Stochastic Systems

Optimization of Inventory and Capacity in Large-Scale Assembly Systems Using Extreme-Value Theory

Mirjam S. Meijer, Dennis Schol, Willem van Jaarsveld, Maria Vlasiou, Bert Zwart

Please scroll down for article—it is on subsequent pages

Optimization of Inventory and Capacity in Large-Scale Assembly Systems Using Extreme-Value Theory

Mirjam S. Meijer,[a] Dennis Schol,[b] Willem van Jaarsveld,[b] Maria Vlasiou,[b,c] Bert Zwart[b,d,*]

[a] Kühne Logistics University, 20457 Hamburg, Germany; [b] Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands;
[c] University of Twente, 7522 NB Enschede, Netherlands; [d] Centrum Wiskunde & Informatica, 1098 XG Amsterdam, Netherlands
*Corresponding author
Contact: mirjam.meijer@klu.org, https://orcid.org/0000-0002-2260-8557 (MSM); c.schol@tue.nl, https://orcid.org/0000-0002-9601-2203
(DS); w.l.v.jaarsveld@tue.nl, https://orcid.org/0000-0003-3620-4067 (WvJ); m.vlasiou@utwente.nl, https://orcid.org/0000-0002-0457-2925
(MV); bert.zwart@cwi.nl, https://orcid.org/0000-0001-9336-0096 (BZ)

Abstract. High-tech systems are typically produced in two stages: (1) production of components using specialized equipment and staff and (2) system assembly/integration. Component production capacity is subject to fluctuations, causing a high risk of shortages of at least one component, which results in costly delays. Companies hedge this risk by strategic investments in excess production capacity and in buffer inventories of components. To optimize these, it is crucial to characterize the relation between component shortage risk and capacity and inventory investments. We suppose that component production capacity and produce demand are normally distributed over finite time intervals, and we accordingly model the production system as a symmetric fork-join queueing network with N statistically identical queues with a common arrival process and independent service processes. Assuming a symmetric cost structure, we subsequently apply extreme value theory to gain analytic insights into this optimization problem. We derive several new results for this queueing network, notably that the scaled maximum of N steady-state queue lengths converges in distribution to a Gaussian random variable. These results translate into asymptotically optimal methods to dimension the system. Tests on a range of problems reveal that these methods typically work well for systems of moderate size.

Keywords: extreme value theory • asymptotic analysis • capacitated inventory systems

1. Introduction

Delivery reliability is a key performance indicator for high-tech manufacturers, such as ASML, Philips, and Airbus. High-tech systems, such as wafer steppers, medical imaging equipment, and aircraft are produced by assembling thousands of components, each produced by highly skilled staff using specialized equipment. This production system facilitates modular design and testing, but it is also vulnerable: the shortage of a single component will result in delivery delays that cause customer grievances, a build-up of inventory of other components, and a severe reduction in turnover and cashflow. For example, in 2021, ASML was hit by material shortages in its supply chain, causing it to cut its revenue guidance (Denton 2021). Also, in other industries with higher demand volumes, for example, car manufacturing, many components are required to assemble the final product, and a single missing item can hinder production of the entire end-product. An example is the shutdown of complete manufacturing lines at several car manufacturers because of shortages of semiconductors (Ewing and Clark 2021).

Two complementary approaches may contribute to guaranteeing a reliable production system by reducing the risk of component shortages: excess component production capacity and inventory buffers. Production capacity and inventory buffers have a qualitatively different role in the mitigation of component shortages. Excess production capacity implies that the expected maximum number of components that can be produced per quarter

exceeds the expected demand per quarter; for example, as a rule, production capacity may be 110% of expected demand. Inventory buffers are components that are produced in anticipation of demand; typically, such anticipative production continues until the inventory buffer reaches a target, for example, of six weeks of demand. Excess production capacity is always available, whereas inventory buffers are consumed when used to absorb production or demand fluctuations.

Joint optimization of excess component production capacity and component buffers is the ultimate goal because investments in excess component production capacity and component buffer inventories run into the hundreds of millions of euros (ASML Holding NV 2021). High-level investment plans for capacity and inventory may be devised for each product line (e.g., ASML's TWINSCAN XT range or Philips' Azurion 7C range), depending on the role of the product line in the company's portfolio and other considerations. Despite the strategic importance of these investments, there is a lack of quantitative methods for determining appropriate investments in capacity and inventory to achieve the desired level of delivery reliability. Indeed, despite decades of research in inventory management, the joint optimization of production capacity and inventory remains a considerable challenge (Bradley and Glynn 2002). Whereas 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. The much more common situation of assembling a system from many components has proved very challenging.

In this paper, we make a step toward overcoming this challenge. We propose a stylized model capturing key features of high-tech manufacturing that is based on interactions with high-tech manufacturers in the Netherlands and that yields new insights into the joint optimization of capacity and inventory for large-scale assembly systems. We focus on a single product line. Typically, a majority of the expensive components used in high-tech products are common to all products in a product line, being unique to that line, and we consider capacity and inventory optimization for those common components. Component shortages result in delays in the start of the assembly/integration process. Given the tight production planning that is common at high-tech manufacturers, such delays, in turn, result in costly delivery delays. Component production is capacitated and subject to random fluctuations. For example, the production capacity of components may be $\mu \pm \sigma$ items per quarter, and we assume a normal distribution for this per-period production capacity (e.g., Bradley and Glynn 2002, Wu and Chao 2014), which is the most natural assumption as the stochastic term represents the error around the mean. We adopt a continuous-time model, and we likewise assume that production capacity in every finite interval is linear with normally distributed white noise; that is, cumulative net production is a Brownian motion with drift $-\beta < 0$ and variance σ^2 (cf. Bradley and Glynn 2002, Harrison 2013). We analyze the steady-state behavior of this system.

To analyze the overall production system, we consider a symmetric fork-join network of N queues driven by a common arrival process and having independent, identical service processes. Because of this common arrival process, total inventory per component, including backlogged items, is equal for all components. However, as a result of variations in the service times, the number of backlogged items may vary per component. We express the optimal component production capacity and inventory in this model in terms of the steady-state delay distribution of the slowest component, which has the form of a maximum of N all-time suprema of Brownian motions, and we subsequently focus on analyzing this delay distribution. In particular, in large-scale systems with many components/queues, one can expect that the maximum delay (which is due to stochasticity of demand and service processes) grows without bound as a function of the size of the system. To analyze and quantify this phenomenon, we derive new analytic results for the delays in this fork-join network as $N \to \infty$. To do so, we make a major assumption, which is that the randomness and cost characteristics of each of the N suppliers are identical, resulting in a symmetric system with identical net service capacities and base-stock levels. The symmetry we impose makes a mathematical treatment of our model within reach. Whereas this is a shortcoming of our work, it already reveals useful insights, and we complement our analytic results with simulation experiments for asymmetric systems.

1.1. Extreme Value Analysis

Original equipment manufacturers (OEMs) typically level the demand to smooth the production process. Accordingly, in our base model, we assume that demand is completely leveled, which corresponds to a fork-join queue with a deterministic arrival stream. Extremes for this network as $N \to \infty$ are obtained using extreme value theory (EVT), and based on those results, 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 (Gans et al. 2003, Borst et al. 2004, van Leeuwaarden et al. 2019), we distinguish three regimes that depend on the growth rates of cost parameters and are determined by the probability γ_N of not having enough inventory. Given that $\gamma_N \to \gamma$, 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.

1.2. Demand Fluctuations

Other than the number of produced components being stochastic, despite efforts to level demand, typically, some demand variation remains. Thus, a natural choice is that the demand has, apart from a linear term, a white noise term as well, which is normally distributed. Therefore, in Section 5, we assume that the cumulative stochastic demand for systems is modeled by a Brownian motion with variance σ_A^2. (cf. Bradley and Glynn 2002 for a single-component manufacturing system). 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). In high-tech manufacturing, normally distributed demand is a suitable assumption especially when considering longer time periods, but it is also a reasonable approximation for shorter periods. As a consequence of these demand variations, component delays become dependent because they face the same stochastic demands from system assembly. The question is now how this affects the maximum delay as the number of queues/components $N \to \infty$. Most of the work in extreme value theory has been done for independent random variables(cf. Resnick 1987, de Haan and Ferreira 2006), and suitable results from extreme value theory are absent for our setting, rendering the analysis of extremes in the dependent case challenging.

1.3. New Extreme Value Limit

Our answer to this challenge is somewhat surprising: in Theorem 5.1, we prove that the scaled maximum queue length converges to a normally distributed random variable as $N \to \infty$. In particular, if $Q_i(\infty, \beta)$ is the invariant queue length at node i,

$$\frac{\max_{i \leq N} Q_i(\infty, \beta) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma \sigma_A}{\sqrt{2}\beta}X, \tag{1}$$

with X standard normal. An intuitive explanation of this result is the following. Using Lindley's recursion, we can write the maximum queue length as a maximum of N suprema. By using subadditivity arguments, we can separate the independent and dependent parts; the independent part converges using standard extreme value results, whereas the dependent part satisfies a central limit theorem. To the best of our knowledge, we are the first who prove a result of this type. A consequence of this convergence result is 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, whereas this quantile function also appears in the limit of the optimal capacity.

1.4. Numerical Experiments

In Section 5.3, numerical experiments show that we typically are most of the time 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 zero as $N \to \infty$; cf. Theorem 5.2. We give an improvement of this approximation by combining our results for deterministic 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.

1.5. Limitations of Simulation

In Section 5.3, we explain the simulation procedure in the case of stochastic demand. We aim to approximate the maximum queue length of the all-time supremum of N dependent Brownian motions. Because the dependence structure between two all-time suprema of Brownian motions is complicated, we cannot resort to an easy simulation procedure, for example, by using copulas. We, namely, need to simulate discretized approximations of all of these N Brownian paths. Subsequently, we need to cut the Brownian path at some finite time point. We then record the largest observations of all of these paths. Subsequently, we compute the maximum of N of these records to obtain one observation of a maximum queue length. Afterward, we need to repeat this procedure to collect data. Finally, we use the collected data to compute empirical means and to estimate quantile functions. This means that the computation time grows with at least N, the size of the fork-join queue. Besides, in this simulation procedure, a lot of discretization and approximation steps are needed, which increase the error. Though the simulation results give a clear indication of the convergence rate of our limit theorem for small fork-join

queueing networks, clearly the procedure is unworkable for a system with a number of servers of the order of thousands, which, as a matter of fact, shows the usefulness of the limit in Theorem 5.1 as an approximation.

1.6. Summary of Results

In this paper, we study an assembly system with N components, in which the demand and the number of produced components are deterministic with some random perturbation, which is assumed to be normally distributed. Thus, the total delay for one component in steady state can be modeled by the all-time supremum of a Brownian motion. We model the system as a fork-join queue. We then use results from EVT to estimate the longest queue, and we minimize the total costs in the system using this approximation (cf. Theorems 4.1, 5.1, and 5.2 for the most important results).

1.7. New Insights

This paper generates new insights in fork-join queues that 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 multicomponent assembly system with dependent delays. Because of the dependencies in delays, evaluating such a system with fixed capacity and inventory is already a difficult problem. We provide several asymptotically optimal expressions for capacity and inventory that are either in closed form or can easily be computed numerically. Our results may help OEMs to optimally allocate budget to capacity and inventory to cost-efficiently ensure timely deliveries to their customers.

1.8. Overview

The remainder of this paper is organized as follows. In Section 2, we provide an overview of relevant literature. We introduce the general mathematical model in Section 3 and subsequently present the optimization problem in which 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. A refinement of the approximations from Section 5 is provided in Section 6, in which we combine the lessons learned in Sections 4 and 5 to obtain better approximations for optimal capacity and inventory. In Section 7, we briefly touch upon the case of asymmetric systems and demonstrate that, even in these settings, our result for symmetric systems remain useful. We give a summary and conclusions in Section 8 and provide most of the proofs in Appendix A.

2. Literature Review

Simultaneous optimization of capacity and inventory is an important problem in supply chain management, but the literature on this topic is limited because of the complexity of the problem (Bradley and Glynn 2002). Considering the interaction between a manufacturer and a single supplier, Chaturvedi and Martínez-de Albéniz (2016) discuss the trade-off between inventory and capacity and how properly diversifying supply sources can reduce inventory and capacity investments. 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 multiserver 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.

In particular, we derive extreme value results for multicomponent assembly systems as the number of components grows large in order to obtain asymptotically optimal capacity and inventory decisions. We are not aware of related studies of extreme values for inventory and capacity optimization, but the approach is conceptually related to studies that apply asymptotic analysis to analyze inventory control problems, and we next review this literature. Such studies typically analyze inventory models that are inherently high dimensional; asymptotic analysis may be used to derive much simpler optimization problems that form an accurate approximation in some relevant asymptotic regime. This approach has led to major progress in the analysis of inventory problems, for example, for lost sales models (Goldberg et al. 2016, Xin and Goldberg 2016), dual sourcing (Xin and Goldberg 2018), and assembly-to-order systems (Reiman and Wang 2015, Doğru et al. 2017) in the presence of large lead times. Assemble-to-order systems with high-volume demand are studied by Plambeck (2008) and Plambeck

and Ward (2008), whereas Zhang et al. (2020) study policies for managing perishable inventory when the market size grows large. A comprehensive overview of advances using asymptotic analysis can be found in Goldberg et al. (2021). Whereas conceptually related, our analysis differs substantially because a queueing model rather than a Markov decision process underlies our problem, and we aim to analyze extremes in the queueing model to optimize certain model parameters. In that sense, our work is related to Glasserman (1997), who provides approximations for setting base-stock levels in single-stage and multistage systems that are asymptotically exact as the target service level or the backorder penalty becomes large. For single-product lost sales inventory systems under periodic review, Huh et al. (2009) show that order-up-to policies are asymptotically optimal when the lost sales penalty is large compared with the holding cost. Bijvank et al. (2014) show the robustness of this result when using the optimal base-stock levels of the corresponding backorder system instead of those of the lost sales system. The asymptotic analysis in this paper is also influenced by related problems for queues with many servers inspired by agent staffing problems in call centers; we refer to Borst et al. (2004), Gans et al. (2003), and van Leeuwaarden et al. (2019) for background.

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, section 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, section 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 because, 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.

We note that our study focuses on the common components of a single high-tech system, which is a considerably simpler problem than general assemble-to-order problems (cf. Atan et al. 2017). Our focus enables us to obtain results for the key trade-off between capacity, inventory, and delivery reliability, sidestepping the difficulties of inventory control in multiproduct assemble-to-order systems with component commonality (see, e.g., Song 1998, Lu and Song 2005, Reiman and Wang 2015, Atan et al. 2017).

Literature concerning simultaneous optimization of capacity and inventory in single-sourced assembly (or assembly-to-order) 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 in which one supplier has uncertainty in the yield. Our main contribution, compared with 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.

To analyze the problem at hand, we examine fork-join queueing networks with N servers for which 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 because of 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, in which $N \to \infty$.

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 and Zhong (2016) give the solution of the Laplace transform of joint first passage times in terms of the solution of a partial differential equation, for which 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 are also studied; cf. Dębicki et al. (2015), in which 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 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 in which the number of servers is finite, whereas we investigate a fork-join queue in which 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 multidimensional reflected Brownian motions. Nguyen (1994) extends this result to a fork-join queue with multiple job types. Lu and Pang (2015, 2017a, b) study fork-join networks. In Lu and Pang (2015), they investigate a fork-join network in which 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, in which 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, for which 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.

3. Model and Preliminaries

The production system of OEMs such as ASML, Philips, or Airbus consists of roughly two stages: (1) component production and (2) assembly/integration of components. This setup is crucial to enable the modular design, production, and testing of components, and substantial value is added in both stages. For these reasons, system integration is only initiated after customers have committed to purchasing the system. We consider a manufacturing system in which a manufacturer assembles a final product from N common components, in which N is a large number, meaning that all components are required whenever a product is assembled. Each component is produced on a single production line that involves highly skilled staff and specialized equipment. In anticipation of uncertain demand, an inventory buffer is built up: production continues until a target inventory position is reached, after which production is switched off until the inventory position drops below this target. Such base-stock policies are widely used for modeling component inventories (e.g., Akçay and Xu 2004, Bollapragada et al. 2004, Karsten et al. 2012). Also, in a high-tech manufacturing environment, in which capacity mainly refers to people working in cleanrooms that can be at work or have a day off instead of expensive machines with high start-up costs, such policies are suitable. Despite these inventory buffers, random delays may occur in the production process for each of the components.

3.1. Model

We adopt a symmetric, continuous-time model and assume that production capacity in every finite time interval is normally distributed, meaning that cumulative production is a Brownian motion with drift. We then look at this system in equilibrium and find a trade-off between investing in the base-stock buffer and investing in capacity. 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 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 analyze the cost-minimization problem, we model this assembly system by a fork-join network of N statistically identical but possibly correlated queues. Demand is represented by the common arrival process of jobs going to each server, and each server, with independent, identical service processes, represents production of a component. 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. As demand at each server is driven

by a common arrival process, the total inventory of a component, including the number of backlogged components is equal for all components. However, as the service times vary, the division between the number of finished components and the number of backlogged components may vary per server. When all servers have a finished component in their warehouse, the end product can be assembled. This system is visualized in Figure 1.

3.2. Brownian Fork-Join Queue

We model queue lengths as reflected Brownian motions, following Harrison (1985) and Abate and Whitt (1987). Other papers using Brownian queues to analyze assembly systems are, for example, Plambeck (2008) and Plambeck and Ward (2008).

Definition 3.1. For all $i \leq N$, the service process at server i is governed by the Brownian motion $\{W_i(t), t \geq 0\}$ with standard deviation σ, and the arrival process is governed by the Brownian motion $\{W_A(t), t \geq 0\}$ with standard deviation σ_A. The queue length at server i at time $t > 0$ equals

$$Q_i(t,\beta) := \sup_{0 < s < t} ((W_i(t) + W_A(t) - \beta t) - (W_i(s) + W_A(s) - \beta s)), \tag{2}$$

with $Q_i(0,\beta) = 0$. For $i, j \leq N$ with $i \neq j$ the Brownian motions $\{W_i(t), t \geq 0\}$ and $\{W_j(t), t \geq 0\}$ are independent and identically distributed.

Formally, the Brownian motions $\{W_i(t), t \geq 0\}$ and $\{W_A(t), t \geq 0\}$ represent fluctuations in the service and arrival processes as they have zero mean. The controllable parameter β represents the excess capacity in each individual queue.

3.3. Base-Stock Level and Capacity

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, that is, the difference between the production rate and arrival rate; in other words, β_i captures the capacity investment of server i. As mentioned before, we assume that, for all servers, the net capacity and the base-stock levels are the same; thus, $\beta_i = \beta_j = \beta$ and $I_i = I_j = I$. The backlog $Q_i(t,\beta)$ represents the number of outstanding orders of component $i \leq N$ at time t with $Q_i(t,\beta)$ given in Definition 3.1. If $\sigma_A^2 > 0$, $(Q_i(t,\beta))_{i \leq N}$ are dependent random variables.

3.4. Transient Inventory Levels and Backorders

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 and, second, the committed inventory that consists of items that are committed to realized demand but put aside because other components are not yet available. That is, the excess supply of component i is given by $(I - Q_i(t,\beta))^+$. Moreover, the number of backorders for component i at time t is equal to $(Q_i(t,\beta) - I)^+$ because, for $Q_i(t,\beta) \leq I$, the shortage is compensated by inventory I, and only the part of $Q_i(t,\beta)$ exceeding I represents actual backorders that cannot be satisfied. Because 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 $\max_{j \leq N}(Q_j(t,\beta) - I)^+$. Therefore, the committed

Figure 1. Fork-Join Queue
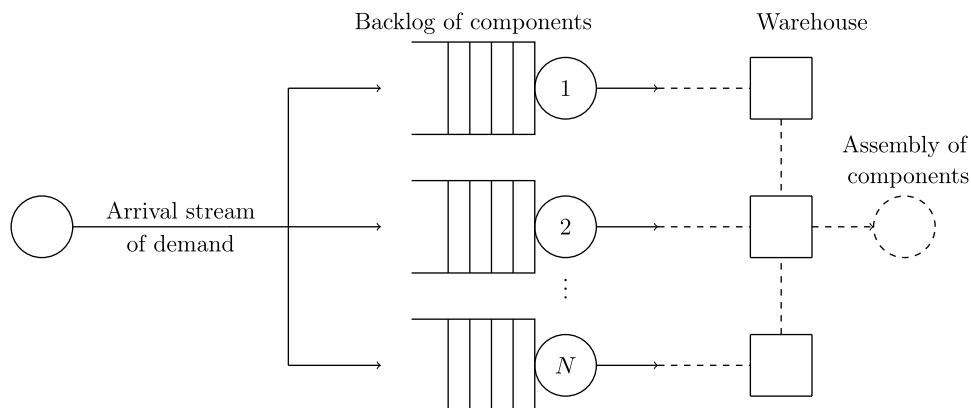
inventory of component i equals the number of backorders in the system minus its own backlog and can be expressed as $\max_{j \leq N}(Q_j(t,\beta) - I)^+ - (Q_i(t,\beta) - I)^+$. The total inventory of component i at time t is, thus, given by

$$I_i(t) = (I - Q_i(t,\beta))^+ + \max_{j \leq N}(Q_j(t,\beta) - I)^+ - (Q_i(t,\beta) - I)^+ = I - Q_i(t,\beta) + \max_{j \leq N}(Q_j(t,\beta) - I)^+, \quad (3)$$

with $I_i(0) = I$. Observe that the total inventory $I_i(t)$ at time t is a function of the number of outstanding orders at time t. The reason why this is true is that the random variable $Q_i(t,\beta)$ does not depend on the total inventory because the servers always produce when there is an incoming task irrespective of whether there are items in stock or not. When there are items in stock, the product is immediately assembled, but servers work in order to reach the target inventory. When there are no items in stock, servers work to finish their component. Hence, whether a server works does not depend on the total inventory, but only on the demand and their own service speed. This means that the total inventory at time t is described as the function given in Equation (3). Thus, in order to know the total inventory on a certain time t, one should know the number of outstanding orders on that given time t when the dynamics of these outstanding orders are described as the dynamics of reflected Brownian motions until time t. Thus, this describes the dynamics of the system.

3.5. Steady-State Limit

Because the backlogs are modeled as reflected Brownian motions with negative drift, the backlogs have a steady-state limit. This limit extends to the largest backlog in the system and the total inventory of component i. We prove this in Lemma 3.1.

Lemma 3.1 (Steady State of Backlogs). *Given $(Q_i(t,\beta), i \leq N)$ with $Q_i(t,\beta)$ defined in (2), we have that $(Q_i(t,\beta), i \leq N) \xrightarrow{d} (Q_i(\infty,\beta), i \leq N)$ with*

$$\left(Q_i(\infty,\beta), i \leq N\right) \overset{d}{=} \left(\sup_{s>0}(W_i(s) + W_A(s) - \beta s), i \leq N\right). \quad (4)$$

In particular,

$$\max_{i \leq N} Q_i(\infty,\beta) \overset{d}{=} \max_{i \leq N} \sup_{s>0}(W_i(s) + W_A(s) - \beta s). \quad (5)$$

Proof. The argument in one dimension is standard (see, e.g., section III.6 of Asmussen 2003); we extend it to our setting. Given $t > 0$, we can define Brownian motions $\{\hat{W}_A(s), s \geq 0\}$ and $\{\hat{W}_i(s), s \geq 0\}$ that satisfy $\hat{W}_A(t-s) = W_A(t) - W_A(s)$ and $\hat{W}_i(t-s) = W_i(t) - W_i(s)$. From this, it follows that, for fixed $t > 0$, we have that

$$(Q_i(t,\beta), i \leq N) = \left(\sup_{0 \leq s \leq t}(\hat{W}_i(s) + \hat{W}_A(s) - \beta s), i \leq N\right)$$

$$\overset{d}{=} \left(\sup_{0 \leq s \leq t}(W_i(s) + W_A(s) - \beta s), i \leq N\right).$$

Now, we obtain the lemma by letting $t \to \infty$, using monotone convergence. $\quad\square$

Combining this result with (3), we obtain an analogous result for the steady-state total inventory. In particular,

$$\sum_{i=1}^{N} I_i(t) \xrightarrow{d} \sum_{i=1}^{N} \left(I - Q_i(\infty,\beta) + \max_{j \leq N}(Q_j(\infty,\beta) - I)^+\right).$$

From now on, we write $Q_i(\beta) := Q_i(\infty,\beta)$.

3.6. Cost Function

We scale the cost of building net capacity to one and let $h^{(N)}$ and $b^{(N)}$ denote (inventory) holding costs and back-order costs, respectively, which may depend on N. Our goal is to minimize the expected total costs of the system in steady state.

Definition 3.2. We define

$$C_N(I, \beta) := \mathbb{E}\left[\sum_{i \leq N}\left[h^{(N)}\left(I - Q_i(\beta) + \max_{j \leq N}(Q_j(\beta) - I)^+\right)\right] + b^{(N)}\max_{j \leq N}(Q_j(\beta) - I)^+\right], \tag{6}$$

with the distribution of $Q_i(\beta)$ given in Equation (5).

Equation (6) simplifies to

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

Then, the expected total costs in the system are equal to $C_N(I, \beta) + \beta N$, where the term βN reflects our normalization of unity net capacity costs per queue. If this term were removed, it would be optimal to choose $\beta = \infty$ and $I = 0$.

Because of the self-similarity of Brownian motion, we can write

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

$$\stackrel{d}{=} \max_{i \leq N} \sup_{t > 0}(W_A(t) + W_i(t) - t).$$

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

$$\min_{(I, \beta)}(C_N(I, \beta) + \beta N) = \min_{(I, \beta)}\left(\frac{1}{\beta}C_N(I\beta, 1) + \beta N\right) = \min_{(I, \beta)}\left(\frac{1}{\beta}C_N(I, 1) + \beta N\right). \tag{7}$$

In the last part of Equation (7), I has the interpretation of the base-stock level at which the net capacity $\beta = 1$. Therefore, from now on, the actual number of products on stock at time 0 equals I/β. 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 3.3.

Definition 3.3. We define

$$F_N(I, \beta) := C_N(I, \beta) + \beta N = \frac{1}{\beta}C_N(I) + \beta N, \tag{8}$$

with $C_N(I) := C_N(I, 1)$ and $C_N(I, \beta)$ given in Equation (6).

Our goal is to solve $\min_{(I, \beta)} F_N(I, \beta)$, focusing on the case in which N is large.

3.7. Preliminary Results
As we have defined the Brownian fork-join queue and the corresponding cost functions, we now state some general results that are valid regardless of whether $\sigma_A = 0$ or $\sigma_A > 0$. In the next lemma, we show that we can write $\min_{(I, \beta)} F_N(I, \beta)$ 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 base-stock level* 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.

Using Lemma 3.2, we can characterize the optimal net capacity and base-stock level. In Lemma 3.3, we provide expressions for the optimal net capacity and costs in terms of the optimal base-stock level, 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 $\max_{i \leq N} Q_i$.

Lemma 3.4. *The optimal base-stock level* 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 $N \to \infty$.

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

This first order approximation is valid regardless of 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

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

The optimal value $\overline{I}_N$ for the associated first order minimization problem $\min_I \overline{C}_N(I)$ is given by $\overline{I}_N = \frac{\sigma^2}{2}\log N$ because $b^{(N)} > 0$. Using this approximation, we see that $C_N(\overline{I}_N) \approx \overline{C}_N(\overline{I}_N) = (1 + o(1))\frac{\sigma^2}{2}Nh^{(N)}\log N$, $\overline{\beta}_N = \sqrt{\overline{C}_N(\overline{I}_N)/N} = (1 + o(1))\sqrt{\frac{\sigma^2}{2}h^{(N)}\log N}$, and $F_N(\overline{I}_N, \overline{\beta}_N) \approx 2\sqrt{N}\sqrt{\frac{\sigma^2}{2}Nh^{(N)}\log N}$. These results can be made rigorous, and the decision rule $\overline{I}_N$ can be shown to be asymptotically optimal, that is, that $F_N(\overline{I}_N, \overline{\beta}_N) = F_N(I_N^*, \beta_N^*)(1 + o(1))$. 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 $\gamma_N \xrightarrow{N \to \infty} 1$. Then,*

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

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 in which 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, whereas the latter case requires the development of a novel limit theorem. Furthermore, we use the result given in Corollary 3.1; this corollary shows how the ratio between the optimal costs and approximate costs can be represented when the approximate base-stock level and net capacity are solutions to a minimization problem as well. This corollary follows trivially from Lemma 3.3.

Corollary 3.1. *Assume we have a function $\tilde{F}_N(I, \beta) : (0, \infty) \times (0, \infty) \to \mathbb{R}$. 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, so 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

In this section, we consider the case in which demand is deterministic. From this, it follows that all N queues are mutually independent.

4.1. Solution and Convergence of the Minimization Problem

We now analyze the minimization of the cost function described in Definition 3.3 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 because we have the expression $\max_{i \leq N} Q_i$ in this function. In Lemma 4.1, we give the optimal base-stock level 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 $n > 0$. Second, we consider the quality-driven regime, in which $\gamma_N \overset{N \to \infty}{\to} 0$. Finally, we investigate the efficiency-driven regime, in which $\gamma_N \overset{N \to \infty}{\to} 1$. 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 zero and variance σ^2. Let $h^{(N)}$ and $b^{(N)}$ be positive sequences. In order to minimize $F_N(I, \beta)$, the optimal base-stock level I_N^* satisfies,*

$$I_N^* = P_N^{-1}(1 - \gamma_N) = \frac{\sigma^2}{2}\log\left(\frac{1}{1 - (1 - \gamma_N)^{\frac{1}{N}}}\right), \tag{9}$$

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 approximate the function F_N. Because $(Q_i, i \leq N)$ are independent and exponentially distributed, we know by standard extreme value theory (cf. de Haan and Ferreira 2006) that $\frac{2}{\sigma^2}\max_{i \leq N} Q_i - \log N \overset{d}{\to} G$ as $N \to \infty$ with $G \sim$ Gumbel. Therefore, for N large, $\max_{i \leq N} Q_i \overset{d}{\approx} \frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N$. 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 4.1, we give the resulting function $\hat{F}_N(I, \beta)$ that is to be minimized.

Definition 4.1.

$$\hat{C}_N(I) := \mathbb{E}\left[Nh^{(N)}(I - Q_i) + (Nh^{(N)} + b^{(N)})\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N - I\right)^+\right], \tag{10}$$

and

$$\hat{F}_N(I, \beta) := \frac{1}{\beta}\hat{C}_N(I) + \beta N. \tag{11}$$

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 of the cost function F_N defined in Definition 3.3, and we write $(\hat{I}_N, \hat{\beta}_N)$ for the minimizers of the cost function $\hat{F}_N$ defined in Definition 4.1. Throughout this paper, we indicate second order approximations by the $\wedge$-symbol.

In Proposition 4.1, we present the base-stock level that minimizes $\hat{F}_N$. This base-stock level turns out to be a quantile of $\frac{\sigma^2}{2}G$ added to $\frac{\sigma^2}{2}\log N$.

Proposition 4.1 (Approximation). *Minimizing $\hat{F}_N(I, \beta)$ with $G \sim$ Gumbel gives solution $(\hat{I}_N, \hat{\beta}_N, \hat{F}_N(\hat{I}_N, \hat{\beta}_N))$ with*

$$\hat{I}_N = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log(-\log(1 - \gamma_N)), \tag{12}$$

and

$$\hat{C}_N(\hat{I}_N) = Nh^{(N)}\left(\hat{I}_N - \frac{\sigma^2}{2}\right) + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\left(\int_{-\log(1-\gamma_N)}^{\infty}\frac{e^{-t}}{t}dt + \Gamma + \log(-\log(1-\gamma_N))\right),\tag{13}$$

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

Combining Equations (12) and (13) 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 base-stock level and net capacity with the costs under the approximate base-stock level and net capacity. We distinguish the balanced, quality-driven, and efficiency-driven regimes.

By using the results from Lemmas A.1 and A.2 in Appendix A.2, we prove the order bounds in the balanced, quality-driven, and efficiency-driven regimes in Theorem 4.1. In the efficiency-driven regime, we impose the additional condition $\gamma_N < 1 - \exp(-N)$ needed to make sure that $\hat{I}_N > 0$. If we, namely, choose $\gamma_N > 1 - \exp(-N)$, we get that $\hat{I}_N < 0$, which is not feasible because $\hat{I}_N$ has the physical meaning of the number of items that needs to be stored.

Theorem 4.1 (Order Bounds). *Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$ if $\gamma_N = \gamma \in (0,1)$ in the balanced regime. Then,*

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(1/(N\log N)),\tag{14}$$

if $\gamma_N \overset{N\to\infty}{\to} 0$ in the quality-driven regime. Then,

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(\gamma_N/(N\log(N/\gamma_N))),\tag{15}$$

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

$$\frac{F_N(I_N^*, \beta_N^*)}{F_N(\hat{I}_N, \hat{\beta}_N)} = 1 - O(1/\log N).\tag{16}$$

Using the order bounds given in Theorem 4.1, we can establish for the three different regimes how $F_N(I_N^*, \beta_N^*)$ scales with N as N becomes large.

Lemma 4.2. *Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$ if $\gamma_N = \gamma \in (0,1)$ in the balanced regime. Then,*

$$F_N(I_N^*, \beta_N^*)$$
$$= 2\sqrt{N}\sqrt{Nh^{(N)}\frac{\sigma^2}{2}(\log N - \log(-\log(1-\gamma)) - 1) + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\mathbb{E}[(G + \log(-\log(1-\gamma)))^+]}$$
$$\quad + O(\sqrt{h^{(N)}}/\sqrt{\log N}),\tag{17}$$

if $\gamma_N \overset{N\to\infty}{\to} 0$ in the quality-driven regime. Then,

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh^{(N)}\frac{\sigma^2}{2}(\log(N/\gamma_N) - 1) + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\gamma_N}$$
$$\quad + O\left(\gamma_N\sqrt{h^{(N)}}/\sqrt{\log(N/\gamma_N)}\right),\tag{18}$$

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

$$F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh^{(N)}\frac{\sigma^2}{2}(\log N - 1) + b^{(N)}\frac{\sigma^2}{2}\log(-\log(1-\gamma_N))} + O(N\sqrt{h^{(N)}}/\sqrt{\log N}).\tag{19}$$

The results given in Theorem 4.1 and Lemma 4.2 are obtained by using the properties stated in Online Lemmas A.1 and A.2. In Online Lemma A.1, 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 on $|C_N(I_N^*) - C_N(\hat{I}_N)|$ and $|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|$ in Online Lemma A.2, which leads to sharp results in Theorem 4.1 and Lemma 4.2.

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 1,000. The results for the balanced, quality-driven, and efficiency-driven regimes are given in Tables 1–3, 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 one.

5. Stochastic Demand

We now extend our framework to the case in which 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. Because 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.

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

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 5.1, 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 5.1. *Let $(W_i, 1 \leq i \leq N)$ be independent Brownian motions with mean zero and variance σ^2 and W_A be a Brownian motion with mean zero and variance σ_A^2. Then,*

$$\frac{\max_{i \leq N} \sup_{s>0}(W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma \sigma_A}{\sqrt{2}\beta} X, \qquad (20)$$

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

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

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

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	23.9296	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	1,838.5	3.29056	1.83846	1,838.5	$3.91 \cdot 10^{-5}$
1,000	3.63731	1.93044	3,860.87	3.63713	1.93042	3,860.87	$1.97 \cdot 10^{-5}$

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	2,493.06	6.21511	2.49306	2,493.06	$5.51467 \cdot 10^{-6}$
1,000	6.90801	2.62833	5,256.66	6.90801	2.62833	5,256.66	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	62.4616	0.000797
50	0.965997	9.35451	935.451	0.927385	9.34122	935.452	$8.65678 \cdot 10^{-6}$
100	1.21527	14.4701	2,894.02	1.19242	14.4615	2,894.02	$1.30518 \cdot 10^{-6}$
200	1.48208	22.0864	8,834.57	1.46889	22.0808	8,834.57	$2.20863 \cdot 10^{-7}$
500	1.85348	38.0553	38,055.3	1.84728	38.0521	38,055.3	$2.51171 \cdot 10^{-8}$
1,000	2.14443	56.945	113,890	2.14098	56.9428	113,890	$5.30189 \cdot 10^{-9}$

A heuristic explanation of the result in Theorem 5.1 is as follows: though $(Q_i, i \leq N)$ are dependent random variables, because we are adding the same Brownian motion W_A, $\max_{i \leq N} W_i(s)$ dominates 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 the 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, $W_A(\tau) \overset{d}{\approx} \frac{\sigma\sigma_A}{\sqrt{2\beta}}\sqrt{\log N}X$ with X standard normally distributed, which results in Equation (20).

The proof of Theorem 5.1 consists of four parts, which are stated in Lemmas 5.1–5.4 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 5.1 for appropriately chosen t, t_1, t_2. More specifically, in Lemma 5.1, 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 5.2 and 5.3, respectively, leading up to Lemma 5.4 that shows the concentration around the critical time $d \log N$, proving a tight upper bound.

Lemma 5.1. *For* $d = \frac{\sigma^2}{2\beta^2}$,

$$\frac{\max_{i \leq N}(W_i(d \log N) + W_A(d \log N)) - \beta d \log N - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma\sigma_A}{\sqrt{2\beta}}X, \tag{21}$$

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

Lemma 5.2. *For $d = \frac{\sigma^2}{2\beta^2}$ and $0 < \epsilon < d$ and for all x,*

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

Lemma 5.3. *For $d = \frac{\sigma^2}{2\beta^2}$ and all $\epsilon > 0$ and $x \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{\max_{i \leq N} \sup_{s \geq (d+\epsilon)\log N}(W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \geq x\right) \overset{N \to \infty}{\to} 0. \tag{23}$$

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

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

$$\leq \mathbb{P}\left(\sigma_A\sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon}X_1 + \sqrt{2\epsilon}\sigma_A|X_2| > x\right), \tag{24}$$

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 5.1. In Lemma 5.5, we prove that convergence holds even in L_1 when X is chosen appropriately.

Lemma 5.5. *Define $X_N := \frac{\sqrt{2}\beta}{\sigma\sigma_A}\frac{W_A(\frac{\sigma^2}{2\beta^2}\log N)}{\sqrt{\log N}}$. Then,*

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

The proof of Lemma 5.5 is also given in Appendix A.3. In the next section, we apply Theorem 5.1 and Lemma 5.5 to solve and approximate the minimization problem. Specifically, Lemma 5.5 gives us an order bound between the optimal base-stock level and the approximate base-stock level.

5.2. Solution and Convergence of the Minimization Problem

We can use the convergence result proven in Theorem 5.1 to prove asymptotics of the minimization of the function F_N. Because $\frac{\sqrt{2}\beta}{\sigma\sigma_A}\frac{\max_{i \leq N} Q_i(\beta) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$ 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, lemma 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 5.2, we can solve the minimization problem, which yields the optimal base-stock level and net capacity given in Lemma 5.6. The proofs concerning the solution and subsequent convergence results are provided in Appendix A.4.

Definition 5.2. We define

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

Lemma 5.6. *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,*

$$I_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}P_N^{A-1}(1 - \gamma_N)\sqrt{\log N}. \tag{25}$$

When we are in the balanced regime, we can approximate the minimization problem given in Definition 5.1, using the convergence result in Theorem 5.1, and prove how far the approximate solution is from the optimal solution. This is done in Proposition 5.1 and Theorem 5.2. In Lemma 5.7, 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 5.1. *For* $(b^{(N)})_{N \geq 1}, (h^{(N)})_{N \geq 1}$ *and* $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$,

$$\hat{I}_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma_N), \tag{26}$$

and

$$\hat{C}_N^A(\hat{I}_N^A) = Nh^{(N)} \left(\frac{\sigma^2}{2} \log N - \frac{\sigma^2 + \sigma_A^2}{2} \right) + (Nh^{(N)} + b^{(N)}) \frac{\sigma \sigma_A \sqrt{\log N} e^{-\frac{1}{2}\Phi^{-1}(1-\gamma_N)^2}}{2\sqrt{\pi}}. \tag{27}$$

Theorem 5.2 (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 5.7 (Balanced Regime). *Assume* $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$ *with* $\gamma_N = \gamma \in (0, 1)$. *Then,*

$$I_N^A = \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} \Phi^{-1}(1 - \gamma) + o(\sqrt{\log N}), \tag{28}$$

and

$$F_N(I_N^A, \beta_N^A) = 2\sqrt{N} \sqrt{\hat{C}_N^A(\hat{I}_N^A)} + o(N\sqrt{h^{(N)}}). \tag{29}$$

The result in Lemma 5.7 only holds for the balanced regime, so a natural question is what we can say about the efficiency- and quality-driven regimes. 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 one. Thus, we expect that the approximation given in (26) also satisfies this convergence result. In order to determine whether this approximation also satisfies the order bound given in Theorem 5.2, a further analysis is needed. The analysis we provide for the balanced regime heavily relies on van der Vaart (1998, lemma 21.2), which says that, if $Y_N \xrightarrow{d} Y$, then for $\gamma \in (0, 1)$, $P_{Y_N}^{-1}(\gamma) \xrightarrow{N \to \infty} P_Y^{-1}(\gamma)$. This gives us the convergence result (28) of the inventory in the balanced regime. In order to be able to prove a similar result for the efficiency-driven regime, we need an improvement of van der Vaart (1998, lemma 21.2), which also holds when $\gamma_N \xrightarrow{N \to \infty} 1$.

However, for the quality-driven regime, this convergence result does not hold because we see in Lemma 4.2 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 5.2, 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 base-stock level. 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 base-stock level and under these asymptotic approximations. Because 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 3.3, 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, section 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. Because 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 that 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 this 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 5.2. To obtain a median-unbiased estimate of the quantile, we use the approach suggested by Zieliński (2009). 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 base-stock level as given in Equation (25). Using the optimal base-stock level, we determine the optimal net capacity given in

Lemma 3.3. Because 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 base-stock level (provided in Proposition 5.1) to the costs under the optimal net capacity and base-stock level 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 $F_N(I, \beta)$.

The procedure described is applicable for N in the order of hundreds; however, it is close to impossible to provide a fast simulation for N in the order of thousands. Hence, to give a useful approximation of the optimal capacity and base-stock level in these cases, we need to use the limit we derived in Theorem 5.1.

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.

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, for which 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 5.2. 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 $\sigma_A = 0.1$, for which 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 base-stock level and net capacity and the costs under the approximations are around 10%. Furthermore, we can 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 and stochastic demand, the approximate problems are given in Definitions 4.1 and 5.1. If σ_A is small, then we know that, on the one hand,

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

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

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

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

$$\max_{i \leq N} Q_i \stackrel{d}{\approx} \frac{\sigma^2}{2} \log N + \frac{\sigma \sigma_A}{\sqrt{2}} \sqrt{\log N} X + \frac{\sigma^2}{2} G. \tag{30}$$

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

When we replace $\max_{i \leq N} Q_i$ with Equation (30) in the minimization problem, we get

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

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,

$$\int_{-\infty}^{\infty} \exp\left(-\exp\left(-\frac{2}{\sigma^2}\left(I_N^M - \frac{\sigma^2}{2}\log N - \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}x \right) \right) \right) \phi(x)dx = 1 - \gamma_N. \tag{31}$$

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

$$\beta_N^M = \frac{\sqrt{\mathbb{E}\left[Nh^{(N)}(I_N^M - Q_i) + (Nh^{(N)} + b^{(N)})\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M \right)^+ \right]}}{\sqrt{N}}. \tag{32}$$

The relevant expectations in this symbolic expression can be computed numerically; see Appendix A.5 for details.

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, in which the last column repeats the results from Section 5.3, we observe that the mixed-behavior approximations show better

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

Table 7. Comparison of Costs Master Solution for $h^{(N)} = 1$, $b^{(N)} = N$

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right)\sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right)\sqrt{\log N}$
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

convergence also when σ_A is larger. Furthermore, when we saw in Section 5.3 that the cost difference increased considerably with the change in $b^{(N)}$, we now do see a slight 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.

7. Analyzing Asymmetric Systems

This paper derives several new, analytical results for joint capacity and inventory optimization for large-scale, symmetric assembly systems. In this section, we provide an informal discussion of the application of such results in asymmetric settings.

For ease of exposition, consider a case in which different components have different holding costs. For other parameters, our assumptions remain in place. In practical settings, component prices might range from a few thousand to hundreds of thousands of euros. Companies seeking to apply advanced methods for optimizing capacity and inventory investments focus on the most expensive components: for inexpensive components, some coarse heuristics would suffice.

Suppose the company seeks to derive separate inventory buffer and capacity rules for two groups of components: expensive and very expensive components. This yields $k = 2$ groups of components. We seek to apply our results on extremes as the total number of components N in these two groups grows large; we keep k and the ratio of components in the two groups fixed. Also, because we seek to derive rules at the group level, it makes sense to assume symmetry within groups, that is, by averaging cost parameters within the groups. For example, consider the following: $N/2$ servers have a holding cost $h_1^{(N)}$ and $N/2$ servers have a holding cost $h_2^{(N)}$. Then, we

Table 8. Comparison of Costs Master Solution for $h^{(N)} = 1$, $b^{(N)} = 3N$

N	σ_A	I_N^M	β_N^M	$F_N(I_N^M, \beta_N^M)$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(I_N^M, \beta_N^M)}\right)\sqrt{\log N}$	$\left(1 - \frac{F_N(I_N^A, \beta_N^A)}{F_N(\hat{I}_N^A, \hat{\beta}_N^A)}\right)\sqrt{\log N}$
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

Table 9. Comparison of Optimal Costs and Costs Under Upper Bound Heuristic, $\sigma = 1$, $\sigma_A = 0$

N	$h_1^{(N)}$	$h_2^{(N)}$	Ratio	$b^{(N)}$	Optimal	Heuristic	Difference, %
10	1	10	1:1	10	42.3 ± 0.1	42.9 ± 0.1	0.14
100	1	10	1:1	100	615.6 ± 1.2	617.4 ± 1.0	0.3
1,000	1	10	1:1	1,000	$7,597.9 \pm 8.2$	$7,643.0 \pm 7.8$	0.6
10	10	100	1:1	1	126.0 ± 0.4	127.0 ± 0.4	0.7
100	100	1,000	1:1	1	$5,967 \pm 10.9$	$6,002 \pm 9.6$	0.6
1,000	1,000	10,000	1:1	1	$236,063 \pm 256$	$236,402 \pm 233$	0.1
10	1	10	1:3	10	53.1 ± 0.2	53.2 ± 0.2	0.2
100	1	10	1:3	100	770.5 ± 1.3	772.9 ± 1.2	0.3
1,000	1	10	1:3	1,000	$9,551.1 \pm 10.7$	$9,581.6 \pm 9.5$	0.3

need to minimize

$$\frac{N}{2}\left(h_1^{(N)}\frac{1}{\beta_1}\left(I_1 - \frac{\sigma^2}{2}\right) + \beta_1\right) + \frac{N}{2}\left(h_2^{(N)}\frac{1}{\beta_2}\left(I_2 - \frac{\sigma^2}{2}\right) + \beta_2\right)$$

$$+ \left(\frac{N}{2}h_1^{(N)} + \frac{N}{2}h_2^{(N)} + b^{(N)}\right)\mathbb{E}\left[\max\left(\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1), \frac{1}{\beta_2}\max_{N/2+1 \le i \le N}(Q_i(1) - I_2)\right)^+\right]. \tag{33}$$

This is over $(I_1, I_2, \beta_1, \beta_2)$. Obviously,

$$\mathbb{E}\left[\max\left(\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1), \frac{1}{\beta_2}\max_{N/2+1 \le i \le N}(Q_i(1) - I_2)\right)^+\right]$$

$$\le \mathbb{E}\left[\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1)^+\right] + \mathbb{E}\left[\frac{1}{\beta_2}\max_{i \le N/2}(Q_i(1) - I_2)^+\right].$$

The cost function in Equation (33) can, therefore, be bounded from above by

$$\frac{N}{2}\left(h_1^{(N)}\frac{1}{\beta_1}\left(I_1 - \frac{\sigma^2}{2}\right) + \beta_1\right) + \frac{N}{2}\left(h_2^{(N)}\frac{1}{\beta_2}\left(I_2 - \frac{\sigma^2}{2}\right) + \beta_2\right)$$

$$+ \left(\frac{N}{2}h_1^{(N)} + \frac{N}{2}h_2^{(N)} + b^{(N)}\right)\left(\mathbb{E}\left[\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1)^+\right] + \mathbb{E}\left[\frac{1}{\beta_2}\max_{i \le N/2}(Q_i(1) - I_2)^+\right]\right).$$

Our analytical results enable us to minimize this upper bound; for instance, choosing $\tilde{h}_{1,2}^{(N)} = h_{1,2}^{(N)}$ and $\tilde{b}_{1,2} = \frac{N}{2}h_{2,1}^{(N)} + b^{(N)}$ yields

$$\frac{N}{2}\left(h_1^{(N)}\frac{1}{\beta_1}\left(I_1 - \frac{\sigma^2}{2}\right) + \beta_1\right) + \frac{N}{2}\left(h_2^{(N)}\frac{1}{\beta_2}\left(I_2 - \frac{\sigma^2}{2}\right) + \beta_2\right)$$

$$+ \left(\frac{N}{2}h_1^{(N)} + \frac{N}{2}h_2^{(N)} + b^{(N)}\right)\left(\mathbb{E}\left[\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1)^+\right] + \mathbb{E}\left[\frac{1}{\beta_2}\max_{i \le N/2}(Q_i(1) - I_2)^+\right]\right)$$

$$= \frac{N}{2}\left(\tilde{h}_1^{(N)}\frac{1}{\beta_1}\left(I_1 - \frac{\sigma^2}{2}\right) + \beta_1\right) + \left(\frac{N}{2}\tilde{h}_1^{(N)} + \tilde{b}_1^{(N)}\right)\mathbb{E}\left[\frac{1}{\beta_1}\max_{i \le N/2}(Q_i(1) - I_1)^+\right]$$

$$+ \frac{N}{2}\left(\tilde{h}_2^{(N)}\frac{1}{\beta_2}\left(I_2 - \frac{\sigma^2}{2}\right) + \beta_2\right) + \left(\frac{N}{2}\tilde{h}_2^{(N)} + \tilde{b}_2^{(N)}\right)\mathbb{E}\left[\frac{1}{\beta_2}\max_{i \le N/2}(Q_i(1) - I_2)^+\right].$$

This is the sum of two functions that can be minimized using the exact solutions that we derived. In Table 9, we compare numerically the actual costs under the capacity and base-stock level that are obtained by minimizing this upper bound with the costs under the optimal capacity and base-stock level. In this table, the ratio indicates how many servers have a holding cost $h_1^{(N)}$ and how many servers have a holding cost $h_2^{(N)}$; the 1:1 ratio

corresponds to the preceding example, whereas the 1:3 ratio can be treated similarly. The table demonstrates that our asymptotic results may be useful when optimizing asymmetric systems as well as symmetric systems.

8. Conclusions

In this study, we define a large-scale assembly system in which N components are assembled into a final product. We study an assembly system with linear demand and production, subject to some random noise. Thus, we impose the natural assumption that this noise is normally distributed. Hence, 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 delays in assembly of the final product, which is equal to the delay of the slowest produced component. Before attempting 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 distinguish two cases: First of all, we covered the case of deterministic demand, resulting in all delays being independent. Second, we investigated the case in which demand is stochastic and, consequently, delays of the components are dependent.

For the deterministic demand scenario, we prove order bounds for three different regimes: balanced, quality driven, and efficiency driven. Additionally, we verify 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 develop a limit theorem that we use to obtain approximate solutions. We show numerically that, even though, theoretically, these approximations perform well, for practical situations, there is still room for improvement. However, this limit theorem is still necessary for systems with N of the order of thousands because it is close to impossible to simulate these systems quickly. Therefore, we provide additional approximations for a mixed-behavior regime, in which we use a combination of the approximations for the deterministic and stochastic demand scenarios. We demonstrate 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, in which the components are not produced in-house by the manufacturer, 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 5.1 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.

Appendix A. Proofs
A.1. Proofs of Section 3

Proof of Lemma 3.2. $F_N(I, \beta) > 0$; hence, F_N has a global infimum, and because $\lim_{\beta \downarrow 0} F_N(I, \beta) = \infty$, $\lim_{\beta \to \infty} F_N(I, \beta) = \infty$ and $\lim_{I \to \infty} F_N(I, \beta) = \infty$, 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

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

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

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 base-stock level 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

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

$$= (b^{(N)} + Nh^{(N)}) f(I) \geq 0,$$

because f is the probability density function of $\max_{i \leq N} Q_i$. This density exists (cf. Dai and Harrison 1992, proposition 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 $F_N(I_N^*, \beta_N^*) = 2\sqrt{NC_N(I_N^*)} = 2N\beta_N^*$. $\square$

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

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

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 3.1, we have

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

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, $C_N(I) > \overline{C}_N(I)$. Therefore,

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

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

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

In the case that $\gamma_N = \gamma \in (0, 1)$, we have, by applying Lemma 3.5, that $|\overline{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| = o((Nh^{(N)} + b^{(N)})\log N)$. Furthermore, $C_N(\bar{I}_N) \sim Nh^{(N)}\frac{\sigma^2}{2}\log N$, and because $\max_{i \leq N} Q_i/\log N \overset{\mathbb{P}}{\to} \sigma^2/2$ as $N \to \infty$, we also have that $I_N^*/\log N \overset{N \to \infty}{\to} \sigma^2/2$. 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 $\gamma_N \overset{N \to \infty}{\to} 1$, we first observe that $\overline{C}_N(\bar{I}_N) = Nh^{(N)}\left(\frac{\sigma^2}{2}\log N - \frac{\sigma^2 + \sigma_A^2}{2}\right) \sim Nh^{(N)}\frac{\sigma^2}{2}\log N$. Furthermore,

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

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

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 $\mathbb{E}\left[\left|\max_{i \leq N} Q_i - \frac{\sigma^2}{2}\log N\right|\right]/\log N \overset{N \to \infty}{\to} 0$. Thus,

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

Finally,

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

$$\geq Nh^{(N)}\left(I_N^* - \frac{\sigma^2 + \sigma_A^2}{2}\right) + (Nh^{(N)} + b^{(N)})\mathbb{E}\left[\max_{i \leq N} Q_i - I_N^*\right]$$

$$\geq -Nh^{(N)}\frac{\sigma^2 + \sigma_A^2}{2} + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\log N - b^{(N)}I_N^*.$$

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

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

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 $P_N^{-1}(x) = \frac{\sigma^2}{2}\log(1/(1 - x^{\frac{1}{N}}))$. □

Proof of Proposition 4.1. 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 base-stock level is given in Equation (12). Furthermore,

$$\mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log N - \hat{I}_N\right)^+\right] = \mathbb{E}\left[\left(\frac{\sigma^2}{2}G + \frac{\sigma^2}{2}\log(-\log(1 - \gamma_N))\right)^+\right]$$

$$= \frac{\sigma^2}{2}\int_{-\log(-\log(1-\gamma_N))}^{\infty} 1 - e^{-e^{-x}}dx.$$

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 (13). □

Lemma A.1. *Define*

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

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

$$\max_{i \leq N} Q_i > \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N, \tag{A.2}$$

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 zero.*

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

$$\mathbb{P}(G_N < x) = \mathbb{P}\left(-\log\left(-\log\left(\left(1 - \exp\left(-\frac{2}{\sigma^2}\max_{i \leq N} Q_i\right)\right)^N\right)\right) < x\right)$$

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

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

To prove (A.2), 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 $\max_{i \le N} Q_i$ and $\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N$ becomes, we first observe that

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

Thus, we need to obtain that $x + \frac{\sigma^2}{2} \log(-\log(1 - e^{-\frac{2}{\sigma^2}x}))$ 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, $\lim_{x \to \infty} x + \frac{\sigma^2}{2} \log(-\log(1 - e^{-\frac{2}{\sigma^2}x})) = 0$. $\square$

Lemma A.2. *Let* $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$; *then,*

$$
|C_N(I_N^*) - C_N(\hat{I}_N)| \le (I_N^* - \hat{I}_N)(Nh^{(N)} + b^{(N)})\left(1 - \gamma_N - \left(1 + \frac{\log(1 - \gamma_N)}{N} \right)^N \right), \tag{A.3}
$$

$$
|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)| \le (I_N^* - \hat{I}_N)Nh^{(N)}\left(1 - \left(1 + \frac{\log(1 - \gamma_N)}{N} \right)^N \right). \tag{A.4}
$$

Proof. Because of the inequality in (A.2), $I_N^* > \hat{I}_N$. Then, we have

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

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

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

Furthermore,

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

Equation (A.3) follows. To prove Equation (A.4), we observe that

$$
\begin{aligned}
&|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)| \\
&\quad = (Nh^{(N)} + b^{(N)})\mathbb{E}\left[\left(\max_{i \le N} Q_i - \hat{I}_N \right)^+ - \left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N - \hat{I}_N \right)^+ \right] \\
&\quad = (Nh^{(N)} + b^{(N)})\mathbb{E}\left[\left(\max_{i \le N} Q_i - \frac{\sigma^2}{2} G_N - \frac{\sigma^2}{2} \log N \right)\mathbb{1}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N \right) \right] \tag{A.5} \\
&\qquad + (Nh^{(N)} + b^{(N)})\mathbb{E}\left[\left(\max_{i \le N} Q_i - \hat{I}_N \right)\mathbb{1}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N < \max_{i \le N} Q_i \right) \right]. \tag{A.6}
\end{aligned}
$$

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

$$
\mathbb{E}\left[(\max_{i \leq N} Q_i - \frac{\sigma^2}{2} G_N - \frac{\sigma^2}{2} \log N)\mathbb{1}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N\right)\right]
$$

$$
\leq (I_N^* - \hat{I}_N)\mathbb{P}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N > \hat{I}_N\right) \tag{A.7}
$$

$$
= (I_N^* - \hat{I}_N)\gamma_N.
$$

Similarly, for (A.6), we observe that, if $\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N$, then $\max_{i \leq N} Q_i < I_N^*$, and thus,

$$
\mathbb{E}\left[(\max_{i \leq N} Q_i - \hat{I}_N)\mathbb{1}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N < \max_{i \leq N} Q_i\right)\right]
$$

$$
\leq (I_N^* - \hat{I}_N)\mathbb{P}\left(\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2} \log N < \hat{I}_N < \max_{i \leq N} Q_i\right) \tag{A.8}
$$

$$
\leq (I_N^* - \hat{I}_N)\left(1 - \left(1 + \frac{\log(1 - \gamma_N)}{N}\right)^N - \gamma_N\right).
$$

Adding the bounds in (A.7) and (A.8) gives the result. $\square$

Proof of Theorem 4.1. First, we assume that $\gamma_N = \gamma \in (0, 1)$. Using Corollary 3.1, 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 (A.2), we have for all I that $C_N(I) > \hat{C}_N(I)$, and thus,

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

We write $f(x) := I_{1/x}^* - \hat{I}_{1/x}$ for $x > 0$. Then, we have that

$$
f(x) = \frac{\sigma^2}{2} \log\left(\frac{1}{1 - (1 - \gamma)^x}\right) + \frac{\sigma^2}{2} \log x + \frac{\sigma^2}{2} \log(-\log(1 - \gamma)))
$$

$$
= \frac{\sigma^2}{2} \log\left(\frac{x}{1 - (1 - \gamma)^x}\right) + \frac{\sigma^2}{2} \log(-\log(1 - \gamma))).
$$

By first noting that $x/(1 - (1 - \gamma)^x) = 1/(1 - e^{-x\log(1 - \gamma)}) \to 1/(-\log(1 - \gamma)) > 0$, we see that $\log(x/(1 - (1 - \gamma)^x)) \to -\log(-\log(1 - \gamma))$ as $x \downarrow 0$. From this, it follows that $f(x) \to 0$ as $x \downarrow 0$, and we can extend the domain of the function f such that $f(0) := 0$ and f is twice differentiable at $x = 0$. By computing the Taylor series of the function f at $x = 0$, we get

$$
f(x) = -\frac{\sigma^2}{4} x\log(1 - \gamma) + O(x^2).
$$

Thus, $(I_N^* - \hat{I}_N) \sim -\sigma^2 \log(1 - \gamma)/(4N)$, as $N \to \infty$. Following (A.4), 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 (A.3), we get $|C_N(I_N^*) - C_N(\hat{I}_N)|/(Nh^{(N)} + b^{(N)}) = O(1/N^2)$. We have

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

$$
\sim Nh^{(N)}\frac{\sigma^2}{2}\log N,
$$

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

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

$$= \frac{\sqrt{C_N(\hat{I}_N) - O((Nh^{(N)} + b^{(N)})/N^2)}}{\sqrt{C_N(\hat{I}_N)}} \frac{\sqrt{C_N(\hat{I}_N) - O(Nh^{(N)}/N)}}{\sqrt{C_N(\hat{I}_N)}}$$

$$= \sqrt{1 - O(1/(N^2 \log N))} \sqrt{1 - O(1/(N \log N))}$$

$$= 1 - O(1/(N \log N)).$$

Now, we assume that $\gamma_N \overset{N\to\infty}{\to} 0$, and 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,

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

From this, it follows that

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

Also,

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

Earlier, we show that, when $\gamma_N = \gamma$, $(I_N^* - \hat{I}_N) = O(1/N)$, now I_N^* is larger because $\mathbb{P}(\max_{i \le N} Q_i < I_N^*) = 1 - \gamma_N \overset{N\to\infty}{\to} 1$. Following the statement in Lemma A.1 that the difference between $\max_{i \le N} Q_i$ and $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N$ decreases as $\max_{i \le N} Q_i$ increases, we can conclude that $(I_N^* - \hat{I}_N) = O(1/N)$. Following the proof before, and by using the order bounds in (A.3) and (A.4), 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 $\gamma_N \overset{N\to\infty}{\to} 1$ and $\gamma_N \le 1 - \exp(-N)$. Then, $\hat{I}_N \ge 0$. Furthermore, when $\gamma_N \overset{N\to\infty}{\to} 1$, we have $\log(-\log(1-\gamma_N)) \overset{N\to\infty}{\to} \infty$, and from this, it follows that

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

Thus,

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

$$= \frac{\sigma^2}{2}Nh^{(N)}\log N + \frac{\sigma^2}{2}b^{(N)}\log(-\log(1-\gamma_N)).$$

Because we consider the efficiency-driven regime, we have $b^{(N)}/(Nh^{(N)}) \overset{N\to\infty}{\to} 0$. 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, $I_N^* - \hat{I}_N = O(1)$, and thus, the bounds in (A.3) and (A.4) are of $O(Nh^{(N)})$. By using the same argument as in the proof for the balanced regime,

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

Proof of Lemma 4.2. Following Equations (A.3) and (A.4) and using the same arguments as in the proof of Theorem 4.1, 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

$$\hat{C}_N(\hat{I}_N) = Nh^{(N)}\frac{\sigma^2}{2}\Big(\log N - \log(-\log(1-\gamma)) - 1\Big)$$

$$+ (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\mathbb{E}[(G + \log(-\log(1-\gamma)))^+].$$

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 $\gamma_N \xrightarrow{N\to\infty} 0$, we have that $-\log(-\log(1-\gamma_N)) \sim -\log(\gamma_N)$, and 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

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

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

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

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

$$= \frac{\sigma^2}{2}Nh^{(N)}\log N + \frac{\sigma^2}{2}b^{(N)}\log(-\log(1-\gamma_N)).$$

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 5.1. Let $b_N = \sqrt{2\log N} - \log(4\pi\log N)/(2\sqrt{2\log N})$. Then,

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

with $G \sim$ Gumbel as $N \to \infty$ (cf. de Haan and Ferreira [2006], example 1.1.7, for a proof). Observe that

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

$$= \frac{1}{\sigma\sqrt{d}}\left(\sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}}\right)\frac{\max_{i\le N}W_i(d\log N) - \sigma\sqrt{2d}\log N + \frac{\sigma\sqrt{d}\log(4\pi\log N)}{2\sqrt{2}}}{\sqrt{\log N}}.$$

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

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

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

Proof of Lemma 5.2. To prove Lemma 5.2, we first observe that

$$\frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s) + W_A(s) - \beta s)) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$$

$$\le \frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s) - \beta s)) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} + \frac{\sup_{0<s<(d-\epsilon)\log N}W_A(s)}{\sqrt{\log N}}. \tag{A.9}$$

We first focus on the first term on the right-hand side of (A.9). We know that $\sup_{0<s<(d-\epsilon)\log N}(W_i(s) - \beta s)$ is a reflected Brownian motion, so we can write down its cumulative distribution function explicitly:

$$\mathbb{P}\left(\sup_{0<s<(d-\epsilon)\log N}(W_i(s) - \beta s) \le x\right)$$

$$= 1 - \Phi\left(\frac{-x - \beta(d-\epsilon)\log N}{\sigma\sqrt{(d-\epsilon)\log N}}\right) - \exp\left(-\frac{2\beta}{\sigma^2}x\right)\Phi\left(\frac{-x + \beta(d-\epsilon)\log N}{\sigma\sqrt{(d-\epsilon)\log N}}\right); \tag{A.10}$$

see Abate and Whitt (1987, equation (1.1)). From this, together with the union bound, it follows that

$$\mathbb{P}\left(\max_{i \leq N} \sup_{0 < s < (d-\epsilon)\log N} (W_i(s) - \beta s) \geq \frac{\sigma^2}{2\beta}\log N + x\sqrt{\log N}\right) \tag{A.11}$$

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

$$= N\Phi\left(\frac{-\beta(2d-\epsilon)\log N - x\sqrt{\log N}}{\sigma\sqrt{(d-\epsilon)\log N}}\right) + \exp\left(-\frac{2\beta}{\sigma^2}x\sqrt{\log N}\right)\Phi\left(\frac{-\epsilon\beta\log N - x\sqrt{\log N}}{\sigma\sqrt{(d-\epsilon)\log N}}\right). \tag{A.12}$$

The cumulative distribution of the normal distribution Φ satisfies $\Phi(-x) = 1 - \Phi(x)$. Furthermore, we have that $1 - \Phi(x) \sim \exp(-x^2/2)/(\sqrt{2\pi}x)$ as $x \to \infty$; see Adler and Taylor (2007, equation (2.1.1)). This asymptotic equivalence gives us that the first term in (A.12) satisfies

$$N\Phi\left(\frac{-\beta(2d-\epsilon)\log N - x\sqrt{\log N}}{\sigma\sqrt{(d-\epsilon)\log N}}\right) = N\exp\left(-\frac{\beta^2(2d-\epsilon)^2}{2\sigma^2(d-\epsilon)}\log N(1 + o(1))\right)$$

$$= N\exp\left(-\frac{(2d-\epsilon)^2}{4d(d-\epsilon)}\log N(1 + o(1))\right).$$

For all $\epsilon \in (0, d)$, we have that $\frac{(2d-\epsilon)^2}{4d(d-\epsilon)} = \frac{4d^2 - 4d\epsilon + \epsilon^2}{4d(d-\epsilon)} > \frac{4d^2 - 4d\epsilon}{4d(d-\epsilon)} = 1$. Thus, we can conclude that

$$N\exp\left(-\frac{(2d-\epsilon)^2}{4d(d-\epsilon)}\log N(1 + o(1))\right) \overset{N\to\infty}{\to} 0.$$

With the asymptotic equivalence from Adler and Taylor (2007, equation (2.1.1)), we get for the second term in (A.12) that

$$\exp\left(-\frac{2\beta}{\sigma^2}x\sqrt{\log N}\right)\Phi\left(\frac{-\epsilon\beta\log N - x\sqrt{\log N}}{\sigma\sqrt{(d-\epsilon)\log N}}\right) = \exp\left(-\frac{\epsilon^2\beta^2}{2\sigma^2(d-\epsilon)}\log N(1 + o(1))\right) \overset{N\to\infty}{\to} 0.$$

For the second term on the right-hand side of (A.9), we argue as follows: by filling in $\beta = 0$ and replacing σ with σ_A in Equation (A.10), one can easily see that

$$\sup_{0 < s < (d-\epsilon)\log N} W_A(s) \overset{d}{=} |W_A((d-\epsilon)\log N)| \overset{d}{=} \sqrt{(d-\epsilon)\log N}|X|,$$

with $X \sim \mathcal{N}(0,1)$. Thus, we can use the upper bound in (A.9) and conclude that

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

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

$$\leq N\mathbb{P}\left(\frac{\sup_{0 < s < (d-\epsilon)\log N}(W_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \geq x - y\right) + \mathbb{P}\left(\frac{\sup_{0 < s < (d-\epsilon)\log N}W_A(s)}{\sqrt{\log N}} \geq y\right)$$

$$\overset{N\to\infty}{\to} \mathbb{P}\left(|X| > \frac{y}{\sqrt{d-\epsilon}}\right).$$

This last expression converges to zero as $y \to \infty$, and the lemma follows. $\square$

Proof of Lemma 5.3. 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,

$$\frac{\max_{i \leq N}\left(\sup_{s \geq (d+\epsilon)\log N}(W_i(s) + W_A(s) - \beta s)\right) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$$

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

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

We have

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

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

$$\frac{\max_{i \leq N}\left(\sup_{s \geq (d+\epsilon)\log N}(W_i(s) - (\beta - \delta)s)\right) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$$

$$\overset{d}{=} \frac{\max_{i \leq N}\left(W_i((d+\epsilon)\log N) + E_i\right) - \left(\frac{\sigma^2}{2\beta} + (\beta - \delta)(d+\epsilon)\right)\log N}{\sqrt{\log N}}.$$

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

$$\mathbb{P}\left(\max_{i \leq N}\left(W_i((d+\epsilon)\log N) + E_i\right) > x\right) \leq N\mathbb{P}\left(W_i((d+\epsilon)\log N) + E_i > x\right)$$

$$\leq N\mathbb{E}[e^{sW_i((d+\epsilon)\log N)}]\mathbb{E}[e^{sE_i}]e^{-sx},$$

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,

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

$$\leq N^{1+\frac{\sigma^2(d+\epsilon)s^2}{2} - s\left(\frac{\sigma^2}{2\beta}+(\beta-\delta)(d+\epsilon)\right)}e^{-sx\sqrt{\log N}}\frac{\frac{2(\beta-\delta)}{\sigma^2}}{\frac{2(\beta-\delta)}{\sigma^2} - s}. \tag{A.13}$$

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

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

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,

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

Moreover, $\sup_{s>0}(W_A(s) - \delta s) \sim \mathrm{Exp}\left(\frac{2\delta}{\sigma_A^2}\right)$. Therefore, $\frac{\sup_{s>0}(W_A(s)-\delta s)}{\sqrt{\log N}} \overset{\mathbb{P}}{\to} 0$. The limit in (23) follows. $\square$

Proof of Lemma 5.4. First, we bound

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

$$\leq \sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N}\frac{W_A(s)}{\sqrt{\log N}} + \frac{\max_{i \leq N}\sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N}(W_i(s) - \beta s) - \frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}}$$

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

We can write

$$\sup_{(d-\epsilon)\log N \leq s < (d+\epsilon)\log N}\frac{W_A(s)}{\sqrt{\log N}} = \frac{W_A((d-\epsilon)\log N)}{\sqrt{\log N}} + \sup_{0 \leq s < 2\epsilon\log N}\frac{W_A'(s)}{\sqrt{\log N}}$$

$$\overset{d}{=} \sigma_A\sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon}X_1 + \sqrt{2\epsilon}\sigma_A|X_2|,$$

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

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

as $N \to \infty$ with $G \sim$ 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 $N \to \infty$. The statement follows. $\quad\square$

Proof of Theorem 5.1. We have the following lower bound:

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

$$\geq \mathbb{P}\left(\frac{\max_{i \leq N} (W_i(d \log N) + W_A(d \log N)) - \beta d \log N - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \geq x \right).$$

From this and Lemma 5.1, we know that

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

By using the union bound, we get

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

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

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

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

Combining this with the results from Lemmas 5.2–5.4 gives

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

$$\leq \mathbb{P}\left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon} X_1 + \sqrt{2\epsilon} \sigma_A |X_2| > x \right),$$

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

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

$$\leq \lim_{\epsilon \downarrow 0} \mathbb{P}\left(\sigma_A \sqrt{\frac{\sigma^2}{2\beta^2} - \epsilon} X_1 + \sqrt{2\epsilon} \sigma_A |X_2| > x \right)$$

$$= 1 - \Phi\left(\frac{x\sqrt{2}\beta}{\sigma \sigma_A} \right).$$

Hence, the statement follows. $\quad\square$

Proof of Lemma 5.5. 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 $0 < \epsilon < d$, and we write

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

First, observe that

$$
\mathbb{E}\left[\left|\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N\right|\right]
\tag{A.14}
$$

$$
\le \mathbb{E}\left[\left|\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i \le N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}}\right|\right]
\tag{A.15}
$$

$$
+ \mathbb{E}\left[\left|\frac{\max_{i \le N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N\right|\right].
\tag{A.16}
$$

Because of Pickands (1968, theorem 3.1), we obtain for the term in (A.16) that

$$
\mathbb{E}\left[\left|\frac{\max_{i \le N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}} - \frac{\sigma \sigma_A}{\sqrt{2}} X_N\right|\right]
$$
$$
= \mathbb{E}\left[\left|\frac{\max_{i \le N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}}\right|\right] \xrightarrow{N \to \infty} 0.
\tag{A.17}
$$

Furthermore, because $Q_i > W_i(d \log N) + W_A(d \log N) - d \log N$, we can rewrite (A.15):

$$
\mathbb{E}\left[\left|\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i \le N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}}\right|\right]
$$
$$
= \mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i \le N} W_i(d \log N) + W_A(d \log N) - \sigma^2 \log N}{\sqrt{\log N}}\right]
\tag{A.18}
$$
$$
= \mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right] - \mathbb{E}\left[\frac{\max_{i \le N} W_i(d \log N) - \sigma^2 \log N}{\sqrt{\log N}}\right].
$$

The second term in (A.18) converges to zero as $N \to \infty$, which follows from the convergence in (A.17). In order to find a converging upper bound for the first term in (A.18), we write

$$
\mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right]
$$

$$
\le \mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(-M \le \frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \le M\right)\right]
\tag{A.19}
$$

$$
+ \mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} > M\right)\right].
\tag{A.20}
$$

For the term in (A.19), we can conclude from Theorem 5.1 together with the dominated convergence theorem that

$$
\mathbb{E}\left[\frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \mathbb{1}\left(-M \le \frac{\max_{i \le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} \le M\right)\right]
$$
$$
\xrightarrow{N \to \infty} \mathbb{E}\left[\frac{\sigma \sigma_A}{\sqrt{2}} X \mathbb{1}\left(-M \le \frac{\sigma \sigma_A}{\sqrt{2}} X \le M\right)\right] = 0,
$$

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

In order to find a converging upper bound for the term in (A.20), we bound

$$
\max_{i \le N} Q_i \le \max_{i \le N} \sup_{s > 0}\left(W_i(s) - \left(1 - 1/\sqrt{\log N}\right)s\right) + \sup_{s > 0}(W_A(s) - s/\sqrt{\log N}) =: Z_N.
$$

Then, we have the bound

$$
\mathbb{E}\left[\frac{\max_{i\le N}Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\max_{i\le N}Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M\right)\right]
$$

$$
\le \mathbb{E}\left[\frac{Z_N - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M/2\right)\right]
$$

$$
+\mathbb{E}\left[\frac{Z_N - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\sup_{s>0}(W_A(s)-s/\sqrt{\log N})}{\sqrt{\log N}}\ge M/2\right)\right].
$$

Because $\sup_{s>0}(W_A(s)-s/\sqrt{\log N})$ is exponentially distributed with mean $\sigma_A^2\sqrt{\log N}/2$, we have that

$$
\mathbb{E}\left[\frac{\sup_{s>0}(W_A(s)-s/\sqrt{\log N})}{\sqrt{\log N}}\right]=\frac{\sigma_A^2}{2}.
$$

Additionally, $\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)$ is the maximum of N i.i.d. exponentials with mean $\sigma^2/(2(1-1/\sqrt{\log N}))$, and it is a standard result that

$$
\mathbb{E}\left[\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)\right]=\frac{\sigma^2}{2(1-1/\sqrt{\log N})}\sum_{i=1}^{N}\frac{1}{i},
$$

see Rényi (1953). From this, it follows that

$$
\mathbb{E}\left[\frac{\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right]\overset{N\to\infty}{\to}\frac{\sigma^2}{2}.
$$

Furthermore, because of the memoryless property of exponential random variables, we have that

$$
\mathbb{E}\left[\frac{\sup_{s>0}(W_A(s)-s/\sqrt{\log N})}{\sqrt{\log N}}\mathbb{1}\left(\frac{\sup_{s>0}(W_A(s)-s/\sqrt{\log N})}{\sqrt{\log N}}\ge M/2\right)\right]
$$
$$
=\exp(-M/\sigma_A^2)\left(\frac{M}{2}+\frac{\sigma_A^2}{2}\right)\overset{M\to\infty}{\to}0,
$$

and

$$
\mathbb{E}\left[\frac{\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\cdot\mathbb{1}\left(\frac{\max_{i\le N}\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M/2\right)\right]
$$

$$
=\mathbb{E}\left[\max_{i\le N}\left(\frac{\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\cdot\mathbb{1}\left(\frac{\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M/2\right)\right)\right]
$$

$$
\le N\mathbb{E}\left[\frac{\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\cdot\mathbb{1}\left(\frac{\sup_{s>0}(W_i(s)-(1-1/\sqrt{\log N})s)-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M/2\right)\right]
$$

$$
=N\exp\left(-\frac{2(1-1/\sqrt{\log N})(\frac{\sigma^2}{2}\log N+\frac{M}{2}\sqrt{\log N})}{\sigma^2}\right)\left(\frac{M}{2}+\frac{\sigma^2}{2(1-1/\sqrt{\log N})}\right)
$$

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

for $M>\sigma^2$. From these results, it follows that,

$$
\lim_{M\to\infty}\limsup_{N\to\infty}\mathbb{E}\left[\frac{\max_{i\le N}Q_i-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\mathbb{1}\left(\frac{\max_{i\le N}Q_i-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\ge M\right)\right]=0.
$$

The lemma follows. $\quad\square$

A.4. Proofs of Section 5.2

Proof of Lemma 5.6. From Lemma 3.2, we know that the optimal inventory I_N^A satisfies

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

We have

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

$$= Nh^{(N)} - (Nh^{(N)} + b^{(N)})\mathbb{P}(\max_{i\leq N} Q_i > I_N^A)$$

$$= Nh^{(N)} - (Nh^{(N)} + b^{(N)})\mathbb{P}\left(\frac{\sqrt{2}}{\sigma\sigma_A}\frac{\max_{i\leq N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} > \frac{\sqrt{2}}{\sigma\sigma_A}\frac{I_N^A - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right).$$

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

Proof of Proposition 5.1. 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 (26) minimizes $\hat{C}_N^A(I)$. We know that

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

$$= \left(\frac{\sigma^2}{2}\log N - \hat{I}_N^A\right)\mathbb{P}\left(\frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X \geq \hat{I}_N^A - \frac{\sigma^2}{2}\log N\right)$$

$$+ \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\sigma^2\log N - 2\hat{I}_N^A)^2}{4\sigma^2\sigma_A^2\log N}\right)$$

$$= -\frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\Phi^{-1}(1 - \gamma_N)\gamma_N$$

$$+ \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Phi^{-1}(1 - \gamma_N)^2\right).$$

The expression in Equation (27) follows. $\square$

Proof of Theorem 5.2. Using Corollary 3.1, 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, lemma 21.2) that $(I_N^A - \hat{I}_N^A)/\sqrt{\log N} \overset{N\to\infty}{\to} 0$. Furthermore, we prove in Lemma 5.5 that $\mathbb{E}\left[\left|\max_{i\leq N} Q_i - \frac{\sigma^2}{2}\log N - \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X\right|/\sqrt{\log N}\right] \overset{N\to\infty}{\to} 0$. From this, it follows that $|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| = o((Nh^{(N)} + b^{(N)})\sqrt{\log N})$. Because $\hat{C}_N^A(\hat{I}_N^A) \sim \frac{\sigma^2}{2}Nh^{(N)}\log N$, we have $\frac{\sqrt{C_N(I_N^A)}}{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}} = 1 - o((Nh^{(N)} + b^{(N)})\sqrt{\log N}/(Nh^{(N)}\log N)) = 1 - o(1/\sqrt{\log N})$.

Second, assume $\hat{C}_N^A(\hat{I}_N^A) < C_N(\hat{I}_N^A)$, and 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 5.7. 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, $|I_N^A - \hat{I}_N^A| = o(\sqrt{\log N})$, and thus, (28) follows. Furthermore, by using the same argument as in Lemma 4.2, (29) follows. $\square$

A.5. Mixed-Behavior Approximations

Though we have a symbolic expression for β_N^M in (32), it is not completely clear how to compute the part

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

in β_N^M. First, observe that we can write

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

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

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

Thus,

$$\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right]$$
$$= \int_{I_N^M}^{\infty}\int_0^{\infty}\mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dsdx$$
$$= \int_0^{\infty}\int_{I_N^M}^{\infty}\mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dxds.$$

It turns out that

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

can be expressed in terms of error functions. Thus, because I_N^M can be numerically found by solving Equation (31), $\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right]$ can be computed numerically as well. Observe that the procedure to obtain I_N^M and β_N^M is efficient and that its running time is independent of the system size N.

References

Abate J, Whitt W (1987) Transient behavior of regulated Brownian motion I: Starting at the origin. *Adv. Appl. Probab.* 19(3):560–598.

Akçay Y, Xu SH (2004) Joint inventory replenishment and component allocation optimization in an assemble-to-order system. *Management Sci.* 50(1):99–116.

Altendorfer K, Minner S (2011) Simultaneous optimization of capacity and planned lead time in a two-stage production system with different customer due dates. *Eur. J. Oper. Res.* 213(1):134–146.

ASML Holding NV (2021) ASML annual report 2020. Accessed July 5, 2021, https://www.asml.com/en/investors/annual-report/2020.

Asmussen S (2003) *Applied Probability and Queues*, vol. 2 (Springer, New York).

Asmussen S, Glynn PW, Pitman J (1995) Discretization error in simulation of one-dimensional reflecting Brownian motion. *Ann. Appl. Probab.* 5(4):875–896.

Atan Z, Rousseau M (2016) Inventory optimization for perishables subject to supply disruptions. *Optim. Lett.* 10(1):89–108.

Atan Z, Ahmadi T, Stegehuis C, de Kok T, Adan I (2017) Assemble-to-order systems: A review. *Eur. J. Oper. Res.* 261(3):866–879.

Atar R, Mandelbaum A, Zviran A (2012) Control of fork-join networks in heavy traffic. *2012 50th Annual Allerton Conf. Comm., Control. Comput.* (IEEE, Piscataway, NJ), 823–830.

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. *Proc. IEEE* 77(1):138–161.

Bijvank M, Huh WT, Janakiraman G, Kang W (2014) Robustness of order-up-to policies in lost-sales inventory systems. *Oper. Res.* 62(5):1040–1047.

Bollapragada R, Rao US, Zhang J (2004) Managing two-stage serial inventory systems under demand and supply uncertainty and customer service level requirements. *IIE Trans.* 36(1):73–85.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.

Bradley JR, Glynn PW (2002) Managing capacity and inventory jointly in manufacturing systems. *Management Sci.* 48(2):273–288.

Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. *J. Appl. Probab.* 14(4):732–739.

Chaturvedi A, Martínez-de Albéniz V (2016) Safety stock, excess capacity or diversification: Trade-offs under supply and demand uncertainty. *Production Oper. Management* 25(1):77–95.

de Haan L, Ferreira A (2006) *Extreme Value Theory: An Introduction* (Springer Science & Business Media, New York).

Dębicki K, Ji L, Rolski T (2020) Exact asymptotics of component-wise extrema of two-dimensional Brownian motion. *Extremes* 23:569–602.

Dębicki K, Hashorva E, Ji L, Tabiś K (2015) Extremes of vector-valued Gaussian processes: Exact asymptotics. *Stochastic Processes Appl.* 125(11):4039–4065.

Denton J (2021) ASML cuts guidance in the face of supply chain issues: The chip stock is falling. Accessed February 7, 2022, https://www.barrons.com/articles/asml-cuts-guidance-supply-chain-issues-51634728074.

Doğru MK, Reiman MI, Wang Q (2017) Assemble-to-order inventory management via stochastic programming: Chained BOMs and the M-system. *Production Oper. Management* 26(3):446–468.

Ewing J, Clark D (2021) Lack of tiny parts disrupts auto factories worldwide. *The New York Times Online* (January 13), https://www.nytimes.com/2021/01/13/business/auto-factories-semiconductor-chips.html.

Flatto L, Hahn S (1984) Two parallel queues created by arrivals with two demands I. *SIAM J. Appl. Math.* 44(5):1041–1053.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Glasserman P (1997) Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* 45(2):244–257.

Goldberg DA, Reiman MI, Wang Q (2021) A survey of recent progress in the asymptotic analysis of inventory systems. *Production Oper. Management* 30(6):1718–1750.

Goldberg DA, Katz-Rogozhnikov DA, Lu Y, Sharma M, Squillante MS (2016) Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *Math. Oper. Res.* 41(3):898–913.

Gopalakrishnan R, Doroudi S, Ward AR, Wierman A (2016) Routing and staffing when servers are strategic. *Oper. Res.* 64(4):1033–1050.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

Harrison JM (1985) *Brownian Motion and Stochastic Flow Systems* (Wiley, New York).

Harrison JM (2013) *Brownian Models of Performance and Control* (Cambridge University Press, Cambridge, UK).

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Sci.* 55(3):404–420.

Karsten F, Slikker M, van Houtum GJ (2012) Inventory pooling games for expensive, low-demand spare parts. *Naval Res. Logist.* 59(5):311–324.

Klein SJd (1988) Fredholm integral equations in queueing analysis. Unpublished PhD thesis, Rijksuniversiteit Utrecht, Netherlands.

Klosterhalfen ST, Minner S, Willems SP (2014) Strategic safety stock placement in supply networks with static dual supply. *Manufacturing Service Oper. Management* 16(2):204–219.

Ko SS, Serfozo RF (2004) Response times in M/M/s fork-join networks. *Adv. Appl. Probab.* 36(3):854–871.

Kou S, Zhong H (2016) First-passage times of two-dimensional Brownian motion. *Adv. Appl. Probab.* 48(4):1045–1060.

Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing Service Oper. 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, New York).

Lu H, Pang G (2015) Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Math. Oper. Res.* 41(2):560–595.

Lu H, Pang G (2017a) Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stoch. 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.

Lu Y, Song JS (2005) Order-based cost optimization in assemble-to-order systems. *Oper. Res.* 53(1):151–169.

Mayorga ME, Ahn HS (2011) Joint management of capacity and inventory in make-to-stock production systems with multi-class demand. *Eur. J. Oper. Res.* 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 Sci.* 62(6):1830–1841.

Nelson R, Tantawi AN (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE Trans. Comput.* 37(6):739–743.

Nguyen V (1993) Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *Ann. Appl. Probab.* 3(1):28–55.

Nguyen V (1994) The trouble with diversity: Fork-join networks with heterogeneous customer population. *Ann. Appl. Probab.* 4(1):1–25.

Pan W, So KC (2016) Component procurement strategies in decentralized assembly systems under supply uncertainty. *IIE Trans.* 48(3):267–282.

Plambeck EL (2008) Asymptotically optimal control for an assemble-to-order system with capacitated component production and fixed transport costs. *Oper. Res.* 56(5):1158–1171.

Plambeck EL, Ward AR (2008) Optimal control of a high-volume assemble-to-order system with maximum leadtime quotation and expediting. *Queueing Systems* 60(1):1–69.

Reddy KN, Kumar A (2020) Capacity investment and inventory planning for a hybrid manufacturing-remanufacturing system in the circular economy. *Internat. J. Production Res.* 59(8):2450–2478.

Reed J, Zhang B (2017) Managing capacity and inventory jointly for multi-server make-to-stock queues. *Queueing Systems* 86:61–94.

Reiman MI, Wang Q (2015) Asymptotically optimal inventory control for assemble-to-order systems with identical lead times. *Oper. Res.* 63(3):716–732.

Resnick SI (1987) *Extreme Values, Regular Variation and Point Processes* (Springer, New York).

Sleptchenko A, van der Heijden MC, van Harten A (2003) Trade-off between inventory and repair capacity in spare part networks. *J. Oper. Res. Soc.* 54(3):263–272.

Song JS (1998) On the order fill rate in a multi-item, base-stock inventory system. *Oper. Res.* 46(6):831–845.

van der Vaart AW (1998) *Asymptotic Statistics,* Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, UK).

van Leeuwaarden 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 Rev.* 61(3):403–440.

Varma S (1990) Heavy and light traffic approximations for queues with synchronization constraints. Unpublished PhD thesis, University of Maryland, College Park, MD.

Wright PE (1992) Two parallel processors with coupled inputs. *Adv. Appl. Probab.* 24(4):986–1007.

Wu J, Chao X (2014) Optimal control of a Brownian production/inventory system with average cost criterion. *Math. Oper. Res.* 39(1):163–189.

Xin L, Goldberg DA (2016) Optimality gap of constant-order policies decays exponentially in the lead time for lost sales models. *Oper. Res.* 64(6):1556–1565.

Xin L, Goldberg DA (2018) Asymptotic optimality of tailored base-surge policies in dual-sourcing inventory systems. *Management Sci.* 64(1):437–452.

Zhang H, Zhang J, Zhang RQ (2020) Simple policies with provable bounds for managing perishable inventory. *Production Oper. Management* 29(11):2637–2650.

Zieliński R (2009) Optimal nonparametric quantile estimators. Toward a general theory. A survey. *Comm. Statist. Theory 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. *Internat. J. Production Res.* 42(22):4673–4689.