

Extreme-value theory for large fork-join queues, with applications to high-tech supply chains

Mirjam S. Meijer

Eindhoven University of Technology m.s.meijer@tue.nl

Dennis Schol

Eindhoven University of Technology c.schol@tue.nl

Willem van Jaarsveld

Eindhoven University of Technology w.l.v.jaarsveld@tue.nl

Maria Vlasiou

University of Twente, Eindhoven University of Technology m.vlasiou@tue.nl

Bert Zwart

Eindhoven University of Technology, CWI bert.zwart@cw.nl

We study extreme values in certain fork-join queueing networks: consider N identical queues with a common arrival process and independent service processes. All arrival and service processes are deterministic with random perturbations following Brownian motions. We prove that as $N \rightarrow \infty$, the scaled maximum of N steady-state queue lengths converges in distribution to a normally distributed random variable.

We explore repercussions of this result for original equipment manufacturers (OEMs) that assemble a large number of components, each produced using specialized equipment, into complex systems. Component production capacity is subject to fluctuations, causing high risk of shortages of at least one component, which results in costly system production delays. OEMs hedge this risk by investing in a combination of excess production capacity and component inventories. We formulate a stylized model of the OEM that enables us to study the resulting trade-off between shortage risk, inventory costs, and capacity costs. Our asymptotic extreme value results translate into asymptotically exact methods for cost-optimal inventory and capacity decisions, some of which are in closed form. We validate our asymptotic results with a set of detailed numerical experiments. These experiments indicate that our results are asymptotically exact, while for transient times they depend on model parameters.

Key words: extreme value theory; fork-join queue; capacitated inventory and assembly systems

1. Introduction

Fork-join queueing networks are a key modeling tool in stochastic operations research, as they capture many situations in which parts of jobs need to be assembled. One can think of applications as supply chains, manufacturing systems, and computer and communication networks. The analysis of these networks poses serious challenges; for example, the requirement that all components of a final product need to be physically present for the assembly process causes dependencies that are hard to analyze. In this paper, we look at a fork-join queueing network that consists of a large number of parallel queues. In large systems, one can expect that delays due to stochasticity of demand and service processes grow without bound as a function of the size of the system. Our aim is to analyze and quantify this phenomenon, as well as its impact on determining the capacity of the system.

To this end, we consider a fork-join network of N statistically identical queues driven by a common arrival process and having independent service processes. All arrival and service processes consist of a deterministic term, perturbed by (independent) Brownian motions. We are interested in the behavior of the maximal queue length in steady state as the number of queues grows large. We examine

separately the cases of purely deterministic arrivals and of perturbed arrivals. Our asymptotic results provide insight into the performance of large fork-join networks. The proof techniques we use are quite generic. For deterministic arrivals, we use standard extreme value theory, while for correlated arrivals, we rely on sample path analysis and conditional limit theorems for large suprema of Brownian motions.

When the arrival process is deterministic, the stationary queue lengths are independent and exponentially distributed. Standard results from extreme value theory imply that the scaled maximum queue length converges to a Gumbel distributed random variable as the number of queues $N \rightarrow \infty$. A goal of this paper is to investigate the impact of this scaling law on the simultaneous optimization of capacity and inventory of this class of assembly systems. Such simultaneous optimization is computationally challenging, but we show that this optimization becomes tractable as $N \rightarrow \infty$. The inventory and capacity induced by the extreme value limit are asymptotically correct and the convergence rate is fast.

When the arrival process is deterministic plus a random perturbation following a Brownian motion, the stationary queue lengths are still exponentially distributed, but no longer independent. The question is now how this affects the maximum queue length as the number of queues $N \rightarrow \infty$. Most of the work in extreme value theory has been done for independent random variables; cf. De Haan and Ferreira (2006), Resnick (1987). It turns out that suitable results from extreme value theory are absent for our setting. Thus, deriving a convergence result for the maximum queue length for perturbed arrivals as $N \rightarrow \infty$ is one of the key technical challenges underlying this paper. Our answer to this challenge is somewhat surprising: the dependence structure causes the scaled maximum queue length to converge to a normally distributed random variable as $N \rightarrow \infty$. That this scaled maximum is in the domain of attraction of the normal distribution is remarkable since for independent random variables, such a scaled maximum can only converge to a Gumbel, a Weibull or a Fréchet distributed random variable. Thus, our result shows that the normal distribution has a non-empty domain of attraction in an extreme-value theory context. An intuitive explanation of this fact, based on asymptotic independence of hitting times, is provided in Section 5.

The above-mentioned theoretical results can be applied to develop structural insights into the dimensioning of assembly systems. In particular, we explore repercussions of our results for high-tech equipment manufacturers (OEMs), for example Airbus and ASML. High-tech equipment is typically assembled-to-order from thousands of specialized *components*. The production of components involves highly skilled staff and specialized equipment: It is *capacitated* and subject to random fluctuations. Component shortages result in delays in system assembly, which results in costly product delivery delays. Also, when an assembly delay occurs because of a missing component, all other components need to be stocked, incurring holding costs.

OEMs spend billions of dollars on spare component production capacity and component inventories in the hope of guaranteeing a reliable production system (ASML Holding N.V. 2021). However, despite decades of research in inventory management, the joint optimization of production capacity and inventory remains a challenge (Bradley and Glynn 2002), and there is a lack of analytical results that may aid OEMs in analyzing the crucial trade-offs that underlie the outcome of their investments. Indeed, while the topic has increasingly been studied (see e.g. Reed and Zhang 2017), the focus of analysis has been on problems with a single component. We consider the much more common situation of assembling a system from many components, and we aim to choose capacity and inventory levels that minimize the sum of holding, capacity and backorder costs.

To appropriately model fluctuations in production capacity in continuous time, the cumulative production of each component is modelled as a Brownian motion with drift. This is a natural extension of normally distributed production capacity in discrete time, which is a common choice in the literature (e.g. Bradley and Glynn 2002, Wu and Chao 2014). OEMs typically *level* the demand to smooth the production process. Accordingly, in our base model we assume that demand

is completely levelled/deterministic. For this base model, in Section 4 we derive easy to calculate expressions for capacity and inventory that are asymptotically optimal as the number of components grows large. We provide order bounds between the costs under optimal and approximate inventory and capacity.

In particular, inspired by the literature on call centers: Borst et al. (2004), Gans et al. (2003) and Van Leeuwen et al. (2019) we distinguish three regimes, which depend on the growth rates of cost parameters and are determined by the probability γ_N of not having enough inventory. Given that $\gamma_N \rightarrow \gamma$, we say that the regime is *balanced* if $\gamma \in (0, 1)$. Furthermore we are in the quality driven regime if $\gamma = 0$ and in the efficiency driven regime if $\gamma = 1$. For the base model, we establish asymptotic cost optimality in all three regimes. For the balanced, quality driven, and efficiency driven regimes, we have convergence rates of $1/(N \log N)$, $\gamma_N/(N \log(N/\gamma_N))$ and $1/\log N$ respectively.

Despite efforts to level demand, typically some demand variation remains. Therefore in Section 5, we assume that the stochastic demand for systems is modelled by a Brownian motion. This implies that the demand over any finite time period is a normal variable, which is a standard assumption in literature (e.g. Klosterhalfen et al. 2014, Atan and Rousseau 2016). As a consequence, component delays become *dependent*, since they face the same stochastic demands from system assembly. Our main technical result for dependent Brownian motions implies that, with proper scaling of holding and backorder costs, the optimal inventory for stochastic demand converges to a scaled version of the quantile function of the normal distribution, while this quantile function also appears in the limit of the optimal capacity. Numerical experiments show that we typically are most of the times 10% off the optimum (e.g. when N is in the range from 10 to 100); cf. Tables 5 and 6. Naturally, the difference goes to 0 as $N \rightarrow \infty$; cf. Theorem 5.11.

We give an improvement of this approximation by combining our results for deterministic demand and stochastic demand. Based on this approximation, we optimize the capacity and inventory decisions and we test the quality of these approximations through numerical experiments. It turns out that these approximations perform well already when considering a limited number of components, and are typically less than 2% off the optimum.

This paper generates novel insights in fork-join queues. These insights lead to new analytical results for an important class of assembly systems: This paper is the first to consider simultaneous optimization of inventory and capacity in a multi-component assembly system with dependent delays. Due to the dependencies in delays, evaluating such a system with fixed capacity and inventory is already a difficult problem, unless you resort to simulation. We provide several asymptotically optimal expressions for capacity and inventory that are either in closed-form or can easily be computed numerically.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of relevant literature. The content of the paper is then structured around the application to high-tech assembly systems, with theoretical results appearing as we need them. In particular, we introduce the general mathematical model in Section 3 and subsequently present the optimization problem where we need to decide on capacity and inventory to minimize costs. We study the assembly system with deterministic demand in Section 4. We provide explicit expressions and approximations for optimal inventory and capacity. The stochastic demand case, with solutions to the minimization problem and convergence results, is studied in more detail in Section 5. That section also includes our key result on the extremal behavior of dependent Brownian fork-join queues, given in Theorem 5.2. A refinement of the approximations from Section 5 is provided in Section 6, where we combine the lessons learnt in Sections 4 and 5 to obtain better approximations for optimal capacity and inventory. We give a summary and conclusions in Section 7 and provide most of the proofs in Appendix A.

2. Literature Review

In this paper, we examine fork-join queueing networks with N servers where the arrival and service streams are almost deterministic with a Brownian component. Our goal is to find and investigate the

maximum queue length as N goes to infinity. The queue lengths are dependent random variables due to the joint interarrivals. Thus, our paper is related to the convergence of extreme values (maximum queue lengths) of dependent random variables. An overview of early results on extreme value theory for dependent random variables is given in Leadbetter et al. (1983). The authors provide conditions when the sequence of random variables may be treated as a sequence of independent random variables; this is the case when the covariance of random variables X_i and X_j decreases when i and j are further apart from each other. They also present a convergence result for the joint all-time suprema of a finite number of dependent stationary processes, they prove in Theorem 11.2.3 that, under some assumptions, the joint all-time suprema of a finite number of dependent stationary processes are mutually independent. This is somewhat related to the problem that we study; however, we do not investigate stationary processes and we only look at the largest of the N all-time suprema, where $N \rightarrow \infty$.

We investigate the extreme values for a sequence of N Brownian motions. To be precise, we examine the joint all-time suprema of N dependent Brownian motions with a negative and linear drift term, when N is large. A lot of work has been done on joint suprema of Brownian motions. For instance, Kou et al. (2016) give 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. Dębicki et al. (2020) 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; cf. Dębicki et al. (2015), where the authors give tail asymptotics of the joint suprema of independent Gaussian processes over a finite time interval. These are just three examples, but the literature is rich with variations around assumptions on independence and dependence or around whether or not drift terms are linear, with joint suprema of two or more than two processes, with suprema over finite and infinite time intervals, and with extensions to other Gaussian processes. In this paper, we specifically examine the maximum of N all-time suprema of dependent Brownian motions. In this respect, the work of Brown and Resnick (1977) comes the closest to our work. In that paper, the authors study process convergence of the scaled maximum of N independent Brownian motions to a stationary limiting process whose marginals are Gumbel distributed. However, we add to this by considering the maximum of the all-time suprema of N dependent Brownian motions.

Our work also relates to the literature on fork-join queues. Specifically, we study asymptotic results for a fork-join queueing system with N servers. Most exact results on fork-join queues are limited to systems with two service stations; cf. Flatto and Hahn (1984), Wright (1992), Baccelli (1985) and Klein (1988). For fork-join queues with more than two servers only approximations of performance measures are given; cf. Ko and Serfozo (2004), Baccelli and Makowski (1989) and Nelson and Tantawi (1988). Most of these papers focus on fork-join queueing systems where the number of servers is finite, while we investigate a fork-join queue where N goes to infinity. Furthermore, in these papers, the focus lies on steady-state distributions and other one-dimensional performance measures. Work on the heavy-traffic process limit has also been done. For example, Varma (1990) derives a heavy-traffic analysis for fork-join queues, and shows weak convergence of several processes, such as the joint queue lengths in front of each server. Furthermore, Nguyen (1993) proves that various appearing limiting processes are in fact multi-dimensional reflected Brownian motions. Nguyen (1994) extends this result to a fork-join queue with multiple job types. Lu and Pang study fork-join networks in Lu and Pang (2015, 2017a,b). In Lu and Pang (2015), they investigate a fork-join network where each service station has multiple servers under nonexchangeable synchronization and operates in the quality-driven regime. They derive functional central limit theorems for the number of tasks waiting in the waiting buffers for synchronization and for the number of synchronized jobs. In Lu and Pang (2017a), they extend this analysis to a fork-join network with a fixed number of service stations, each having many servers, where the system operates in the Halfin-Whitt regime. In Lu and Pang (2017b), the authors investigate these heavy-traffic limits for

a fixed number of infinite-server stations, where services are dependent and could be disrupted. Finally, we mention Atar et al. (2012), who investigate the control of a fork-join queue in heavy traffic by using feedback procedures.

Besides the literature on extreme value theory and fork-join queues, our work relates to the supply chain management literature. Simultaneous optimization of capacity and inventory is an important problem in supply chain management, but the literature on this topic is limited due to complexity of the problem (Bradley and Glynn 2002). Sleptchenko et al. (2003) study simultaneous optimization of spare-part inventory and repair capacity. In the last decade, simultaneous optimization of capacity and inventory in a single supplier-manufacturer relationship has been studied increasingly (e.g. Reed and Zhang 2017, Reddy and Kumar 2020). Reed and Zhang (2017) show that the square-root staffing rule of Halfin and Whitt (1981) is a valuable tool in optimizing inventory and capacity in a multi-server make-to-stock queue. Altendorfer and Minner (2011) study simultaneous optimization of inventory and planned lead-time and Mayorga and Ahn (2011) study the joint optimization of inventory and temporarily available additional capacity. Our work differs fundamentally from these studies, as we consider the assembly of multiple components that face the same (stochastic) demand instead of the interaction between a manufacturer and a single supplier.

Brownian motion models are common in the literature on inventory control. Optimal control of inventory that can be described by a Brownian motion is described by Harrison (2013, §7), who provides optimality conditions for both discounted and average cost criteria. Closely related to our work is the Brownian Motion Model presented by Bradley and Glynn (2002, §3) to study the trade-off between capacity and inventory. They provide closed-form approximations to the optimal capacity and base-stock levels in a system with a single item. We consider an assembly system in which multiple components are merged into one end-product. This is an essential difference, since in our model inventory does not only buffer against uncertain demand, but a component may also need to be stored when other components are not yet available.

A review of literature studying inventory control in a multi-supplier setting is provided by Svoboda et al. (2020). However, this mainly concerns multi-sourced items that can be delivered by any of the available suppliers. Masih-Tehrani et al. (2011) add an additional dimension to these multi-sourced systems by considering stochastically dependent manufacturing capacities. They state that disruptions affecting one supplier are likely to have an effect on the other suppliers as well.

Bernstein and DeCroix (2006) and Bollapragada et al. (2004) study base-stock policies in a single-sourced assembly system with multiple suppliers. In these systems, multiple components, each sourced from a single supplier, need to be merged into a final product. Bernstein and DeCroix (2006) investigate the effect of using information on pipeline inventories in a decentralized system. Bollapragada et al. (2004) consider the performance of base-stock policies in case both demand and the supplier's capacity are uncertain. Literature concerning simultaneous optimization of capacity and inventory in single-sourced assembly systems with multiple components is limited. Zou et al. (2004) study how supply chain efficiency can be increased by synchronizing processing times and delivery quantities. Pan and So (2016) consider the simultaneous optimization of component prices and production quantities in a two-supplier setting where one supplier has uncertainty in the yield. Our main contribution compared to the work of Zou et al. (2004) and Pan and So (2016) is that we provide approximations of the optimal capacity and base-stock levels that only require two moments.

3. Problem formulation

We consider a manufacturing system in which a manufacturer assembles a final product from N components, each of which is produced on a single production line, where N is a large number. Random delays may occur in the production process for each of the components. To efficiently satisfy demand of the end-product, which may either be deterministic or stochastic, we need to decide how much capacity to establish for each component and how many finished components to

keep on inventory as a buffer. Even though it is costly to establish capacity and to hold inventory, not being able to satisfy demand gives rise to backorder costs. Therefore, we need to find capacity and inventory levels that minimize total expected costs.

To formulate the cost-minimization problem, we model this assembly system by a fork-join queue. Demand is represented by the arrival stream of jobs going to each server and each server represents a component production line. The backlog of each component is represented by a queue of jobs that have not been served yet. After completion of a job, the finished component is stored in a warehouse. When all servers have a finished component in their warehouse, the end-product can be assembled. This system is visualized in Figure 1.

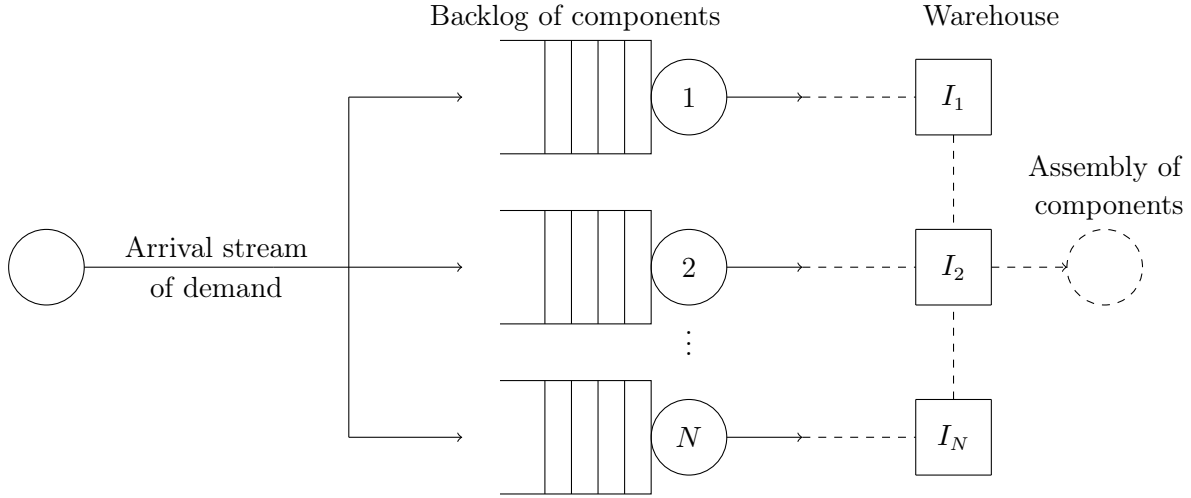


Figure 1 Fork-join queue

To buffer against uncertainties in the supply and demand processes, we introduce a base-stock level I_i for each component $i \leq N$. We define $\beta_i > 0$ as the net capacity for component i , i.e. the difference between the production rate and arrival rate, in other words, β_i captures the capacity investment of server i . $Q_i(\beta_i)$ is the number of outstanding orders of component $i \leq N$. We model this as $Q_i(\beta_i) = \sup_{s>0} (W_i(s) + W_A(s) - \beta_i s)$, where $(W_i, i \leq N)$ are independent Brownian motions with mean 0 and variance σ^2 that represent fluctuations that occur during the production process of component i and where W_A is a Brownian motion with mean 0 and variance σ_A^2 representing the fluctuations in the number of demands. One can see $Q_i(\beta_i)$ as a two-moment or heavy traffic approximation of the steady state queue length in front of server i . If $\sigma_A^2 > 0$, $(Q_i(\beta_i))_{i \leq N}$ are dependent random variables.

We proceed by developing an expression for the total system costs, which requires expressions for the inventory and backorders. The inventory of component i consists of two parts: first, the excess supply that works as a buffer against uncertain demand; second, the committed inventory that consists of items that are committed to realized demand but put aside because other components are not yet available. I.e., the excess supply of component i is given by $(I_i - Q_i(\beta_i))^+$. Moreover, the number of backorders for component i is equal to $(Q_i(\beta_i) - I_i)^+$, since for $Q_i(\beta_i) \leq I_i$ the shortage is compensated by inventory I_i and only the part of $Q_i(\beta_i)$ exceeding I_i represents actual backorders that cannot be satisfied. Since all components need to be available to assemble the final product, the number of backorders in the system is equal to the number of backorders of the component with the largest backlog and is thus given by $\max_{i \leq N} (Q_i(\beta_i) - I_i)^+$. Therefore, the committed inventory of component i equals the number of backorders in the system minus its own backlog and can be

expressed as $\max_{i \leq N} (Q_i(\beta_i) - I_i)^+ - (Q_i(\beta_i) - I_i)^+$. The total inventory of component i is thus given by

$$(I_i - Q_i(\beta_i))^+ + \max_{i \leq N} (Q_i(\beta_i) - I_i)^+ - (Q_i(\beta_i) - I_i)^+ = I_i - Q_i(\beta_i) + \max_{i \leq N} (Q_i(\beta_i) - I_i)^+.$$

We scale the cost of building net capacity to one and let $h^{(N)}$ and $b^{(N)}$ denote holding costs and backorder costs, respectively, which may depend on N . Our goal is to minimize the expected total costs of the system. If we define

$$\begin{aligned} C_N(I, \beta) &= \mathbb{E} \left[\sum_{i \leq N} \left[h^{(N)} \left(I - Q_i(\beta) + \left(\max_{i \leq N} Q_i(\beta) - I \right)^+ \right) \right] + b^{(N)} \left(\max_{i \leq N} Q_i(\beta) - I \right)^+ \right] \\ &= \mathbb{E} \left[N h^{(N)} (I - Q_i(\beta)) + (N h^{(N)} + b^{(N)}) \left(\max_{i \leq N} Q_i(\beta) - I \right)^+ \right], \end{aligned} \quad (1)$$

then, if $\beta_i = \beta$ and $I_i = I$ for given I and β , the expected total costs in the system are equal to $C_N(I, \beta) + \beta N$. In the centralized optimization problem, this expression is minimized with respect to I and β . In Appendix A.1, we show that it suffices to consider symmetric solutions where both I_i and β_i are constant in i when $(Q_i(\beta_i))_{i \leq N}$ are independent random variables or when we minimize over one drift parameter. For these two cases, we exploit the self-similarity property of Brownian motions, which makes it more convenient to simplify $C_N(I, \beta)$. Due to the self-similarity of Brownian motion, we can write

$$\beta \max_{i \leq N} \sup_{s > 0} (W_i(s) - \beta s) = \beta \max_{i \leq N} \sup_{t > 0} \left(W_i \left(\frac{t}{\beta^2} \right) - \beta \frac{t}{\beta^2} \right) \stackrel{d}{=} \max_{i \leq N} \sup_{t > 0} (W_i(t) - t).$$

This means that $\max_{i \leq N} Q_i(\beta) \stackrel{d}{=} \frac{1}{\beta} \max_{i \leq N} Q_i(1)$. Therefore, after rescaling the variable I , we can write

$$\min_{(I, \beta)} \left(C_N(I, \beta) + \beta N \right) = \min_{(I, \beta)} \left(\frac{1}{\beta} C_N(I\beta, 1) + \beta N \right) = \min_{(I, \beta)} \left(\frac{1}{\beta} C_N(I, 1) + \beta N \right). \quad (2)$$

In the last part of Equation (2), I has the interpretation of the base-stock level where the net capacity $\beta = 1$. Therefore, from now on, the actual number of products on stock at time 0 equals I/β . Furthermore, we will simply refer to I as "inventory" hereafter. Similarly, the actual unsatisfied demands of component i equals $Q_i(1)/\beta$ and we write $Q_i = Q_i(1)$. This allows us to write the cost function $F_N(I, \beta)$ to be optimized as given in Definition 3.1.

DEFINITION 3.1. We define

$$F_N(I, \beta) := \frac{1}{\beta} C_N(I) + \beta N, \quad (3)$$

with $C_N(I) := C_N(I, 1)$ and $C_N(I, \beta)$ given in Equation (1).

Our goal is to solve $\min_{(I, \beta)} F_N(I, \beta)$, focusing on the case where N is large. Before we focus on this regime, we first derive some additional properties of this problem, which are valid for each N . In the next lemma, we show that we can write this minimization problem as two separate minimization problems.

LEMMA 3.2. Let $(b^{(N)})_{N \geq 1}, (h^{(N)})_{N \geq 1}$ be sequences such that $h^{(N)} > 0$ and $b^{(N)} > 0$ for all N . Let (I_N, β_N) minimize $F_N(I, \beta)$. Then the optimal inventory I_N minimizes $C_N(I)$ and the optimal β_N minimizes $\frac{1}{\beta} C_N(I_N) + \beta N$. Furthermore, the function $C_N(I)$ is convex with respect to I , and the function $\frac{1}{\beta} C_N(I) + \beta N$ is convex with respect to β .

The proofs of this section can be found in Appendix A.1.

Using Lemma 3.2, we can characterize the optimal net capacity and inventory decisions. In Lemma 3.3 we provide expressions for the optimal net capacity and costs in terms of the optimal inventory decision, which is given in Lemma 3.4.

LEMMA 3.3. Given $I_N^* = \arg \min_I C_N(I)$, minimizing $F_N(I, \beta)$ with respect to β yields $\beta_N^* = \sqrt{\frac{C_N(I_N^*)}{N}}$. Furthermore, the corresponding costs are $F_N(I_N^*, \beta_N^*) = 2N\beta_N^* = 2\sqrt{C_N(I_N^*)N}$.

The optimal value of I can be expressed as a quantile of the distribution of $\max_i Q_i$:

LEMMA 3.4. I_N^* is the unique solution of

$$\mathbb{P}\left(\max_{i \leq N} Q_i \leq I_N^*\right) = \frac{b^{(N)}}{Nh^{(N)} + b^{(N)}}.$$

The main technical issue is that the distribution of this maximum is in general not very tractable, especially when N is large. The main theme of our work is to consider approximations of this distribution using extreme value theory, to analyze their quality if N is large.

To explain our ideas, we mention the following first-order approximation of $\max_{i \leq N} Q_i$:

LEMMA 3.5. $\max_{i \leq N} Q_i$ satisfies the first-order approximation

$$\frac{\max_{i \leq N} Q_i}{\log N} \xrightarrow{L_1} \frac{\sigma^2}{2},$$

as $N \rightarrow \infty$.

The lemma easily follows from more refined results that are proven later on in this paper.

This first-order approximation is valid regardless whether $\sigma_A = 0$ or $\sigma_A > 0$. In the subsequent two sections, we consider more refined extreme-value theory approximations covering both cases. It turns out that the second-order behavior of the maximum is qualitatively different when σ_A becomes strictly positive. This has, in turn, an impact on the structure of the optimal solution of our cost minimization problem when N grows large.

To better understand this structure, we heuristically analyze the first-order approximation of the cost minimization problem and apply it to approximate I_N^* and β_N^* . First, we use the approximation $\max_{i \leq N} Q_i \approx \frac{\sigma^2}{2} \log N$ to write

$$C_N(I) \approx \bar{C}_N(I) = Nh^{(N)} \left(I - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \left(\frac{\sigma^2}{2} \log N - I \right)^+.$$

The optimal value \bar{I}_N for the associated first-order minimization problem $\min_I \bar{C}_N(I)$ is given by $\bar{I}_N = \frac{\sigma^2}{2} \log N$, since $b^{(N)} > 0$. Using this approximation, we see that $C_N(\bar{I}_N) \approx \bar{C}_N(\bar{I}_N) = (1 + o(1)) \frac{\sigma^2}{2} Nh^{(N)} \log N$, $\bar{\beta}_N = \sqrt{\bar{C}_N(\bar{I}_N)/N} = (1 + o(1)) \sqrt{\frac{\sigma^2}{2} h^{(N)} \log N}$, and $F_N(\bar{I}_N, \bar{\beta}_N) \approx 2\sqrt{N} \sqrt{\frac{\sigma^2}{2} Nh^{(N)} \log N}$. These results can be made rigorous and the decision rule \bar{I}_N can be shown to be asymptotically optimal, i.e. that $F_N(\bar{I}_N, \bar{\beta}_N) = F_N(I_N^*, \beta_N^*)(1 + o(1))$. To prove this, we need to specify how the cost parameters $h^{(N)}$ and $b^{(N)}$ scale with N . For this, we consider three regimes. These regimes relate to the quantile $b^{(N)}/(Nh^{(N)} + b^{(N)})$ of $\max_i Q_i$ at which I_N^* attains its optimal solution. Assume that $b^{(N)}/(Nh^{(N)} + b^{(N)})$ converges to a constant $1 - \gamma$. We classify the three regimes in a similar way as is done in the analysis of large call centers; cf. Borst et al. (2004):

- We are in the *balanced regime* if $\gamma \in (0, 1)$.
- If $\gamma = 0$, for large systems, the inventory is always sufficiently high to ensure that the manufacturer can assemble the end-product. We call this the *quality-driven regime*.

• Finally, if $\gamma = 1$, inventories are much lower, and we call this the *efficiency-driven regime*. When we are in the balanced or efficiency-driven regime we can prove how far the costs under the first order approximation are from the real optimal costs. This is established in Lemma 3.6:

LEMMA 3.6. Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, with $\gamma_N = \gamma \in (0, 1)$ or $\gamma_N \xrightarrow{N \rightarrow \infty} 1$. Then

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\tilde{I}_N, \tilde{\beta}_N)} = 1 - o(1).$$

In the next two sections, we carry out a more elaborate program using more refined extreme value estimates of $\max_{i \leq N} Q_i$. This analysis gives sharper order bounds than those given in Lemma 3.6. In particular, in the following sections we consider the minimization in two distinct cases. First, in Section 4, we look at the case where demand is assumed to be deterministic, such that $W_A = 0$. Thereafter, in Section 5, we consider the stochastic demand case. In the former case, we utilize existing results in extreme value theory, while the latter case requires the development of a novel limit theorem. Furthermore, we use the result given in Corollary 3.7; this corollary shows how the ratio between the optimal costs and approximate costs can be represented, when the approximate inventory and net capacity are a solution to a minimization problem as well. This corollary follows trivially from Lemma 3.3.

COROLLARY 3.7. Assume we have a function $\tilde{F}_N(I, \beta) : (0, \infty) \times (0, \infty) \rightarrow \mathbb{R}$. Furthermore, assume that the function \tilde{F}_N has the form

$$\tilde{F}_N(I, \beta) = \frac{1}{\beta} \tilde{C}_N(I) + \beta N,$$

where \tilde{C}_N is a positive function with domain $(0, \infty)$. Moreover, assume that the minimum value $\tilde{F}_N(\tilde{I}_N, \tilde{\beta}_N) = 2N\tilde{\beta}_N = 2\sqrt{\tilde{C}_N(\tilde{I}_N)N}$, where \tilde{I}_N and $\tilde{\beta}_N$ are minimizers, then

$$\frac{F(I_N^*, \beta_N^*)}{F(\tilde{I}_N, \tilde{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\tilde{C}_N(\tilde{I}_N)}}{C_N(\tilde{I}_N) + \tilde{C}_N(\tilde{I}_N)}.$$

4. The basic model: deterministic arrival stream

4.1. Solution and convergence of the minimization problem

We now analyze the minimization of the cost function described in Definition 3.1 for the special case with $W_A = 0$ representing deterministic demand. Although we can simplify the minimization problem significantly, by using the self-similarity of Brownian motions and by writing the minimization problem as two separate minimization problems, as shown in Lemma 3.2, the function F_N still has a difficult form, since we have the expression $\max_{i \leq N} Q_i$ in this function. In Lemma 4.1 we give the optimal inventory in order to minimize costs. We assume that the holding and backlog costs $h^{(N)}$ and $b^{(N)}$ are positive sequences, and we distinguish three cases. First of all, we consider the balanced regime $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)}) = \gamma \in (0, 1)$ for all $N > 0$. Secondly, we consider the quality driven regime, where $\gamma_N \xrightarrow{N \rightarrow \infty} 0$. Finally, we investigate the efficiency driven regime, where $\gamma_N \xrightarrow{N \rightarrow \infty} 1$. All proofs for this section can be found in Appendix A.2. We present numerical results for the three regimes in Section 4.2.

LEMMA 4.1. Let $Q_i = \sup_{s > 0} (W_i(s) - s)$, with $(W_i, 1 \leq i \leq N)$ independent Brownian motions with mean 0 and variance σ^2 . Let $h^{(N)}$ and $b^{(N)}$ be positive sequences. In order to minimize $F_N(I, \beta)$, the optimal inventory I_N^* satisfies,

$$I_N^* = P_N^{-1}(1 - \gamma_N) = \frac{\sigma^2}{2} \log \left(\frac{1}{1 - (1 - \gamma_N)^{\frac{1}{N}}} \right), \quad (4)$$

with P_N^{-1} the quantile function of $\mathbb{P}(\max_{i \leq N} Q_i < x)$ and $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$.

To get a better understanding of the limiting behavior of the solution to $\min_{(I,\beta)} F_N(I,\beta)$, we would like to approximate the function F_N . Since $(Q_i, i \leq N)$ are independent and exponentially distributed, we know by standard extreme value theory (cf. De Haan and Ferreira (2006)) that $\frac{\sigma^2}{2} \max_{i \leq N} Q_i - \log N \xrightarrow{d} G$, as $N \rightarrow \infty$, with $G \sim \text{Gumbel}$. Therefore, for N large, $\max_{i \leq N} Q_i \stackrel{d}{\approx} \frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N$. We get a new minimization problem when we replace $\max_{i \leq N} Q_i$ with this approximation $\frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N$. In Definition 4.2 we give the resulting function $\hat{F}_N(I,\beta)$ that is to be minimized.

DEFINITION 4.2.

$$\hat{C}_N(I) := \mathbb{E} \left[Nh^{(N)}(I - Q_i) + \left(Nh^{(N)} + b^{(N)} \right) \left(\frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N - I \right)^+ \right], \quad (5)$$

and

$$\hat{F}_N(I,\beta) := \frac{1}{\beta} \hat{C}_N(I) + \beta N. \quad (6)$$

In the remainder of this section, we investigate whether minimizing $\hat{F}_N(I,\beta)$ results in costs that are close to those when we minimize $F_N(I,\beta)$. Note that we write (I_N^*, β_N^*) for the minimizers for the cost function F_N defined in Definition 3.1, and we write $(\hat{I}_N, \hat{\beta}_N)$ for the minimizers for the cost function \hat{F}_N defined in Definition 4.2. Thus, throughout this paper, we indicate second-order approximations by the \wedge -symbol.

In Proposition 4.3, we present the inventory that minimizes \hat{F}_N . This inventory turns out to be a quantile of $\frac{\sigma^2}{2} G$ added to $\frac{\sigma^2}{2} \log N$.

PROPOSITION 4.3 (APPROXIMATION). Minimizing $\hat{F}_N(I,\beta)$ with $G \sim \text{Gumbel}$, gives solution $(\hat{I}_N, \hat{\beta}_N, \hat{F}_N(\hat{I}_N, \hat{\beta}_N))$, with

$$\hat{I}_N = \frac{\sigma^2}{2} \log N - \frac{\sigma^2}{2} \log(-\log(1 - \gamma_N)), \quad (7)$$

and

$$\hat{C}_N(\hat{I}_N) = Nh^{(N)} \left(\hat{I}_N - \frac{\sigma^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \left(\int_{-\log(1-\gamma_N)}^{\infty} \frac{e^{-t}}{t} dt + \Gamma + \log(-\log(1 - \gamma_N)) \right), \quad (8)$$

where $\Gamma \approx 0.577$ is Euler's constant and $\gamma_N = Nh^{(N)} / (Nh^{(N)} + b^{(N)})$.

Combining Equations (7) and (8) with the results in Lemma 3.3 gives the solution $(\hat{I}_N, \hat{\beta}_N, \hat{F}_N(\hat{I}_N, \hat{\beta}_N))$.

We compare the costs under the optimal inventory and net capacity with the costs under the approximate inventory and net capacity. We distinguish the balanced regime, quality driven regime and efficiency driven regime. We first present two lemmas that are needed to prove order bounds between the costs under the optimal inventory and net capacity, and the costs under the approximate inventory and net capacity. In Lemma 4.6 we show that we can define a random variable that follows a Gumbel distribution, and is on the same probability space as $\max_{i \leq N} Q_i$. In Lemma 4.7 we present bounds on $|C_N(I_N^*) - C_N(\hat{I}_N)|$ and $|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|$. Finally, by using the results from Lemmas 4.6 and 4.7, we prove the order bounds in the balanced, quality driven and efficiency driven regime in Theorem 4.4. In the efficiency driven regime, we impose the additional condition $\gamma_N < 1 - \exp(-N)$ needed to make sure that $\hat{I}_N > 0$.

THEOREM 4.4 (ORDER BOUNDS). Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, if $\gamma_N = \gamma \in (0, 1)$, in the *balanced regime*, then

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(1/(N \log N)), \quad (9)$$

if $\gamma_N \xrightarrow{N \rightarrow \infty} 0$, in the *quality driven regime*, then

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(\gamma_N/(N \log(N/\gamma_N))), \quad (10)$$

and if $\gamma_N \xrightarrow{N \rightarrow \infty} 1$ and $\gamma_N < 1 - \exp(-N)$, in the *efficiency driven regime*, then

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(1/\log N). \quad (11)$$

Using the order bounds given in Theorem 4.4, we can establish for the three different regimes how $F_N(I_N^*, \beta_N^*)$ scales with N as N becomes large.

LEMMA 4.5. Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, if $\gamma_N = \gamma \in (0, 1)$ in the *balanced regime*, then

$$\begin{aligned} & F_N(I_N^*, \beta_N^*) \\ &= 2\sqrt{N} \sqrt{Nh^{(N)} \frac{\sigma^2}{2} (\log N - \log(-\log(1 - \gamma)) - 1) + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \mathbb{E}[(G + \log(-\log(1 - \gamma)))^+]} \\ &+ O(\sqrt{h^{(N)}}/\sqrt{\log N}), \end{aligned} \quad (12)$$

if $\gamma_N \xrightarrow{N \rightarrow \infty} 0$ in the *quality driven regime*, then

$$\begin{aligned} F_N(I_N^*, \beta_N^*) &= 2\sqrt{N} \sqrt{Nh^{(N)} \frac{\sigma^2}{2} (\log(N/\gamma_N) - 1) + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \gamma_N} \\ &+ O(\gamma_N \sqrt{h^{(N)}}/\sqrt{\log(N/\gamma_N)}), \end{aligned} \quad (13)$$

and if $\gamma_N \xrightarrow{N \rightarrow \infty} 1$ and $\gamma_N < 1 - \exp(-N)$ in the *efficiency driven regime*, then

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N} \sqrt{Nh^{(N)} \frac{\sigma^2}{2} (\log N - 1) + b^{(N)} \frac{\sigma^2}{2} \log(-\log(1 - \gamma_N))} + O(N\sqrt{h^{(N)}}/\sqrt{\log N}). \quad (14)$$

The results given in Theorem 4.4 and Lemma 4.5 are obtained by using the properties stated in Lemmas 4.6 and 4.7. In Lemma 4.6 we show that we can write a Gumbel distributed random variable that is on the same probability space as $\max_{i \leq N} Q_i$. This gives us a very powerful result; namely that $\max_{i \leq N} Q_i$ and G_N are ordered and that their difference decreases as $\max_{i \leq N} Q_i$ becomes large. Consequently, we obtain very sharp bounds in Lemma 4.7 which leads to sharp results in Theorem 4.4 and Lemma 4.5.

LEMMA 4.6. Define

$$G_N := -\log \left(-\log \left(\left(1 - \exp \left(-\frac{2}{\sigma^2} \max_{i \leq N} Q_i \right) \right)^N \right) \right), \quad (15)$$

then $\mathbb{P}(G_N < x) = e^{-e^{-x}}$, for all N . Moreover,

$$\max_{i \leq N} Q_i > \frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N, \quad (16)$$

and $\max_{i \leq N} Q_i - \frac{\sigma^2}{2} G_N - \frac{\sigma^2}{2} \log N$ strictly decreases as a function of $\max_{i \leq N} Q_i$ with limit 0.

LEMMA 4.7. Let $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, then

$$\left| C_N(I_N^*) - C_N(\hat{I}_N) \right| \leq (I_N^* - \hat{I}_N)(Nh^{(N)} + b^{(N)}) \left(1 - \gamma_N - \left(1 + \frac{\log(1 - \gamma_N)}{N} \right)^N \right), \quad (17)$$

$$\left| \hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N) \right| \leq (I_N^* - \hat{I}_N)Nh^{(N)} \left(1 - \left(1 + \frac{\log(1 - \gamma_N)}{N} \right)^N \right). \quad (18)$$

4.2. Numerical experiments

We now provide some numerical results to illustrate the solutions to the minimization problem and their characteristics discussed in Section 4.1. In all experiments, we let $\sigma = 1$ and let N vary from 10 to 1000. The results for the balanced regime, quality driven regime and efficiency driven regime are given in Tables 1, 2 and 3, respectively. We can observe that in all regimes the approximate solutions are close to the optimal solutions. Most importantly, already for small N , the fraction of the costs corresponding to the optimal solution over the costs corresponding to the approximate solution nearly equals 1.

Table 1 Balanced Regime, $h^{(N)} = 1, b^{(N)} = N$ such that $\gamma_N = \frac{1}{2}$.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	\hat{I}_N	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} \right) N \log N$
10	1.35178	1.19648	23.9315	1.33455	1.19328	23.9315	0.001807
50	2.14273	1.49338	149.338	2.13927	1.49286	149.338	0.000379
100	2.48757	1.60499	320.997	2.48584	1.60475	320.997	0.000192
200	2.83328	1.70944	683.775	2.83242	1.70932	683.775	$9.68 \cdot 10^{-5}$
500	3.29091	1.8385	1838.5	3.29056	1.83846	1838.5	$3.91 \cdot 10^{-5}$
1000	3.63731	1.93044	3860.87	3.63713	1.93042	3860.87	$1.97 \cdot 10^{-5}$

Table 2 Quality Driven Regime, $h^{(N)} = 1, b^{(N)} = N^2$ such that $\gamma_N = \frac{1}{1+N}$.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	\hat{I}_N	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} \right) \frac{N}{\gamma_N} \log \frac{N}{\gamma_N}$
10	2.32898	1.52962	30.5925	2.3266	1.52924	30.5925	0.000617
50	3.91708	1.97978	197.978	3.91698	1.97976	197.978	$2.52 \cdot 10^{-5}$
100	4.60768	2.14684	429.368	4.60766	2.14684	429.368	$6.31162 \cdot 10^{-6}$
200	5.29957	2.30221	920.886	5.29956	2.30221	920.886	$1.21801 \cdot 10^{-6}$
500	6.21511	2.49306	2493.06	6.21511	2.49306	2493.06	$5.51467 \cdot 10^{-6}$
1000	6.90801	2.62833	5256.66	6.90801	2.62833	5256.66	0.000176

Table 3 Efficiency Driven Regime, $h^{(N)} = N, b^{(N)} = 1$ such that $\gamma_N = \frac{N^2}{N^2+1}$.

N	I_N^*	β_N^*	$F_N(I_N^*, \beta_N^*)$	\hat{I}_N	$\hat{\beta}_N$	$F_N(\hat{I}_N, \hat{\beta}_N)$	$\left(1 - \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} \right) \log N$
10	0.497572	3.12224	62.4448	0.386624	3.08439	62.4665	0.000797
50	0.965997	9.35451	935.451	0.927385	9.34122	935.453	$8.65678 \cdot 10^{-6}$
100	1.21527	14.4701	2894.02	1.19242	14.4615	2894.02	$1.30518 \cdot 10^{-6}$
200	1.48208	22.0864	8834.57	1.46889	22.0808	8834.57	$2.20863 \cdot 10^{-7}$
500	1.85348	38.0553	38055.3	1.84728	38.0521	38055.3	$2.51171 \cdot 10^{-8}$
1000	2.14443	56.945	113890	2.14098	56.9428	113890	$5.30189 \cdot 10^{-9}$

5. Stochastic Demand

We now extend our framework to the case where demand is stochastic. This means that stochasticity not only arises from the production process of the individual components, but also results from uncertain demands. Consequently, delays may no longer only be caused by low production of a specific component, but may also occur when there is a sudden peak in demand. Since all components need to be available to assemble the end-product and satisfy demand, delays of the different components are now correlated. We use the same strategy when demand is stochastic as in the basic model with deterministic demand. However, we can no longer approximate the maximum queue length distribution with the Gumbel distribution. In Section 5.1 we show that for N large, $\max_{i \leq N} Q_i \approx \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X$ with X a standard normal random variable. Using this approximation, we obtain a new minimization problem, in which we minimize $\hat{F}_N^A(I, \beta)$ as given in Definition 5.1 with respect to I and β .

DEFINITION 5.1.

$$\hat{C}_N^A(I) = \mathbb{E} \left[N h^{(N)}(I - Q_i) + \left(N h^{(N)} + b^{(N)} \right) \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X - I \right)^+ \right],$$

and

$$\hat{F}_N^A(I, \beta) = \frac{1}{\beta} \hat{C}_N^A(I) + \beta N.$$

In Section 5.2 we elaborate on the solution and convergence of the minimization problem.

5.1. Extreme value limit

In this section, we focus on the maximum of N dependent random variables. In Theorem 5.2 we prove that a scaled version of $\max_{i \leq N} Q_i(\beta)$ converges in distribution to a normally distributed random variable, as N goes to infinity.

THEOREM 5.2. Let $(W_i, 1 \leq i \leq N)$ be independent Brownian motions with mean 0 and variance σ^2 , and W_A be a Brownian motion with mean 0 and variance σ_A^2 . Then

$$\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma \sigma_A}{\sqrt{2}\beta} X, \quad (19)$$

with $X \sim \mathcal{N}(0, 1)$. In other words, for all $x \in \mathbb{R}$

$$\mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} > x \right) \xrightarrow{N \rightarrow \infty} 1 - \Phi \left(\frac{x \sqrt{2}\beta}{\sigma \sigma_A} \right),$$

with Φ the cumulative distribution function of a standard normal random variable.

A heuristic explanation of the result in Theorem 5.2 is as follows: though $(Q_i, i \leq N)$ are dependent random variables, since we are adding the same Brownian motion W_A , $\max_{i \leq N} W_i(s)$ will dominate more and more over W_A as N becomes larger. Consequently, W_A does not affect the time at which the supremum of $\max_{i \leq N} W_i(s) + W_A(s) - \beta s$ is attained. Hence, for N large $\max_{i \leq N} Q_i(\beta) \approx \max_{i \leq N} \sup_{s > 0} (W_i(s) - \beta s) + W_A(\tau)$, with τ the hitting time of the supremum of $\max_{i \leq N} (W_i(s) - \beta s)$. Based on theory on conditional expectations of Lévy processes we know that the conditional expectation of the hitting time $\tau(x)$ to reach a point x is linear with x , to be precise, for $N = 1$, it is known that $\mathbb{E}[\tau(x) | \tau(x) < \infty] = x/\beta$. Combining this with the fact that $\max_{i \leq N} \sup_{s > 0} (W_i(s) - \beta s) \sim \frac{\sigma^2}{2\beta} \log N$, we expect that the supremum of $\max_{i \leq N} (W_i(s) - \beta s)$ is reached at $\tau \approx \frac{1}{\beta} \cdot \frac{\sigma^2}{2\beta} \log N =$

$\frac{\sigma^2}{2\beta^2} \log N$. Therefore, $W_A(\tau) \approx \frac{\sigma\sigma_A}{\sqrt{2\beta}} \sqrt{\log N} X$, with X standard normally distributed, which results in Equation (19).

The proof of Theorem 5.2 consists of four parts, which are stated in Lemmas 5.3, 5.4, 5.5 and 5.6 for which the proofs are provided in Appendix A.3. For a process X we have for all $t > 0$ that

$$\mathbb{P}\left(\sup_{s>0} X(s) > x\right) \geq \mathbb{P}(X(t) > x).$$

Furthermore, for every $0 < t_1 < t_2$,

$$\mathbb{P}\left(\sup_{s>0} X(s) > x\right) \leq \mathbb{P}\left(\sup_{0<s<t_1} X(s) > x\right) + \mathbb{P}\left(\sup_{t_1\leq s<t_2} X(s) > x\right) + \mathbb{P}\left(\sup_{s\geq t_2} X(s) > x\right).$$

We prove that these lower and upper bounds are tight for the process given in Theorem 5.2 for appropriately chosen t, t_1, t_2 . More specifically, in Lemma 5.3 we prove the asymptotic behavior at the critical time $d \log N$ where $d = \frac{\sigma^2}{2\beta^2}$, resulting in the tight lower bound. We show that times before and after this critical time have no influence in Lemmas 5.4 and 5.5, respectively, leading up to Lemma 5.6 that shows the concentration around the critical time $d \log N$, proving a tight upper bound.

LEMMA 5.3. For $d = \frac{\sigma^2}{2\beta^2}$,

$$\frac{\max_{i\leq N}(W_i(d \log N) + W_A(d \log N)) - \beta d \log N - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma\sigma_A}{\sqrt{2\beta}} X, \quad (20)$$

with $X \sim \mathcal{N}(0, 1)$, as $N \rightarrow \infty$.

LEMMA 5.4. For $d = \frac{\sigma^2}{2\beta^2}$ and $0 < \epsilon < d$, and for all x ,

$$\mathbb{P}\left(\frac{\max_{i\leq N} \sup_{0<s<(d-\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x\right) \xrightarrow{N\rightarrow\infty} 0. \quad (21)$$

LEMMA 5.5. For $d = \frac{\sigma^2}{2\beta^2}$ and all $\epsilon > 0$, and $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{\max_{i\leq N} \sup_{s\geq(d+\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x\right) \xrightarrow{N\rightarrow\infty} 0. \quad (22)$$

LEMMA 5.6. For $d = \frac{\sigma^2}{2\beta^2}$ and $\epsilon > 0$ and for all x ,

$$\begin{aligned} & \limsup_{N\rightarrow\infty} \mathbb{P}\left(\frac{\max_{i\leq N} \sup_{(d-\epsilon)\log N \leq s \leq (d+\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x\right) \\ & \leq \mathbb{P}\left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2}} - \epsilon X_1 + \sqrt{2\epsilon}\sigma_A |X_2| > x\right), \end{aligned} \quad (23)$$

with $X_1, X_2 \sim \mathcal{N}(0, 1)$ and independent.

In Appendix A.3 we show how these lemmas can be used to prove Theorem 5.2. In Lemma 5.7, we prove that convergence holds even in L_1 , when X is chosen appropriately.

LEMMA 5.7. Define $X_N := \frac{\sqrt{2\beta}}{\sigma\sigma_A} \frac{W_A\left(\frac{\sigma^2}{2\beta^2} \log N\right)}{\sqrt{\log N}}$. Then,

$$\mathbb{E} \left[\left| \frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2\beta}} X_N \right| \right] \xrightarrow{N \rightarrow \infty} 0.$$

The proof of Lemma 5.7 is also given in Appendix A.3. In the next section, we apply Theorem 5.2 and Lemma 5.7 to solve and approximate the minimization problem. Specifically, Lemma 5.7 gives us an order bound between the optimal inventory and the approximate inventory.

5.2. Solution and Convergence of the Minimization Problem

We can use the convergence result proven in Theorem 5.2 to prove asymptotics of the minimization of the function F_N . Since $\frac{\sqrt{2\beta}}{\sigma\sigma_A} \frac{\max_{i \leq N} Q_i(\beta) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}}$ is a continuous random variable, we know that its quantile function converges to the quantile function of a standard normal random variable; cf. Van der Vaart (1998, p. 305, Lem. 21.2). So we can use this to derive asymptotics of the minimization problem of F_N .

Using $P_N^A(z)$ as described in Definition 5.8, we can solve the minimization problem, which yields the optimal inventory and net capacity given in Lemma 5.9. The proofs concerning the solution and subsequent convergence results are provided in Appendix A.4.

DEFINITION 5.8. We define

$$P_N^A(z) = \mathbb{P} \left(\frac{\sqrt{2} \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sigma\sigma_A \sqrt{\log N}} \leq z \right).$$

LEMMA 5.9. Let $(b^{(N)})_{N \geq 1}, (h^{(N)})_{N \geq 1}$ be sequences such that $h^{(N)} > 0$ and $b^{(N)} > 0$ for all N , and $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$. Let (β_N^A, I_N^A) minimize $F_N(I, \beta)$. Then

$$I_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} P_N^{A-1}(1 - \gamma_N) \sqrt{\log N}. \quad (24)$$

When we are in the balanced regime, we can approximate the minimization problem given in Definition 5.1, using the convergence result in Theorem 5.2, and prove how far the approximate solution is from the optimal solution. This is done in Proposition 5.10 and Theorem 5.11. In Lemma 5.12 we show how the optimal costs scale with N when we are in the balanced regime. The proofs are given in Appendix A.4.

PROPOSITION 5.10. For $(b^{(N)})_{N \geq 1}, (h^{(N)})_{N \geq 1}$ and $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$,

$$\hat{I}_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma_N), \quad (25)$$

and

$$\hat{C}_N^A(\hat{I}_N^A) = Nh^{(N)} \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \frac{\sigma\sigma_A \sqrt{\log N} e^{-\frac{1}{2}\Phi^{-1}(1-\gamma_N)^2}}{2\sqrt{\pi}}. \quad (26)$$

THEOREM 5.11 (ORDER BOUND). Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, with $\gamma_N = \gamma \in (0, 1)$. Then

$$\left| \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)} - 1 \right| = o \left(\frac{1}{\sqrt{\log N}} \right).$$

LEMMA 5.12 (BALANCED REGIME). Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, with $\gamma_N = \gamma \in (0, 1)$. Then

$$I_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma) + o(\sqrt{\log N}), \quad (27)$$

and

$$F_N(I_N^A, \beta_N^A) = 2\sqrt{N} \sqrt{\hat{C}_N^A(\hat{I}_N^A)} + o(N\sqrt{h^{(N)}}). \quad (28)$$

The result in Lemma 5.12 only holds for the balanced regime, so a natural question is what we can say about the efficiency and the quality driven regime. As is shown in Lemma 3.6, in the efficiency driven regime, the first order approximation $\bar{I}_N = \frac{\sigma^2}{2} \log N$ gives that the ratio of the approximate costs and the optimal costs converge to 1. Thus we expect the approximation given in (25) will also satisfy this convergence result. In order to determine whether this approximation also satisfies the order bound given in Theorem 5.11, a further analysis is needed. The analysis we provide for the balanced regime heavily relies on Van der Vaart (1998, p. 305, Lem. 21.2), which says that if $Y_N \xrightarrow{d} Y$, then for $\gamma \in (0, 1)$, $P_{Y_N}^{-1}(\gamma) \xrightarrow{N \rightarrow \infty} P_Y^{-1}(\gamma)$. This gives us the convergence result (27) of the inventory in the balanced regime. In order to be able to prove a similar result for the efficiency driven regime, we should need an improvement of Van der Vaart (1998, p. 305, Lem. 21.2) which also holds when $\gamma_N \xrightarrow{N \rightarrow \infty} 1$.

However, for the quality driven regime, this convergence result does not hold, because we see in Lemma 4.5 that $I_N^A \approx \frac{\sigma^2}{2} \log(N/\gamma_N)$. In order to find a sharp order bound such as given in Theorem 5.11 we should resort to the analysis of tail asymptotics, which is beyond the scope of this study.

5.3. Numerical Experiments

In Section 5.2, we provided expressions to calculate the asymptotically optimal net capacity and inventory. The question remains how large the number of components has to be for these approximations to be of use. Therefore, we now examine the expected costs under both the optimal net capacity and inventory and under these asymptotic approximations. Since it is not straightforward to calculate $\mathbb{E}[(\max_{i \leq N} Q_i - I)^+]$ for dependent Q_i , to evaluate the cost function given in Definition 3.1 we resort to simulation. First, we explain the details of our simulation experiment, after which we discuss the numerical results.

In our simulation, we aim to determine the maximum delay over all components, so $\max_{i \leq N} Q_i$. For this, we use the algorithm proposed by Asmussen et al. (1995, §4.5), who describe an exact algorithm for simulating a reflected Brownian motion at the grid points. At every grid point, we draw normal random variables with the required drift and variance for the supply and demand processes and update the maximum. We use a step size of 0.001 for the grid points. Since we cannot simulate over an infinite horizon, we have to determine when to terminate the simulation. The maximum value is expected to be attained at a time which is smaller than $\hat{t} = \frac{\sigma^2 + \sigma_A^2}{2} \sum_{j=1}^N \frac{1}{j}$. To simulate well beyond this point, we run the simulation until $t = 2\hat{t}$.

Using the above method to simulate $\max_{i \leq N} Q_i$, we can estimate $P_N^{A^{-1}}(1 - \gamma_N)$ with $P_N^A(z)$ as described in Definition 5.8. To obtain a median-unbiased estimate of the quantile, we use the approach suggested by Zieliński (2009, p. 982-983). For this, we sample $\max_{i \leq N} Q_i$ 100 times and randomly choose between the observations $(1 - \gamma_N) \cdot 100$ and $(1 - \gamma_N) \cdot 100 + 1$, with weights depending on the value of the fractile. Our estimate is equal to the median over 100 iterations. Once we have our estimate of $P_N^{A^{-1}}(1 - \gamma_N)$, we determine the value of the optimal inventory as given in Equation (24). Using the optimal inventory we determine the optimal net capacity given in Lemma 3.3. Since this also requires the expectation of $(\max_{i \leq N} Q_i - I)^+$, we determine this value by taking the average based on 10,000 simulations.

Next, we compare the costs under our asymptotic approximations of the net capacity and inventory (provided in Proposition 5.10) to the costs under the optimal net capacity and inventory obtained from the simulation. We again sample $(\max_{i \leq N} Q_i - I)^+$ based on 10,000 new simulations and determine the costs of the different policies using cost function $F_N(I, \beta)$.

In order to assess the performance of the approximations and its sensitivity to various model parameters, we perform a full factorial experiment. In our experiment, we vary the number of components, demand variability and backorder costs. The setup of the experiment is given in Table 4. We set $h^{(N)} = 1$ and $\sigma = 1$ in all experiments. In total we have 24 instances. The results are given in Tables 5 and 6 for $b^{(N)} = N$ and $b^{(N)} = 3N$, respectively.

Table 4 Parameter settings for experiments

Parameter	Values
N	10, 50, 100
σ_A	0.1, 0.5, 0.75, 1
$b^{(N)}$	$N, 3N$

Table 5 Comparison of costs approximate solution for $h^{(N)} = 1, b^{(N)} = N$

N	σ_A	I_N^A	β_N^A	$F_N(I_N^A, \beta_N^A)$	\hat{I}_N^A	$\hat{\beta}_N^A$	$F_N(\hat{I}_N^A, \hat{\beta}_N^A)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right) \sqrt{\log N}$
10	0.1	1.327	1.1583	23.1894	1.151	0.855514	24.5143	0.0820
50	0.1	2.122	1.47611	147.534	1.956	1.25004	150.337	0.0369
100	0.1	2.455	1.58865	318.588	2.303	1.38516	322.994	0.0293
10	0.5	1.486	1.25448	25.333	1.151	0.976909	26.9363	0.0903
50	0.5	2.338	1.59412	159.934	1.956	1.3744	164.689	0.0571
100	0.5	2.715	1.71664	343.937	2.303	1.51094	352.91	0.0546
10	0.75	1.714	1.36908	27.191	1.151	1.00605	29.7614	0.1311
50	0.75	2.638	1.70591	171.443	1.956	1.41834	180.556	0.0998
100	0.75	2.980	1.83438	367.348	2.303	1.55865	383.319	0.0894
10	1	1.990	1.47358	29.8393	1.151	1.0037	34.6552	0.2109
50	1	3.006	1.84276	185.25	1.956	1.43941	201.314	0.1578
100	1	3.394	1.97602	393.668	2.303	1.58534	421.505	0.1417

There are several important observations to be made from Table 5. First of all, we can observe that for $N = 10$ the difference in costs between the simulated optimal solution and the asymptotic solution is around 10% for most cases, the case $N = 10$ and $\sigma_A = 1$ is an outlier, where the difference is around 15%. As N increases to 50, the difference decreases. Furthermore, the difference becomes larger when σ increases. In the last column, we verify the convergence result from Theorem 5.11. We observe that the difference decreases as N increases, and that increasing σ_A causes the difference to increase.

When we consider the results for $b^{(N)} = 3N$ given in Table 6, we observe that the difference between the asymptotic and optimal costs is considerably higher than for $b^{(N)} = N$. Especially for $N = 10$, the difference is around 15% of the optimum, except for $N = 10$ and $\sigma_A = 0.1$, where the difference is around 20%. However, for a larger number of components, the difference is around 10% of the optimum. Interestingly, for the case $\sigma_A = 1$, the difference between $b^{(N)} = N$ and $b^{(N)} = 3N$ is relatively small.

Overall, in most of our experiments the difference between the costs under the optimal inventory and net capacity and the costs under the approximations are around 10%. Furthermore, we can

Table 6 Comparison of costs approximate solution for $h^{(N)} = 1$, $b^{(N)} = 3N$

N	σ_A	I_N^A	β_N^A	$F_N(I_N^A, \beta_N^A)$	\hat{I}_N^A	$\hat{\beta}_N^A$	$F_N(\hat{I}_N^A, \hat{\beta}_N^A)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right) \sqrt{\log N}$
10	0.1	1.726	1.31058	25.9539	1.224	0.884692	31.2239	0.2561
50	0.1	2.533	1.5931	159.026	2.050	1.27624	173.141	0.1612
100	0.1	2.883	1.69656	341.44	2.405	1.41084	367.575	0.1526
10	0.5	2.067	1.43331	28.3311	1.513	1.0992	31.2606	0.1422
50	0.5	2.987	1.74381	173.875	2.428	1.48993	183.166	0.1003
100	0.5	3.370	1.86469	371.779	2.814	1.62542	387.809	0.0887
10	0.75	2.449	1.57036	31.4004	1.694	1.18023	35.5139	0.1758
50	0.75	3.418	1.89842	190.571	2.664	1.58369	205.174	0.1408
100	0.75	3.899	2.01955	404.306	3.070	1.72277	429.58	0.1263
10	1	2.913	1.72878	34.6096	1.875	1.23092	40.7704	0.2293
50	1	4.158	2.06968	207.553	2.899	1.65341	230.281	0.1952
100	1	4.567	2.20696	439.681	3.326	1.79761	479.663	0.1789

conclude that for small variations in demand and low backorder costs, the asymptotic approach performs well in terms of costs already for a reasonable number of components. Also, the performance improves by increasing N . Finally, the performance of the approximations highly depends on the backorder costs relative to the holding costs.

6. Mixed-behavior approximations

The numerical results in Section 5.3 show that the approximations are in most of the cases around 10-15% off the optimal value. In this section, we show how we can further improve the approximations.

Under deterministic demand and stochastic demand, the approximate problems are given in Definition 4.2 and Definition 5.1. If σ_A is small, then we know that on the one hand,

$$\max_{i \leq N} Q_i \approx \frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N,$$

because Q_i and Q_j are only slightly correlated. But on the other hand,

$$\max_{i \leq N} Q_i \approx \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} \log N \approx \frac{\sigma^2}{2} \log N.$$

Since the Gumbel term is missing here, this could be the reason that this approximation is not working for small N . Thus, it could be beneficial to look at the combination of these two approximations. Then, we have

$$\max_{i \leq N} Q_i \approx \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G. \quad (29)$$

When we replace $\max_{i \leq N} Q_i$ with Equation (29) in the minimization problem, we get

$$\min_{I, \beta} \left(\frac{1}{\beta} \mathbb{E} \left[N h^{(N)} (I - Q_i) + (N h^{(N)} + b^{(N)}) \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G - I \right)^+ \right] + \beta N \right).$$

The optimal I_N^M satisfies $\mathbb{P} \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G < I_N^M \right) = 1 - \gamma_N$. Thus,

$$\int_{-\infty}^{\infty} \exp \left(- \exp \left(- \frac{2}{\sigma^2} \left(I_N^M - \frac{\sigma^2}{2} \log N - \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} x \right) \right) \right) \phi(x) dx = 1 - \gamma_N. \quad (30)$$

Now, I_N^M can be computed through standard numerical methods such as the bisection method. Furthermore, the optimal net capacity β_N^M satisfies

$$\beta_N^M = \frac{\sqrt{\mathbb{E} \left[Nh^{(N)}(I_N^M - Q_i) + (Nh^{(N)} + b^{(N)}) \left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G - I_N^M \right)^+ \right]}}{\sqrt{N}}.$$

Though we have a symbolic expression for β_N^M , it is not completely clear how to compute

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G - I_N^M \right)^+ \right] \\ &= \int_{I_N^M}^{\infty} \mathbb{P} \left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G > x \right) dx. \end{aligned}$$

We can write

$$\begin{aligned} & \mathbb{P} \left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G > x \right) \\ &= \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} + G > \frac{2}{\sigma^2} x - \log N \right) \\ &= \int_{-\infty}^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N - z \right) \exp(-\exp(-z) - z) dz. \end{aligned}$$

Now, we write $z = -\log s$. Then,

$$\begin{aligned} & \int_{-\infty}^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N - z \right) \exp(-\exp(-z) - z) dz \\ &= \int_0^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N + \log s \right) \exp(-s) ds. \end{aligned}$$

Thus,

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G - I_N^M \right)^+ \right] \\ &= \int_{I_N^M}^{\infty} \int_0^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N + \log s \right) \exp(-s) ds dx \\ &= \int_0^{\infty} \int_{I_N^M}^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N + \log s \right) \exp(-s) dx ds. \end{aligned}$$

It turns out that

$$\int_{I_N^M}^{\infty} \mathbb{P} \left(\frac{\sigma_A \sqrt{2}}{\sigma} \sqrt{\log NX} > \frac{2}{\sigma^2} x - \log N + \log s \right) \exp(-s) dx$$

can be expressed in terms of error functions. Thus, since I_N^M can be numerically found by solving Equation (30), $\mathbb{E} \left[\left(\frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log NX} + \frac{\sigma^2}{2} G - I_N^M \right)^+ \right]$ can be computed numerically as well. Observe that the running time of the procedure to obtain I_N^M and β_N^M is independent of the system size N and is efficient.

6.1. Numerical results mixed-behavior approximations

Using the same simulation procedure as described in Section 5.3, we evaluate the performance of these adjusted approximations. The results for the cases of $h^{(N)} = 1$, $b^{(N)} = N$ and $h^{(N)} = 1$, $b^{(N)} = 3N$ are given in Tables 7 and 8, respectively.

From the simulation results we can conclude that these adjusted approximations result in costs that are much closer to the optimal costs, already for small N . When comparing the last two columns, where the last column repeats the results from Section 5.3, we observe that the mixed-behavior approximations show better convergence, also when σ_A is larger. Furthermore, where we saw in Section 5.3 that the cost difference increased considerably with the change in $b^{(N)}$, we now do see an increase, but the difference is still small for a larger value of $b^{(N)}$. Therefore, we can conclude that these mixed-behavior approximations perform well especially when demand variations are no more than 75% of the variations in component production, even with a small number of components.

Table 7 Comparison of costs master solution for $h^{(N)} = 1$, $b^{(N)} = N$

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right) \sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\bar{I}_N^A, \bar{\beta}_N^A)}\right) \sqrt{\log N}$
10	0.1	1.33785	1.1945	23.2022	0.000837	0.082011
50	0.1	2.14487	1.49567	147.567	0.000442	0.036877
100	0.1	2.49244	1.60808	318.638	0.000337	0.029273
10	0.5	1.38072	1.21129	25.4342	0.006038	0.090320
50	0.5	2.19829	1.53814	160.497	0.006938	0.057107
100	0.5	2.54871	1.65808	345.247	0.008143	0.054563
10	0.75	1.40013	1.2128	27.6956	0.027647	0.131055
50	0.75	2.216	1.56166	174.269	0.032074	0.099827
100	0.75	2.5656	1.68745	372.643	0.030493	0.089412
10	1	1.41255	1.19665	31.5428	0.081950	0.210871
50	1	2.22627	1.57136	192.722	0.076684	0.157827
100	1	2.57434	1.70384	407.343	0.072043	0.141724

Table 8 Comparison of costs master solution for $h^{(N)} = 1$, $b^{(N)} = 3N$

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right) \sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\bar{I}_N^A, \bar{\beta}_N^A)}\right) \sqrt{\log N}$
10	0.1	1.78238	1.34746	25.9965	0.002487	0.256113
50	0.1	2.59271	1.62088	159.162	0.001690	0.161243
100	0.1	2.94168	1.72533	341.49	0.000314	0.152581
10	0.5	1.94345	1.38309	28.3671	0.001926	0.142201
50	0.5	2.83775	1.68955	174.284	0.004642	0.100327
100	0.5	3.21861	1.8044	372.617	0.004826	0.088703
10	0.75	2.09429	1.41142	32.0055	0.028689	0.175760
50	0.75	3.04648	1.74512	193.854	0.033496	0.140773
100	0.75	3.44819	1.86761	410.624	0.033019	0.126256
10	1	2.25658	1.43095	36.5165	0.079240	0.229298
50	1	3.26538	1.79271	216.91	0.085321	0.195211
100	1	3.68765	1.92281	456.859	0.080689	0.178876

7. Conclusions

In this study, we defined a large scale assembly system in which N components are assembled into a final product. The delays per component are written as an all-time supremum of a Brownian motion minus a drift term. We aimed to minimize the total costs in the system with respect to the inventory and net capacity per component. The costs in the system consist of inventory holding costs for each component and penalty costs for delay of assembling the final product, which is equal to the delay of the slowest produced component. Before we tried to solve the minimization problem, we simplified the minimization problem, using the self-similarity property of a Brownian motion, into two separate minimization problems. We distinguished two cases, first of all we covered the case of deterministic demand, resulting in all delays being independent. Secondly, we investigated the case that demand is stochastic and consequently delays of the components are dependent.

For the deterministic demand scenario, we proved order bounds for three different regimes: balanced, quality driven and efficiency driven. Additionally, we verified numerically that already for a limited number of components, our approximations result in costs that are very close to the costs corresponding to the optimal solution. For the stochastic demand scenario, we developed a novel limit theorem that we use to obtain approximate solutions. We showed numerically that even though theoretically these approximation perform well, for practical situations there is still room for improvement. Therefore, we provided additional approximations for a mixed-behavior regime, where we use a combination of the approximations for the deterministic and stochastic demand scenarios. We demonstrated numerically that these approximations perform very well already for a practical number of components.

Future work could extend the model to a decentralized minimization problem, where the components are not produced in-house by the OEM but are sourced at outside suppliers that have their own objectives, which results in an asymptotic analysis of a game theoretical equilibrium, cf. Nair et al. (2016), Gopalakrishnan et al. (2016) and Kumar and Randhawa (2010). Additionally, we expect that we can extend the result in Theorem 5.2 to general Lévy processes. However, the cost minimization problem relies heavily on the self-similarity property of Brownian motions. Thus, to solve the minimization problem for Lévy processes, other techniques are needed.

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. The research is also supported by the NWO programs MEERVOUD [Vlasiou: 632.003.002] and Talent VICI [Zwart: 639.033.413].

References

- Altendorfer K, Minner S (2011) Simultaneous optimization of capacity and planned lead time in a two-stage production system with different customer due dates. *European Journal of Operational Research* 213(1):134–146.
- ASML Holding NV (2021) ASML annual report 2020. <https://www.asml.com/en/investors/annual-report/2020>.
- Asmussen S, Glynn PW, Pitman J (1995) Discretization error in simulation of one-dimensional reflecting Brownian motion. *The Annals of Applied Probability* 875–896.
- Atan Z, Rousseau M (2016) Inventory optimization for perishables subject to supply disruptions. *Optimization Letters* 10(1):89–108.
- Atar R, Mandelbaum A, Zviran A (2012) Control of fork-join networks in heavy traffic. *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 823–830 (IEEE).
- Baccelli F (1985) Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. *RR-0426, INRIA. ffinria-00076130* .
- Baccelli F, Makowski AM (1989) Queueing models for systems with synchronization constraints. *Proceedings of the IEEE* 77(1):138–161.

- Bernstein F, DeCroix GA (2006) Inventory policies in a decentralized assembly system. *Operations Research* 54(2):324–336.
- Bollapragada R, Rao US, Zhang J (2004) Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Science* 50(12):1729–1743.
- Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Operations Research* 52(1):17–34.
- Bradley JR, Glynn PW (2002) Managing capacity and inventory jointly in manufacturing systems. *Management Science* 48(2):273–288.
- Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. *Journal of Applied Probability* 732–739.
- Dai J, Harrison JM (1992) Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability* 65–86.
- Dębicki K, Hashorva E, Ji L, Tabiś K (2015) Extremes of vector-valued Gaussian processes: Exact asymptotics. *Stochastic Processes and their Applications* 125(11):4039–4065.
- Dębicki K, Ji L, Rolski T (2020) Exact asymptotics of component-wise extrema of two-dimensional Brownian motion. *Extremes* 23:569–602.
- Flatto L, Hahn S (1984) Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics* 44(5):1041–1053.
- Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.
- Gopalakrishnan R, Doroudi S, Ward AR, Wierman A (2016) Routing and staffing when servers are strategic. *Operations Research* 64(4):1033–1050.
- de Haan L, Ferreira A (2006) *Extreme value theory: an introduction* (Springer Science & Business Media).
- Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.
- Harrison JM (2013) *Brownian Models of Performance and Control* (Cambridge University Press), URL <http://dx.doi.org/10.1017/CB09781139087698>.
- Klein SJD (1988) *Fredholm integral equations in queueing analysis*. Ph.D. thesis, Rijksuniversiteit Utrecht.
- Klosterhalfen ST, Minner S, Willems SP (2014) Strategic safety stock placement in supply networks with static dual supply. *Manufacturing and Service Operations Management* 16(2):204–219.
- Ko SS, Serfozo RF (2004) Response times in M/M/s fork-join networks. *Advances in Applied Probability* 36(3):854–871.
- Kou S, Zhong H, et al. (2016) First-passage times of two-dimensional Brownian motion. *Advances in Applied Probability* 48(4):1045–1060.
- Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing & Service Operations Management* 12(3):511–526.
- Leadbetter MR, Lindgren G, Rootzén H (1983) *Extremes and related properties of random sequences and processes* (Springer Science & Business Media).
- van Leeuwen JS, Mathijsen BW, Zwart B (2019) Economies-of-scale in many-server queueing systems: Tutorial and partial review of the qed halfin–whitt heavy-traffic regime. *SIAM Review* 61(3):403–440.
- Lu H, Pang G (2015) Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Mathematics of Operations Research* 41(2):560–595.
- Lu H, Pang G (2017a) Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stochastic Systems* 6(2):519–600.
- Lu H, Pang G (2017b) Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services. *Queueing Systems* 85(1-2):67–115.
- Masih-Tehrani B, Xu SH, Kumara S, Li H (2011) A single-period analysis of a two-echelon inventory system with dependent supply uncertainty. *Transportation Research Part B: Methodological* 45(8):1128–1151.

- Mayorga ME, Ahn HS (2011) Joint management of capacity and inventory in make-to-stock production systems with multi-class demand. *European Journal of Operational Research* 212(2):312–324.
- Nair J, Wierman A, Zwart B (2016) Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base. *Management Science* 62(6):1830–1841.
- Nelson R, Tantawi AN (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers* 37(6):739–743.
- Nguyen V (1993) Processing networks with parallel and sequential tasks: Heavy traffic analysis and brownian limits. *The Annals of Applied Probability* 28–55.
- Nguyen V (1994) The trouble with diversity: Fork-join networks with heterogeneous customer population. *The Annals of Applied Probability* 1–25.
- Pan W, So KC (2016) Component procurement strategies in decentralized assembly systems under supply uncertainty. *IIE Transactions* 48(3):267–282.
- Pickands III J (1968) Moment convergence of sample extremes. *The Annals of Mathematical Statistics* 39(3):881–889.
- Reddy KN, Kumar A (2020) Capacity investment and inventory planning for a hybrid manufacturing-remanufacturing system in the circular economy. *International Journal of Production Research* 1–29.
- Reed J, Zhang B (2017) Managing capacity and inventory jointly for multi-server make-to-stock queues. *Queueing Systems* 86(1-2):61–94.
- Resnick SI (1987) *Extreme values, regular variation and point processes* (Springer).
- Sleptchenko A, van der Heijden MC, van Harten A (2003) Trade-off between inventory and repair capacity in spare part networks. *Journal of the Operational Research Society* 54(3):263–272.
- Svoboda J, Minner S, Yao M (2020) Typology and literature review on multiple supplier inventory control models. *European Journal of Operational Research* .
- van der Vaart AW (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press), URL <http://dx.doi.org/10.1017/CB09780511802256>.
- Varma S (1990) *Heavy and light traffic approximations for queues with synchronization constraints*. Ph.D. thesis, University of Maryland.
- Wright PE (1992) Two parallel processors with coupled inputs. *Advances in Applied Probability* 24(4):986–1007.
- Wu J, Chao X (2014) Optimal control of a Brownian production/inventory system with average cost criterion. *Mathematics of Operations Research* 39(1):163–189.
- Zieliński R (2009) Optimal nonparametric quantile estimators. Towards a general theory. A survey. *Communications in Statistics-Theory and Methods* 38(7):980–992.
- Zou X, Pokharel S, Piplani R (2004) Channel coordination in an assembly system facing uncertain demand with synchronized processing time and delivery quantity. *International Journal of Production Research* 42(22):4673–4689.

Appendix A: Proofs

A.1. Proofs of Section 3

LEMMA A.1. (i) In the independent case $\sigma_A = 0$:

$$\begin{aligned} & \min_{(\beta_1, \dots, \beta_N), (I_1, \dots, I_N)} \sum_{i=1}^N \mathbb{E} [h^{(N)}(I_i - Q_i(\beta_i)) + \beta_i] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta_j) - I_j)^+ \right] \\ &= \min_{(\beta, I)} \mathbb{E} [Nh^{(N)}(I - Q_i(\beta))] + \beta N + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta) - I)^+ \right], \end{aligned}$$

(ii) in the dependent case $\sigma_A > 0$:

$$\min_{\beta, (I_1, I_2, \dots, I_N)} \sum_{i=1}^N \mathbb{E} [h^{(N)}(I_i - Q_i(\beta)) + \beta] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta) - I_j)^+ \right]$$

$$= \min_{(\beta, I)} \mathbb{E} [Nh^{(N)}(I - Q_i(\beta))] + \beta N + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta) - I)^+ \right].$$

Proof In the independent case, we can write, by using the self-similarity property of Brownian motions, that

$$\begin{aligned} & \sum_{i=1}^N \mathbb{E} [h^{(N)}(I_i - Q_i(\beta_i)) + \beta_i] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta_j) - I_j)^+ \right] \\ &= \sum_{i=1}^N \mathbb{E} \left[h^{(N)} \left(I_i - \frac{1}{\beta_i} Q_i(1) \right) + \beta_i \right] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} \left(\frac{1}{\beta_j} Q_j(1) - I_j \right)^+ \right]. \end{aligned}$$

We write $\eta_i = 1/\beta_i$. Thus,

$$\begin{aligned} & \sum_{i=1}^N \mathbb{E} \left[h^{(N)} \left(I_i - \frac{1}{\beta_i} Q_i(1) \right) + \beta_i \right] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} \left(\frac{1}{\beta_j} Q_j(1) - I_j \right)^+ \right] \\ &= \sum_{i=1}^N \mathbb{E} \left[h^{(N)} (I_i - \eta_i Q_i(1)) + \frac{1}{\eta_i} \right] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (\eta_j Q_j(1) - I_j)^+ \right]. \end{aligned}$$

It is easy to see that $\sum_{i=1}^N \mathbb{E} [h^{(N)}(I_i - \eta_i Q_i(1)) + 1/\eta_i]$ is convex with respect to $(\eta_1, \dots, \eta_N, I_1, \dots, I_N)$, with $\eta_j, I_j > 0$. In order to examine whether $\mathbb{E} \left[\max_{j \leq N} (\eta_j Q_j(1) - I_j)^+ \right]$ is convex we should prove convexity of $\eta_j Q_j(1) - I_j$, because taking the expectation of a convex function and taking maxima of convex functions preserve convexity. Since $\eta_j Q_j(1) - I_j$ is linear in both η_j and I_j , convexity holds. Now, assume

$$C = \min_{(\beta_1, \beta_2, \dots, \beta_N), (I_1, I_2, \dots, I_N)} \sum_{i=1}^N \mathbb{E} [h^{(N)}(I_i - Q_i(\beta_i)) + \beta_i] + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta_j) - I_j)^+ \right]$$

with minimizers $(\beta_1^{(l)}, \dots, \beta_N^{(l)})$ and $(I_1^{(l)}, \dots, I_N^{(l)})$. Assume there exists i, j such that $\beta_i^{(l)} \neq \beta_j^{(l)}$ or $I_i^{(l)} \neq I_j^{(l)}$. Then, because of the symmetry of the problem with respect to the N servers, all the permutations of the minimizers give solutions. Assume there are k permutations, where the l -th permutation has minimizers $(\beta_1^{(l)}, \dots, \beta_N^{(l)})$ and $(I_1^{(l)}, \dots, I_N^{(l)})$. Now, define β_i and I_i such that they satisfy $1/\beta_i = \frac{1}{k} \sum_{l=1}^k 1/\beta_i^{(l)}$, and $I_i = \frac{1}{k} \sum_{l=1}^k I_i^{(l)}$. Because of the symmetry of the cost function around the N servers, we have that $\beta_i = \beta_j = \beta$, and $I_i = I_j = I$. Since we have a convex function with respect to I_i and $1/\beta_i$,

$$C \geq \mathbb{E} [Nh^{(N)}(I - Q_i(\beta))] + \beta N + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{j \leq N} (Q_j(\beta) - I)^+ \right].$$

Thus $I_i = I$, and $\beta_i = \beta$ are minimizers. An analogous derivation holds for the dependent case where we only minimize over one drift parameter. \square

REMARK 1. In the dependent case where all servers choose a different drift parameter, we have that $\sup_{s>0} (W_i(s) + W_A(s) - \beta_i s) = \sup_{s>0} (\hat{W}_i(s) + \hat{W}_A(s) - s)/\beta_i$ where $\hat{W}_i(s) = W_i(s/\beta_i^2)\beta_i$ and $\hat{W}_A(s) = W_A(s/\beta_i^2)\beta_i$. However, $\mathbb{E} [W_A(s/\beta_i^2)\beta_i W_A(s/\beta_j^2)\beta_j] = \sigma_A^2 \beta_i \beta_j s / \max(\beta_i, \beta_j)^2 \neq \sigma_A^2 s$ when $\beta_i \neq \beta_j$. Thus, when we have different drift parameters β_i and β_j , the joint distribution of $\sup_{s>0} (W_i(s) + W_A(s) - \beta_i s)$ and $\sup_{s>0} (W_j(s) + W_A(s) - \beta_j s)$ is not the same as the joint distribution of $\sup_{s>0} (W_i(s) + W_A(s) - s)/\beta_i$ and $\sup_{s>0} (W_j(s) + W_A(s) - s)/\beta_j$. So to prove Lemma A.1 when the drifts are different, other techniques are needed.

Proof of Lemma 3.2 $F_N(I, \beta) > 0$, hence F_N has a global infimum, and since $\lim_{\beta \downarrow 0} F_N(I, \beta) = \infty$, $\lim_{\beta \rightarrow \infty} F_N(I, \beta) = \infty$ and $\lim_{I \rightarrow \infty} F_N(I, \beta) = \infty$, F_N has a global minimum. Now, assume $F_N(I_N, \beta_N) = \min_{(I, \beta)} F_N(I, \beta)$. Assume that there exists an \hat{I}_N such that

$$\mathbb{E} \left[Nh^{(N)} \left(\hat{I}_N - Q_i + \left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ \right) + b^{(N)} \left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ \right] < \mathbb{E} \left[Nh^{(N)} \left(I_N - Q_i + \left(\max_{i \leq N} Q_i - I_N \right)^+ \right) + b^{(N)} \left(\max_{i \leq N} Q_i - I_N \right)^+ \right].$$

Then $F_N(\hat{I}_N, \beta_N) < F_N(I_N, \beta_N)$. This contradicts the statement that (I_N, β_N) gives the minimum of F_N . Hence, the optimal inventory minimizes $C_N(I)$. The proof that β_N minimizes $\frac{1}{\beta} C_N(I_N) + \beta N$ goes analogously.

To prove that $C_N(I)$ is convex with respect to I , we observe that

$$\begin{aligned} \frac{d^2}{dI^2} C_N(I) &= (b^{(N)} + Nh^{(N)}) \frac{d^2}{dI^2} \mathbb{E} \left[\left(\max_{i \leq N} Q_i - I \right)^+ \right] = (b^{(N)} + Nh^{(N)}) \frac{d^2}{dI^2} \int_I^\infty \mathbb{P} \left(\max_{i \leq N} Q_i > x \right) dx \\ &= (b^{(N)} + Nh^{(N)}) f(I) \geq 0, \end{aligned}$$

because f is the probability density function of $\max_{i \leq N} Q_i$. This density exists; cf. Dai and Harrison (1992, Prop. 2a). In conclusion, we have a convex minimization problem. Moreover, $\frac{d^2}{d\beta^2} \left(\frac{1}{\beta} C_N(I_N) + \beta N \right) = \frac{2}{\beta^3} C_N(I_N) > 0$. Thus $\frac{1}{\beta} C_N(I_N) + \beta N$ is also convex with respect to β . \square

Proof of Lemma 3.3 $F_N(I, \beta)$ has the form $F_N(I, \beta) = \frac{1}{\beta} C_N(I) + \beta N$, thus in order to minimize $F_N(I_N^*, \beta)$, we know by Lemma 3.2 that we need to solve $\frac{d}{d\beta} F_N(I_N^*, \beta) = -\frac{1}{\beta^2} C_N(I_N^*) + N = 0$. Thus, $\beta_N^* = \frac{\sqrt{C_N(I_N^*)}}{\sqrt{N}}$, and $F_N(I_N^*, \beta_N^*) = 2\sqrt{N C_N(I_N^*)} = 2N\beta_N^*$. \square

Proof of Lemma 3.4 To solve $\min_I C_N(I)$ we have to solve $\frac{d}{dI} C_N(I) = 0$, this gives for the optimal inventory I_N^* that

$$Nh^{(N)} - (Nh^{(N)} + b^{(N)}) \mathbb{P} \left(\max_{i \leq N} Q_i > I_N^* \right) = 0.$$

Hence $I_N^* = P_N^{-1} \left(\frac{b^{(N)}}{Nh^{(N)} + b^{(N)}} \right)$, with P_N^{-1} the quantile function of $\max_{i \leq N} Q_i$. \square

Proof of Lemma 3.6 Following Corollary 3.7, we have

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\bar{I}_N, \bar{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)} \sqrt{C_N(\bar{I}_N)}}{C_N(\bar{I}_N) + \bar{C}_N(\bar{I}_N)}.$$

Furthermore, observe that

$$\mathbb{E} \left[\max_{i \leq N} Q_i \right] \geq \mathbb{E} \left[\max_{i \leq N} \sup_{s > 0} (W_i(s) - s) + W_A(\tau) \right] = \frac{\sigma^2}{2} \sum_{i=1}^N \frac{1}{i} \geq \frac{\sigma^2}{2} \log N,$$

where τ is the first hitting time of the supremum of $\max_{i \leq N} (W_i(t) - t)$. From this it follows that for $I < \frac{\sigma^2}{2} \log N$, $\frac{\sigma^2}{2} \log N - I < \mathbb{E}[\max_{i \leq N} Q_i - I] < \mathbb{E}[(\max_{i \leq N} Q_i - I)^+]$. For $I > \frac{\sigma^2}{2} \log N$, $(\frac{\sigma^2}{2} \log N - I)^+ = 0 < \mathbb{E}[(\max_{i \leq N} Q_i - I)^+]$. In conclusion, $C_N(I) > \bar{C}_N(I)$. Therefore,

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\bar{I}_N, \bar{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)} \sqrt{C_N(\bar{I}_N)}}{C_N(\bar{I}_N) + \bar{C}_N(\bar{I}_N)} \geq \frac{\sqrt{C_N(I_N^*)} \sqrt{C_N(\bar{I}_N)}}{C_N(\bar{I}_N)}.$$

We have $|C_N(I_N^*) - C_N(\bar{I}_N)| \leq (2Nh^{(N)} + b^{(N)})|I_N^* - \bar{I}_N|$, and

$$|\bar{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| \leq (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N \right| \right].$$

In the case that $\gamma_N = \gamma \in (0, 1)$, we have by applying Lemma 3.5 that $|\bar{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| = o((Nh^{(N)} + b^{(N)}) \log N)$. Furthermore, $C_N(\bar{I}_N) \sim Nh^{(N)} \frac{\sigma^2}{2} \log N$, and since $\max_{i \leq N} Q_i / \log N \xrightarrow{\mathbb{P}} \sigma^2/2$, as $N \rightarrow \infty$, we also have that $I_N^* / \log N \xrightarrow{N \rightarrow \infty} \sigma^2/2$. Thus $|C_N(I_N^*) - C_N(\bar{I}_N)| = o((Nh^{(N)} + b^{(N)}) \log N)$, and the lemma follows.

In the case that $\gamma_N \xrightarrow{N \rightarrow \infty} 1$, we first observe that $\bar{C}_N(\bar{I}_N) = Nh^{(N)} \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) \sim Nh^{(N)} \frac{\sigma^2}{2} \log N$. Furthermore,

$$\begin{aligned} C_N(\bar{I}_N) &= Nh^{(N)} \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N \right)^+ \right] \\ &\leq Nh^{(N)} \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N \right| \right]. \end{aligned}$$

Thus,

$$\frac{C_N(\bar{I}_N)}{Nh^{(N)} \log N} \leq \frac{\sigma^2}{2} + o(1) + \frac{1}{\gamma_N} \frac{\mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N \right| \right]}{\log N}.$$

By Lemma 3.5, we know that $\mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N \right| \right] / \log N \xrightarrow{N \rightarrow \infty} 0$. Thus

$$\limsup_{N \rightarrow \infty} C_N(\bar{I}_N) / (Nh^{(N)} \log N) \leq \sigma^2/2.$$

Finally,

$$\begin{aligned} C_N(I_N^*) &= Nh^{(N)} \left(I_N^* - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - I_N^* \right)^+ \right] \\ &\geq Nh^{(N)} \left(I_N^* - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\max_{i \leq N} Q_i - I_N^* \right] \\ &\geq -Nh^{(N)} \frac{\sigma^2 + \sigma_A^2}{2} + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \log N - b^{(N)} I_N^*. \end{aligned}$$

$I_N^* = O(\log N)$, and $b^{(N)} / (Nh^{(N)}) \xrightarrow{N \rightarrow \infty} 0$, therefore, $\liminf_{N \rightarrow \infty} C_N(I_N^*) / (Nh^{(N)} \log N) \geq \sigma^2/2$. Combining these results gives

$$\liminf_{N \rightarrow \infty} \frac{F_N(I_N^*, \beta_N^*)}{F_N(\bar{I}_N, \bar{\beta}_N)} \geq \liminf_{N \rightarrow \infty} \frac{\sqrt{C_N(I_N^*)} \sqrt{C_N(\bar{I}_N)}}{C_N(\bar{I}_N)} = 1.$$

□

A.2. Proofs of Section 4

Proof of Lemma 4.1 In Lemma 3.4, it is shown that $I_N^* = P_N^{-1}(1 - \gamma_N)$, with P_N^{-1} the quantile function of $\max_{i \leq N} Q_i$. Because $(Q_i, i \leq N)$ are independent and exponentially distributed,

$$\mathbb{P} \left(\max_{i \leq N} Q_i \leq P_N^{-1}(x) \right) = x = \left(1 - e^{-\frac{2}{\sigma^2} P_N^{-1}(x)} \right)^N.$$

From this it follows that $P_N^{-1}(x) = \frac{\sigma^2}{2} \log \left(1 / \left(1 - x^{\frac{1}{N}} \right) \right)$.

□

Proof of Proposition 4.3 Minimizing $\hat{F}_N(\hat{I}_N, \hat{\beta}_N)$ goes analogously as minimizing $F_N(I_N, \beta_N)$ in Lemma 4.1. Hence $\hat{I}_N = \hat{P}_N^{-1}(1 - \gamma_N)$. Thus, we have to solve

$$\mathbb{P}\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N \leq \hat{P}_N^{-1}(x)\right) = \mathbb{P}\left(G \leq \frac{2}{\sigma^2}\hat{P}_N^{-1}(x) - \log N\right) = e^{-e^{-\left(\frac{2}{\sigma^2}\hat{P}_N^{-1}(x) - \log N\right)}} = x.$$

Therefore, $\hat{P}_N^{-1}(x) = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log(-\log x)$. Hence, the optimal inventory is given in Equation (7). Furthermore,

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N - \hat{I}_N\right)^+\right] &= \mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log(-\log(1 - \gamma_N))\right)^+\right] \\ &= \frac{\sigma^2}{2} \int_{-\log(-\log(1 - \gamma_N))}^{\infty} 1 - e^{-e^{-x}} dx. \end{aligned}$$

By using partial integration and substitution we can write

$$\frac{\sigma^2}{2} \int_{-\log(-\log(1 - \gamma_N))}^{\infty} 1 - e^{-e^{-x}} dx = \frac{\sigma^2}{2} \left(\int_{-\log(1 - \gamma_N)}^{\infty} \frac{e^{-t}}{t} dt + \Gamma + \log(-\log(1 - \gamma_N)) \right).$$

Hence, this gives us the expression of $\hat{C}_N(\hat{I}_N)$ in (8). \square

Proof of Lemma 4.6 To prove that G_N follows a Gumbel distribution, we first observe that $\mathbb{P}(\max_{i \leq N} Q_i < x) = (1 - \exp(-\frac{2}{\sigma^2}x))^N$. Therefore, $(1 - \exp(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i))^N \sim \text{Unif}[0, 1]$. Then,

$$\begin{aligned} \mathbb{P}(G_N < x) &= \mathbb{P}\left(-\log\left(-\log\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i\right)\right)^N\right)\right) < x\right) \\ &= \mathbb{P}\left(-\log\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i\right)\right)^N\right) > e^{-x}\right) \\ &= \mathbb{P}\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i\right)\right)^N < e^{-e^{-x}}\right) = e^{-e^{-x}}. \end{aligned}$$

To prove (16), we need to show that for all $x > 0$ and N

$$x > -\frac{\sigma^2}{2} \log\left(-\log\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}x\right)\right)^N\right)\right) + \frac{\sigma^2}{2} \log N.$$

This is equivalent to the inequality $x > -\frac{\sigma^2}{2} \log(-\log(1 - \exp(-\frac{2}{\sigma^2}x)))$, which is equivalent to $1 - e^{-\frac{2}{\sigma^2}x} < e^{-e^{-\frac{2}{\sigma^2}x}}$, with $x > 0$. This is equivalent to $e^{-y} > 1 - y$ for $y \in (0, e^{-1}]$. Observe that for $y = 0$, we have equality, and we have for $y > 0$ that $(e^{-y})' > -1 = (1 - y)'$. The statement follows. To prove that the larger $\max_{i \leq N} Q_i$ becomes, the smaller the difference between $\max_{i \leq N} Q_i$ and $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N$ becomes, we first observe that

$$\begin{aligned} \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N &= -\frac{\sigma^2}{2} \log\left(-\log\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i\right)\right)^N\right)\right) + \frac{\sigma^2}{2} \log N \\ &= -\frac{\sigma^2}{2} \log\left(-\log\left(1 - e^{-\frac{2}{\sigma^2}\max_{i \leq N} Q_i}\right)\right). \end{aligned}$$

Thus we need to obtain that $x + \frac{\sigma^2}{2} \log(-\log(1 - e^{-\frac{2x}{\sigma^2}}))$ is strictly decreasing in x for $x > 0$. Taking the first derivative gives the inequality

$$\frac{e^{-\frac{2x}{\sigma^2}}}{\left(1 - e^{-\frac{2x}{\sigma^2}}\right) \log\left(1 - e^{-\frac{2x}{\sigma^2}}\right)} + 1 < 0.$$

This is equivalent to the inequality $-y/((1-y)\log(1-y)) > 1$ for $y \in (0, 1)$, which can be rewritten to $\log y > 1 - 1/y$, which is a basic logarithm inequality. Finally, $\lim_{x \rightarrow \infty} x + \frac{\sigma^2}{2} \log(-\log(1 - e^{-\frac{2x}{\sigma^2}})) = 0$. \square

Proof of Lemma 4.7 Due to the inequality in (16), $I_N^* > \hat{I}_N$, then, we have

$$\begin{aligned} C_N(I_N^*) - C_N(\hat{I}_N) &= Nh^{(N)}(I_N^* - \hat{I}_N) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - I_N^* \right)^+ - \left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ \right] \\ &= Nh^{(N)}(I_N^* - \hat{I}_N) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\hat{I}_N - I_N^* \right) \mathbb{1} \left(\max_{i \leq N} Q_i > I_N^* \right) \right] \\ &\quad - (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ \mathbb{1} \left(\hat{I}_N < \max_{i \leq N} Q_i < I_N^* \right) \right]. \end{aligned}$$

We have $\mathbb{P}(\max_{i \leq N} Q_i > I_N^*) = \gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, thus

$$Nh^{(N)}(I_N^* - \hat{I}_N) + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\hat{I}_N - I_N^* \right) \mathbb{1} \left(\max_{i \leq N} Q_i > I_N^* \right) \right] = 0.$$

Furthermore,

$$\begin{aligned} \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ \mathbb{1} \left(\hat{I}_N < \max_{i \leq N} Q_i < I_N^* \right) \right] &\leq (I_N^* - \hat{I}_N) \mathbb{P} \left(\hat{I}_N < \max_{i \leq N} Q_i < I_N^* \right) \\ &= (I_N^* - \hat{I}_N) \left(1 - \gamma_N - \left(1 + \frac{\log(1 - \gamma_N)}{N} \right)^N \right). \end{aligned}$$

Equation (17) follows. To prove Equation (18), we observe that

$$\begin{aligned} &|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)| \\ &= (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \hat{I}_N \right)^+ - \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N - \hat{I}_N \right)^+ \right] \\ &= (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \frac{\sigma^2}{2} G_N - \frac{\sigma^2}{2} \log N \right) \mathbb{1} \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N \right) \right] \quad (31) \\ &+ (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \hat{I}_N \right) \mathbb{1} \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N < \max_{i \leq N} Q_i \right) \right]. \quad (32) \end{aligned}$$

Because G_N and $\max_{i \leq N} Q_i$ are on the same probability space, we have $\mathbb{P} \left(\max_{i \leq N} Q_i = I_N^* \mid \frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N = \hat{I}_N \right) = 1$. Furthermore, $x + \frac{\sigma^2}{2} \log(-\log(1 - e^{-\frac{2x}{\sigma^2}}))$ is decreasing in x . Thus, we can bound

$$\begin{aligned} &\mathbb{E} \left[\left(\max_{i \leq N} Q_i - \frac{\sigma^2}{2} G_N - \frac{\sigma^2}{2} \log N \right) \mathbb{1} \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N \right) \right] \\ &\leq (I_N^* - \hat{I}_N) \mathbb{P} \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N \right) \\ &= (I_N^* - \hat{I}_N) \gamma_N. \end{aligned} \quad (33)$$

Similarly, for (32), we observe that if $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N$, then $\max_{i \leq N} Q_i < I_N^*$, thus,

$$\begin{aligned} & \mathbb{E} \left[\left(\max_{i \leq N} Q_i - \hat{I}_N \right) \mathbb{1} \left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \max_{i \leq N} Q_i \right) \right] \\ & \leq (I_N^* - \hat{I}_N) \mathbb{P} \left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \max_{i \leq N} Q_i \right) \\ & \leq (I_N^* - \hat{I}_N) \left(1 - \left(1 + \frac{\log(1-\gamma_N)}{N} \right)^N - \gamma_N \right). \end{aligned} \quad (34)$$

Adding the bounds in (33) and (34) gives the result. \square

Proof of Theorem 4.4 First of all, we assume that $\gamma_N = \gamma \in (0, 1)$. Using Corollary 3.7, we have

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = \frac{2\sqrt{C_N(I_N^*)}\sqrt{\hat{C}_N(\hat{I}_N)}}{C_N(\hat{I}_N) + \hat{C}_N(\hat{I}_N)}.$$

Because of the inequality in (16), we have for all I that $C_N(I) > \hat{C}_N(I)$, thus

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} > \frac{2\sqrt{C_N(I_N^*)}\sqrt{\hat{C}_N(\hat{I}_N)}}{2C_N(\hat{I}_N)}.$$

By computing the Taylor series around $x = 0$, we have

$$\begin{aligned} I_{1/x}^* &= \frac{\sigma^2}{2} \log \left(\frac{1}{1 - (1-\gamma)^x} \right) = -\frac{\sigma^2}{2} \log x - \frac{\sigma^2}{2} \log(-\log(1-\gamma)) - \frac{\sigma^2}{4} x \log(1-\gamma) + O(x^2) \\ &= \hat{I}_{1/x} - \frac{\sigma^2}{4} x \log(1-\gamma) + O(x^2). \end{aligned}$$

Thus, $(I_N^* - \hat{I}_N) \sim -\sigma^2 \log(1-\gamma)/(4N)$. Following (18), we can conclude that $|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|/(Nh^{(N)}) = O(1/N)$. We can do the same for $\mathbb{P}(\hat{I}_N < \max_{i \leq N} Q_i < I_N^*)$, and get

$$\left(1 - \gamma - \left(1 + \frac{\log(1-\gamma)}{N} \right)^N \right) \sim \frac{1}{2N} (1-\gamma) \log(1-\gamma)^2.$$

Thus, after applying the inequality in (17), we get $|C_N(I_N^*) - C_N(\hat{I}_N)|/(Nh^{(N)} + b^{(N)}) = O(1/N^2)$. We have

$$\begin{aligned} \hat{C}_N(\hat{I}_N) &= Nh^{(N)} \frac{\sigma^2}{2} (\log N - \log(-\log(1-\gamma)) - 1) + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \mathbb{E}[(G + \log(-\log(1-\gamma)))^+] \\ &\sim Nh^{(N)} \frac{\sigma^2}{2} \log N, \end{aligned}$$

because $(Nh^{(N)} + b^{(N)})/(Nh^{(N)}) = 1/\gamma$, and $-\log(-\log(1-\gamma))$ and $\mathbb{E}[(G_N + \log(-\log(1-\gamma)))^+]$ are of $O(1)$. In conclusion, we have

$$\begin{aligned} \frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} &> \frac{\sqrt{C_N(I_N^*)} \sqrt{\hat{C}_N(\hat{I}_N)}}{\sqrt{C_N(\hat{I}_N)} \sqrt{C_N(\hat{I}_N)}} \\ &= \frac{\sqrt{C_N(I_N^*) - O((Nh^{(N)} + b^{(N)})/N^2)} \sqrt{C_N(\hat{I}_N) - O(Nh^{(N)}/N)}}{\sqrt{C_N(\hat{I}_N)} \sqrt{C_N(\hat{I}_N)}} \\ &= \frac{\sqrt{1 - O(1/(N^2 \log N))} \sqrt{1 - O(1/(N \log N))}}{\sqrt{C_N(\hat{I}_N)} \sqrt{C_N(\hat{I}_N)}} \\ &= 1 - O(1/(N \log N)). \end{aligned}$$

Now, we assume that $\gamma_N \xrightarrow{N \rightarrow \infty} 0$, then we have that $-\log(-\log(1 - \gamma_N)) \sim -\log(\gamma_N)$, thus $\hat{I}_N \sim \frac{\sigma^2}{2} \log(N/\gamma_N)$. Also,

$$\mathbb{E}[(G_N + \log(-\log(1 - \gamma_N)))^+] \sim \mathbb{E}[(G_N + \log(\gamma_N))^+] \sim \gamma_N.$$

From this it follows that $\hat{C}_N(\hat{I}_N) \sim Nh^{(N)} \frac{\sigma^2}{2} \log(N/\gamma_N)$. Furthermore,

$$\mathbb{P}\left(\max_{i \leq N} Q_i > \hat{I}_N\right) = 1 - \left(1 + \frac{\log(1 - \gamma_N)}{N}\right)^N \leq N \mathbb{P}(Q_i > \hat{I}_N) = -\log(1 - \gamma_N) = \gamma_N(1 + O(\gamma_N/2)).$$

From this it follows that

$$\left(1 - \gamma_N - \left(1 + \frac{\log(1 - \gamma_N)}{N}\right)^N\right) \leq -\log(1 - \gamma_N) - \gamma_N = \frac{\gamma_N^2}{2}(1 + o(1)).$$

Also

$$\mathbb{P}\left(\max_{i \leq N} Q_i < I_N^*\right) = \mathbb{P}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N\right) = 1 - \gamma_N \xrightarrow{N \rightarrow \infty} 1.$$

Earlier, we showed that when $\gamma_N = \gamma$, $(I_N^* - \hat{I}_N) = O(1/N)$, now I_N^* is larger, because $\mathbb{P}(\max_{i \leq N} Q_i < I_N^*) = 1 - \gamma_N \xrightarrow{N \rightarrow \infty} 1$. Following the statement in Lemma 4.6 that the difference between $\max_{i \leq N} Q_i$ and $\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N$ decreases as $\max_{i \leq N} Q_i$ increases, we can conclude that $(I_N^* - \hat{I}_N) = O(1/N)$. Following the proof before, and by using the order bounds in (17) and (18), we have that

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(\gamma_N/(N \log(N/\gamma_N))).$$

Finally, we consider the case that $\gamma_N \xrightarrow{N \rightarrow \infty} 1$ and $\gamma_N \leq 1 - \exp(-N)$. Then, $\hat{I}_N \geq 0$. Furthermore, when $\gamma_N \xrightarrow{N \rightarrow \infty} 1$, we have $\log(-\log(1 - \gamma_N)) \xrightarrow{N \rightarrow \infty} \infty$, from this it follows that

$$\mathbb{E}[(G_N + \log(-\log(1 - \gamma_N)))^+] \sim \log(-\log(1 - \gamma_N)).$$

Thus

$$\begin{aligned} \hat{C}_N(\hat{I}_N) &\sim \frac{\sigma^2}{2} Nh^{(N)} (\log N - \log(-\log(1 - \gamma_N))) + \frac{\sigma^2}{2} (Nh^{(N)} + b^{(N)}) \log(-\log(1 - \gamma_N)) \\ &= \frac{\sigma^2}{2} Nh^{(N)} \log N + \frac{\sigma^2}{2} b^{(N)} \log(-\log(1 - \gamma_N)). \end{aligned}$$

Since we consider the efficiency driven regime, we have $b^{(N)}/(Nh^{(N)}) \xrightarrow{N \rightarrow \infty} 0$. Also, it is easy to deduce that when $\gamma_N < 1 - \exp(-N)$, we have $\log(-\log(1 - \gamma_N)) < \log N$. Thus $\hat{C}_N(\hat{I}_N) \sim \frac{\sigma^2}{2} Nh^{(N)} \log N$. Furthermore, $I_N^* - \hat{I}_N = O(1)$, thus the bounds in (17) and (18) are of $O(Nh^{(N)})$. By using the same argument as in the proof for the balanced regime,

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(1/\log N).$$

□

Proof of Lemma 4.5 Following Equations (17) and (18) and using the same arguments as in the proof of Theorem 4.4, we can find the same order bound for $F_N(I_N^*, \beta_N^*)/\hat{F}_N(\hat{I}_N, \hat{\beta}_N) = \sqrt{C_N(I_N^*)}/\sqrt{\hat{C}_N(\hat{I}_N)}$.

In the case that $\gamma_N = \gamma \in (0, 1)$, we have

$$\begin{aligned} \hat{C}_N(\hat{I}_N) &= Nh^{(N)} \frac{\sigma^2}{2} (\log N - \log(-\log(1-\gamma)) - 1) \\ &\quad + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \mathbb{E} \left[(G + \log(-\log(1-\gamma)))^+ \right]. \end{aligned}$$

Thus $\hat{F}_N(\hat{I}_N, \hat{\beta}_N)/(N \log N) = 2\sqrt{N} \sqrt{\hat{C}_N(\hat{I}_N)}/(N \log N) = O(\sqrt{h^{(N)}}/\sqrt{\log N})$.

When $\gamma_N \xrightarrow{N \rightarrow \infty} 0$, we have that $-\log(-\log(1-\gamma_N)) \sim -\log(\gamma_N)$, thus $\hat{I}_N \sim \frac{\sigma^2}{2} \log(N/\gamma_N)$. Also,

$$\mathbb{E} \left[(G_N + \log(-\log(1-\gamma_N)))^+ \right] \sim \mathbb{E} \left[(G_N + \log(\gamma_N))^+ \right] \sim \gamma_N.$$

From this it follows that

$$\hat{C}_N(\hat{I}_N) \sim Nh^{(N)} \frac{\sigma^2}{2} (\log(N/\gamma_N) - 1) + (Nh^{(N)} + b^{(N)}) \frac{\sigma^2}{2} \gamma_N.$$

Therefore, $2\sqrt{N} \sqrt{\hat{C}_N(\hat{I}_N)}/(N \log(N/\gamma_N)) = O(\gamma_N \sqrt{h^{(N)}}/\sqrt{\log(N/\gamma_N)})$.

When $\gamma_N \xrightarrow{N \rightarrow \infty} 1$, we have

$$\begin{aligned} \hat{C}_N(\hat{I}_N) &\sim \frac{\sigma^2}{2} Nh^{(N)} (\log N - \log(-\log(1-\gamma_N))) + \frac{\sigma^2}{2} (Nh^{(N)} + b^{(N)}) \log(-\log(1-\gamma_N)) \\ &= \frac{\sigma^2}{2} Nh^{(N)} \log N + \frac{\sigma^2}{2} b^{(N)} \log(-\log(1-\gamma_N)). \end{aligned}$$

Thus, $2\sqrt{N} \sqrt{\hat{C}_N(\hat{I}_N)}/\log N = O(N \sqrt{h^{(N)}}/\sqrt{\log N})$. \square

A.3. Proofs of Section 5.1

Proof of Lemma 5.3 Let $b_N = \sqrt{2 \log N} - \log(4\pi \log N)/(2\sqrt{2 \log N})$. Then

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

with $G \sim \text{Gumbel}$, as $N \rightarrow \infty$, cf. de Haan and Ferreira (2006, p. 11, Ex. 1.1.7) for a proof. Observe that

$$\begin{aligned} &b_N \left(\frac{\max_{i \leq N} W_i(d \log N)}{\sigma \sqrt{d \log N}} - b_N \right) \\ &= \frac{1}{\sigma \sqrt{d}} \left(\sqrt{2 \log N} - \frac{\log(4\pi \log N)}{2\sqrt{2 \log N}} \right) \frac{\max_{i \leq N} W_i(d \log N) - \sigma \sqrt{2d} \log N + \frac{\sigma \sqrt{d} \log(4\pi \log N)}{2\sqrt{2}}}{\sqrt{\log N}}. \end{aligned}$$

Furthermore, $\beta d + \frac{\sigma^2}{2\beta} = \sigma \sqrt{2d} = \frac{\sigma^2}{\beta}$. From this it follows that

$$\frac{\max_{i \leq N} W_i(d \log N) - \beta d \log N - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0.$$

Moreover, $\frac{W_A(d \log N)}{\sqrt{\log N}} \stackrel{d}{=} \frac{\sigma \sigma_A}{\sqrt{2\beta}} X$, with $X \sim \mathcal{N}(0, 1)$. The statement follows. \square

Proof of Lemma 5.4 To prove Lemma 5.4, we first observe that

$$\begin{aligned}
& \frac{\max_{i \leq N} \left(\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) + W_A(s) - \beta s) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\
& \leq \frac{\max_{i \leq N} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\
& + \frac{\sup_{0 < s < (d-\epsilon) \log N} \left(W_A(s) - \frac{1}{\log \log N} s \right)}{\sqrt{\log N}} \\
& \leq \frac{\max_{i \leq N} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} + \frac{\sup_{s > 0} \left(W_A(s) - \frac{1}{\log \log N} s \right)}{\sqrt{\log N}}.
\end{aligned}$$

Furthermore, we can write

$$\begin{aligned}
& \mathbb{P} \left(\frac{\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} > x \right) \\
& = \mathbb{P} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) > x \sqrt{\log N} + \frac{\sigma^2}{2\beta} \log N \right).
\end{aligned}$$

We know that $\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right)$ is a reflected Brownian motion, so we can write down its cumulative distribution function explicitly:

$$\begin{aligned}
& \mathbb{P} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) \leq x \right) \\
& = 1 - \Phi \left(\frac{-x - \left(\beta - \frac{1}{\log \log N} \right) (d-\epsilon) \log N}{\sigma \sqrt{(d-\epsilon) \log N}} \right) \\
& - \exp \left(-\frac{2 \left(\beta - \frac{1}{\log \log N} \right) x}{\sigma^2} \right) \Phi \left(\frac{-x + \left(\beta - \frac{1}{\log \log N} \right) (d-\epsilon) \log N}{\sigma \sqrt{(d-\epsilon) \log N}} \right).
\end{aligned}$$

It turns out that

$$\mathbb{P} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) < x \sqrt{\log N} + \frac{\sigma^2}{2\beta} \log N \right)^N \xrightarrow{N \rightarrow \infty} 1,$$

for all x . One can see this heuristically by observing that

$$\max_{i \leq N} \sup_{s < (d-\epsilon) \log N} (W_i(s) - \beta s) \approx \max_{i \leq N} (W_i((d-\epsilon) \log N) - \beta (d-\epsilon) \log N),$$

because the hitting time of the supremum of $\max_{i \leq N} (W_i(s) - \beta s)$ is approximately $d \log N$. Thus, up to that time $\max_{i \leq N} (W_i(s) - \beta s)$ is increasing. We know that $\max_{i \leq N} W_i((d-\epsilon) \log N) \approx \sqrt{2(d-\epsilon)\sigma \log N}$. Therefore,

$$\begin{aligned}
& \frac{\max_{i \leq N} \sup_{s < (d-\epsilon) \log N} (W_i(s) - \beta s)}{\log N} \xrightarrow{\mathbb{P}} \frac{\sigma \sqrt{\sigma^2 - 2\beta^2 \epsilon}}{\beta} - \frac{\sigma^2}{2\beta} + \beta \epsilon \\
& = \frac{\sigma^2}{2\beta} + \left(\frac{\sigma \sqrt{\sigma^2 - 2\beta^2 \epsilon}}{\beta} - \frac{\sigma^2}{\beta} + \beta \epsilon \right) \leq \frac{\sigma^2}{2\beta} - C\epsilon^2, \quad (35)
\end{aligned}$$

with $C > 0$. Hence,

$$\begin{aligned} & \frac{\max_{i \leq N} \sup_{s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right)}{\log N} \\ & \leq \frac{\max_{i \leq N} \sup_{s < (d-\epsilon) \log N} (W_i(s) - \beta s)}{\log N} + \frac{(d-\epsilon) \log N}{(\log \log N) \log N} \\ & \xrightarrow{\mathbb{P}} \frac{\sigma^2}{2\beta} + \left(\frac{\sigma \sqrt{\sigma^2 - 2\beta^2 \epsilon}}{\beta} - \frac{\sigma^2}{\beta} + \beta \epsilon \right) \leq \frac{\sigma^2}{2\beta} - C\epsilon^2. \end{aligned}$$

Thus,

$$\frac{\max_{i \leq N} \sup_{s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \leq -C\epsilon^2 \sqrt{\log N},$$

for N large. Therefore,

$$\mathbb{P} \left(\sup_{0 < s < (d-\epsilon) \log N} \left(W_i(s) - \left(\beta - \frac{1}{\log \log N} \right) s \right) < x \sqrt{\log N} + \frac{\sigma^2}{2\beta} \log N \right)^N \xrightarrow{N \rightarrow \infty} 1.$$

Furthermore, $\sup_{s > 0} \left(W_A(s) - \frac{s}{\log \log N} \right) \sim \text{Exp} \left(\frac{2}{\sigma_A^2 \log \log N} \right)$. Therefore, $\frac{\sup_{s > 0} \left(W_A(s) - \frac{s}{\log \log N} \right)}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0$, as $N \rightarrow \infty$. The statement follows. \square

Proof of Lemma 5.5 Let $\epsilon > 0$ be given. Choose $\delta < \min \left(\frac{2(\beta^3 \epsilon + \beta \sigma^2)}{2\beta^2 \epsilon + \sigma^2} - 2\sqrt{\frac{\beta^2 \sigma^2}{2\beta^2 \epsilon + \sigma^2}}, \frac{2\beta^3 \epsilon}{2\beta^2 \epsilon + \sigma^2}, \beta \right)$ and positive. Then

$$\begin{aligned} & \frac{\max_{i \leq N} \left(\sup_{s \geq (d+\epsilon) \log N} (W_i(s) + W_A(s) - \beta s) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\ & \leq \frac{\max_{i \leq N} \left(\sup_{s \geq (d+\epsilon) \log N} (W_i(s) - (\beta - \delta)s) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} + \frac{\sup_{s \geq (d+\epsilon) \log N} (W_A(s) - \delta s)}{\sqrt{\log N}} \\ & \leq \frac{\max_{i \leq N} \left(\sup_{s \geq (d+\epsilon) \log N} (W_i(s) - (\beta - \delta)s) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} + \frac{\sup_{s > 0} (W_A(s) - \delta s)}{\sqrt{\log N}}. \end{aligned}$$

We have

$$\sup_{s \geq (d+\epsilon) \log N} (W_i(s) - (\beta - \delta)s) \stackrel{d}{=} W_i((d+\epsilon) \log N) - (\beta - \delta)(d+\epsilon) \log N + \sup_{s > 0} (W'_i(s) - (\beta - \delta)s),$$

with $(W'_i, i \leq N)$ independent Brownian motions with mean 0 and variance σ^2 . We write $E_i = \sup_{s > 0} (W'_i(s) - (\beta - \delta)s)$. Hence, $E_i \sim \text{Exp} \left(\frac{2(\beta - \delta)}{\sigma^2} \right)$. So

$$\begin{aligned} & \frac{\max_{i \leq N} \left(\sup_{s \geq (d+\epsilon) \log N} (W_i(s) - (\beta - \delta)s) \right) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\ & \stackrel{d}{=} \frac{\max_{i \leq N} (W_i((d+\epsilon) \log N) + E_i) - \left(\frac{\sigma^2}{2\beta} + (\beta - \delta)(d+\epsilon) \right) \log N}{\sqrt{\log N}}. \end{aligned}$$

By using the union bound and Chernoff's bound, we get that

$$\begin{aligned} \mathbb{P} \left(\max_{i \leq N} (W_i((d+\epsilon) \log N) + E_i) > x \right) & \leq N \mathbb{P} (W_i((d+\epsilon) \log N) + E_i > x) \\ & \leq N \mathbb{E} \left[e^{s W_i((d+\epsilon) \log N)} \right] \mathbb{E} \left[e^{s E_i} \right] e^{-sx}, \end{aligned}$$

for all $s > 0$. $\mathbb{E}[e^{sW_i((d+\epsilon)\log N)}] = e^{\frac{s^2(\sigma\sqrt{(d+\epsilon)\log N})^2}{2}} = N^{\frac{\sigma^2(d+\epsilon)s^2}{2}}$ and $\mathbb{E}[e^{sE_i}] = \frac{2(\beta-\delta)}{\sigma^2} / \left(\frac{2(\beta-\delta)}{\sigma^2} - s\right)$. Hence,

$$\begin{aligned} & \mathbb{P}\left(\max_{i \leq N} (W_i((d+\epsilon)\log N) + E_i) > x\sqrt{\log N} + \left(\frac{\sigma^2}{2\beta} + (\beta-\delta)(d+\epsilon)\right)\log N\right) \\ & \leq N^{1+\frac{\sigma^2(d+\epsilon)s^2}{2}-s\left(\frac{\sigma^2}{2\beta}+(\beta-\delta)(d+\epsilon)\right)} e^{-sx\sqrt{\log N} \frac{2(\beta-\delta)}{\frac{2(\beta-\delta)}{\sigma^2}-s}}. \end{aligned} \quad (36)$$

Now, we choose $s^* = \frac{\beta}{2\beta^2\epsilon+\sigma^2} + \frac{\beta-\delta}{\sigma^2}$. Because $\delta < \frac{2\beta^3\epsilon}{2\beta^2\epsilon+\sigma^2}$, $s^* < \frac{2(\beta-\delta)}{\sigma^2}$. Also,

$$1 + \frac{\sigma^2(d+\epsilon)s^{*2}}{2} - s^* \left(\frac{\sigma^2}{2\beta} + (\beta-\delta)(d+\epsilon)\right) < 0,$$

because $\delta < \frac{2(\beta^3\epsilon+\beta\sigma^2)}{2\beta^2\epsilon+\sigma^2} - 2\sqrt{\frac{\beta^2\sigma^2}{2\beta^2\epsilon+\sigma^2}}$. Therefore

$$\mathbb{P}\left(\max_{i \leq N} (W_i((d+\epsilon)\log N) + E_i) > x\sqrt{\log N} + \left(\frac{\sigma^2}{2\beta} + (\beta-\delta)(d+\epsilon)\right)\log N\right) \xrightarrow{N \rightarrow \infty} 0.$$

Moreover, $\sup_{s>0}(W_A(s) - \delta s) \sim \text{Exp}\left(\frac{2\delta}{\sigma_A^2}\right)$. Therefore, $\frac{\sup_{s>0}(W_A(s) - \delta s)}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0$. The limit in (22) follows. \square

Proof of Lemma 5.6 First of all, we bound

$$\begin{aligned} & \frac{\max_{i \leq N} \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\ & \leq \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N} \frac{W_A(s)}{\sqrt{\log N}} + \frac{\max_{i \leq N} \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N} (W_i(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \\ & \leq \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N} \frac{W_A(s)}{\sqrt{\log N}} + \frac{\max_{i \leq N} \sup_{s>0} (W_i(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}}. \end{aligned}$$

We can write

$$\begin{aligned} \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N} \frac{W_A(s)}{\sqrt{\log N}} &= \frac{W_A((d-\epsilon)\log N)}{\sqrt{\log N}} + \sup_{0 \leq s < 2\epsilon \log N} \frac{W'_A(s)}{\sqrt{\log N}} \\ &\stackrel{d}{=} \sigma_A \sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon X_1 + \sqrt{2\epsilon}\sigma_A |X_2|}, \end{aligned}$$

with $X_1, X_2 \sim \mathcal{N}(0, 1)$ and independent, and W'_A a Brownian motion with mean 0 and variance σ_A^2 . Furthermore, we have that

$$\frac{2\beta}{\sigma^2} \left(\max_{i \leq N} \sup_{s>0} (W_i(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N \right) \xrightarrow{d} G,$$

as $N \rightarrow \infty$, with $G \sim \text{Gumbel}$. Therefore,

$$\frac{\max_{i \leq N} \sup_{s>0} (W_i(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0,$$

as $N \rightarrow \infty$. The statement follows. \square

Proof of Theorem 5.2 We have the following lower bound:

$$\begin{aligned} & \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \geq \mathbb{P} \left(\frac{\max_{i \leq N} (W_i(d \log N) + W_A(d \log N)) - \beta d \log N - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right). \end{aligned}$$

From this and Lemma 5.3, we know that

$$\liminf_{N \rightarrow \infty} \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \geq 1 - \Phi \left(\frac{x\sqrt{2}\beta}{\sigma\sigma_A} \right).$$

By using the union bound, we get

$$\begin{aligned} & \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \leq \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{0 < s < (d-\epsilon) \log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \quad + \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{(d-\epsilon) \log N \leq s < (d+\epsilon) \log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \quad + \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s \geq (d+\epsilon) \log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right). \end{aligned}$$

Combining this with the results from Lemmas 5.4, 5.5 and 5.6 gives

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \leq \mathbb{P} \left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2}} - \epsilon X_1 + \sqrt{2\epsilon} \sigma_A |X_2| > x \right), \end{aligned}$$

with $X_1, X_2 \sim \mathcal{N}(0, 1)$ and independent. This upper bound holds for all $\epsilon > 0$, therefore

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right) \\ & \leq \lim_{\epsilon \downarrow 0} \mathbb{P} \left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2}} - \epsilon X_1 + \sqrt{2\epsilon} \sigma_A |X_2| > x \right) \\ & = 1 - \Phi \left(\frac{x\sqrt{2}\beta}{\sigma\sigma_A} \right). \end{aligned}$$

Hence, the statement follows. \square

Proof of Lemma 5.7 Because of the self-similarity property, we can assume without loss of generality that $\beta = 1$. Let $d = \frac{\sigma^2}{2}$, and $X_N = \frac{\sqrt{2}}{\sigma\sigma_A} \frac{W_A(d \log N)}{\sqrt{\log N}}$. It is easy to see that $X_N \sim \mathcal{N}(0, 1)$. Let $0 < \epsilon < d$, we write

$$Q_i = \sup_{s > 0} (W_i(s) + W_A(s) - s),$$

$$Q_i^{(1,N)} = \sup_{0 < s < (d-\epsilon) \log N} (W_i(s) + W_A(s) - s),$$

$$Q_i^{(2,N)} = \sup_{(d-\epsilon) \log N < s < (d+\epsilon) \log N} (W_i(s) + W_A(s) - s),$$

and

$$Q_i^{(3,N)} = \sup_{s > (d+\epsilon) \log N} (W_i(s) + W_A(s) - s).$$

We want to prove that

$$\mathbb{E} \left[\left| \frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N \right| \right] \xrightarrow{N \rightarrow \infty} 0. \quad (37)$$

First observe that

$$\mathbb{E} \left[\left| \frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N \right| \right] \quad (38)$$

$$\leq \mathbb{E} \left[\left| \frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right| \right] \quad (39)$$

$$+ \mathbb{E} \left[\left| \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N \right| \right]. \quad (40)$$

Because $Q_i > W_i(d \log N) + W_A(d \log N) - d \log N$, we can rewrite (39):

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right| \right] \\ &= \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right]. \end{aligned} \quad (41)$$

Moreover, due to Pickands III (1968, Th. 3.1):

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N \right| \right] \\ &= \mathbb{E} \left[\left| \frac{\max_{i \leq N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right| \right] \xrightarrow{N \rightarrow \infty} 0. \end{aligned} \quad (42)$$

If $x = \max(x^{(1)}, x^{(2)}, x^{(3)})$, with $x^{(1)}, x^{(2)}, x^{(3)} \geq 0$, then $x \leq x^{(1)} + x^{(2)} + x^{(3)}$. Thus,

$$\mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \quad (43)$$

$$\leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(1,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \quad (44)$$

$$+ \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \quad (45)$$

$$+ \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(3,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right]. \quad (46)$$

For (44), we have

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(1,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(1,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] + \mathbb{E} \left[\left(-\frac{\max_{i \leq N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right]. \end{aligned} \quad (47)$$

By (42), the second term converges to 0. For the first term, following Lemma 5.4, we observe that

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(1,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} (\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s)) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right. \right. \\ & \quad \left. \left. + \frac{\sup_{s > 0} (W_A(s) - s/\log \log N) - \frac{W_A(d \log N)}{\sqrt{\log N}}}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} (\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s)) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \end{aligned} \quad (48)$$

$$+ \mathbb{E} \left[\left(\frac{\sup_{s > 0} (W_A(s) - s/\log \log N)}{\sqrt{\log N}} \right)^+ \right]. \quad (49)$$

The term in (49) converges to 0. Furthermore,

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} (\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s)) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \\ & = \int_0^\infty \mathbb{P} \left(\frac{\max_{i \leq N} (\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s)) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} > x \right) dx \\ & = \int_0^\infty \mathbb{P} \left(\max_{i \leq N} \sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s) - W_A(d \log N) > x \sqrt{\log N} + \frac{\sigma^2}{2} \log N \right) dx \\ & \leq \int_0^\infty N \mathbb{P} \left(\sup_{0 < s < (d-\epsilon) \log N} (W_i(s) - (1 - 1/\log \log N)s) \right. \\ & \quad \left. > x/2 \sqrt{\log N} + \left(\frac{\sigma^2}{2} - \frac{1}{2} (\sigma^2 - \sigma \sqrt{\sigma^2 - 2\epsilon} - \epsilon) \right) \log N \right) dx \\ & + \int_0^\infty \mathbb{P} \left(-W_A(d \log N) > x/2 \sqrt{\log N} + \left(\frac{1}{2} (\sigma^2 - \sigma \sqrt{\sigma^2 - 2\epsilon} - \epsilon) \right) \log N \right) dx \\ & \xrightarrow{N \rightarrow \infty} 0; \end{aligned}$$

see the inequality (35) in the proof of Lemma 5.4 for details. For the term in (45), we have that

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \\ & = \mathbb{E} \left[\frac{\max_{i \leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right] \\ & = \mathbb{E} \left[\frac{\max_{i \leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right]. \end{aligned}$$

By Pickands III (1968, Th. 3.1), we have that

$$\mathbb{E} \left[\frac{\max_{i \leq N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right] \xrightarrow{N \rightarrow \infty} 0.$$

Furthermore, (here we use the same bounds as in Lemma 5.6)

$$\begin{aligned} & \mathbb{E} \left[\frac{\max_{i \leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right] \\ & \leq \mathbb{E} \left[\frac{\max_{i \leq N} \sup_{s > 0} (W_i(s) - s) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right] + \mathbb{E} \left[\frac{W_A((d - \epsilon) \log N) + \sup_{0 < 2\epsilon \log N} \tilde{W}_A(s)}{\sqrt{\log N}} \right] \\ & \xrightarrow{N \rightarrow \infty} 0 + \sqrt{2\epsilon} \sigma_A \mathbb{E}[|X_N|] = \sqrt{2\epsilon} \sigma_A \sqrt{\frac{2}{\pi}}. \end{aligned} \quad (50)$$

Similar as in (47), we have

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(3,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\max_{i \leq N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(3,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] + \mathbb{E} \left[\left(-\frac{\max_{i \leq N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} \right)^+ \right]. \end{aligned}$$

The second term goes to 0, following the proof of Lemma 5.5, for the first term we have

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} Q_i^{(3,N)} - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} \sup_{s > (d+\epsilon) \log N} (W_i(s) - (1-\delta)s) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right. \right. \\ & \quad \left. \left. + \frac{\sup_{s > (d+\epsilon) \log N} (W_A(s) - \delta s) - W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \\ & \leq \mathbb{E} \left[\left(\frac{\max_{i \leq N} \sup_{s > (d+\epsilon) \log N} (W_i(s) - (1-\delta)s) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right)^+ \right] \end{aligned} \quad (51)$$

$$+ \mathbb{E} \left[\left(\frac{\sup_{s > (d+\epsilon) \log N} (W_A(s) - \delta s) - W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right]. \quad (52)$$

For (51), we have, by using the inequality from Equation (36) with $s^* = 1/(2\epsilon + \sigma^2) + (1-\delta)/\sigma^2$, that

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\max_{i \leq N} \sup_{s > (d+\epsilon) \log N} (W_i(s) - (1-\delta)s) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} \right)^+ \right] \\ & = \int_0^\infty \mathbb{P} \left(\frac{\max_{i \leq N} \sup_{s > (d+\epsilon) \log N} (W_i(s) - (1-\delta)s) - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} > x \right) dx \\ & \leq \int_0^\infty N^{1 + \frac{\sigma^2(d+\epsilon)s^{*2}}{2} - s^* \left(\frac{\sigma^2}{2} + (1-\delta)(d+\epsilon) \right)} e^{-sx\sqrt{\log N}} \frac{\frac{2(1-\delta)}{\sigma^2}}{\frac{2(1-\delta)}{\sigma^2} - s^*} dx \\ & = N^{1 + \frac{\sigma^2(d+\epsilon)s^{*2}}{2} - s^* \left(\frac{\sigma^2}{2} + (1-\delta)(d+\epsilon) \right)} \frac{\frac{2(1-\delta)}{\sigma^2}}{\frac{2(1-\delta)}{\sigma^2} - s^*} \frac{1}{s^* \sqrt{\log N}} \xrightarrow{N \rightarrow \infty} 0. \end{aligned}$$

For (52), we observe that

$$\begin{aligned}
& \limsup_{N \rightarrow \infty} \mathbb{E} \left[\left(\frac{\sup_{s > (d+\epsilon) \log N} (W_A(s) - \delta s) - W_A(d \log N)}{\sqrt{\log N}} \right)^+ \right] \\
&= \limsup_{N \rightarrow \infty} \mathbb{E} \left[\left(\frac{\tilde{W}_A(\epsilon \log N) - \delta(d+\epsilon) \log N + \sup_{s > 0} (W_A(s) - \delta s)}{\sqrt{\log N}} \right)^+ \right] \\
&\leq \mathbb{E} \left[\left(\frac{\tilde{W}_A(\epsilon \log N)}{\sqrt{\log N}} \right)^+ \right] + \mathbb{E} \left[\left(\frac{-\delta(d+\epsilon) \log N}{\sqrt{\log N}} \right)^+ \right] + \mathbb{E} \left[\left(\frac{\sup_{s > 0} (W_A(s) - \delta s)}{\sqrt{\log N}} \right)^+ \right] \\
&\xrightarrow{N \rightarrow \infty} \sigma_A \sqrt{\epsilon} \sqrt{\frac{1}{2\pi}}.
\end{aligned} \tag{53}$$

Concluding, after we collect the non-zero answers which are given in (50) and (53) we get

$$\limsup_{N \rightarrow \infty} \mathbb{E} \left[\left| \frac{\max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N \right| \right] \leq \sqrt{2\epsilon} \sigma_A \sqrt{\frac{2}{\pi}} + \sigma_A \sqrt{\epsilon} \sqrt{\frac{1}{2\pi}} \xrightarrow{\epsilon \downarrow 0} 0.$$

□

A.4. Proofs of Section 5.2

Proof of Lemma 5.9 From Lemma 3.2, we know that the optimal inventory I_N^A satisfies

$$\frac{d}{dI} \mathbb{E} \left[Nh^{(N)} \left(I_N^A - Q_i + \left(\max_{i \leq N} Q_i - I_N^A \right)^+ \right) + b^{(N)} \left(\max_{i \leq N} Q_i - I_N^A \right)^+ \right] = 0.$$

We have

$$\begin{aligned}
& \frac{d}{dI} \mathbb{E} \left[Nh^{(N)} \left(I_N^A - Q_i + \left(\max_{i \leq N} Q_i - I_N^A \right)^+ \right) + b^{(N)} \left(\max_{i \leq N} Q_i - I_N^A \right)^+ \right] \\
&= Nh^{(N)} - (Nh^{(N)} + b^{(N)}) \mathbb{P} \left(\max_{i \leq N} Q_i > I_N^A \right) \\
&= Nh^{(N)} - (Nh^{(N)} + b^{(N)}) \mathbb{P} \left(\frac{\sqrt{2} \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N}{\sigma \sigma_A \sqrt{\log N}} > \frac{\sqrt{2} I_N^A - \frac{\sigma^2}{2} \log N}{\sigma \sigma_A \sqrt{\log N}} \right).
\end{aligned}$$

Therefore, I_N^A satisfies $\frac{\sqrt{2}}{\sigma \sigma_A} (I_N^A - \frac{\sigma^2}{2} \log N) / \sqrt{\log N} = P_N^{A-1} (1 - \gamma_N)$. □

Proof of Proposition 5.10 We have to find I and β such that $F_N(I, \beta)$ is minimized. As before, we know that the optimal \hat{I}_N^A should satisfy

$$Nh^{(N)} - (Nh^{(N)} + b^{(N)}) \mathbb{P} \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X > \hat{I}_N^A \right) = 0.$$

Thus, \hat{I}_N^A as given in (25) minimizes $\hat{C}_N^A(I)$. We know that

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X - \hat{I}_N^A \right)^+ \right] &= \int_{\frac{\hat{I}_N^A - \frac{\sigma^2}{2} \log N}{\frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N}}}^{\infty} \left(\frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} x - \hat{I}_N^A \right) \phi(x) dx \\
&= \left(\frac{\sigma^2}{2} \log N - \hat{I}_N^A \right) \mathbb{P} \left(\frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X \geq \hat{I}_N^A - \frac{\sigma^2}{2} \log N \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\sigma^2 \log N - 2\hat{I}_N^A)^2}{4\sigma^2\sigma_A^2 \log N}\right) \\
& = -\frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma_N) \gamma_N \\
& + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \Phi^{-1}(1 - \gamma_N)^2\right).
\end{aligned}$$

The expression in Equation (26) follows. \square

Proof of Theorem 5.11 Using Corollary 3.7, we have

$$\frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)} = \frac{2\sqrt{C_N(I_N^A)}\sqrt{\hat{C}_N^A(\hat{I}_N^A)}}{C_N(\hat{I}_N^A) + \hat{C}_N^A(\hat{I}_N^A)}.$$

First, assume $\hat{C}_N^A(\hat{I}_N^A) > C_N(\hat{I}_N^A)$. Then, $F_N(I_N^A, \beta_N^A)/F_N(\hat{I}_N^A, \hat{\beta}_N^A) > \sqrt{C_N(I_N^A)/\hat{C}_N^A(\hat{I}_N^A)}$. We have

$$|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| \leq (2Nh^{(N)} + b^{(N)})|I_N^A - \hat{I}_N^A| + (Nh^{(N)} + b^{(N)}) \mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N - \frac{\sigma\sigma_A}{\sqrt{2}} X \right| \right].$$

We know by van der Vaart (1998, p. 305, Lem. 21.2), that $(I_N^A - \hat{I}_N^A)/\sqrt{\log N} \xrightarrow{N \rightarrow \infty} 0$. Furthermore, we prove in Lemma 5.7 that $\mathbb{E} \left[\left| \max_{i \leq N} Q_i - \frac{\sigma^2}{2} \log N - \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} X \right| / \sqrt{\log N} \right] \xrightarrow{N \rightarrow \infty} 0$. From this it follows that $|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| = o((Nh^{(N)} + b^{(N)})\sqrt{\log N})$. Since $\hat{C}_N^A(\hat{I}_N^A) \sim \frac{\sigma^2}{2} Nh^{(N)} \log N$, we have $\frac{\sqrt{C_N(I_N^A)}}{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}} = 1 - o((Nh^{(N)} + b^{(N)})\sqrt{\log N}/(Nh^{(N)} \log N)) = 1 - o(1/\sqrt{\log N})$.

Secondly, assume $\hat{C}_N^A(\hat{I}_N^A) < C_N(\hat{I}_N^A)$, then

$$\frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)} > \frac{\sqrt{C_N(I_N^A)\hat{C}_N^A(\hat{I}_N^A)}}{C_N(\hat{I}_N^A)} = \frac{\sqrt{C_N(I_N^A)}}{\sqrt{C_N(\hat{I}_N^A)}} \frac{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}}{\sqrt{C_N(\hat{I}_N^A)}}.$$

With an analogous derivation, we obtain the same order bound. \square

Proof of Lemma 5.12 We have $\hat{I}_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma)$. Furthermore, $|I_N^A - \hat{I}_N^A| = o(\sqrt{\log N})$, thus (27) follows. Furthermore, by using the same argument as in Lemma 4.5, (28) follows. \square