Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 164 (2023) 99–138

stochastic processes and their applications

www.elsevier.com/locate/spa

Tail asymptotics for the delay in a Brownian fork-join queue

Dennis Schol[a,*], Maria Vlasiou[b], Bert Zwart[c]

[a] Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[b] University of Twente, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[c] Eindhoven University of Technology, CWI, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 8 August 2022; received in revised form 27 March 2023; accepted 24 June 2023
Available online 3 July 2023

Abstract

We study the tail behavior of $\max_{i\leq N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s)$ as $N\to\infty$, with $(W_i, i\leq N)$ i.i.d. Brownian motions and W_A an independent Brownian motion. This random variable can be seen as the maximum of N mutually dependent Brownian queues, which in turn can be interpreted as the backlog in a Brownian fork-join queue. In previous work, we have shown that this random variable centers around $\frac{\sigma^2}{2\beta}\log N$. Here, we analyze the rare event that this random variable reaches the value $(\frac{\sigma^2}{2\beta}+a)\log N$, with $a>0$. It turns out that its probability behaves roughly as a power law with N, where the exponent depends on a. However, there are three regimes, around a critical point $a^\star$; namely, $0<a<a^\star$, $a=a^\star$, and $a>a^\star$. The latter regime exhibits a form of asymptotic independence, while the first regime reveals highly irregular behavior with a clear dependence structure among the N suprema, with a nontrivial transition at $a=a^\star$.

Keywords: Brownian queues; Fork-join queues; Extreme-value theory; Tail asymptotics

1. Introduction

Fork-join queues are a useful modeling tool for congestion in complex networks, such as assembly systems, communication networks, and supply chains. Such networks can be large and assembly is only possible upon availability of all parts. Thus, the bottleneck of the system is caused by the slowest production line in the system. This setting motivates us to investigate

* Corresponding author.
E-mail addresses: dennisschol@live.nl (D. Schol), m.vlasiou@tue.nl (M. Vlasiou), bert.zwart@cwi.nl (B. Zwart).

https://doi.org/10.1016/j.spa.2023.06.013
0304-4149/

such delays in a stylized version of a large fork-join queueing system. In this setting, a key quantity of interest is the behavior of the longest queue. We assume that arrival and service processes are Brownian, as it is a standard result in queueing theory that queueing systems in heavy-traffic can be approximated by reflected Brownian motions. Furthermore, when the arrival and service processes are deterministic with some white noise perturbation, it is also a natural choice to model this with Brownian motions. We analyze the steady-state behavior of this system. Hence, we can model the backlog in queue i by $Q^{\beta}_{i,A} = \sup_{s>0}(W_i(s)+W_A(s)-\beta s)$, where W_A is a Brownian motion term with standard deviation σ_A that represents the fluctuations in the arrival process, W_i is a Brownian motion term with standard deviation σ that represents the fluctuations in the service process, and $\beta > 0$ represents the drift of the queue. Furthermore, we assume that $(W_i, i \le N)$ are i.i.d. Brownian motions, and for all i, W_i and W_A are mutually independent. These are natural choices as well, because these assumptions indicate that servers' work speeds are mutually independent, and independent with respect to the interarrival times.

Because the bottleneck in the system is the slowest production line, we are interested in the longest queue length, and we investigate the random variable $\bar{Q}^{\beta}_N = \max_{i\le N} Q^{\beta}_{i,A}$. We see that this random variable is a maximum of N dependent random variables, due to the common arrival process W_A. As we try to model systems with many servers, we are typically interested in the behavior of this random variable as $N \to \infty$. In [16], it is shown that $\bar{Q}^{\beta}_N$ is in the domain of attraction of the normal distribution:

$$\mathbb{P}\left(\bar{Q}^{\beta}_N > \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \xrightarrow{N\to\infty} \mathbb{P}\left(\frac{\sigma\sigma_A}{\sqrt{2}\beta}X > x\right), \tag{1}$$

with $X \stackrel{d}{=} \mathcal{N}(0,1)$. This means that $\bar{Q}^{\beta}_N$ centers around $\frac{\sigma^2}{2\beta}\log N$ and deviates with order $\sqrt{\log N}$.

This convergence result provides a prediction of the typical delay. However, one might also be interested in the question how likely it is that the delay will be much longer, as delays may cause large costs. Obviously, the probability $\mathbb{P}(\bar{Q}^{\beta}_N > y_N) \xrightarrow{N\to\infty} 0$, when $y_N - \frac{\sigma^2}{2\beta}\log N$ grows to infinity at a rate faster than $\sqrt{\log N}$, but the question is how fast this probability converges to 0. In this study, we focus on the probability

$$\mathbb{P}\left(\bar{Q}^{\beta}_N > \left(\frac{\sigma^2}{2\beta} + a\right)\log N\right),$$

with $a > 0$. As we show later on, the exact behavior of this tail probability depends on the choice of a, where we can distinguish three regimes: $0 < a < a^\star$, $a = a^\star$, and $a > a^\star$, with $a^\star$ an explicitly identified constant in $(0,\infty)$. The logarithmic asymptotics for these three regimes are given in Theorem 1, while sharper asymptotics for the cases $a > a^\star$, $a = a^\star$, and $0 < a < a^\star$ are given in Theorems 2, 3, and 4, respectively. It easily follows from the proofs that when y_N is of larger order than $\log N$, the convergence behavior of $\mathbb{P}(\bar{Q}^{\beta}_N > y_N)$ is the same as for the case $a > a^\star$, cf. Corollary 2.

Our work is related to the literature on extreme values of Gaussian processes. In this paper, we examine exceedance probabilities of the order $(\frac{\sigma^2}{2\beta} + a)\log N$ with $a > 0$. More work has been done on joint suprema of Brownian motions. For instance, [11] gives the solution of the Laplace transform of joint first passage times in terms of the solution of a partial differential equation, where the Brownian motions are dependent. Further, [6] analyze the tail asymptotics of the all-time suprema of two dependent Brownian motions. The joint suprema of a finite number of Brownian motions is also studied [5], where the authors give tail asymptotics of the joint suprema of independent Gaussian processes over a finite time interval. These are just three examples — more results may be found in [15,20].

Our work also relates to the literature on fork-join queues. Exact results on fork-join queues with two service stations can be found in [2,7,9,23]. Approximations for systems with an arbitrary but fixed number of servers can be found in [3,10,17]. In [22] a heavy-traffic analysis for fork-join queues is derived; see also [18,19]. More recent work in this direction may be found in [12–14,21]. Our work adds to the existing literature, as we analyze the largest of N queues as $N \to \infty$. Literature on such extreme-value results is rare. More specifically, we derive a large deviation principle for the longest of N dependent Brownian queues as $N \to \infty$, to obtain this, we use and extend the results obtained in [6], in which the case $N = 2$ is investigated.

This paper is organized as follows. In Section 2, we present our main results, which contain an interesting phase transition in the way a large supremum occurs depending on the value of a. We explain the reason behind this phase transition in detail. The rest of the paper is devoted to proofs. In Section 3, we give a proof of Theorem 1, which focuses on logarithmic asymptotics. In Section 4, we present some auxiliary lemmas. In Sections 5.1, 5.2, and 5.3, we provide the proofs of Theorems 2, 3, and 4, respectively, which deal with asymptotic estimates that are sharper than Theorem 1.

2. Main results

In this section, we present our main results and also provide some intuition. We first introduce some additional notation.

Definition 1. The sequence $(W_i, i \leq N)$ is a sequence of i.i.d. Brownian motions with standard deviation σ, $\{W_A(t), t \geq 0\}$ is a Brownian motion with standard deviation σ_A, $\{W_i(t), t \geq 0\}$ and $\{W_A(t), t \geq 0\}$ are mutually independent for all i, the steady-state queue length in front of server i is given by

$$Q^{\beta}_{i,A} := \sup_{s>0}(W_i(s) + W_A(s) - \beta s), \tag{2}$$

and the maximum queue length equals

$$\bar{Q}^{\beta}_{N} := \max_{i \leq N} Q^{\beta}_{i,A}. \tag{3}$$

Next, we write the supremum of a Brownian motion $\{W_i(t) + W_A(t) - \beta t, t \geq 0\}$ over an interval (u, v) as

$$Q^{\beta}_{i,A}(u, v) := \sup_{u<s<v}(W_i(s) + W_A(s) - \beta s), \tag{4}$$

and the maximum of N of these identically distributed random variables as

$$\bar{Q}^{\beta}_{N}(u, v) := \max_{i \leq N} Q^{\beta}_{i,A}(u, v). \tag{5}$$

Furthermore, we write $Q^{\beta}_{i,A}(u) = Q^{\beta}_{i,A}(u, \infty)$ and $\bar{Q}^{\beta}_{N}(u) = \bar{Q}^{\beta}_{N}(u, \infty)$.

We give additional shorthand notation that we use later on.

Definition 2.

$$f_N(a) := \left(\frac{\sigma^2}{2\beta} + a\right) \log N, \tag{6}$$

$$\lambda(a) := 1 - \sigma/\sqrt{2a\beta + \sigma^2}, \tag{7}$$

$$T_N(a,k) := f_N(a)/\beta + k\sqrt{\log N}, \tag{8}$$

$$T_N(a) := T_N(a,0). \tag{9}$$

Finally, we define

$$\gamma(a) := \begin{cases} \dfrac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2} & \text{if } 0 < a < a^\star, \\ \dfrac{2a\beta - \sigma_A^2}{\sigma^2 + \sigma_A^2} & \text{if } a \geq a^\star, \end{cases} \tag{10}$$

with

$$a^\star := \frac{\sigma_A^4}{\sigma^2 2\beta} + \frac{\sigma_A^2}{\beta}.$$

The function $\gamma(a)$ appears in the limit of the logarithmic asymptotics of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. As can be seen from (10), from $a = a^\star$ onwards, the function $\gamma(a)$ is linear. Moreover, we see that $\gamma(a)$ is continuous everywhere, also for $a = a^\star$. In Fig. 1, we plot $-\gamma(a)$ for certain choices of the parameters σ, σ_A, β, and $a^\star$.

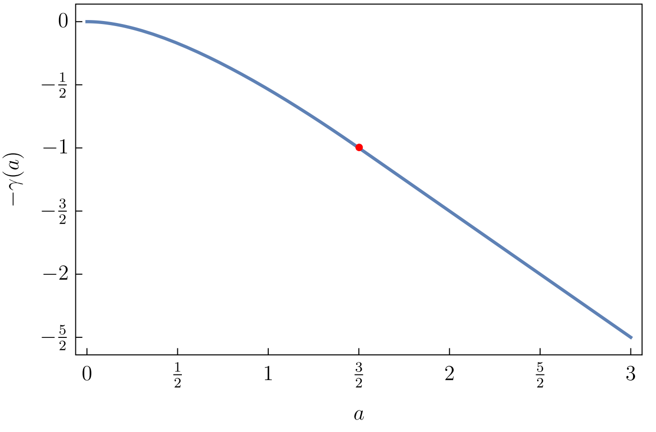

Fig. 1. $\sigma = 1, \sigma_A = 1, \beta = 1, a^\star = 3/2$.

Throughout this paper, we analyze the fork-join queueing system as defined in Definitions 1 and 2. Our first result, Theorem 1, provides the logarithmic asymptotics of the tail probability of the maximum steady-state queue length $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$.

Theorem 1. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a > 0$, we have that*

$$\frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \overset{N\to\infty}{\longrightarrow} -\gamma(a). \tag{11}$$

We give the proof of Theorem 1 in Section 3. To provide some intuition, the form of the function $\gamma(a)$ suggests there are at least two regimes: the case where $0 < a < a^\star$, and the case where $a \geq a^\star$. These two cases reveal interesting information on the tail behavior of the maximum queue length $\bar{Q}_N^\beta$.

Case $a > a^\star$. First, we give some intuitive explanation for the case $a > a^\star$. The maximum steady-state queue length is the maximum of N dependent exponentially distributed random variables. We can use the memoryless property of the exponential distribution to get some heuristic insights into the behavior of the maximum steady-state queue length. Define $\tau := \inf\{t > 0 : \max_{i\le N} W_i(t)+W_A(t)-\beta t \ge f_N(a^\star)\}$ and let $i^\star \in \{j \le N : W_j(\tau)+W_A(\tau)-\beta\tau = \max_{i\le N} W_i(\tau)+W_A(\tau)-\beta\tau\}$. Then, we get

$$\begin{aligned}
&\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \\
&\quad= \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a)\right) \\
&\quad= \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a) \,\Big|\, \tau<\infty\right)\mathbb{P}(\tau<\infty) \\
&\quad\ge \mathbb{P}\left(\sup_{s>\tau}(W_{i^\star}(s)+W_A(s)-\beta s) > f_N(a) \,\Big|\, \tau<\infty\right)\mathbb{P}(\tau<\infty).
\end{aligned} \tag{12}$$

Now, due to the fact that Brownian motions have independent increments, we can write $\sup_{s>\tau}(W_{i^\star}(s)+W_A(s)-\beta s) \stackrel{d}{=} W_{i^\star}(\tau)+W_A(\tau)-\beta\tau+\sup_{s>0}(\hat{W}_{i^\star}(s)+\hat{W}_A(s)-\beta s)$, with $\{\hat{W}_{i^\star}(t), t\ge 0\}$ and $\{\hat{W}_A(t), t\ge 0\}$ independent copies of $\{W_{i^\star}(t), t\ge 0\}$ and $\{W_A(t), t\ge 0\}$, respectively. Thus, the lower bound in (12) simplifies to

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star)\right) \\
&\quad\times\mathbb{P}\left(\sup_{s>0}(W_{i^\star}(s)+W_A(s)-\beta s) > (a-a^\star)\log N\right).
\end{aligned}$$

Therefore, when we compare this lower bound with the convergence result given in (11), we get that

$$\begin{aligned}
&\mathbb{P}\left(\bar{Q}_N^\beta > f_N(a)\right) \\
&\quad= \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a)\right) \\
&\quad\ge \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star)\right) \\
&\qquad\times\mathbb{P}\left(\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > (a-a^\star)\log N\right) \\
&\quad= \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star)\right)\exp\left(-\frac{2\beta(a-a^\star)}{\sigma^2+\sigma_A^2}\log N\right) \\
&\quad\approx N^{-\gamma(a^\star)}\exp\left(-\frac{2\beta(a-a^\star)}{\sigma^2+\sigma_A^2}\log N\right) \\
&\quad= N^{-\gamma(a)},
\end{aligned} \tag{13}$$

with the "≈" sign indicating that we use the logarithmic asymptotics from (11) to approximate $\mathbb{P}(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star))$ with $N^{-\gamma(a^\star)}$, while ignoring lower-order terms. Thus, we see that when we use the result from (11) for $a = a^\star$, then this lower bound is sharp in the logarithmic sense for $a > a^\star$. Furthermore, this derivation heuristically explains that the function $\gamma(a)$ is linear for $a \ge a^\star$.

The second intuitive observation is that for $a \geq a^\star$, $N^{-\gamma(a)} = N\mathbb{P}(Q^\beta_{i,A} > f_N(a))$. Obviously, since $a \geq 0$, the union bound gives that

$$\mathbb{P}(\bar{Q}^\beta_N > f_N(a)) \leq N\mathbb{P}(Q^\beta_{i,A} > f_N(a)) = N^{-\frac{2a\beta-\sigma_A^2}{\sigma^2+\sigma_A^2}}. \tag{14}$$

The fact that the union bound is sharp when $a \geq a^\star$ indicates that for $a \geq a^\star$, the N queues are almost asymptotically independent; i.e.,

$$\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a)\right)$$
$$\approx \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}\big(W_i(s)+W_{A,i}(s)-\beta s\big) > f_N(a)\right),$$

with the "≈" sign again indicating that we use the logarithmic asymptotics from (11), but we ignore lower-order terms. Here, the arrival processes $(W_{A,i}, i \leq N)$ are independent Brownian motions, and $\{W_{A,i}(t), t \geq 0\}$ and $\{W_i(t), t \geq 0\}$ are mutually independent. In Section 5.2, we see that the boundary case $a = a^\star$ does show some dependent behavior, but this dependence structure cannot be deduced from the logarithmic asymptotics.

Case $0 < a < a^\star$. Finally, the case $0 < a < a^\star$ is more involved. The function $\gamma(a)$ involves a in a nonlinear fashion. As we observe in Eq. (14), due to the fact that the exponent of the tail probability of an exponentially distributed random variable is linear in a, we expect that the logarithmic asymptotics are also linear in a. Thus, the structure of $\gamma(a)$ shows that the dependent part W_A influences the tail asymptotics, and we have that

$$\liminf_{N\to\infty}\mathbb{P}\left(\#\{j\leq N : \sup_{s>0}(W_j(s)+W_A(s)-\beta s) > f_N(a)\} > 1 \,\middle|\, \bar{Q}^\beta_N > f_N(a)\right) > 0.$$

The reason that we see this is that in order to get that the maximum steady-state queue length $\bar{Q}^\beta_N$ reaches the level $f_N(a)$, the arrival process $\{W_A(t) - \lambda(a)\beta t, t \geq 0\}$ must reach a high level around $\lambda(a)f_N(a)$, which is a rare event; see Eq. (39). Furthermore, one of the N service processes needs to reach a level around $(1-\lambda(a))f_N(a)$; however, this is not a rare event. Even more, the event that a finite number of service processes reaches a level around $(1-\lambda(a))f_N(a)$ has a non-zero probability; see Eq. (40).

The function $\gamma(a)$ has more characteristics that can be explained from the limit in (1). For example, $\gamma(0) = 0$, which is to be expected as we know from (1) and (6) that for $x = 0$

$$\mathbb{P}(\bar{Q}^\beta_N > f_N(0)) \xrightarrow{N\to\infty} \frac{1}{2}.$$

We further have that $(\log N)\gamma(x/\sqrt{\log N}) \xrightarrow{N\to\infty} \frac{x^2\beta^2}{\sigma^2\sigma_A^2}$. It thus follows that for N large,

$$N^{-\gamma(x/\sqrt{\log N})} \approx N^{-\frac{x^2\beta^2}{\sigma^2\sigma_A^2\log N}} = \exp\left(-\frac{x^2\beta^2}{\sigma^2\sigma_A^2}\right),$$

which is the exponent of the limiting distribution given in (1).

To prove the logarithmic asymptotics in Theorem 1, it suffices to look at random variables of the type $\max_{i\leq N}(W_i(T_N) + W_A(T_N) - \beta T_N)$ instead of the random variable $\bar{Q}^\beta_N = \max_{i\leq N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s)$, where the appropriate choice of T_N is $T_N(a)$; see Eq. (9). We show this in more detail in the proof of Lemma 1. For $a > a^\star$, the logarithmic asymptotics are relatively straightforward to derive because we see a notion of asymptotic independence, as explained above. In the proof of Lemma 1, we show that when $0 < a \le a^\star$,

$$\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a))) \approx \log(\mathbb{P}(\max_{i\le N} W_i(T_N(a)) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a))f_N(a))) + \log(\mathbb{P}(W_A(T_N(a)) - \lambda(a)\beta T_N(a) > \lambda(a)f_N(a))), \tag{15}$$

when N is large, and we show that the term $\log(\mathbb{P}(\max_{i\le N} W_i(T_N(a)) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a))f_N(a)))$ becomes negligible as $N\to\infty$.

We now turn to precise asymptotics, which are stated in Theorems 2, 3, and 4 for the cases $a > a^\star$, $a = a^\star$, and $0 < a < a^\star$, respectively. The proofs of these theorems can be found in Sections Section 5.1, 5.2, and 5.3.

Theorem 2. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a > a^\star$, we have that*

$$N^{\gamma(a)}\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \overset{N\to\infty}{\longrightarrow} 1. \tag{16}$$

The theorem shows that for $a > a^\star$, the tail probability of the steady-state maximum queue length has the same asymptotic behavior as the one for independently and identically distributed arrival processes for each queue.

Theorem 3. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a = a^\star$, we have that*

$$N^{\gamma(a^\star)}\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \overset{N\to\infty}{\longrightarrow} \frac{1}{2}. \tag{17}$$

To give a heuristic explanation of why we have a transition point at $a = a^\star$, we argue as follows. Because the all-time supremum of a Brownian motion is exponentially distributed it is easy to see that for $a = a^\star$,

$$\sup_{s>0}(W_A(s) - \lambda(a^\star)\beta s) \overset{d}{=} \sup_{s>0}(W_i(s) - (1-\lambda(a^\star))\beta s) \overset{d}{=} \sup_{s>0}(W_i(s) + W_A(s) - \beta s),$$

where $\lambda(a)$ is given in Eq. (7). Similarly, after a straightforward calculation, we observe that for $0 < a < a^\star$,

$$\sup_{s>0}(W_A(s) - \lambda(a)\beta s) \ge_{st.} \sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s),$$

and for $a > a^\star$,

$$\sup_{s>0}(W_A(s) - \lambda(a)\beta s) \le_{st.} \sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s),$$

with $X \ge_{st.} Y$ meaning that $\mathbb{P}(X \ge x) \ge \mathbb{P}(Y \ge x)$ for all x. For $0 < a < a^\star$, large values of $\bar{Q}_N^\beta$ are predominantly caused by fluctuations of $\{W_A(t) - \lambda(a)\beta t, t \ge 0\}$; we show this rigorously in Section 5.3. In contrast, for $a > a^\star$, fluctuations are caused by a combination of the arrival process and one of the service processes, and therefore we see a notion of asymptotic independence.

To explain in more detail why we have a constant 1/2 at the boundary case $a = a^\star$, we first let $\hat{Q}_{i,A}^\beta$ be an independent copy of $Q_{i,A}^\beta$. Furthermore, observe that since the all-time supremum of a Brownian motion with negative drift is exponentially distributed, $\mathbb{P}(\sup_{s>0}(W_A(s) - \lambda(a^\star)\beta s) > \lambda(a^\star)f_N(a^\star)) = N^{-\gamma(a^\star)}$. Moreover, if the event $\sup_{s>0}(W_A(s) - \lambda(a^\star)\beta s)$

$> \lambda(a^\star) f_N(a^\star)$ happens, it most likely occurs at time $T_N(a^\star)$. By using the union bound and that all suprema follow the same distribution, we may therefore write

$$\begin{aligned}
&\mathbb{P}(\bar{Q}^\beta_N(T_N(a^\star)) > f_N(a^\star) | W_A\big(T_N(a^\star)\big) - \lambda(a^\star)\beta T_N(a^\star) = \lambda(a^\star) f_N(a^\star))\\
&\quad = \mathbb{P}\left(\max_{i\le N}\left(W_i\big(T_N(a^\star)\big) - (1-\lambda(a^\star))\beta T_N(a^\star) + \hat{Q}^\beta_{i,A}\right) > (1-\lambda(a^\star)) f_N(a^\star)\right)\\
&\quad \approx N\mathbb{P}\left(W_i\big(T_N(a^\star)\big) - (1-\lambda(a^\star))\beta T_N(a^\star) + \hat{Q}^\beta_{i,A} > (1-\lambda(a^\star)) f_N(a^\star)\right)\\
&\quad = N\mathbb{P}\left(\sup_{s>T_N(a^\star)}(W_i(s) - (1-\lambda(a^\star))\beta s) > (1-\lambda(a^\star)) f_N(a^\star)\right) \overset{N\to\infty}{\longrightarrow} \frac{1}{2}.
\end{aligned}$$

The reason that we see a factor 1/2 emerging in the limit follows from the fact that we take the supremum over the set $(T_N(a^\star), \infty)$. As the all-time suprema of Brownian motions are exponentially distributed, it is easy to see that

$$N\mathbb{P}\left(\sup_{s>0}(W_i(s) - (1-\lambda(a^\star))\beta s) > (1-\lambda(a^\star)) f_N(a^\star)\right) \overset{N\to\infty}{\longrightarrow} 1.$$

Typical hitting times of this supremum are of the form $T_N(a^\star) + k\sqrt{\log N}$, with $k \in \mathbb{R}$. We will see in the proofs that the density of these hitting times will deviate symmetrically around $T_N(a^\star)$; see Lemma 4. This heuristically explains that when we take the supremum over the set $(T_N(a^\star), \infty)$, we obtain the limit of 1/2. If we condition on $\max_{i\le N}\sup_{s>0}(W_i(s) - (1 - \lambda(a^\star))\beta s) = (1-\lambda(a^\star)) f_N(a^\star)$, we obtain the same expression after using the same heuristic argument.

Our final result is an improvement of the logarithmic asymptotics for the case $0 < a < a^\star$.

Theorem 4. *For the model given in Definition 1 with the additional notation given in Definition 2, and $0 < a < a^\star$, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\bar{Q}^\beta_N > f_N(a)) > 0, \tag{18}$$

and

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\bar{Q}^\beta_N > f_N(a)) < \infty. \tag{19}$$

Remark 1. We can prove lower and upper bounds that are sharper than logarithmic. However, we do not specify these bounds, but from the proof of Theorem 4 it becomes clear that

$$\begin{aligned}
&\liminf_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\bar{Q}^\beta_N > f_N(a))\\
&\ge \int_{-\infty}^{\infty} \frac{\beta^2\left(\sigma\left(\sigma - \sqrt{2a\beta+\sigma^2}\right) + 2a\beta\right)\exp\left(-\frac{\beta^4k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(2a\beta+\sigma^2\right)^{3/2}}\\
&\quad\times\left(1-\exp\left(-\frac{\exp\left(-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}}\right)\right)dk,
\end{aligned}$$

and

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a))$$
$$\leq \int_{-\infty}^{\infty}\left(\frac{\sigma_A \exp\left(-\frac{\beta^4k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2(2a\beta+\sigma^2)^2}\right)}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}\right.$$
$$\left.+\frac{\sigma_A\left(\sigma^2+\sigma_A^2\right)\exp\left(-\frac{2\beta^4k^2\left(\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)+a\beta\left(\sqrt{2a\beta+\sigma^2}-2\sigma\right)\right)}{\sigma_A^2(2a\beta+\sigma^2)^{5/2}}\right)}{2\sqrt{\pi}\sigma\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)}\right)$$
$$\times\frac{\beta^2 e^{-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}}}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}dk+1.$$

We give a proof of Theorem 4 in Section 5.3. As already suggested in Theorem 1, for the case $0 < a < a^\star$ we observe more irregular behavior, which manifests itself already in the values of $\gamma(a)$. In Theorem 4, we observe that the second term is not a constant, as was the case for the values $a > a^\star$ and $a = a^\star$, but is $(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}$. To obtain heuristic insights, we argue that

$$\mathbb{P}\left(\sup_{s>0}(W_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N\right) = \exp\left(-\frac{2\lambda(a)\beta}{\sigma_A^2}(\lambda(a)f_N(a)+r_N)\right)$$
$$= N^{-\gamma(a)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}, \tag{20}$$

with $r_N = \frac{\sigma\sqrt{2a\beta+\sigma^2}}{4\beta}\log\log N$. Furthermore, we have for all k that

$$\mathbb{P}\left(\max_{i\leq N} W_i\big(T_N(a,k)\big)-(1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a)-r_N\right) = \Theta(1), \tag{21}$$

where $z_N = \Theta(1)$ means that $\liminf_{N\to\infty} z_N > 0$ and $\limsup_{N\to\infty} z_N < \infty$. Combining these two results together with the definition of $\bar{Q}_N^{\beta}$ in (3), we see that

$$\mathbb{P}\left(\bar{Q}_N^{\beta} > f_N(a)\right)$$
$$\geq \mathbb{P}\left(\sup_{s>0}(W_A(s)-\lambda(a)\beta s) > \lambda(a)f_N(a)+r_N,\right.$$
$$\left.\max_{i\leq N} W_i(\tau^{(N)})-(1-\lambda(a))\beta\tau^{(N)} > (1-\lambda(a))f_N(a)-r_N\right), \tag{22}$$

where $\tau^{(N)} = \inf\{t \geq 0 : W_A(t)-\lambda(a)\beta t > \lambda(a)f_N(a)+r_N\}$. We show later on that $\tau^{(N)}$, conditioned on being finite, is of the form $T_N(a,K)$, with $T_N(a,\cdot)$ defined in (8) and K being a random variable. Because

$$
\begin{aligned}
&\mathbb{P}\Bigg(\sup_{s>0}(W_A(s)-\lambda(a)\beta s)>\lambda(a)f_N(a)+r_N,\\
&\qquad \max_{i\le N} W_i(\tau^{(N)})-(1-\lambda(a))\beta\tau^{(N)}>(1-\lambda(a))f_N(a)-r_N\Bigg)\\
&\quad=\mathbb{P}\left(\sup_{s>0}(W_A(s)-\lambda(a)\beta s)>\lambda(a)f_N(a)+r_N\right)\\
&\qquad\times\mathbb{P}\left(\max_{i\le N} W_i(\tau^{(N)})-(1-\lambda(a))\beta\tau^{(N)}>(1-\lambda(a))f_N(a)-r_N\,\middle|\,\tau^{(N)}<\infty\right),
\end{aligned}
\tag{23}
$$

we retrieve (18) after combining the results from (20)–(23). Thus, it turns out that for $0 < a < a^\star$, r_N plays a key role. As explained in Section 5.2, in the case $0 < a < a^\star$, $\{W_A(t)-\lambda(a)\beta t, t\ge 0\}$ dominates, which explains why the tail asymptotics of the maximum queue length $\bar{Q}_N^\beta$ are the same as the tail asymptotics of $\sup_{s>0}(W_A(s)-\lambda(a)\beta s)$, and the behavior of $\max_{i\le N} W_i\big(T_N(a,k)\big)-(1-\lambda(a))\beta T_N(a,k)$ is typical.

The main approach of proving the lower and upper bounds in (18) and (19), as well as the limits in (16) and (17), is by analyzing lower and upper bounds on the tail probability of the steady-state maximum queue length $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. These bounds are derived by utilizing the union bound, Bonferroni's inequality, and a careful construction of hitting times. These hitting times are needed to estimate the time when the supremum most likely hits the desired level and to adequately separate the independent part W_i and the dependent part W_A from each other. We also rely on some existing asymptotic estimates in the literature from extreme-value theory, and on [6], which investigates the case $N = 2$. Finally, we develop a number of auxiliary technical estimates related to the asymptotic behavior of convolutions of normally and exponentially distributed random variables.

These techniques, when put together, are effective in the case $a = a^\star$ and $a > a^\star$ in order to obtain exact asymptotics. In the case $0 < a < a^\star$, we are able to improve upon Theorem 1 and characterize the asymptotic behavior of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$ up to a constant. To derive precise asymptotics in this case seems beyond the scope of the techniques developed in this paper.

3. Proof of the logarithmic asymptotics

In this section, we give a proof of Theorem 1, establishing logarithmic asymptotics for the maximum queue length. Our approach is to derive logarithmic lower and upper bounds of the maximum queue length by formalizing the heuristic idea given in (15), and show that they coincide. These bounds are presented in Lemmas 1 and 2.

Lemma 1. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a > 0$, we have that*

$$\liminf_{N\to\infty}\frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N}\ge -\gamma(a). \tag{24}$$

Proof. Recall that $\lambda(a) = 1-\sigma/\sqrt{2a\beta+\sigma^2}$ and $T_N(a) = f_N(a)/\beta$. By choosing $s = f_N(a)/\beta$ and splitting $-\beta s$ into two terms, observe that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\le N}\sup_{s>0}\left(W_i(s)+W_A(s)-\beta s\right) > f_N(a)\right) \\
&\quad\ge \mathbb{P}\Bigg(\max_{i\le N} W_i\big(T_N(a)\big) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a)) f_N(a), \\
&\qquad\qquad W_A\big(T_N(a)\big) - \lambda(a)\beta T_N(a) > \lambda(a) f_N(a)\Bigg) \qquad (25)\\
&\quad= \mathbb{P}\left(\max_{i\le N} W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right)\mathbb{P}\left(W_A\big(T_N(a)\big) > 2\lambda(a) f_N(a)\right). \qquad (26)
\end{aligned}$$

The expression in (26) is due to the fact that for all i, $\{W_i(t), t \ge 0\}$ and $\{W_A(t), t \ge 0\}$ are independent. We now analyze the two probabilities in (26) separately. Since $\{W_i(t), t \ge 0\}$ and $\{W_j(t), t \ge 0\}$ are i.i.d. for all i and j, for the first probability in (26) we get from Bonferroni's inequality that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\le N} W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right) \\
&\quad\ge N\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right) \qquad (27)\\
&\qquad-\binom{N}{2}\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right)^2.
\end{aligned}$$

Furthermore, it is easy to see that

$$\mathbb{P}\left(\sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a)) f_N(a)\right) = \frac{1}{N} \qquad (28)$$

and that

$$\begin{aligned}
&\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right) \\
&\quad\le \mathbb{P}\left(\sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a)) f_N(a)\right),
\end{aligned}$$

and therefore we bound the second term on the right-hand side of (27) as

$$\begin{aligned}
&\binom{N}{2}\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right)^2 \\
&\quad\le \frac{N^2}{2}\mathbb{P}\left(\sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s) > (1-\lambda(a)) f_N(a)\right) \\
&\qquad\times\ \mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right) \\
&\quad= \frac{N}{2}\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right).
\end{aligned}$$

Thus, the lower bound given in (27) can be further bounded by

$$\mathbb{P}\left(\max_{i\le N} W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right) \ge \frac{N}{2}\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a)) f_N(a)\right).$$

As we aim to derive logarithmic asymptotics, we see that

$$\log\left(\frac{N}{2}\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\right)\right)$$
$$\sim \log N + \log\left(\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\right)\right),$$

as $N \to \infty$, with $f(x) \sim g(x)$ as $x \to \infty$ meaning that $\lim_{x\to\infty} f(x)/g(x) = 1$. In addition, recall that for a normally distributed random variable X with standard deviation σ, $\log(\mathbb{P}(X > x)) \sim -x^2/(2\sigma^2)$, as $x \to \infty$. Thus, we get that

$$\log\left(\mathbb{P}\left(W_i\big(T_N(a)\big) > 2(1-\lambda(a))f_N(a)\right)\right) \sim -\frac{(2(1-\lambda(a))f_N(a))^2}{2\sigma^2 T_N(a)} = -\log N,$$

as $N \to \infty$, following the definitions of $\lambda(a)$, $f_N(a)$, and $T_N(a)$. Concluding,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\left(\max_{i\le N} W_i\big(T_N(a)\big) - (1-\lambda(a))\beta T_N(a) > (1-\lambda(a))f_N(a)\right)\right)}{\log N} \ge 0. \quad (29)$$

For the second probability in (26), the logarithmic asymptotics can be easily computed since $W_A\big(f_N(a)\big)$ is normally distributed. We obtain that

$$\frac{\log\left(\mathbb{P}\left(W_A\big(T_N(a)\big) > 2\lambda(a)f_N(a)\right)\right)}{\log N} \overset{N\to\infty}{\longrightarrow} -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}. \quad (30)$$

Thus, after combining these two results in (29) and (30) with (26), we have that,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a)\right)\right)}{\log N} \ge -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}, \quad (31)$$

irrespective of the choice of a. Now, observe that for $a > 0$,

$$\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2} \ge \frac{2a\beta - \sigma_A^2}{\sigma^2+\sigma_A^2},$$

with equality for $a = a^\star$. This means that only for $0 < a \le a^\star$, the lower bound in (31) is sharp enough. For $a > a^\star$, we apply the inequality in (13) to obtain for all $c > 0$ that

$$\mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star+c)\right)$$
$$\ge \mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star)\right)\exp\left(-\frac{2\beta c\log N}{\sigma^2+\sigma_A^2}\right). \quad (32)$$

Combining this result with the inequality in (31), we get that for all $c > 0$,

$$\liminf_{N\to\infty} \frac{\log\left(\mathbb{P}\left(\max_{i\le N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s) > f_N(a^\star+c)\right)\right)}{\log N}$$
$$\ge -\gamma(a^\star) - \frac{2\beta c}{\sigma^2+\sigma_A^2} = -\gamma(a^\star+c).$$

Combining the lower bounds in (31) and (32) gives the lower bound in (24). □

Lemma 2. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a > 0$, we have that*

$$\limsup_{N\to\infty} \frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \leq -\gamma(a). \tag{33}$$

Proof. We have by the union bound in (14) that

$$\limsup_{N\to\infty} \frac{\log(\mathbb{P}(\bar{Q}_N^\beta > f_N(a)))}{\log N} \leq -\frac{2a\beta - \sigma_A^2}{\sigma^2 + \sigma_A^2}. \tag{34}$$

This upper bound implies the upper bound given in (33) for $a \geq a^\star$. Turning to the case $0 < a < a^\star$, we can bound the tail probability of the maximum queue length by using subadditivity, the union bound, and by integrating over possible values of $\sup_{s>0}(W_A(s) - \lambda(a)\beta s)$, and we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\bar{Q}_N^\beta > f_N(a)\right) &(35)\\
&\leq \mathbb{P}\left(\max_{i\leq N}\sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s) + \sup_{s>0}(W_A(s) - \lambda(a)\beta s) > f_N(a)\right)\\
&\leq \int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2} N\log N \mathbb{P}\left(\sup_{s>0}(W_i(s) - (1-\lambda(a))\beta s) > f_N(a) - y\log N\right)\\
&\quad\times \exp\left(-\frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy + \mathbb{P}\left(\sup_{s>0}(W_A(s) - \lambda(a)\beta s) > \lambda(a) f_N(a)\right) &(36)\\
&= \int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2} N\log N\\
&\quad\times \exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(f_N(a) - y\log N) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy\\
&\quad+ \mathbb{P}\left(\sup_{s>0}(W_A(s) - \lambda(a)\beta s) > \lambda(a) f_N(a)\right). &(37)
\end{aligned}$$

Because the function $\exp(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(f_N(a) - y\log N) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2})$ with $y \in [0, \lambda(a)(\frac{\sigma^2}{2\beta}+a)]$ is maximized when $y = \lambda(a)(\frac{\sigma^2}{2\beta}+a)$ and equals $N^{-\frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}-1}$, we get that

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{\log\left(\int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \frac{2\lambda(a)\beta}{\sigma_A^2}\log N \times N\exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(f_N(a) - y\log N) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy\right)}{\log N}\\
&= 1 + \limsup_{N\to\infty} \frac{\log\left(\int_0^{\lambda(a)(\frac{\sigma^2}{2\beta}+a)} \exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(f_N(a) - y\log N) - \frac{2\lambda(a)\beta y\log N}{\sigma_A^2}\right) dy\right)}{\log N}\\
&\leq -\frac{2a\beta + 2\sigma^2 - 2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}. &(38)
\end{aligned}$$

Now that we have found an upper bound for the integral in (37), we are left with the expression

$\mathbb{P}(\sup_{s>0}(W_A(s)-\lambda(a)\beta s)>\lambda(a)f_N(a))$ in (37). For this expression, it holds that

$$\mathbb{P}\left(\sup_{s>0}(W_A(s)-\lambda(a)\beta s)>\lambda(a)f_N(a)\right)=N^{-\frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}}.$$

Combining the upper bounds in (34) and (37) gives the logarithmic upper bound on the maximum queue length in (33). □

4. Useful lemmas

In the previous section, we gave a proof of the logarithmic asymptotics for the maximum queue length $\bar{Q}_N^{\beta}$. In order to be able to prove sharper results on the tail asymptotics, we need some auxiliary results; the goal of this section is to derive these. We begin by giving an overview of the results in this section.

First, observe that for a Brownian motion $\{W(t), t \geq 0\}$, we have that

$$\sup_{s>T}(W(s)-\beta s)\stackrel{d}{=}W(T)-\beta T+\sup_{s>0}(\hat{W}(s)-\beta s),$$

where $\{\hat{W}(t), t \geq 0\}$ is an independent copy of $\{W(t), t \geq 0\}$. From this, it follows that if we take the supremum of a Brownian motion starting at a positive time, this is in distribution the same as adding a normally distributed random variable to an exponentially distributed random variable. The tail asymptotics of this convolution equal the tail asymptotics of the normally distributed part, the exponentially distributed part, or a more complicated mixture of the two, depending on the starting time T, the standard deviation of $W(s)$ and the drift β. In Lemma 3, these three cases are studied in more detail.

Second, our main strategy to investigate the tail asymptotics involves the use of hitting times. Observe that we have a maximum of N mutually dependent random variables. Based on the results in Section 3, we are able to make an educated guess where the supremum is attained. Following the proof of Lemma 1, we see that for $T_N(a)$ given in (9),

$$\begin{aligned}&\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}(W_i(s)+W_A(s)-\beta s)>f_N(a)\right)\\&\approx\mathbb{P}\left(\max_{i\leq N}(W_i(T_N(a))+W_A(T_N(a))-\beta T_N(a))>f_N(a)\right),\end{aligned}$$

in the sense that the left-hand and the right-hand side have the same logarithmic asymptotics. So the hitting time, conditioned on being finite, is approximately equal to $T_N(a)$. Next, observe that for $0<a\leq a^{\star}$,

$$\mathbb{P}\left(\sup_{s>0}(W_A(s)-\lambda(a)\beta s)>\lambda(a)f_N(a)\right)=\exp\left(-\frac{2\lambda(a)\beta}{\sigma_A^2}\lambda(a)f_N(a)\right)=N^{-\gamma(a)},\quad(39)$$

and

$$\begin{aligned}&\mathbb{P}\left(\max_{i\leq N}\sup_{s>0}(W_i(s)-(1-\lambda(a))\beta s)>(1-\lambda(a))f_N(a)\right)\\&=1-\left(1-\exp\left(-\frac{2(1-\lambda(a))\beta}{\sigma^2}(1-\lambda(a))f_N(a)\right)\right)^N=\Theta(1).\end{aligned}\quad(40)$$

Since the expectation of the hitting time, conditioned on being finite, of a level x, equals this value x divided by the drift (see [4, Eq. (2.0.1) & (2.0.2)(1), p. 301]), it is easy to see that in both (39) and (40) the conditional expectation of the hitting time equals $T_N(a)$. Thus, this heuristically explains why the processes $\{W_A(t)-\lambda(a)\beta t, t \geq 0\}$ and $\{W_i(t)-(1-\lambda(a))\beta t, t \geq 0\}$ are important. In Definition 3, we define the hitting-time densities of these processes and in Lemma 4 we show that after proper scaling these densities converge to the densities of normally distributed random variables, corrected with a constant.

Finally, we need to analyze limits of the type

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} \mathbb{P}\left(\sup_{s>\tau^{(N)}} X_i(s) > y_N \middle| \tau^{(N)} = t \right) f_{\tau^{(N)}}(t)dt, \tag{41}$$

where $\tau^{(N)}$ is a hitting time and $f_{\tau^{(N)}}$ its density, with $f_{\tau^{(N)}}(t) = 0$ for $t < 0$. In Lemma 5, we show that under certain assumptions, we can interchange the integral and the limit, when the integrand is a product of two functions, as is the case in (41). The proof of this interchange is similar to the proof of the dominated convergence theorem.

Lemma 3 (*Convolution of Normal and Exponential Distributions*). *Let $X \stackrel{d}{=} \mathcal{N}(0,1)$ and $E \stackrel{d}{=} Exp(1)$ be independent random variables. Let $(\eta_N, N \geq 1)$, $(x_N, N \geq 1)$ be sequences with $\eta_N > 0$, $x_N \stackrel{N\to\infty}{\longrightarrow} \infty$, and $x_N/\eta_N \stackrel{N\to\infty}{\longrightarrow} \infty$. Furthermore, let $\mu > 0$ and $c \in \mathbb{R}$. Then*

1. *if $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \stackrel{N\to\infty}{\longrightarrow} c$,*

$$\mathbb{P}\left(\eta_N X + \frac{1}{\mu}E > x_N\right) \sim \frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} + \frac{1}{2}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}(1+erf(c)), \tag{42}$$

as $N \to \infty$, with

$$erf(c) = \frac{2}{\sqrt{\pi}}\int_0^c \exp(-t^2)dt.$$

2. *if $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \stackrel{N\to\infty}{\longrightarrow} \infty$,*

$$\mathbb{P}\left(\eta_N X + \frac{1}{\mu}E > x_N\right) \sim \frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} + e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}, \tag{43}$$

as $N \to \infty$,

3. *and if $\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N} \stackrel{N\to\infty}{\longrightarrow} -\infty$,*

$$\mathbb{P}\left(\eta_N X + \frac{1}{\mu}E > x_N\right) \sim \frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} - \frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\frac{\eta_N e^{-\frac{\left(x_N-\mu\eta_N^2\right)^2}{2\eta_N^2}}}{x_N-\mu\eta_N^2}, \tag{44}$$

as $N \to \infty$.

Proof. We have

$$\mathbb{P}\left(\eta_N X + \frac{1}{\mu}E > x_N\right) = \mathbb{P}\left(\eta_N X > x_N\right) + \int_{-\infty}^{x_N/\eta_N} \mathbb{P}\left(\frac{1}{\mu}E > x_N - \eta_N z\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz. \tag{45}$$

Using the fact that $1+\operatorname{erf}(-z) \sim \frac{e^{-z^2}}{\sqrt{\pi} z}$, as $z \to \infty$, the first term on the right-hand side of (45) satisfies

$$\mathbb{P}\left(\eta_N X > x_N\right) \sim \frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi} x_N},$$

as $N \to \infty$. Furthermore,

$$\int_{-\infty}^{x_N/\eta_N} \mathbb{P}\left(\frac{1}{\mu} E > x_N - \eta_N z\right) \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz = \frac{1}{2} e^{\frac{1}{2}\mu\left(\mu\eta_N^2 - 2x_N\right)} \left(\operatorname{erf}\left(\frac{x_N - \mu\eta_N^2}{\sqrt{2}\eta_N}\right) + 1\right).$$

The lemma follows by using that $\operatorname{erf}(z) \to 1$, as $z \to \infty$, and once more that $1+\operatorname{erf}(-z) \sim \frac{e^{-z^2}}{\sqrt{\pi} z}$, as $z \to \infty$; see [1, 7.1.13, 7.1.16 & 7.1.23]. □

For the remainder of this paper, we use τ to indicate stochastic hitting times.

Definition 3. For $a > 0$, $r \in \mathbb{R}$, and $i \in \{1, 2, \ldots, N\}$, we define the random variable $\tau_{i,N}^{a,-r}$ by

$$\tau_{i,N}^{a,-r} := \inf\{t \geq 0 : W_i(t) - (1-\lambda(a))\beta t > (1-\lambda(a)) f_N(a) - r\},$$

and the function $f_{\tau_{i,N}^{a,-r}}$ as its density, with $f_{\tau_{i,N}^{a,-r}}(t) = 0$ for $t < 0$.

Similarly, we define the random variable $\tilde{\tau}_{A,N}^{a,r}$ by

$$\tilde{\tau}_{A,N}^{a,r} := \inf\{t \geq 0 : W_A(t) - \lambda(a)\beta t > \lambda(a) f_N(a) + r\},$$

and the function $f_{\tilde{\tau}_{A,N}^{a,r}}$ as its density, with $f_{\tilde{\tau}_{A,N}^{a,r}}(t) = 0$ for $t < 0$.

Lemma 4 (*Convergence of Hitting-time Density*). *For the density function $f_{\tau_{i,N}^{a,-r}}$ given in Definition 3 and $T_N(a,k)$ given in Eq. (8), we have that*

$$\begin{aligned} N\sqrt{\log N} f_{\tau_{i,N}^{a,-r}}\left(T_N(a,k)\right) &\xrightarrow{N\to\infty} \frac{\beta^2}{\sqrt{\pi}\left(2a\beta + \sigma^2\right)} \\ &\times \exp\left(\frac{\beta\left(8a^2\beta^2 r - \beta^3 k^2 \sigma\sqrt{2a\beta + \sigma^2} + 8a\beta r\sigma^2 + 2r\sigma^4\right)}{\sigma\left(2a\beta + \sigma^2\right)^{5/2}}\right) \\ &= \frac{\beta^2}{\sqrt{\pi}\left(2a\beta + \sigma^2\right)} \exp\left(\frac{-\beta^4 k^2}{\left(2a\beta + \sigma^2\right)^2}\right) \\ &\times \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right). \end{aligned} \tag{46}$$

Proof. The density $f_{\tau_{i,N}^{a,-r}}(t)$ satisfies

$$f_{\tau_{i,N}^{a,-r}}(t) = \frac{(1-\lambda(a)) f_N(a) - r}{\sqrt{2\pi}\sigma t^{3/2}} \exp\left(-\frac{((1-\lambda(a)) f_N(a) - r + (1-\lambda(a))\beta t)^2}{2\sigma^2 t}\right), \tag{47}$$

for $t > 0$, and 0 otherwise; see [4, Eq. (2.0.2), p. 301]. Due to the fact that $T_N(a,k) = f_N(a)/\beta + k\sqrt{\log N}$, for all $k \in \mathbb{R}$, there exists N_k, such that for $N > N_k$, $T_N(a,k) > 0$. Following the notation given in Definition 2, we have that the prefactor of the density of the hitting time in the point $T_N(a,k)$ equals

$$\frac{(1-\lambda(a))f_N(a)-r}{\sqrt{2\pi}\sigma T_N(a,k)^{3/2}} = \frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)\log N - r}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N + k\sqrt{\log N})^{3/2}}$$
$$\sim \frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)\log N}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N)^{3/2}},$$

as $N\to\infty$. When we simplify this last term further, we get

$$\frac{\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)}{\sqrt{2\pi}\sigma((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}))^{3/2}\sqrt{\log N}} = \frac{\frac{1}{\sqrt{2a\beta+\sigma^2}}}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}}$$
$$= \frac{1}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{2a\beta+\sigma^2}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}}.$$

Because we can write $\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}} = \frac{1}{\sqrt{2}\beta}\sqrt{2a\beta+\sigma^2}$, we get

$$\frac{1}{\sqrt{2\pi}\frac{1}{\beta}\sqrt{2a\beta+\sigma^2}\sqrt{\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}}\sqrt{\log N}} = \frac{\beta^2}{\sqrt{\pi}(2a\beta+\sigma^2)\sqrt{\log N}}.$$

So, we can conclude that $\sqrt{\log N}$ times the first term of the density $f_{\tau^{a,-r}_{i,N}}(t)$ in (47) converges to $\frac{\beta^2}{\sqrt{\pi}(2a\beta+\sigma^2)}$, as $N\to\infty$, which is the prefactor of the limit in (46). So, in order to prove the limit in (46), we are left with proving that

$$N\exp\left(-\frac{((1-\lambda(a))f_N(a)-r+(1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)}\right)$$
$$\xrightarrow{N\to\infty} \exp\left(\frac{\beta\left(8a^2\beta^2 r - \beta^3k^2\sigma\sqrt{2a\beta+\sigma^2}+8a\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}}\right). \tag{48}$$

The numerator of the exponent on the left-hand side of (48) equals

$$((1-\lambda(a))f_N(a)-r+(1-\lambda(a))\beta T_N(a,k))^2.$$

Because of the form of $f_N(a)$ and $T_N(a,k)$ as given in Definition 2, we can write

$$((1-\lambda(a))f_N(a)-r+(1-\lambda(a))\beta T_N(a,k))^2$$
$$= c_1(\log N)^2 + c_2(\log N)^{3/2} + c_3\log N + c_4\sqrt{\log N} + r^2, \tag{49}$$

with c_1, c_2, c_3, c_4 constant in N. In order to determine the value of c_1 we should gather all the terms in

$$(1-\lambda(a))f_N(a)-r+(1-\lambda(a))\beta T_N(a,k)$$

that scale as $\log N$. We have

$$(1-\lambda(a))f_N(a)-r+(1-\lambda(a))\beta T_N(a,k)$$
$$= \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N - r + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta\left(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}\right)\log N$$

$$
\begin{aligned}
&+ \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} \\
&= \frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + o(\log N).
\end{aligned}
$$

Therefore,

$$
c_1 = \left(\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\right)^2 = \frac{4\sigma^2}{2a\beta+\sigma^2}\left(\frac{\sigma^2}{2\beta}+a\right)^2 = \frac{2\sigma^2}{\beta}\left(\frac{\sigma^2}{2\beta}+a\right).
$$

Now, to determine the value of c_2 in (49), we have

$$
\begin{aligned}
&(1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k) \\
&= \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N - r + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta\left(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta}\right)\log N \\
&\quad + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} \\
&= \frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} - r.
\end{aligned}
$$

Therefore, c_2 equals

$$
c_2 = 2\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k = 4\frac{\sigma^2}{2a\beta+\sigma^2}\left(\frac{\sigma^2}{2\beta}+a\right)\beta k = 2\sigma^2 k.
$$

Observe that

$$
c_1(\log N)^2 + c_2(\log N)^{3/2} = 2\sigma^2 T_N(a,k)\log N.
$$

Thus, the exponent on the left-hand side of (48) can be rewritten as

$$
-\frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)} = -\log N + O(1),
$$

and we can conclude that

$$
\begin{aligned}
&N\exp\left(-\frac{((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)}\right) \\
&= N\exp(-\log N + O(1)) = O(1).
\end{aligned}
$$

The only term in (49) that is still of importance, is the term c_3. We have

$$
\begin{aligned}
&-((1-\lambda(a))f_N(a) - r + (1-\lambda(a))\beta T_N(a,k))^2 \\
&= -\left(\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right)\log N + \frac{\sigma}{\sqrt{2a\beta+\sigma^2}}\beta k\sqrt{\log N} - r\right)^2.
\end{aligned}
$$

The terms that scale as $\log N$ are as follows:

$$
c_3\log N = -\left(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}\left(\frac{\sigma^2}{2\beta}+a\right) + \frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2\right)\log N.
$$

Thus,

$$\frac{-\left(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2\right)\log N}{2\sigma^2 T_N(a,k)}$$
$$=\frac{-\left(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2\right)\log N}{2\sigma^2((\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})\log N+k\sqrt{\log N})}.$$

This expression converges to

$$\frac{-\left(-2r\frac{2\sigma}{\sqrt{2a\beta+\sigma^2}}(\frac{\sigma^2}{2\beta}+a)+\frac{\sigma^2}{2a\beta+\sigma^2}\beta^2k^2\right)}{2\sigma^2(\frac{\sigma^2}{2\beta^2}+\frac{a}{\beta})}$$
$$=\frac{\beta\left(8a^2\beta^2r-\beta^3k^2\sigma\sqrt{2a\beta+\sigma^2}+8a\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}},$$

as $N\to\infty$, which is exactly the exponent in the limit of (48). Putting everything together, the limit in (46) follows. □

Corollary 1. *For the density function $f_{\tau_{i,N}^{a,-r}}$ given in Definition 3 and $T_N(a,k)$ given in Eq. (8) we have that*

$$\lim_{N\to\infty}\int_{-\infty}^{\infty}N\sqrt{\log N}f_{\tau_{i,N}^{a,-r}}\left(T_N(a,k)\right)dk=\int_{-\infty}^{\infty}\lim_{N\to\infty}N\sqrt{\log N}f_{\tau_{i,N}^{a,-r}}\left(T_N(a,k)\right)dk. \quad (50)$$

Proof. Observe that for N large enough such that $(1-\lambda(a))f_N(a)-r>0$,

$$\int_{-\infty}^{\infty}N\sqrt{\log N}f_{\tau_{i,N}^{a,-r}}\left(T_N(a,k)\right)dk$$
$$=N\mathbb{P}\left(\sup_{s>0}(W_i(s)-(1-\lambda(a))\beta s)>(1-\lambda(a))f_N(a)-r\right)$$
$$=\exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right),$$

due to the fact that $\sup_{s>0}(W_i(s)-(1-\lambda(a))\beta s)$ is exponentially distributed with parameter $2(1-\lambda(a))\beta/\sigma^2$. Additionally,

$$\int_{-\infty}^{\infty}\frac{\beta^2\exp\left(\frac{\beta\left(8a^2\beta^2r-\beta^3k^2\sigma\sqrt{2a\beta+\sigma^2}+8a\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}dk$$
$$=\int_{-\infty}^{\infty}\frac{\beta^2\exp\left(-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}\exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right)dk.$$

The first term in this integral is the density of a normally distributed random variable. Therefore, we get that

$$\int_{-\infty}^{\infty} \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)} \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right) dk = \exp\left(\frac{2(1-\lambda(a))\beta r}{\sigma^2}\right). \quad \square$$

Lemma 5 (*Convergence of Integrals of Sequences of Functions*)**.** *Assume we have sequences of positive integrable functions $v_N(x)$ and $w_N(x)$ that satisfy the following:*

- $v_N(x) \overset{N\to\infty}{\longrightarrow} v(x)$,
- $\int_{-\infty}^{\infty} v_N(x)dx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)dx$,
- $w_N(x) \overset{N\to\infty}{\longrightarrow} w(x)$,
- *There exists a constant $c > 0$ such that $w_N(x) < c$ for all x and N.*

Then

$$\int_{-\infty}^{\infty} v_N(x)w_N(x)dx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)w(x)dx. \tag{51}$$

Proof. First, by using Fatou's lemma, we obtain that

$$\liminf_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)w_N(x)dx \geq \int_{-\infty}^{\infty} v(x)w(x)dx.$$

Furthermore, observe that $v_N(x)c - v_N(x)w_N(x) > 0$ for all x and N. Now, from Fatou's lemma, it follows that

$$\liminf_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)c - v_N(x)w_N(x)dx \geq \int_{-\infty}^{\infty} v(x)c - v(x)w(x)dx.$$

Because $\int_{-\infty}^{\infty} v_N(x)cdx \overset{N\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} v(x)cdx$, we get that

$$\limsup_{N\to\infty} \int_{-\infty}^{\infty} v_N(x)w_N(x)dx \leq \int_{-\infty}^{\infty} v(x)w(x)dx.$$

The lemma follows. $\square$

In Definition 4, we give shorthand notation of some probability measures that we use later on.

Definition 4.

$$P_{i,j}^{(N)} := \mathbb{P}\left(\min(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0})\mathbb{1}(\tau_{i,N}^{a^\star,0} < \infty), Q_{j,A}^{\beta}(\tau_{j,N}^{a^\star,0})\mathbb{1}(\tau_{j,N}^{a^\star,0} < \infty)) > f_N(a)\right), \tag{52}$$

$$Q_{i,j}^{(N)}(k,l) := \mathbb{P}(\min(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0}), Q_{j,A}^{\beta}(\tau_{j,N}^{a^\star,0})) > f_N(a)|\tau_{i,N}^{a^\star,0} = T_N(a^\star,k), \tau_{j,N}^{a^\star,0} = T_N(a^\star,l)), \tag{53}$$

$$\mathbb{P}^{(k<l)}(A) := \mathbb{P}(A|\tau_{i,N}^{a^\star,0} = T_N(a^\star,k) < \tau_{j,N}^{a^\star,0} = T_N(a^\star,l)), \tag{54}$$

and

$$\mathbb{P}_{i,a,-r,k}^{(N)}(A) := \mathbb{P}(A|\tau_{i,N}^{a,-r} = T_N(a,k)). \tag{55}$$

5. Proofs of the sharper asymptotics

In this section, we prove sharper asymptotics of the tail behavior of $\mathbb{P}(\bar{Q}_N^\beta > f_N(a))$. Recall the definition of $\tau_{i,N}^{a,-r}$ and $\tilde{\tau}_{A,N}^{a,r}$ given in Definition 3, and observe that

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) = \mathbb{P}(\max_{i\le N} Q_{i,A}^\beta(\tau_{i,N}^{a,-r} \wedge \tilde{\tau}_{A,N}^{a,r})\mathbb{1}(\tau_{i,N}^{a,-r} \wedge \tilde{\tau}_{A,N}^{a,r} < \infty) > f_N(a)). \tag{56}$$

This equation is valid, because for $0 < t < \tau_{i,N}^{a,-r} \wedge \tilde{\tau}_{A,N}^{a,r}$, we see that $W_i(t) - (1-\lambda(a))\beta t < (1-\lambda(a))f_N(a) - r$ and $W_A(t) - \lambda(a)\beta t < \lambda(a)f_N(a) + r$. Thus, $W_i(t) + W_A(t) - \beta t < f_N(a)$. Now, using (56), we obtain lower and upper bounds of the form

$$\begin{aligned}
&\max\Big(\mathbb{P}\big(\max_{i\le N} Q_{i,A}^\beta(\tau_{i,N}^{a,-r})\mathbb{1}(\tau_{i,N}^{a,-r} < \infty) > f_N(a)\big), \mathbb{P}\big(\bar{Q}_N^\beta(\tilde{\tau}_{A,N}^{a,r})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r} < \infty) > f_N(a)\big)\Big) \\
&\le \mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \\
&\le \mathbb{P}\big(\max_{i\le N} Q_{i,A}^\beta(\tau_{i,N}^{a,-r})\mathbb{1}(\tau_{i,N}^{a,-r} < \infty) > f_N(a)\big) + \mathbb{P}\big(\bar{Q}_N^\beta(\tilde{\tau}_{A,N}^{a,r})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r} < \infty) > f_N(a)\big),
\end{aligned} \tag{57}$$

which we can exploit. Other important inequalities that we use are the union bound and Bonferroni's inequality. In the case of identically distributed random variables X_i, these bounds simplify to

$$N\mathbb{P}(X_i > x) - \binom{N}{2}\mathbb{P}(\min(X_i, X_j) > x) \le \mathbb{P}(\max_{i\le N} X_i > x) \le N\mathbb{P}(X_i > x),$$

which is the case for our problem. Dębicki et al. [6] have derived the tail asymptotics of $\min(Q_{i,A}^\beta, Q_{j,A}^\beta)$. In Lemma 7, we show how we use [6, Thm. 2.3] on the tails of $\min(Q_{i,A}^\beta, Q_{j,A}^\beta)$ together with Bonferroni's inequality such that these are applicable in our proof of the case $a > a^\star$.

Now that we can write upper and lower bounds in which hitting times play a role, we condition on the hitting times and get sequences of the form as given in (41).

By using Lemma 5, we obtain that

$$\begin{aligned}
&\lim_{N\to\infty} \int_{-\infty}^{\infty} \mathbb{P}\left(\sup_{s>\tau^{(N)}} X_i(s) > y_N \middle| \tau^{(N)} = t\right) f_{\tau^{(N)}}(t)dt \\
&= \int_{-\infty}^{\infty} \lim_{N\to\infty} \mathbb{P}\left(\sup_{s>\tau^{(N)}} X_i(s) > y_N \middle| \tau^{(N)} = t\right) f_{\tau^{(N)}}(t)dt.
\end{aligned}$$

To obtain limits of the form as given in (41), we use Lemmas 3 and 4.

5.1. *The case $a > a^\star$*

In this section, we prove Theorem 2 on exact asymptotics of the maximum queue length when $a > a^\star$. As is stated in (16), $\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \sim N^{-\gamma(a)}$, as $N \to \infty$, when $a > a^\star$. Since the union bound in (14) gives us that $N^{\gamma(a)}\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \le 1$, we only need to show that

$$\liminf_{N\to\infty} N^{\gamma(a)}\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \ge 1.$$

In order to prove the lim inf, we first observe that $\bar{Q}_N^\beta > \max_{i\le N} Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N})\mathbb{1}(\tau^{a^\star,0}_{i,N} < \infty)$, and we know by using Bonferroni's inequality that

$$\begin{aligned}
&\mathbb{P}\,(\max_{i\le N} Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N})\mathbb{1}(\tau^{a^\star,0}_{i,N} < \infty) > f_N(a))\\
&\quad\ge N\mathbb{P}\left(Q^\beta_{1,A}(\tau^{a^\star,0}_{1,N})\mathbb{1}(\tau^{a^\star,0}_{1,N} < \infty) > f_N(a)\right)\\
&\quad\quad - \binom{N}{2}\mathbb{P}\left(\min(Q^\beta_{1,A}(\tau^{a^\star,0}_{1,N})\mathbb{1}(\tau^{a^\star,0}_{1,N} < \infty), Q^\beta_{2,A}(\tau^{a^\star,0}_{2,N})\mathbb{1}(\tau^{a^\star,0}_{2,N} < \infty)) > f_N(a)\right),
\end{aligned} \tag{58}$$

where $\tau^{a^\star,0}_{i,N}$ and $\tau^{a^\star,0}_{j,N}$ are hitting times defined in Lemma 4. In Lemma 7, we show that the first term is leading, and the second order term is of smaller order. In order to prove this, we first give a convenient upper bound for

$$\mathbb{P}^{(k<l)}\left(\min(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})) > f_N(a)\right)$$

in Lemma 6, with $\mathbb{P}^{(k<l)}\,(A)$ given in Eq. (54) in Definition 4.

For the remainder of this paper, let $\{\hat{W}(t), t \ge 0\}$ be an independent copy of the Brownian motion $\{W(t), t \ge 0\}$, and $\hat{Q}^\beta_{i,A}(s,t)$ an independent copy of $Q^\beta_{i,A}(s,t)$.

Lemma 6. *Let $a > a^\star$ and $\mathbb{P}^{(k<l)}(A)$ be given in Eq. (54). Furthermore, $\tau^{a^\star,0}_{i,N}$ is given in Definition 3 and $\hat{Q}^\beta_{i,A}$ is an independent copy of $Q^\beta_{i,A}$. Then for all $\delta > 0$ there exists an $N_\delta > 0$ such that for all $N \ge N_\delta$*

$$\begin{aligned}
&\mathbb{P}^{(k<l)}\left(\min(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})) > f_N(a)\right)\\
&\quad\le 4\mathbb{P}^{(k<l)}\Bigg((1+\delta)W_A(\tau^{a^\star,0}_{i,N}) + \min(\hat{Q}^\beta_{i,A}, \hat{Q}^\beta_{j,A})\\
&\quad\quad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg).
\end{aligned}$$

Proof. First, we have that

$$\begin{aligned}
&\mathbb{P}^{(k<l)}\left(\min(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})) > f_N(a)\right)\\
&\quad\le \mathbb{P}^{(k<l)}\left(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}, \tau^{a^\star,0}_{j,N}) > f_N(a)\right)\\
&\quad\quad + \mathbb{P}^{(k<l)}\left(\min(Q^\beta_{i,A}(\tau^{a^\star,0}_{j,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})) > f_N(a)\right),
\end{aligned} \tag{59}$$

because

$$\min(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})) < \max(Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}, \tau^{a^\star,0}_{j,N}), \min(Q^\beta_{i,A}(\tau^{a^\star,0}_{j,N}), Q^\beta_{j,A}(\tau^{a^\star,0}_{j,N})))$$

when $\tau^{a^\star,0}_{i,N} < \tau^{a^\star,0}_{j,N} < \infty$. Now, recall from Definition 3 that

$$\begin{aligned}
&Q^\beta_{i,A}(\tau^{a^\star,0}_{i,N}, \tau^{a^\star,0}_{j,N})\\
&\quad= \sup_{\tau^{a^\star,0}_{i,N} < s < \tau^{a^\star,0}_{j,N}} (W_i(s) + W_A(s) - \beta s)\\
&\quad\overset{d}{=} (1-\lambda(a^\star))f_N(a^\star) + W_A(\tau^{a^\star,0}_{i,N}) - \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N} + \hat{Q}^\beta_{i,A}(0, \tau^{a^\star,0}_{j,N} - \tau^{a^\star,0}_{i,N}).
\end{aligned}$$

Thus, for the first term on the right-hand side of (59) we have

$$\begin{aligned}
&\mathbb{P}^{(k<l)}\left(Q^{\beta}_{i,A}(\tau^{a^\star,0}_{i,N},\tau^{a^\star,0}_{j,N}) > f_N(a)\right)\\
&= \mathbb{P}^{(k<l)}\Bigg(W_A(\tau^{a^\star,0}_{i,N}) + \hat{Q}^{\beta}_{i,A}(0,\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N})\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg)\\
&\le \mathbb{P}^{(k<l)}\Bigg(W_A(\tau^{a^\star,0}_{i,N}) + \left|\hat{W}_i(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) + \hat{W}_A(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N})\right|\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg). \qquad (60)
\end{aligned}$$

For any x and y, it holds that $x+|y| = \max(x+y, x-y)$. Therefore, by the union bound, we can bound the probability in (60) as

$$\begin{aligned}
&\mathbb{P}^{(k<l)}\Bigg(W_A(\tau^{a^\star,0}_{i,N}) + \left|\hat{W}_i(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) + \hat{W}_A(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N})\right|\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg) \qquad (61)\\
&\le 2\mathbb{P}^{(k<l)}\Bigg(W_A(\tau^{a^\star,0}_{i,N}) + \hat{W}_i(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) + \hat{W}_A(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N})\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg) \qquad (62)\\
&\le 2\mathbb{P}^{(k<l)}\left((1+\delta)W_A(\tau^{a^\star,0}_{i,N}) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\right) \qquad (63)\\
&\le 2\mathbb{P}^{(k<l)}\Bigg((1+\delta)W_A(\tau^{a^\star,0}_{i,N}) + \min(\hat{Q}^{\beta}_{i,A},\hat{Q}^{\beta}_{j,A})\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{i,N}\Bigg), \qquad (64)
\end{aligned}$$

for $\delta > 0$ and $N > N_\delta$. The upper bound in (63) holds for $N > N_\delta$ with N_δ large enough, since under the measure $\mathbb{P}^{(k<l)}$ given in (54), $\tau^{a^\star,0}_{i,N} = f_N(a^\star)/\beta + k\sqrt{\log N} \sim \left(\frac{\sigma^2}{2\beta^2} + \frac{a^\star}{\beta}\right)\log N$ as $N\to\infty$, and $\tau^{a^\star,0}_{j,N} - \tau^{a^\star,0}_{i,N} = (l-k)\sqrt{\log N} = O(\sqrt{\log N})$. The upper bound in (64) holds because we add a positive random variable. For the second term on the right-hand side of (59), first observe that $\mathbb{P}(\min(X,Y) > z) = \mathbb{P}(X > z, Y > z)$. Second, under the assumption that $\tau^{a^\star,0}_{i,N} < \tau^{a^\star,0}_{j,N} < \infty$, we can write

$$\begin{aligned}
Q^{\beta}_{i,A}(\tau^{a^\star,0}_{j,N}) \overset{d}{=}\; &(1-\lambda(a^\star))f_N(a^\star) + \hat{W}_i(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N})\\
&- (1-\lambda(a^\star))\beta(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) + W_A(\tau^{a^\star,0}_{j,N}) - \lambda(a^\star)\beta\tau^{a^\star,0}_{j,N} + \hat{Q}^{\beta}_{i,A}.
\end{aligned}$$

Thus, by applying similar techniques as for the analysis of the first term in (59), we obtain that

$$\begin{aligned}
&\mathbb{P}^{(k<l)}\left(\min(Q^{\beta}_{i,A}(\tau^{a^\star,0}_{j,N}), Q^{\beta}_{j,A}(\tau^{a^\star,0}_{j,N})) > f_N(a)\right)\\
&\quad= \mathbb{P}^{(k<l)}\Bigg(W_A(\tau^{a^\star,0}_{j,N}) + \hat{W}_i(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) - (1-\lambda(a^\star))\beta(\tau^{a^\star,0}_{j,N}-\tau^{a^\star,0}_{i,N}) + \hat{Q}^{\beta}_{i,A}\\
&\qquad > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{j,N},\\
&\qquad W_A(\tau^{a^\star,0}_{j,N}) + \hat{Q}^{\beta}_{j,A} > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta\tau^{a^\star,0}_{j,N}\Bigg).
\end{aligned}$$

This joint probability satisfies the following bound:

$$
\begin{aligned}
&\mathbb{P}^{(k<l)}\Big(W_A(\tau_{j,N}^{a^\star,0}) + \hat{W}_i(\tau_{j,N}^{a^\star,0} - \tau_{i,N}^{a^\star,0}) - (1-\lambda(a^\star))\beta(\tau_{j,N}^{a^\star,0} - \tau_{i,N}^{a^\star,0}) + \hat{Q}_{i,A}^{\beta} \\
&\quad > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}, \\
&\quad W_A(\tau_{j,N}^{a^\star,0}) + \hat{Q}_{j,A}^{\beta} > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}\Big) \\
&\le \mathbb{P}^{(k<l)}\Big(W_A(\tau_{j,N}^{a^\star,0}) + \hat{W}_i(\tau_{j,N}^{a^\star,0} - \tau_{i,N}^{a^\star,0}) + \hat{Q}_{i,A}^{\beta} \\
&\quad > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}, \\
&\quad W_A(\tau_{j,N}^{a^\star,0}) + \hat{Q}_{j,A}^{\beta} > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0} \Big).
\end{aligned}
$$

We can bound this further and get

$$
\begin{aligned}
&\mathbb{P}^{(k<l)}\Big(W_A(\tau_{j,N}^{a^\star,0}) + \hat{W}_i(\tau_{j,N}^{a^\star,0} - \tau_{i,N}^{a^\star,0}) + \hat{Q}_{i,A}^{\beta} > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) \\
&+ \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}, \\
&\quad W_A(\tau_{j,N}^{a^\star,0}) + \hat{Q}_{j,A}^{\beta} > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}\Big) \\
&\le \mathbb{P}^{(k<l)}\Big(W_A(\tau_{j,N}^{a^\star,0}) + \max(\hat{W}_i(\tau_{j,N}^{a^\star,0} - \tau_{i,N}^{a^\star,0}), 0) + \min(\hat{Q}_{i,A}^{\beta}, \hat{Q}_{j,A}^{\beta}) \\
&\quad > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{j,N}^{a^\star,0}\Big) \\
&\le 2\mathbb{P}^{(k<l)}\Big((1+\delta) W_A(\tau_{i,N}^{a^\star,0}) + \min(\hat{Q}_{i,A}^{\beta}, \hat{Q}_{j,A}^{\beta}) \\
&\quad > f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta\tau_{i,N}^{a^\star,0} \Big),
\end{aligned}
$$

for $\delta > 0$ and $N > N_\delta$. Combining this bound with the bound in (64) completes the proof of the lemma. □

Lemma 7. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a > a^\star$, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a)}\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a)) \ge 1.$$

The general idea of the proof of Lemma 7 is to make rigorous that the lower bound on the maximum queue length $\bar{Q}_N^{\beta}$ given in (58) is approximately the same as $N\mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0})\mathbb{1}(\tau_{i,N}^{a^\star,0} < \infty) > f_N(a))$ when N is large. Thus the last term in (58) is asymptotically negligible. We use the result from Lemma 6 to establish this. Observe now that, following Definition 3,

$$
\begin{aligned}
Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0}) &\stackrel{d}{=} W_i(\tau_{i,N}^{a^\star,0}) + W_A(\tau_{i,N}^{a^\star,0}) - \beta\tau_{i,N}^{a^\star,0} + \hat{Q}_{i,A}^{\beta} \\
&= (1-\lambda(a^\star)) f_N(a^\star) + W_A(\tau_{i,N}^{a^\star,0}) - \lambda(a^\star)\beta\tau_{i,N}^{a^\star,0} + \hat{Q}_{i,A}^{\beta}.
\end{aligned}
$$

Furthermore, observe that due to Eq. (28), $\mathbb{P}(\tau_{i,N}^{a^\star,0} < \infty) = 1/N$. From this, it follows that

$$N\mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0})\mathbb{1}(\tau_{i,N}^{a^\star,0} < \infty) > f_N(a)) = \mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0}) > f_N(a)|\tau_{i,N}^{a^\star,0} < \infty).$$

Therefore, in order to prove a sharp lower bound on the tail asymptotics of the maximum queue length, we prove by using Fatou's lemma that

$$\begin{aligned}\liminf_{N\to\infty} N^{\gamma(a)}\mathbb{P}(&W_A(\tau_{i,N}^{a^\star,0}) - \lambda(a^\star)\beta\tau_{i,N}^{a^\star,0} + \hat{Q}_{i,A}^{\beta}\\ &> f_N(a) - (1-\lambda(a^\star))f_N(a^\star)|\tau_{i,N}^{a^\star,0} < \infty) \geq 1.\end{aligned}$$

In order to prove this, we show that $\hat{Q}_{i,A}^{\beta}$ is most likely to hit a level $g_N(a,x,k)$ (to be specified later), and $W_A(\tau_{i,N}^{a^\star,0}) - \lambda(a^\star)\beta\tau_{i,N}^{a^\star,0}$ is most likely to hit the level $f_N(a) - (1-\lambda(a^\star))f_N(a^\star) - g_N(a,x,k)$.

We now turn to a formal proof of Lemma 7.

Proof. Following Eq. (52) in Definition 4, we can simplify the inequality in (58) to

$$\mathbb{P}(\max_{i\leq N} Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0})\mathbb{1}(\tau_{i,N}^{a^\star,0} < \infty) > f_N(a)) \geq NP_{i,i}^{(N)} - \binom{N}{2}P_{i,j}^{(N)}. \tag{65}$$

Now, before we analyze (65) in more detail, observe that we can express $\mathbb{P}(\tau_{i,N}^{a^\star,0} < \infty, \tau_{j,N}^{a^\star,0} < \infty)$ as

$$\begin{aligned}\mathbb{P}(\tau_{i,N}^{a^\star,0} < \infty, \tau_{j,N}^{a^\star,0} < \infty) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)\log N dkdl\\ &= \frac{1}{N^2}.\end{aligned}$$

Then, by using Eq. (53) in Definition 4, we get that

$$\begin{aligned}NP_{i,i}^{(N)} =& N\int_{-\infty}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big)\sqrt{\log N}\, Q_{i,i}^{(N)}(k,k)dk\\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N\, Q_{i,i}^{(N)}(k,k)dkdl.\end{aligned}$$

Also, observe that $\binom{N}{2} < N^2/2$, and that

$$\frac{N^2}{2}P_{i,j}^{(N)} = \frac{N^2}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)\log N\, Q_{i,j}^{(N)}(k,l)dkdl.$$

In conclusion, we can write the inequality in (65) as

$$\mathbb{P}(\max_{i\leq N} Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,0})\mathbb{1}(\tau_{i,N}^{a^\star,0} < \infty) > f_N(a)) \tag{66}$$

$$\geq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N\left(Q_{i,i}^{(N)}(k,k) - \frac{Q_{i,j}^{(N)}(k,l)}{2}\right)dkdl \tag{67}$$

$$\begin{aligned}&= \int_{-\infty}^{\infty}\int_{-\infty}^{l} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N\left(Q_{i,i}^{(N)}(k,k) - \frac{Q_{i,j}^{(N)}(k,l)}{2}\right)dkdl\\ &\quad + \int_{-\infty}^{\infty}\int_{l}^{\infty} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big)N^2\log N\left(Q_{i,i}^{(N)}(k,k) - \frac{Q_{i,j}^{(N)}(k,l)}{2}\right)dkdl.\end{aligned}$$

Since we want to prove a sharp lower bound on the tail asymptotics of the maximum queue length $\bar{Q}_N^{\beta}$ we can use the expression in (67). We want to prove the convergence of a lower bound of this integral by using Fatou's lemma. Therefore, we focus on the integrand first and prove convergence for the integrand as $N \to \infty$. Assume that $k \leq l$, and observe that $Q_{i,i}^{(N)}(k,k) - Q_{i,j}^{(N)}(k,l)/2 > 0$. Thus,

$$Q_{i,i}^{(N)}(k,k) - \frac{1}{2} Q_{i,j}^{(N)}(k,l) = \left(Q_{i,i}^{(N)}(k,k) - \frac{Q_{i,j}^{(N)}(k,l)}{2} \right)^{+}.$$

The density of $W_A\big(T_N(a^\star,k)\big)$ equals

$$\frac{\exp\left(-x^2/(2\sigma_A^2 T_N(a^\star,k))\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}.$$

We write $a = a^\star + \epsilon$, with $\epsilon > 0$. Let

$$\begin{aligned} g_N(a,x,k) = {} & f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k) \\ & - \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N - x\sqrt{\log N}. \end{aligned}$$

Observe that

$$\begin{aligned} g_N(a,x,k) + \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N} = {} & f_N(a) - (1-\lambda(a^\star)) f_N(a^\star) \\ & + \lambda(a^\star)\beta T_N(a^\star,k). \end{aligned}$$

Furthermore,

$$\begin{aligned} N^{\gamma(a)} Q_{i,i}^{(N)}(k,k) = {} & N^{\gamma(a)} \mathbb{P}\left(W_A\big(T_N(a^\star,k)\big) + \hat{Q}_{i,A}^{\beta} > g_N(a,x,k) + \frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N} \right) \\ = {} & \int_{-\infty}^{\infty} N^{\gamma(a)} \mathbb{P}(\hat{Q}_{i,A}^{\beta} > g_N(a,x,k)) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)} \right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} dx. \end{aligned}$$

We can simplify this expression further and get with a similar analysis as given in the proof of Lemma 4, that

$$\begin{aligned} & N^{\gamma(a)} \mathbb{P}(\hat{Q}_{i,A}^{\beta} > g_N(a,x,k)) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)} \right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} \\ & = N^{\gamma(a)} \exp\left(-\frac{2\beta}{\sigma^2+\sigma_A^2} g_N(a,x,k) \right) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)} \right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} \\ & \xrightarrow{N\to\infty} \frac{\beta\sigma}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right) - 2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4} \right). \end{aligned}$$

Furthermore, following Lemma 4, we have that

$$f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big) N^2 \log N \xrightarrow{N\to\infty} \frac{\beta^2 \exp\left(-\frac{\beta^4 k^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \times \frac{\beta^2 \exp\left(-\frac{\beta^4 l^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)}.$$

Also, following Lemma 6, we have that

$$Q_{i,j}^{(N)}(k,l) \le 4\mathbb{P}\Bigg((1+\delta)W_A\big(T_N(a^\star,k)\big) + \min(\hat{Q}_{i,A}^{\beta}, \hat{Q}_{j,A}^{\beta}) > f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k) \Bigg),$$

for all $\delta > 0$ for $N > N_\delta$. Let $0 < \delta < \frac{\beta\sigma^4\epsilon}{2\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}$ and let

$$h_N(a,x,k) = f_N(a) - (1-\lambda(a^\star))f_N(a^\star) + \lambda(a^\star)\beta T_N(a^\star,k) - (1+\delta)\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right).$$

From Dębicki et al. [6, Thm. 2.3], we know that

$$\mathbb{P}\left(\min(\hat{Q}_{i,A}^{\beta}, \hat{Q}_{j,A}^{\beta}) > x\right)\exp\left(\frac{2\beta}{\sigma^2/2+\sigma_A^2}x\right) \xrightarrow{x\to\infty} 0. \tag{68}$$

We have that

$$N^{\gamma(a)}\exp\left(-\frac{2\beta}{\sigma^2/2+\sigma_A^2}h_N(a,x,k)\right) \frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N + x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}} \xrightarrow{N\to\infty} 0.$$

Thus, when $k \le l$, then

$$\liminf_{N\to\infty} N^{\gamma(a)} f_{\tau_{i,N}^{a^\star,0}}\big(T_N(a^\star,k)\big) f_{\tau_{j,N}^{a^\star,0}}\big(T_N(a^\star,l)\big) N^2\log N\left(Q_{i,i}^{(N)}(k,k) - \frac{Q_{i,j}^{(N)}(k,l)}{2}\right)^+$$
$$\ge \frac{\beta^2\exp\left(-\frac{\beta^4k^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta^2\exp\left(-\frac{\beta^4l^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta\sigma\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}.$$

The case $k > l$ can be treated analogously. Finally, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\beta^2\exp\left(-\frac{\beta^4k^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)} \frac{\beta^2\exp\left(-\frac{\beta^4l^2}{(2a^\star\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)}$$

$$\times \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} dxdkdl = 1,$$

because this is an integral over the whole domain of a product of three densities of normally distributed random variables. By applying Fatou's lemma, Lemma 7 follows. □

Corollary 2. *Let* $(y_N, N \geq 1)$ *be a sequence such that* $\liminf_{N\to\infty} y_N/\log N = \infty$*, then the tail probability of the steady-state maximum queue length satisfies*

$$\mathbb{P}(\bar{Q}_N^\beta > y_N) \sim N\mathbb{P}(Q_{i,A}^\beta > y_N),$$

as $N \to \infty$.

Proof. By using the union bound, we have that $\mathbb{P}(\bar{Q}_N^\beta > y_N) \leq N\mathbb{P}(Q_{i,A}^\beta > y_N)$. Furthermore, by using Bonferroni's inequality, we obtain that $\mathbb{P}(\bar{Q}_N^\beta > y_N) \geq N\mathbb{P}(Q_{i,A}^\beta > y_N) - N^2/2\mathbb{P}(Q_{i,A}^\beta > y_N, Q_{j,A}^\beta > y_N)$. Now, using the limit in (68), we see that

$$\limsup_{N\to\infty} \frac{N^2/2\mathbb{P}(Q_{i,A}^\beta > y_N, Q_{j,A}^\beta > y_N)}{N\mathbb{P}(Q_{i,A}^\beta > y_N)} \leq \limsup_{N\to\infty} \frac{1}{2} \frac{N\exp\left(-\frac{2\beta}{\sigma^2/2+\sigma_A^2}y_N\right)}{\exp\left(-\frac{2\beta}{\sigma^2+\sigma_A^2}y_N\right)} = 0.$$

The corollary follows. □

5.2. The case $a = a^\star$

In Section 3, we showed that we have at least two regimes, namely $0 < a < a^\star$, and $a \geq a^\star$. It turns out, that when we investigate sharper asymptotics, the case $a = a^\star$ deserves special attention. In the present section, we establish that in the case $a = a^\star$, $\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \sim \frac{1}{2}N^{-\gamma(a^\star)}$, thus the prefactor is 1/2 instead of 1 as in the case $a > a^\star$. To make the heuristics given in Section 2 rigorous, we proceed by deriving asymptotic lower and upper bounds, in two separate lemmas. As in Section 5.1, we prove that the lim inf converges to the desired limit. We do this in Lemma 8. The proof of this Lemma is similar to the proof of Lemma 7. However, the simple union bound $N\mathbb{P}(Q_{i,A}^\beta > f_N(a^\star)) \sim N^{-\gamma(a^\star)}$ is not tight for $a = a^\star$. Thus, we also need to prove that the lim sup is tight. We provide this proof in Lemma 9.

Lemma 8. *For the model given in Definition 1 with the additional notation given in Definition 2, and* $a = a^\star$*, we have that*

$$\liminf_{N\to\infty} N^{\gamma(a^\star)}\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \geq \frac{1}{2}.$$

Proof. First, we have the lower bound

$$\mathbb{P}(\bar{Q}_N^\beta > f_N(a^\star)) \geq \mathbb{P}(\max_{i\leq N} Q_{i,A}^\beta(\tau_{i,N}^{a^\star,r})\mathbb{1}(\tau_{i,N}^{a^\star,r} < \infty) > f_N(a^\star)).$$

As in (65) we can bound this further by Bonferroni's inequality to

$$N\mathbb{P}\left(Q_{i,A}^\beta(\tau_{i,N}^{a^\star,r})\mathbb{1}(\tau_{i,N}^{a^\star,r} < \infty) > f_N(a^\star)\right)$$

$$-\binom{N}{2}\mathbb{P}\left(\min(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty),Q^{\beta}_{j,A}(\tau^{a^\star,r}_{j,N})\mathbb{1}(\tau^{a^\star,r}_{j,N}<\infty))>f_N(a^\star)\right)$$
$$\geq\left(N-\frac{N^2}{2}\mathbb{P}\left(\tau^{a^\star,r}_{j,N}<\infty\right)\right)\mathbb{P}\left(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty)>f_N(a^\star)\right). \tag{69}$$

The last step is true because

$$\begin{aligned}&\mathbb{P}\left(\min(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty),Q^{\beta}_{j,A}(\tau^{a^\star,r}_{j,N})\mathbb{1}(\tau^{a^\star,r}_{j,N}<\infty))>f_N(a^\star)\right)\\&=\mathbb{P}\left(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty)>f_N(a^\star),Q^{\beta}_{j,A}(\tau^{a^\star,r}_{j,N})\mathbb{1}(\tau^{a^\star,r}_{j,N}<\infty)>f_N(a^\star)\right)\\&\leq\mathbb{P}\left(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty)>f_N(a^\star),\tau^{a^\star,r}_{j,N}<\infty\right)\\&=\mathbb{P}\left(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty)>f_N(a^\star)\right)\mathbb{P}(\tau^{a^\star,r}_{j,N}<\infty).\end{aligned}$$

Since $\mathbb{P}(\tau^{a^\star,r}_{j,N}<\infty)=\exp(-2(1-\lambda(a^\star))\beta r/\sigma^2)/N$, we can simplify the expression in (69) to

$$\left(1-\frac{\exp\left(-\frac{2(1-\lambda(a^\star))\beta r}{\sigma^2}\right)}{2}\right)N\mathbb{P}\left(Q^{\beta}_{i,A}(\tau^{a^\star,r}_{i,N})\mathbb{1}(\tau^{a^\star,r}_{i,N}<\infty)>f_N(a^\star)\right). \tag{70}$$

Following the same strategy as in the proof of Lemma 7, we have that

$$\begin{aligned}g_N(a^\star,x,k)&=f_N(a^\star)-(1-\lambda(a^\star))f_N(a^\star)+\lambda(a^\star)\beta T_N(a^\star,k)\\&\quad-\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N-x\sqrt{\log N}=\left(-x+\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N}.\end{aligned}$$

Now, for $x<\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$, it follows that

$$\begin{aligned}&N^{\gamma(a^\star)}\mathbb{P}(\hat{Q}^{\beta}_{i,A}>g_N(a^\star,x,k)-r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2(\sigma^2+\sigma_A^2)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\\&=N^{\gamma(a^\star)}\\&\quad\times\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2(\sigma^2+\sigma_A^2)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2T_N(a^\star,k)}-\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\left(-x+\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N}-r\right)\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\end{aligned} \tag{71}$$

By using the definition of $T_N(a^\star,k)$ in (8), we see that $\sqrt{\log N}/(\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)})\xrightarrow{N\to\infty}\beta\sigma/(\sqrt{\pi}\sigma_A(\sigma^2+\sigma_A^2))$. Furthermore, $\gamma(a^\star)\log N$ plus the exponent on the right-hand side of (71) equals

$$\begin{aligned}&\gamma(a^\star)\log N-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2T_N(a^\star,k)}\\&\quad-\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\left(-x+\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\sqrt{\log N}-r\right)\end{aligned}$$

$$\overset{N\to\infty}{\longrightarrow} -\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}+\frac{2\beta r}{\sigma^2+\sigma_A^2},$$

with a similar proof as in the proof of Lemma 4. Thus,

$$N^{\gamma(a^\star)}\mathbb{P}(\hat{Q}_{i,A}^\beta > g_N(a^\star,x,k)-r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$\overset{N\to\infty}{\longrightarrow} \frac{\beta\sigma}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}\exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\beta k\sigma_A^2\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)\exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right),$$

when $x < \sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$. When $x > \sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$, we see that $g_N(a^\star,x,k) = (-x+\sigma_A^2\beta k/(\sigma^2+\sigma_A^2))\sqrt{\log N} \overset{N\to\infty}{\longrightarrow} -\infty$, thus $\mathbb{P}(\hat{Q}_{i,A}^\beta > g_N(a^\star,x,k)-r) \overset{N\to\infty}{\longrightarrow} 1$. In this case, we get that

$$N^{\gamma(a^\star)}\mathbb{P}(\hat{Q}_{i,A}^\beta > g_N(a^\star,x,k)-r)\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$= N^{\gamma(a^\star)}\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$= \frac{\sqrt{\log N}}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}\exp\left(-\frac{2\beta\sigma_A^2\log N\left(\sigma_A^2(x-\beta k)+\sigma^2 x\right)+\beta^2\sigma^2x^2\sqrt{\log N}}{\sigma_A^2\left(2\beta^2k\sigma^2+(\sigma^2+\sigma_A^2)^2\sqrt{\log N}\right)}\right)$$

$$\overset{N\to\infty}{\longrightarrow} 0,$$

for $x > \sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$.

Thus, by combining this result with the result from Lemma 4, for $x < \sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$,

$$f_{\tau_{i,N}^{a^\star,r}}\left(T_N(a^\star,k)\right)N\sqrt{\log N}N^{\gamma(a^\star)}\mathbb{P}\left(\hat{Q}_{i,A}^\beta > g_N(a^\star,x,k)-r\right)$$

$$\times\frac{\sqrt{\log N}\exp\left(-\frac{\left(\frac{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)}{\beta\sigma^2}\log N+x\sqrt{\log N}\right)^2}{2\sigma_A^2 T_N(a^\star,k)}\right)}{\sqrt{2\pi}\sigma_A\sqrt{T_N(a^\star,k)}}$$

$$\overset{N\to\infty}{\longrightarrow} \frac{\beta^2\sigma^2\exp\left(-\frac{\beta\left(\beta^3k^2\sigma^4+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}$$

$$\times \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\sigma_A^2\beta k\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)$$
$$=: L_1(x,k).$$

The function $L_1(x,k)$ satisfies

$$\begin{aligned}
L_1(x,k) &= \frac{\beta^2\sigma^2 \exp\left(-\frac{\beta\left(\beta^3k^2\sigma^4+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2} \\
&\quad\times \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\sigma_A^2\beta k\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \exp\left(\frac{2\beta r}{\sigma^2+\sigma_A^2}\right) \\
&= \frac{\beta^2\sigma^2 \exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2} \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x\left(\sigma^2+\sigma_A^2\right)-2\sigma_A^2\beta k\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)} \\
&= \frac{\beta^2\sigma^2 \exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2} \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x-2\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}.
\end{aligned} \tag{72}$$

Thus, $L_1(x,k)$ can be written as a product of two densities of normally distributed random variables. When we consider the last term in (72) as a function of x, we get that the function

$$\frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x-2\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)\right)^2}{\sigma_A^2\left(\sigma^2+\sigma_A^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(\sigma^2+\sigma_A^2\right)}$$

is the density of a normally distributed random variable with mean $2\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)$ and standard deviation $\sigma_A(\sigma^2+\sigma_A^2)/(\sqrt{2}\beta\sigma)$. From this, it follows that

$$\begin{aligned}
&\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} \frac{\beta\sigma \exp\left(-\frac{\beta^2\sigma^2\left(x-2\beta k\sigma_A^2/(\sigma^2+\sigma_A^2)\right)^2}{\sigma_A^2(\sigma^2+\sigma_A^2)^2}\right)}{\sqrt{\pi}\sigma_A(\sigma^2+\sigma_A^2)} dx \\
&\quad= \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 + \frac{2\sigma_A^2\beta k}{\sigma^2+\sigma_A^2} \le \frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right) \\
&\quad= \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \le -\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right),
\end{aligned}$$

with X_1 standard normally distributed. Furthermore, when we consider the first term in (72) as a function of k, we get that the function

$$\frac{\beta^2\sigma^2 \exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}(\sigma^2+\sigma_A^2)^2}$$

is the density of a normally distributed random variable with mean 0 and standard deviation $(\sigma^2+\sigma_A^2)^2/(\sqrt{2}\beta^2\sigma^2)$. Therefore, we can conclude that the integral

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} L_1(x,k)dxdk \\
&\quad = \int_{-\infty}^{\infty} \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \le -\frac{\sigma_A^2\beta k}{\sigma^2+\sigma_A^2}\right)\frac{\beta^2\sigma^2\exp\left(-\frac{\beta^4k^2\sigma^4}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}dk \\
&\quad = \mathbb{P}\left(\frac{\sigma_A(\sigma^2+\sigma_A^2)}{\sqrt{2}\beta\sigma}X_1 \le -\frac{\sigma_A^2\beta}{\sigma^2+\sigma_A^2}\frac{(\sigma^2+\sigma_A^2)^2}{\sqrt{2}\beta^2\sigma^2}X_2\right) \\
&\quad = \frac{1}{2},
\end{aligned}$$

with X_2 standard normally distributed, and X_1 and X_2 mutually independent. Now, by applying Fatou's lemma, we have that

$$\begin{aligned}
&\liminf_{N\to\infty} N^{\gamma(a^\star)}N\mathbb{P}\left(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,r})\mathbb{1}(\tau_{i,N}^{a^\star,r}<\infty) > f_N(a^\star)\right) \\
&\quad \ge \int_{-\infty}^{\infty}\int_{-\infty}^{\sigma_A^2\beta k/(\sigma^2+\sigma_A^2)} L_1(x,k)dxdk \\
&\quad = \frac{1}{2}.
\end{aligned}$$

Thus, by applying this result to the expression in (70), we get that

$$\liminf_{N\to\infty} N^{\gamma(a^\star)}\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a^\star)) \ge \frac{1}{2}\left(1-\frac{\exp\left(-\frac{2(1-\lambda(a^\star))\beta r}{\sigma^2}\right)}{2}\right) \xrightarrow{r\to\infty} \frac{1}{2}. \quad \square$$

Lemma 9. *For the model given in Definition 1 with the additional notation given in Definition 2, and $a=a^\star$, we have that*

$$\limsup_{N\to\infty} N^{\gamma(a^\star)}\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a^\star)) \le \frac{1}{2}.$$

Proof. Let $\tilde{\tau}_{A,N}^{a^\star,r} = \inf\{t : W_A(t) - \lambda(a^\star)\beta t > \lambda(a^\star)f_N(a^\star) + r\}$. Following Eq. (56) and the upper bound in (57), we have that

$$\begin{aligned}
\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a^\star)) &\le \mathbb{P}(\bar{Q}_N^{\beta}(\tilde{\tau}_{A,N}^{a^\star,r})\mathbb{1}(\tilde{\tau}_{A,N}^{a^\star,r}<\infty) > f_N(a^\star)) \\
&\quad + \mathbb{P}(\max_{i\le N} Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})\mathbb{1}(\tau_{i,N}^{a^\star,-r}<\infty) > f_N(a^\star)).
\end{aligned} \tag{73}$$

Now, observe that we can bound the first term on the right-hand side of (73) as

$$\mathbb{P}(\bar{Q}_N^{\beta}(\tilde{\tau}_{A,N}^{a^\star,r})\mathbb{1}(\tilde{\tau}_{A,N}^{a^\star,r} < \infty) > f_N(a^\star)) \leq \mathbb{P}(\tilde{\tau}_{A,N}^{a^\star,r} < \infty) = N^{-\gamma(a^\star)} \exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right). \tag{74}$$

Furthermore, by using Eq. (55) in Definition 4, we can bound the second term on the right-hand side of (73) as

$$\begin{aligned} &N^{\gamma(a^\star)}\mathbb{P}(\max_{i\leq N} Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})\mathbb{1}(\tau_{i,N}^{a^\star,-r} < \infty) > f_N(a^\star)) \\ &\leq N^{\gamma(a^\star)} N\mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})\mathbb{1}(\tau_{i,N}^{a^\star,-r} < \infty) > f_N(a^\star)) \\ &= \int_{-\infty}^{\infty} N^{\gamma(a^\star)} N\mathbb{P}_{i,a^\star,-r,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r}) > f_N(a^\star)) f_{\tau_{i,N}^{a^\star,-r}}(T_N(a^\star,k))\sqrt{\log N}dk. \end{aligned} \tag{75}$$

Now, we examine the parts of the integrand of this integral, and we apply Lemma 5. First, note that, following Definition 3,

$$\mathbb{P}_{i,a^\star,-r,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r}) > f_N(a^\star)) = \mathbb{P}_{i,a^\star,-r,k}^{(N)}(W_A(\tau_{i,N}^{a^\star,-r}) + \hat{Q}_{i,A}^{\beta} > \lambda(a^\star) f_N(a^\star) + r + \lambda(a^\star)\beta\tau_{i,N}^{a^\star,-r}).$$

We can analyze this probability using Lemma 3 by taking $x_N = 2\lambda(a^\star) f_N(a^\star) + \lambda(a^\star)\beta k\sqrt{\log N} + r$, $\eta_N = \sigma_A\sqrt{T_N(a^\star,k)}$, and $\mu = 2\beta/(\sigma^2+\sigma_A^2)$. Write

$$\begin{aligned} \frac{x_N - \mu\eta_N^2}{\sqrt{2}\eta_N} &= \frac{2\lambda(a^\star) f_N(a^\star) + \lambda(a^\star)\beta k\sqrt{\log N} + r - \frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2 T_N(a^\star,k)}{\sqrt{2}\sqrt{\sigma_A^2 T_N(a^\star,k)}} \\ &= \frac{r - \lambda(a^\star)\beta k\sqrt{\log N}}{\sqrt{2}\sqrt{\sigma_A^2 T_N(a^\star,k)}} \\ &\xrightarrow{N\to\infty} -\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}. \end{aligned}$$

The first term on the right-hand side of (42) in Lemma 3 satisfies

$$\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} \sim \frac{\sigma \exp\left(-\frac{\beta\left(\beta^3k^2\sigma_A^2\sigma^2+2r\left(\sigma^2+\sigma_A^2\right)^3\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{2\sqrt{\pi}\sigma_A}\frac{N^{-\gamma(a^\star)}}{\sqrt{\log N}},$$

as $N\to\infty$, and the second term satisfies

$$\begin{aligned} \frac{1}{2}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\left(1+\mathrm{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right) &\sim \frac{1}{2}\exp\left(-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right) \\ &\quad\times\left(1+\mathrm{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)N^{-\gamma(a^\star)}, \end{aligned}$$

as $N\to\infty$. So, we can conclude that

$$\begin{aligned} \mathbb{P}_{i,a^\star,-r,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r}) > f_N(a^\star)) &\sim \frac{1}{2}\exp\left(-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right) \\ &\quad\times\left(1+\mathrm{erf}\left(-\frac{\beta^2\sigma\sigma_A k}{(\sigma^2+\sigma_A^2)^2}\right)\right)N^{-\gamma(a^\star)}, \end{aligned}$$

as $N \to \infty$. Second, following Lemma 4, the density of the hitting time $\tau_{i,N}^{a^\star,-r}$ appears in the integrand in (75), and satisfies

$$Nf_{\tau_{i,N}^{a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N} \xrightarrow{N\to\infty} \frac{\beta^2 \exp\left(\frac{\beta\left(8{a^\star}^2\beta^2 r-\beta^3k^2\sigma\sqrt{2a^\star\beta+\sigma^2}+8a^\star\beta r\sigma^2+2r\sigma^4\right)}{\sigma\left(2a^\star\beta+\sigma^2\right)^{5/2}}\right)}{\sqrt{\pi}\left(2a^\star\beta+\sigma^2\right)}$$

$$=\frac{\beta^2\sigma^2\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}\right)}{\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}.$$

Thus, for the integrand in (75) we have that

$$N^{\gamma(a^\star)}N\mathbb{P}_{i,a^\star,-r,k}^{(N)}\left(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})>f_N(a^\star)\right)f_{\tau_{i,N}^{a^\star,-r}}\left(T_N(a^\star,k)\right)\sqrt{\log N}$$
$$\xrightarrow{N\to\infty}\frac{\beta^2\sigma^2\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_Ak}{(\sigma^2+\sigma_A^2)^2}\right)\right)\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)}{2\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}.$$

When we integrate this result we get

$$\int_{-\infty}^{\infty}\frac{\beta^2\sigma^2\left(1+\operatorname{erf}\left(-\frac{\beta^2\sigma\sigma_Ak}{(\sigma^2+\sigma_A^2)^2}\right)\right)\exp\left(\frac{\beta\left(2r\left(\sigma^2+\sigma_A^2\right)^3-\beta^3k^2\sigma^4\right)}{\left(\sigma^2+\sigma_A^2\right)^4}-\frac{2\beta r}{\sigma^2+\sigma_A^2}\right)}{2\sqrt{\pi}\left(\sigma^2+\sigma_A^2\right)^2}dk=\frac{1}{2}.$$

Now, we argue that the fourth condition of Lemma 5 holds, more specifically that $N^{\gamma(a^\star)}\mathbb{P}_{i,a^\star,-r,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})>f_N(a^\star))$ is bounded for all N and k. First, we have for $a=a^\star$ that

$$\sup_{s>0}(W_A(s)-\lambda(a^\star)\beta s)\stackrel{d}{=}\sup_{s>0}(W_i(s)-(1-\lambda(a^\star))\beta s)\stackrel{d}{=}\sup_{s>0}(W_i(s)+W_A(s)-\beta s).$$

Thus,

$$\begin{aligned}
&N^{\gamma(a^\star)}\mathbb{P}_{i,a^\star,-r,k}^{(N)}\left(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})>f_N(a^\star)\right)\\
&=N^{\gamma(a^\star)}\mathbb{P}_{i,a^\star,-r,k}^{(N)}\left(\sup_{s>\tau_{i,N}^{a^\star,-r}}(W_i(s)+W_A(s)-\beta s)>f_N(a^\star)\right)\\
&=N^{\gamma(a^\star)}\\
&\quad\times\mathbb{P}_{i,a^\star,-r,k}^{(N)}\left(W_A(\tau_{i,N}^{a^\star,-r})-\lambda(a^\star)\beta\tau_{i,N}^{a^\star,-r}+\sup_{s>0}(\hat{W}_i(s)+\hat{W}_A(s)-\beta s)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&=N^{\gamma(a^\star)}\mathbb{P}_{i,a^\star,-r,k}^{(N)}\left(\sup_{s>\tau_{i,N}^{a^\star,-r}}(\hat{W}_A(s)-\lambda(a^\star)\beta s)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&\le N^{\gamma(a^\star)}\mathbb{P}\left(\sup_{s>0}\left(W_A(s)-\lambda(a^\star)\beta s\right)>\lambda(a^\star)f_N(a^\star)+r\right)\\
&=N^{\gamma(a^\star)}\exp\left(-\frac{2\lambda(a^\star)\beta}{\sigma_A^2}(\lambda(a^\star)f_N(a^\star)+r)\right)
\end{aligned}$$

$$= \exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right).$$

Thus, the fourth condition of Lemma 5 holds. Furthermore, we have that

$$\lim_{N\to\infty}\int_{-\infty}^{\infty} N f_{\tau_{i,N}^{a^\star,-r}}\big(T_N(a^\star,k)\big)\sqrt{\log N}dk = \int_{-\infty}^{\infty}\lim_{N\to\infty} N f_{\tau_{i,N}^{a^\star,-r}}\big(T_N(a^\star,k)\big)\sqrt{\log N}dk,$$

which means that the second condition of Lemma 5 also holds. Now, we can use Lemma 5 to conclude that the upper bound in (75) is asymptotically bounded;

$$\limsup_{N\to\infty} N^{\gamma(a^\star)} N\mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a^\star,-r})\mathbb{1}(\tau_{i,N}^{a^\star,-r}<\infty) > f_N(a^\star)) \leq \frac{1}{2}. \tag{76}$$

Now, after combining the bounds in (74) and (76),

$$\limsup_{N\to\infty} N^{\gamma(a^\star)}\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a^\star)) \leq \frac{1}{2} + \exp\left(-\frac{2\lambda(a^\star)\beta r}{\sigma_A^2}\right) \xrightarrow{r\to\infty} \frac{1}{2}. \quad \square$$

5.3. *The case* $0 < a < a^\star$

In Theorem 1, we have shown that $\gamma(a) = \frac{2a\beta+2\sigma^2-2\sigma\sqrt{2a\beta+\sigma^2}}{\sigma_A^2}$. Therefore, we expect highly dependent behavior for the tail asymptotics of the maximum queue length, as this indicates that the union upper bound $\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a)) \leq N\mathbb{P}(Q_{i,A}^{\beta} > f_N(a))$ is not sharp when $0 < a < a^\star$, as is explained in the proof of Lemma 2.

Proof of Theorem 4. First, we prove Eq. (18). We write

$$r_N := \frac{\sigma\sqrt{2a\beta+\sigma^2}}{4\beta}\log\log N.$$

Let $\tilde{\tau}_{A,N}^{a,r_N} = \inf\{t \geq 0 : W_A(t) - \lambda(a)\beta t > \lambda(a) f_N(a) + r_N\}$. Let $f_{\tilde{\tau}_{A,N}^{a,r_N}}$ be its density. Observe that

$$\begin{aligned}
&\mathbb{P}(\bar{Q}_N^{\beta} > f_N(a))\\
&\quad\geq \mathbb{P}(\bar{Q}_N^{\beta}(\tilde{\tau}_{A,N}^{a,r_N},\tilde{\tau}_{A,N}^{a,r_N})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r_N}<\infty) > f_N(a))\\
&\quad= \int_{-\infty}^{\infty}\mathbb{P}\left(\max_{i\leq N} W_i\big(T_N(a,k)\big) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N\right)\\
&\qquad\times f_{\tilde{\tau}_{A,N}^{a,r_N}}\big(T_N(a,k)\big)\sqrt{\log N}dk.
\end{aligned} \tag{77}$$

As in the proof of Lemma 9, we analyze the components of the integrand of (77) separately. By following a similar derivation as in Lemma 4, we see that the hitting-time density $f_{\tilde{\tau}_{A,N}^{a,r_N}}\big(T_N(a,k)\big)$ in (77), with $\tilde{\tau}_{A,N}^{a,r_N}$ defined in Definition 3 and the hitting-time density given in [4, Eq. (2.0.2), p. 301], satisfies

$$
\begin{aligned}
&N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} f_{\tilde{\tau}_{A,N}^{a,r_N}}\left(T_N(a,k)\right)\sqrt{\log N} \\
&\quad = N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} \frac{\lambda(a) f_N(a)+r_N}{\sqrt{2\pi}\sigma_A T_N(a,k)^{3/2}} \\
&\qquad \times \exp\left(-\frac{(\lambda(a)f_N(a)+r_N+\lambda(a)\beta T_N(a,k))^2}{2\sigma_A^2 T_N(a,k)}\right)\sqrt{\log N} \\
&\quad \overset{N\to\infty}{\longrightarrow} \frac{\beta^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)\exp\left(-\frac{\beta^4k^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{\sqrt{\pi}\sigma_A\left(2a\beta+\sigma^2\right)}.
\end{aligned}
\tag{78}
$$

Moreover, it is proven in [8, Ex. 1.1.7, p. 11] that for

$$
b_N = \sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}},
$$

we have that

$$
b_N\left(\frac{\max_{i\le N} W_i(d\log N)}{\sigma\sqrt{d\log N}} - b_N\right) \xrightarrow{d} G,
$$

as $N\to\infty$, with $G\sim$ Gumbel. From this, it follows that the term $\mathbb{P}(\max_{i\le N} W_i\left(T_N(a,k)\right) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N)$ in (77) satisfies

$$
\begin{aligned}
&\mathbb{P}\left(\max_{i\le N} W_i\left(T_N(a,k)\right) - (1-\lambda(a))\beta T_N(a,k) > (1-\lambda(a))f_N(a) - r_N\right) \\
&\quad \overset{N\to\infty}{\longrightarrow} 1-\exp\left(-\frac{\exp\left(-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}}\right).
\end{aligned}
\tag{79}
$$

Thus, the product of the limits in (78) and (79) gives the tail asymptotics of the integrand in (77). Now, by applying Fatou's lemma, we obtain a sharper than logarithmic lower bound on the asymptotics for the maximum queue length, and is given in (18).

In order to prove (19), we use the upper bound given in (57) and observe that

$$
\mathbb{P}(\bar{Q}_N^\beta > f_N(a)) \le \mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,N}^{a,r_N})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r_N} < \infty) > f_N(a)) \tag{80}
$$

$$
+\,\mathbb{P}(\max_{i\le N} Q_{i,A}^\beta(\tau_{i,N}^{a,-r_N})\mathbb{1}(\tau_{i,N}^{a,-r_N} < \infty) > f_N(a)). \tag{81}
$$

We can bound the expression in (80) as follows:

$$
\mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,N}^{a,r_N})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r_N} < \infty) > f_N(a)) \le \mathbb{P}(\tilde{\tau}_{A,N}^{a,r_N} < \infty) = N^{-\gamma(a)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}. \tag{82}
$$

Therefore,

$$
\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\bar{Q}_N^\beta(\tilde{\tau}_{A,N}^{a,r_N})\mathbb{1}(\tilde{\tau}_{A,N}^{a,r_N} < \infty) > f_N(a)) \le 1.
$$

Hence, because of the bounds given in (80) and (81), to prove that (19) holds, it is left to show that

$$\limsup_{N\to\infty} N^{\gamma(a)}(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}\mathbb{P}(\max_{i\le N} Q_{i,A}^{\beta}(\tau_{i,N}^{a,-r_N})\mathbb{1}(\tau_{i,N}^{a,-r_N}<\infty)>f_N(a))<\infty.$$

To prove this, observe that, by using the union bound and by conditioning on the hitting time $\tau_{i,N}^{a,-r_N}$ the expression in (81) satisfies

$$\begin{aligned}
&\mathbb{P}(\max_{i\le N} Q_{i,A}^{\beta}(\tau_{i,N}^{a,-r_N})\mathbb{1}(\tau_{i,N}^{a,-r_N}<\infty)>f_N(a))\\
&\le N\mathbb{P}(Q_{i,A}^{\beta}(\tau_{i,N}^{a,-r_N})\mathbb{1}(\tau_{i,N}^{a,-r_N}<\infty)>f_N(a))\\
&=\int_{-\infty}^{\infty} N\mathbb{P}_{i,a,-r_N,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a,-r_N})>f_N(a))f_{\tau_{i,N}^{a,-r_N}}\big(T_N(a,k)\big)\sqrt{\log N}dk. \qquad (83)
\end{aligned}$$

Now, we can use Lemma 5 to show convergence of the integral in (83). By following a similar analysis as in Lemma 4 and by using the expression of the hitting-time density given in [4, Eq. (2.0.2), p. 301], we have that

$$\begin{aligned}
&N\frac{1}{\sqrt{\log N}}\sqrt{\log N}f_{\tau_{i,N}^{a,-r_N}}\big(T_N(a,k)\big)\\
&=N\frac{(1-\lambda(a))f_N(a)-r_N}{\sqrt{2\pi}\sigma T_N(a,k)^{3/2}}\exp\left(-\frac{((1-\lambda(a))f_N(a)-r_N+(1-\lambda(a))\beta T_N(a,k))^2}{2\sigma^2 T_N(a,k)}\right)\\
&\xrightarrow{N\to\infty}\frac{\beta^2\exp\left(-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}\right)}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}.
\end{aligned}$$

Furthermore,

$$\int_{-\infty}^{\infty}\frac{\beta^2 e^{-\frac{\beta^4k^2}{(2a\beta+\sigma^2)^2}}}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)}dk=\int_{-\infty}^{\infty}\frac{N}{\sqrt{\log N}}\sqrt{\log N}f_{\tau_{i,N}^{a,-r_N}}\big(T_N(a,k)\big)dk=1. \qquad (84)$$

Thus, the first and second condition in Lemma 5 hold. To complete the proof, we now only need to analyze

$$\mathbb{P}_{i,a,-r_N,k}^{(N)}(Q_{i,A}^{\beta}(\tau_{i,N}^{a,-r_N})>f_N(a))=\mathbb{P}_{i,a,-r_N,k}^{(N)}(W_A(\tau_{i,N}^{a,-r_N})+\hat{Q}_{i,A}^{\beta}>\lambda(a)f_N(a)+r_N+\lambda(a)\beta\tau_{i,N}^{a,-r_N}), \qquad (85)$$

which is a component in the integrand in (83). We show that this expression satisfies the third and fourth condition of Lemma 5 by proving pointwise convergence and by proving that this probability is uniformly bounded by a constant. To do this, first observe that the random variable in (85) has the form of the sum of a normally distributed random variable and an exponentially distributed random variable. Hence we can follow the framework of Lemma 3 in order to analyze this probability. We take $x_N=2\lambda(a)f_N(a)+\lambda(a)\beta k\sqrt{\log N}+r_N$, $\eta_N=\sigma_A\sqrt{T_N(a,k)}$, and $\mu=2\beta/(\sigma^2+\sigma_A^2)$. Now, the expression in (85) can be written in the form of Eq. (45). Furthermore, observe that

$$\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}=\frac{2\lambda(a)f_N(a)+\lambda(a)\beta k\sqrt{\log N}+r_N-\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2T_N(a,k)}{\sqrt{2}\sqrt{\sigma_A^2T_N(a,k)}}\xrightarrow{N\to\infty}-\infty.$$

Thus, for $0 < a < a^\star$, we are in the third situation of Lemma 3. Following the same analysis as in the proof of Lemma 4, we see that the first term in (44) satisfies

$$\frac{\eta_N e^{-\frac{x_N^2}{2\eta_N^2}}}{\sqrt{2\pi}x_N} \sim \frac{\sigma_A \exp\left(-\frac{\beta^4 k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)}\frac{1}{\sqrt{\log N}},$$

as $N \to \infty$. Furthermore, we have for all $t > 0$ that

$$\mathbb{P}\left(W_A(t) - \lambda(a)\beta t > x\right) \le \mathbb{P}\left(W_A(x/(\lambda(a)\beta)) > 2x\right).$$

From this, it follows that the first part in (45) satisfies

$$\begin{aligned}
&\mathbb{P}\left(\eta_N X > x_N\right)\\
&\quad= \mathbb{P}\left(W_A(\tau_{i,N}^{a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau_{i,N}^{a,-r_N}\,\middle|\,\tau_{i,N}^{a,-r_N} = T_N(a,k)\right)\\
&\quad\le \mathbb{P}\left(W_A(\tau_{i,N}^{a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau_{i,N}^{a,-r_N}\,\middle|\,\tau_{i,N}^{a,-r_N} = \frac{f_N(a)}{\beta} + \frac{r_N}{\lambda(a)\beta}\right)\\
&\quad\sim \frac{\sigma_A}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)}\frac{1}{\sqrt{\log N}},
\end{aligned}$$

as $N \to \infty$. So there exists an $\epsilon > 0$ and an N_ϵ such that for $N > N_\epsilon$ and all $k > -f_N(a)/(\beta\sqrt{\log N})$,

$$\begin{aligned}
&(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{\gamma(a)}\sqrt{\log N}\,\mathbb{P}^{(N)}_{i,a,-r_N,k}\left(W_A(\tau_{i,N}^{a,-r_N}) > \lambda(a)f_N(a) + r_N + \lambda(a)\beta\tau_{i,N}^{a,-r_N}\right)\\
&\quad\le \frac{\sigma_A}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)} + \epsilon.
\end{aligned} \tag{86}$$

The second term in (44) satisfies

$$\begin{aligned}
&-\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\frac{\eta_N e^{-\frac{\left(x_N-\mu\eta_N^2\right)^2}{2\eta_N^2}}}{x_N-\mu\eta_N^2}\\
&\quad\sim \frac{\sigma_A\left(\sigma^2+\sigma_A^2\right)\exp\left(-\frac{2\beta^4k^2\left(\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)+a\beta\left(\sqrt{2a\beta+\sigma^2}-2\sigma\right)\right)}{\sigma_A^2\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{2\sqrt{\pi}\sigma\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)}\\
&\qquad\times(\log N)^{-\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}} N^{-\gamma(a)}\frac{1}{\sqrt{\log N}},
\end{aligned} \tag{87}$$

as $N \to \infty$. In this case, first observe that in Eq. (45) the exact expression of the convolution term equals

$$\int_{-\infty}^{x_N/\eta_N}\mathbb{P}\left(\frac{1}{\mu}E > x_N - \eta_N z\right)\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz = \frac{1}{2}\left(\operatorname{erf}\left(\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}\right)+1\right)e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}.$$

Second, observe that this can be further rewritten into

$$\begin{aligned}
&\frac{1}{2}\left(\operatorname{erf}\left(\frac{x_N-\mu\eta_N^2}{\sqrt{2}\eta_N}\right)+1\right)e^{\frac{1}{2}\mu\left(\mu\eta_N^2-2x_N\right)}\\
&=\mathbb{P}^{(N)}_{i,a,-r_N,k}\left(W_A(\tau^{a,-r_N}_{i,N})>\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2\tau^{a,-r_N}_{i,N}-\lambda(a)f_N(a)-r_N-\lambda(a)\beta\tau^{a,-r_N}_{i,N}\right)\\
&\quad\times\exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2T_N(a,k)-2\lambda(a)f_N(a)-2\lambda(a)\beta T_N(a,k)-2r_N\right)\right).
\end{aligned}$$

Thus, the expression that we are investigating is a product of a tail probability of a Gaussian random variable and an exponential function. With an analogous derivation as for the first term in (44), due to the expression in (87) we can bound for all $t>0$

$$\begin{aligned}
&(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}N^{\gamma(a)}\sqrt{\log N}\mathbb{P}\left(W_A(t)>\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2t-\lambda(a)f_N(a)-r_N-\lambda(a)\beta t\right)\\
&\quad\times\exp\left(\frac{1}{2}\frac{2\beta}{\sigma^2+\sigma_A^2}\left(\frac{2\beta}{\sigma^2+\sigma_A^2}\sigma_A^2t-2\lambda(a)f_N(a)-2\lambda(a)\beta t-2r_N\right)\right). \qquad (88)
\end{aligned}$$

Hence, due to the upper bounds for (86) and (88), we have that the third and fourth condition of Lemma 5 are satisfied. Thus, in the end, we know that

$$\begin{aligned}
&(\log N)^{\frac{\lambda(a)}{1-\lambda(a)}\frac{\sigma^2}{2\sigma_A^2}}N^{\gamma(a)}N\mathbb{P}^{(N)}_{i,a,-r_N,k}(Q^{\beta}_{i,A}(\tau^{a,-r_N}_{i,N})>f_N(a))f_{\tau^{a,-r_N}_{i,N}}\left(T_N(a,k)\right)\sqrt{\log N}\\
&\xrightarrow{N\to\infty}\left(\frac{\sigma_A\exp\left(-\frac{\beta^4k^2\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)^2}{\sigma_A^2\left(2a\beta+\sigma^2\right)^2}\right)}{2\sqrt{\pi}\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)}\right.\\
&\qquad\left.+\frac{\sigma_A\left(\sigma^2+\sigma_A^2\right)\exp\left(-\frac{2\beta^4k^2\left(\sigma^2\left(\sqrt{2a\beta+\sigma^2}-\sigma\right)+a\beta\left(\sqrt{2a\beta+\sigma^2}-2\sigma\right)\right)}{\sigma_A^2\left(2a\beta+\sigma^2\right)^{5/2}}\right)}{2\sqrt{\pi}\sigma\left(\sigma\left(\sigma-\sqrt{2a\beta+\sigma^2}\right)+\sigma_A^2\right)}\right)\\
&\qquad\times\frac{\beta^2e^{-\frac{\beta^4k^2}{\left(2a\beta+\sigma^2\right)^2}}}{\sqrt{\pi}\left(2a\beta+\sigma^2\right)},
\end{aligned}$$

and we apply Lemma 5 to conclude that (19) holds. □

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is part of the research program Complexity in high-tech manufacturing, (partly) financed by the Dutch Research Council (NWO) through contract 438.16.121.

References

[1] Milton Abramowitz, Irene A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Tenth, Vol. 55, US Government printing office, 1972.

[2] François Baccelli, Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case, Technical Report RR–0426, INRIA, 1985.

[3] François Baccelli, Armand M. Makowski, Queueing models for systems with synchronization constraints, Proc. IEEE 77 (1) (1989) 138–161.

[4] Andrei N. Borodin, Paavo Salminen, Handbook of Brownian Motion-Facts and Formulae, Springer Science & Business Media, 2015.

[5] Krzysztof Dębicki, Enkelejd Hashorva, Lanpeng Ji, Kamil Tabiś, Extremes of vector-valued Gaussian processes: Exact asymptotics, Stoch. Process. Appl. 125 (11) (2015) 4039–4065.

[6] Krzysztof Dębicki, Lanpeng Ji, Tomasz Rolski, Exact asymptotics of component-wise extrema of two-dimensional Brownian motion, Extremes 23 (2020) 569—602.

[7] Leopold Flatto, Sann Hahn, Two parallel queues created by arrivals with two demands I, SIAM J. Appl. Math. 44 (5) (1984) 1041–1053.

[8] Laurens de Haan, Ana Ferreira, Extreme Value Theory: An Introduction, Springer Science & Business Media, 2006.

[9] Stephanus J. de Klein, Fredholm Integral Equations in Queueing Analysis (Ph.D. thesis), Rijksuniversiteit Utrecht, 1988.

[10] Sung-Seok Ko, Richard F. Serfozo, Response times in M/M/s fork-join networks, Adv. Appl. Probab. 36 (3) (2004) 854–871.

[11] Steven Kou, Haowen Zhong, et al., First-passage times of two-dimensional Brownian motion, Adv. Appl. Probab. 48 (4) (2016) 1045–1060.

[12] Hongyuan Lu, Guodong Pang, Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic, Math. Oper. Res. 41 (2) (2015) 560–595.

[13] Hongyuan Lu, Guodong Pang, Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime, Stoch. Syst. 6 (2) (2017a) 519–600.

[14] Hongyuan Lu, Guodong Pang, Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services, Queueing Syst. 85 (1–2) (2017b) 67–115.

[15] Michel Mandjes, Large Deviations for Gaussian Queues: Modelling Communication Networks, John Wiley & Sons, 2007.

[16] Mirjam Meijer, Dennis Schol, Willem van Jaarsveld, Maria Vlasiou, Bert Zwart, Extreme-value theory for large fork-join queues, with applications to high-tech supply chains, 2021, https://Arxiv.Org/Abs/2105.09189.

[17] Randolph Nelson, Asser N. Tantawi, Approximate analysis of fork/join synchronization in parallel queues, IEEE Trans. Comput. 37 (6) (1988) 739–743.

[18] Viên Nguyen, Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits, Ann. Appl. Probab. (1993) 28–55.

[19] Viên Nguyen, The trouble with diversity: Fork-join networks with heterogeneous customer population, Ann. Appl. Probab. (1994) 1–25.

[20] Vladimir Piterbarg, Asymptotic Methods in the Theory of Gaussian Processes and Fields, Vol. 148, American Mathematical Soc., 1996.

[21] Dennis Schol, Maria Vlasiou, Bert Zwart, Large fork-join queues with nearly deterministic arrival and service times, Math. Oper. Res. 47 (2) (2021) 1335–1364, http://dx.doi.org/10.1287/moor.2021.1171.

[22] Subir Varma, Heavy and Light Traffic Approximations for Queues with Synchronization Constraints (Ph.D. thesis), University of Maryland, 1990.

[23] Paul E. Wright, Two parallel processors with coupled inputs, Adv. Appl. Probab. 24 (4) (1992) 986–1007.