TAIL ASYMPTOTICS FOR DELAY IN A HALF-LOADED GI/GI/2 QUEUE WITH HEAVY-TAILED JOB SIZES

JOSE BLANCHET AND KARTHYEK MURTHY

Abstract. We obtain asymptotic bounds for the tail distribution of steady-state waiting time in a two server queue where each server processes incoming jobs at a rate equal to the rate of their arrivals (that is, the half-loaded regime). The job sizes are taken to be regularly varying. When the incoming jobs have finite variance, there are basically two types of effects that dominate the tail asymptotics. While the quantitative distinction between these two manifests itself only in the slowly varying components, the two effects arise from qualitatively very different phenomena (arrival of one extremely big job (or) two big jobs). Then there is a phase transition that occurs when the incoming jobs have infinite variance. In that case, only one of these effects dominate the tail asymptotics, the one involving arrival of one extremely big job.

1. Introduction

The tail behaviour of the distribution of steady-state delay in multiserver queues processing jobs with heavy-tailed sizes has attracted substantial attention in stochastic operations research. Most of the literature has focused on the case in which the traffic intensity, ρ, (that is, the ratio between the mean service requirement and the mean interarrival time) is not an integer and there are qualitative reasons, as we shall discuss, that make the integer case significantly more delicate to analyze. Our contribution in this work is to provide the first asymptotic upper and lower bounds for the tail distribution, that match up to a constant factor, for the integer case. In that process, we identify the occurrence of a few surprising phenomena that are not common in the asymptotic analysis of multiserver queues. We concentrate on the two server queue because it provides a vehicle to study the qualitative phenomenon that is of interest to us.

As mentioned earlier, most of the literature concentrates on the case in which ρ is not an integer. A series of conjectures relating tail distribution of steady-state delay to the traffic intensity has been made in [22]. These conjectures turned out to be basically correct for the case of regularly varying job sizes and were verified for the case of a two-server queue in [11], where more general asymptotic bounds for subexponential distributions are provided. In [12], the authors provide bounds (up to constants) that verify the conjecture in [22] for general multiserver queues with regularly varying job sizes and non-integer traffic intensity. There is a related body of literature aimed at studying stability properties, such as the existence of the mean steady-state delay, in terms of the traffic intensity of the system and tail properties of the incoming traffic. The relations found in this literature, see [20] and [21], again are also derived only for the case of non-integer traffic intensity and are consistent with the relations found for the tail distributions mentioned earlier (which can be used to derive the existence of moments).

In order to discuss our contributions in more detail, let us introduce some notation. Let V denote the amount of time required to service a generic job arriving to the queue and let $\bar{B}(x) = \mathbb{P}\{V > x\}$. We assume that $\bar{B}(\cdot)$ is regularly varying with index $\alpha > 1$, that is,

$$\bar{B}(x) = x^{-\alpha} L(x),$$

for some function $L(\cdot)$ satisfying $\lim_{x\to\infty} L(tx)/L(x) = 1$ for each $t > 0$; such a function $L(\cdot)$ is said to be slowly varying. Jobs are assumed to arrive as a Poisson stream (or, more generally, a renewal stream) with rate equal to $\mathbb{E}V$ and service requirements that are identical copies of V. Under this setting, the traffic intensity ρ equals 1. Let us write W to denote the steady-state waiting time of the two-server queue that processes jobs according to FCFS (first-come-first-serve) discipline. Our first result establishes that if $\alpha > 2$, then

$$\mathbb{P}\{W > b\} = \Theta\left(b^2\bar{B}(b^2) + b^2\bar{B}^2(b)\right), \tag{1}$$

as $b \to \infty$. Here recall that $f(b) = \Theta(g(b))$ if and only if $f(b) \le c_1 g(b)$ and $g(b) \le c_2 f(b)$ for some positive constants c_1 and c_2 that are independent of b. To get a sense of how subtle the difference between the terms appearing in (1) are, it is instructive to consider the example $L(x) = \log(1+x)$, where the second term appearing in the right hand side of (1) dominates the asymptotic behaviour. On the other hand, if $L(x) = 1/\log(1+x)$, the first term in the right hand side of (1) dominates the asymptotic behaviour. Finally, if $L(x) \sim c$ for some $c > 0$ (the asymptotically Pareto case) both terms contribute substantially.

Further, let us contrast the result in (1) with that derived in [11]. For the case $\rho < 1$, it was found that

$$\mathbb{P}\{W > b\} = \Theta\left(b^2\bar{B}^2(b)\right), \tag{2}$$

whereas for the case $\rho \in (1,2)$, [11] obtained that

$$\mathbb{P}\{W > b\} = \Theta\left(b\bar{B}(b)\right), \tag{3}$$

as $b \to \infty$[1]. Since there is a sharp difference between the cases $\rho < 1$ and $\rho \in (1,2)$ as in (2) and (3), it has been of great interest to identify what happens when ρ equals 1. We resolve this in our work by noting that (1) is much closer to the case $\rho < 1$ than it is to the case $\rho > 1$. Although, quantitatively, the rates of convergence between the two terms in (1) might differ only by a multiplicative function which varies slowly, the qualitative picture behind the mechanism that gives rise to them is dramatically different. The first term in the right hand side of (1) arises from the same type of phenomena behind the tail behaviour in the case $\rho < 1$.

In Section 2, in addition to introducing the notation required to precisely state our results, we discuss at length the intuition behind both the asymptotic results (2) and (3), as well as our asymptotic expression (1). At this point, it suffices to say that the phenomena underlying the development of (2) and (3) are a combination of two features, first, arrival of large jobs whose effects persist for long time scales, and, second, the impact of such effects, which is measured using the Law of Large Numbers. In contrast, the development of (1) involves not only the combination of these two features, but, in addition, one has to account for the impact of effects which occur at the scales governed by the Central Limit Theorem.

We identify another interesting phenomenon when the job sizes have infinite variance: If $\rho = 1$ and $\alpha \in (1,2)$, it turns out that the asymptotics are governed by

$$\mathbb{P}\{W > b\} = \Theta\left(b^{\alpha}\bar{B}(b^{\alpha})\right),$$

suggesting that the tail behaviour is closer to the case $\rho > 1$ than to the case $\rho < 1$. This is a sharp transition from the system behaviour when $\mathrm{Var}[V] < \infty$, where the tail asymptotic is closer to the $\rho < 1$ case. Such surprising transitions in system behaviour seem to be unique to the integer traffic intensity case.

In summary, the qualitative development behind our asymptotic bounds introduces a combination of elements that are not typical in the asymptotic analysis of multiserver queues. After developing necessary intuition behind the results (1), (2) and (3) in Section 2, we derive the

[1]From here on, we avoid the quantification $b \to \infty$ whenever it is evident from the context

respective lower and upper bounds in (1) in Sections 3 and 4. Apart from unraveling surprising transitions in the system behaviour that seem to happen only when the traffic intensity is an integer, an important contribution of this paper is in the use of regenerative ratio representation and Lyapunov bound techniques to characterize tail behaviour of steady-state delay in multi-server queues. An alternate proof for the upper bound, that takes inspiration from a completely different approach due to [11] and [12], is reported in [16]. However, in [11] and [12], it is crucial to have ρ not equal to an integer so that certain upper bound processes might be defined. So, we believe that our alternate approach presented in [16] might add useful ideas to the traditional techniques used in the asymptotic analysis of multiserver queues.

2. The main result and its intuition

We consider a two-server queue that processes incoming jobs under the first-come-first-serve discipline. Jobs are indexed by the order of arrival. Job 0 arrives at time 0, and for $n \geq 1$, job n arrives at time $T_1 + \ldots + T_n$. Job n requires service for time V_n. Here the sequence of interarrival times $(T_n : n \geq 1)$ and service times $(V_n : n \geq 0)$ are taken to be i.i.d. copies, respectively, of the generic interarrival and service time variables T and V. As mentioned in the Introduction, we assume that $\mathbb{E}T = \mathbb{E}V$, and hence the traffic intensity ρ, which is the ratio between $\mathbb{E}V$ and $\mathbb{E}T$, equals 1. To make the computations easier, we assume, without loss of generality, that $\mathbb{E}T = 1$ (otherwise, time can always be rescaled to make this hold). Additionally, we make the following assumptions on the distributions of V and T.

Assumption 1. *The tail distribution of V admits the representation,*

$$\bar{B}(x) := \mathbb{P}\{V > x\} = x^{-\alpha} L(x),$$

for some $\alpha > 1$ and a function $L(\cdot)$ slowly varying at infinity, that is, $\lim_{x\to\infty} L(tx)/L(x) = 1$ for every $t > 0$.

Assumption 2. $\mathbb{P}\{T > x\} = o(\bar{B}(x))$.

Assumption 2 is quite natural given that typically one models interarrival times as exponentially distributed random variables. We also use the notation

$$X_{n+1} = V_n - T_{n+1} \text{ for } n \geq 0.$$

Since T is non-negative, the right-tail of $X := V - T$ is asymptotically similar to that of V (see, for example, Corollary 1.11 in Chapter IX of [1]). In other words,

$$\mathbb{P}\{X > x\} \sim \bar{B}(x) \text{ as } x \to \infty. \tag{4}$$

The ordered workload vector of the servers as seen by the n^{th} job during its arrival, denoted by $\mathbf{W}_n = (W_n^{(1)}, W_n^{(2)})$, satisfies the well-known Kiefer-Wolfowitz recursion:

$$W_{n+1}^{(1)} = \left(W_n^{(1)} + V_n - T_{n+1}\right)^+ \wedge \left(W_n^{(2)} - T_{n+1}\right)^+ \text{ and} \tag{5a}$$

$$W_{n+1}^{(2)} = \left(W_n^{(1)} + V_n - T_{n+1}\right)^+ \vee \left(W_n^{(2)} - T_{n+1}\right)^+. \tag{5b}$$

Since $\rho < 2$, the queue is stable in the sense that the weak limit (limit in distribution) of $\mathbf{W}_n$, denoted by $\mathbf{W}_\infty$, exists and we are interested in deriving bounds for the tail probabilities of the steady-state waiting time

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b\right\},$$

for large values of b. Our main result is the following.

Theorem 1. *Suppose that $\rho = 1$ and Assumptions 1 and 2 are in force. If $\alpha > 2$, then*

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \Theta\left(b^2\bar{B}(b^2) + b^2\bar{B}^2(b)\right), \text{ as } b \to \infty. \tag{6}$$

If $\alpha \in (1,2)$, under the additional assumption that $\bar{B}(x) \sim cx^{-\alpha}$ for some $c > 0$, we have that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \Theta\left(b^\alpha \bar{B}(b^\alpha)\right), \text{ as } b \to \infty. \tag{7}$$

We now proceed to discuss how this result contrasts with what is known in the literature and thereby expose the intuition behind it.

2.1. **Discussion of earlier results in the literature.** As indicated in the Introduction, the tail asymptotics of steady-state delay is known depending on the case $\rho < 1$ (or) $\rho \in (1,2)$, and is given by (2) and (3), respectively. In order to see the mechanism behind these two asymptotics, let us assume without loss of generality that $\mathbb{E}T = 1$ (if not, time can be rescaled to make this assumption hold). Additionally, let us assume that the generic interarrival time T has unbounded support (for example, T is exponentially distributed), and consider the regenerative ratio representation

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \frac{\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right]}{\mathbb{E}_{\mathbf{0}}\left(\tau_0\right)}, \tag{8}$$

where $\tau_0 = \inf\{n \geq 1 : W_n^{(2)} = 0\}$ denotes the first time when the Kiefer-Wolfowitz process $\mathbf{W}_n$ enters the set $\{(0,0)\}$. Since $\rho < 2$ and T has unbounded support, the state $(0,0)$ is recurrent, thus leading to the regenerative ratio representation (8) For simplicity, throughout our discussions, we shall assume that T has unbounded support. This assumption is merely technical. It can be relaxed at the price of using a slightly more complicated regenerative representation. For further details on the representation (8) and details on relaxing the assumption on support of T, see, for example, [4], [9], [13], [14], or [10]. Moreover, our alternate proof of the upper bound presented in [16] does not rely on this assumption.

In order to study (8), define the stopping times

$$\tau_b^{(i)} := \inf\{n \geq 0 : W_n^{(i)} > b\}, \quad i = 1, 2.$$

First, let us consider the event $\{\tau_b^{(1)} < \tau_0\}$, which is the event that there is at least one customer who waits more than b units of time in a busy period. Moreover, since $\tau_b^{(2)} < \tau_b^{(1)}$, it is instructive to first consider the event $\{\tau_b^{(2)} < \tau_0\}$, which can be seen, intuitively, to be caused by the arrival of a big job of size larger than b within the initial $O(1)$ units of time in the busy period. Due to this reasoning, one can write

$$\mathbb{P}_{\mathbf{0}}\left\{\tau_b^{(2)} < \tau_0\right\} = \Theta\left(\mathbb{P}\{V > b\}\right) \text{ and } \mathbb{P}\left\{W_{\tau_b^{(2)}}^{(2)} > x \,\middle|\, \tau_b^{(2)} < \tau_0\right\} \approx \mathbb{P}\left\{V > x \,\middle|\, V > b\right\}. \tag{9}$$

Therefore, one can approximately characterize the process $\mathbf{W}$, immediately after the arrival of the first big job of size larger than b, as below:

$$\frac{1}{b}\mathbf{W}_{\tau_b^{(2)}} = \frac{1}{b}\left(W_{\tau_b^{(2)}}^{(1)}, W_{\tau_b^{(2)}}^{(2)}\right) \approx (0, Z), \tag{10}$$

where Z satisfies $\mathbb{P}\{Z > x\} = \lim_{b\to\infty}\mathbb{P}\left\{V > bx \,\middle|\, V > b\right\} = x^{-\alpha}$ for $x \geq 1$. As per recursions (5a) and (5b), the server that gets to process this big job cannot process any new arrivals until both the workloads become comparable again at some time in the future, which we refer as τ_{eq}.

During this period where one of the servers is effectively blocked from processing new arrivals (call it the blocked server and the other server as active server), the dynamics of the queue is given by:

$$\mathbf{W}_n = \left(\left(W_{n-1}^{(1)} + V_{n-1} - T_n \right)^+, W_{n-1}^{(2)} - T_n \right), \quad \tau_b^{(2)} < n < \tau_{eq}.$$

The dynamics of the active server matches with that of the single server queue, and hence the waiting time experienced by the k^{th} job after the big jump can be be roughly approximated, in distribution, by maximum of k steps of a random walk with increments that are i.i.d. copies of $V - T$. Observe that the aforementioned random walk has drift equal to $\rho - 1$, which can be positive, zero (or) negative, respectively, based on whether $\rho > 1, \rho = 1$ (or) $\rho < 1$. As a consequence, the maximum of the random walk, in the respective cases, can be of magnitude $O(k), O(\sqrt{k})$ (or) $O(1)$ in k units of time (this can be seen by invoking Law of Large Numbers and Central Limit Theorem for i.i.d. sums). Therefore, due to (10), the workload until time τ_{eq} can be approximately written as

$$W_{\tau_b^{(2)}+k}^{(1)} \approx \begin{cases} c_1 k & \text{if } \rho > 1, \\ c_2\sqrt{k} & \text{if } \rho = 1, \\ O(1) & \text{if } \rho < 1 \end{cases} \qquad \text{and } W_{\tau_b^{(2)}+k}^{(2)} \approx bZ - k \tag{11}$$

for some positive constants c_1 and c_2. Because of this clear difference in behaviour of $W^{(1)}$ based on the value of ρ, we need to consider cases $\rho \in (1, 2), \rho < 1$ and $\rho = 1$ separately. We once again stress that our discussion in this section is completely heuristic, aiming to emphasize the intuition behind the results. While cases $\rho \in (1, 2)$ and $\rho < 1$ are treated rigorously in [11], future sections in this paper are devoted to the rigorous treatment of the case $\rho = 1$.

2.1.1. *Case 1: $\rho \in (1, 2)$.* If $\rho \in (1, 2)$, then one server is not enough to keep the system stable. As a result, when one server is blocked for $O(bZ)$ units of time due to the arrival of a big job, the active server effectively becomes a single server processing all the arrivals, and hence the workload $W^{(1)}$ gradually increases with time as in (11). Recall that τ_{eq} is the time where both the servers have roughly equal workload, and therefore due to (11), we solve for τ_{eq} by setting

$$c_1 \left(\tau_{eq} - \tau_b^{(2)} \right) \approx bZ - \left(\tau_{eq} - \tau_b^{(2)} \right).$$

As a result, $W^{(1)}$ increases roughly up to time

$$\tau_{eq} \approx \tau_b^{(2)} + \frac{bZ}{c_1 + 1},$$

when both $W^{(1)}$ and $W^{(2)}$ become comparable, after which both the servers jointly process incoming arrivals according to (5a) and (5b), resulting in a total decrease of workload at rate $2 - \rho$. In this mechanism, for any job to be delayed by more than b units of time, it must happen that $c_1 k \geq b$ for some $k \leq bZ/(c_1 + 1)$, and therefore,

$$\lim_{b\to\infty} \mathbb{P}_{\mathbf{0}} \left\{ \tau_b^{(1)} < \tau_0 \;\middle|\; \tau_b^{(2)} < \tau_0 \right\} = \lim_{b\to\infty} \mathbb{P} \left\{ c_1 \frac{bZ}{c_1 + 1} \geq b \right\} = \mathbb{P} \left\{ Z > 1 + \frac{1}{c_1} \right\} > 0. \tag{12}$$

If we let N_1 to denote the number of jobs that experience at least b units of delay up to time τ_{eq} and N_2 to denote the respective count after τ_{eq}, then the above heuristics suggest that

$$N_1 = \frac{\left(W^{(1)}_{\tau_{eq}} - b\right)^+}{c_1} = \left(\frac{bZ}{c_1+1} - \frac{b}{c_1}\right)^+ \text{ and}$$
$$N_2 = \frac{\left(W^{(1)}_{\tau_{eq}} - b\right)^+}{2-\rho} = \frac{1}{2-\rho}\left(\frac{c_1 bZ}{c_1+1} - b\right)^+.$$

Therefore, due to (12), we obtain that

$$\begin{aligned}\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right)\right] &= \mathbb{E}_0\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right) \;\middle|\; \tau^{(1)}_b < \tau_0\right] \times \mathbb{P}\left\{\tau^{(1)}_b < \tau_0\right\} \\ &\approx \mathbb{E}\left[N_1 + N_2 \;\middle|\; Z > 1 + \frac{1}{c_1}\right] \times \mathbb{P}\left\{Z > 1 + \frac{1}{c_1}\right\} \times \Theta\left(\mathbb{P}\{V > b\}\right) \\ &= \Theta\left(b \times \mathbb{P}\left\{Z > 1 + \frac{1}{r}\right\} \times \bar{B}(b)\right).\end{aligned}$$

As a result, from (8), we obtain that

$$\mathbb{P}\left\{W^{(1)}_\infty > b\right\} = \Theta\left(b\bar{B}(b)\right),$$

which is precisely same as (3). This final form of asymptotic is rigorously established in [11], albeit, using a different reasoning.

2.1.2. *Case 2:* $\rho < 1$. If $\rho < 1$, conditional on the occurrence of $\{\tau^{(2)}_b < \tau_0\}$, it is no longer true that the event $\{\tau^{(1)}_b < \tau_0\}$ happens with positive probability as $b \to \infty$ (compare this with (12) when $\rho \in (1,2)$). The reason is that if $\rho < 1$, the system is stable and the workload remains $O(1)$, as in (11), even if one removes one server and force it to operate as a single server system. As a result, we need to invoke heavy-tailed large deviations behaviour, which dictates that arrival of one more job of size larger than b is required, typically, to experience waiting time larger than b. This requirement is dealt as follows: Conditional on the occurrence of $\{\tau^{(2)}_b < \tau_0\}$, as in (11), we have

$$W^{(1)}_{\tau^{(2)}_b + k} = O(1) \text{ and } W^{(2)}_{\tau^{(2)}_b + k} \approx bZ - k.$$

Here, the workload $W^{(2)}$ becomes smaller than b if $k > b(Z-1)$, and therefore, the cheapest way to observe large delays (of duration at least b) is to have a $K \leq b(Z-1)$ such that the $(\tau^{(2)}_b + K)^{th}$ job requires service for duration larger than b. Following the same line of reasoning behind (10), we approximate the size of the second big job by $b\hat{Z}$, where $\hat{Z}$ is an independent copy of Z. As a result, we arrive at the following distributional approximation :

$$\mathbf{W}^{(1)}_{\tau^{(1)}_b} \approx \min\left(bZ - K_1, b\hat{Z}\right) \text{ and } \mathbf{W}^{(2)}_{\tau^{(1)}_b} \approx \max\left(bZ - K_1, b\hat{Z}\right).$$

Next, the number of jobs that get delayed by more than b units of time (which depends on K) is approximately given by

$$N(K) := \frac{\min(bZ - K, b\hat{Z}) - b}{2-\rho},$$

where $K \le b(Z-1)$. As a result,

$$\begin{aligned}
\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)}>b\right)\right] &\approx \mathbb{E}\left[N(K)\, I\left(0\le K\le b(Z-1)\right) \;\middle|\; \tau_b^{(2)}<\tau_0\right]\times \mathbb{P}\left\{\tau_b^{(2)}<\tau_0\right\}\\
&= \mathbb{E}\left[\sum_{k=1}^{b(Z-1)} \min\left(b(Z-1)-k, b(\hat{Z}-1)\right) I\left(V_{\tau_b^{(2)}+k}>b\right) \;\middle|\; \tau_b^{(2)}<\tau_0\right]\times\Theta\left(\mathbb{P}\{V>b\}\right)\\
&= \Theta\left(b^2\mathbb{P}\left\{V>b\right\}^2\right),
\end{aligned}$$

and therefore, $\mathbb{P}\{W_\infty^{(1)}>b\}=\Theta(b^2\bar{B}^2(b))$, which coincides with (2).

2.2. **Intuitive discussion of Theorem 1: The case $\rho = 1$.** Our goal in this discussion is to communicate the following insights:

1) Contrary to Case 1 and Case 2, the conditional distribution of the Kiefer-Wolfowitz vector $\mathbf{W}$ given that $\{\tau_b^{(1)}<\tau_0\}$ does not fully explain the mechanism behind the asymptotic results in Theorem 1.
2) Unlike Cases 1 and 2, it is not enough to account for the impact of the large service times using linear dynamics which evolve according to the Law of Large Numbers.

We shall first concentrate on the situation where the job sizes V have finite variance, more precisely, the case $\alpha>2$. The case $\alpha\in(1,2)$ can be understood using similar ideas. We shall leverage off the type of arguments that were given for Case 1 and Case 2. Since $\rho=1$ sits right in the middle we shall consider two mechanisms, one involving two jumps (analogous to Case 2), and one involving one jump (analogous to Case 1).

Delays due to two jumps: Conditional on $\{\tau_b^{(2)}<\tau_0\}$, similar to cases 1 and 2, the dynamics of the active server and the blocked server, as in (11), are given respectively by

$$W^{(1)}_{\tau_b^{(2)}+k}\approx c_2\sqrt{k} \text{ and } W_{\tau_b^{(2)}+k}\approx bZ-k, \tag{13}$$

for k such that $\tau_b^{(2)}+k\le\tau_{eq}$. As discussed previously, fluctuations of order $\sqrt{k}$ arise in workload due to the Central Limit Theorem, and this phenomenon, as we shall see below, gains relevance only when $\rho=1$. Since our interest here is in studying delays due to the occurrence of two big jumps, as in Case 2, if there exists a $K<b(Z-1)$ such that $(K+\tau_b^{(2)})$-th customer brings a job of size $b-O(\sqrt{b})$ or larger, then at least one job gets delayed by b units or more. The contribution to $\mathbb{P}\{\tau_b^{(1)}<\tau_0\}$ due to the occurrence of 2 jumps can be calculated as below:

(14)

$$\begin{aligned}
P_{2\text{ jumps}}(b) &:= \mathbb{P}\left\{\tau_b^{(2)}<\tau_0\right\}\times\Theta\left(\mathbb{P}\left\{V_{\tau_b^{(2)}+k}>b-\sqrt{b}\text{ for some } k\le b(Z-1) \;\middle|\; \tau_b^{(2)}<\tau_0\right\}\right)\\
&=\Theta\left(\mathbb{P}\left\{V>b\right\}\times\sum_{k=1}^{\infty}\mathbb{P}\left\{bZ>k, V_k>b-\sqrt{b}\right\}\right)\\
&=\Theta\left(\mathbb{P}\left\{V>b\right\}^2\times\sum_{k=1}^{\infty}\mathbb{P}\left(bZ>k\right)\right)=\Theta\left(b\bar{B}^2(b)\right).
\end{aligned}$$

Following the same line of reasoning as in Case 2, we obtain the following contribution to $\mathbb{E}_{\mathbf{0}}[\sum_{k=0}^{\tau_0-1} I(W_k^{(1)} > b)]$ due to 2 jumps:

$$Q_{2\text{ jumps}}(b) = \Theta\left(b^2\bar{B}^2(b)\right). \tag{15}$$

Delay due to 1 jump: Similar to Case 1, when $\rho = 1$, the active server accumulates work, albeit at a slower rate, as given in (13). Since the workload of a critically loaded single server queue grows like $O(\sqrt{k})$ in k units of time, it is intuitive to expect that if there is a big jump of size exceeding b^2 in the first $O(1)$ units of time of the busy period, subsequently one of the servers gets blocked for more than b^2 units of time, and the active server which faces all the incoming traffic accumulates workload of size larger than b, with non-vanishing probability, in those b^2 units of time. Therefore, similar to (12), we have that

$$\lim_{b\to\infty} \mathbb{P}\left\{\tau_b^{(1)} < \tau_0 \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right\} > 0.$$

Therefore, due to (9), the contribution to $\mathbb{P}\{\tau_b^{(1)} < \tau_0\}$ due to the arrival of only one big job is given by

$$P_{1\text{ jump}}(b) \approx \mathbb{P}\left\{\tau_b^{(1)} < \tau_0 \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right\} \times \mathbb{P}\left\{\tau_{b^2} < \tau_0\right\} = \Theta\left(\bar{B}(b^2)\right),$$

which is negligible compared to the right hand side of (14). As a result, we have that

$$\mathbb{P}\left\{\tau_b^{(1)} < \tau_0\right\} \sim P_{2\text{ jumps}}(b) = \Theta\left(b\bar{B}^2(b)\right).$$

However, accounting for the number of jobs that experience at least b units of delay dramatically changes the contribution of this single huge jump in the computation of steady-state delay probabilities. In particular, a single jump of size exceeding b^2 blocks one of the servers for $V \mid V > b^2 \approx b^2 Z$ units of time, and if we perform calculations similar to Case 1, we shall obtain that $\Theta(b^2)$ jobs experience delays larger than b. As a consequence, we have the following contribution in the single, huge jump regime:

$$Q_{1\text{ jump}} := \mathbb{E}\left[\sum_{i=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right] \times \mathbb{P}\left\{\tau_b^{(2)} < \tau_0\right\} = \Theta\left(b^2\mathbb{P}\left\{V > b^2\right\}\right),$$

which might not be negligible to the corresponding contribution due to 2 jumps derived in (15). In fact, as demonstrated in an example in the Introduction, this contribution due to single huge jump could be larger than its counterpart for 2 jumps based on the slowly varying function $L(\cdot)$ (consider the example $L(x) = \log(1+x)$). As a result, we have two competing components in the expression for steady-state probability of delay in (6).

We conclude with a heuristic explanation of the mechanism involving one jump for the case $\alpha \in (1,2)$ if $\rho = 1$. In this case, once a server is blocked for k units of time, the active server operates as a critical single-server queue, processing jobs requiring services with infinite variance, and due to the generalized Central Limit Theorem, the workload of the critical queue exhibits fluctuations of order $O(k^{1/\alpha})$. Therefore, if the initial huge jump, which occurs within $O(1)$ units of time at the beginning of the busy period, is of size larger than b^α, then this huge job blocks one of the servers for more than b^α units of time, and as a result, $\Theta(b^\alpha)$ jobs wait for a duration larger than b. Reasoning as in the finite variance case, the contribution to steady-state delay due to the arrival of one huge job is $\Theta(b^\alpha\mathbb{P}\{V > b^\alpha\}) = \Theta(b^\alpha b^{-\alpha^2}L(b^\alpha))$. On the other hand, the contribution arising from two jumps as in Case 2, namely, according to (2), remains $\Theta(b^{2-2\alpha}L(b)^2)$, which is negligible compared to $\Theta(b^\alpha b^{-\alpha^2}L(b^\alpha))$ because $\alpha \in (1,2)$ implies $2\alpha - 2 > \alpha^2 - \alpha$. Hence, we arrive at the estimate (7) in Theorem 1.

3. Proof of lower bound

The objective of this section is to prove the following result.

Proposition 1. *Suppose that Assumption 2 holds, and that $\rho = 1$. Then, if Assumption 1 holds with $\alpha > 2$, there exists $c_1 > 0$ and $b_0 > 0$ such that for all $b > b_0$.*

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq c_1 \left(b^2 \bar{B}(b^2) + b^2 \bar{B}^2(b)\right).$$

On the other hand, if $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$ for some $c > 0$ and $\alpha \in (1, 2)$, then there exists $c_1 > 0$ and $b_0 > 0$ such that for all $b > b_0$,

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq c_1 \left(b^\alpha \bar{B}(b^\alpha)\right).$$

We now provide the proof of Proposition 1.

Case 1: (Under the assumption that $\alpha > 2$). We first derive a lower bound based on a single big jump of size exceeding b^2. Let $N_A(t)$ denote the number of jobs that arrive in the interval $(0, t]$. Let $b > 2$ and consider the event, D_1, with the following properties:

1) The coordinate $W_1^{(2)} > 6b^2$ (that is, Job 0 blocks one of the servers for $\Omega\left(b^2\right)$ time units).
2) The total amount of work brought by all the jobs that arrive in the time interval $(0, 2b^2]$ does not exceed $3b^2$. In other words, $V_1 + \ldots + V_{N_A(2b^2)} \leq 3b^2$.
3) Every job that arrives in the time interval $[b^2, 2b^2]$ experiences delay for at least b units of time before getting processed. That is,
$$\min_{N_A(b^2) \leq n \leq N_A(2b^2)} W_n^{(1)} > b.$$

On the set D_1, the dynamics of the queue described by recursions (5a) and (5b) reduces to

$$W_n^{(1)} = (W_{n-1} + X_n)^+ \text{ and } W_n^{(2)} = W_1^{(2)} - (T_2 + \ldots + T_n)$$

for $2 \leq n \leq N_A(2b^2)$. Further, if we let $S_1 := W_1^{(1)} = 0$ and $S_n := X_2 + \ldots + X_n$ for $n \geq 2$, then the following holds on the set D_1 :

$$\min_{N_A(b^2) \leq n \leq N_A(2b^2)} W_n^{(1)} \geq \min_{N_A(b^2) \leq n \leq N_A(2b^2)} S_n.$$

As a result,

$$\begin{aligned}
\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right)\right] &\geq \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right); D_1\right] \\
&\geq \mathbb{E}_{\mathbf{0}}\left[N_A\left(2b^2\right) - N_A\left(b^2\right); W_1^{(2)} > 6b^2, \min_{N_A(b^2) \leq n \leq N_A(2b^2)} S_n > 1.5b, \sum_{i=1}^{N_A(2b^2)} V_i \leq 3b^2\right] \\
&\geq \mathbb{P}\left\{X_1 > 6b^2\right\} \mathbb{E}\left[N_A\left(2b^2\right) - N_A\left(b^2\right); D_1'\right],
\end{aligned}$$

where the event

$$D_1' := \left\{N_A\left(b^2\right) \geq 0.5b^2,\ N_A\left(2b^2\right) \in \left[1.5b^2, 2.5b^2\right], \min_{0.5b^2 \leq n \leq 2.5b^2} S_n > b, \sum_{i=1}^{\lceil 2.5b^2 \rceil} V_i \leq 3b^2\right\}$$

has probability at least

$$\mathbb{P}\left\{\inf_{0.5 \leq t \leq 2.5} \sigma B(t) > 1\right\}(1 - o(1))$$

because of functional CLT and the facts that $N_A(x)/x \to 1$ and $(V_1 + \ldots + V_n)/n \to 1$ with probability one. Here $B(\cdot)$ is a standard Brownian motion and σ^2 denotes the variance of X. Additionally, due to the regenerative ratio representation (8) and the regularly varying nature of the tail of X (recall that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$ as $x \to \infty$), we conclude that

$$\begin{aligned} \mathbb{P}\left\{W_\infty^{(1)} > b\right\} &\geq \frac{b^2\mathbb{P}\left\{X > 6b^2\right\} \times \mathbb{P}\left(D_1'\right)}{\mathbb{E}\tau_0} \\ &\geq c_1 b^2 \bar{B}\left(b^2\right) \end{aligned} \tag{16}$$

for some $c_1 > 0$ and all b large enough.

Case 2: (Also under the assumption that $\alpha > 2$). We now derive a lower bound based on the occurrence of two jumps, each of size exceeding b. Let $b > 2$ and consider the event, D_2, with the following properties:

1) The coordinate $W_1^{(2)} > 5b$ (that is, Job 0 blocks one of the servers for $\Omega(b)$ time units).
2) Apart from Job 0, only one of the $N_A(b)$ jobs that arrive in the time interval $(0, b]$ bring a service requirement of size exceeding $5b$.
3) The number of customers who arrive during the time intervals $(0, b]$ and $(b, 2b]$ are numbers between $0.5b$ and $1.5b$. Alternatively, $N_A(b) \in [0.5b, 1.5b]$ and $N_A(2b) - N_A(b) \in [0.5b, 1.5b]$.

So, on the set D_2 we have that at least $N_A(2b) - N_A(b) \geq 0.5b$ jobs experience a waiting time more than b units of time, and hence

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right); D_2\right] \geq 0.5b\mathbb{P}\left(D_2\right).$$

However, since $N_A(x)/x \to \infty$, we have that

$$\mathbb{P}\left\{N_A(b) \in [0.5b, 1.5b],\ N_A(2b) - N_A(b) \in [0.5b, 1.5b]\right\} \sim 1,$$

as $b \to \infty$. As a result,

$$\begin{aligned} \mathbb{P}\left(D_2\right) &\geq (1 - o(1)) \sum_{k \leq 0.5b} \mathbb{P}_{\mathbf{0}}\left\{W_1^{(2)} > 5b,\ V_k > 5b, \bigcap_{i \leq 1.5b, i \neq j} \{V_i < 5b\}\right\} \\ &\geq 0.5b\mathbb{P}\left\{X_1 > 5b\right\}\bar{B}(5b)\left(1 - \bar{B}(5b)\right)^{1.5b}(1 - o(1)) \\ &\geq b\bar{B}^2(5b)(1 - o(1)) \end{aligned}$$

Then, as in Case 1, due to the regenerative ratio representation (8) and the regularly varying nature of $\bar{B}(\cdot)$, we conclude that there exists a constant c_2 such that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq \frac{0.5b \times b\bar{B}^2(5b)}{\mathbb{E}\tau_0}(1 - o(1)) \geq c_1 b^2 \bar{B}(b)^2. \tag{17}$$

Combining (16) and (17) we obtain the statement of Proposition 1 for the case $\alpha > 2$.

Case 3: We now consider the assumption that $\alpha \in (1, 2)$ and $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$. The strategy is similar to Case 1. Define an event, D_3, satisfying the following properties:

1) The coordinate $W_1^{(2)} > 6b^\alpha$ (that is, Job 0 blocks one of the servers for $\Omega(b^\alpha)$ time units).
2) The total amount of work brought by all the jobs that arrive in the time interval $(0, 2b^\alpha]$ does not exceed $3b^\alpha$. In other words, $V_1 + \ldots + V_{N_A(2b^\alpha)} \leq 3b^\alpha$.

3) Every job that arrives in the time interval $[b^\alpha, 2b^\alpha]$ experiences delay for at least b units of time before getting processed. That is,
$$\min_{N_A(b^2)\le n\le N_A(2b^2)} W_n^{(1)} > b.$$

Then, following the same steps as in Case 1, we obtain that
$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=1}^{\tau_0} I\left(W_k^{(1)} > b\right)\right] \ge \bar{B}\left(6b^\alpha\right)\mathbb{E}\left[N_A\left(2b^\alpha\right) - N_A\left(b^\alpha\right); D_3'\right]$$
where the event
$$D_3' := \left\{N_A\left(b^\alpha\right) \ge 0.5b^\alpha,\ N_A\left(2b^\alpha\right) \in \left[1.5b^\alpha, 2.5b^\alpha\right], \min_{0.5b^\alpha\le n\le 2.5b^\alpha} S_n > b, \sum_{i=1}^{\lceil 2.5b^\alpha\rceil} V_i \le 3b^\alpha\right\}$$
has non-vanishing probability as $b \to \infty$ because $b^{-1}S_{[tb^\alpha]}$ converges weakly in $D[0,\infty)$, to a Stable process $Z(\cdot)$. As a result, we obtain
$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=1}^{\tau_0} I\left(W_k^{(1)} > b\right)\right] \ge \bar{B}\left(6b^\alpha\right) \times \mathbb{P}\left\{\inf_{1\le t\le 3} Z(t) > 1\right\}(1 - o(1)).$$
This observation, along with the regenerative ratio representation (8), concludes the proof of Proposition 1.

4. Proof of upper bound

The objective of this section is to prove the following proposition.

Proposition 2. *Suppose that Assumption 2 holds, and that $\rho = 1$. Then, if Assumption 1 holds with $\alpha > 2$, there exist $c_1 > 0$ and $b_0 > 0$ such that for all $b > b_0$,*
$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \le c_1\left(b^2\bar{B}(b^2) + b^2\bar{B}^2(b)\right).$$
On the other hand, if $\bar{B}(x) \sim cx^{-\alpha}$, as $x \to \infty$, for some $c > 0$ and $\alpha \in (1, 2)$, then one can find positive constants c_1 and b_0 such that for all $b > b_0$,
$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \le c_1\left(b^\alpha\bar{B}(b^\alpha)\right).$$

The rest of this section is devoted to the proof of Proposition 2. First, pick $\delta_-, \delta, \delta_+$ such that $0 < \delta_- < \delta < \delta_+ < 1$. In addition to the stopping times
$$\tau_x^{(i)} = \inf\left\{n \ge 0 : W_n^{(i)} > x\right\},$$
which are defined for $x > 0, i = 1$ and 2, let us define
$$\bar{\tau}_{b\delta_+}^{(2)} = \inf\left\{n \ge \tau_{b\delta_-}^{(2)} : W_n^{(2)} \le b\delta_+\right\}.$$
Additionally, let
$$B_1(b) := \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)} > \tau_{b\delta}^{(1)}\right)\right] \quad \text{and}$$
$$B_2(b) := \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)} \le \tau_{b\delta}^{(1)}\right)\right].$$

Then, it follows from the regenerative ratio representation (8) that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \frac{B_1(b) + B_2(b)}{\mathbb{E}_{\mathbf{0}}[\tau_0]}. \tag{18}$$

The term $B_1(b)$ corresponds to the case where all the actions happen: once there is a large jump in $W^{(2)}$ which takes it beyond $b\delta_-$, one of the servers gets blocked for a long time, and the other server which faces the entire traffic in that duration piles up work more than $b\delta$. On the other hand, the term $B_2(b)$ corresponds to the case where the first jump is wasted: that is, there is not enough buildup in $W^{(1)}$ after the occurrence of first jump in $W^{(2)}$. The rigorous procedure of obtaining upper bounds for $B_1(b)$ and $B_2(b)$ is divided into several parts:

Part 1) First, we obtain an upper bound for $\mathbb{E}_{\mathbf{w}}[\tau_0]$ uniformly over all initial conditions $\mathbf{w} = (w_1, w_2)$. This shall be useful in obtaining upper bounds for both $B_1(b)$ and $B_2(b)$ because of the simple observation that

$$\mathbb{E}_{\mathbf{w}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right)\right] \leq \mathbb{E}_{\mathbf{w}}[\tau_0].$$

Additionally, in an attempt to obtain a stochastic description of the workload $W^{(2)}$ after it exceeds δb_-, we derive a stochastic domination result for $W^{(2)}_{\tau_b^{(2)}}$ which shall be useful.

Part 2) We reduce the contribution of the first term $B_1(b)$ into a large deviations problem for zero-mean random walks with regularly varying increments. We use the stochastic domination result obtained in Part 1) along with another domination argument, in terms of a suitably defined critically loaded single-server queue, to account for all of what happens after the first jump. In turn, the introduction of the single-server queue sets the stage for the use of uniform large deviations for random walks. The analysis of part 2) emphasizes the convenience of partitioning the numerator in (8) into $B_1(b)$ and $B_2(b)$.

Part 3.a) This is the portion of the argument that requires $\alpha > 2$. It invokes classical results for uniform large deviations of regularly varying random walks available due to Nagaev (uniform in the sense that the asymptotics jointly account both the Brownian approximations in the CLT scaling regime and the large deviations approximations in scaling regimes beyond that of CLT). The execution of part 2) involves routine estimations of one dimensional integrals using basic properties of regularly varying distributions. We obtain the required upper bound for $B_1(b)$ after some elementary simplifications.

Part 3.b) The analysis here is entirely parallel to that of part 3.a), except that the uniform estimates involve an approximation using an α-stable process (instead of Brownian motion as in Part 3.a)).

Part 4) is devoted to obtaining an upper bound for the residual term $B_2(b)$. This is accomplished by first performing calculations that result in an intermediate bound for $B_2(b)$ in terms of expected number of jobs that wait for duration longer than b after the first jump. The second calculation involves obtaining a good upper bound for $\mathbb{P}_{\mathbf{w}}\{\tau_b^{(2)} < \tau_0\}$ uniformly over initial conditions $\mathbf{w} \in \{(w_1, w_2) : w_2 < b\delta_+\}$.

In the following subsections we shall estimate the contributions of $B_1(b)$ and $B_2(b)$ following the outline presented above. In order to streamline the presentation, we present proofs of some of the results in the appendix.

Part 1) Some useful upper bounds. Recall our earlier definition $X := V - T$. As mentioned previously, the goal of this subsection is to provide some generic bounds which will be useful in deriving upper bounds for both $B_1(b)$ and $B_2(b)$.

Lemma 1. *Suppose that $\rho = 1$. Then there exist positive constants C_1 and C_0 such that for all $\mathbf{w} = (w_1, w_2)$ satisfying $0 \le w_1 \le w_2$,*

$$\mathbb{E}_{\mathbf{w}}[\tau_0] \le C_1 w_2 + C_0.$$

Remark 1. The conclusion of Lemma 1 holds true for every $\rho < 2$. Our proof for Lemma 1 can be easily modified to accommodate every $\rho < 2$.

Lemma 2. *For every $x \ge b$ and $\mathbf{w} = (w_1, w_2)$ with $0 \le w_1 \le w_2 < b$,*

$$\mathbb{P}_{\mathbf{w}}\left\{W^{(2)}_{\tau_b^{(2)}} > x \;\middle|\; \tau_b^{(2)} < \tau_0\right\} \le \mathbb{P}\left\{X + b > x \mid X > b\right\}.$$

In other words, $W^{(2)}_{\tau_b^{(2)}}$ given $\tau_b^{(2)} < \tau_0$ is stochastically dominated by $X + b$ given $X > b$.

If $\rho < 2$, it is intuitive to expect the servers to effectively drain work whenever $W^{(2)}$ is large. Lemma 3, whose proof is given in Appendix B, asserts the same when $\rho = 1$.

Lemma 3. *There exist positive constants C and ε such that*

$$\mathbb{E}_{(w_1,w_2)}\left[W_1^{(1)} + W_1^{(2)}\right] < (w_1 + w_2) - \varepsilon$$

as long as $w_2 \ge C$.

Lemma 1 follows as a corollary of Lemma 3 via a standard Lyapunov argument.

Proof of Lemma 1. Let $A := \inf\{(w_1, w_2) : w_1 \le w_2 \le C\}$ and $T_A := \inf\{n \ge 1 : W_n \in A\}$. Additionally, let $V((w_1, w_2)) = (w_1 + w_2)/\varepsilon$ for $0 \le w_1 \le w_2$. Here C and ε are chosen as in Lemma 3. It follows from recursions (5a) and (5b) that

$$\sup_{(w_1,w_2)\in A} \mathbb{E}_{(w_1,w_2)}\left[V\left(W_1^{(1)}, W_2^{(2)}\right)\right] \le \frac{\mathbb{E}\left[(C + V - T)^+ + (C - T)^+\right]}{\varepsilon} =: C_2 < \infty.$$

This observation, in conjunction with Lemma 3 and Theorem 11.3.4 of [15], results in

$$\mathbb{E}_{(w_1,w_2)}[T_A] \le \frac{w_1 + w_2}{\epsilon} + C_2 \le \frac{2}{\epsilon} w_2 + C_2 \tag{19}$$

for every $0 \le w_1 \le w_2$. Moreover, since $\inf_{\mathbf{w}\in A} \mathbb{P}_{\mathbf{w}}\{W_1^{(2)} = 0\} \ge \mathbb{P}\{T > C\} > 0$, it follows from a simple geometric trials argument that $\sup_{\mathbf{w}\in A} \mathbb{E}_{\mathbf{w}}[\tau_0] < \infty$. This observation, along with (19), proves the claim. □

Proof of Lemma 2. Note that

$$\mathbb{P}_{\mathbf{w}}\left\{W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0\right\} = \sum_{k=1}^{\infty} \mathbb{P}_{\mathbf{w}}\left\{W_k^{(2)} > x, \tau_0 > k, \tau_b^{(2)} = k\right\}.$$

If $\tau_b^{(2)} = k$, it follows from recursion (5b) that $W_k^{(2)} = W_{k-1}^{(1)} + X_k$. Here recall that $X_k = V_{k-1} - T_k$. Therefore,

$$\mathbb{P}_{\mathbf{w}}\left\{W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0\right\} = \sum_{k=1}^{\infty} \mathbb{P}_{\mathbf{w}}\left\{X_k > x - W_{k-1}^{(1)}, \tau_0 > k-1, \tau_b^{(2)} > k-1\right\}$$
$$= \sum_{k=1}^{\infty} \mathbb{E}_{\mathbf{w}}\left[I\left(\tau_0 > k-1, \tau_b^{(2)} > k-1\right)\bar{F}\left(x - W_{k-1}^{(1)}\right)\right] \tag{20}$$

Observe that $W_{k-1}^{(2)} < b$ whenever $\tau_b^{(2)} > k-1$. Additionally, since x is taken to be larger than b,

$$\frac{\bar{F}\left(x - W_{k-1}^{(1)}\right)}{\bar{F}\left(b - W_{k-1}^{(1)}\right)} \leq \frac{\bar{F}(x-b)}{\bar{F}(b)} \wedge 1 = \mathbb{P}\left\{X + b > x \mid X > b\right\}$$

on the set $\{\tau_b^{(2)} > k-1\}$. Therefore,

$$\mathbb{P}_{\mathbf{w}}\left\{W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0\right\}$$
$$\leq \mathbb{P}\left\{X + b > x \mid X > b\right\} \times \sum_{k=1}^{\infty} \mathbb{E}_{\mathbf{w}}\left[I\left(\tau_0 > k-1, \tau_b^{(2)} > k-1\right)\bar{F}\left(b - W_{k-1}^{(1)}\right)\right]$$
$$= \mathbb{P}\left\{X + b > x \mid X > b\right\}\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\},$$

where the last expression was obtained by letting $x = b$ in the second line in (20). The last inequality is equivalent to the statement of Lemma 2, and this concludes the proof. $\square$

Part 2) Reduction to a zero-mean random walk problem. Recall our earlier definition $X_n := V_{n-1} - T_n$ for $n \geq 1$, where $(V_n : n \geq 1)$ are i.i.d. copies of V and $(T_n : n \geq 1)$ are i.i.d. copies of T. Additionally, we had set $V_0 := 0$. Further, define $S_0 := 0$, $S_n := X_1 + \ldots + X_n$, and

$$N_A(t) := \sup\left\{n \geq 0 : T_1 + \ldots + T_n \leq t\right\} \vee 0$$

for $t \geq 0$. Here we follow the usual convention that $\sup \emptyset = -\infty$. Therefore, $N_A(0) = 0$. Note that $N_A(t)$ is the number of customers that arrive in the time interval $(0, t]$. In addition to the above definitions, let $X := V - T$ and define

$$B_3(b) := \mathbb{E}\left[I\left(\max_{0 \leq n \leq N_A(X)+1} 2\left|S_n\right| > (\delta - \delta_-)\, b\right)\left(X + \max_{0 \leq n \leq N_A(X)+1}|S_n|\right) \;\middle|\; X > b\delta_+\right].$$

Our objective in this subsection is to show the following result.

Lemma 4. *Suppose that Assumptions 1 and 2 hold, and that $\rho = 1$. Then,*

$$B_1(b) = O\left(\mathbb{P}_{\mathbf{0}}\left\{\tau_{b\delta_+}^{(2)} < \tau_0\right\} \times B_3(b)\right).$$

Let $\mathcal{F}_n$ denote the $\sigma-$algebra generated by the random variables V_k and $T_k, k \leq n$. Then

$$B_1(b) = \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}_{b\delta_+}^{(2)} > \tau_{b\delta}^{(1)}\right)\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b\right) \;\middle|\; \mathcal{F}_{\tau_{b\delta}^{(1)}}\right]\right].$$

Since $W_k^{(1)}$ is smaller than b for $k < \tau_{b\delta}^{(1)}$, on the set $\{\tau_b^{(1)} < \tau_0\}$, we have

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) \;\middle|\; \mathcal{F}_{\tau_{b\delta}^{(1)}}\right] = \mathbb{E}_{\mathbf{W}_{\tau_{b\delta}^{(1)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] \leq \mathbb{E}_{\mathbf{W}_{\tau_{b\delta}^{(1)}}}\left[\tau_0\right].$$

Then, due to Lemma 1,

$$B_1(b) \le \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}, \tau^{(1)}_{b} < \tau_0\right)\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right)\right].$$

First, observe that whenever $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$, we must also have that $\tau^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}$. Otherwise, from the definition of $\bar{\tau}^{(2)}_{b\delta_+}$, it follows that $\bar{\tau}^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}$ which in turn occurs earlier than $\tau^{(1)}_{b\delta}$, and this contradicts our blanket assumption $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$. Therefore,

$$\begin{aligned} B_1(b) &\le \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}, \tau^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}, \tau^{(1)}_{b} < \tau_0\right)\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right)\right] \\ &\le \mathbb{E}_{\mathbf{0}}\left[I\left(\tau^{(2)}_{b\delta_+} \le \tau^{(1)}_{b\delta_-} \wedge \tau_0\right)\mathbb{E}_{\mathbf{0}}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right) \;\middle|\; \mathcal{F}_{\tau^{(2)}_{b\delta_+}}\right]\right]. \end{aligned}$$

As a consequence of strong Markov property of $\mathbf{W}$, we have that

$$B_1(b) \le \mathbb{E}_{\mathbf{0}}\left[I\left(\tau^{(2)}_{b\delta_+} \le \tau^{(1)}_{b\delta_-} \wedge \tau_0\right)\mathbb{E}_{\mathbf{W}_{\tau^{(2)}_{b\delta_+}}}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]\right]$$

If we set $\xi_1 := W^{(1)}_{\tau^{(2)}_{b\delta_+}}$ and $\xi_2 := W^{(2)}_{\tau^{(2)}_{b\delta_+}}$, again due to the Markov property of $\mathbf{W}$,

$$B_1(b) \le \mathbb{E}\left[I(\xi_1 < b\delta_-)\mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]\right]\mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_+} < \tau_0\right\}, \tag{21}$$

where ξ_2, by definition, is larger than $b\delta_+$.

Evaluation of the inner expectation. We analyse the inner expectation

$$\chi(\xi_1,\xi_2) := \mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]$$

in (21) by restarting the queuing system with initial conditions $\mathbf{W}_0 = (\xi_1,\xi_2)$. Whenever $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$, due to recursions (5a) and (5b), the dynamics of the queue until $\tau^{(1)}_{b\delta}$ is described by

$$W^{(1)}_n = \left(W^{(1)}_{n-1} + X_n\right)^+ \text{ and } W^{(2)}_n = W^{(2)}_{n-1} - T_n$$

for $1 \le n < \tau^{(1)}_{b\delta}$, in conjunction with $W^{(1)}_0 = \xi_1 < b\delta_-$ and $W^{(2)}_0 = \xi_2 > b\delta_+$. As a result,

$$T_1 + \ldots + T_{\tau^{(1)}_{b\delta}-1} = \xi_2 - W^{(2)}_{\tau^{(1)}_{b\delta}-1} \le \xi_2 - b\delta_+$$

whenever $\bar{\tau}^{(2)}_{b\delta^+} > \tau^{(1)}_{b\delta}$. Therefore, $\tau^{(1)}_{b\delta} \le N_A(\xi_2 - b\delta^+) + 1$, which in turn implies that

$$\max_{0<n\le N_A(\xi_2-b\delta^+)+1} W^{(1)}_n > b\delta.$$

Consequently,

$$\chi(\xi_1,\xi_2) \le \mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 \max_{0<n\le N_A(\xi_2-b\delta^+)+1} W^{(2)}_n + C_0\right) I\left(\max_{0<n\le N_A(\xi_2-b\delta^+)+1} W^{(1)}_n > b\delta\right)\right].$$

The following result is verified in Appendix B.

Lemma 5. *Suppose that* $\mathbf{W}_0 = (w_1, w_2)$*, and recall the definitions* $X_n := V_{n-1} - T_n,\ S_0 := 0$ *and* $S_n := X_1 + \ldots + X_n$*. Then, for all* $n \geq 0$*,*

$$\max_{0<k\leq n} W_k^{(i)} \leq 2 \max_{0\leq k\leq n} |S_k| + w_i,\ \ i = 1, 2.$$

As a consequence of Lemma 5,

$$\chi(\xi_1, \xi_2) \leq C_3 \mathbb{E}\left[\left(\max_{0\leq n\leq N_A(\xi_2 - b\delta^+)+1} 2S_n + \xi_2\right) I\left(\max_{0\leq n\leq N_A(\xi_2-b\delta^+)+1} 2S_n + \xi_1 > b\delta\right) \ \Bigg| \ \ \xi_1, \xi_2\right].$$

where the constant C_0 been absorbed in another suitable constant C_3. Then it is immediate from (21) that

$$B_1(b) \leq C_3\mathbb{E}\left[\left(\max_{0\leq n\leq N_A(\xi_2-b\delta^+)+1} 2S_n + \xi_2\right) I\left(\max_{0\leq n\leq N_A(\xi_2-b\delta^+)+1} 2S_n > (\delta - \delta_-)\, b\right)\right] \mathbb{P}_{\mathbf{0}}\left\{\tau_{b\delta_+}^{(2)} < \tau_0\right\}.$$

Here, recall that $\xi_2 := W^{(2)}_{\tau^{(2)}_{b\delta_+}}$, which is stochastically dominated by the conditional distribution of $X + b\delta_+$ given that $X > b\delta_+$ (due to Lemma 2). Since $B_2(b)$ is a non-decreasing function of ξ_2, we use the above stochastic dominance to yield

$$\begin{aligned} B_1(b) \leq C_3 \mathbb{P}_{\mathbf{0}}&\left\{\tau_{b\delta_+}^{(2)} < \tau_0\right\} \times \\ &\mathbb{E}\left[\left(\max_{0\leq n\leq N_A(X)+1} 2S_n + X + b\delta_+\right) I\left(\max_{0\leq n\leq N_A(X)+1} 2S_n > (\delta-\delta_-)\, b\right) \ \Bigg| \ X > b\delta_+\right]. \end{aligned}$$

Lemma 4 follows from the above inequality once we observe that $X + b\delta_+ \leq 2X$ when $X > b\delta_+$. This completes the proof of Lemma 4. ☐

Part 3.a) Simplifications using uniform large deviations: the $\alpha > 2$ case. Using classical results borrowed from the literature on large deviations for zero-mean random walks, we aim to prove the following result in this subsection.

Lemma 6. *Suppose that Assumptions 1 and 2 are in force,* $\alpha > 2$ *and* $\rho = 1$*. Then,*

$$B_3(b) = O\left(b^2 \bar{B}(b) + b^2 \frac{\bar{B}(b^2)}{\bar{B}(b)}\right).$$

We begin by recalling results on uniform large deviations for regularly varying random walks. For example, the following large deviations result which holds under Assumptions 1 and 2 assuming that $\alpha > 2$, is well-known

$$\mathbb{P}\{S_m > b\} = \left(\bar{\Phi}\left(\frac{b}{\sqrt{m}\sigma}\right) + m\mathbb{P}\{X_1 > b\}\right)(1 + o(1)),\ \text{as}\ m \to \infty, \tag{22}$$

uniformly for $b > \sqrt{m}$, where $\bar{\Phi}(\cdot)$ is the tail of a standard normal distribution. The asymptotic approximation (22) is due to A. V. Nagaev (see Theorem 1.9 of [17] or Corollary 7 of [19]).

For our purposes, we need an extension of (22), in which S_n is replaced by $\max_{0\leq k\leq m} |S_k|$. However, we do not need exact asymptotic results as in (22), but only an asymptotic upper bound. This is the content of the following result, which is proved in Appendix B as an immediate consequence of Corollary 1 of [18]. (For related uniform sample path large deviations results see [6], and the related Theorem 5 of [5].)

Lemma 7. *Suppose that V satisfies Assumption 1 with $\alpha > 2$, and T satisfies Assumption 2. Recall that $X_1, X_2, \ldots$ are i.i.d. copies of $X = V - T$. Then, there exists a positive integer m_0 such that for all $x \geq m^{1/2}$ and $m > m_0$*

$$\mathbb{P}\left\{\max_{0\leq k\leq m}|S_k| > x\right\} \leq 3\left(\mathbb{P}\left\{\max_{0\leq t\leq 1}\sigma\left|B\left(t\right)\right| > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\left\{|X| > x\right\}\right),$$

where $\sigma^2 = Var[X]$ and $B\left(\cdot\right)$ is a standard Brownian motion.

We also need the following pair of standard results on regular variation: Karamata's theorem (refer Theorem 1 in Chapter VIII.9 of [8]) and Potter's bounds (see, for example, Theorem 1.1.4 of [6])

Proposition 3 (Karamata's theorem)**.** *Suppose that $v(t) = t^{-\alpha}L(t)$ for some slowly varying function $L(\cdot)$ and α satisfying $\alpha - \beta > 1$. Then*

$$\int_x^\infty u^\beta v(u)du \sim \frac{x^{\beta+1}v(x)}{\alpha-\beta-1}, \text{ as } x \to \infty. \tag{23}$$

On the other hand, if $\alpha - \beta < 1$, then

$$\int_0^x u^\beta v(u)du \sim \frac{x^{\beta+1}v(x)}{1-\alpha+\beta}, \text{ as } x \to \infty. \tag{24}$$

Proposition 4 (Potter's bounds)**.** *If $v(t) = t^{-\alpha}L\left(t\right)$ for some $\alpha > 0$ and some slowly varying function $L\left(\cdot\right)$ then, for any $\varepsilon \in \left(0, \min(\alpha, 1)\right)$, there exists a $t_\varepsilon > 0$ such that for all t and c satisfying $t \geq t_\varepsilon$ and $ct \geq t_\varepsilon$,*

$$(1-\varepsilon)\min\{c^{-\alpha+\varepsilon}, c^{-\alpha-\varepsilon}\} \leq \frac{v(ct)}{v(t)} \leq (1+\varepsilon)\max\{c^{-\alpha+\varepsilon}, c^{-\alpha-\varepsilon}\}. \tag{25}$$

We establish Lemma 6 in two parts. The first task involves analysing the relatively easier term, which has the running maximum appearing only in the indicator function.

Lemma 8. *Under Assumption 1 with $\alpha > 2$, and Assumption 2,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1}2\left|S_n\right| > \left(\delta-\delta_-\right)b\right)X \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}\left(b\right) + b^2\frac{\bar{B}\left(b^2\right)}{\bar{B}(b)}\right).$$

Following this, we estimate the term in which the running maximum appears both multiplying and inside the indicator.

Lemma 9. *Under Assumption 1 with $\alpha > 2$, and Assumption 2,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1}2\left|S_n\right| > \left(\delta-\delta_-\right)b\right)\max_{0\leq n\leq N_A(X)+1}\left|S_n\right| \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}\left(b\right)\right).$$

Lemma 10, whose proof is given in Appendix B, will be useful in proving Lemmas 8 and 9.

Lemma 10. *If $v(x) = x^{-\alpha}l(x)$ for some $\alpha > 2$ and a function $l(\cdot)$ slowly varying at infinity, then for every $c > 0$,*

$$\int_b^\infty v(t)\exp\left(-c\frac{b^2}{t}\right)dt = O\left(b^2 v\left(b^2\right)\right).$$

Proof of Lemma 8. Letting $c = (\delta - \delta_-)/2$, observe that

$$\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)} 2\left|S_n\right| > (\delta-\delta_-)\,b,\ N_A(X)+1\le 2X\right)X\ \Big|\ X > b\delta_+\right]$$

$$\le \int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le n\le 2t}\left|S_n\right| > cb\right\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}}$$

(26)

$$\le 3\int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le s\le 1}\sigma B(s) > \frac{cb}{\sqrt{2t}}\right\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}} + 3\int_{b\delta_+}^{\frac{c^2b^2}{2}} 2t^2\mathbb{P}\{|X|>cb\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}}$$

because of the application of the uniform asymptotic presented in Lemma 7 in the region $2t \le c^2b^2$ and Central Limit Theorem in the region $2t > c^2b^2$. Recall from (4) that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$, and subsequently, due to Karamata's theorem (23), we obtain

$$\int_{b\delta_+}^{\infty} t^2\mathbb{P}\{X\in dt\} \le \mathbb{E}\left[X^2 I\left(X>b\delta_+\right)\right] = O\left(\int_{b\delta_+}^{\infty} s\mathbb{P}\{X>s\}\,ds\right) = O\left(b^2\bar{B}(b)\right)$$

and therefore

$$\frac{\mathbb{P}\left(|X|>cb\right)}{\mathbb{P}\{X>b\delta_+\}}\int_b^{\infty} t^2\mathbb{P}\{X\in dt\} = O\left(b^2\bar{B}(b)\right). \tag{27}$$

To deal with the first term in (26), we do integration by parts (by taking $u = \mathbb{P}\{\max_{0\le s\le 1} B(s) > cb/\sqrt{2\sigma t}\}$ and $v = \int_t^{\infty}\mathbb{P}\{X>u\}du - t\mathbb{P}\{X>t\}$) to obtain

$$\int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le s\le 1}\sigma B(s) > \frac{cb}{\sqrt{2t}}\right\}\mathbb{P}\{X\in dt\} = O\left(b\int_{b\delta_+}^{\infty}\frac{\mathbb{P}\{X>t\}}{\sqrt{t}}\exp\left(-\frac{cb^2}{4\sigma t}\right)dt\right),$$

which, in turn, is $O(b\times b\bar{B}(b^2))$ because of Lemma 10. Therefore, due to (26) and (27), along with the observation that $\mathbb{P}\{X > b\delta_+\} = \Theta(\bar{B}(b))$ (due to regular variation), we obtain

(28)

$$\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)+1} 2\left|S_n\right| > (\delta-\delta_-)\,b, N_A(X)+1\le 2X\right)X\ \Big|\ X>b\delta_+\right] = O\left(b^2\bar{B}(b) + b^2\frac{\bar{B}(b^2)}{\bar{B}(b)}\right).$$

On the other hand, given that $N_A(t)/t \to 1$ as $t\to\infty$, the event $\{N_A(t) > 2t-1\}$ corresponds to a large deviations event with exponentially small probability for large values of t. Therefore, we have that

$$\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)} 2\left|S_n\right| > (\delta-\delta_-)\,b, N_A(X)+1 > 2X\right)X\ \Big|\ X>b\delta_+\right]$$

$$\le \int_{b\delta_+}^{\infty} t\mathbb{P}\{N_A(t) > 2t-1\}\mathbb{P}\{X\in dt\} = O\left(\exp\left(-\gamma b\right)\right),$$

for a suitable $\gamma > 0$. This observation, along with (28), concludes the proof of Lemma 8. ☐

The proof of Lemma 9, where running maximum appears twice, is similar, but more involved, and is presented in Appendix B, so that we can continue with central arguments in the main body of the paper. Before moving to Part 3.b) of the proof, it is important to note that Lemma 6 stands proved as an immediate consequence of Lemmas 8 and 9.

Part 3.b) Simplifications using uniform large deviations: the $\alpha \in (1,2)$ case. We shall leverage much of the reasoning behind Part 3.a) and prove the following result:

Lemma 11. *Suppose that $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$ for some $c > 0$ and $\alpha \in (1,2)$. Also, suppose that Assumption 2 holds. Then,*

$$B_3(b) = O\left(\frac{b^\alpha \bar{B}(b^\alpha)}{\bar{B}(b)}\right).$$

We begin with a uniform convergence result which is a special case of Theorem 3.8.2 of [6]:

Lemma 12. *Suppose that $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$ for some $c > 0$ and $\alpha \in (1,2)$. Also, suppose that Assumption 2 holds.Then, there exists a positive integer m_0 such that for all $m \geq m_0$,*

$$\mathbb{P}\left\{\max_{0\leq n\leq m} |S_n| > x\right\} \leq 3\mathbb{P}\left\{Z_* > \frac{x}{(cm)^{1/\alpha}}\right\},$$

where $Z_* := \max_{0\leq s\leq 1} Z(s)$ is the maximum of an α-stable process $(Z(t) : 0 \leq t \leq 1)$ satisfying $\mathbb{P}\{Z(1) > x\} \sim x^{-\alpha}$ as $x \to \infty$. Additionally, for such a stable process $Z(\cdot)$, we have that

$$\mathbb{P}\{Z_* > x\} \sim x^{-\alpha} \text{ as } x \to \infty.$$

The adaptation of Theorem 3.8.2 of [6] to the case where maximum of $|S_n|$ appears (instead of maximum of S_n) is similar to the argument in the proof of Lemma 7 in Part 3.a), and therefore is omitted. The dominant contribution to $B_3(b)$ is accounted for in the following result:

Lemma 13. *Suppose that $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$ for some $c > 0$ and $\alpha \in (1,2)$. Also, suppose that Assumption 2 holds. Then,*

$$\mathbb{E}\left[I\left(\max_{0\leq k\leq N_A(X)+1} 2|S_n| > (\delta - \delta_-)b\right)X \;\middle|\; X > b\delta_+\right] = O\left(\frac{b^\alpha \bar{B}(b^\alpha)}{\bar{B}(b)}\right).$$

Proof of Lemma 13. As a consequence of Lemma 12, we get

$$\begin{aligned}\mathbb{E}\left[I\left(\max_{0\leq k\leq 2X} 2|S_n| > (\delta - \delta_-)b\right)XI(X > b\delta_+)\right] &\leq 3\mathbb{E}\left[\mathbb{P}\left\{Z_* > \frac{(\delta-\delta_-)b}{2(2cX)^{\frac{1}{\alpha}}}\right\}XI(X > b\delta_+)\right]\\ &= 3\mathbb{E}\left[XI\left(X > (b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^\alpha\right)\right] \qquad (29)\end{aligned}$$

where $\bar{c} := (\delta - \delta_-)/(2^{\alpha+1}c)^{1/\alpha}$. Additionally, for all large enough x, there exists a constant C such that $\mathbb{E}[XI(X > x)] \leq Cx^{-(\alpha-1)}$, because of Karamata's theorem and the observation that $\mathbb{P}\{X > x\} \sim cx^{-\alpha}$ as $x \to \infty$. Therefore, for all b large enough, we obtain

$$\begin{aligned}\mathbb{E}\left[XI\left(X > (b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^\alpha\right)\right] &\leq C\mathbb{E}\left[\left((b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^\alpha\right)^{-(\alpha-1)}\right]\\ &\leq C(b\delta_+)^{-(\alpha-1)}\mathbb{P}\left\{Z_* > \frac{\bar{c}b^{1-\frac{1}{\alpha}}}{\delta_+^{\frac{1}{\alpha}}}\right\} + C(\bar{c}b)^{-\alpha(\alpha-1)}\mathbb{E}\left[Z_*^{\alpha^2-\alpha}\right],\end{aligned}$$

which, in turn, is $O(b^\alpha \bar{B}(b^\alpha))$ because $\mathbb{E}[Z_*^{\alpha^2-\alpha}] < \infty$ when $\alpha \in (1,2)$. Therefore, due to (29),

$$\mathbb{E}\left[I\left(\max_{0\leq k\leq 2X} 2|S_n| > (\delta - \delta_-)b,\ N_A(X) + 1 \leq 2X\right)XI(X > b\delta_+)\right] = O\left(b^\alpha \bar{B}(b^\alpha)\right).$$

On the other hand, the event $\{N_A(b)+1>2b\}$ is a large deviations event with probabilities exponentially decaying in b, and as argued in the proof of Lemma 8,

$$\mathbb{E}\left[I\left(\max_{0\leq k\leq 2X} 2\left|S_n\right| > (\delta-\delta_-)\, b,\ N_A(X)+1>2X\right) X I\left(X>b\delta_+\right)\right] = O\left(\exp(-\gamma b)\right),$$

for a suitable $\gamma>0$. These two observations, after adjusting for the conditioning by dividing by $\mathbb{P}\{X>b\delta_+\}=\Theta(\bar{B}(b))$, prove Lemma 13. $\square$

Lemma 14. *Suppose that $\bar{B}\left(x\right)\sim cx^{-\alpha}$ as $x\to\infty$ for some $c>0$ and $\alpha\in(1,2)$. Also, suppose that Assumption 2 holds. Then,*

$$\mathbb{E}\left[I\left(\max_{0\leq k\leq N_A(X)+1} 2\left|S_n\right| > (\delta-\delta_-)\, b\right)\max_{0\leq k\leq N_A(X)+1}\left|S_n\right| \;\middle|\; X>b\delta_+\right] = O\left(b^2\bar{B}\left(b\right)\right).$$

As in Part 3.a), the proof of Lemma 14 is furnished in Appendix B. The main result of this section, Lemma 11, which aims to prove that $B_3(b)=O(b^\alpha\bar{B}(b^\alpha)/\bar{B}(b))$ is an immediate consequence of Lemmas 13 and 14, along with the observation that $b^2\bar{B}^2(b)=o(b^\alpha\bar{B}(b^\alpha))$ when $\alpha<2$.

Part 4) Estimation of $B_2(b)$. The objective of this subsection is to prove Lemma 15, and subsequently, complete the proof of Proposition 2.

Lemma 15. *Suppose that Assumptions 1 and 2 hold, and that $\rho=1$. Then,*

$$B_2\left(b\right)=O\left(b^2\bar{B}\left(b\right)^2\right).$$

It follows from the definition of $B_2(b)$ that

$$\begin{aligned}B_2(b) &= \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)}>b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)}\leq\tau_{b\delta}^{(1)},\tau_{b\delta_-}^{(2)}<\tau_0\right)\right]\\ &= \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}_{b\delta_+}^{(2)}\leq\tau_{b\delta}^{(1)},\tau_{b\delta_-}^{(2)}<\tau_0\right)\mathbb{E}_{\mathbf{0}}\left[\sum_{k=\bar{\tau}_{b\delta_+}^{(2)}}^{\tau_0-1} I\left(W_k^{(1)}>b\right)\;\middle|\;\mathcal{F}_{\bar{\tau}_{b\delta_+}^{(2)}}\right]\right]\end{aligned}$$

Then due to the Markov property of $\mathbf{W}$, we get

$$B_2(b)=\mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}_{b\delta_+}^{(2)}\leq\tau_{b\delta}^{(1)},\tau_{b\delta_-}^{(2)}<\tau_0\right)\mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)}>b\right)\right]\right]. \tag{30}$$

Evaluation of inner expectation. Due to a similar conditioning with respect to $\mathcal{F}_{\tau_b^{(2)}}$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)}>b\right)\right] &= \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[I\left(\tau_b^{(2)}<\tau_0\right)\mathbb{E}_{\mathbf{W}_{\tau_b^{(2)}}}\left[\sum_{k=\tau_b^{(2)}}^{\tau_0-1} I\left(W_k^{(1)}>b\right)\right]\right]\\ &\leq \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[I\left(\tau_b^{(2)}<\tau_0\right)\mathbb{E}_{\mathbf{W}_{\tau_b^{(2)}}}\left[\tau_0\right]\right],\end{aligned}$$

which, due to Lemma 1, admits the following upper bound:

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right)\right] &\le \mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[I\left(\tau^{(2)}_b < \tau_0\right)\left(C_1 W^{(2)}_{\tau^{(2)}_b} + C_0\right)\right] \\
&= C_1 \mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[W^{(2)}_{\tau^{(2)}_b} \;\middle|\; \tau^{(2)}_b < \tau_0\right] \mathbb{P}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left\{\tau^{(2)}_b < \tau_0\right\} + C_0 \mathbb{P}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left\{\tau^{(2)}_b < \tau_0\right\} \\
&\le C_2 b \mathbb{P}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left\{\tau^{(2)}_b < \tau_0\right\} \qquad (31)
\end{aligned}$$

for some positive constant C_2, because, due to Lemma 2,

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[W^{(2)}_{\tau^{(2)}_b} \;\middle|\; \tau^{(2)}_b < \tau_0\right] \le \mathbb{E}\left[X + b \;\middle|\; X > b\right] = O(b).$$

Next, we obtain a bound for $\mathbb{P}_{\mathbf{w}}\{\tau^{(2)}_b < \tau_0\}$, and use it in (31).

Lemma 16. *Suppose that Assumptions 1 and 2 hold, and that* $\rho = 1$*. Let* $\delta_+ < 1/2$*, then there exists a constant* $C > 0$ *such that for all* $\mathbf{w} = (w_1, w_2)$ *satisfying* $w_1 \le w_2 < b\delta_+$ *we have*

$$\mathbb{P}_{\mathbf{w}}\left\{\tau^{(2)}_b < \tau_0\right\} \le C\left(w_2 + 1\right)\bar{B}\left(b\right).$$

The proof of Lemma 16 is instructive, as it employs Lyapunov bound techniques to derive the above uniform bound. The arguments involved are different from the rest of the paper, and the whole of Appendix A is dedicated to expose the techniques clearly. Now, we aim to complete the proof of Lemma 15. Due to Lemma 16 and (31),

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right)\right] \le CC_2 b\left(W^{(2)}_{\bar{\tau}^{(2)}_{b\delta_+}} + 1\right)\bar{B}(b) \le CC_2 b\left(b\delta_+ + 1\right)\bar{B}(b),$$

and therefore, due to (30) and a similar application of Lemma 16 with $\mathbf{w} = \mathbf{0}$,

$$B_2(b) \le CC_2 \mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_-} < \tau_0\right\} b\left(b\delta_+ + 1\right)\bar{B}(b) = O\left(b^2 \bar{B}^2(b)\right).$$

This concludes the proof of Lemma 15.

Proof of Proposition 2. We apply Lemma 16 with $\mathbf{w} = \mathbf{0}$ to the bound for $B_1(b)$ in Lemma 4 as well. This, due to Lemmas 6 and 11, results in

$$B_1(b) = O\left(\bar{B}\left(b\delta_+\right) \times B_3(b)\right) = \begin{cases} O\left(b^2\bar{B}^2(b) + b^2\bar{B}\left(b^2\right)\right) & \text{if } \alpha > 2, \\ O\left(b^\alpha \bar{B}\left(b^\alpha\right)\right) & \text{if } \alpha \in (1,2). \end{cases}$$

Additionally, we have that $B_2(b) = O(b^2\bar{B}(b)^2)$ and $\mathbb{E}_{\mathbf{0}}[\tau_0] < \infty$, respectively, from Lemmas 15 and 1. Therefore, from (18), we arrive at the statement of Proposition 2, and this concludes the proof. □

Appendix A. Lyapunov bound techniques for a uniform bound on $\mathbb{P}_{\mathbf{w}}\{\tau^{(2)}_b < \tau_0\}$

We use the Lyapunov bound technique that has been employed in [2], [3], and [7]. The strategy is to define a Markov kernel $Q_\theta\left(\mathbf{w}, \cdot\right)$ (indexed by some parameter θ) and a non-negative function $H_b\left(w_1, w_2\right)$ satisfying the following conditions:

(L1) For every $\mathbf{w} = (w_1, w_2)$ such that $w_2 < b$,

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w}, \mathbf{W}_1\right) H_b\left(\mathbf{W}_1\right)\right] \leq H_b\left(\mathbf{w}\right),$$

where $\mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in \cdot\}$ is the nominal transition kernel induced by recursions (5a) and (5b), $r_\theta(\mathbf{w}, \mathbf{x}) := \mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in d\mathbf{x}\}/Q_\theta(\mathbf{w}, d\mathbf{x})$ is the corresponding Radon-Nikodym derivative with respect to $Q_\theta(\mathbf{w}, \cdot)$, and $\mathbb{E}^{\theta}_{\mathbf{w}}[\cdot]$ is the expectation associated with the probability measure in path space for the Markov evolution induced by $Q_\theta(\mathbf{w}, \cdot)$.

(L2) Whenever $\mathbf{w} = (w_1, w_2)$ is such that $w_2 > b$, $H_b(w_1, w_2) \geq 1$.

If conditions (L1) and (L2) are satisfied, then following the analysis in Part (iii) of Theorem 2 of [2], we have that

$$\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\} \leq \mathbb{E}^{\theta}_{\mathbf{w}}\left[\prod_{n=0}^{\tau_b^{(2)}-1} r_\theta\left(\mathbf{W}_n, \mathbf{W}_{n+1}\right) H_b\left(\mathbf{W}_{\tau_b^{(2)}}\right) I\left(\tau_b^{(2)} < \tau_0\right)\right] \leq H_b\left(w_1, w_2\right). \tag{32}$$

The construction of $Q_\theta(\mathbf{w}, \cdot)$ and $H_b(\cdot)$ follows the intuition explained in [2] and [3]: We wish to select $Q_\theta(\mathbf{w}, \cdot)$ as closely as possible to the conditional distribution of the process $\{\mathbf{W}_n : n \geq 0\}$ given that $\{\tau_b^{(2)} < \tau_0\}$, because in that case, it happens that (32) is automatically satisfied with equality. Additionally, we shall find a suitable non-negative function $G_b(\cdot)$ so that $H_b(w_1, w_2) = G_b(w_1 + w_2)$ satisfies the Lyapunov inequality (L1).
For ease of notation, let us write

$$l := w_1 + w_2,\ L := W_1^{(1)} + W_2^{(2)} \text{ and } \Delta := L - l.$$

In order to construct $Q_\theta(\mathbf{w}, \cdot)$ and $G_b(\cdot)$, first define the Markov transition kernel

$$\begin{aligned} Q'(\mathbf{w}, A) = {} & \mathbb{P}_{\mathbf{w}}\left\{\mathbf{W}_1 \in A \mid X_1 > a(b-l)\right\} p(\mathbf{w}) \\ & + \mathbb{P}_{\mathbf{w}}\left\{\mathbf{W}_1 \in A \mid X_1 \leq a(b-l), W_1^{(2)} > 0\right\}(1 - p(\mathbf{w})), \end{aligned}$$

where $p(\mathbf{w})$ will be specified momentarily, and the choice $a \in (0,1)$ is arbitrary. On the set $\{\tau_b^{(2)} < \tau_0\}$, given $\mathbf{w} = (w_1, w_2)$ with $w_1 \leq w_2 < b$, we have that the nominal kernel $\mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in \cdot\}$ is absolutely continuous with respect to $Q'(\mathbf{w}, \cdot)$. Now, for $z \geq 0$, define

$$h_b(z) = \int_0^{z+\kappa_0} \mathbb{P}\{X > b - z + t\}\, dt = \int_{b-z}^{b+\kappa_0} \mathbb{P}\{X > u\}\, du.$$

Next, write

$$G_b(l) = \min(\kappa_1 h_b(l), 1)$$

and set

$$p(\mathbf{w}) = \frac{\mathbb{P}\{X > a(b-l)\}}{\kappa_2 h_b(l)}$$

where κ_2 is a number larger than

$$\sup_{x>0} \frac{\mathbb{P}\{X > ax\}}{\int_x^{x+l+\kappa_0} \mathbb{P}\{X > u\}\, du} < \infty.$$

Finally, define $\theta = (\kappa_0, \kappa_1, \kappa_2)$ and write

$$Q_\theta(\mathbf{w}, \cdot) = Q'(\mathbf{w}, \cdot)\, I(G_b(l) < 1) + K(\mathbf{w}, \cdot)\, I(G_b(l) = 1).$$

Recall the notation $l = w_1 + w_2$ and $L = W_1^{(1)} + W_1^{(2)}$. Condition (L1) is verified via the following proposition:

Proposition 5. *For every* $\mathbf{w} = (w_1, w_2)$ *such that* $w_2 < b$*, we have that*

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w}, \mathbf{W}_1\right) G_b(L)\right] \le G_b(l).$$

For proving Proposition 5, we consider only the case $G_b(l) < 1$. When $G_b(l) = 1$, the inequality is satisfied trivially. The following results are crucial in the proof of Proposition 5.

Lemma 17. *There exist positive constants* μ *and* C *such that*

$$\mathbb{E}_{(w_1,w_2)}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] < -\mu$$

whenever $w_2 > C$*.*

Proof. First, observe that

$$\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] = \mathbb{E}_{\mathbf{w}}\left[\Delta\right] - \mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} = 0\right)\right].$$

Additionally, note that $\Delta = -(w_1 + w_2)$ when $W_1^{(2)} = 0$. Therefore,

$$\mathbb{E}_{(w_1,w_2)}\left[\Delta I\left(W_1^{(2)} = 0\right)\right] = -\left(w_1 + w_2\right)\mathbb{P}\left\{w_1 + V - T \le 0,\ w_2 - T \le 0\right\}.$$

Therefore, due to Lemma 3,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] &\le \mathbb{E}_{\mathbf{w}}\left[\Delta\right] + \left(w_1 + w_2\right)\mathbb{P}\left\{w_2 - T \le 0\right\} \\ &\le -\epsilon + 2w_2\mathbb{P}\left\{T > w_2\right\},\end{aligned}$$

where $w_2\mathbb{P}\{T > w_2\}$ can be made arbitrarily small by choosing $C > w_2$ large enough. Hence the claim stands verified. □

Lemma 18. *Recall that* $l = w_1 + w_2$*. The following holds as* $(b - l) \to \infty$ *:*

$$\begin{aligned}&\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w}, \mathbf{W}_1\right)\frac{G_b\left(L\right)}{G_b(l)} I\left(X_1 \le a\left(b - l\right)\right)\right] \\ &\quad\le \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\left\{X > b - l\right\}}{h_b(l)}\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right]\left(1 + o(1)\right).\end{aligned}$$

Proof. Since

$$G_b\left(L_1\right) = G_b\left(l\right) + \int_0^1 G_b'\left(l + u\Delta\left(1\right)\right)\Delta\left(1\right) du,$$

we introduce a uniform random variable U, independent of everything else, to write

$$\begin{aligned}&\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w}, \mathbf{W}_1\right)\frac{G_b\left(L\right)}{G_b(l)} I\left(X_1 \le a\left(b - l\right)\right)\right] \\ &\quad= \mathbb{E}_{\mathbf{w}}\left[\frac{G_b\left(L\right)}{G_b(l)} I\left(X_1 \le a\left(b - l\right), W_1^{(2)} > 0\right)\right] \\ &\quad= \mathbb{E}_{\mathbf{w}}\left[\left(1 + \frac{G_b'\left(l + U\Delta\right)}{G_b(l)}\Delta\right) I\left(X_1 \le a\left(b - l\right), W_1^{(2)} > 0\right)\right]\end{aligned}$$

(33)

$$\quad\le \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\left\{X > b - l\right\}}{h_b(l)}\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\left\{X > b - l - U\Delta\right\}}{\mathbb{P}\left\{X > b - l\right\}}\Delta I\left(X_1 \le a\left(b - l\right), W_1^{(2)} > 0\right)\right].$$

We have used $G_b'\left(l + U\Delta\left(1\right)\right) = \kappa_1\mathbb{P}\left\{X > b - l - U\Delta\left(1\right)\right\}$ to write the last step. Additionally, whenever $X_1 \le a\left(b - l\right)$, observe that

$$\Delta = \left(w_1 + X_1\right)^+ - w_1 + \left(w_2 - T_1\right)^+ - w_2 \le X_1^+ \le a\left(b - l\right),$$

and therefore, $\mathbb{P}\{X > b - l - U\Delta\} \le \mathbb{P}\{X > (1-a)(b-l)\} \le m_{1-a}\mathbb{P}\{X > b-l\}$, where

$$m_t := \sup_{x>0} \frac{\mathbb{P}\{X > tx\}}{\mathbb{P}\{X > x\}} < \infty,$$

for every $t > 0$. Here, the finiteness of m_t follows from the regularly varying nature of the tail distribution of X (recall that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$ as $x \to \infty$). As a result, we have the following uniform bound for various values of b and l :

$$\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\{X > b - l - U\Delta\}}{\mathbb{P}\{X > b-l\}}\Delta I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \le m_{1-a}\mathbb{E}X^+. \tag{34}$$

Consequently, due to dominated convergence theorem, we obtain that

$$\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\{X > b - l - U\Delta\}}{\mathbb{P}\{X > b-l\}}\Delta I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \sim \mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right],$$

as $(b-l) \to \infty$. Now, the statement of Lemma 18 is immediate from (33) and the above stated convergence. □

Proof of Proposition 5. As mentioned before, we consider $G_b(l) < 1$. First, observe that

$$\begin{aligned}\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}I(X_1 > a(b-l))\right] &= \mathbb{E}_{\mathbf{w}}\left[\frac{G_b(L)}{G_b(l)}I(X_1 > a(b-l))\right]\\ &\le \frac{\mathbb{P}\{X > a(b-l)\}}{\kappa_1 h_b(l)}\end{aligned} \tag{35}$$

because $G_b(\cdot) \le 1$. For a respective bound on the complementary event $\{X_1 \le a(b-l)\}$, it is easy to see that our strategy must use Lemmas 17 and 18 in the following way: Given $\delta > 0$, there exists a constant C_δ large enough such that for all initial conditions $\mathbf{w} = (w_1, w_2)$ satisfying $w_2 > C$ and $b - l > C_\delta$,

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}I(X_1 \le a(b-l))\right] \le \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} - (1-\delta)\mu\frac{\mathbb{P}\{X > b-l\}}{h_b(l)}.$$

Combining this bound with (35), we obtain

$$\begin{aligned}\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}\right] &\le \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\{X > a(b-l)\}}{h_b(l)}\left(\frac{1}{\kappa_1} - \frac{(1-\delta)\mu}{m_a}\right)\\ &\le 1 + \kappa_2 p(\mathbf{w})\left(\frac{1}{\kappa_1} - \frac{(1-\delta)\mu}{m_a}\right)\end{aligned}$$

which is, in turn, smaller than 1 for κ_1 suitably large. In addition to this, in the region $\{(w_1, w_2) : w_2 > C, b - l < C_\delta\}$, we simply let $G_b(l) = 1$ by again choosing κ_1 large enough. This flexibility in the choice of κ_1 yields us

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}\right] \le 1 \tag{36}$$

for initial conditions $\mathbf{w} = (w_1, w_2)$ satisfying $w_2 > C$. Now, turning our attention to the values of $\mathbf{w}$ such that $w_2 \le C$, we see that $l = w_1 + w_2 \le 2C$, and as a consequence of the regularly varying nature of the tail of X, we obtain

$$\frac{\mathbb{P}\{X > b-l\}}{h_b(l)} = \left(\int_0^{l+\kappa_0} \frac{\mathbb{P}\{X > b-l+u\}}{\mathbb{P}\{X > b-l\}}du\right)^{-1} = \frac{1+o(1)}{l+\kappa_0}$$

as $b \to \infty$. Then, it is immediate from (33) that whenever $w_2 \le C$,

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w},\mathbf{W}_1\right)\frac{G_b\left(L\right)}{G_b(l)}I\left(X_1 \le a\left(b-l\right)\right)\right] \le \mathbb{P}_{(C,C)}\left\{W_1^{(2)} > 0\right\} + \frac{m_{1-a}\mathbb{E}\left[X_1^+\right]}{\kappa_0}\left(1+o(1)\right).$$

Combining this bound with the one obtained in (35), we get

$$\mathbb{E}^{\theta}_{\mathbf{w}}\left[r_\theta\left(\mathbf{w},\mathbf{W}_1\right)\frac{G_b\left(L\right)}{G_b(l)}\right] \le \frac{p\left(\mathbf{w}\right)\kappa_2}{\kappa_1} + \mathbb{P}_{(C,C)}\left\{W_1^{(2)} > 0\right\} + \frac{m_{1-a}\mathbb{E}\left[X_1^+\right]}{\kappa_0}\left(1+o(1)\right),$$

which can also be made smaller than 1 by picking κ_0 and κ_1 large enough. Thus, for all initial conditions $\mathbf{w}$, we have a consistent choice of parameters $(\kappa_0,\kappa_1,\kappa_2)$ that satisfies (L1). □

Since $G_b(l) = 1$ whenever $w_1 + w_2 \ge b - C_\delta$, we also have $G_b(l) = 1$ if $w_2 > b$. This verifies condition (L2). Since both (L1) and (L2) are satisfied, it follows from (32) that if $w_2 < b\delta_+$ for some $\delta_+ < 1/2$, then

$$\begin{aligned}\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\} &\le \kappa_1 h_b\left(l\right) = \kappa_1\int_{b-l}^{b+\kappa_0}\mathbb{P}\left\{X > u\right\}du \\ &\le \kappa_1\mathbb{P}\left\{X > b-l\right\}\left(\kappa_0 + l\right) \le \kappa_1\bar{B}\left(b\left(1-2\delta_+\right)\right)\left(\kappa_0 + 2w_2\right)\left(1+o(1)\right).\end{aligned}$$

The right hand side of the previous inequality is equivalent to the statement of Lemma 16, so we conclude the proof.

Appendix B. Proofs for other estimates

Proof of Lemma 3. First, observe that

$$\begin{aligned}\mathbb{E}\left[(w_1+V-T)^+ - w_1\right] &= \mathbb{E}\left[V-T\right] - \mathbb{E}\left[(V-T)I\left(w_1+V-T<0\right)\right] - w_1\mathbb{P}\left\{w_1+V-T<0\right\} \\ &= -\mathbb{E}\left[(V-T)I\left(w_1+V-T<0\right)\right] - w_1\mathbb{P}\left\{w_1+V-T<0\right\}, \text{ and} \\ \mathbb{E}\left[(w_2-T)^+ - w_2\right] &= -\mathbb{E}T + \mathbb{E}\left[TI\left(w_2-T<0\right)\right] - w_2\mathbb{P}\left\{w_2-T<0\right\}.\end{aligned}$$

Then, it follows from the definition of $\mathbf{W}_1$ in recursions (5a) and (5b) that

$$\begin{aligned}&\mathbb{E}_{(w_1,w_2)}\left[\left(W_1^{(1)}+W_1^{(2)}\right)-(w_1+w_2)\right] = \mathbb{E}\left[(w_1+V-T)^+ - w_1\right] + \mathbb{E}\left[(w_2+T)^+ - w_2\right] \\ &\quad= -\mathbb{E}\left[VI\left(w_1+V-T<0\right)\right] - w_1\mathbb{P}\left\{w_1+V-T<0\right\} - \mathbb{E}\left[TI\left(w_1+V-T\ge 0\right)\right] \\ &\qquad - w_2\mathbb{P}\{w_2-T<0\} + \mathbb{E}\left[TI\left(w_2-T<0\right)\right]\end{aligned}$$

which is negative if $\mathbb{E}[TI(T>w_2)]$ is small enough, and this can be achieved by choosing $C < w_2$ large enough. This completes the proof. □

Proof of Lemma 5. From recursions (5a) and (5b), it is evident that for every $1 \le k \le n$,

$$W_k^{(i)} \le \left(W_{k-1}^{(i)} + X_k\right)^+, \quad i = 1,2.$$

We repeatedly expand the recursion, as below, to obtain

$$\begin{aligned}W_k^{(i)} &\le \max\left\{0,\ W_{k-1}^{(i)} + X_k\right\} \\ &\le \max\left\{0,\ X_k,\ W_{k-2}^{(i)} + X_{k-1} + X_k\right\} \\ &\le \max\left\{0,\ X_k,\ X_{k-1}+X_k,\ X_{k-2}+X_{k-1}+X_k,\ldots,\ W_0^{(i)} + X_1 + \ldots + X_{k-1} + X_k\right\} \\ &\le S_k - \min_{0\le j\le k} S_j + w_i,\end{aligned}$$

where we have used that $S_0 := 0, S_j := X_1 + \ldots + X_j$ and $w_i \geq 0$. Then

$$\max_{0<k\leq n} W_k^{(i)} \leq \max_{0\leq k\leq n} S_k + \max_{0<k\leq n} \max_{0\leq j\leq k} (-S_j) + w_i \leq 2 \max_{0\leq k\leq n} |S_k| + w_i,$$

and this proves the result. □

Proof of Lemma 7. According to Corollary 1 of [18], we have that

$$\mathbb{P}\left\{\max_{0\leq n\leq m} S_n > x\right\} = \left(\mathbb{P}\left\{\max_{0\leq t\leq 1} \sigma B(t) > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\{X > x\}\right)(1 + o(1)). \tag{37}$$

uniformly over $y \geq m^{1/2}$, as $m \to \infty$ (actually, [18] states that the asymptotic is valid assuming $x/m^{1/2} \to \infty$ but the case $x/m^{1/2} = O(1)$ follows from the Central Limit Theorem). Also, from the development in [18], because $\mathbb{P}\{T > x\} = o\left(\bar{B}(x)\right)$, for each $\varepsilon > 0$, there is a positive integer m_ε such that for all $m > m_\varepsilon$,

$$\mathbb{P}\left\{\max_{0\leq n\leq m} (-S_n) > x\right\} \leq (1+\varepsilon)\left(\mathbb{P}\left\{\max_{0\leq t\leq 1} \sigma B(t) > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\{-X > x\}\right). \tag{38}$$

Additionally, since

$$\mathbb{P}\left\{\max_{0\leq n\leq m} |S_n| > x\right\} \leq \mathbb{P}\left\{\max_{0\leq n\leq m} S_n > x\right\} + \mathbb{P}\left\{\max_{0\leq n\leq m} (-S_n) > x\right\}$$

the statement of Lemma 7 immediately follows from (37) and (38). □

Proof of Lemma 9. Let

$$I_1(b) := \mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2|S_n| > (\delta - \delta_-)b, N_A(X) + 1 \leq 2X\right) \max_{0\leq n\leq N_A(X)+1} |S_n| \;\middle|\; X > b\delta_+\right],$$

$$I_2(b) := \mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2|S_n| > (\delta - \delta_-)b, N_A(X) + 1 > 2X\right) \max_{0\leq n\leq N_A(X)+1} |S_n| \;\middle|\; X > b\delta_+\right].$$

Then our objective is to show that $I_1(b) + I_2(b) = O(b^2\bar{B}(b))$. This is an immediate consequence of the following two results.

Lemma 19. *Under Assumption 1 with $\alpha > 2$, and Assumption 2,*

$$I_1(b) = O\left(b^2\bar{B}(b)\right).$$

Lemma 20. *Under Assumption 1 with $\alpha > 2$, and Assumption 2,*

$$I_2(b) = O(\exp(-\nu b)),$$

for a suitable $\nu > 0$.

Proof of Lemma 19. First, observe that

$$I_1(b) \leq \int_{b\delta_+}^{\infty} \mathbb{E}\left[I\left(\max_{0\leq n\leq 2t} |S_n| > \frac{\delta - \delta_-}{2} b\right) \max_{0\leq n\leq 2t} |S_n|\right] \frac{\mathbb{P}\{X \in dt\}}{\mathbb{P}\{X > b\delta_+\}}.$$

Additionally, letting $c = (\delta - \delta_-)/2$, observe that

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq 2t} |S_n| > cb\right) \max_{0\leq n\leq 2t} |S_n|\right] = cb\mathbb{P}\left\{\max_{0\leq n\leq 2t} |S_n| > cb\right\} + \int_{cb}^{\infty} \mathbb{P}\left\{\max_{0\leq n\leq 2t} |S_n| > u\right\} du$$

Therefore, due to (4),

$$I_1(b) = O\left(\frac{1}{\bar{B}(b\delta_+)} \int_{b\delta_+}^{\infty} \left(b\mathbb{P}\left\{\max_{0\leq n\leq 2t} |S_n| > cb\right\} + \int_{cb}^{\infty} \mathbb{P}\left\{\max_{0\leq n\leq 2t} |S_n| > u\right\} du\right) \mathbb{P}\{X \in dt\}\right). \tag{39}$$

Due to the applicability of the uniform asymptotic presented in Lemma 7 in the region $2t \le c^2b^2$, and because of the applicability of Central Limit Theorem in the region $2t > c^2b^2$, we obtain

$$\begin{aligned}
&\int_{b\delta_+}^{\infty} \mathbb{P}\left\{\max_{0\le n\le 2t} |S_n| > cb\right\} \mathbb{P}\{X \in dt\} \\
&\quad = O\left(\int_{b\delta_+}^{\infty} \mathbb{P}\left\{\max_{0\le s\le 1} \sigma\,|B(s)| > \frac{cb}{\sqrt{2t}}\right\} \mathbb{P}\{X\in dt\} + \int_{b\delta_+}^{\frac{c^2b^2}{2}} t\mathbb{P}\left\{|X| > cb\right\} \mathbb{P}\{X\in dt\}\right) \\
&\quad = O\left(b\int_{b\delta_+}^{\infty} \frac{\mathbb{P}\left\{X > t\right\}}{\sqrt{t^3}} \exp\left(-\frac{c^2b^2}{4\sigma^2 t}\right) dt + \mathbb{P}\{|X| > cb\}\mathbb{E}\left[XI\left(X \in \left[b\delta_+, \frac{c^2b^2}{2}\right]\right)\right]\right)
\end{aligned}$$

due to integration by parts. Now, one can apply Lemma 10 to evaluate the first integration, and Karamata's theorem for the second integration, together with the observation that $\mathbb{P}\{|X| > x\} = O(\bar{B}(x))$, to obtain

$$\int_{b\delta_+}^{\infty} \mathbb{P}\left\{\max_{0\le n\le 2t} |S_n| > cb\right\} \mathbb{P}\{X\in dt\} = O\left(\bar{B}\left(b^2\right) + \bar{B}(b)\times b\bar{B}(b)\right) \tag{40}$$

On similar lines of reasoning using Lemma 7, again via careful integration by parts and subsequent application of Lemma 10 and Karamata's theorem, one can derive

$$\begin{aligned}
&\int_{b\delta_+}^{\infty}\int_{cb}^{\infty} \mathbb{P}\left\{\max_{0\le n\le 2t} |S_n| > u\right\} du\mathbb{P}\{X\in dt\} \\
&\quad = O\left(\int_{b\delta_+}^{\infty}\int_{cb}^{\infty} \mathbb{P}\left\{\max_{0\le s\le 1} \sigma\,|B(s)| > \frac{u}{\sqrt{2t}}\right\} du\ \mathbb{P}\{X\in dt\} + \int_{b\delta_+}^{\infty}\int_{\sqrt{2t}}^{\infty} t\mathbb{P}\left\{|X| > u\right\} du\ \mathbb{P}\{X\in dt\}\right) \\
&\quad = O\left(\int_{b\delta_+}^{\infty} \frac{\mathbb{P}\{X>t\}}{\sqrt{t}} \exp\left(-\frac{c^2b^2}{4\sigma^2 t}\right) dt + \int_{\sqrt{2b\delta_+}}^{\infty}\int_{b\delta_+}^{\frac{u^2}{2}} t\mathbb{P}\{X\in dt\}\mathbb{P}\{|X|>u\}\right) \\
&\quad = O\left(b\bar{B}\left(b^2\right) + b\bar{B}(b)\times b\bar{B}(b)\right).
\end{aligned}$$

This bound, along with (39), (40) and the observation that $\bar{B}(b\delta_+) = \Theta(\bar{B}(b))$, prove Lemma 19. □

Proof of Lemma 20. Since $T_1+\ldots+T_{N_A(t)} \le t$ (follows from the definition of $N_A(t)$) and $V_0 := 0$,

$$\begin{aligned}
\mathbb{E}\left[I\left(N_A(t)+1 > 2t\right) \max_{0\le n\le N_A(t)+1} |S_n|\right] &\le \mathbb{E}\left[I\left(N_A(t) > 2t-1\right)\sum_{n=1}^{N_A(t)+1}\left(V_{n-1}+T_n\right)\right] \\
&\le \mathbb{E}\left[I\left(N_A(t) > 2t-1\right)\left(\sum_{n=1}^{N_A(t)} V_n + t + T_{N_A(t)+1}\right)\right] \\
&\le \mathbb{E}V\times\mathbb{E}\left[N_A(t)I\left(N_A(t) > 2t-1\right)\right] + \left(t+\mathbb{E}T\right)\mathbb{P}\left\{N_A(t) > 2t-1\right\} \\
&\le C_1 t\mathbb{P}\left\{N_A(t) > 2t-1\right\} + C_2\int_{2t-1}^{\infty} \mathbb{P}\left\{N_A(t) > s\right\} ds
\end{aligned}$$

for suitable positive constants C_1 and C_2 independent of t. Here, note that the penultimate inequality is simply due to the independence between V_n and T_n for $n \geq 1$. Therefore,

$$\begin{aligned} I_2(b) &\leq \int_{b\delta_+}^{\infty} \mathbb{E}\left[I\left(N_A(t)+1 > 2t\right) \max_{0 \leq n \leq N_A(t)+1} |S_n| \right] \frac{\mathbb{P}\left\{X \in dt\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \\ &\leq C_1 \int_{b\delta_+}^{\infty} t\mathbb{P}\left\{N_A(t) > 2t-1\right\} \frac{\mathbb{P}\left\{X \in dt\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} + C_2 \int_{b\delta_+}^{\infty}\int_{2t-1}^{\infty} \mathbb{P}\left\{N_A(t) > s\right\} ds \frac{\mathbb{P}\left\{X \in dt\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \\ &\leq \mathbb{P}\left\{N_A\left(b\delta_+\right) > 2b\delta_+ - 1\right\} \left(C_1 \int_{b\delta_+}^{\infty} t \frac{\mathbb{P}\left\{X \in dt\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} + C_2 \int_{2b\delta_+ - 1}^{\infty} \frac{\mathbb{P}\left\{X > \frac{s}{2}\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} ds \right), \end{aligned} \tag{41}$$

where we have used a simple change of order of integration to arrive at the above conclusion. Since $N_A(x)/x \to 1$ as $x \to \infty$, the event $\{N_A(b\delta_+) > 2b\delta_+ - 1\}$ is a large deviations event with probability exponentially decaying in b, whereas the sum appearing in the parenthesis in (41) is $O(b)$ due to Karamata's theorem. This proves the claim that $I_2(b) = O(\exp(-\nu b))$ for a suitable constant $\nu > 0$. □

As mentioned earlier, Lemmas 19 and 20, together complete the proof of Lemma 9. □

Proof of Lemma 10. Due to Potter's bounds (25), given $\varepsilon > 0$, we have

$$\int_b^{\infty} \frac{v(t)}{v\left(b^2\right)} \exp\left(-\frac{cb^2}{t}\right) dt = O\left(\int_b^{\infty} \left(\frac{b^2}{t}\right)^{\alpha-\varepsilon} \exp\left(-\frac{cb^2}{t}\right) dt \right)$$

for all suitably large values of b. Changing variables $u = b^2/t$, we obtain

$$\int_b^{\infty} \frac{v(t)}{v\left(b^2\right)} \exp\left(-\frac{cb^2}{t}\right) dt = O\left(b^2 \int_0^{\infty} u^{\alpha-2-\epsilon} \exp(-cu) du \right) = O\left(b^2\right)$$

for all ϵ small enough such that $\alpha - 2 - \epsilon > 0$, and this verifies the claim. □

Proof of Lemma 14. Letting $\bar{c} = (\delta - \delta_-)/2$, observe that

$$\begin{aligned} &\mathbb{E}\left[I\left(\max_{0 \leq n \leq 2X} |S_n| > \bar{c}b \right) \max_{0 \leq n \leq 2X} |S_n| \;\middle|\; X > b\delta_+ \right] \\ &\quad = \bar{c}b\mathbb{P}\left\{ \max_{0 \leq n \leq 2X} |S_n| > \bar{c}b \;\middle|\; X > b\delta_+ \right\} + \int_{\bar{c}b}^{\infty} \mathbb{P}\left\{ \max_{0 \leq n \leq 2X} |S_n| > u \;\middle|\; X > b\delta_+ \right\} du \\ &\quad \leq 3\bar{c}b\mathbb{P}\left\{ Z_* > \frac{\bar{c}b}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+ \right\} + 3\int_{\bar{c}b}^{\infty} \mathbb{P}\left\{ Z_* > \frac{u}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+ \right\} du \end{aligned} \tag{42}$$

because of Lemma 12. Since $\mathbb{P}\{X > x\} \sim cx^{-\alpha}$ as $x \to \infty$, after simple integration using Karamata's theorem (24), one can show that

$$\bar{c}b\mathbb{P}\left\{ Z_* > \frac{\bar{c}b}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+ \right\} = \bar{c}b\mathbb{E}\left[\frac{\mathbb{P}\left\{ X > \frac{1}{2c}\left(\frac{\bar{c}b}{Z_*}\right)^{\alpha} \right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \wedge 1 \right] = O\left(b^2\bar{B}(b)\right), \text{ and}$$

$$\int_{\bar{c}b}^{\infty} \mathbb{P}\left\{ Z_* > \frac{u}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+ \right\} du = \int_{\bar{c}b}^{\infty} \left(\frac{\mathbb{P}\left\{ X > \frac{u^{\alpha}}{2cZ_*^{\alpha}} \right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \wedge 1 \right) du = O\left(b^2\bar{B}(b)\right).$$

Therefore, due to (42), we obtain

$$\mathbb{E}\left[I\left(\max_{0\le n\le 2X}|S_n|>\bar{c}b,\ N_A(X)+1\le 2X\right)\max_{0\le n\le 2X}|S_n|\ \middle|\ X>b\delta_+\right]=O\left(b^2\bar{B}(b)\right).$$

On the other hand, the component corresponding to the large deviations event $\{N_A(X)+1\ge 2X\}$ is handled similar to Lemma 20, and this upper bounding procedure results in

$$\mathbb{E}\left[I\left(\max_{0\le n\le 2X}|S_n|>\bar{c}b,\ N_A(X)+1>2X\right)\max_{0\le n\le 2X}|S_n|\ \middle|\ X>b\delta_+\right]=O\left(\exp(-\nu b)\right),$$

for some $\nu>0$. The last two upper bounds are enough to conclude the statement of Lemma 14. ☐

References

[1] Asmussen, S. *Ruin Probabilities.* Advanced series on statistical science & applied probability. World Scientific Publishing Company, Incorporated, 2000.

[2] Blanchet, J., and Glynn, P. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab. 18*, 4 (08 2008), 1351–1378.

[3] Blanchet, J., Glynn, P., and Liu, J. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *Queueing Systems 57*, 2-3 (2007), 99–113.

[4] Borovkov, A. *Asymptotic methods in queuing theory.* Wiley Series in Probability and Statistics: Probability and Statistics Section Series. J. Wiley, 1984.

[5] Borovkov, A. A., and Borovkov, K. A. On probabilities of large deviations for random walks. I. Regularly varying distribution tails. *Theory of Probability and Its Applications 46*, 2 (2002), 193–213.

[6] Borovkov, A. A., and Borovkov, K. A. *Asymptotic analysis of random walks*, vol. 118 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 2008. Heavy-tailed distributions, Translated from the Russian by O. B. Borovkova.

[7] Denisov, D., Korshunov, D., and Wachtel, V. Potential analysis for positive recurrent markov chains with asymptotically zero drift: Power-type asymptotics. *Stochastic Processes and their Applications 123*, 8 (2013), 3027 – 3051.

[8] Feller, W. *An Introduction to Probability Theory and Its Applications Volume II.* Wiley, 1971.

[9] Foss, S. The method of renovating events and its applications in queueing theory. In *Semi-Markov Models.* Springer, 1986, pp. 337–350.

[10] Foss, S., and Kalashnikov, V. V. Regeneration and renovation in queues. *Queueing Systems 8*, 1 (1991), 211–223.

[11] Foss, S., and Korshunov, D. Heavy tails in multi-server queue. *Queueing Syst. 52*, 1 (2006), 31–48.

[12] Foss, S., and Korshunov, D. On large delays in multi-server queues with heavy tails. *Mathematics of Operations Research 37*, 2 (2012), 201–218.

[13] Kalashnikov, V. Stability estimates for renovative processes. *Eng. Cybern 17* (1980), 85–89.

[14] Kalashnikov, V., and Rachev, S. *Mathematical methods for construction of queueing models.* The Wadsworth & Brooks/Cole operations research series. Wadsworth & Brooks/Cole, 1990.

[15] Meyn, S. P., and Tweedie, R. L. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993.

[16] Murthy, K. *Rare Events in Heavy-tailed Stochastic Systems: Algorithms and Analysis.* PhD thesis, Tata Institute of Fundamental Research, 2015.

[17] Nagaev, S. V. Large deviations of sums of independent random variables. *The Annals of Probability 7*, 5 (10 1979), 745–789.

[18] Pinelis, I. A problem on large deviations in a space of trajectories. *Theory of Probability and Its Applications 26*, 1 (1981), 69–84.

[19] Rozovskii, L. Probabilities of large deviations of sums of independent random variables with common distribution function in the domain of attraction of the normal law. *Theory of Probability and Its Applications 34*, 4 (1989), 625–644.

[20] Scheller-Wolf, A., and Sigman, K. Delay moments for fifo GI/GI/s queues. *Queueing Systems 25*, 1-4 (1997), 77–95.

[21] Scheller-Wolf, A., and Vesilo, R. Sink or swim together: Necessary and sufficient conditions for finite moments of workload components in FIFO multiserver queues. *Queueing Syst. Theory Appl. 67*, 1 (Jan. 2011), 47–61.

[22] Whitt, W. The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. *Queueing Systems 36*, 1-3 (2000), 71–87.

Columbia University, Department of Industrial Engineering & Operations Research, 340 S. W. Mudd Building, 500 West 120 Street, New York, NY 10027, United States.

E-mail address: {jose.blanchet, kra2130}@columbia.edu