + \frac{y/\mu_A - b(t)}{1/\mu_B - 1/\mu_A} = \frac{y - b(t)\mu_B}{1-\rho}.$$

When $a(t) > b(t)\mu_A$, we have $(y - b(t)\mu_B)/(1-\rho) > (y - a(t))/(1-\rho)$, and thus the time interval in the cycle where $V_s \geq a(t), Q_s \geq b(t)$ has length

$$\Big(\frac{y - a(t)}{1-\rho} - b(t)\mu_A\Big)^+.$$

Combined with the results of Asmussen [5], we are led heuristically to expect that

$$\begin{aligned}
&\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\} \\
&\sim \frac{1}{\mathsf{E}C}\frac{\mathsf{E}C}{\mu_A}\int_{a(t)}^{\infty}\Big(\frac{y-a(t)}{1-\rho} - b(t)\mu_A\Big)^+ B(dy) \\
&= \frac{1}{\mu_A - \mu_B}\int_{a(t)}^{\infty}(y - c(t))^+ B(dy) \\
&= \frac{1}{\mu_A - \mu_B}\int_{c(t)}^{\infty}\overline{B}(y)\,dy \sim \mathsf{P}\{V \geq c(t)\}.
\end{aligned}$$

One further indication of this analysis is that Q given $V > a(t)$ should be of the order of magnitude $h(a(t))$ (this follows since $y - a(t)$ is of order $h(a(t))$). In fact, it follows from Theorem 4.1 below that

$$\lim_{x\to\infty}\limsup_{t\to\infty}\mathsf{P}\{Q \geq xh(t) \mid V \geq t\} = 0,$$

whereas (4.6) may be rewritten

$$\lim_{t\to\infty}\mathsf{P}\{V \geq dt \mid Q \geq t\} = 1, \quad d < \mu_B.$$

In the regularly varying case ($h(x) = x/\alpha$), this is similar to the light-tailed case, but for other subexponential distributions ($h(x) = o(x)$) the conclusion is that Q is typically smaller than V; in particular, the distributional Little's law underlying the proof of (1.4) in [7] does not hold in the sample path sense.

4.2. Lower bounds and counterexamples

In fact, (4.8) is not always correct.

Proposition 4.1. *If* $h(c(t))/\sqrt{b(t)} \to 0$, *then*

$$\lim_{t\to\infty} \frac{\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}}{\mathsf{P}\{V \geq c(t)\}} = \infty.$$

Note in particular that $h(c(t))/\sqrt{b(t)} \to 0$ holds if $a(t) = at$, $b(t) = bt$ and $h(x)/x^{1/2} \to 0$.

But (4.8) is always a lower bound.

Proposition 4.2. *Without conditions beyond* (4.7),

$$\liminf_{t\to\infty} \frac{\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}}{\mathsf{P}\{V \geq c(t)\}} \geq 1.$$

For the proof, define $\omega(a) = \inf\{t > 0 : V_t > a\}$,

$$\tau_Q(t) = \inf\{s : Q_s \geq b(t)\}, \quad \tau_V(t;y) = \inf\{s : V_s \leq a(t) \mid V_0 = y\},$$

and let P^y refer to the probability measure with $V_0 = y$, $Q_0 = 0$, $y > a(t)$. It follows from [5], (2.2), that we can write

$$p_1(a, a_0) \leq \mathsf{P}\{\omega(a) < C,\ V_{\omega(a)-} \leq a_0\} \leq p_2(a, a_0), \tag{4.10}$$

where

$$\lim_{a_0\to\infty} \liminf_{a\to\infty} \frac{p_1(a, a_0)}{\overline{B}(a)\,\mathsf{E}\,C/\mu_A} = 1, \qquad \lim_{a_0\to\infty} \limsup_{a\to\infty} \frac{p_2(a, a_0)}{\overline{B}(a)\,\mathsf{E}\,C/\mu_A} = 1$$

(note that $\mathsf{E}\,C/\mu_A$ is the expected number of arrivals during a cycle). In a cycle where $V_s \geq a(t)$ for some s, we must have $\omega(a(t)) < C$ and so for $y > a(t) > a_0$,

$$\mathsf{P}\left\{\omega(a(t)) < C,\ V_{\omega(a(t))} > y, V_{\omega(a(t))-} \leq a_0\right\} = \mathsf{P}\left\{\omega(y) < C,\ V_{\omega(y)-} \leq a_0\right\},$$

which can be bounded below and above by $p_1(y, a_0)$, respectively $p_2(y, a_0)$. For lower bounds, we bound $Q_{\omega(a)-}$ below by 0 and obtain (first let $t \to \infty$ and next $a_0 \to \infty$)

$$\liminf_{t\to\infty} \frac{\mathsf{P}\{V \geq a(t),\ Q \geq b(t)\}}{(\mu_A)^{-1}\,\mathsf{E}\int_{a(t)}^{\infty} B(dy)\,\mathsf{E}^y \int_0^C I\{V_s \geq a(t),\ Q_s \geq b(t)\}\,ds} \geq 1. \tag{4.11}$$

Proof of Proposition 4.2. Consider first the case of finite variance. Combining the heuristics behind Figure 1 with the central limit theorem (which requires

$\mathsf{E}U^2 < \infty$ for V_s), it then follows that given $\varepsilon > 0$, one can choose z, K such that

$$\mathsf{P}^y\Big\{Q_s \geq b(t) \quad \text{for all } s \in \Big(b(t)\mu_A + z\sqrt{b(t)\mu_A/(1-\rho)},\ \frac{y-a(t)}{1-\rho}\Big]\Big\} > 1-\varepsilon,$$

$$\mathsf{P}^y\Big\{V_s \geq a(t) \quad \text{for all } s \in \Big[0,\ \frac{y-a(t)}{1-\rho} - z\sqrt{y-a(t)}\Big]\Big\} > 1-\varepsilon$$

for all $y \geq a(t) + K$. Hence by (4.11), we have asymptotically that

$$\begin{aligned}
&\mathsf{P}\{V \geq a(t),\, Q \geq b(t)\} \\
&\quad \geq \frac{1-2\varepsilon}{\mu_A} \int_{a(t)+K}^{\infty} \Big(\frac{y-a(t)}{1-\rho} - z\sqrt{y-a(t)} - b(t)\mu_A \\
&\qquad\qquad - z\sqrt{b(t)\mu_A/(1-\rho)}\Big)^+ B(dy) \\
&\quad \geq \frac{1-2\varepsilon}{\mu_A - \mu_B} \int_{a(t)+K}^{\infty} \Big(y - c(t) - z\sqrt{y-c(t)}\Big)^+ B(dy), \qquad (4.12)
\end{aligned}$$

using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Since $y - c(t) - z\sqrt{y-c(t)} \geq 0$ for $y \geq c(t) + z$, and $c(t) + z > a(t) + K$ for all large t, it follows by (4.4) that (4.12) has the same asymptotics as

$$\frac{1-2\varepsilon}{\mu_A - \mu_B} \int_{c(t)+z}^{\infty} \big(c(t) - z\sqrt{y-c(t)}\big)\, B(dy).$$

Further, for fixed ε,

$$\begin{aligned}
&\limsup_{c\to\infty} \frac{\int_{c+K_1}^{\infty} \sqrt{y-c}\, B(dy)}{\int_c^{\infty} (y-c) B(dy)} \\
&\quad \leq \limsup_{c\to\infty} \frac{\varepsilon^{-1}\int_c^{c+\varepsilon^{-2}} B(dy) + \varepsilon \int_{c+\varepsilon^{-2}}^{\infty} (y-c)B(dy)}{\int_c^{\infty}(y-c)B(dy)} = \varepsilon^{-1}\cdot 0 + \varepsilon,
\end{aligned}$$

so that by (1.3)

$$\begin{aligned}
&\liminf_{t\to\infty} \frac{\mathsf{P}\{V \geq a(t),\, Q \geq b(t)\}}{\mathsf{P}\{V \geq c(t)\}} \\
&\quad \geq \liminf_{t\to\infty} \frac{(1-2\varepsilon)\int_{c(t)+K_1}^{\infty} (y-c(t))^+\, B(dy)}{\int_{c(t)}^{\infty} (y-c(t))^+\, B(dy)} - \varepsilon = 1 - 3\varepsilon.
\end{aligned}$$

Let $\varepsilon \downarrow 0$.

In the case of infinite variance (which can only occur in the case of regular variation with $1 < \alpha \leq 2$), the same argument goes through if one replaces $b(t)\mu_A + z\sqrt{b(t)}$ and $(y - a(t))/(1-\rho) - z\sqrt{y - a(t)}$ by $(1+\delta)b(t)\mu_A$, respectively $(1-\delta)(y-a(t))/(1-\rho)$. □

Proof of Proposition 4.1. Consider the events

$$A_t = \Big\{Q_s \geq b(t) \text{ for all } s \in \Big(bt\mu_A - z\sqrt{bt}, \frac{y - at}{1-\rho}\Big]\Big\},$$

$$B_t = \Big\{V_s \geq a(t) \text{ for all } s \in \Big[0, \frac{y - at}{1-\rho}\Big]\Big\}.$$

Since $\mathsf{P}^y\{Q_{b(t)\mu_A - z\sqrt{b(t)}} \geq b(t)\}$ has a limit as $t \to \infty$ of the form $\Phi(-z\sigma)$, it is easily seen that we can choose $z > 0$ such that $\mathsf{P}^y\{A_t\} \geq 1/4$. Similarly, $\mathsf{P}^y\{B_t\}$ depends only on $y - a(t)$ and goes to $1/2$ as $y - a(t) \to \infty$ which (note that A_t means atypically large early values of Q_s) is more than sufficient to ensure $\mathsf{P}^y\{B_t \mid A_t\} \geq 1/4$ for t large and $y \geq at + K$. Hence

$$\begin{aligned}
&\mathsf{P}\{V \geq a(t), Q \geq b(t)\} \\
&\geq \frac{1}{16\mu_A} \int_{a(t)+K}^{\infty} \Big(\frac{y - a(t)}{1-\rho} - b(t)\mu_A + z\sqrt{b(t)}\Big)^+ B(dy) \\
&= \frac{1}{16(\mu_A - \mu_B)} \int_{c'(t)}^{\infty} (y - c'(t))^+ B(dy) = \frac{1}{16(\mu_A - \mu_B)} \int_{c'(t)}^{\infty} \overline{B}(y)\, dy,
\end{aligned}$$

where $c'(t) = c(t) - z(1-\rho)\sqrt{b(t)}$. Writing $c''(t) = c(t) - xh(c(t))$, the self-neglecting property (4.3) of h implies that $h(c''(t))/h(c(t)) \to 1$ and hence, using the fact that $z(1-\rho)\sqrt{b(t)} \geq xh(c(t))$ for all large t,

$$\begin{aligned}
\liminf_{t\to\infty} \frac{\int_{c'(t)}^{\infty} \overline{B}(y)\, dy}{\int_{c(t)}^{\infty} \overline{B}(y)\, dy} &\geq \liminf_{t\to\infty} \frac{\int_{c''(t)}^{\infty} \overline{B}(y)\, dy}{\int_{c''(t)+xh(c(t))}^{\infty} \overline{B}(y)\, dy} \\
&= \liminf_{t\to\infty} \frac{\int_{c''(t)}^{\infty} \overline{B}(y)\, dy}{\int_{c''(t)+xh(c''(t))}^{\infty} \overline{B}(y)\, dy} = \frac{1}{\mathsf{P}\{Z > x\}}.
\end{aligned}$$

The assertion now follows by letting $x \uparrow \infty$. □

4.3. Upper bounds and positive results

Here is our main result for the joint asymptotics in the heavy-tailed case.

Theorem 4.1. *Assume $b(t) = xh(a(t))$. Then (4.8) holds without conditions beyond (4.7), i.e.*

$$\mathsf{P}\{V \geq a(t),\, Q \geq xh(a(t))\} \sim \mathsf{P}\{Z > x(\mu_A - \mu_B)\}\, \mathsf{P}\{V \geq a(t)\}.$$

Corollary 4.1. *Assume $a/b > \mu_B$ and B is regularly varying. Then*

$$\mathsf{P}\{V \geq at,\, Q \geq bt\} \sim \mathsf{P}\{V \geq ct\} \sim \frac{L(t)}{\alpha c^{\alpha-1} t^{\alpha-1}},$$

where $c = a + b(\mu_A - \mu_B)$.

Corollary 4.2. *If $h(x)/x \to 0$, then*

$$\mathsf{P}\{V \geq a(t),\, Q \geq xh(a(t))\} \sim e^{-x(\mu_A - \mu_B)}\, \mathsf{P}\{V \geq a(t)\}.$$

Proof of Theorem 4.1. Assume without loss of generality that $a(t) = t$, $b(t) = xh(t)$, so that $c(t) = t + xh(t)(\mu_A - \mu_B)$. Note that we can choose $\delta > 0$ such that

$$\mathsf{P}\{Q_{\omega(t)-} > a_0(1-\delta)/\mu_B \mid V_{\omega(t)-} \leq a_0\} < \varepsilon.$$

Redefining $\tau_Q(t)$ as $\tau_Q(t) = \inf\{s : Q_s \geq b(t) \mid Q_0 = a_0(1-\delta)/\mu_B\}$ and noting that the expected time after $\tau_V(t; y)$ where $V_s \geq t$ is bounded uniformly in y and t, we thus have the asymptotic upper bound

$$\frac{1}{\mathsf{E}C}\, \mathsf{E}\Big[\int_0^C I\{V_s \geq t,\, Q_s \geq b(t)\}\, ds;\ V_{\omega(t)} \leq a_0\Big]$$
$$\leq \frac{1}{\mu_A} \int_t^\infty \mathsf{E}\,(\tau_V(t;y) - \tau_Q(t))^+\, B(dy) + O(\overline{B}(t)).$$

Further, by (4.4),

$$\mathsf{E}\Big[\int_0^C I\{V_s \geq t,\, Q_s \geq b(t)\}\, ds;\ V_{\omega(t)} > a_0\Big]$$
$$\leq \mathsf{P}\{V_{\omega(t)} > a_0 \mid \omega(t) < C\} \int_t^\infty \mathsf{E}^y \int_0^C I\{V_s \geq t\}\, ds\, B(dy)$$
$$= o_{a_0}(1) \int_t^\infty O(y-t) B(dy) = o_{a_0}(1)\, O(\mathsf{P}\{V \geq t\})$$
$$= o_{a_0}(1)\, O(\mathsf{P}\{V \geq c(t)\}).$$

Thus, combining with the lower bound in Proposition 4.2, it suffices to show for a fixed a_0 that

$$\limsup_{t\to\infty} \frac{(\mu_A)^{-1}\int_t^\infty \mathsf{E}\left(\tau_V(t;y)-\tau_Q(t)\right)^+ B(dy)}{\mathsf{P}\{V\geq c(t)\}} \leq 1\,. \tag{4.13}$$

To this end, write

$$\frac{1}{\mu_A}\int_t^\infty \mathsf{E}(\tau_V(t;y)-\tau_Q(t))^+ B(dy) = I_1+I_2+I_3, \tag{4.14}$$

where I_1, I_2, I_3 denote the contributions from the y-intervals $[t, c_1(t))$, $[c_1(t), c_2(t))$, respectively $[c_2(t), \infty)$, where $c_1(t) = c(t)-\varepsilon h(t)$, $c_2(t) = c(t)+\varepsilon h(t)$.

Consider first $y \in [t, c_1(t))$. Then

$$\begin{aligned}\mathsf{E}(\tau_V(t;y)-\tau_Q(t))^+ &\leq \mathsf{E}\left[\tau_V(t;c_1(t))-\tau_Q(t);\ \tau_V(t;c_1(t))>\tau_Q(t)\right]\\ &\leq \left[\mathsf{E}\,\tau_V(t;c_1(t))^2\cdot \mathsf{P}\{\tau_V(t;c_1(t))>\tau_Q(t)\}\right]^{1/2}.\end{aligned}$$

Here $\mathsf{E}\,\tau_V(t;c_1(t))^2 = O((c_1(t)-t)^2) = O(h(t)^2)$ if $\mathsf{E}\,U^2<\infty$,

$$\frac{\tau_V(t;c_1(t))}{(c_1(t)-t)/(1-\rho)} \xrightarrow{\mathsf{P}} 1, \qquad \frac{\tau_Q(t)}{xh(t)\mu_A} \xrightarrow{\mathsf{P}} 1$$

(cf. [21] for similar discrete time estimates). Since

$$\frac{(c_1(t)-t)/(1-\rho)}{(xh(t)\mu_A)} = 1-\frac{\varepsilon}{x(\mu_A-\mu_B)} < 1,$$

we get $\mathsf{P}\{\tau_V(t;c_1(t))>\tau_Q(t)\}\to 0$ so that

$$I_1 \leq \int_t^\infty O(h(t))o(1)\,B(dy) \;=\; o(h(t)\overline{B}(t)).$$

In the regularly varying case with $\mathsf{E}\,U^2=\infty$, i.e. $1<\alpha\leq 2$, we have

$$\mathsf{E}\,\tau_V(t;c_1(t))^{\alpha'} = O((c_1(t)-t)^{\alpha'}) = O(h(x)^{\alpha'})$$

for all $\alpha'\in(1,\alpha)$ (see again [21]) and we can use the Hölder inequality instead of the Cauchy–Schwarz to obtain the same conclusion.

Similarly (but easier), for some constants k_1, k_2 we get

$$\begin{aligned}I_2 &\leq k_1\int_{c_1(t)}^{c_2(t)} \mathsf{E}\,\tau_V(t;c_2(t))\,B(dy)\\ &\leq k_2h(t)\int_{c_1(t)}^{c_2(t)} B(dy) \;\sim\; k_2h(t)\overline{B}(t)\,\mathsf{P}\{x\mu_A-\varepsilon\leq Z<x\mu_A+\varepsilon\}.\end{aligned}$$

Finally, for $y \in [c_2(t), \infty)$, we have

$$\begin{aligned}\mathsf{E}(\tau_V(t;y) - \tau_Q(t))^+ &= \mathsf{E}(\tau_V(t;y) - \tau_Q(t)) + \mathsf{E}(\tau_Q(t) - \tau_V(t;y))^+ \\ &\leq \frac{y-t}{1-\rho} - xh(t)\mu_A + O(1) \\ &\quad + \left[\mathsf{E}\,\tau_Q(t)^2 \cdot \mathsf{P}\{\tau_Q(t) > \tau_V(t; c_2(t))\}\right]^{1/2},\end{aligned}$$

which by similar estimates as for I_1 (using $\mathsf{E}\,T^2 < \infty$) can be written as

$$\frac{y-t}{1-\rho} - xh(t)\mu_A + O(1) + O(h(t))\,o(1)$$

so that

$$I_3(t) \leq \frac{1}{\mu_A - \mu_B} \int_{c(t)}^{\infty} (y - c(t))\,B(dy) \; + \; O(\overline{B}(t)) \; + \; o(h(t))\overline{B}(c(t)).$$

Hence the limsup in (4.13) can be bounded by

$$0 \; + \; k_3\,\mathsf{P}\{x\mu_A - \varepsilon \leq Z < x\mu_A + \varepsilon\} \; + \; 1 + 0 + 0.$$

Letting $\varepsilon \downarrow 0$ completes the proof. □

A. Appendix

It remains to establish the conditional central limit theorems needed to handle the boundary cases in Theorem 3.2. Assume from now on that a, b are arbitrary nonnegative constants. Let μ_A, μ_B and σ_A^2, σ_B^2 denote the mean and variance of a random variable having distribution A, respectively B, and put

$$m_V = b\mu_B, \qquad \sigma_V = \sqrt{b}\sigma_B, \qquad m_Q = \frac{a}{\Lambda_B'(\gamma)}$$

and

$$\sigma_Q = \frac{\Lambda_B''(\gamma)}{at} \cdot \frac{3\omega - 1}{\omega - 1}, \qquad \text{where } \omega = \frac{\Lambda_B'(\gamma)}{\Lambda_A'(-\gamma)}.$$

It is assumed in this section that $\Lambda_B''(\gamma) < \infty$.

Proposition A.1.

(i) *Let V^{bt} be a random variable with distribution function* $\mathsf{P}\{V \leq \cdot \mid Q \geq bt\}$, *and let ξ be standard* Normal(0, 1). *Then as $t \to \infty$,*

$$\frac{V^{bt} - m_V t}{\sigma_V \sqrt{t}} \xrightarrow{\mathcal{D}} \xi. \tag{A.1}$$

(ii) *Let* Q^{at} *be a random variable with distribution function* $\mathsf{P}\{Q \le \cdot \mid V \ge at\}$, *and let* ξ *be standard* Normal(0, 1). *Then as* $t \to \infty$,

$$\frac{Q^{at} - m_Q t}{\sigma_Q \sqrt{t}} \xrightarrow{\mathcal{D}} \xi. \tag{A.2}$$

Proof of Proposition A.1. (i) First, fix $N \in \mathbb{N}_+$ and consider $\mathsf{P}\{V \ge at \mid Q = \lceil bt + N \rceil\}$, where $V = U_r + \sum_{i=1}^{Q} U_i'$. By the central limit theorem,

$$\sum_{i=1}^{\lceil bt+N \rceil} (U_i' - \mu_B)/\sigma_B\sqrt{bt} \xrightarrow{\mathcal{D}} \xi$$

as $t \to \infty$, where ξ is N(0,1); and it needs to be shown that an analogous result holds for V.

To this end, let $\tilde{U}_{bt}$ be a random variable with distribution $G_{bt}(u) \stackrel{\text{def}}{=} \mathsf{P}\{U_r \le u \mid Q = \lceil bt + N \rceil\}$, independent of $\{U_i'\}$. Then by (2.10), $F_{bt}(\cdot) = 1 - G_{bt}(\cdot)$ is dominated by $(1 - G(\cdot))/C_W$, where G is defined as in (2.8). It follows that $\mathrm{Var}(\tilde{U}_t) \le \Lambda_B''(\gamma)/C_W < \infty$. Consequently, by Theorem 7.2 of [10],

$$\frac{\left(\tilde{U}_{bt} + \sum_{i=1}^{\lceil bt+N \rceil} U_i'\right) - b\mu_b t}{\sigma_B\sqrt{bt}} \xrightarrow{\mathcal{D}} \xi, \tag{A.3}$$

where ξ is N(0,1). Since this result holds for any $N \in \mathbb{N}_+$, it follows that

$$\mathsf{P}\left\{V - m_V t \le z\sigma_V\sqrt{t} \mid Q \in \left[\lceil bt \rceil, \lceil bt + N \rceil\right)\right\} \to \Phi(z), \tag{A.4}$$

where $m_V = b\mu_B$, $\sigma_V = \sqrt{b}\sigma_B$, $N \in \mathbb{N}_+$, and $\Phi(\cdot)$ is the standard N(0,1) distribution function.

The desired result then follows from (A.4), since as $N \to \infty$ we may asymptotically neglect

$$\mathsf{P}\left\{V - \mu_V t \le z\sigma_V\sqrt{t},\ Q \ge \lceil bt + N \rceil\right\} \le \mathsf{P}\left\{Q \ge \lceil bt + N \rceil\right\} \sim z^N C_Q z^{\lceil bt \rceil}$$

compared to

$$\mathsf{P}\left\{V - \mu_V t \le z\sigma_V\sqrt{t},\ Q \in \left[\lceil bt \rceil, \lceil bt + N \rceil\right)\right\} \sim \Phi(z) C_Q z^{\lceil bt \rceil}.$$

(ii) Let N_A, N_B be the renewal processes generated by the interarrival times, respectively service times, and let

$$\overline{\mu}_A = \frac{1}{\Lambda_A'(-\gamma)}, \quad \overline{\sigma}_A = \frac{\Lambda_A''(-\gamma)}{\Lambda_A'(-\gamma)^3}, \quad \overline{\mu}_B = \frac{1}{\Lambda_B'(\gamma)}, \quad \overline{\sigma}_B = \frac{\Lambda_B''(\gamma)}{\Lambda_B'(\gamma)^3}$$

(the means and variances associated with the P_γ-distribution of $N_A(1)$, $N_B(1)$, cf. the proof of Theorem 3.2).

If $V'(s) = \sum_{i=1}^{N_A(s)} U_i - s$ is the workload process modified by removing the reflection at 0, then

$$\begin{aligned} V'(s) - s\Big(\frac{\overline{\mu}_A}{\overline{\mu}_B} - 1\Big) &= \sum_{i=1}^{N_A(s)} U_i - s \cdot \frac{\overline{\mu}_A}{\overline{\mu}_B} = \sum_{i=\overline{\mu}_A s}^{N_A(s)} U_i + \sum_{i=1}^{\overline{\mu}_A s} \Big(U_i - \frac{1}{\overline{\mu}_B}\Big) \\ &= (N_A(s) - \overline{\mu}_A s)\frac{1}{\overline{\mu}_B} + \sum_{i=1}^{\overline{\mu}_A s} \Big(U_i - \frac{1}{\overline{\mu}_B}\Big) + o(\sqrt{s}) \end{aligned}$$

in P_γ-distribution. From this and the standard central limit theorem for renewal processes, we conclude that in P_γ-distribution

$$\frac{1}{\sqrt{s}} \begin{pmatrix} N_A(s) - \overline{\mu}_A s \\ N_B(s) - \overline{\mu}_B s \\ \overline{\mu}_B \left\{V'(s) - s\left(\overline{\mu}_A/\overline{\mu}_B - 1\right)\right\} \end{pmatrix} \to N_3(0, \Sigma), \tag{A.5}$$

where

$$\Sigma = \begin{pmatrix} \overline{\sigma}_A & 0 & \overline{\sigma}_A \\ 0 & \overline{\sigma}_B & \overline{\sigma}_B \left(1 - \frac{\overline{\mu}_A}{\overline{\mu}_B}\right) \\ \overline{\sigma}_A & \overline{\sigma}_B \left(1 - \frac{\overline{\mu}_A}{\overline{\mu}_B}\right) & \overline{\sigma}_A + \overline{\sigma}_B \cdot \frac{\overline{\mu}_A}{\overline{\mu}_B} \end{pmatrix}. \tag{A.6}$$

To check the expression for Σ, let Cova denote the asymptotic covariance in the sense of the central limit theorem and note that

$$\begin{aligned} \mathrm{Cova}\Big(\sum_{i=1}^{\overline{\mu}_A s} U_i\,,\, N_B(s) - \overline{\mu}_B s\Big) &= \overline{\mu}_B\,\mathrm{Cova}\Big(\sum_{i=1}^{\overline{\mu}_A s} U_i\,,\, N_B(s) \cdot \frac{1}{\overline{\mu}_B} - \sum_{i=1}^{N_B(s)} U_i\Big) \\ &= -\overline{\mu}_B\,\mathrm{Cova}\Big(\sum_{i=1}^{\overline{\mu}_A s} U_i\,,\, \sum_{i=1}^{N_B(s)} U_i\Big) \\ &= -\overline{\mu}_B\,\mathrm{Cova}\Big(\sum_{i=1}^{\overline{\mu}_A s} U_i\,,\, \sum_{i=1}^{\overline{\mu}_B s} U_i\Big) \\ &= -\frac{\overline{\sigma}_B}{\overline{\mu}_B^2}(\overline{\mu}_A s - \overline{\mu}_B s), \end{aligned}$$

where we used Anscombe's theorem to replace $N_B(s)$ by $\overline{\mu}_B s$.

Let $\mathfrak{T} = \inf\{s : V'(s) \geq at\}$. Then $\mathfrak{T}/(at/(\overline{\mu}_A/\overline{\mu}_B - 1)) \to 1$ in P_γ-probability, and so by using Anscombe's theorem once more we obtain from (A.6) that

$$\frac{1}{\sqrt{at/(\overline{\mu}_A/\overline{\mu}_B - 1)}} \begin{pmatrix} N_A(\mathfrak{T}) - \overline{\mu}_A \mathfrak{T} \\ N_B(\mathfrak{T}) - \overline{\mu}_B \mathfrak{T} \\ at - \mathfrak{T}(\overline{\mu}_A/\overline{\mu}_B - 1) \end{pmatrix} \to N_3(0, \Sigma)$$

in P_γ-distribution and hence

$$\mathsf{P}_\gamma\left\{N_A(\mathfrak{T}) - N_B(\mathfrak{T}) - m_Q t \leq z\sigma_Q\sqrt{t}\right\} \rightarrow \Phi(z),$$

where $m_Q = a\overline{\mu}_B$ and

$$\sigma_Q^2 = \frac{1}{at/(\overline{\mu}_A/\overline{\mu}_B - 1)} \cdot \overline{\sigma}_B\Big(3\frac{\overline{\mu}_A}{\overline{\mu}_B} - 1\Big).$$

Note that $\sigma_Q > 0$, since in P_γ-distribution $\overline{\mu}_A/\overline{\mu}_B > 1$. Using a mixing argument as in [2] then yields

$$\mathsf{P}_\gamma\left\{N_A(\mathfrak{T}) - N_B(\mathfrak{T}) - m_Q t \leq z\sigma_Q\sqrt{t} \mid \mathfrak{T} < C\right\} \rightarrow \Phi(z),$$

where C denotes the busy cycle. By a regenerative argument as in [2], this finally yields

$$\mathsf{P}^{at}\left\{N_A^*(\tau) - N_B^*(\tau) - m_Q t \leq z\sigma_Q\sqrt{t}\right\} \rightarrow \Phi(z),$$

where N_A^*, N_B^*, and τ are defined as in the proof of Theorem 3.2 (ii). □

References

[1] S. Asmussen(1981) Equilibrium properties of the M/G/1 queue. *Z. Wahrsch. verw. Geb.* **58**, 267–281.

[2] S. Asmussen(1982) Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue. *Adv. Appl. Probab.* **14**, 143–170.

[3] S. Asmussen(1987) *Applied Probability and Queues.* John Wiley and Sons, New York.

[4] S. Asmussen(1992) Phase-type representations in random walk and queueing problems. *Ann. Prob.* **20**, 772–789.

[5] S. Asmussen(1997) Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage times. *Ann. Appl. Probab.* **8**, 354–374.

[6] S. Asmussen and C. Klüppelberg(1996) Large deviations results for subexponential tails, with applications to insurance risk. *Stoch. Process. Appl.* **64**, 103–125.

[7] S. Asmussen, C. Klüppelberg and K. Sigman (1999) Sampling at subexponential times, with queueing applications. *Stoch. Process. Appl.* **79**, 265–286.

[8] S. Aspandiiarov and E.A. Pechersky (1997) A large deviations problem for compound Poisson processes in queueing theory. *Markov Processes Relat. Fields* **3**, 333–366.

[9] R.R. Bahadur and R.R. Rao (1960) On deviations of the sample mean. *Ann. Math. Stat.* **31**, 1015-1027.

[10] P. Billingsley (1968) *Convergence of Probability Measures.* John Wiley and Sons, New York.

[11] A.A. Borovkov (1976) *Stochastic Processes in Queueing Theory.* Springer-Verlag, New York.

[12] G.L. Choudhury and W. Whitt (1996) Long-tail buffer-content distributions in broadband networks. *Performance Evaluation* **30**, 177–190.

[13] J.W. Cohen (1982) *The Single Server Queue.* North-Holland, Amsterdam.

[14] H. Cramér (1938) Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités scientifiques et industrielles* **736**, 5–23.

[15] H.E. Daniels (1954) Saddlepoint approximations in statistics. *Ann. Math. Stat.* **25**, 631–650.

[16] P. Embrechts, C. Klüppelberg and T. Mikosch (1997) *Modelling Extremal Events for Finance and Insurance.* Springer-Verlag, Heidelberg.

[17] P. Embrechts and N. Veraverbeke (1982) Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Math. and Econ.* **1**, 55–72.

[18] W. Feller (1971) *An Introduction to Probability Theory and Its Applications*, Vol. II. John Wiley and Sons, New York.

[19] P.W. Glynn and W. Whitt (1994) Logarithmic asymptotics for steady-state probabilities in a single-server queue. *Adv. Appl. Probab.*, Takács issue, 131–156.

[20] C. Goldie and S.I. Resnick (1988) Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution. *Adv. Appl. Probab.* **20**, 706–718.

[21] A. Gut (1988) *Stopped Random Walks.* Springer-Verlag, New York.

[22] T. Höglund (1979) A unified formulation of the Central Limit Theorem for small and large deviations from the mean. *Z. Wahrsch. verw. Geb.* **49**, 105–117.

[23] J.L. Jensen (1988) Uniform saddlepoint approximations. *Adv. Appl. Probab.* **20**, 622–634.

[24] J.L. Jensen (1991) Uniform saddlepoint approximations and log-concave densities. *J. Royal Stat. Soc. Ser. B* **53**, 157–172.

[25] J.L. Jensen (1995) *Saddlepoint Approximations.* Clarendon Press, Oxford.

[26] T. Mikosch and A.V. Nagaev (1998) Large deviations of heavy-tailed sums with applications in insurance. *Extremes* **1**, 81–110.

[27] V.V. Petrov (1965) On the probabilities of large deviations for sums of independent random variables. *Theory Probab. and Appl.* **10**, 287–298.

[28] R.T. Rockafellar (1970) *Convex Analysis.* Princeton University Press, Princeton.

[29] W. Whitt (1998) The impact of a heavy-tailed service-time distribution upon the M/GI/S waiting-time distribution. Preprint.