On the Distribution of Sojourn Times in Tandem Queues

FLORIN CIUCU AND SIMA MEHRI, University of Warwick, UK

The paper studies the (end-to-end) waiting and sojourn times in tandem queues with general arrivals and light-tailed service times. It is shown that the tails of the corresponding distributions are subject to polynomial-exponential upper bounds, whereby the degrees of the polynomials depend on both the number of bottleneck queues and the 'light-tailedness' of the service times. Closed-form bounds constructed for a two-queue tandem with exponential service times are shown to be numerically sharp, improve upon alternative large-deviations bounds by many orders of magnitude, and recover the exact results in the case of Poisson arrivals.

Additional Key Words and Phrases: Tandem Queues; Stochastic Bounds

ACM Reference Format:
Florin Ciucu and Sima Mehri. 2025. On the Distribution of Sojourn Times in Tandem Queues. 1, 1 (April 2025), 34 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

A landmark result in queueing theory is the product-form structure of the number of jobs N_i at the individual queues in steady-state, i.e.,

$$\mathbb{P}\left(N_1 = n_1, \ldots, N_M = n_M\right) = \prod_i \mathbb{P}\left(N_i = n_i\right) \ .$$

This property was first proved in Jackson networks with Poisson arrivals, exponentially distributed service times, FIFO scheduling, and probabilistic jobs' routing. Several generalizations (e.g., BCMP or Kelly networks) allow for instance for more general service time distributions or other scheduling algorithms (see the surveys by Disney and König [16], Nelson [31], or Balsamo and Marin [5]).

What is challenging in Jackson networks, and even more so in queueing networks with general arrivals, is to characterize *end-to-end metrics* (e.g., the end-to-end waiting times). For instance, even in the $M/M/1 \rightarrow M/M/1$ case, unlike the local (per-queue) sojourn times at the two queues which are independent (Reich [34]), the corresponding local waiting times are not (Burke [9])[1]. The joint distribution of local waiting times is known in the $M/M/1 \rightarrow M/M/1$ case (Karpelevitch and Kreinin [23]), as well as the distribution of the end-to-end waiting time in a more general case when the second queue is served by multiple servers (Krämer [26]). These results were generalized for a large class of networks with Poisson arrivals by Ayhan and Baccelli [3], but only in terms of providing the joint Laplace-Stieltjes transform (LST) which generally requires numerical inversions; explicit expressions for the distribution of local waiting times in such networks were later obtained by Ayhan and Seo [4].

[1]The *local* sojourn time (aka delay) of a job at some queue is the sum of the corresponding *local* waiting time and service time. The (end-to-end) sojourn/waiting times of a job are the sums of the corresponding *local* sojourn/waiting times along its whole network path.

Author's Contact Information: Florin Ciucu and Sima Mehri, University of Warwick, UK.

ACM XXXX-XXXX/2025/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Similar results on sojourn times exist with a notable exception. A very general result is a product-form expression for the joint LST over a so-called quasi overtake-free path in both open and closed product-form networks (e.g., BCMP). This is stated in the survey of Boxma and Daduna [7] (see Theorem 2.4 therein) and generalizes prior results for open or closed networks; the same survey also addresses the issue of numerical computations of inverting the LST and several approximation techniques in both product- and non product-form networks. For a more focused survey on numerical computations of sojourn times' quantiles see Harrison and Knottenbelt [21]. As mentioned earlier, unlike the local waiting times in an $M/M/1 \rightarrow M/M/1$ tandem, the local sojourn times are independent. This exceptional property immediately extends to feedforward Jackson networks, and as an immediate consequence the sojourn time has an Erlang distribution when all service rates are equal. Walrand and Varaiya [38] generalized this result to open Jackson networks subject to non-overtaking paths (for a more comprehensive survey see Disney and König [16]).

The big challenge in queueing networks is to address the case of non-Poisson arrivals. One versatile approach to relax the Poisson assumption characteristic to Jackson, BCMP, or Kelly networks is large deviations. Ganesh [20] studied the sojourn time $\mathcal{S}$ in tandem queues under essentially the same general conditions as in this paper and proved the *logarithmic asymptotics*

$$\lim_{x\to\infty} \frac{\ln \mathbb{P}\left(\mathcal{S} > x\right)}{x} = -\theta$$

for some asymptotic decay rate θ; the particular case of two queues was treated by Foss [19].

An alternative approach is (stochastic) network calculus (see Chang [10] or Jiang and Liu [22]). While intrinsically also based on large deviations, network calculus explicitly computes the pre-factor of the exponential instead of discarding it through the limit $\lim_{x\to\infty} \frac{\ln \mathbb{P}(\mathcal{S}>x)}{x}$. In this way, non-asymptotic bounds on the tails of $\mathcal{S}$ and $\mathcal{W}$ follow in a more or less straightforward manner. Such results can be obtained for broad classes of arrivals, including non-renewal or heavy-tailed processes (e.g., Liebeherr *et al.* [27]), but their proverbial drawback is the poor numerical accuracy.

This paper studies the tails $\mathbb{P}\left(\mathcal{W} > x\right)$ and $\mathbb{P}\left(\mathcal{S} > x\right)$ in tandem queues (networks) with non-Poisson arrivals and finite x. We consider the $GI/G/1 \rightarrow \cdot/G/1 \rightarrow \cdots \rightarrow \cdot/G/1$ tandem with M queues, general arrivals, and light-tailed service times (i.e., having a moment generating function which is finite around zero). At a high level, our approach follows the standard $GI/G/1$ analysis of formulating and analyzing a fixed-point integral/renewal equation. The crucial difference is that instead of using the very distribution $\mathbb{P}(\mathcal{W} \leq x)$ as the integrand, we first decompose $\mathcal{W}$, and also $\mathcal{S}$, into M maxima of (nested) random walks and use their joint distribution as the integrand.

This new approach enables solving a relaxed version of the fixed-point integral equation, by changing the equality into an inequality; the solution lends itself to polynomial-exponential bounds on the tails $\mathbb{P}\left(\mathcal{W} > x\right)$ and $\mathbb{P}\left(\mathcal{S} > x\right)$. Closed-form bounds are obtained in the $GI/M/1 \rightarrow \cdot/M/1$ case with exponential service times, recovering the exact results in the case of Poisson arrivals. Numerical results for deterministic and Erlang arrivals illustrate that the bounds are not only sharp, but also that their polynomial-exponential structure is instrumental in capturing the concave behavior of the tails (on a linear-log scale); in turn, alternative bounds obtained using large-deviations are loose by many orders of magnitude. Additional closed-form bounds are derived for a $GI/H_n/1 \rightarrow \cdot/H_n/1$ tandem with hyperexponential service times which are subject to coefficients of variation greater than one; numerical results show that increasing the coefficients of variation only marginally degrades the bounds' accuracy.

In the following we fully treat the case of a tandem with $M = 2$ queues. Then, in § 3, we derive closed-form bounds for $GI/M/1 \rightarrow \cdot/M/1$ and $GI/H_n/1 \rightarrow \cdot/H_n/1$ tandems, and compare them against simulations and alternative bounds based on large-deviations. In § 4 we present the main results for the general case of a tandem with M queues, which is more challenging from a notational

perspective than the $M = 2$ case, and defer the proofs to Appendix § B; the Appendix also contains the derivations of the large-deviations-based bounds for the $M = 2$ case, remaining proofs, and auxiliary technical results. Conclusions and some open questions are discussed in § 5.

2 A Tandem of Two Queues

We start with the special case of a tandem $GI/G/1 \to \cdot/G/1$ with two queues, each with infinite capacity. There are $n+1$ jobs denoted by $0, 1, \ldots, n$ and traversing the tandem. Job 0 arrives at time 0 to an empty system, whereas the interarrival time between jobs $k-1$ and k is denoted by X_{n+3-k} for $k = 1, \ldots, n$; the arrival time of job k is thus $\sum_{i=1}^{k} X_{n+3-i}$. The service times of job k at the first and second queues are Y_{n+2-k} and Z_{n+1-k}, respectively. All sequences (X_n), (Y_n), and (Z_n) are i.i.d. and mutually independent.

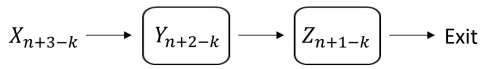


Fig. 1. Inter-arrival and service times for job k in a tandem of two queues

For keeping track of notation, note that job 1 arrives precisely at time X_{n+2}, whereas its service times are Y_{n+1} and Z_n. At the other extreme, the last job n arrives X_3 time units after the previous job $n-1$, and its service times are Y_2 and Z_1.[2] We shall focus on the waiting and sojourn times of job n in the stationary limit $n \to \infty$, by first deriving its exit time for finite n.

The service times are assumed to be *light-tailed*, i.e., they admit finite moment generating functions. For instance, at the first queue, we assume that

$$\theta^+ := \sup\{\theta > 0 : \mathbb{E}[e^{\theta(Y-X)}] < \infty\} \in (0, \infty] .$$

Here, X and Y are shorthand notations for X_1 and Y_1. If $\mathbb{E}[e^{\theta^+(Y-X)}] \geq 1$, which covers most cases of interest, then $\mathbb{E}[e^{\theta(Y-X)}] = 1$ admits a unique solution which dictates the asymptotic decay rates of waiting/sojourn times (Kingman [24]). Otherwise, if $\mathbb{E}[e^{\theta^+(Y-X)}] < 1$, then the asymptotic decay rate would be θ^+; for an example of a service time distribution with this property see Appendix § D.

Before providing a novel representation of the waiting and sojourn times of job n in the stationary limit $n \to \infty$, we quickly highlight what makes it particulary very challenging to deal with end-to-end metrics.

2.1 Sums of Waiting Times vs Product-Form Results

The fundamental difficulty in treating sums of local waiting times can be understood from Burke's proof [9] on their lack of independence; a similar argument was informally made by Harrison and Knottenbelt [21].

Consider the $M/M/1 \to M/M/1$ special case in steady-state and denote by $N_1(t)$ and $N_2(t)$ the number of jobs in the two queues at (*the same*) time t, and by W_1 and W_2 the local waiting times of *the same* arbitrary job. Burke showed that

$$\mathbb{P}(W_2 = 0 \mid W_1 = 0) > \mathbb{P}(W_2 = 0)$$

by explicitly computing the left term; the right term is simply $\mathbb{P}(N_2(t) = 0)$ by using the Arrival Theorem for Jackson networks (see, e.g., Walrand [37], p. 73).

[2]The notation/indexing X_3, Y_2, Z_1 for the main components of job n is deliberate in order to simplify the notation for the exit, waiting, and sojourn times of job n in the tandem. The apparently superfluous index n for a generic job k is also purposely used to simplify the notations/expressions for the same metrics.

Denoting by X and Y the arrival and service times of a job experiencing zero waiting time in the first queue, the previous relation becomes

$$\mathbb{P}\left(N_2(X+Y)=0 \mid N_1(X)=0\right) > \mathbb{P}(N_2(X+Y)=0)\,,$$

i.e., $N_1(X)$ and $N_2(X+Y)$ are not independent, even if X is a stopping time independent of Y.

It should now be apparent that a key difficulty in jointly dealing with W_1 and W_2 is the equivalence with jointly and implicitly dealing with N_1 and N_2 at *different* (random) times where independence lacks. This is unlike the product-form result dealing with N_1 and N_2 at the *same* time, i.e.,

$$\mathbb{P}(N_1(t)=n_1, N_2(t)=n_2) = \mathbb{P}(N_1(t)=n_1)\mathbb{P}(N_2(t)=n_2)\,.$$

2.2 A novel representation for waiting and sojourn times in a tandem

Denote for all jobs $k=0,1,\ldots,n$

$$\tau_k^{(1)} := \max_{n+2-k\le j\le n+2} X_{n+2}+\cdots+X_{j+1}+Y_j+\cdots+Y_{n+2-k}$$

$$\tau_k^{(2)} := \max_{n+1-k\le i<j\le n+2} X_{n+2}+\cdots+X_{j+1}+Y_j+\cdots+Y_{i+1}+Z_i+\cdots+Z_{n+1-k}\,.$$

By convention, all empty sums are set to 0. It can be quickly shown that $\tau_k^{(1)}$ is the exit time of job k from the first queue using induction and Lindley's recursion

$$\tau_{k+1}^{(1)} = \max\{\tau_k^{(1)}, X_{n+2}+\cdots+X_{n+2-k}\}+Y_{n+1-k}\,.$$

Note that the second term in the 'max' and Y_{n+1-k} are the arrival and service times of job $k+1$ at the first queue, respectively. Similarly, $\tau_k^{(2)}$ is the exit time of job k from the second queue (i.e., from the tandem) using induction and Lindley's recursion

$$\tau_{k+1}^{(2)} = \max\{\tau_k^{(2)}, \tau_{k+1}^{(1)}\}+Z_{n-k}\,.$$

Therefore, the exit time of job n from the tandem is

$$\tau_n := \max_{1\le i<j\le n+2} X_{n+2}+\cdots+X_{j+1}+Y_j+\cdots+Y_{i+1}+Z_i+\cdots+Z_1\,. \tag{1}$$

Note that $\tau_n=\tau_n^{(2)}$. The previous sum may be more conveniently read backwards as $Z_1+\cdots+Z_i+Y_{i+1}+\cdots+Y_j+X_{j+1}+\cdots+X_{n+2}$, while also recalling that Z_1 is the service time of job n.

Since $X_3+\cdots+X_{n+2}$ is the arrival time of job n to the tandem, it then immediately follows that the waiting time of job n in the tandem has the same distribution as

$$\mathcal{W}_n := \tau_n - (Z_1+Y_2+X_3+\cdots+X_{n+2})\,.$$

Here, $X_3+\cdots+X_{n+2}$, Y_2, and Z_1 are the arrival time and service times at the two queues of job n, respectively.

Let us now introduce the notations $u_+ := \max\{u,0\}$ and $(U,V) \simeq (Y-X, Z-X)$, where '$\simeq$' stands for equality in distribution[3]. Define now the maxima of random walks

$$T_k^1 := \max_{k\le i<\infty} Y_k+U_{k+1}+\cdots+U_i$$

$$T_k^2 := \max_{k\le i<j<\infty} Z_k+V_{k+1}+\cdots+V_i+U_{i+1}+\cdots+U_j\,,$$

for $k\ge 1$. These are crucial for our main results in that they will feature as the central terms in the expressions of the stationary waiting and sojourn times $\mathcal{W}$ and $\mathcal{S}$ when taking $n\to\infty$. Recall that empty sums are set to 0, e.g., $T_k^1=\max\{Y_k, Y_k+U_{k+1}, Y_k+U_{k+1}+U_{k+2},\ldots\}$ where the first

[3]We make the convention to drop subscripts when clear from the context, e.g., $(U,V)\simeq(Y-X,Z-X)$ stands for $(U_i,V_i)\simeq(Y_i-X_i,Z_i-X_i)$ for all i's.

term corresponds to $i = k$, in which case $U_{k+1} + \cdots + U_i = 0$. Let us mention that T_k^1, except for the leading term Y_k, is essentially the maximum of the random walk featuring in the analysis of the stationary waiting time W_1 at the first queue, i.e., $\mathbb{P}(W_1 > x) = \mathbb{P}\left(\max_{1 \le i < \infty} U_1 + \cdots + U_i > x\right)$; the introduction of Y_k is purposely to simplify our main result from Theorem 1. In turn, T_k^2, except for the leading term Z_k, plays the same role as T_k^1 but for the waiting time across the tandem (see the large-deviations analysis from § A.1).

Note the recurrences:

$$\begin{aligned} T_k^1 &= \max\{Y_k, Y_k + U_{k+1}, Y_k + U_{k+1} + U_{k+2}, \dots\} \\ &= Y_k + \left(T_{k+1}^1 - X_{k+1}\right)_+ \end{aligned} \tag{2}$$

and

$$\begin{aligned} T_k^2 = \max\{&Z_k + U_{k+1}, Z_k + U_{k+1} + U_{k+2}, Z_k + U_{k+1} + U_{k+2} + U_{k+3}, \dots, \\ &Z_k + V_{k+1} + U_{k+2}, Z_k + V_{k+1} + U_{k+2} + U_{k+3}, Z_k + V_{k+1} + U_{k+2} + U_{k+3} + U_{k+4}, \dots, \\ &Z_k + V_{k+1} + V_{k+2} + U_{k+3}, Z_k + V_{k+1} + V_{k+2} + U_{k+3} + U_{k+4}, \dots, \\ &\dots\} \\ = \max\{&T_{k+1}^1, T_{k+1}^2\} + Z_k - X_{k+1}\,, \end{aligned} \tag{3}$$

by using $U_{k+1} = Y_{k+1} - X_{k+1}$, $V_{k+1} = Z_{k+1} - X_{k+1}$, and regrouping terms.

Under the stability condition $\mathbb{E}[X_1] > \max\{\mathbb{E}[Y_1], \mathbb{E}[Z_1]\}$, it is known that the distribution of $\mathcal{W}_n$ converges as $n \to \infty$ to a unique stationary distribution (Loynes [29]) corresponding to that of

$$\mathcal{W} := \max\left\{0, T_3^1 - X_3 + (Z_2 - Y_2)_+, T_3^2 - X_3 + Z_2 - Y_2\right\} \tag{4}$$

described in terms of the previous two maxima of random walks T_3^1 and T_3^2; this follows immediately by taking the limit $n \to \infty$ in the expression of the exit time τ_n. Similarly, the distribution of the sojourn time of job n, i.e., $\mathcal{W}_n + Y_2 + Z_1$ also converges to a unique stationary distribution corresponding to that of

$$\begin{aligned} \mathcal{S} &:= \mathcal{W} + Z_1 + Y_2 = \max\{Y_2, T_3^1 - X_3 + \max\{Z_2, Y_2\}, T_3^2 - X_3 + Z_2\} + Z_1 \\ &= \max\{T_2^1, T_2^2\} + Z_1\,. \end{aligned} \tag{5}$$

2.3 Alternative Interpretation and Representation

In addition to the previous "random walk" interpretation of the terms T_k^1 and T_k^2, let us next provide a more intuitive "queueing" interpretation. The term T_2^1 featuring in the expression of $\mathcal{S}$ from (5) is essentially the sojourn time of job n at the first queue in the stationary limit $n \to \infty$. In turn, the term T_2^2 is the difference between the exit time of job $n-1$ from the second queue and the arrival time of job n at the first queue, again, in the stationary limit $n \to \infty$; recall also that the term Z_1 from (5) is the service time of job n at the second queue. For a visualization of all the possible relevant scenarios of arrivals and departures involving the jobs $n-1$ and n see Fig. 2; note that, in case (a), the value of T_2^2 is negative. More generally, the term T_k^j, for $j = 1, 2$, is the difference between the exit time of job $n + 3 - k - j$ from queue j and the arrival time of job $n + 2 - k$ at queue 1, in the stationary limit $n \to \infty$.

While the queueing interpretation of T_2^1 is clear from the "Lindley-like" recursion of T_k^1 from (2), the corresponding interpretation of T_2^2 in connection to the recurrence of T_k^2 from (3) is not. For this reason, in Fig. 2.(c), we have additionally included a departure scenario of job $n-2$ from queue 2, in which case $\max\{T_3^1, T_3^2\} = T_3^1$ and clearly $T_2^2 = T_3^1 + Z_2 - X_3$. If job $n-2$ departs queue 2 before job $n-1$ arrives at queue 1 then, again, $\max\{T_3^1, T_3^2\} = T_3^1$. Otherwise, if job $n-2$ departs from

queue 2 after job $n-1$ arrives at queue 2, then $\max\{T_3^1, T_3^2\} = T_3^2$ and clearly $T_2^2 = T_3^2 + Z_2 - X_3$. The same argument, involving job $n-2$, can be repeated in the scenarios from (a) and (b).

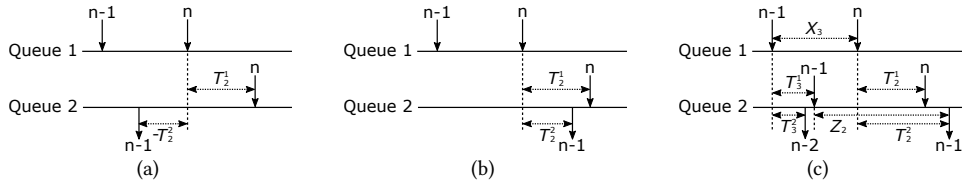


Fig. 2. Queueing interpretation of the representation of $\mathcal{S}$ from (5) (arrivals / departures are depicted by arrows pointing to / leaving the horizontal (time) lines). In (a): job $n-1$ departs from queue 2 before job n arrives at queue 1; in (b): job $n-1$ departs from queue 2 in between the arrival times of job n at the two queues; in (c): job $n-1$ departs from queue 2 after the arrival time of job n at queue 2; in addition from (a) and (b), a departure case of job $n-2$ from queue 2 is included to illustrate the recurrence of T_k^2 from (3) for $k=2$.

Lastly, we mention that $\mathcal{S}$ typically appears in the literature (e.g., Ganesh [20] or Foss [19]) as

$$\max_{0\le i\le j\le\infty} Z_0 + \cdots + Z_i + Y_i + \cdots + Y_j - (X_0 + \cdots + X_{j-1}) \,,$$

by slightly re-indexing (X_n), (Y_n), and (Z_n). The reason for explicitly isolating the maxima of random walks T_2^1 and T_2^2 in our formulation from (5) will become apparent in the main result of the paper, i.e., Theorem 1.(a), which establishes that the joint distribution of T_1^1 and T_1^2 is the unique solution of a fixed-point integral equation. Variations of the main result in terms of integral inequalities (Theorem 1.(b,c)) will further enable, in Corollary 2, the construction of upper and lower bounds on the tails of $\mathcal{W}$ and $\mathcal{S}$.

2.4 Main Result: An Integral Equation

First we denote $u \wedge v := \min\{u, v\}$ and $u \vee v := \max\{u, v\}$, and define the compact set

$$\mathcal{D}_2 := \{(u \wedge v, v) \in \bar{\mathbb{R}}^2 : u \ge 0\} = \{(v, v) : v \le 0\} \cup \{(u, v) : u \le v \le 0\}$$

which is a closed subset of the compact set $\bar{\mathbb{R}}^2$; by notation, $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$

Theorem 1. *Let $U = Y - X$ and $V = Z - X$ be two random variables satisfying $\mathbb{P}(U > 0) > 0$ and $\mathbb{P}(V > 0) > 0$, and where X, Y, and Z are non-negative and independent.*

(a) The integral equation

$$\mathbb{E}\left[\mathbb{1}_{\{u\ge Y\}}\psi\left((u-U)\wedge(v-V), v-V\right)\right] = \psi(u,v) \tag{6}$$

admits a unique solution in the class of bounded functions $\psi : \mathcal{D}_2 \to \mathbb{R}$ having the limit $\psi(\infty,\infty) = \lim_{u,v\to\infty}\psi(u,v) = 1$. This is given by

$$\psi(u,v) := \mathbb{P}(T_1^1 \le u, T_1^2 \le v) \,.$$

(b) Assume that the function $\gamma : \mathcal{D}_2 \to (-\infty, K_\gamma]$, for some finite K_γ, satisfies for all $(u,v) \in \operatorname{supp}(\gamma \vee 0)$

$$\mathbb{E}\left[\mathbb{1}_{\{u\ge Y\}}\gamma\left((u-U)\wedge(v-V), v-V\right)\right] \ge \gamma(u,v) \,. \tag{7}$$

If $\gamma(\infty,\infty) := \limsup_{u,v\to\infty}\gamma(u,v) = 1$ then $\psi \ge \gamma$.

(c) Assume that the function $\eta : \mathcal{D}_2 \to [0,\infty)$ satisfies for all $(u,v) \in \operatorname{supp}(\psi)$

$$\mathbb{E}\left[\mathbb{1}_{\{u\ge Y\}}\eta\left((u-U)\wedge(v-V), v-V\right)\right] \le \eta(u,v) \,. \tag{8}$$

If $\eta(\infty,\infty) := \liminf_{u,v\to\infty}\eta(u,v) = 1$ then $\psi \le \eta$.

Recall the notation convention to drop subcripts when clear from the context, e.g., in $U = Y - X$, Y and X refer to service and interarrival times, as in the model.

As shown shortly in Corollary 2, the problem of finding upper and lower bounds on the tails of $\mathcal{W}$ and $\mathcal{S}$ reduces to the problem of finding the functions γ and η in (b) and (c), respectively.

Proof. For part (a) we first prove that the given ψ satisfies (6); uniqueness will follow after proving (b). We have

$$\begin{aligned}\psi(u,v) &= \mathbb{P}(T_1^1 \le u, T_1^2 \le v)\\ &= \mathbb{P}\left(T_2^1 \le (u - Y_1 + X_2) \wedge (v - Z_1 + X_2), T_2^2 \le v - Z_1 + X_2, Y_1 \le u\right),\\ &= \mathbb{P}\left(T_1^1 \le (u - U) \wedge (v - V), T_1^2 \le v - V, Y \le u\right),\end{aligned}$$

where (U, V) is independent of (T_1^1, T_1^2). So

$$= \mathbb{E}\left[\mathbb{1}_{\{u \ge Y\}} \psi\left((u - U) \wedge (v - V), v - V\right)\right].$$

To prove part (b), i.e., $\psi \ge \gamma$, define first the function

$$f(u,v) := \limsup_{(x,y)\to(u,v)} \left(\gamma(x,y) - \psi(x,y)\right) \ \forall (u,v) \in \mathcal{D}_2\,,$$

which is upper-semi continuous and attains its maximum on the compact set $\mathcal{D}_2$ (see § C.**1**,**2**). Let

$$K := \max_{(u,v)\in\mathcal{D}_2} f(u,v)\,.$$

If $K \le 0$ the proof is complete; assume otherwise that $K > 0$. Define

$$\mathcal{K} := \{(u,v) \in \mathcal{D}_2 : f(u,v) = K\}\,,$$

which is a closed subset of $\mathcal{D}_2$, and

$$a := \min\{u \in \bar{\mathbb{R}} : \exists v \in \bar{\mathbb{R}} : (u,v) \in \mathcal{K}\}$$

$$b := \min\{v \in \bar{\mathbb{R}} : (a,v) \in \mathcal{K}\}\,.$$

Since $f(a,b) = K > 0$, there exists a sequence $(a_n, b_n) \in \operatorname{supp}(\gamma \vee 0)$ such that $(a_n, b_n) \to (a,b)$ as $n \to \infty$ and

$$\begin{aligned}K = f(a,b) &= \lim_{n\to\infty} (\gamma - \psi)(a_n, b_n)\\ &\le \limsup_{n\to\infty} \mathbb{E}\left[\mathbb{1}_{\{a_n \ge Y\}} (\gamma - \psi)\left((a_n - U) \wedge (b_n - V), b_n - V\right)\right]\end{aligned}$$

Since $\gamma - \psi \le K_\gamma < \infty$, we can further use Fatou's lemma

$$\begin{aligned}&\le \mathbb{E}\left[\limsup_{n\to\infty} \mathbb{1}_{\{a_n \ge Y\}} (\gamma - \psi)\left((a_n - U) \wedge (b_n - V), b_n - V\right)\right]\\ &\le \mathbb{E}\left[\limsup_{n\to\infty} \mathbb{1}_{\{a_n \ge Y\}} \left((\gamma - \psi)\left((a_n - U) \wedge (b_n - V), b_n - V\right) \vee 0\right)\right]\\ &\le \mathbb{E}\left[\mathbb{1}_{\{a \ge Y\}} \left(f\left((a - U) \wedge (b - V), b - V\right) \vee 0\right)\right] \le K \cdot \mathbb{P}(a \ge Y)\,,\end{aligned}$$

using the definitions of K and f; note that $((a_n - U) \wedge (b_n - V), b_n - V) \in \mathcal{D}_2$ on $\{a_n \ge Y\}$. It then follows that $\mathbb{P}(a \ge Y) = 1$, such that necessarily

$$f\left((a - U) \wedge (b - V), b - V\right) = K \tag{9}$$

holds almost surely (a.s.) for the inequalities above to hold as equalities.

Next we claim that $(a,b)=(\infty,\infty)$. Assume by contradiction that $a<\infty$. Then (9) and $\mathbb{P}(U>0)>0$ contradict with the choice of a, and hence $a=\infty$. Similarly, assume by contradiction that $b<\infty$. Then (9) and $\mathbb{P}(V>0)>0$ contradict with the choice of b, and hence $b=\infty$ as well. Finally,

$$K=f(\infty,\infty)=\limsup_{u,v\to\infty}(\gamma-\psi)(u,v)=0$$

from the limiting conditions on γ and ψ, which contradicts the assumption that $K>0$. Hence $\psi\geq\gamma$.

We can now prove the uniqueness of ψ solving for (6). Let ψ_1 and ψ_2 be two bounded solutions satisfying $\psi_i(\infty,\infty)=\lim_{u,v\to\infty}\psi_i(u,v)=1$. Applying the second part of the theorem with $\psi=\psi_i$ and $\gamma=\psi_{3-i}$ (note that the proof only needs that ψ satisfies (6), is bounded, and $\psi(\infty,\infty)=\lim_{u,v\to\infty}\psi(u,v)=1$) we obtain that $\psi_i\geq\psi_{3-i}$ for $i=1,2$, and hence $\psi_1=\psi_2$.

Finally, for part (c), we need to show that $\psi\leq\eta$. Define

$$f(u,v):=\limsup_{(x,y)\to(u,v)}\left(\psi(x,y)-\eta(x,y)\right).$$

As in part (b), f is upper-semi continuous and takes its maximum on the compact set $\mathcal{D}_2$, i.e.,

$$K:=\max_{(u,v)\in\mathcal{D}_2}f(u,v).$$

If $K\leq 0$ the proof is complete. Otherwise, assume that $K>0$ and define

$$\mathcal{K}:=\{(u,v)\in\mathcal{D}_2:f(u,v)=K\}$$

which is closed subset of $\bar{\mathbb{R}}^2$, and

$$a:=\min\{u\in\bar{\mathbb{R}}:\exists v\in\bar{\mathbb{R}}:(u,v)\in\mathcal{K}\}$$

$$b:=\min\{v\in\bar{\mathbb{R}}:(a,v)\in\mathcal{K}\}.$$

Since $f(a,b)=K>0$, there exists a sequence $(a_n,b_n)\in\mathrm{supp}\,(\psi)$, such that $(a_n,b_n)\to(a,b)$ and

$$\begin{aligned}K=f(a,b)&=\lim_{n\to\infty}(\psi-\eta)(a_n,b_n)\\&\leq\limsup_{n\to\infty}\mathbb{E}\left[\mathbb{1}_{\{a_n\geq Y\}}(\psi-\eta)\left((a_n-U)\wedge(b_n-V),b_n-V\right)\right]\end{aligned}$$

Since $\psi-\eta\leq 1$, we can apply Fatou's lemma

$$\begin{aligned}&\leq\mathbb{E}\left[\limsup_{n\to\infty}\mathbb{1}_{\{a_n\geq Y\}}(\psi-\eta)\left((a_n-U)\wedge(b_n-V),b_n-V\right)\right]\\&\leq\mathbb{E}\left[\limsup_{n\to\infty}\mathbb{1}_{\{a_n\geq Y\}}\left((\psi-\eta)\left((a_n-U)\wedge(b_n-V),b_n-V\right)\vee 0\right)\right]\\&\leq\mathbb{E}\left[\mathbb{1}_{\{a\geq Y\}}\left(f\left((a-U)\wedge(b-V),b-V\right)\vee 0\right)\right]\leq K\cdot\mathbb{P}(a\geq Y),\end{aligned}$$

using the definitions of K and f. It then follows that $\mathbb{P}(a\geq Y)=1$ such that necessarily

$$f\left((a-U)\wedge(b-V),b-V\right)=K\text{ a.s.}\tag{10}$$

Next we claim that $(a,b)=(\infty,\infty)$. Otherwise, if $a<\infty$ then (10) and $\mathbb{P}(U>0)>0$ contradict with the choice of a and hence $a=\infty$. Similarly, if $b<\infty$ then (10) and $\mathbb{P}(V>0)>0$ contradict with the choice of b and hence $b=\infty$ too. Finally,

$$K=f(\infty,\infty)=\limsup_{u,v\to\infty}(\psi-\eta)(u,v)=0$$

from the limiting conditions on ψ and η contradicting the assumption that $K>0$. Hence $\psi\leq\eta$.

□

We are now able to make the connection between the generic functions γ and η from Parts (b) and (c) of Theorem 1, and bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$.

Corollary 2. (Generic Upper and Lower Bounds) *Consider the functions ψ, γ, and η as in Theorem 1. Then the waiting time of a job $n \to \infty$ satisfies for all $x \geq 0$*

$$\begin{aligned}
&1 - \mathbb{E}\left[\mathbb{1}_{\{x+X_3+Y_2\geq Z_2\}}\eta(x + X_3 - (Z_2 - Y_2)_+, x + X_3 - (Z_2 - Y_2))\right] \leq \mathbb{P}(\mathcal{W} > x) \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X_3+Y_2\geq Z_2\}}\psi(x + X_3 - (Z_2 - Y_2)_+, x + X_3 - (Z_2 - Y_2))\right] \\
&\leq 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X_3+Y_2\geq Z_2\}}\gamma(x + X_3 - (Z_2 - Y_2)_+, x + X_3 - (Z_2 - Y_2))\right] .
\end{aligned}$$

The corresponding sojourn time satisfies

$$\begin{aligned}
1 - \mathbb{E}\left[\mathbb{1}_{\{x\geq Z_1\}}\eta(x - Z_1, x - Z_1)\right] \leq \mathbb{P}(\mathcal{S} > x) &= 1 - \mathbb{E}\left[\psi(x - Z_1, x - Z_1)\right] \\
&\leq 1 - \mathbb{E}\left[\mathbb{1}_{\{x\geq Z_1\}}\gamma(x - Z_1, x - Z_1)\right] .
\end{aligned}$$

This corollary shows that closed-form bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$ can be obtained once constructing explicit functions γ and η satisfying the conditions from parts (b) and (c) of Theorem 1, respectively.

Proof. From $\mathcal{W}$'s representation from (4) it follows for all $x \geq 0$

$$\begin{aligned}
\mathbb{P}(\mathcal{W} > x) &= 1 - \mathbb{P}(\mathcal{W} \leq x) = 1 - \mathbb{P}(\max\left\{0, T_3^1 - X_3 + (Z_2 - Y_2)_+, T_3^2 - X_3 + Z_2 - Y_2\right\} \leq x) \\
&= 1 - \mathbb{P}\left(T_3^1 \leq x + X_3 - (Z_2 - Y_2)_+, T_3^2 \leq x + X_3 - (Z_2 - Y_2)\right) \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X_3+Y_2\geq Z_2\}}\psi(x + X_3 - (Z_2 - Y_2)_+, x + X_3 - (Z_2 - Y_2))\right] ,
\end{aligned}$$

using that $T_3^1 \geq 0$. Since $(x + X_3 - (Z_2 - Y_2)_+, x + X_3 - (Z_2 - Y_2)) \in \mathcal{D}_2$ and $\gamma \leq \psi \leq \eta$ on $\mathcal{D}_2$, the upper and lower bounds on $\mathbb{P}(\mathcal{W} > x)$ follow immediately.

Lastly, we can write for the sojourn time $\mathcal{S}$ from (5) for $x \geq 0$

$$\begin{aligned}
\mathbb{P}(\mathcal{S} > x) &= 1 - \mathbb{P}(\mathcal{S} \leq x) = 1 - \mathbb{E}\left[\psi(x - Z_1, x - Z_1)\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{x\geq Z_1}\psi(x - Z_1, x - Z_1)\right] ,
\end{aligned}$$

using that $\psi(u, v) = 0$ for $u < 0$ (note that $T_1^1 \geq 0$ a.s.). The rest follows from $\gamma \leq \psi \leq \eta$ on $\mathcal{D}_2$. □

2.5 The Integral Equation (6) vs Related Work

At the core of Theorem 1, which enables the generic construction of upper and lower bounds on the tails of $\mathcal{W}$ and $\mathcal{S}$, stands the integral equation (6) which can be rewritten as a two-dimensional renewal equation

$$\psi(u, v) = \int_0^\infty \int_0^u \int_0^\infty \psi\left((u + x - y) \wedge (v + x - z), v + x - z\right) dF(x, y, z) , \tag{11}$$

where F is the joint distribution of (X, Y, Z); the 'renewalness' stems from expressing the underlying maxima of random walks in T_1^1 and T_1^2 in terms of T_2^1 and T_2^2, by extracting the first increments and further using stationarity.

From a conceptual point of view, (11) relates to the standard $GI/G/1$ analysis. Indeed, solving for the distribution $\mathbb{P}(W_1 \leq x)$ of the local waiting time in the first queue from Fig. 1 reduces to solving for

$$\mathbb{P}(W_1 \leq x) = \int_{-\infty}^x \mathbb{P}(W_1 \leq x - u) dG(u) , \tag{12}$$

where G is the distribution of $U := Y - X$, by applying the renewal argument and Lebesgue Convergence Theorem.

The crucial difference between (12) and (11) stands in the integrand itself. While (12) uses the very distribution $\mathbb{P}(W_1 \leq x)$ which is being sought after, the integrand in (11) is based on the joint distribution $\mathbb{P}(T_1^1 \leq u, T_1^2 \leq v)$ stemming from the underlying maxima of random walks in $\mathcal{W}$; as shown earlier in Theorem 1.(a), this joint distribution is also the *unique* solution of (11).

Despite a vast amount of related literature (see, e.g., Cohen [15]) there is no exact and closed-form solution to (12), partly due to outstanding numerical challenges associated to Wiener-Hopf type of integral equations. The equation does however lend itself to a generic and especially simple construction of upper and lower bounds. Indeed, by assuming the existence of a function $\gamma(x)$ satisfying for all $x \geq 0$

$$\int_{-\infty}^{x} \gamma(u) dG(u) \geq \gamma(x) \,, \tag{13}$$

then

$$\mathbb{P}(W_1 > x) \leq 1 - \gamma(x) \,.$$

The proof is immediate using the same renewal argument and induction (see Kingman [25]). One such function is $\gamma(x) = 1 - e^{-\theta x}$, where $\theta > 0$ satisfies $\mathbb{E}[e^{\theta U}] = 1$ (or, more generally, $\theta = \sup\{r > 0 : \mathbb{E}[e^{rU}] \leq 1\}$); the corresponding bound $\mathbb{P}(W_1 > x) \leq e^{-\theta x}$ is known as the Kingman's bound for $GI/G/1$ queues, a.k.a. the Lundberg's inequality in Financial Mathematics (e.g., Mandjes and Boxma [30]), and which can alternatively be obtained using a martingale-based argument (Kingman [24]).

It is instructive to apply the integral equation (6) for the $GI/G/1$ queue by letting $Z = 0$ (i.e., instantaneous service times at the second queue from Fig. 1). Then, according to Theorem 1.(b) and Corollary 2, the derivation of an upper bound reduces to finding a function $\gamma(u, v)$ satisfying

$$\mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} \gamma((u - (Y - X)) \wedge (v + X), v + X)\right] \geq \gamma(u, v) \,, \tag{14}$$

for all $(u, v) \in \text{supp}\,(\gamma \vee 0) \subseteq \mathcal{D}_2 = \{(u \wedge v, v) \in \bar{\mathbb{R}}^2 : u \geq 0\}$ and $\gamma(\infty, \infty) = 1$. As expected, Kingman's bound is recovered by letting

$$\gamma(u, v) = 1 - e^{-\theta u} \,.$$

Similarly, both (13) and (14) recover the tighter bound

$$\mathbb{P}(W_1 > x) \leq \frac{1}{\inf_{u \geq 0} \mathbb{E}\left[e^{\theta(U-u)} \mid U > u\right]} e^{-\theta x}$$

which is exact in the $GI/M/1$ case (Ross [35]); this can be simply done by multiplying the exponential in $\gamma(x)$ and $\gamma(u, v)$ by the prefactor from the bound.

2.6 Existence of Polynomial-Exponential Bounds

Before explicitly constructing functions $\gamma : \mathcal{D}_2 \to \mathbb{R}$ which can lend themselves to sharp (upper) bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$, we first prove their polynomial-exponential structure. For this very purpose, in the proof of the next "Existence" theorem, we are not concerned *yet* with the sharpness of the polynomial's coefficients.

Recall the notation $(U, V) \simeq (Y - X, Z - X)$, where X stands for the interarrival times, and Y and Z stand for the service times at the two nodes.

Theorem 3. (Existence of polynomial-exponential Upper Bounds) *Define*

$$\theta_1 := \sup\{r > 0 : \mathbb{E}[e^{rU}] \leq 1\}, \qquad \theta_2 := \sup\{r > 0 : \max\{\mathbb{E}[e^{rU}], \mathbb{E}[e^{rV}]\} \leq 1\}$$

$$I_U(r) := \begin{cases} 1 & \text{if } \mathbb{E}[e^{rU}] = 1 \\ 0 & \text{otherwise} \end{cases}, \qquad I_V(r) := \begin{cases} 1 & \text{if } \mathbb{E}[e^{rV}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

for random variables U, V. Assume that there exists a constant K such that for all $v \geq 0$

$$\mathbb{E}\left[(V-v)e^{\theta_2(V-v)} \mid V > v\right] \leq K < \infty\,.$$

Then there exist a positive constant $P_1 \geq 0$ and a polynomial $P_2 : \mathbb{R}^2 \to \mathbb{R}$ of degree $I_V(\theta_2)$ and satisfying $P_2(u,v) \geq 0\ \forall v \geq u \geq 0$, such that

$$\gamma(u,v) := \mathbb{1}_{\{v \geq u \geq 0\}}\left[1 - P_1 e^{-\theta_1 u} - P_2(u,v)e^{-\theta_2 v}\right]\ \forall (u,v) \in \mathcal{D}_2$$

satisfies the requirements from part (b) of Theorem 1.

Note that $\gamma(u,u) = 0$ for $u < 0$. The polynomial-exponential structure of γ involves a mix of two exponentials and corresponding polynomials. If the queues are homogeneous, implying that U and V have the same law, then $\theta_1 = \theta_2$ and hence a single exponential. Otherwise, the polynomial-exponential structure is more nuanced and depends on the location of the 'bottleneck'. For instance, if the distributions of U and V are in the same class, but $E[U] > E[V]$, then the first queue is the bottleneck, $\theta_1 = \theta_2$, and $I_V(\theta_2) = 0$, i.e., the degree of P_2 is 0. Otherwise, if $E[U] < E[V]$, then the second node is the bottleneck; under the additional condition $\mathbb{E}[e^{\theta^+ U}] \geq 1$ (recall the description on 'light-tailedness' from the beginning of § 2), then $\theta_1 > \theta_2$ and $I_V(\theta_2) = 1$; thus, the polynomial-exponential structure involves two exponentials whereas the degree of P_2 is 1.

We also mention that the existence of a matching polynomial-exponential structure for η, needed for lower bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$, is still open.

Proof. We proceed in two steps.

Step 1: First we prove that there exist a constant $Q_1 \geq 0$ and a polynomial $Q_2 : \mathbb{R} \to \mathbb{R}$ of degree $I_V(\theta_2)$ with non-negative coefficients such that for all $u \geq 0, v \geq 0$

$$Q_1 e^{-\theta_1 u} \geq \mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} Q_1 e^{\theta_1(U-u)}\right] + \mathbb{P}(Y > u) \tag{15}$$

$$Q_2(v)e^{-\theta_2 v} \geq \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}}\left(Q_1 e^{\theta_1(V-v)} + Q_2(v-V)e^{\theta_2(V-v)}\right)\right] + \mathbb{P}(V > v)\,. \tag{16}$$

Proof: Inequality (15) holds immediately for

$$Q_1 := \left(\inf_{u \geq 0} \mathbb{E}\left[e^{\theta_1(U-u)} \mid Y > u\right]\right)^{-1}$$

by splitting $\mathbb{1}_{\{u \geq Y\}} = 1 - \mathbb{1}_{\{Y > u\}}$ and rewriting $\mathbb{E}[\mathbb{1}_{\{Y>u\}} e^{\theta_1(U-u)}] = E[e^{\theta_1(U-u)} \mid Y > u]\mathbb{P}(Y > u)$.

Let $Q_2(v) := A_0 + A_1 v$. To also prove (16) it is sufficient to show that there exist the non-negative constants A_0, A_1 such that

$$\begin{aligned} &A_0\left\{e^{-\theta_2 v} - \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} e^{\theta_2(V-v)}\right]\right\} + A_1\left\{v e^{-\theta_2 v} - \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}}(v-V)e^{\theta_2(V-v)}\right]\right\} \\ &\geq \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} Q_1 e^{\theta_1(V-v)}\right] + \mathbb{P}(V > v)\,. \end{aligned} \tag{17}$$

For $v \geq 0$, and using that $\theta_1 \geq \theta_2$, the right side can be bounded as

$$\begin{aligned} &\mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} Q_1 e^{\theta_1(V-v)}\right] + \mathbb{P}(V > v) \\ &\leq \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} Q_1 e^{\theta_2(V-v)}\right] + \mathbb{P}(V > v) \\ &= \mathbb{E}\left[Q_1 e^{\theta_2(V-v)}\right] + \mathbb{P}(V > v) - \mathbb{E}\left[Q_1 e^{\theta_2(V-v)} \mid V > v\right]\mathbb{P}(V > v) \\ &\leq Q_1 \mathbb{E}[e^{\theta_2 V}]e^{-\theta_2 v} + \mathbb{P}(V > v)\,. \end{aligned}$$

In turn, on the left side of (17), we bound the coefficient of A_0 in the opposite direction

$$e^{-\theta_2 v} - \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} e^{\theta_2 (V-v)}\right] \geq \left(1 - \mathbb{E}\left[e^{\theta_2 V}\right]\right) e^{-\theta_2 v} + \mathbb{P}(V > v)$$

and similarly the coefficient of A_1

$$\begin{aligned} & v e^{-\theta_2 v} - \mathbb{E}\left[\mathbb{1}_{\{v \geq V\}} (v - V) e^{\theta_2 (V-v)}\right] \\ & \quad = v e^{-\theta_2 v} - \mathbb{E}\left[(v - V) e^{\theta_2 (V-v)}\right] + \mathbb{E}\left[\mathbb{1}_{\{V > v\}} (v - V) e^{\theta_2 (V-v)}\right] \\ & \quad \geq \left(1 - \mathbb{E}\left[e^{\theta_2 V}\right]\right) v e^{-\theta_2 v} + \mathbb{E}\left[V e^{\theta_2 V}\right] e^{-\theta_2 v} - K\mathbb{P}(V > v) \ . \end{aligned}$$

Therefore, it is sufficient to determine the coefficients A_0 and A_1 satisfying the tighter version of (16)

$$\begin{aligned} & A_0 \left\{\left(1 - \mathbb{E}\left[e^{\theta_2 V}\right]\right) e^{-\theta_2 v} + \mathbb{P}(V > v)\right\} \\ & \qquad + A_1 \left\{\left(1 - \mathbb{E}\left[e^{\theta_2 V}\right]\right) v e^{-\theta_2 v} + \mathbb{E}\left[V e^{\theta_2 V}\right] e^{-\theta_2 v} - K\mathbb{P}(V > v)\right\} \\ & \quad \geq Q_1 \mathbb{E}[e^{\theta_2 V}] e^{-\theta_2 v} + \mathbb{P}(V > v) \ . \end{aligned} \tag{18}$$

There are two cases. If $\mathbb{E}\left[e^{\theta_2 V}\right] < 1$ then

$$A_1 := 0 \quad \text{and} \quad A_0 := \max\left\{\frac{Q_1 \mathbb{E}[e^{\theta_2 V}]}{1 - \mathbb{E}[e^{\theta_2 V}]}, 1\right\}$$

satisfy (17) and $Q_2(v) = A_0$ has degree $I_V(\theta_2) = 0$.

In the other case, if $\mathbb{E}\left[e^{\theta_2 V}\right] = 1$, then

$$A_1 := \frac{Q_1}{\mathbb{E}[V e^{\theta_2 V}]} \quad \text{and} \quad A_0 := 1 + A_1 K$$

also satisfy (17); moreover $Q_2(v) = A_0 + A_1 v$ has degree $I_V(\theta) = 1$. Note that $A_1 > 0$ because the function $f(\theta) := \mathbb{E}[e^{\theta V}]$ is convex, $f(0) = f(\theta_2) = 1$, and hence $f'(\theta_2) = \mathbb{E}[V e^{\theta_2 V}] > 0$.

Step 2: Let $P_1 := Q_1$ and $P_2(u, v) := Q_2(v)$ from Step 1. Then γ satisfies (7).

Proof: Note first that the 'marginal' function of γ for the first queue

$$\gamma_1(u) := (1 - P_1 e^{-\theta_1 u}) \mathbb{1}_{\{u \geq 0\}}$$

satisfies for all $u \geq 0$

$$\mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} \gamma_1(u - U)\right] = \mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} \left(1 - Q_1 e^{\theta_1 (U-u)}\right)\right] \geq 1 - Q_1 e^{-\theta_1 u} = \gamma_1(u) \tag{19}$$

using the definition of Q_1. Note also that

$$\text{supp}\,(\gamma \vee 0) = \{(u, v) \in \mathcal{D}_2 : 1 > Q_1 e^{-\theta_1 u} + Q_2(v) e^{-\theta_2 v}\} \ .$$

We can now prove condition (7) for γ. For all $(u, v) \in \mathcal{D}_2$

$$\begin{aligned} & \mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} \gamma\left((u - U) \wedge (v - V), v - V\right)\right] \\ & = \mathbb{E}\left[\mathbb{1}_{\{u \geq Y, v \geq V\}} \left[1 - Q_1 e^{-\theta_1 ((v-V) \wedge (u-U))} - Q_2(v - V) e^{-\theta_2 (v-V)}\right]\right] \end{aligned}$$

Since $Q_1 \geq 0$ and using $\max\{x, y\} \leq x + y$ for $x, y \geq 0$ we can continue with

$$\begin{aligned}
&\geq \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left(1 - Q_1 e^{\theta_1(U-u)} - \left(Q_1 e^{\theta_1(V-v)} + Q_2(v-V)e^{\theta_2(V-v)}\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v<V\}}\left(-1 + Q_1 e^{\theta_1(U-u)}\right)\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y\}}\left(1 - Q_1 e^{\theta_1(U-u)}\right)\right] \\
&\quad - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left(Q_1 e^{\theta_1(V-v)} + Q_2(v-V)e^{\theta_2(V-v)}\right)\right] .
\end{aligned}$$

Using $Q_1 \geq 0$ for the first expectation and (19) for the second, we can further continue

$$\begin{aligned}
&\geq -\mathbb{P}(u \geq Y, v < V) + \left(1 - Q_1 e^{-\theta_1 u}\right) - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left(Q_1 e^{\theta_1(V-v)} + Q_2(v-V)e^{\theta_2(V-v)}\right)\right] \\
&\geq -\mathbb{P}(v < V) - \mathbb{E}\left[\mathbb{1}_{\{v\geq V\}}\left(Q_1 e^{\theta_1(V-v)} + Q_2(v-V)e^{\theta_2(V-v)}\right)\right] + \left(1 - Q_1 e^{-\theta_1 u}\right) \\
&\geq 1 - Q_1 e^{-\theta_1 u} - Q_2(v)e^{-\theta_2 v} = \gamma(u, v) .
\end{aligned}$$

In the second inequality we discarded the event $\{u \geq Y\}$ and in the last we used (16) from Step 1.

The remaining conditions from Theorem 1.(b), i.e., γ is bounded and $\gamma(\infty, \infty) = 1$ hold trivially. □

3 Closed-Form Polynomial-Exponential Bounds + Numerics

Here we provide closed-form bounds for two types of tandem queues with general inter-arrival times distributions and equal service rates: $GI/M/1 \to \cdot/M/1$, with exponential service times, and $GI/H_n/1 \to \cdot/H_n/1$ with hyperexponential service times. In the former case we also provide alternative bounds obtained using large-deviations and investigate the numerical accuracy. In the latter case we further investigate the impact of the service times' coefficient of variation on the bounds' accuracy.

Denote $U := Y - X$ and $V := Z - X$, where X, Y, and Z are independent, and assume that $\mathbb{E}[e^{\theta U}] = \mathbb{E}[e^{\theta V}] = 1$ for some $\theta > 0$; existence follows from the stability condition $\rho < 1$, where ρ is the utilization factor. Let for $R \in \{Y, Z\}$, $r \geq 0$, and $j \geq 0$

$$K_j^R(r) := \mathbb{E}[(R-r)^j e^{\theta(R-r)} \mid R > r] .$$

In order to invoke Corollary 2, which provides the generic structure of the bounds in terms of the function $\gamma(\cdot, \cdot)$, we first need a closed-form construction of $\gamma(\cdot, \cdot)$. According to Theorem 1.(b) and the constructed polynomial-exponential structure of the functions $\gamma(u, v)$ from the "Existence" Theorem 3, we need to find the constants A, B, C, D such that, by defining $\gamma : \mathcal{D}_2 \to \mathbb{R}$

$$\gamma(u, v) := \mathbb{1}_{\{v\geq u\geq 0\}}\left[1 - Ae^{-\theta u} - (B + Cu + Dv)e^{-\theta v}\right] , \tag{20}$$

the following inequalities hold: $A \geq 0$, $B + Cu + Dv \geq 0$ $\forall v \geq u \geq 0$, and (7) from Theorem 1, i.e.,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left[1 - Ae^{-\theta[(u-U)\wedge(v-V)]} - \left(B + C\left[(u-U)\wedge(v-V)\right] + D(v-V)\right)e^{-\theta(v-V)}\right]\right] \\
&\quad \geq 1 - Ae^{-\theta u} - (B + Cu + Dv)e^{-\theta v} ,
\end{aligned} \tag{21}$$

for all $(u, v) \in \operatorname{supp}(\gamma \vee 0)$.

To find the parameters A, B, C, D for (21) to hold, we provide next a sufficient set of inequalities which are much easier to handle separately.

Lemma 4. *The following five inequalities are sufficient for* (21) *to hold:*

$$AK_0^Y(u)\mathbb{E}[e^{-\theta X}] \geq 1$$

$$CK_1^Z(v-u+Y) + A(1-K_0^Z(v-u+Y)) \geq 0$$

$$C\mathbb{E}\left[Ue^{\theta V}\right] + D\mathbb{E}\left[Ve^{\theta V}\right] \geq 0$$

$$B + C\mathbb{E}\left[(u-U)e^{\theta V} \mid Y > u\right] + D\mathbb{E}\left[(v-V)e^{\theta V}\right] \geq 0$$

$$(A+B)K_0^Z(v+X) - (C+D)K_1^Z(v+X) \geq 1\,.$$

If all inequalities hold as equalities then $\gamma = \psi$.

The five inequalities are obtained by expanding (21) and grouping terms from its left and right sides; for convenience, we depict each group in a different colour; for the proof see Appendix § A. As already hinted, the above grouping is likely sub-optimal, i.e., different groupings or other more direct approaches not involving simplifying groups may lend themselves to better A, B, C, D in the sense of obtaining tighter bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$. Also, different (polynomial-exponential) expressions of $\gamma(u,v)$ than the one from (20), which was inspired by the existence result from Theorem 3, may be more efficient.

3.1 $GI/M/1 \to \cdot/M/1$

For the $GI/M/1 \to \cdot/M/1$ model, i.e., $Y, Z \simeq Exp(\mu)$ and $\mathbb{E}[X] = \frac{1}{\mu\rho}$, let $\theta > 0$ such that $\mathbb{E}[e^{\theta(Y-X)}] = 1$, or, equivalently, $\mathbb{E}[e^{-\theta X}] = (\mu-\theta)/\mu$. Observing first that

$$K_j^Y(x) = \frac{\mu \times j!}{(\mu-\theta)^{j+1}}\,,$$

and denoting $\alpha := \mathbb{E}\left[Xe^{-\theta X}\right]$, three of the four parameters of $\gamma(u,v)$ from (20) follow immediately[4] from the inequalities 1, 2, 3 from Lemma 4:

$$A = 1, \qquad C = \frac{\theta(\mu-\theta)}{\mu}, \qquad D = C \times \left(\frac{\alpha\frac{\mu}{\mu-\theta} - \frac{1}{\mu}}{\frac{1}{\mu-\theta} - \alpha\frac{\mu}{\mu-\theta}} \vee 0\right)\,.$$

The remaining parameter B can be obtained by treating separately the remaining inequalities 4 and 5 from Lemma 4: from the fourth we define

$$B_1 := C\left(\frac{1}{\mu} - \alpha\frac{\mu}{\mu-\theta}\right) + D\left(\frac{1}{\mu-\theta} - \alpha\frac{\mu}{\mu-\theta}\right) = C\left(\frac{1}{\mu} - \alpha\frac{\mu}{\mu-\theta}\right) \vee 0$$

and from the fifth we define

$$B_2 := \frac{\mu-\theta}{\mu} + \frac{C+D}{\mu-\theta} - A = \frac{D}{\mu-\theta}\,.$$

We can finally define $B := B_1 \vee B_2 = B_1 \mathbb{1}_{\{D=0\}} + B_2 \mathbb{1}_{\{D>0\}}$.

We can now apply Corollary 2 to obtain the bounds on the tail of $\mathcal{W}$, i.e.,

$$\mathbb{P}(\mathcal{W} > x) \leq 1 - \mathbb{E}\Big\{\mathbb{1}_{\{x+X+Y\geq Z\}}\Big[1 - Ae^{-\theta(x+X-(Z-Y)_+)} - (B + C(x+X-(Z-Y)_+) + D(x+X-(Z-Y)))e^{-\theta(x+X-(Z-Y))}\Big]\Big\}\,.$$

[4]For more complete derivations see the treatment of the more general $GI/H_n/1 \to \cdot/H_n/1$ case.

The bound becomes

$$\mathbb{P}(\mathcal{W} > x) \le \mathbb{E}\Bigg[\frac{e^{-\mu(x+X)}}{2} + A\left(\frac{(2\mu-\theta)e^{-\theta x}}{2\mu} - \frac{\mu e^{-\mu(x+X)}}{2(\mu-\theta)}\right) + B\left(\frac{\mu e^{-\theta x}}{\mu+\theta} - \frac{\mu e^{-\mu(x+X)}}{2(\mu-\theta)}\right) + C\left(\frac{\mu^2(x+X)e^{-\theta(x+X)}}{\mu^2-\theta^2} + \frac{\mu e^{-\mu(x+X)}}{2(\mu-\theta)^2} - \frac{e^{-\theta x}}{2(\mu-\theta)}\right) + D\left(\frac{\mu^2(x+X)e^{-\theta(x+X)}}{\mu^2-\theta^2} - \frac{2\mu\theta e^{-\theta x}}{(\mu^2-\theta^2)(\mu+\theta)} + \frac{\mu e^{-\mu(x+X)}}{2(\mu-\theta)^2}\right)\Bigg] . \tag{22}$$

In turn,

$$\begin{aligned}\mathbb{P}(\mathcal{S} > x) &\le 1 - \mathbb{E}\left[\mathbb{1}_{\{x\ge Z\}}\left(1-(A+B)+(C+D)(x-Z)\right)e^{-\theta(x-Z)}\right] \\ &= e^{-\mu x} + \frac{\mu}{\mu-\theta}(A+B)\left(e^{-\theta x}-e^{-\mu x}\right) + \frac{\mu}{(\mu-\theta)^2}(C+D)\left(((\mu-\theta)x-1)e^{-\theta x}+e^{-\mu x}\right) .\end{aligned}$$

Denoting $\beta := \mathbb{E}\left[e^{-\mu X}\right]$ and collecting terms we obtain the polynomial-exponential bounds:

$$\mathbb{P}(\mathcal{W} > x) \le \begin{cases} \left(1 - \frac{2\theta^2}{\mu(\mu+\theta)} + \frac{\theta(\mu-\theta)}{\mu+\theta}x\right)e^{-\theta x} + \beta\left(\frac{\theta\mu\alpha}{2(\mu-\theta)} - \frac{\theta}{2\mu}\right)e^{-\mu x} \\ \left(1 - \frac{2\theta}{\mu} + \frac{2\theta^2(2-\alpha\mu)}{(\mu+\theta)^2(1-\alpha\mu)} + \frac{\theta^2(\mu-\theta)}{\mu(\mu+\theta)(1-\alpha\mu)}x\right)e^{-\theta x} \end{cases}$$

$$\mathbb{P}(\mathcal{S} > x) \le \begin{cases} (1+\theta x)e^{-\theta x} + \theta\left(\frac{1}{\mu} - \frac{\alpha\mu}{\mu-\theta}\right)\left(e^{-\theta x} - e^{-\mu x}\right) \\ (1 + \frac{\theta^2}{\mu(1-\alpha\mu)}x)e^{-\theta x} , \end{cases} \tag{23}$$

where the top branches hold when $\alpha \le \frac{\mu-\theta}{\mu^2}$, which is equivalent to $D = 0$, and the bottom otherwise.

We observe that the bounds on $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$ for $\alpha \le \frac{\mu-\theta}{\mu^2}$ have a *mixed* polynomial-exponential structure as they involve two exponentials (in θ and μ) along with corresponding polynomials.

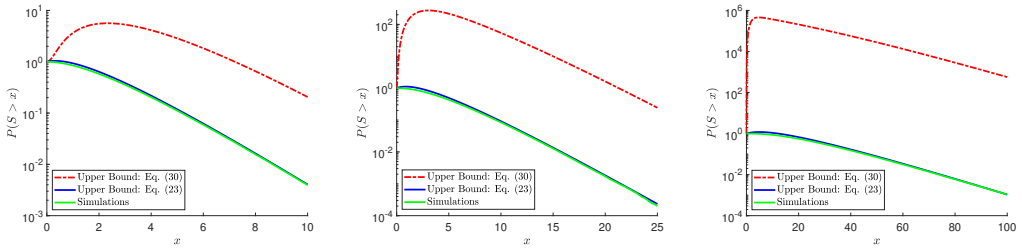
In the $M/M/1 \to \cdot/M/1$ special case, we have $\alpha = \frac{\mu-\theta}{\mu^2}$, $B = D = 0$, and the five inequalities in Lemma 4 become equalities. So $\gamma = \psi$ and

$$\mathbb{P}(\mathcal{W} > x) = \left(1 - \frac{2\theta^2}{\mu(\mu+\theta)} + \frac{x(\mu-\theta)\theta}{\mu+\theta}\right)e^{-\theta x} , \qquad \mathbb{P}(\mathcal{S} > x) = (1+\theta x)e^{-\theta x} .$$

The former result first appeared in [26].

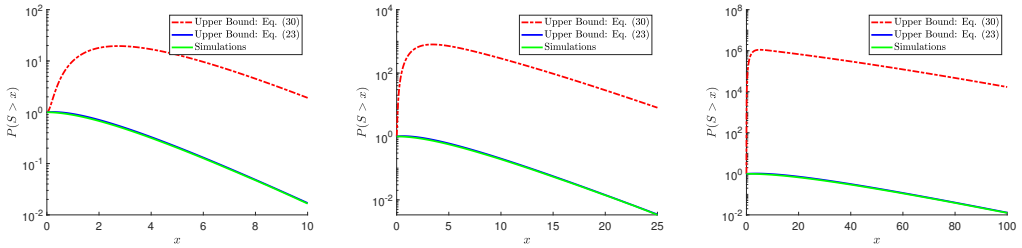
Next we numerically illustrate the sojourn-time bound from (23) against the corresponding one from (30) based on large-deviations (for the derivations see Appendix § A.1). We also include simulation results obtained from 10^6 runs and sample paths of 10^4 points (i.e., the number of jobs); the running time for this setting is 10^{14} due to the nested 'max' in the expression of $\mathcal{S}$ from (5), requiring itself 10^8 time.

Fig. 3 displays the results for the $D/M/1 \to \cdot/M/1$ tandem, for several values of the utilization factor ρ; the service rate is $\mu = 1$ and the inter-arrival times are $X = \frac{1}{\rho\mu}$. The (extremely) poor accuracy of the large-deviations-based bounds is particularly pronounced at high utilizations, despite the D/M setting; this is due not only to the nested 'max', and consequently the nested infinite application of the Union Bound, but also to the relatively small space in running the optimization '$\inf_{\{0<\theta<\mu:\beta<1\}}$', as opposed to scenarios with lower ρ. In turn, in addition to the overall accuracy of the bounds from (23), we highlight their ability to capture the initial concave 'bend' which is clearly visible in (c) around small x; this feature is precisely due to the polynomial-exponential structure of the bounds.



(a) $\rho = 0.5$ (b) $\rho = 0.75$ (c) $\rho = 0.95$

Fig. 3. The sojourn time CCDF $\mathbb{P}(\mathcal{S} > x)$ for the $D/M/1 \rightarrow \cdot/M/1$ tandem: Large-Deviations Upper Bounds from (30) vs. Polynomial-Exponential Upper Bounds from (23) vs. Simulations



(a) $\rho = 0.5$ (b) $\rho = 0.75$ (c) $\rho = 0.95$

Fig. 4. The sojourn time CCDF $\mathbb{P}(\mathcal{S} > x)$ for the $Erlang(2)/M/1 \rightarrow \cdot/M/1$ tandem: Large-Deviations Upper Bounds from (30) vs. Polynomial-Exponential Upper Bounds from (23) vs. Simulations

Fig. 4 illustrates results for the $Erlang(2)/M/1 \rightarrow \cdot/M/1$ tandem, as a special case with Gamma distributed input, and reveals similar observations as in the D/M case.

3.2 $GI/H_n/1 \rightarrow \cdot/H_n/1$

We now address the case of hyperexponential service times, i.e., $Y, Z \simeq \sum_{i=1}^{n} p_i Exp(\mu_i)$ where $p_1, \ldots, p_n \in [0,1]$, $\sum_i p_i = 1$, and, without loss of generality, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. The coefficient of variation is

$$CV_Y = \left(\sqrt{\sum_i \frac{2p_i}{\mu_i^2} - \mathbb{E}[Y]^2} \right) \mathbb{E}[Y] \, ,$$

where $\mathbb{E}[Y] = \sum_i \frac{p_i}{\mu_i}$; it is known that $CV_Y \geq 1$.

Let $\theta > 0$ such that $\mathbb{E}[e^{\theta(Y-X)}] = 1$, or, equivalently, $\mathbb{E}[e^{-\theta X}] = \left(\sum_i \frac{p_i \mu_i}{\mu_i - \theta} \right)^{-1}$; note that $\theta \leq \min_i \mu_i$. For brevity, we only derive bounds on $\mathbb{P}(S > x)$.

First, using elementary integration, we obtain

$$K_j^Y(x) = \frac{\sum_i \frac{p_i \mu_i e^{-\mu_i x} \times j!}{(\mu_i - \theta)^{j+1}}}{\sum_i p_i e^{-\mu_i x}} \; \forall j \geq 0 \forall x \geq 0 \, .$$

The rest follows according to the same procedure described in the $GI/M/1 \rightarrow \cdot/M/1$ case, i.e., match the parameters A, B, C, and D to satisfy the five inequalities from Lemma 4. From the first

we define

$$A := \frac{\mathbb{E}\left[e^{\theta Y}\right]}{\inf_{x\geq 0} K_0^Y(x)} = \frac{\sum_i \frac{p_i\mu_i}{\mu_i-\theta}}{\inf_{x\geq 0} \sum_i \frac{p_i\mu_i}{\mu_i-\theta} e^{-\mu_i x} \Big/ \sum_i p_i e^{-\mu_i x}} = 1\,.$$

Next, from the second inequality we define

$$C := \sup_{x\geq 0} \frac{K_0^Y(x)-1}{K_1^Y(x)} = \sup_{x\geq 0} \frac{\frac{\sum_i \frac{p_i\mu_i}{\mu_i-\theta} e^{-\mu_i x}}{\sum_i p_i e^{-\mu_i x}} - 1}{\frac{\sum_i \frac{p_i\mu_i}{(\mu_i-\theta)^2} e^{-\mu_i x}}{\sum_i p_i e^{-\mu_i x}}} = \theta \frac{\sum_i \frac{p_i}{\mu_i-\theta}}{\sum_i \frac{p_i\mu_i}{(\mu_i-\theta)^2}}\,.$$

Furthermore, from the third inequality we define

$$D := -C\frac{\mathbb{E}\left[(Y-X)e^{\theta(Z-X)}\right]}{\mathbb{E}\left[(Z-X)e^{\theta(Z-X)}\right]} \vee 0 = C \times \left(\frac{\alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta} - \sum_i \frac{p_i}{\mu_i}}{\frac{\sum_i \frac{p_i\mu_i}{(\mu_i-\theta)^2}}{\sum_i \frac{p_i\mu_i}{\mu_i-\theta}} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta}} \vee 0 \right),$$

where $\alpha := \mathbb{E}\left[Xe^{-\theta X}\right]$. Note the additional requirement that $D \geq 0$ to satisfy $B + Cu + Dv \geq 0\ \forall v \geq u \geq 0$, as mentioned just before Lemma 4.

The remaining parameter B can be obtained by treating separately the inequalities 4 and 5. From the former we define

$$B_1 := C \sup_{u\geq 0} \mathbb{E}\left[(Y-X-u)e^{\theta(Z-X)} \mid Y > u\right] + D \sup_{v\geq 0} \mathbb{E}\left[(Z-X-v)e^{\theta(Z-X)}\right]\,.$$

Using $\mathbb{E}\left[e^{\theta(Z-X)}\right] = 1$ and the independence of X, Y, and Z, the term inside the first supremum can be written as

$$\begin{aligned}\frac{\mathbb{E}\left[Ye^{\theta(Z-X)}\mathbb{1}_{Y>u}\right]}{\mathbb{P}(Y>u)} - \frac{\mathbb{E}\left[Xe^{\theta(Z-X)}\mathbb{1}_{Y>u}\right]}{\mathbb{P}(Y>u)} - u &= \frac{u\sum_i p_i e^{-\mu_i u} + \sum_i \frac{p_i}{\mu_i} e^{-\mu_i u}}{\sum_i p_i e^{-\mu_i u}} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta} - u \\ &= \frac{\sum_i \frac{p_i}{\mu_i} e^{-\mu_i u}}{\sum_i p_i e^{-\mu_i u}} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta}\,.\end{aligned}$$

Using $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ the value of the supremum is $\frac{1}{\mu_1} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta}$; note that the underlying function in u is nondecreasing. In turn, the second supremum is obviously attained at $v = 0$ and takes the value

$$\frac{\sum_i p_i\mu_i/(\mu_i-\theta)^2}{\sum_i p_i\mu_i/(\mu_i-\theta)} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta}\,.$$

We then obtain

$$B_1 = C\left(\frac{1}{\mu_1} - \alpha \sum_i \frac{p_i\mu_i}{\mu_i-\theta} \wedge \sum_i \frac{p_i}{\mu_i}\right).$$

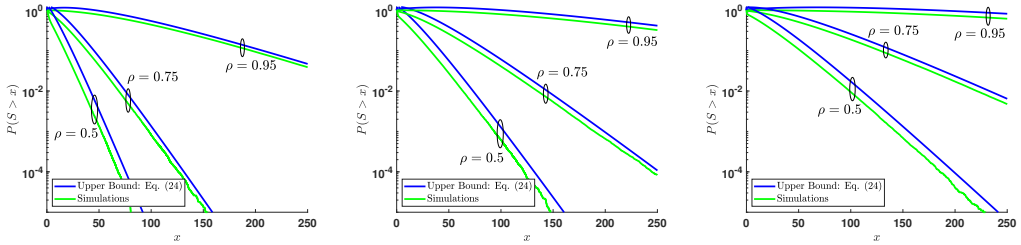
Finally, from inequality 5 we define

$$B_2 := \sup_{u\geq 0} \frac{1 + (C+D)\sum_i \frac{p_i\mu_i}{(\mu_i-\theta)^2} e^{-\mu_i u} \Big/ \sum_i p_i e^{-\mu_i u}}{\sum_i \frac{p_i\mu_i}{\mu_i-\theta} e^{-\mu_i u} \Big/ \sum_i p_i e^{-\mu_i u}} - 1 = \frac{C+D}{\mu_1-\theta} - \frac{\theta}{\mu_1}\,,$$

using again that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$; note also that the function inside the 'sup' is nondecreasing in u such that the 'sup' is attained by taking $u \to \infty$.

Defining now $B := B_1 \vee B_2$, which satisfies the additional requirement that $B \geq 0$ because $B_1 \geq C\left(\frac{1}{\mu_1} - \sum_i \frac{p_i}{\mu_i}\right) \geq 0$, we obtain from Corollary 2

$$\mathbb{P}(\mathcal{S} > x) \leq 1 - \mathbb{E}\left[\mathbb{1}_{\{x \geq Z\}}\left(1 - (A+B) + (C+D)(x-Z)\right)e^{-\theta(x-Z)}\right]$$
$$= \sum_{i=1}^{n} p_i \left\{ e^{-\mu_i x} + \frac{\mu_i}{\mu_i - \theta}(A+B)\left(e^{-\theta x} - e^{-\mu_i x}\right) + \frac{\mu_i}{(\mu_i-\theta)^2}(C+D)\left(((\mu_i - \theta)x - 1)e^{-\theta x} + e^{-\mu_i x}\right)\right\}. \tag{24}$$



(a) $CV_Y = 2$ (b) $CV_Y = 3$ (c) $CV_Y = 4$

Fig. 5. The sojourn time CCDF $\mathbb{P}(\mathcal{S} > x)$ for the $Erlang(2)/H_2/1 \rightarrow \cdot/H_2/1$ tandem: Polynomial-Exponential Upper Bounds from (24) vs. Simulations

Fig. 5 shows numerical results for several values of the coefficient of variation CV_Y and the utilization factor ρ in the case of the $Erlang(2)/H_2/1 \rightarrow \cdot/H_2/1$ tandem. The service rates are set to $\mathbb{E}[Y] = \mathbb{E}[Z] = 1$, $p_2 = 0.9$, and $\mu_2 = 1.69$, $\mu_2 = 3$, and $\mu_2 = 11.5$, corresponding to $CV_Y = 2$, $CV_Y = 3$, and $CV_Y = 4$, respectively; note that $\mu_1 = \frac{p_1\mu_2}{\mu_2 - p_2}$. While the bounds' accuracy slightly degrades in comparison to the $Erlang(2)/M/1 \rightarrow \cdot/M/1$ tandem, by increasing CV_Y from 1 to 2, we note that further increasing CV_Y does not (visually) decrease the bounds' accuracy. We mention that similar observations hold a $D/H_2/1 \rightarrow \cdot/H_2/1$ tandem (the numerics are omitted for brevity).

4 Generalization: Tandem of M Queues

Here we generalize the previous results to a tandem $GI/G/1 \rightarrow \cdot/G/1 \rightarrow \cdots \rightarrow \cdot/G/1$ of M queues. The extension is more or less straightforward, except for the increase in notational complexity due to the need of keeping track of the queues using an additional index.

There are $n+1$ jobs denoted by $0, 1, \ldots, n$ and traversing the tandem. Job 0 arrives at time 0 to an empty system, and the interarrival time between jobs $k-1$ and k is $X_{M+n-k+1}$ for $k = 1, \ldots, n$; the arrival time of job k is thus $\sum_{i=1}^{k} X_{M+n-i+1}$. The service times of job k at the queues $j \in \{1, 2, \ldots, M\}$ are light-tailed and denoted[5] by $Y^{(j)}_{M+n-j+1-k}$; see Fig. 6. All sequences are i.i.d. and mutually independent, and we assume the stability condition $\mathbb{E}[X_M] > \max_j \mathbb{E}[Y^{(j)}_{M-j+1}]$.

As in the $M = 2$ particular case, we point out that the apparent cumbersome indexing for the jobs' interarrival and service times lends itself to simple/clean expressions for the exit, waiting, and sojourn times, to be derived in Appendix § B.1; in turn, a simpler indexing in the model would make the notation for the waiting and sojourn times, and of most incoming derivations in the paper, as extremely cumbersome.

[5]Superscript indexes mainly stand for the queues' positions in the tandem. Recall also the convention to drop subscripts when clear from the context; for instance, X would stand for a r.v. with the same distribution as X_1, i.e., $X \simeq X_1$.

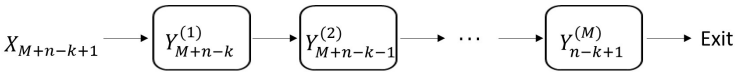


Fig. 6. Inter-arrival and service times for job k in a tandem of M queues

Define the compact set

$$\mathcal{D}_M := \left\{(v_1 \wedge v_2, \ldots, v_{M-1} \wedge v_M, v_M) : v_1 \geq 0, (v_1, \ldots, v_M) \in \bar{\mathbb{R}}^M\right\} ,$$

which is a closed subset of the compact set $\bar{\mathbb{R}}^M$. Denote also the maxima of random walks

$$T^j_k := \max_{k \leq i_j < i_{j-1} < \cdots < i_1 < \infty} Y^{(j)}_k + V^{(j)}_{k+1} \cdots + V^{(j)}_{i_j} + V^{(j-1)}_{i_j+1} + \cdots + V^{(j-1)}_{i_{j-1}} + \cdots + V^{(1)}_{i_2+1} + \cdots + V^{(1)}_{i_1} ,$$

for $1 \leq j \leq M$ and $k \geq 1$, where $V^{(i)} = Y^{(i)} - X$; according to the dropping subscripts convention, $V^{(i)} \simeq Y^{(i)} - X$ stands for $V^{(i)}_k \simeq Y^{(i)}_k - X_k \ \forall k$. Note also that $V^{(1)}$ and $V^{(2)}$ play the roles of the U and V from the $M = 2$ case.

Next we present the main result which establishes that the joint distribution of the maxima of random walks T^j_1, for $j = 1, \ldots, M$, which drive the representations of $\mathcal{W}$ and $\mathcal{S}$ (see Appendix § B.1), is the unique solution of a fixed-point integral equation.

Theorem 5. (Main Result: Generic Construction of Upper and Lower Bounds) *Let $V^{(i)} = Y^{(i)} - X$ be random variables satisfying $\mathbb{P}(V^{(i)} > 0) > 0$ for $i = 1, \ldots, M$, and where $Y^{(1)}, \ldots, Y^{(M)}$, and X are independent. Let $(V^{(1)}_1, \ldots, V^{(M)}_1)$, $(V^{(1)}_2, \ldots, V^{(M)}_2), \ldots$ be i.i.d. copies of $(V^{(1)}, \ldots, V^{(M)})$. Denote the vector*

$$\mathcal{V}(v_1, \ldots, v_M) := \left(\bigwedge_{1 \leq i \leq 2} \left(v_i - V^{(i)}\right), \ldots, \bigwedge_{M-1 \leq i \leq M} \left(v_i - V^{(i)}\right), v_M - V^{(M)} \right) .$$

(a) The integral equation

$$\mathbb{E}\left[\mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \psi\left(\mathcal{V}(v_1, \ldots, v_M)\right)\right] = \psi(v_1, \ldots, v_M) \tag{25}$$

admits a unique solution in the class of bounded functions $\psi : \mathcal{D}_M \to \mathbb{R}$ having the limit $\psi(\infty, \ldots, \infty) := \lim_{v_1, \ldots, v_M \to \infty} \psi(v_1, \ldots, v_M) = 1$. This is given by

$$\psi(v_1, \ldots, v_M) := \mathbb{P}\left(T^1_1 \leq v_1, \ldots, T^M_1 \leq v_M\right) .$$

(b) Assume that the function $\gamma : \mathcal{D}_M \to (-\infty, K_\gamma]$, for some finite K_γ, satisfies for all $(v_1, \ldots, v_M) \in$ supp $(\gamma \vee 0)$

$$\mathbb{E}\left[\mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \gamma\left(\mathcal{V}(v_1, \ldots, v_M)\right)\right] \geq \gamma(v_1, \ldots, v_M) . \tag{26}$$

If $\gamma(\infty, \ldots, \infty) := \limsup_{v_1, \ldots, v_M \to \infty} \gamma(v_1, \ldots, v_M) = 1$ then $\psi \geq \gamma$.

(c) Assume that the function $\eta : \mathcal{D}_M \to [0, \infty)$ satisfies for all $(v_1, \ldots, v_M) \in$ supp (ψ)

$$\mathbb{E}\left[\mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \eta\left(\mathcal{V}(v_1, \ldots, v_M)\right)\right] \leq \eta(v_1, \ldots, v_M) \tag{27}$$

If $\eta(\infty, \ldots, \infty) := \liminf_{v_1, \ldots, v_M \to \infty} \eta(v_1, \ldots, v_M) = 1$ then $\psi \leq \eta$.

As in the $M = 2$ case, the problem of finding upper and lower bounds on the tails of $\mathcal{W}$ and $\mathcal{S}$ reduces to the problem of finding the functions γ and η in (b) and (c), respectively; the connection is provided in the next result.

Corollary 6. (Generic Upper and Lower Bounds) *Consider the functions ψ, γ, and η as in Theorem 5. Then the waiting time $\mathcal{W}$ and sojourn time $\mathcal{S}$ of a job $n \to \infty$ satisfies for all $x \geq 0$*

$$\begin{aligned}
&1 - \mathbb{E}\left[\mathbb{1}_{\{x+X+S_M^1 \geq S_M^2\}} \eta\left(x + X + S_M^1 - S_M^1 \vee S_M^2, \ldots, x + X + S_M^1 - S_M^M\right)\right] \\
&\leq \mathbb{P}(\mathcal{W} > x) \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X+S_M^1 \geq S_M^2\}} \psi\left(x + X + S_M^1 - S_M^1 \vee S_M^2, \ldots, x + X + S_M^1 - S_M^M\right)\right] \\
&\leq 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X+S_M^1 \geq S_M^2\}} \gamma\left(x + X + S_M^1 - S_M^1 \vee S_M^2, \ldots, x + X + S_M^1 - S_M^M\right)\right]
\end{aligned}$$

and

$$\begin{aligned}
&1 - \mathbb{E}\left[\mathbb{1}_{\{x \geq S_{M-1}^2\}} \eta\left(x - S_{M-1}^2, x - S_{M-1}^2 \vee S_{M-1}^3, \ldots, x - S_{M-1}^M\right)\right] \\
&\leq \mathbb{P}(\mathcal{S} > x) \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{\{x \geq S_{M-1}^2\}} \psi\left(x - S_{M-1}^2, x - S_{M-1}^2 \vee S_{M-1}^3, \ldots, x - S_{M-1}^M\right)\right] \\
&\leq 1 - \mathbb{E}\left[\mathbb{1}_{\{x \geq S_{M-1}^2\}} \gamma\left(x - S_{M-1}^2, x - S_{M-1}^2 \vee S_{M-1}^3, \ldots, x - S_{M-1}^M\right)\right].
\end{aligned}$$

Next we prove the existence of polynomial-exponential bounds on the tails of $\mathcal{W}$ and $\mathcal{S}$.

Theorem 7. (Existence of polynomial-exponential Upper Bounds) *Denote $\theta_i := \sup\{r > 0 : \forall 1 \leq j \leq i : \mathbb{E}[e^{rV^{(j)}}] \leq 1\}$ for $i \in \{1, 2, \ldots, M\}$ and let*

$$I_i(\theta) := \begin{cases} 1 & \text{if } \mathbb{E}[e^{\theta V^{(i)}}] = 1 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

and

$$d_1(\theta) := 0, \quad d_i(\theta) := I_2(\theta) + \ldots + I_i(\theta), \qquad 2 \leq i \leq M\,.$$

Suppose that for all $v \geq 0$ and all $i \in \{1, 2, \ldots, M\}$ and $j \leq (2\lfloor (d_i(\theta_i) - 1)/2 \rfloor + 1) \vee 0$

$$\mathbb{E}\left[\left(V^{(i)} - v\right)^j e^{\theta_i (V^{(i)} - v)} \mid V^{(i)} > v\right] \leq K_{i,j} < +\infty\,,$$

for some positive constants $K_{i,j}$. Then there exist the polynomials $Q_i : \mathbb{R} \to \mathbb{R}$ with degrees $d_i(\theta_i)$ and $Q_i(v) \geq 0$ for all $1 \leq i \leq M$ and $v \geq 0$, such that $\gamma : \bar{\mathbb{R}}^M \to (-\infty, 1]$, defined by

$$\gamma(v_1, \ldots, v_M) := \mathbb{1}_{\{(v_1,\ldots,v_M) \in [0,\infty]^M\}} \left[1 - \sum_{i=1}^M Q_i(v_i) e^{-\theta_i v_i}\right],$$

satisfies (26) *for all $(v_1, \ldots, v_M) \in \operatorname{supp}(\gamma \vee 0)$. In particular, the restricted function*

$$\gamma|_{\mathcal{D}_M}(v_1, \ldots, v_M) := \mathbb{1}_{\{(v_1,\ldots,v_M) \in [0,\infty]^M \cap \mathcal{D}_M\}} \left[1 - \sum_{i=1}^M Q_i(v_i) e^{-\theta_i v_i}\right]$$

satisfies (26) *for all $(v_1, \ldots, v_M) \in \operatorname{supp}(\gamma|_{\mathcal{D}_M} \vee 0)$.*

Note that $2\lfloor \frac{d-1}{2} \rfloor + 1$ is the largest odd integer smaller or equal than an integer d. The degree of the polynomial $Q_M(v_M)$, which dictates the behaviors of $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$ according to Corollary 6, depends on the indicator functions from (28). At one extreme, if all queues are homogeneous and $\mathbb{E}[e^{\theta^+(Y-X)}] \geq 1$ (recall the discussion on 'light-tailedness'), then the degree of Q_M is $M - 1$, as all indicators I_i but the first are 1. At another extreme, if there is a single bottleneck, and $\mathbb{E}[e^{\theta^+(Y-X)}] \geq 1$, then the degree of P_M depends on the position of the bottleneck: if it comes

first then the degree is 0, otherwise it is 1 regardless the position (e.g., second or last). Another extreme in the homogeneous case is when $\mathbb{E}[e^{\theta^+(Y-X)}] < 1$, i.e., the service times are subject to a 'very light-tailed' distribution as in Appendix § D, in which case the degree of P_M would be 0.

5 Conclusions

This paper aims to improve queueing bounds available for general arrival processes but proverbially very loose. Unlike prior work dedicated to single queues (e.g., [2, 13, 14, 17, 32, 35]), the target here is a tandem queueing network which is renowned to pose extraordinary technical difficulties.

The breakthrough is the formulation of the stationary waiting and sojourn times in terms of random walks' maxima, whose joint distribution obeys a fundamental fixed-point integral equation. Relaxing this equation as an inequality lends itself to closed-form polynomial-exponential upper-bounds on the tails $\mathbb{P}(\mathcal{W} > x)$ and $\mathbb{P}(\mathcal{S} > x)$. In the $GI/M/1 \rightarrow \cdot/M/1$ tandem case, numerical results indicate that the obtained bounds are very sharp both in heavy and light traffic, and improve upon alternative large-deviations-based bounds by many orders of magnitude.

More general closed-form bounds were derived for a $GI/H_n/1 \rightarrow \cdot/H_n/1$ tandem with hyperexponential service times; numerical results revealed a slight accuracy loss when considering service times with coefficient of variations larger than one. This indicates that the specific polynomial-exponential structure of the bounds, as provided by the "Existence" Theorem 3, or the matching procedure of the underlying parameters from Lemma 4 can potentially be significantly improved. A more fundamental open question concerns the existence, or disproval, of matching lower bounds for the upper ones from Theorem 3; should the former be true, then obtaining exact closed-form distributions for $\mathcal{W}$ and $\mathcal{S}$ would be within reach.

Acknowledgements

This work has been funded by the Engineering and Physical Sciences Research Council (EPSRC) through the project EP/T031115/1. We thank Ziv Scully for shepherding our paper and the anonymous Sigmetrics reviewers for their thoughtful comments.

References

[1] Joseph Abate and Ward Whitt. 1997. Asymptotics for M/G/1 Low-Priority Waiting-Time Tail Probabilities. *Queueing Syst. Theory Appl.* 25, 1/4 (Jan. 1997), 173–233. doi:10.1023/A:1019104402024

[2] Søren Asmussen. 2003. *Applied Probability and Queues.* Springer.

[3] Hayriye Ayhan and François L. Baccelli. 2001. Expansions for Joint Laplace Transform of Stationary Waiting Times in $(max, +)$-Linear Systems with Poisson Input. *Queueing Syst. Theory Appl.* 37, 1/3 (2001), 291–328. doi:10.1023/A:1011008704491

[4] Hayriye Ayhan and Dong-Won Seo. 2002. Tail Probability of Transient and Stationary Waiting Times in $(max, +)$-Linear systems. *IEEE Trans. Automat. Control* 47, 1 (2002), 151–157. doi:10.1109/9.981736

[5] Simonetta Balsamo and Andrea Marin. 2007. *Queueing Networks.* 34–82. doi:10.1007/978-3-540-72522-0_2

[6] Robert S. Borden. 1998. *A Course in Advanced Calculus.* Dover Publications.

[7] Onno J. Boxma and Hans Daduna. 1990. Sojourn Times in Queueing Networks. In *Stochastic Analysis of Computer and Communication Systems (Editor H. Takagi).* North-Holland Publishing Company, 401–450.

[8] Almut Burchard, Jörg Liebeherr, and Florin Ciucu. 2011. On Superlinear Scaling of Network Delays. *IEEE/ACM Transactions on Networking* 19, 4 (Aug. 2011), 1043–1056. doi:10.1109/TNET.2010.2095505

[9] Paul J. Burke. 1964. The Dependence of Delays in Tandem Queues. *The Annals of Mathematical Statistics* 35, 2 (1964), 874 – 875. doi:10.1214/aoms/1177703590

[10] Cheng-Shang Chang. 2000. *Performance Guarantees in Communication Networks.* Springer Verlag.

[11] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt. 1996. Squeezing the Most out of ATM. *IEEE Transactions on Communications* 44, 2 (Feb. 1996), 203–217.

[12] Florin Ciucu, Almut Burchard, and Jörg Liebeherr. 2005. A Network Service Curve Approach for the Stochastic Analysis of Networks. In *ACM Sigmetrics.* 279–290.

[13] Florin Ciucu, Sima Mehri, and Amr Rizk. 2024. On Ultra-Sharp Queueing Bounds. In *IEEE Infocom.*

[14] Florin Ciucu and Felix Poloczek. 2018. Two Extensions of Kingman's GI/G/1 Bound. *Proc. of the ACM on Measurement and Analysis of Computing Systems - ACM Sigmetrics / IFIP Performance* 2, 3 (Dec. 2018), 43:1–43:33.
[15] Jacob W. Cohen. 1982. *The Single Server Queue (2nd Edition).* Elsevier Science.
[16] Ralph L. Disney and Dieter König. 1985. Queueing Networks: A Survey of Their Random Processes. *SIAM Rev.* 27, 3 (1985), 335–403. doi:10.1137/1027109
[17] Nick G. Duffield. 1994. Exponential Bounds for Queues with Markovian Arrivals. *Queueing Systems* 17, 3-4 (Sept. 1994), 413–430.
[18] Markus Fidler. 2006. An End-to-End Probabilistic Network Calculus with Moment Generating Functions. In *IEEE International Workshop on Quality of Service (IWQoS).* 261–270.
[19] Serguei G. Foss. 2007. On the Exact Asymptotics for the Stationary Sojourn Time Distribution in a Tandem of Queues with Light-Tailed Service Times. *Probl. Inf. Transm.* 43, 4 (Dec. 2007), 353–366.
[20] Ayalvadi J. Ganesh. 1998. Large Deviations of the Sojourn Time for Queues in Series. *Annals of Operations Research* 79 (1998), 3–26.
[21] Peter G. Harrison and William J. Knottenbelt. 2004. Quantiles of Sojourn Times. In *Computer System Performance Modeling in Perspective (Editor E. Gelenbe).* Imperial College Press, 155–193.
[22] Yuming Jiang and Yong Liu. 2008. *Stochastic Network Calculus.* Springer.
[23] Fridrikh I. Karpelevitch and Alexander Ya. Kreinin. 1992. Joint Distributions in Poissonian Tandem Queues. *Queueing Syst. Theory Appl.* 12, 3-4 (1992), 273–286. doi:10.1007/BF01158803
[24] John F. C. Kingman. 1964. A Martingale Inequality in the Theory of Queues. *Cambridge Philosophical Society* 60, 2 (April 1964), 359–361.
[25] John F. C. Kingman. 1970. Inequalities in the Theory of Queues. *Journal of the Royal Statistical Society, Series B* 32, 1 (1970), 102–110.
[26] W. Krämer. 1973. Total Waiting Time Distribution Function and the Fate of a Customer in a System with Two Queues in Series. In *International Teletraffic Congress (ITC 7).*
[27] Jörg Liebeherr, Almut Burchard, and Florin Ciucu. 2012. Delay Bounds in Communication Networks with Heavy-Tailed and Self-Similar Traffic. *IEEE Transactions on Information Theory* 58, 2 (Feb. 2012), 1010–1024.
[28] Haining Liu. 1993. *Buffer Size and Packet Loss in Tandem Queueing Network.* Ph. D. Dissertation. University of California at San Diego.
[29] R. M. Loynes. 1964. The Stability of a System of Queues in Series. *Mathematical Proc. of the Cambridge Philosophical Society* 60, 3 (July 1964), 569–574. doi:10.1017/S0305004100038044
[30] Michel R. H. Mandjes and Onno J. Boxma. 2023. *The Cramér-Lundberg Model and its Variants: A Queueing Perspective.* Springer.
[31] Randolph D. Nelson. 1993. The Mathematics of Product Form Queuing Networks. *ACM Comput. Surv.* 25, 3 (Sept. 1993), 339–369. doi:10.1145/158439.158906
[32] Zbigniew Palmowski and Tomasz Rolski. 1996. A Note on Martingale Inequalities for Fluid Models. *Statistics & Probability Letters* 31, 1 (Dec. 1996), 13–21.
[33] Felix Poloczek and Florin Ciucu. 2014. Scheduling Analysis with Martingales. *Performance Evaluation (Special Issue: IFIP Performance 2014)* 79 (Sept. 2014), 56 – 72.
[34] Edgar Reich. 1963. Note on Queues in Tandem. *The Annals of Mathematical Statistics* 34, 1 (1963), 338–341. http://www.jstor.org/stable/2991315
[35] Sheldon M. Ross. 1974. Bounds on the Delay Distribution in GI/G/1 queues. *Journal of Applied Probability* 11, 2 (June 1974), 417–421.
[36] Michel Talagrand. 1996. Majorizing Measures: The Generic Chaining. *Annals of Probability* 24, 3 (July 1996), 1049–1103.
[37] Jean C. Walrand. 1989. *An Introduction to Queueing Networks.* Prentice Hall.
[38] Jean C. Walrand and Pravin Varaiya. 1980. Sojourn Times and the Overtaking Condition in Jacksonian Networks. *Advances in Applied Probability* 12, 4 (1980), 1000–1018. http://www.jstor.org/stable/1426753

A Closed-Form Polynomial-Exponential Bounds

Proof. (of Lemma 4) Inequality (21) is equivalent to

$$AF_A + BF_B + CF_C + DF_D \geq F_1 \,, \tag{29}$$

where for $v \geq u$,

$$
\begin{aligned}
F_B(u,v) &:= e^{-\theta v} - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}} e^{\theta(V-v)}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y\}} e^{\theta(V-v)}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v<V\}} e^{\theta(V-v)}\right] \\
&= e^{-\theta v}\mathbb{P}(Y > u) + \mathbb{E}[K_0^Z(v+X)\mathbb{1}_{\{Z>v+X, Y\leq u\}}] \,.
\end{aligned}
$$

The last equality is obtained by using standard properties of conditional expectation and the independence of X, Y, Z:

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v<V\}} e^{\theta(V-v)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Z>X+v\}}\mathbb{1}_{\{Y\leq u\}} e^{\theta(Z-X-v)} \mid X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Z>X+v\}} e^{\theta(Z-X-v)} \mid X\right]\mathbb{1}_{\{Y\leq u\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{\theta(Z-X-v)} \mid Z > X+v, X\right]\mathbb{P}(Z > X+v \mid X)\mathbb{1}_{\{Y\leq u\}}\right] \\
&= \mathbb{E}\left[K_0^Z(X+v)\mathbb{P}(Z > X+v \mid X)\mathbb{1}_{\{Y\leq u\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[K_0^Z(X+v)\mathbb{1}_{\{Z>X+v\}}\mathbb{1}_{\{Y\leq u\}} \mid X\right]\right] \\
&= \mathbb{E}\left[K_0^Z(X+v)\mathbb{1}_{\{Z>X+v\}}\mathbb{1}_{\{Y\leq u\}}\right] \,.
\end{aligned}
$$

In the next to last line we used the measurability of $K_0^Z(X+v)$ and $\mathbb{P}(Z > X+v \mid X)$ with respect to the σ-field generated by X. The indicator $\mathbb{1}_{\{Y\leq u\}}$ easily expands to obtain the last two lines from the previous equation (for the formation of the convenient 'coloured' groups). We will use the same argument in the expansions of F_A, F_C, and F_D:

$$
\begin{aligned}
F_A(u,v) &:= e^{-\theta u} - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}} e^{-\theta((u-U)\wedge(v-V))}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y \text{ or } v<V\}} e^{-\theta u+\theta U}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left\{e^{-\theta u+\theta U} - e^{-\theta((u-U)\wedge(v-V))}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y\}} e^{-\theta u+\theta U}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V>v\}} e^{-\theta u+\theta U}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, 0\geq V-v\geq U-u\}}\left\{e^{-\theta u+\theta U} - e^{-\theta v+\theta V}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y\}} e^{-\theta u+\theta U}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V-v\geq U-u\}} e^{-\theta u+\theta U}\right] \\
&\quad - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V-v\geq U-u\}} e^{-\theta v+\theta V}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V>v\}} e^{-\theta v+\theta V}\right] \\
&= \mathbb{E}[K_0^Y(u)\mathbb{1}_{\{Y>u\}} e^{-\theta X}] \\
&\quad + \mathbb{E}\left[(1-K_0^Z(v-u+Y))\mathbb{1}_{\{Z\geq v-u+Y, Y\leq u\}} e^{\theta(Y-X-u)}\right] \\
&\quad + \mathbb{E}[K_0^Z(v+X)\mathbb{1}_{\{Z>v+X, Y\leq u\}}] \,.
\end{aligned}
$$

$$
\begin{aligned}
F_C(u,v) &:= ue^{-\theta v} - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}\left((u-U)\wedge(v-V)\right)e^{\theta(V-v)}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y \text{ or } v<V\}}(u-U)e^{-\theta v+\theta V}\right] + \mathbb{E}\left[Ue^{\theta(V-v)}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, 0\geq V-v\geq U-u\}}\left(u-v-U+V\right)e^{\theta(V-v)}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{u<Y\}}(u-U)e^{-\theta v+\theta V}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V>v\}}(v-V)e^{-\theta v+\theta V}\right] \\
&\quad + \mathbb{E}\left[Ue^{\theta(V-v)}\right] + \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, V-v\geq U-u\}}\left(u-v-U+V\right)e^{\theta(V-v)}\right] \\
&= \mathbb{E}[Ue^{\theta V}]e^{-\theta v} + \mathbb{E}[(u-U)e^{\theta V} \mid Y>u]\mathbb{P}(Y>u)e^{-\theta v} \\
&\quad - \mathbb{E}[K_1^Z(v+X)\mathbb{1}_{\{Z>v+X, Y\leq u\}}] \\
&\quad + \mathbb{E}\left[K_1^Z(v-u+Y)\mathbb{1}_{\{Z\geq v-u+Y, Y\leq u\}}e^{\theta(Y-X-u)}\right].
\end{aligned}
$$

$$
\begin{aligned}
F_D(u,v) &:= ve^{-\theta v} - \mathbb{E}\left[\mathbb{1}_{\{u\geq Y, v\geq V\}}(v-V)e^{\theta(V-v)}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{Y>u \text{ or } V>v\}}(v-V)e^{-\theta v+\theta V}\right] + \mathbb{E}\left[Ve^{\theta(V-v)}\right] \\
&= e^{-\theta v}\mathbb{E}\left[(v-V)e^{\theta V}\right]\mathbb{P}(Y>u) + \mathbb{E}\left[Ve^{\theta V}\right]e^{-\theta v} \\
&\quad - \mathbb{E}[K_1^Z(v+X)\mathbb{1}_{\{Z>v+X, Y\leq u\}}].
\end{aligned}
$$

Lastly

$$
\begin{aligned}
F_1(u,v) &:= \mathbb{P}(Y>u) + \mathbb{P}(V>v, Y\leq u) \\
&= \mathbb{P}(Y>u) + \mathbb{P}(Z>v+X, Y\leq u).
\end{aligned}
$$

□

A.1 A Large-Deviations / Network Calculus Approach

Here we present alternative bounds using the standard large-deviations / network calculus approach, which crucially relies on the Union Bound

$$
\mathbb{E}[\max\{X,Y\}] \leq \mathbb{E}[X] + \mathbb{E}[Y] \qquad \text{or} \qquad \mathbb{P}(A\cup B) \leq \mathbb{P}(A) + \mathbb{P}(B),
$$

for positive r.v. X and Y, or events A and B. One advantage of this approach is that it yields the *exact* asymptotic decay rates (e.g., for $\mathbb{P}(\mathcal{S} > x)$, see Ganesh [20]). Another, as shown in several applications of network calculus, is that it enables the analysis of queueing networks with broad classes of arrivals, service times, or scheduling algorithms, and can further lead to the exact asymptotic scaling of sojourn times ([8, 10, 12, 18, 22, 28]).

The drawback of this class of results is poor numerical tightness, particularly in non-asymptotic regimes (i.e., for finite values of x in the case of $\mathbb{P}(\mathcal{S} > x)$). This issue was brought up in the context of the (large-deviations-based) effective bandwidth literature from the late 80's - 90's. Choudhury *et al.* [11] revealed large numerical discrepancies, of several orders of magnitude, between effective bandwidth results and simulations in the case of Markovian arrivals. More recently, similar numerical issues have been reported about network calculus results concerning single queues only; in addition, it was shown that relying on Kingman's GI/G/1 bound (recall § 2.5), as opposed to the Union Bound, can largely fix the issue of numerical tightness in single queues

(Poloczek and Ciucu [33]) and broad arrival patterns in heavy-traffic; at the other extreme, in the case of light-traffic, ultra-sharp bounds have been recently obtained in Ciucu *et al.* [13].

The large numerical inaccuracies mainly stem from the obliviousness of the Union Bound to the underlying correlations; this issue is particularly pronounced in non-Poisson/non-memoryless type of events, as indicated by Talagrand [36]. Moreover, the underlying numerical errors accumulate over an infinite number of applications of the Union Bound, as waiting/sojourn times involve whole sample-paths.

For example, when aiming for the sojourn time $\mathcal{S}$, the network calculus / large-deviations approach proceeds by first computing the bounds

$$\begin{cases} \mathbb{P}\left(\max_{3\le i<\infty} U_3+\cdots+U_i > x\right) \le (\beta+\beta^2+\beta^3+\ldots)e^{-\theta x} = \frac{\beta}{1-\beta}e^{-\theta x} \\ \mathbb{P}\left(\max_{2\le i<j<\infty} V_3+\cdots+V_i+U_{i+1}+\cdots+U_j > x\right) \\ \qquad\qquad \le (\beta+2\beta^2+3\beta^3+\ldots)e^{-\theta x} = \frac{\beta}{(1-\beta)^2}e^{-\theta x} \end{cases}$$

by repeatedly using the Chernoff and Union Bounds, where $\beta := \mathbb{E}\left[e^{\theta(Y-X)}\right]$ and $\theta > 0$ is chosen such that $\beta < 1$ (to guarantee the convergence of the infinite series). The former bound concerns the maximum of a random walk, an event characteristic to single queues, and follows by infinitely applying the Union Bound – at the expense of disregarding correlations within the partial sums $U_3+\cdots+U_i$. In turn, the latter concerns an event involving a double-maximum (over i and j), which is characteristic to a tandem of two queues; the bound itself follows from a nested infinite application of the Union Bound, while also disregarding correlations within the underlying partial sums. The numerical inaccuracies associated with the former bound, as reported in the single-queues literature, naturally exacerbate in the case of the latter bound targeting tandem queues.

In the case when Y and Z are exponentially distributed with rate μ, we obtain by applying the Union Bound one more time, along with double integration, that

$$\begin{aligned} \mathbb{P}(\mathcal{S} > x) &\le \mathbb{P}(Y+Z>x) + \mathbb{P}\left(\max_{3\le i<\infty} U_3+\cdots+U_i > x-Y-Z \ge 0\right) \\ &\quad + \mathbb{P}\left(\max_{2\le i<j<\infty} V_3+\cdots+V_i+U_{i+1}+\cdots+U_j > x-Y-Z \ge 0\right) \\ &\le (1+\mu x)e^{-\mu x} + \inf_{\{0<\theta<\mu:\beta<1\}} \frac{\beta(2-\beta)}{(1-\beta)^2}\frac{\mu^2}{(\mu-\theta)^2}\left(e^{-\theta x} - (1+(\mu-\theta)x)e^{-\mu x}\right) . \end{aligned} \tag{30}$$

Due to the underlying transcendental nature of the bound, the optimal value of θ requires a numerical search.

B Generalization: Tandem of M Queues

B.1 A Novel Representation of Waiting and Sojourn Times

Using induction and Lindley's recursion, the exit time of job n from the tandem is

$$\begin{aligned} \tau_n := \max_{1\le i_M<i_{M-1}<\ldots<i_1\le M+n} & Y_1^{(M)}+\cdots+Y_{i_M}^{(M)}+Y_{i_M+1}^{(M-1)}+\cdots+Y_{i_{M-1}}^{(M-1)} \\ & +\cdots+Y_{i_2+1}^{(1)}+\cdots+Y_{i_1}^{(1)}+X_{i_1+1}+\cdots+X_{M+n} . \end{aligned}$$

The expression is an immediate generalization of the exit time shown for $M=2$ queues (see (1)). Note that, when $M=2$, $Y_i^{(2)}$ corresponds to Z_i (the service times at the second queue), whereas $Y_i^{(1)}$ corresponds to Y_i (the service times at the first queue).

(Re)denoting $V_k^{(i)} \simeq Y_k^{(i)} - X_k$ let us redefine for convenience the maxima of random walks

$$T_k^j := \max_{k \le i_j < i_{j-1} < \cdots < i_1 < \infty} Y_k^{(j)} + V_{k+1}^{(j)} \cdots + V_{i_j}^{(j)} + V_{i_j+1}^{(j-1)} + \cdots + V_{i_{j-1}}^{(j-1)} + \cdots + V_{i_2+1}^{(1)} + \cdots + V_{i_1}^{(1)}$$

for $1 \le j \le M$ and $k \ge 1$. These are subject to the recursions

$$T_k^j = Y_k^{(j)} - X_{k+1} + T_{k+1}^{j-1} \vee T_{k+1}^j \ \forall j = 2, \ldots, M \tag{31}$$

and $T_k^1 = Y_k^{(1)} + \left(T_{k+1}^1 - X_{k+1}\right) \vee 0$ for all $k \ge 1$. The derivations follow immediately as in the $M = 2$ case by conveniently regrouping terms (see § 2.2).

To obtain $\mathcal{W}$ and $\mathcal{S}$ we also need to define

$$S_k^M := Y_1^{(M)} + \cdots + Y_k^{(M)} \ \forall k \ge 1 \,,$$

$$S_k^j := \max_{1 \le i_M < \ldots < i_{j+1} < k} Y_1^{(M)} + \cdots + Y_{i_M}^{(M)} + Y_{i_M+1}^{(M-1)} + \cdots + Y_{i_{M-1}}^{(M-1)} + \cdots + Y_{i_{j+1}+1}^{(j)} + \cdots + Y_k^{(j)}$$

for $1 \le j \le M-1$ and $k \ge M - j + 1$. These are partial sums of the service times across the queues, and are subject to the recursions $S_M^1 = S_{M-1}^2 + Y_M^{(1)}$, $S_M^j = S_{M-1}^j \vee S_{M-1}^{j+1} + Y_M^{(j)}$ for $j = 2, \ldots, M-1$, and $S_M^M = S_{M-1}^M + Y_M^{(M)}$, obtained by regrouping terms.

Using also the recursion from (31), we can express the waiting time of jobs in the limit $n \to \infty$ as

$$\begin{aligned} \mathcal{W} &:= \lim_{n \to \infty} \left\{ \tau_n - \left[Y_1^{(M)} + Y_2^{(M-1)} + \cdots + Y_M^{(1)} + X_{M+1} + X_{M+2} + \cdots + X_{M+n} \right] \right\} \\ &= \max_{1 \le j \le M} \left\{ T_{M+1}^j - X_{M+1} + S_M^j \vee S_M^{(j+1) \wedge M} - S_M^1 \right\} \vee 0 \,, \end{aligned} \tag{32}$$

which is subject to a unique stationary distribution (Loynes [29]).

In turn, the sojourn time of job $n \to \infty$ is the sum of the waiting and the local service times, i.e.,

$$\mathcal{S} := \mathcal{W} + Y_M^{(1)} + \cdots + Y_1^{(M)} = \max_{1 \le j \le M} \left\{ T_M^j + S_{M-1}^{j \vee 2} \vee S_{M-1}^{(j+1) \wedge M} \right\} \,.$$

This follows from the expression of $\mathcal{W}$ from (32)

To further familiarize with notation, note that the superscript pairs $(j \vee 2, (j+1) \wedge M)$ for $j = 1, \ldots, M$ span the pairs $\{(2,2), (2,3), (3,4), \ldots, (M-1, M), (M, M)\}$. In the case when $M = 2$

$$\mathcal{S} = \max\{T_2^1 + S_1^2 \vee S_1^2, T_2^2 + S_1^2 \vee S_1^2\} \,,$$

which recovers the expression of $\mathcal{S}$ from (5) since $S_1^2 = Y_1^{(2)}$ and $Y_1^{(2)} = Z_1$ in the notation from the $M = 2$ case. In turn, in the case when $M = 3$, we would have

$$\mathcal{S} = \max\{T_3^1 + S_2^2 \vee S_2^2, T_3^2 + S_2^2 \vee S_2^3, T_3^3 + S_2^3 \vee S_2^3\} \,,$$

where $S_2^2 = Y_1^{(3)} + Y_2^{(2)}$ and $S_2^3 = Y_1^{(3)}$.

B.2 Proofs

Proof. (of Theorem 5) For (a) we can write for all $(v_1, \dots, v_M) \in \mathcal{D}_M$

$$\begin{aligned}
&\psi(v_1, \dots, v_M) = \mathbb{P}(T_1^1 \le v_1, \dots, T_1^M \le v_M) \\
&= \mathbb{P}\Bigg(Y_1^{(1)} \le v_1, Y_1^{(1)} - X_2 + T_2^1 \le v_1, Y_1^{(2)} - X_2 + T_2^1 \vee T_2^2 \le v_2, \dots, \\
&\qquad Y_1^{(M)} - X_2 + T_2^{M-1} \vee T_2^M \le v_M \Bigg) \\
&= \mathbb{P}\Bigg(Y_1^{(1)} \le v_1, T_2^1 \le \bigwedge_{1 \le i \le 2} \left(v_i + X_2 - Y_1^{(i)} \right), \dots, \\
&\qquad T_2^{M-1} \le \bigwedge_{M-1 \le i \le M} \left(v_i + X_2 - Y_1^{(i)} \right), T_2^M \le v_M + X_2 - Y_1^{(M)} \Bigg) \\
&= \mathbb{E}\left[\mathbb{1}_{\{v_1 \ge Y^{(1)}\}} \psi \left(\bigwedge_{1 \le i \le 2} \left(v_i - V^{(i)} \right), \dots, \bigwedge_{M-1 \le i \le M} \left(v_i - V^{(i)} \right), v_M - V^{(M)} \right) \right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{v_1 \ge Y^{(1)}\}} \psi\left(\mathcal{V}(v_1, \dots, v_M) \right) \right].
\end{aligned}$$

To show the uniqueness of ψ we first prove (b) and (c) and denote

$$\phi_1 := \gamma - \psi, \quad \phi_2 := \psi - \eta$$

and

$$f_i(v_1, \dots, v_M) := \limsup_{(u_1, \dots, u_M) \to (v_1, \dots, v_M)} \phi_i(u_1, \dots, u_M),$$

which are upper semi-continuous and attain their maximums on the compact set $\mathcal{D}_M$ (see Appendix C)

$$K_i := \max_{(v_1, \dots, v_M) \in \mathcal{D}_M} f_i(v_1, \dots, v_M), \qquad i = 1, 2.$$

Since $f_i(\infty, \dots, \infty) = 0$ it follows that $K_i \ge 0$. If $K_i = 0$ the proof is complete; assume otherwise that $K_i > 0$. Let

$$\mathcal{K}_i := \{(v_1, \dots, v_M) \in \mathcal{D}_M : f_i(v_1, \dots, v_M) = K_i\}$$

and define for $i = 1, 2$ and $j = 2, \dots, M$

$$a_1^{(i)} := \min\{v_1 \in \bar{\mathbb{R}} : \exists (v_2, \dots, v_M) \in \bar{\mathbb{R}}^{M-1} : (v_1, \dots, v_M) \in \mathcal{K}_i\}$$

$$a_j^{(i)} := \min\{v_j \in \bar{\mathbb{R}} : \exists (v_{j+1}, \dots, v_M) \in \bar{\mathbb{R}}^{M-j}, (a_1^{(i)}, \dots, a_{j-1}^{(i)}, v_j, \dots, v_M) \in \mathcal{K}_i\},$$

which are well-defined since f_i is upper semi-continuous; also,

$$(a_1^{(1)}, \dots, a_M^{(1)}) \in \mathcal{K}_1 \subseteq \operatorname{supp}(\gamma \vee 0), \quad (a_1^{(2)}, \dots, a_M^{(2)}) \in \mathcal{K}_2 \subseteq \operatorname{supp}(\psi).$$

We can now write

$$\begin{aligned}
K_i &= f_i(a_1^{(i)}, \ldots, a_M^{(i)}) \\
&= \limsup_{(v_1,\ldots,v_M)\to(a_1^{(i)},\ldots,a_M^{(i)})} \phi_i(v_1, \ldots, v_M) \\
&\leq \limsup_{(v_1,\ldots,v_M)\to(a_1^{(i)},\ldots,a_M^{(i)})} \mathbb{E}\left[\mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \phi_i\left(\mathcal{V}(v_1, \ldots, v_M)\right)\right] \\
&\leq \mathbb{E}\left[\limsup_{(v_1,\ldots,v_M)\to(a_1^{(i)},\ldots,a_M^{(i)})} \mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \phi_i\left(\mathcal{V}(v_1, \ldots, v_M)\right)\right] \\
&\leq \mathbb{E}\left[\limsup_{(v_1,\ldots,v_M)\to(a_1^{(i)},\ldots,a_M^{(i)})} \mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \left(\phi_i\left(\mathcal{V}(v_1, \ldots, v_M)\right) \vee 0\right)\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{\{a_1^{(i)} \geq Y^{(1)}\}} \left(f_i\left(\mathcal{V}\left(a_1^{(i)}, \ldots, a_M^{(i)}\right)\right) \vee 0\right)\right] \\
&\leq K_i \cdot \mathbb{P}(a_1^{(i)} \geq Y^{(1)})
\end{aligned}$$

Hence $\mathbb{P}(a_1^{(i)} \geq Y^{(1)}) = 1$ and

$$\begin{aligned}
&f_i\left(\mathcal{V}\left(a_1^{(i)}, \ldots, a_M^{(i)}\right)\right) \\
&= f_i\left(\bigwedge_{1\leq j\leq 2}\left(a_j^{(i)} - V^{(j)}\right), \ldots, \bigwedge_{M-1\leq j\leq M}\left(a_j^{(i)} - V^{(j)}\right), a_M^{(i)} - V^{(M)}\right) \qquad (33) \\
&= K_i \text{ a.s.}
\end{aligned}$$

We now prove by contradiction that $(a_1^{(i)}, \ldots, a_M^{(i)}) = (\infty, \ldots, \infty)$. Letting

$$j_0 := \min\{j \in \{1, 2, \ldots, m\}, a_j^{(i)} < \infty\}$$

it then follows from (33) and the choice of $a_{j_0}^{(i)}$ that $\mathbb{P}(V^{(j_0)} > 0) = 0$, thus contradicting the assumption that $\mathbb{P}(V^{(j_0)} > 0) > 0$. Therefore

$$K_i = f_i(\infty, \ldots, \infty) = 0\,,$$

which contradicts with the assumption that $K_i > 0$, and hence $\psi \leq \gamma$ and $\eta \leq \psi$.

We can now prove the uniqueness of ψ solving for (25). Let ψ_1 and ψ_2 be two bounded solutions satisfying

$$\psi_i(\infty, \ldots, \infty) = \lim_{v_1,\ldots,v_M\to\infty} \psi_i(v_1, \ldots, v_M) = 1\,.$$

Applying the second part of the theorem with $\psi = \psi_i$ and $\gamma = \psi_{3-i}$ (note that the proof only needs that ψ satisfies (25), is bounded, and $\psi(\infty, \ldots, \infty) = \lim_{v_1,\ldots,v_M\to\infty} \psi(v_1, \ldots, v_M) = 1$) we obtain that

$$\psi_i \geq \psi_{3-i}$$

for $i = 1, 2$, and hence $\psi_1 = \psi_2$. For more details about various steps in the proof see the proof for $M = 2$ from Appendix § 2.4. □

Proof. (of Corollary 6) We have for $x \geq 0$

$$\begin{aligned}\mathbb{P}(\mathcal{W} > x) &= 1 - \mathbb{P}\left(\max_{1\leq j\leq M}\{T^j_{M+1} + S^j_M \vee S^{(j+1)\wedge M}_M\} \leq x + X_{M+1} + S^1_M\right)\\ &= 1 - \mathbb{P}\left(\forall 1 \leq j \leq M, T^j_{M+1} \leq x + X_{M+1} + S^1_M - S^j_M \vee S^{(j+1)\wedge M}_M\right)\\ &= 1 - \mathbb{E}\left[\mathbb{1}_{\{x+X+S^1_M \geq S^2_M\}}\psi\left(x + X + S^1_M - S^1_M \vee S^2_M, \ldots, x + X + S^1_M - S^M_M\right)\right],\end{aligned}$$

from the stationarity of T^j_k for all $k \in \mathbb{N}$. Since

$$\left(x + X + S^1_M - S^1_M \vee S^2_M, \ldots, x + X + S^1_M - S^M_M\right) \in \mathcal{D}_M$$

and $\eta \leq \psi \leq \gamma$ on $\mathcal{D}_M$, the upper and lower bounds on $\mathbb{P}(\mathcal{W} > x)$ follow immediately. In turn, in the case of the sojourn time, we have for $x \geq 0$

$$\begin{aligned}\mathbb{P}(\mathcal{S} > x) &= 1 - \mathbb{P}\left(\max_{1\leq j\leq M}\{T^j_M + S^{j\vee 2}_{M-1} \vee S^{(j+1)\wedge M}_{M-1}\} \leq x\right)\\ &= 1 - \mathbb{P}\left(\forall 1 \leq j \leq M, T^j_M \leq x - S^{j\vee 2}_{M-1} \vee S^{(j+1)\wedge M}_{M-1}, x \geq S^2_{M-1}\right)\\ &= 1 - \mathbb{E}\left[\mathbb{1}_{\{x\geq S^2_{M-1}\}}\psi\left(x - S^2_{M-1}, x - S^2_{M-1} \vee S^3_{M-1}, \ldots, x - S^M_{M-1}\right)\right].\end{aligned}$$

Since

$$\left(x - S^2_{M-1}, x - S^2_{M-1} \vee S^3_{M-1}, \ldots, x - S^M_{M-1}\right) \in \mathcal{D}_M$$

and $\eta \leq \psi \leq \gamma$ on $\mathcal{D}_M$, the upper and lower bounds on $\mathbb{P}(\mathcal{S} > x)$ follow immediately. □

Proof. (of Corollary 7) We proceed in two steps.

Step 1: First we prove by induction on $i \geq 1$ that there exist the polynomials $Q_i : \mathbb{R} \to \mathbb{R}, i \in \{1, 2, \ldots, M\}$ with degrees at most $d_i(\theta_i)$, respectively, having non-negative values on $[0, \infty)$, such that for all $v \geq 0$

$$Q_1 e^{-\theta_1 v} \geq \mathbb{E}\left[\mathbb{1}_{\{v\geq Y^{(1)}\}} Q_1 e^{\theta_1(V^{(1)}-v)}\right] + \mathbb{P}(v < Y^{(1)}) \tag{34}$$

and for all $2 \leq i \leq M$ and $v \geq 0$

$$Q_i(v)e^{-\theta_i v} \geq \mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(i)}\}} \sum_{l=i-1}^{i} Q_l\left(v - V^{(i)}\right) e^{\theta_l(V^{(i)}-v)}\right] + \mathbb{P}(v < V^{(i)}). \tag{35}$$

Proof of Step 1: The case $i = 1$ follows immediately by letting

$$Q_1 := \left(\inf_{v\geq 0} \mathbb{E}\left[e^{\theta_1(V^{(1)}-v)} \mid Y^{(1)} > v\right]\right)^{-1}.$$

For some $k \geq 2$ we next assume the existence of the polynomials Q_i for $i \leq k-1$. We need to prove that there exists $Q_k(v) := \sum_{j=0}^{d_k(\theta_k)} A_j v^j$ such that (35) holds for $i = k$. It is thus sufficient to show that there exists the non-negative constants $A_{d_k(\theta_k)}, \ldots, A_0$ such that

$$\begin{aligned}&\sum_{j=0}^{d_k(\theta_k)} A_j \left\{v^j e^{-\theta_k v} - \mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}}\left(v - V^{(k)}\right)^j e^{\theta_k(V^{(k)}-v)}\right]\right\}\\ &\qquad \geq \mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}} Q_{k-1}\left(v - V^{(k)}\right) e^{\theta_{k-1}(V^{(k)}-v)}\right] + \mathbb{P}(v < V^{(k)}).\end{aligned} \tag{36}$$

Next we bound both sides and then show the existence of the A_j's for the tighter inequality. An upper bound on the first term from the right side is

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}}Q_{k-1}\left(v-V^{(k)}\right)e^{\theta_{k-1}(V^{(k)}-v)}\right]\\
&\leq \mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}}\left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}Q_{k-1}\left(v-V^{(k)}\right)\right\}e^{\theta_k(V^{(k)}-v)}\right]\\
&= \mathbb{E}\left[\left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}Q_{k-1}\left(v-V^{(k)}\right)\right\}e^{\theta_k(V^{(k)}-v)}\right]\\
&\quad-\mathbb{E}\left[\left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}Q_{k-1}\left(v-V^{(k)}\right)\right\}e^{\theta_k(V^{(k)}-v)}\mid V^{(k)}>v\right]\\
&\quad\times\mathbb{P}(V^{(k)}>v)\\
&\leq \left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}R_{d_{k-1}(\theta_{k-1})}(v)\right\}e^{-\theta_k v}\\
&\quad-\inf_{v\geq 0}\mathbb{E}\left[\left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}Q_{k-1}\left(v-V^{(k)}\right)\right\}e^{\theta_k(V^{(k)}-v)}\mid V^{(k)}>v\right]\\
&\quad\times\mathbb{P}(V^{(k)}>v)\\
&\leq \left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}R_{d_{k-1}(\theta_{k-1})}(v)\right\}e^{-\theta_k v}\\
&\quad+\left(C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+C\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}\sum_{l=0}^{\lfloor(d_{k-1}(\theta_{k-1})-1)/2\rfloor\vee 0}K_k^{2l+1}\right)\mathbb{P}(V^{(k)}>v)\\
&\leq \left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}R_{d_{k-1}(\theta_{k-1})}(v)\right\}e^{-\theta_k v}+C\mathbb{P}(V^{(k)}>v)\,,
\end{aligned}$$

where $R_{d_{k-1}(\theta_{k-1})}$ is a polynomial of degree at most $d_{k-1}(\theta_{k-1})$ and C is some positive constant. In the last inequality we used the induction hypothesis on Q_{k-1} and only bounded the odd powers of $(v-V^{(k)})$ using the assumption on the conditional expectations; the other terms are positive.

Next we lower bound the terms in brackets from the left side of (36). For $j\geq 1$

$$\begin{aligned}
&v^j e^{-\theta_k v}-\mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}}\left(v-V^{(k)}\right)^j e^{\theta_k(V^{(k)}-v)}\right]\\
&= v^j e^{-\theta_k v}-\mathbb{E}\left[\left(v-V^{(k)}\right)^j e^{\theta_k(V^{(k)}-v)}\right]+\mathbb{E}\left[\mathbb{1}_{\{V^{(k)}>v\}}\left(v-V^{(k)}\right)^j e^{\theta_k(V^{(k)}-v)}\right]\\
&\geq \left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)v^j e^{-\theta_k v}+j\mathbb{E}\left[V^{(k)}e^{\theta_k V^{(k)}}\right]v^{j-1}e^{-\theta_k v}+\tilde{R}_{j-2}(v)e^{-\theta_k v}\\
&\quad+\inf_{v\geq 0}\mathbb{E}\left[\left(v-V^{(k)}\right)^j e^{\theta_k(V^{(k)}-v)}\mid V^{(k)}>v\right]\mathbb{P}(V^{(k)}>v)\\
&\geq \left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)v^j e^{-\theta_k v}+j\mathbb{E}\left[V^{(k)}e^{\theta_k V^{(k)}}\right]v^{j-1}e^{-\theta_k v}\\
&\quad+\tilde{R}_{j-2}(v)e^{-\theta_k v}-K_k^j\mathbb{1}_{\{j\in 2\mathbb{Z}+1\}}\mathbb{P}(V^{(k)}>v)\,,
\end{aligned}$$

where $\tilde{R}_{j-2}$ is a polynomial of degree at most $j-2$; in the last inequality we only bounded the conditional expectation with K_k^j when j is odd, by accounting for the underlying sign. Also, for $j=0$,

$$\begin{aligned}
&e^{-\theta_k v}-\mathbb{E}\left[\mathbb{1}_{\{v\geq V^{(k)}\}}e^{\theta_k(V^{(k)}-v)}\right]\\
&\geq \left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)e^{-\theta_k v}+\inf_{v\geq 0}\mathbb{E}\left[e^{\theta_k(V^{(k)}-v)}\mid V^{(k)}>v\right]\mathbb{P}(V^{(k)}>v)\\
&\geq \left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)e^{-\theta_k v}+\mathbb{P}(V^{(k)}>v)\,.
\end{aligned}$$

It is thus sufficient to find the coefficients A_j's satisfying the tighter inequality:

$$\begin{aligned} A_0 &\left\{\left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)e^{-\theta_k v}+\mathbb{P}(V^{(k)}>v)\right\} \\ &+\sum_{j=1}^{d_k(\theta_k)} A_j\Bigg\{\left(1-\mathbb{E}\left[e^{\theta_k V^{(k)}}\right]\right)v^j e^{-\theta_k v}+j\mathbb{E}\left[V^{(k)}e^{\theta_k V^{(k)}}\right]v^{j-1}e^{-\theta_k v} \\ &\qquad +\tilde{R}_{j-2}(v)e^{-\theta_k v}-K_k^j\mathbb{1}_{\{j\in 2\mathbb{Z}+1\}}\mathbb{P}(V^{(k)}>v)\Bigg\} \\ &\geq \left\{C\mathbb{1}_{\{\theta_{k-1}>\theta_k\}}+\mathbb{1}_{\{\theta_{k-1}=\theta_k\}}R_{d_{k-1}(\theta_{k-1})}(v)\right\}e^{-\theta_k v}+(C+1)\mathbb{P}(V^{(k)}>v)\,. \end{aligned} \tag{37}$$

There are two main cases:

If $\theta_k < \theta_{k-1}$ (Case 1), i.e., node k is the first bottleneck in the sequence $(1,2,\dots,k)$, then $d_k(\theta_k) = I_k(\theta_k) \in \{0,1\}$; in this case, $I_k(\theta_k) = 0$ happens when $\mathbb{E}[e^{\theta(Y^{(k)}-X)}] < 1\ \forall\theta > 0$. If $\mathbb{E}[e^{\theta_k V^{(k)}}] < 1$ (Case 1.1) then we first choose $A_1 = 0$. It is easy to see that $A_0 > 0$ sufficiently large is sufficient to satisfy (37). If $\mathbb{E}[e^{\theta_k V^{(k)}}] = 1$ (Case 1.2) then $d_k(\theta_k) = 1$ and the existence of A_0 and A_1 is obvious.

In the other main case, i.e., $\theta_k = \theta_{k-1}$ (Case 2), the terms containing $\mathbb{P}(V^{(k)} > v)$ are properly bounded by choosing a sufficiently large $A_0 > 0$. If $\mathbb{E}[e^{\theta_k V^{(k)}}] < 1$ (Case 2.1) then the coefficient of A_j is a polynomial of degree j with positive dominant coefficient, and since the right side of (37) has degree $d_{k-1}(\theta_{k-1}) = d_k(\theta_k)$ (because $\mathbb{E}[e^{\theta_k V^{(k)}}] < 1 \Rightarrow I_k(\theta_k) = 0$) there exist non-negative $A_{d_k(\theta_k)},\dots,A_0$ satisfying (37). Similarly, if $\mathbb{E}[e^{\theta_k V^{(k)}}] = 1$, the coefficient of A_j is a polynomial of degree $j-1$ with positive dominant coefficient and since the right side of (37) has degree $d_{k-1}(\theta_{k-1}) = d_k(\theta_k) - 1$ (because $\mathbb{E}[e^{\theta_k V^{(k)}}] = 1 \Rightarrow I_k(\theta_k) = 1$) it follows that there exist non-negative $A_{d_k(\theta_k)},\dots,A_0$ satisfying (37). Step 1 is thus complete.

Step 2: Let the Q_j's from Step 1 and define

$$\gamma(v_1,\dots,v_M) := \mathbb{1}_{\{(v_1,\dots,v_M)\in[0,\infty]^M\}}\left[1-\sum_{j=1}^{M}Q_j(v_j)e^{-\theta_j v_j}\right]\,.$$

Then γ satisfies (26) for all $(v_1,\dots,v_M) \in \operatorname{supp}(\gamma \vee 0)$.

Proof of Step 2: We prove by induction on $m = 1,\dots,M$ that the marginal functions γ_m of γ (i.e., restricted to the first m components), defined by

$$\gamma_m(v_1,\dots,v_m) := \mathbb{1}_{\{(v_1,\dots,v_m)\in[0,\infty]^m\}}\left[1-\sum_{j=1}^{m}Q_j(v_j)e^{-\theta_j v_j}\right]$$

satisfy (26) for all $(v_1,\dots,v_m) \in \operatorname{supp}(\gamma_m \vee 0)$[6]. For $m = 1$,

$$\gamma_1(v_1) := (1-Q_1e^{-\theta_1 v_1})\mathbb{1}_{\{v_1\geq 0\}}$$

satisfies for all $v_1 \geq 0$

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_{\{v_1\geq Y^{(1)}\}}\gamma_1(v_1-V^{(1)})\right] &= \mathbb{E}\left[\mathbb{1}_{\{v_1\geq Y^{(1)}\}}\left(1-Q_1e^{-\theta_1 v_1+\theta_1 V^{(1)}}\right)\right] \\ &\geq 1-Q_1e^{-\theta_1 v_1} = \gamma_1(v_1)\,, \end{aligned}$$

according to the construction from (34).

[6]With abuse of notation, in Step 2, when referring for some m to (26) or other expressions in M, we mean the corresponding restrictions as if $M = m$.

For the induction step we assume that the statement holds for '$m-1$' and we need to prove it for 'm'. Because Q_j's satisfy Step 1 for $j = 1, \ldots, m-1$ and the random variables $V^{(1)}, \ldots, V^{(m-1)}$, we have that

$$\gamma_{m-1}(v_1, \ldots, v_{m-1}) := \mathbb{1}_{\{(v_1,\ldots,v_{m-1}) \in [0,\infty]^{m-1}\}} \left[1 - \sum_{j=1}^{m-1} Q_j(v_j) e^{-\theta_j v_j}\right]$$

satisfies (26) for all $(v_1, \ldots, v_{m-1}) \in \text{supp}\,(\gamma_{m-1} \vee 0)$. Noting that

$$\text{supp}\,(\gamma_m \vee 0) := \{(v_1, \ldots, v_m) \in [0,\infty]^m : 1 > \sum_{j=1}^{m} Q_j(v_j) e^{-\theta_j v_j}\}$$

it follows that if $(v_1, \ldots, v_m) \in \text{supp}\,(\gamma_m \vee 0)$ then $(v_1, \ldots, v_{m-1}) \in \text{supp}\,(\gamma_{m-1} \vee 0)$. For all $(v_1, \ldots, v_m) \in \text{supp}\,(\gamma_m \vee 0)$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{v_1 \geq Y^{(1)}\}} \gamma_m\left(\bigwedge_{1\leq i\leq 2}\left(v_i - V^{(i)}\right), \ldots, \bigwedge_{m-1\leq i\leq m}\left(v_i - V^{(i)}\right), v_m - V^{(m)}\right)\right] \\
&= \mathbb{E}\Bigg[\mathbb{1}_{\{v_1 \geq Y^{(1)}, \forall 2\leq i\leq m: v_i \geq V^{(i)}\}} \\
&\quad \Bigg\{1 - \sum_{j=1}^{m-1} Q_j\left(\bigwedge_{j\leq i\leq j+1}\left(v_i - V^{(i)}\right)\right) e^{-\theta_j\left(\wedge_{j\leq i\leq j+1}\left(v_i - V^{(i)}\right)\right)} \\
&\quad - Q_m\left(v_m - V^{(m)}\right) e^{-\theta_m\left(v_m - V^{(m)}\right)}\Bigg\}\Bigg]
\end{aligned}$$

Since $Q_{m-1}(x) \geq 0\ \forall x \geq 0$, by construction, we can continue

$$\begin{aligned}
&\geq \mathbb{E}\Bigg[\mathbb{1}_{\{v_1 \geq Y^{(1)}, \forall 2\leq i\leq m: v_i \geq V^{(i)}\}} \\
&\quad \Bigg\{1 - \sum_{j=1}^{m-1} Q_j\left(\bigwedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i - V^{(i)}\right)\right) e^{-\theta_j\left(\wedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i - V^{(i)}\right)\right)} \\
&\quad - \sum_{j=m-1}^{m} Q_j\left(v_m - V^{(m)}\right) e^{-\theta_j\left(v_m - V^{(m)}\right)}\Bigg\}\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{1}_{\{v_1 \geq Y^{(1)}, v_m < V^{(m)}, \forall 2\leq i\leq m-1, v_i \geq V^{(i)}\}} \\
&\quad \Bigg\{-1 + \sum_{j=1}^{m-1} Q_j\left(\bigwedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i - V^{(i)}\right)\right) e^{-\theta_j\left(\wedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i - V^{(i)}\right)\right)}\Bigg\}\Bigg]
\end{aligned}$$

$$+\mathbb{E}\Bigg[\mathbb{1}_{\{v_1\geq Y^{(1)},\forall 2\leq i\leq m-1,v_i\geq V^{(i)}\}} \left\{1-\sum_{j=1}^{m-1}Q_j\Bigg(\bigwedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i-V^{(i)}\right)\Bigg)e^{-\theta_j\left(\wedge_{j\leq i\leq (j+1)\wedge(m-1)}\left(v_i-V^{(i)}\right)\right)}\right\}\Bigg]$$

$$-\mathbb{E}\left[\mathbb{1}_{\{v_m\geq V^{(m)}\}}\sum_{j=m-1}^{m}Q_j\left(v_m-V^{(m)}\right)e^{-\theta_j\left(v_m-V^{(m)}\right)}\right]$$

Using the positivity of the Q_j's for the first expectation and the induction hypothesis for the second we can continue

$$\geq -\mathbb{P}\left(v_m<V^{(m)}\right)+\left(1-\sum_{j=1}^{m-1}Q_j(v_j)e^{-\theta_j v_j}\right)$$

$$-\mathbb{E}\left[\mathbb{1}_{\{v_m\geq V^{(m)}\}}\sum_{j=m-1}^{m}Q_j\left(v_m-V^{(m)}\right)e^{-\theta_j\left(v_m-V^{(m)}\right)}\right]$$

$$\geq 1-\sum_{j=1}^{m}Q_j(v_j)e^{-\theta_j v_j}=\gamma_m(v_1,\ldots,v_m)\ .$$

In the last inequality we applied Step 1 on the event $v_m\geq V^{(m)}$.

Finally, the remaining conditions from part (b) of Theorem 5, i.e., γ is bounded and $\gamma(\infty,\ldots,\infty)=1$ hold trivially. □

C Semi-Continuous Functions

In our proofs we need two results from analysis. Let $\mathcal{D}$ be a compact subset of $\bar{\mathbb{R}}^M$, $\phi:\mathcal{D}\to\bar{\mathbb{R}}$ be an arbitrary function, and define $f:\mathcal{D}\to\mathbb{R}$

$$f(x):=\limsup_{y\to x}\phi(y)\ \forall x\in\mathcal{D}\ .$$

1. The first result is that f is an upper semi-continuous function, i.e., for every convergent sequence $x_n\to x$ in $\mathcal{D}$,

$$f(x)\geq\limsup_{x_n\to x}f(x_n)\ .$$

Indeed, from the definition of $f(x_n)$, there exists $y_n\in\mathcal{D}$ such that $|y_n-x_n|<n^{-1}$ and $\phi(y_n)\geq f(x_n)-n^{-1}$. Since $x_n\to x$, we get $|y_n-x|\leq|y_n-x_n|+|x_n-x|\to 0$ as $n\to\infty$, and hence $y_n\to x$ as well. From $\phi(y_n)\geq f(x_n)-n^{-1}$ it then follows that

$$f(x)=\limsup_{y\to x}\phi(y)\geq\limsup_{n\to\infty}\phi(y_n)\geq\limsup_{n\to\infty}f(x_n)\ ,$$

thus proving that f is upper semi-continuous.

2. The second result is that any upper semi-continuous function on compact domain attains its maximum, according to Weierstrass Maximum Theorem (Borden [6], p. 40). In particular, f attains its maximum, i.e., if $K:=\sup_{x\in\mathcal{D}}f(x)$ then

$$\mathcal{K}:=\{x\in\mathcal{D}:f(x)=K\}$$

is non-empty and closed. Furthermore,

$$a_1 := \inf\{x_1 \in \bar{\mathbb{R}} : \exists (x_2, \ldots, x_M) \in \bar{\mathbb{R}}^{M-1} : (x_1, \ldots, x_M) \in \mathcal{K}\}$$
$$a_j := \inf\{x_j \in \bar{\mathbb{R}} : \exists (x_j, \ldots, x_M) \in \bar{\mathbb{R}}^{M-j+1} : (a_1, \ldots, a_{j-1}, x_j, \ldots, x_M) \in \mathcal{K}\}$$

are well defined and $a := (a_1, \ldots, a_M) \in \mathcal{K}$. Moreover, if $K > 0$, then there exists a sequence $x_n \to a$ such that $\phi(x_n) > 0$ and $x_n \in \text{supp}\,(\phi \vee 0)$.

Note the standard notation for a function's support, i.e.,

$$\text{supp}\,(\phi) := \{x \in \mathcal{D} : \phi(x) \neq 0\} \text{ and } \text{supp}\,(\phi \vee 0) := \{x \in \mathcal{D} : \phi(x) > 0\}\ .$$

D Light-Tailed Distributions

Let a random variable Y with density $f_Y(y) = \alpha \mathbb{1}_{y \geq 0} e^{-\mu y}/(1 + y^2)$, for $\mu > 0$ and a suitable value of α. Let also a random variable X independent of Y and satisfying $X > \mathbb{E}[Y] \vee (\ln(\alpha\pi/2)/\mu)$ a.s. Then

$$\theta^+ := \sup\{\theta \in \mathbb{R} : \mathbb{E}[e^{\theta(Y-X)}] < \infty\} = \mu$$

and

$$\mathbb{E}[e^{\theta^+(Y-X)}] = \mathbb{E}[e^{\mu(Y-X)}] = \alpha \mathbb{E}[e^{-\mu X}] \int_0^\infty \frac{1}{1+x^2} dx = \alpha\pi \mathbb{E}[e^{-\mu X}]/2 < 1\ .$$

We note that the distribution of Y belongs to the class of Type-II distributions (see Abate and Whitt [1]).