Optimization and Coordination in High-tech Supply Chains

Optimization and Coordination in High-tech Supply Chains

This thesis is part of the PhD thesis series of the Beta Research School for Operations Management and Logistics (onderzoeksschool-beta.nl) in which the following universities cooperate: Eindhoven University of Technology, Ghent University, Maastricht University, Tilburg University, University of Twente, VU Amsterdam, Wageningen University and Research, KU Leuven and Universiteit Hasselt.

This work is part of the research programme Complexity in high-tech manufacturing with project number 439.16.121, which is (partly) financed by the Dutch Research Council (NWO).

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-5433-1

Cover picture: *Detroit Industry Murals* (1933) by the Mexican painter Diego Rivera at the Detroit Institute of Art depicting laborers working at Ford Motor Company's River Rouge Plant (www.alamy.com).

Printed by: Proefschriftmaken.nl

TU/e EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

Beta
Research School for Operations
Management and Logistics

Optimization and Coordination in High-tech Supply Chains

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op 2 februari 2022 om 16:00 uur

door

Mirjam Susanne Meijer

geboren te Zwolle

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. F. Langerak
1^{e} promotor:	dr. W.L. van Jaarsveld
2^{e} promotor:	prof. dr. A.G. de Kok
leden:	prof. dr. C.S. Tang (University of California, Los Angeles)
	prof. dr. A.P. Zwart
	prof. dr. S. Transchel (Kühne Logistics University)
	dr. J. Arts (University of Luxembourg)
	dr. Z. Atan

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

1 Introduction

High-tech original equipment manufacturers (OEMs) produce and assemble state-of-the-art systems, consisting of many complex and expensive components or modules. Such high-tech systems are produced in a joint effort by teams of highly specialized engineers employed by the OEM, who acts as system integrator, and dozens of suppliers. An example is the production of lithography machines for the semi-conductor industry by ASML. To produce its end-product, ASML sources components at many suppliers. Over the past decades ASML has introduced multiple new generations of their machines featuring new technologies. Upon introduction of a new generation, the OEM works together with its suppliers to engineer the required components. An important supplier of ASML is VDL/ETG, who produces critical components such as the wafer handler. To assemble their lithography machines and deliver them to the end-customer, it is important for ASML that VDL/ETG and all other suppliers have sufficient capacity to produce the required components. However, some suppliers may be hesitant to invest as unused capacity is costly. Therefore, it is important to coordinate and optimize supply chain decisions. Other examples of high-tech supply chains include the aircraft industry, with Boeing and Airbus sourcing components at many different suppliers, including Yuasa and Saft who produce aircraft batteries, and production of medical imaging equipment by Philips or Siemens.

To assemble the final product and deliver it to the customer, the components supplied by the different suppliers all need to be available. Because of the sheer size and complexity of the supply chain, it is impossible to oversee this entire operation and therefore production is not controlled centrally. Instead, the process is somehow orchestrated by coordinating component inventory replenishment and production planning between upstream parties that produce a sub-component of the system and downstream parties that use these sub-components. Each party manages its operations according to agreements with the other supply chain parties as well as its own objectives. In order to effectively and efficiently operate the supply chain and deliver the end-product to the customer, it is important to align objectives between the different parties and coordinate between their operations.

Supply chain management has been widely studied over the past decades. Recent reviews of literature focusing on assemble-to-order systems in specific and multi-echelon systems in general are provided by Atan et al. (2017) and de Kok et al. (2018), respectively. Cachon (2003) discusses different contracts that aim to coordinate the supply chain, i.e. contracts that guarantee that the optimal control policy in case of centralized control is also attained in the decentralized case where different parties optimize their own objectives. However, some characteristics of high-tech supply chains have profound consequences on application of available theory. For example, most research on penalty contracts, in which the OEM enforces a penalty on the supplier in case the supplier fails to deliver, assumes no limitations on the penalty costs paid by the supplier and hence it is always possible to coordinate the supply chain (see e.g. Sieke et al., 2012). In high-tech supply chains, however, suppliers are typically unable to take over all shortage risk from the OEM. Consequently, penalty contracts may not be feasible as the required penalty may be too high for a supplier to pay. Another stream in literature focuses on contracting based on established capacity (e.g. Brown and Lee, 2003; Roels and Tang, 2017). However, in high-tech supply chains the OEM often cannot verify the exact capacity established by the supplier, since the OEM does not have the knowledge to infer exact production capacity resulting from installed equipment and technically skilled staff. This hinders application of such capacity contracts in high-tech supply chains.

On the other hand, some characteristics of high-tech supply chains offer new opportunities for coordinating the supply chain or optimizing inventory decisions.

In high-tech supply chains there is a certain level of interdependence between an OEM and a supplier of a critical component. For the supplier, the OEM is an important customer that generates a large share of their business. Since critical components are often only sourced at a single supplier, as specific skills and investments are needed, the OEM is dependent on the performance of this supplier for receiving this component. This relationship can be utilized for improving supply chain coordination. Also, the large scale of high-tech assembly systems consisting of many components suggests potential for applying asymptotic analysis, where a parameter, in this case the number of suppliers, approaches an extreme value. We investigate how these characteristics can be used to improve supply chain management in high-tech supply chains.

1.1. Research problems and contributions

We identify three research problems that are relevant for improving coordination and control of high-tech supply chains. These problems are discussed in Sections 1.1.1, 1.1.2 and 1.1.3, which introduce the problems and discuss their relevance, both in theory and in practice.

1.1.1 Overcoming double marginalization by considering multiple product generations

It is a well-known result that simple wholesale price contracts, where a buyer offers a wholesale price to a supplier and the supplier determines how much to deliver, lead to a large loss compared to optimal supply chain profit. This phenomenon is known as double marginalization and results from the fact that the decisions of either party are aimed at optimizing its own profits as opposed to overall supply chain profits. However, since this type of contract is easy to implement in practice, it is commonly used, also in high-tech industries.

Contracts that are aimed at reducing double marginalization are often either difficult to implement in high-tech supply chains, as both parties need to agree on multiple cost or quantity factors early on in the design process, or not enforceable in practice, such as penalty contracts. To illustrate this, we show in Chapter 2 that under a coordinating penalty contract the penalty charged for failing to satisfy demand may be a multiple of the supplier's best-case profit. Instead, we study contracts that are easily implemented in practice and improve supply chain coordination. In high-tech manufacturing, the OEM and a supplier often work together for extensive periods. When an OEM works with a reliable supplier that is performing well, the OEM is likely to continue working with this supplier as new generations of a product are introduced. Since one of the main drivers for a supplier to perform well is the risk of loosing an important customer, we investigate how this can be incorporated in supply chain contracting to counter double marginalization. Specifically, we study whether the prospect of continued collaboration in case of sufficient capacity can be used to motivate a supplier to increase capacity investments.

Research Objective 1 *Develop a supply chain contract that improves coordination by offering the prospect of contract renewal.*

The importance of long-term supplier-buyer relationships in high-tech supply chains has been recognized in literature (Jin and Wu, 2007). However, research on using contract renewal as a motivation for the supplier is limited. Taylor and Plambeck (2007a,b) study informal relational contracts that are aimed at creating an incentive for the supplier to invest by establishing a trust-based relationship over multiple periods. The drawback of these contracts is that they may be complex and difficult to implement in a high-tech setting. In Chapter 2, by introducing a renewal contract where the decision of continuing the collaboration explicitly depends on the supplier's capacity investment, we offer a simple contract that only requires agreements on a wholesale price.

In Chapter 2, we assume that there are many suppliers available. Based on this, once the OEM does not renew the contract the supplier looses an important customer that will never return. Since in reality high-tech components often result from co-development between a supplier and the OEM and require complex technical skills and equipment from the supplier, there is likely a limited number of suppliers with the required capabilities. This means that the OEM deals with oligopolistic suppliers, which increases the suppliers' power. In Chapter 3, we investigate how this affects the applicability of renewal contracts.

Research Objective 2 *Analyze the potential of contingent renewal contracts with performance dependent renewal probability in case of oligopolistic suppliers.*

When the number of suppliers with the required capabilities is limited, once the OEM switches to an alternative supplier, the incumbent supplier expects the OEM to return at some point. This results in competition between potential suppliers, as the incumbent supplier wants to invest in capacity to satisfy demand and retain the OEM's business, but once the OEM switches to an alternative supplier, the time it takes until the OEM returns to the incumbent supplier depends on the capacity investment of the alternative supplier. Competition between supplier's is widely studied in literature, but the focus is mostly on quantity or price competition between suppliers that are simultaneously offering substitute components (Tsitsiklis and Xu, 2014; Hu et al., 2017). Li and Wan (2017) study how fostering supplier competition can be used to support performance. In this case, we focus on a supplier that is the sole supplier of a required component, but risks loosing the OEM's business to a competitor in case of bad performance. The performance of that competitor then influences when the OEM may return to the incumbent supplier. Supplier switching has been studied by Wagner and Friedl (2007) and Pfeiffer (2010), who develop conditions for a buyer, in this case the OEM, for switching. Merckx and Chaturvedi (2020) additionally analyze the trade-off between fostering supplier competition by offering short-term contracts and stimulating performance of the incumbent supplier by offering a single long-term contract. We contribute to this analysis by designing a simple wholesale price contract, where the probability of switching is endogenously determined by the capacity decision of the incumbent supplier.

1.1.2 Synchronization between in-house production and supplier sourcing

In Chapters 2 and 3 we focus on the supply process of a single component. Since high-tech end-products consist of many components and unavailability of one of the components has costly consequences for the production of the end-product, it is important to coordinate between the ordering policy of different supplier-sourced components and the production of components that are produced in-house by the OEM. We study an assembly system with a combination of a fixed lead-time component sourced at an outside supplier and a component that is produced in-house once a customer order is placed. Our objective is to find an optimal inventory policy for the lead-time component.

Research Objective 3 *Determine an optimal inventory policy for synchronizing orders of a component sourced at a supplier with the in-house production process of a high-tech OEM.*

We propose a state-dependent base-stock policy for ordering the fixed lead-time component from the supplier while taking into account the number of outstanding orders for the in-house produced component. We show that this policy is optimal and verify numerically that it generates considerable savings compared to a static policy that disregards this information. These results hold both in continuous time and discrete time, demonstrating applicability in many practical settings.

Synchronization between order policies of different items has been studied. Important results on coordination of orders for items with deterministic lead-time are given by Rosling (1989). Coordination of multiple items with stochastic lead-times was first studied by Benjaafar and ElHafsi (2006). The problem we study in Chapter 4 contains a combination of these cases on which literature is lacking, namely assembly systems consisting of components with a fairly predictable lead-time as well as components with stochastic lead-times arising from a capacitated supply system.

1.1.3 Capacity and inventory decisions in large-scale assembly systems

Since high-tech end-products often consist of many components that need to be available at the time of assembly, we finally study the problem of simultaneously determining capacity and inventory in a large-scale assembly system with many components. When one component is missing, this leads to costly delays in production of the end-product. The probability of delays occurring can be reduced by increasing capacity or keeping inventory, which both also have associated costs. We formulate a stylized model that enables us to study the trade-off between shortage risk, inventory costs, and capacity costs with the following objective:

Research Objective 4 *Derive capacity and inventory decisions for components in a large-scale assembly system that are asymptotically optimal as the number of components goes to infinity.*

Shortages can occur for example when there are disruptions in the component production process or when a peak in demand occurs. Since the demand for all components results from demand of the end-product, delays of the different components are correlated. We use asymptotic analysis to obtain an extreme value result for the delay of the component with the largest backlog as the number of components grows large. We translate these results to asymptotically exact methods for cost-optimal inventory and capacity decisions and numerically evaluate their performance for any realistic number of components.

Simultaneous optimization of capacity and inventory is considered to be a difficult problem (Bradley and Glynn, 2002). Consequently, this is often tackled by studying a sequential optimization problem instead. Recently, simultaneous optimization of capacity and inventory has received more attention (Reed and Zhang, 2017; Reddy and Kumar, 2020). An approach similar to ours is used by Bradley and Glynn (2002), yet they do not encounter the problem of correlated delays as only a single item is considered. We thus contribute by introducing asymptotically exact methods for determining capacity and inventory for a large-scale assembly system with dependent delays. We show numerically that these methods result in costs that are close to the optimal costs already for a practical number of components. Additionally, we introduce an important technical result by showing that the dependence resulting from a common arrival process causes the scaled maximum queue length to converge to a normally distributed random variable as the number of components grows large.

1.2. Outline

This thesis analyzes the problems introduced in Section 1.1 in four Chapters. All chapters can be read individually. Chapter 2, adapted from Meijer et al. (2021d), introduces a contingent renewal contract and compares it to existing contracts. In Chapter 3, which is based on Meijer et al. (2021b), we extend the application of the contingent renewal contract to the case where there is only a limited number of oligopolistic suppliers available. Relevant results from Chapter 2 are highlighted in Chapter 3, such that the chapter can be understood without complete knowledge of the contents of Chapter 2. In Chapter 4, which is based on Meijer et al. (2021c), we apply a combination of queuing theory and Markov processes to derive an optimal order policy for the synchronization problem introduced in Section 1.1.2. In Chapter 5, we provide approximations for both inventory and capacity that are asymptotically optimal as the number of components grows large. This chapter is based on Meijer et al. (2021a). Finally, this theses is concluded in Chapter 6 where we summarize the main findings and their practical implications, followed by some directions for future research.

2

Direct versus Indirect Penalties for Supply Contracts in High-tech Industry

Unlike consumer goods industry, a high-tech manufacturer (OEM) often amortizes new product development costs over multiple generations, where demand for each generation is based on advance orders (i.e., known demand) and additional uncertain demand. Also, high-tech OEMs usually source from a single supplier. Relative to the high retail price, the costs for a supplier to produce high-tech components are low. Consequently, incentives are misaligned: the OEM faces relatively high under-stock costs and the supplier faces high over-stock costs.

In this chapter, we examine supply contracts that are intended to align the incentives between a high-tech OEM and a supplier so that the supplier will invest adequate and yet non-verifiable capacity to meet the OEM's demand. When focusing on a single generation, the manufacturer can coordinate a decentralized supply chain and extract all surplus by augmenting a traditional wholesale price contract with a "contingent penalty" should the supplier fail to fulfill the OEM's demand. When the resulting penalty is too high to be enforceable, we consider a new class of "contingent renewal" wholesale price contracts with a stipulation: the OEM will renew the contract with the incumbent supplier for the next generation *only when*

This chapter is based on Meijer et al. (2021d).

the supplier can fulfill the demand for the current generation. By using non-renewal as an implicit penalty, we show that the contingent renewal contract can coordinate the supply chain. While the OEM can capture the bulk of the supply chain profit, this innovative contract does not enable the OEM to extract the entire surplus.

2.1. Introduction

This chapter examines a supply contract problem arising from a high-tech supply chain that involves an original equipment manufacturer (OEM) who designs and manufactures state-of-the-art systems and a focal supplier who makes a critical component for the system. Examples of high-tech industries (and corresponding OEMs) include commercial aircraft (Boeing, Airbus), medical imaging equipment (Philips, Siemens, GE), and lithography systems for the semiconductor industry (ASML, CANON). Suppliers who make critical components include Yuasa (who makes the batteries for Boeing's 787), VDL/ETG (who makes the wafer handler for ASML), and Neways (who makes control systems for Philips). To reduce development cost and development time in the high-tech industry, most OEMs act as system "integrators": they initiate and develop different generations of a product, which they engineer together with hundreds of suppliers (Tang and Zimmerman, 2009). However, based on our discussion with ASML in the Netherlands, we learned that, unlike mass produced consumer products such as apparel and home furnishings, the sourcing and supplier management for high-tech components are fundamentally different as follows:

1. **Multiple product generations**. Because the research and development cost of a new product involves billions of euros, each new product has multiple generations so that the OEM can continue to improve the design for each generation instead of starting from scratch. Therefore, accounting for the life of different generations, the life cycle of a new product can last for decades. For example, Boeing's 747 was launched in 1970 and retired in 2018.

2. **Advance orders and a highly uncertain additional demand**. Like any high-tech product, the demand for a new product (or a new generation of the product) is highly uncertain because some customers are risk-averse in adopting new technologies especially when they are uncertain about product

performance. To reduce demand uncertainty, many OEMs (e.g., Boeing, ASML) encourage customers to place their orders in advance even though some customers prefer to order after product launch. For example, Boeing received some 900 advance orders for the Boeing 787, but many airlines only ordered after the aircraft was in production (Nolan, 2009). Hence, the *base demand* associated with the advance orders is known, but a highly uncertain *additional demand* remains prevalent.

3. **Single-sourced.** Because the production of high-tech components requires component-specific equipment and technology-specific technical staff, the production *capacity* of a high-tech component is also generation-specific.[1] Also, due to high demand uncertainty, suppliers are reluctant to participate unless it is sole-sourced. For these reasons, the OEM sources each component from a single supplier for each generation. In fact, such supply contracts are often renewed, and OEMs may work with the same supplier for multiple product generations.[2]

4. **Non-verifiable supply capacity.** While the OEM can audit the equipment acquired by the supplier for a specific generation, the actual production capacity is difficult for the OEM to verify. Production often involves technical staff and engineers to operate the required equipment, but the available staff can also be assigned to work on other OEMs' orders. Therefore, the actual production capacity is difficult to judge for the OEM. While the actual capacity is not verifiable ex-ante, the OEM will know the supplier's capacity only when the supplier failed to meet the ex-post realized demand (placed by the OEM).

5. **Very high under-stock cost.** In the high-tech industry, the selling price of each system is in millions of euros. At the same time, the research and development cost is in billions of euros, while the unit cost of a component is relatively very low: it can range from a few hundred euros to hundreds of thousands of euros.

Other issues such as varying product quality (see e.g. Transchel et al., 2016) may also play a role, but are outside the scope of this chapter. The above context creates the

[1] This is because new technology is involved for each new generation, the extended use of equipment for older generation production is normally infeasible or impractical.

[2] Also, due to the underlying advanced technology, OEMs usually work with a single supplier for the development of each component to foster close cooperation with the intention of a longer-term relationship. Examples include Boeing with Alcoa, ASML with VDL/ETG, Philips with Neways, etc.

following challenges faced by many OEMs (e.g., ASML) in the high-tech industry. First, due to high selling price and low unit production cost (excluding the research and development cost), the OEM incurs a very high under-stock cost and would like the supplier to invest ample capacity to meet both the base demand and additional demand. However, under a standard wholesale price contract the supplier has little incentive to invest in ample capacity because such capacity is costly and because the wholesales price is relatively low. Such misaligned incentives have regularly caused shortages of components, which in turn have caused failures of OEMs to meet uncertain demand.[3] Second, due to non-verifiable supply capacity ex-ante, the OEM cannot contract directly based on the reserved capacity as examined in the literature (Brown and Lee, 2003; Roels and Tang, 2017).

These two challenges motivate us to develop a new class of supply contracts that is intended to entice the supplier to invest sufficient capacity to coordinate the decentralized supply chain by aligning the incentives of the OEM and the supplier when capacity is not verifiable ex-ante. In addition to the known base demand, we assume that the additional uncertain demand follows an exponential distribution to ensure tractability. We first examine the classic wholesale price contract arising from a single product generation that may occur when the OEM treats the sourcing decision of different product generations separately. When the OEM pays the supplier a wholesale price for each unit delivered, we re-confirm a well-known result: a traditional wholesale price contract can coordinate the supply chain only when the wholesale price equals the OEM's profit margin, which the OEM will not oblige (Cachon, 2003). However, when the OEM offers a wholesale price while imposing a "shortfall penalty" (i.e., the supplier fails to meet the OEM's demand), we find that the "augmented" wholesale price contract can enable the OEM to coordinate the supply chain while capturing all the profit. Hence, such an augmented contract is optimal to the OEM for managing the supply contract for a single product generation.

While the augmented wholesale price contract is optimal and can coordinate the decentralized supply chain for a single product generation, it may not be practical when the penalty is too high to be enforced upon financially constrained suppliers. Our analysis reveals that this happens primarily when the cost of underage is extremely high, a situation that occurs in the high-tech setting as selling prices

[3] For example, a shortage of aircraft-grade fasteners, allegedly resulting from Alcoa's insufficient capacity investments, caused headline delivery delays of the Boeing 787 (Reuters, 2007).

are much higher than the supplier's capacity and production costs. In those cases, instead of imposing a direct shortfall penalty, we consider a new class of supply contracts that takes the sourcing of multiple product generations into consideration by offering the option of renewing the contract. The aim of studying the augmented wholesale price contract with penalty is thus two-fold. First, by analyzing this contract we illustrate how penalties can be used as motivation, which aids the understanding of the remainder of the chapter, but have shortcomings when applied in high-tech supply chains. Second, the augmented wholesale price contract with penalty serves a the basis on which we build our renewal contract, where we frame non-renewal as indirect penalty.

Specifically, we consider a class of "contingent renewal" wholesale price contracts that can be described as follows: in addition to the wholesale price, the OEM will renew the contract with the incumbent supplier for the production of the next generation only when it has sufficient capacity to fulfill the OEM's demand for the current generation. Observe that this contingent contract creates an "indirect penalty" associated with non-renewal that affects the supplier's future profit,[4] and this contingent contract is enforceable because the OEM has the option to work with different suppliers for different generations (especially when the supplier capacity is generation-specific). Alternatively, we could frame renewal as a reward, with non-renewal being the baseline, as the OEM does not have legal commitment to source from the same supplier for subsequent generations. We use the term penalty for consistency.

By expanding our model to capture the characteristics of multiple product generations, we find that the contingent renewal supply contract can enable the OEM to coordinate the decentralized supply chain; however, the OEM cannot extract the entire surplus from the supplier. Even so, we find that the OEM can capture the bulk of the total profit of the entire supply chain when the selling price is much higher than the supplier's capacity and production costs and when the supplier has a high valuation of future profits or when a substantial fraction of total demand is ordered in advance.

This chapter is organized as follows. We review relevant literature in Section 2.2. In Section 2.3, we focus our analysis for the single product generation case and

[4] Discussions with high-tech manufacturers have revealed that long-term cooperations (and the possibility to terminate them) are key to entice suppliers to comply.

show that, by augmenting the wholesale price contract with a contingent penalty, the OEM can coordinate the supply chain optimally by extracting the entire surplus from the supplier. Section 2.4 extends our analysis to the multi-generation case. By developing a wholesale price contract with endogenous renewal probability, the OEM can coordinate the supply chain but it cannot extract the entire surplus from the supplier. However, the coordinating contingent wholesale price contract enables the OEM to capture the bulk of the total supply chain profit under certain conditions. In Section 2.5 we show that our results continue to hold when we consider for example other demand distributions. The chapter concludes in Section 2.6. All proofs are provided in Appendix 2.A.

2.2. Literature review

Some of the contracts studied in this chapter may be renewed over multiple periods. Taylor and Plambeck (2007b,a) study informal, relational contracts that create incentives for the supplier to invest in production capacity repeatedly over multiple periods. Taylor and Plambeck (2007b) derive the optimal self-enforcing relational contract. However, this contract may be complex and therefore difficult to implement, for which reason an intermediate-complexity relational contract that performs well in many parameter settings is considered as well. Since these contracts are often still difficult to implement in practice, Taylor and Plambeck (2007a) consider simple relational contracts that consist of agreements on price only or on price and quantity. They specify conditions under which a price-only contract is most applicable and when a price-quantity contract is more efficient and compare the performance of the most suitable one to that of the optimal relational contract from Taylor and Plambeck (2007b). They show that for large discount factors, the performance of the simple relational contracts is close to that of the optimal relational contract, but for moderate capacity cost and discount factors the loss compared to the optimal relational contract is substantial. Similar to Taylor and Plambeck (2007b,a), Sun and Debo (2014) show that informal long-term relationships can be sustained when the discount factor of future profits is sufficiently high, taking into account the effect of turbulent markets.

While Taylor and Plambeck focus on comparing various relational contracts, we focus on assessing the relative merits of simple contracts and their applicability inspired by a practical application in high-tech supply chains. We show that wholesale price contracts with penalty can coordinate the supply chain but may not always be enforceable. Although our renewable wholesale price contracts are similar to the price-only relational contracts, our contracts perform better especially when the valuation of future profits is high (cf. Taylor and Plambeck).

Our model explicitly distinguishes between a known base demand (known before product launch) and an uncertain tail demand (revealed after product launch). We derive new analytic results for this model to generate clear insights that are particularly relevant in the high-tech setting. In this setting retail prices are typically high, and for the corresponding asymptotic limit we derive a simple closed-form expression of the division of profit over the OEM and supplier. In this limit, the renewable wholesale price contracts result in the OEM capturing a large share of the profit, which motivates OEMs to adopt this very simple contract in practice. Additionally, we show that a higher known base demand results in a higher share of profit captured by the manufacturer.

Our work is also related to the capacity reservation literature (see e.g. Barnes-Schuster et al., 2002; Brown and Lee, 2003; Erkoc and Wu, 2005; Ren et al., 2010; Roels and Tang, 2017). Unlike these single-period models, our high-tech industry context lends itself to long-term partnerships involving repeated interactions with the supplier (Jin and Wu, 2007). While Serel (2007) considers a multi-period capacity reservation contract between a manufacturer and a long-term supplier, we consider the case when capacity cannot be verified and when the contract renewal hinges on the supplier's performance. Long-term supply contracts are also considered by Frascatore and Mahmoodi (2008). Frascatore and Mahmoodi (2008) conclude that the supplier can be induced to create higher capacity by offering a contract that spans multiple periods and that including a penalty can encourage the supplier to invest in the supply chain optimal capacity. Our model differs significantly from the setting considered by Frascatore and Mahmoodi (2008), as they consider a relation spanning a fixed number of periods while we consider a potentially infinite collaboration where continuation depends on the supplier's investment decisions. Furthermore, we show that direct penalties may not be enforceable in the considered setting. Decisions on continuing a supply chain relationship are studied by Pfeiffer (2010) in a setting with information asymmetry. It is found that the threat of switching between suppliers can be used as an instrument to reduce information asymmetry.

Our model is related to Vendor Managed Inventory (VMI) programs in which the supplier is responsible for replenishing the manufacturer's inventory. The past decades VMI has been investigated increasingly (see e.g. Fry et al., 2001; Lee and Cho, 2018; Hu et al., 2018). Corbett (2001) and Chintapalli et al. (2017) study how supply chain efficiency can be increased by delegating replenishment decisions to the supplier, as inefficiency due to sub-optimal order quantities is reduced. However, these studies focus on one-off interactions or are based on deterministic demand. Guan and Zhao (2010) consider repeated interaction in a stochastic demand setting and find that parties are more willing to share private information when engaging in repeated interactions. Our study is fundamentally different from this stream of work. In our context, the contract renewal probability is endogenously dependent on the supplier's capacity decision and the contract will not be renewed should the supplier fail to fulfill the OEM's uncertain demand.

Finally, the study by Sieke et al. (2012) is closely related to our study. They study supply chain coordination using service level contracts that enforce pre-specified service levels with financial penalty payments. It is concluded that for the considered types of contracts for every service level there exists a contract that coordinates the supply chain, assuming there are no limitations on the penalty costs. We show that in practice there often are limitations on the penalty costs and suggest an alternative type of contracts that calls for non-renewal should the supplier fail to fulfill the OEM's uncertain demand.

2.3. Contracting for one product generation

In this section, we consider a contracting issue arising from the development of one product generation between the OEM and the supplier. This setting occurs when the underlying product has only one generation or when the OEM deals with the supply contract for different product generations separately (i.e., one generation at a time). In Section 2.4 we shall extend our analysis to the case when the OEM takes the development of multiple product generations into consideration when contracting with a supplier.

2.3.1 Centralized supply chain: a benchmark

To establish a benchmark, let us begin by analyzing a supply chain under centralized control in which the OEM and the supplier are managed by a central body. The centralized supply chain makes the capacity decision x "before" demand D is realized. The cost for investing in the requisite capacity is $c \geq 0$ per unit, and the demand (for the single production generation) is $D = b + A$, where $b \geq 0$ is the "base demand" that is known in advance (e.g., advance orders) and A represents the uncertain "additional" demand that is realized after the product is launched, with $A \sim Exp(\lambda)$.[5] (In Section 5.2, we examine numerically the case when A follows an Erlang-n distribution as well as a general demand distribution, for which we obtain similar structural results.) To capture the notion as explained in Section 2.1 that many customers are reluctant to place their orders of high-tech products in advance, we shall assume that the expected additional demand is greater than the base demand so that $\mathbb{E}[A] = \frac{1}{\lambda} \geq b$.

Given the capacity x established at c per unit "before" the demand is realized, the centralized supply chain can produce up to x units at cost k per unit "after" the demand D is realized. If $D \leq x$, then the gross revenue is equal to $(r - k)D$, where r is the exogenously given retail price. To capture the notion that high tech products have a high net profit margin $(r - k - c)$, we shall assume that $r - k - c > c$.[6]

By assuming that unmet demand is lost and the established capacity x has no salvage value,[7] the profit of the centralized supply chain $\Pi(x)$ can be written as:

$$\Pi(x) = -cx + \mathbb{E}[(r-k)\min\{D, x\}] = -cx + \mathbb{E}[(r-k)\min\{b + A, x\}] \qquad (2.1)$$

By considering the first order condition and by using the fact that $A \sim Exp(\lambda)$ along with our assumptions that $r - k - c > c$ and $\frac{1}{\lambda} \geq b$, we obtain:

Proposition 2.1 *When the supply chain is controlled centrally, the optimal capacity*

[5] The decomposition of demand into advance orders and uncertain late orders is for example seen in the case of Boeing, where over 900 advance orders were received before the Boeing 787 was actually produced, while some airlines would place the order only after the new aircraft is in production (Nolan, 2009).

[6] Without the assumption that $r - k - c > c$, additional conditions are necessary in Proposition 2.2 to satisfy the supplier's participation constraint. All other results presented in this chapter continue to hold without this assumption.

[7] Our model can be extended to the case when capacity has salvage value. For ease of exposition, we scale the salvage value to zero.

$x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$ *(where* log *refers to the natural logarithm with base e) and the corresponding optimal supply chain profit* $\Pi^* = (r-k-c)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right) > 0$.

Besides the base demand b, the optimal capacity x^* includes some "additional capacity" $\frac{1}{\lambda}\log\left(\frac{r-k}{c}\right) > 0$ that is intended to satisfy the uncertain additional demand A. This additional capacity is increasing in $\mathbb{E}[A] = \frac{1}{\lambda}$ and the gross margin $(r-k)$, but it is decreasing in the capacity unit cost c. Armed with the "first-best solution" x^* as a benchmark, we now consider the case when the underlying supply chain is decentralized.

2.3.2 Decentralized supply chain: wholesale price contracts

In a decentralized supply chain, the OEM delegates the capacity investment decision to an external supplier and offers a wholesale contract to the supplier based on a wholesale price w (a decision variable). After the OEM has decided on and communicated w, the supplier decides on capacity x, where x is "not verifiable" by the OEM as explained in Section 2.1. In this case, the supplier establishes capacity x at cost c per unit before the demand D is realized. After the demand is realized, the supplier can produce up to x at cost k per unit. Hence, the supplier's profit for any given wholesale price w is:

$$\tilde{\pi}_s(x,w) = -cx + \mathbb{E}[(w-k)\min\{D,x\}] = -cx + \mathbb{E}[(w-k)\min\{b+A,x\}]. \quad (2.2)$$

By considering the first order condition and the fact that $A \sim Exp(\lambda)$, the supplier's optimal capacity is:

$$\tilde{x}(w) = b + \frac{1}{\lambda}\log\left(\frac{w-k}{c}\right) \quad (2.3)$$

if and only if $w \geq c+k$.[8] Substituting Equation (2.3) into Equation (2.2), the corresponding optimal supplier profit is:

$$\tilde{\pi}_s(w) \equiv \tilde{\pi}_s(\tilde{x}(w),w) = (w-c-k)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{w-k}{c}\right). \quad (2.4)$$

[8] Clearly, if the wholesale price is below cost (i.e., $w < c+k$), $\tilde{x}(w) = 0$.

The following lemma asserts that the supplier's participation constraint (i.e., $\tilde{\pi}_s(w) \geq 0$) holds[9] when the wholesale price is greater than the effective unit cost $c+k$. Hence, the supplier participation constraint holds when $w \geq c+k$.

Lemma 2.1 *The supplier's profit $\tilde{\pi}_s(w) = (w-c-k)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{w-k}{c}\right) \geq 0$ if $w \geq k+c$.*

By comparing $\tilde{x}(w)$ given in (2.3) and x^* given in Proposition 1, we get:

Corollary 2.1 *Under the wholesale price contract, one can coordinate the decentralized supply chain so that $\tilde{x}(w) = x^*$ if and only if $w = r$.*

In view of Corollary 2.1, the OEM can coordinate the supply chain by setting $w = r$ to entice the supplier to reserve the capacity $\tilde{x}(r) = x^*$; however, the OEM will earn nothing which is undesirable.

2.3.2.1 Optimal wholesale price

Instead of focusing on supply chain coordination by setting $w = r$, let us consider the case when the OEM optimizes its own profit.

By anticipating the supplier's optimal capacity investment $\tilde{x}(w) = b + \frac{1}{\lambda}\log\left(\frac{w-k}{c}\right)$ as given in (2.3), the OEM's profit associated with any $w \geq c+k$ (to ensure supplier participation) is:

$$\begin{aligned}\tilde{\pi}_m(w) &= \mathbb{E}\left[(r-w)\min\{D, \tilde{x}(w)\}\right] \\ &= \mathbb{E}\left[(r-w)\min\left\{b+A, b+\frac{1}{\lambda}\log\left(\frac{w-k}{c}\right)\right\}\right] \\ &= (r-w)\left(b+\frac{1}{\lambda}\left(1-\frac{c}{w-k}\right)\right). \qquad (2.5)\end{aligned}$$

Hence, the OEM's problem is: $\max_{w \geq c+k} \ \tilde{\pi}_m(w)$. By noting that our assumptions that $r-k-c > c$ and $\frac{1}{\lambda} \geq b$ imply $r-k > (\lambda b+1)c$, the first-order condition yields the optimal wholesale price as follows:

[9] To ease our exposition and without loss of generality, we scale the minimum acceptable profit for the supplier to accept a contract to 0, which is standard approach being used in the supply contract literature (e.g., Cachon (2003)). For completeness, we shall extend our analysis to the case when this minimum acceptable profit is any $Z \geq 0$ in Section 2.5.1.

Proposition 2.2 *In a decentralized supply chain, the optimal wholesale price $\tilde{w}$ satisfies*

$$\tilde{w} = k + \sqrt{\frac{(r-k)c}{b\lambda + 1}},$$

and the corresponding capacity (selected by the supplier) is:

$$\tilde{x} \equiv \tilde{x}(\tilde{w}) = b + \frac{1}{\lambda} \log \left(\sqrt{\frac{r-k}{(b\lambda+1)c}} \right).$$

Because $r - k > (\lambda b + 1)c$, it is easy to check that $\tilde{w} > c + k$ so that the supplier's participation constraint is satisfied. Also, Proposition 2 has the following implications. First, direct comparison between $\tilde{x}$ and x^* given in Proposition 1 reveals that $\tilde{x} < x^*$. Hence, when operating under the wholesale price contract $\tilde{w}$ that maximizes the OEM's profit, the supplier will invest capacity level $\tilde{x}$ that is lower than the first-best solution x^* for the centralized case. Second, it is easy to check that both the OEM's optimal wholesale price $\tilde{w}$ and the supplier's "additional" capacity $\tilde{x} - b$ are increasing in the total margin $r - k$ on the end product and the expected additional demand $\frac{1}{\lambda}$, but decreasing in the base demand b.

By substituting the OEM's optimal wholesale price $\tilde{w}$ and the supplier's capacity $\tilde{x}$ given in Proposition 2 into (2.4) and (2.5), we find:

Proposition 2.3 *When the OEM offers its optimal wholesale price $\tilde{w}$ to the supplier in a decentralized supply chain, the supplier's optimal profit $\tilde{\pi}_s$ and the OEM's optimal profit $\tilde{\pi}_m$ satisfy:*

$$\tilde{\pi}_s = \left(\sqrt{\frac{(r-k)c}{b\lambda+1}} - c \right) b + \frac{\sqrt{\frac{(r-k)c}{b\lambda+1}}}{\lambda} - \frac{c}{\lambda} - \frac{c}{\lambda} \log \left(\sqrt{\frac{r-k}{(b\lambda+1)c}} \right) > 0, \mathit{and}$$

$$\tilde{\pi}_m = \left(r - k - \sqrt{\frac{(r-k)c}{b\lambda+1}} \right) \left(b + \frac{1}{\lambda} \left(1 - \sqrt{\frac{(b\lambda+1)c}{r-k}} \right) \right) > 0.$$

Also, $\tilde{\pi}_s + \tilde{\pi}_m < \Pi^*$, where Π^* is the optimal profit of the centralized supply chain as stated in Proposition 1.

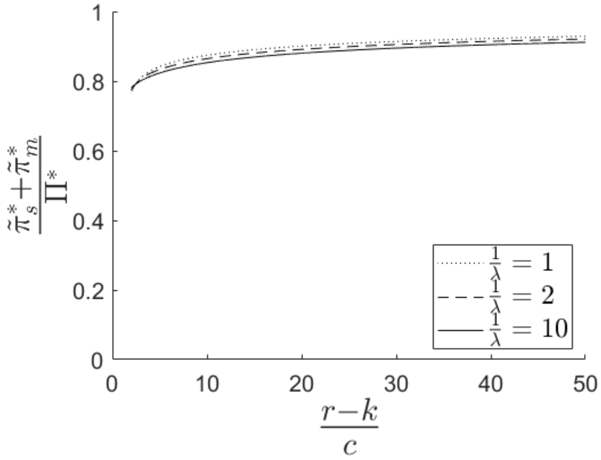

Figure 2.1: Efficiency of the wholesale price contract $\tilde{w}$.

Propositions 2 and 3 reveal that, when operating a decentralized supply chain by using a wholesale price $\tilde{w}$, the supplier will invest at a lower capacity level $\tilde{x}$ (than the first best solution x^*) so that the total profit of the decentralized supply chain is lower than that of the centralized supply chain; i.e., $\tilde{\pi}_s + \tilde{\pi}_m < \Pi^*$.

We now numerically examine the efficiency of the wholesale price contract $\tilde{w}$ (measured in terms of $\frac{\tilde{\pi}_s+\tilde{\pi}_m}{\Pi^*}$) in Figure 2.1. Specifically, we set $b = 1$, but we vary the ratio between gross margin and the capacity investment cost per unit $\frac{r-k}{c}$ from 2 to 50,[10] and set expected additional demand $\mathbb{E}[A] = \frac{1}{\lambda} = 1, 2, 10$. Observe from Figure 2.1 that the wholesale price contract $\tilde{w}$ can be rather inefficient when expected additional demand $\frac{1}{\lambda}$ becomes large, which is the case in the high-tech supply chain context we consider. When $\frac{r-k}{c}$ increases, the inefficiency reduces somewhat, which is mainly due to the fact that the $r - k$ term becomes dominant.

In summary, we find that the wholesale price contract that coordinates the supply chain by setting $w = r$ is not practical, and the wholesale price contract that optimizes the OEM's profit as stated in Proposition 2 is deemed inefficient (Figure 2.1). These observations motivate us to consider ways to "refine" the wholesale price contract in order to improve its efficiency in the next section.

[10] Because of our assumption $(r - k - c) > c$, $\frac{r-k}{c} > 2$.

2.3.3 Augmented wholesale price contracts with lump-sum contingent penalty

Recognizing that the traditional wholesale price contract cannot coordinate the supply chain (unless we set $w = r$) and it is inefficient when we set $w = \tilde{w}$ to maximize the OEM's profit, we now consider an "augmented" wholesale price contract that can be described as follows. In addition to the wholesale price w, the OEM imposes a "lump-sum contingent" penalty ρ that the supplier is liable to pay to the OEM when its capacity x is insufficient to fulfill the realized demand D; i.e., when $D > x$. Once the wholesale price and penalty are known, the supplier determines its optimal capacity. We shall show that this augmented wholesale price contract is optimal: it can coordinate the supply chain and it enables the OEM to attain the first best profit Π^* as in the centralized system.

To analyze the augmented wholesale price contract (w, ρ) via backward induction, let us first determine the supplier's expected profit. In preparation, let us define an indicator function $1_{\{D>x\}}$ that equals 1 if $D > x$ and equals 0, otherwise. Because $D = b + A$, the contingent lump-sum penalty can be expressed as $\rho \cdot 1_{\{b+A>x\}}$. By incorporating this penalty into the supplier's profit given in Equation (2.2), the supplier's profit for any given augmented wholesale price contract (w, ρ) and any capacity x is $\hat{\pi}_s(x, w, \rho)$, where:

$$\hat{\pi}_s(x, w, \rho) = -cx + \mathbb{E}[(w-k)\min(b+A, x) - \rho 1_{\{b+A>x\}}]. \tag{2.6}$$

By differentiating $\hat{\pi}_s(x, w, \rho)$ in Equation (2.6) with respect to x and by considering the first-order condition, the optimal supplier's capacity $\hat{x}(w, \rho)$ for any given augmented wholesale price contract (w, ρ) is:

$$\hat{x}(w, \rho) = b + \frac{1}{\lambda}\log\left(\frac{w-k+\rho\lambda}{c}\right). \tag{2.7}$$

Anticipating the supplier's capacity $\hat{x}(w, \rho)$, the OEM's profit is $\hat{\pi}_m(w, \rho)$, where:

$$\hat{\pi}_m(w, \rho) = \mathbb{E}[(r-w)\min\{b+A, \hat{x}(w, \rho)\} + \rho 1_{\{b+A \geq \hat{x}(w,\rho)\}}] \tag{2.8}$$

By comparing the supplier's capacity $\hat{x}(w, \rho)$ against the first best solution x^* given in Proposition 1, we can identify the following conditions for the augmented wholesale price contract (w, ρ) to coordinate the decentralized supply chain so that $\hat{x}(w, \rho) = x^*$:

Proposition 2.4 *Any augmented wholesale price contract (w, ρ) that satisfies $w + \rho\lambda = r$ can enable the OEM to coordinate the decentralized supply chain so that $\hat{x}(w, \rho) = x^*$. However, among all coordinated augmented contracts, it is optimal for the OEM to set the wholesale price $\hat{w} = k + c + \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ and the contingent lump-sum penalty $\hat{\rho} = \frac{1}{\lambda}\left(r - k - c - \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)\right) \geq 0$ so that the OEM can extract the entire surplus from the supplier; i.e., $\hat{\pi}_s = 0$, $\hat{\pi}_m = \Pi^*$, and $\hat{\pi}_s + \hat{\pi}_m = \Pi^*$.*

Proposition 2.4 reveals that there are infinitely many augmented wholesale price contracts that can coordinate a decentralized supply chain so that the supplier will invest its capacity $\hat{x}(w, \rho) = x^*$. Also, there exists a coordinating contract $(\hat{w}, \hat{\rho})$ that can enable the OEM to extract the entire surplus from the supplier so that her profit is equal to the profit of the entire centrally controlled supply chain. By achieving the first best solution and the highest possible profit, we can conclude that the coordinating contract $(\hat{w}, \hat{\rho})$ is the optimal contract for the decentralized supply chain with 100% contract efficiency (i.e., $\frac{\hat{\pi}_s + \hat{\pi}_m}{\Pi^*} = 1$).

Proposition 2.4 specifies the optimal augmented wholesale price contract for the OEM; however, the implementation of such an optimal contract would depend on the underlying business environment. For instance, the lump-sum penalty $\hat{\rho} = \frac{1}{\lambda}\left(r - k - c - \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)\right)$ can be too high for a financially constrained supplier to pay so that such an optimal augmented wholesale price contract is not enforceable. This situation can occur in the high-tech sector when the OEM's retail price r is much larger than $c + k$ as illustrated in the following numerical example.

A numerical example. Consider the case when $r = 10^7$, $c = 10^5$, $k = 0$, the base demand $b = 50$ and the expected additional demand $\mathbb{E}[A] = \frac{1}{\lambda} = 100$ so that our assumptions $r - k - c > c$ and $\frac{1}{\lambda} \geq b$ hold. By substituting these parameter values into the expressions from Proposition 2.4, we obtain: $\hat{x} = 510$, $\hat{w} \approx 0.4 \cdot 10^6$, and $\hat{\rho} = 959 \cdot 10^6$. Now consider a specific realization of the additional demand where $A = 460$ so that the realized demand $D = b + A = 510 = \hat{x}$. This demand realization represents the best-case scenario for the supplier, under which he can use its entire capacity to fulfill demand without subjecting to any penalty. The profit for the supplier under this specific demand realization is equal to $156.7 \cdot 10^6$. However, considering all possible demand realizations, because $Prob\{A > 460\} \approx 0.01$, there is a 1% chance that $D = b + A > 510 = \hat{x}$. When this happens, the supplier is subject to a penalty $\hat{\rho} = 959 \cdot 10^6$ that is six times his best-case profit of $156.7 \cdot 10^6$ and the supplier is unlikely to be able to pay. Therefore, this example illustrates that, when the retail price r is much higher than $k + c$, the optimal augmented wholesale price contract may not be enforceable even though it is optimal for the OEM.

The above numerical example reveals that, when $(r - k - c)$ is very high, the optimal augmented wholesale price contract may not be enforceable because a "direct" lump-sum penalty is too high for a financially constrained supplier to pay. One could worry that the risk imposed by the supplier is too high in general, since the supplier is held responsible for not meeting uncertain market demands. However, we need to put this in perspective by considering the bigger picture. Suppose that the supplier delivers a range of relatively inexpensive components to the OEM. For each component the supplier and OEM enter an appropriate contract with contingent penalty. Then obviously, for each single component there may be stockouts that the supplier does not have full control over, but still has to pay for. These stockouts are caused by outside influences beyond the control of the supplier, such as exogenous stochastic demand. However, the supplier still controls the risk. Under the wholesale price contract with contingent lump-sum penalty this results in low stockout risks, rendering it likely that the supplier's actual penalty payments are low compared to his earnings. So overall the risk is low. Another option would be switching to a per-unit penalty. In the following section we will investigate whether this would make the augmented wholesale price contract more applicable in high-tech supply chains.

2.3.4 Augmented wholesale price contracts with contingent unit penalty

Instead of a lump-sum shortfall penalty that may appear to be harsh, let us consider the case when the OEM imposes a per-unit shortfall penalty. In this case, the expected penalty no longer equals $\rho\mathbb{E}[1_{\{b+A>x\}}]$, but is equal to $\rho_1\mathbb{E}\left[(b + A - x)^+\right]$ with ρ_1 representing the "per-unit shortfall penalty". The supplier's profit function given by Equation (2.6) is thus adjusted to

$$\hat{\pi}_s(x, w, \rho_1) = -cx + \mathbb{E}[(w - k)\min(b + A, x) - \rho_1(b + A - x)^+] \qquad (2.9)$$

By considering the first order condition, the optimal supplier's capacity $\hat{x}(w, \rho_1)$ for any given augmented wholesale price contract (w, ρ_1) is:

$$\hat{x}(w, \rho_1) = b + \frac{1}{\lambda} \log \left(\frac{w - k + \rho_1}{c} \right). \tag{2.10}$$

In anticipation of the supplier's capacity $\hat{x}(w, \rho_1)$, the OEM's profit is given by:

$$\hat{\pi}_m(w, \rho_1) = \mathbb{E}[(r - w) \min\{b + A, \hat{x}(w, \rho_1)\} + \rho_1(b + A - \hat{x}(w, \rho_1))^+] \tag{2.11}$$

Proposition 2.5 *The augmented wholesale price contract (w, ρ_1) with $\hat{w} = k + c + \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ and $\hat{\rho}_1 = r - k - c - \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ coordinates the supply chain while allowing the OEM to extract the entire surplus.*

Proposition 2.5 shows that besides the coordinating augmented contract with lump-sum shortfall penalty ρ presented in Proposition 2.4, there also exists a coordinating augmented wholesale contract with a per-unit shortfall penalty ρ_1 that allows the OEM to capture the entire surplus. Hence, we can conclude that $(\hat{w}, \hat{\rho}_1)$ is also an optimal contract for the decentralized supply chain with 100% contract efficiency. However, returning to the numerical example, in the following we will show that this contract is equally unlikely to be enforceable in a high-tech setting as the lump-sum penalty contract.

A numerical example. Let us revisit the example discussed in Section 2.3.3, with $r = 10^7$, $c = 10^5$, $k = 0$, base demand $b = 50$ and expected additional demand $\mathbb{E}[A] = \frac{1}{\lambda} = 100$. By substituting these parameter values into the expressions obtained in Proposition 2.5, we find $\hat{x} = 510$ and $\hat{w} \approx 0.4 \cdot 10^6$. We now also have a per-unit penalty $\hat{\rho}_1 = 9.59 \cdot 10^6$, meaning that for every component that the supplier is unable to supply, he incurs a penalty that is nearly 24 times as high as the wholesale price he receives for every unit supplied.

This numerical example illustrates that when $(r - k - c)$ is very high, the optimal augmented wholesale price contract may not be enforceable, whether a lump-sum penalty is used or a per-unit penalty. Upon discussing with an OEM, we discovered a different form of "indirect" penalty that is enforceable by the OEM when it sources components of multiple product generations over time. We explore this indirect penalty next.

2.4. Renewable wholesale price contracts for multiple generations

We now consider the case when the OEM develops multiple generations of a high-tech product. As explained in Section 2.1, because different generations are based on different technologies, the supplier's capacity is generation-specific: the extended use of the capacity designated for one generation to the next generation is not possible. Therefore, to produce components for a new product generation, the supplier needs to invest in new capacity. Because the capacity is generation-specific, the OEM has the option to work with different suppliers for different generations if the incumbent supplier's performance is unsatisfactory.[11]

We learnt from an OEM in the Netherlands that, even though there is an implicit understanding that the OEM would normally renew its contract with the incumbent supplier for the next generation, there is no explicit commitment for contract renewals and there are no explicit conditions for contract non-renewals. This revelation motivates us to examine a class of wholesale price contracts with "contingent renewals": the OEM will renew the contract with the incumbent supplier for the next generation only if the supplier can fulfill the OEM's demand for the current generation. By specifying the condition for renewal/non-renewal explicitly, the OEM can use contract non-renewal as an "indirect" penalty that the OEM can enforce (as opposed to the lump-sum penalty that may not be enforceable).

In this section, we shall extend our single generation model presented in Section 2.3 to the multi-generation case by incorporating the issue of contingent contract renewals as described above. Our intent is to examine the coordinating capability and the efficiency of the contingent wholesale price contract. To obtain tractable results, we shall assume that the demand for each generation $D_t = b_t + A_t$ with $b_t = b$ for every product generation t and A_t are i.i.d. exponential random variables with mean $\mathbb{E}[A_t] = \frac{1}{\lambda}$. To account for inflation or other factors that may affect costs over time, we scale costs for every generation t with a multiplicative factor $\beta \geq 1$. Hence, by letting $c_1 = c$, $k_1 = k$ and $r_1 = r$, the adjusted costs are: $c_t = \beta^{t-1}c$,

[11] In the high-tech industry, it is common practice for the OEM to work with the incumbent supplier to ensure continuity and smooth transition between product generations unless the supplier's performance is unsatisfactory.

$k_t = \beta^{t-1}k$ and $r_t = \beta^{t-1}r$. Additionally, we introduce $0 < \delta < 1$ to discount future profits. Assume that $\delta\beta < 1$.

2.4.1 Centralized supply chain with renewals

We first establish a benchmark by considering a centrally controlled supply chain for producing multiple generations with renewals. Because all costs scale with the same factor β, the optimal capacity decision is generation-independent. Hence, for any capacity x invested for any generation t, the profit obtained by the centralized supply chain for this generation is equal to $\beta\Pi(x)$, with $\Pi(x)$ as stated in Equation (2.1). Then the net present value (NPV) of the total supply chain profit with renewals over all generations is:

$$\begin{aligned}\Pi^{\theta}(x) = \sum_{t=0}^{\infty} \delta^t \beta^t \Pi(x) = \frac{1}{1-\delta\beta}\Pi(x) = \frac{1}{1-\theta}\Pi(x) \\ = \frac{1}{1-\theta} \cdot \left(-cx + \mathbb{E}[(r-k)\min\{b+A, x\}]\right), \quad (2.12)\end{aligned}$$

where $\theta = \delta\beta$. By considering the first order condition and by using the fact that $A \sim Exp(\lambda)$ along with our assumptions that $r - k - c > c$ and $\frac{1}{\lambda} \geq b$, we can apply Proposition 2.1 to show that the optimal capacity for any generation is $x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$. Hence, due to identically distributed demand per generation and all costs scaling with factor β, the same optimization problem is faced for every generation. Therefore, the optimal capacity decision is the same as in the single-generation centralized supply chain. Similarly, the optimal discounted total profit for the centralized supply chain (associated with the discount factor θ) is denoted by Π^{θ}, where $\Pi^{\theta} = \frac{1}{1-\theta}\left((r-k-c)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)\right) \geq 0$.

2.4.2 Wholesale price contracts with exogenous renewal probability

We now consider a decentralized supply chain in which an OEM establishes a supply contract with a focal supplier. Similar to the decentralized case for a single generation, the OEM determines the wholesale price, after which the supplier decides on capacity. In addition, the OEM can decide whether or not to renew the contract with the incumbent supplier. Thus, the sequence of events is as follows:

1. OEM determines wholesale price

2. Supplier invests in production capacity

3. Demand is realized and supplier produces required components up to the maximum capacity

4. OEM decides whether or not to renew contract with supplier

In Section 2.4.3, we shall analyze the contingent wholesale price contract under which the renewal for the next generation depends on whether the supplier's capacity can satisfy the OEM's demand for the current generation. To explicate our analysis, let us first consider a base case in which contract renewal is based on an "exogenously" given real-valued probability R that is independent of the supplier's capacity decision x. Once a contract is not renewed, we assume that the supplier will never be allowed to work with the OEM in the future. This assumption implies that the number of generations Y that the incumbent supplier can work with the OEM follows a geometric distribution so that $Y \sim Geom(1-R)$ and $Prob\{Y=t\} = R^{t-1}(1-R)$ for $t = 1, 2, \ldots$. We assume that the wholesale price scales with the same factor β so that $w_1 = w$ and $w_t = \beta^{t-1} w_1$. Hence, the supplier's capacity decision x is generation-independent because the distribution of demand D remains the same for all generations and all cost parameters scale with the same factor β. Consequently, the expected profit for the supplier in each generation t is $\beta^{t-1}\tilde{\pi}_s(x,w)$, with $\tilde{\pi}_s(x,w)$ as stated in Equation (2.2). By combining these observations, the NPV of the supplier's expected profit over Y product generations can be expressed as:

$$\pi_s^\theta(x) = \mathbb{E}[\sum_{t=1}^{Y} \theta^{t-1}\tilde{\pi}_s(x,w)] = \sum_{t=1}^{\infty} P(Y \geq t)\theta^{t-1}\tilde{\pi}_s(x,w)$$
$$= \sum_{t=1}^{\infty}(R\theta)^{t-1}\tilde{\pi}_s(x,w) = \frac{\tilde{\pi}_s(x,w)}{1-\theta R}. \quad (2.13)$$

It follows from Equation (2.13) that the term $\frac{1}{1-\theta R}$ is independent of x, so we can conclude that, for any given R, the NPV of the supplier's profit is maximized when he maximizes his single-period profits $\tilde{\pi}_s(x,w)$. Combining this observation with our analysis presented in Section 2.3.2, we can conclude that, when the contract renewal probability R is exogenously given, it is optimal for the supplier to set its capacity according to $\tilde{x}(w) = b + \frac{1}{\lambda}\log(\frac{w-k}{c})$ as stated in Equation (2.3), where $\tilde{x}(w)$ is independent of R. Since the supplier's capacity investment does not affect the renewal decision, there is no incentive for the supplier to invest in additional capacity and the optimal capacity remains the same as in the single-generation case. Also, we can use the same approach as presented in Corollary 2.1 to show that a wholesale price contract with exogenously given renewal probability R can coordinate the supply chain (i.e., $\tilde{x}(w) = x^*$) only when we set $w = r$, which the OEM will not oblige.

Instead of coordinating the supply chain, the OEM can seek to maximize her own profit. Because the distribution of demand D is generation-independent and all costs scale with parameter β, we can use the same approach to show that the optimal wholesale price $w_1 = \tilde{w}$ is given in Proposition 2.2. Thus the corresponding contract is inefficient (cf. Proposition 2.3).

Cognizant of the shortcomings of the wholesale price contracts with exogenous renewal probability R, we now examine the coordinating capability and the efficiency of the wholesale price contracts with renewal probability $R(x)$ that is "endogenously" determined by the supplier.

2.4.3 Wholesale price contracts with endogenous renewal probability

We now extend our analysis presented in the previous section to the case when the contract will be renewed for the next generation only if the supplier's capacity x can meet with the OEM's demand D for the current generation. This way, the renewal probability $R(x)$ is now "endogenously" dependent on the capacity x to be selected by the supplier. Based on our assumptions that $D = b + A$ and $A \sim Exp(\lambda)$ are generation-independent, it is easy to check that, for any given supplier capacity x, the renewal probability $R(x) = Prob\{D \leq x\} = Prob\{A \leq (x-b)\} = 1 - e^{-\lambda(x-b)}$ for $x \geq b$, and $R(x) = 0$; otherwise. Hence, for any given supplier capacity x, we can use the same approach as presented in Section 2.4.2 to show that the number of generations $Y(x)$ that the incumbent supplier can work with the OEM will follow a geometric distribution so that $Y \sim Geom(1 - R(x))$, where the renewal probability $R(x)$ is defined above. As before, because the supplier's capacity decision x is generation-independent (because the distribution of demand D is generation-independent and all cost parameters scale with β), the expected profit for the supplier in each generation t is $\beta^{t-1}\tilde{\pi}_s(x,w)$, with $\tilde{\pi}_s(x,w)$ as stated in Equation (2.2). These observations imply that the NPV of the supplier's expected profit over $Y(x)$ product generations can be expressed as:

$$\begin{aligned}\tilde{\pi}_s^{\theta}(x) = \mathbb{E}[\sum_{t=1}^{Y(x)} \theta^{t-1}\tilde{\pi}_s(x,w)] &= \sum_{t=1}^{\infty} P(Y(x) \geq t)\theta^{t-1}\tilde{\pi}_s(x,w) \\ &= \sum_{t=1}^{\infty}(R(x)\theta)^{t-1}\tilde{\pi}_s(x,w) = \frac{\tilde{\pi}_s(x,w)}{1-\theta R(x)}, \end{aligned} \tag{2.14}$$

where the renewal probability $R(x) = 1 - e^{-\lambda(x-b)}$ for $x \geq b$, and $R(x) = 0$; otherwise.

By considering the first-order condition and the assumption that $w \geq c+k$, we can determine the supplier's optimal capacity $\tilde{x}(w,\theta)$ as follows:

Proposition 2.6 *For any given wholesale price contract w (that has $w \geq c+k$) with endogenous renewal probability $R(x)$, the supplier's optimal capacity satisfies:*

$$\tilde{x}(w,\theta) = b + \frac{1}{\lambda}\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{b\lambda(w-k-c)}{c}}\right)\right) \tag{2.15}$$

where $W(\cdot)$ is known as the Lambert W function.[12]

Analogous to the supplier's optimal capacity $\tilde{x}(w)$ given in Equation (2.3) for the single-generation case, the supplier's optimal capacity $\tilde{x}(w,\theta)$ is equal to the base demand b plus some additional capacity to cover the uncertain additional demand A. Unlike $\tilde{x}(w)$, observe from Equation (2.15) that the additional capacity

$$\frac{1}{\lambda}\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{b\lambda(w-k-c)}{c}}\right)\right)$$

created by $\tilde{x}(w,\theta)$ under the "contingent" wholesale price contract depends on the base demand b. The reason lies in the fact that, when b increases, the supplier values contract renewals more because it can obtain a higher profit through the base demand. Consequently, as the base demand becomes bigger, the supplier has stronger incentive to invest in more capacity to increase its renewal probability. Besides the impact of the base demand b, it is easy to show that as the combined

[12] The Lambert W function is the inverse function of $f(x) = xe^x$.

discount factor θ increases, the supplier will value contract renewals more; hence, the supplier will increase its capacity as θ increases so that $\tilde{x}(w, \theta)$ is increasing in $\theta \in (0, 1)$. Thus high-tech manufacturers can benefit from seeking out suppliers that have a focus on the long-term profit.

Next, through direct comparison between $\tilde{x}(w)$ given in Equation (2.3) for the single-generation case (which corresponds to the multi-generation case with exogenous renewal probability R and yet $\tilde{x}(w)$ is independent of R as explained in Section 2.4.2) and $\tilde{x}(w, \theta)$ given in Equation (2.15) for the multi-generation case with endogenous renewal probability $R(\tilde{x}(w, \theta))$, we get:

Proposition 2.7 *For any given wholesale price contract w (that has $w \geq c + k$), the supplier's optimal capacity $\tilde{x}(w, \theta)$ corresponding to endogenous renewal probability $R(\tilde{x}(w, \theta))$ is higher than its optimal capacity $\tilde{x}(w)$ corresponding to any exogenous renewal probability R (i.e., $\tilde{x}(w, \theta) > \tilde{x}(w)$).*

While the supplier's capacity (i.e., both $\tilde{x}(w, \theta)$ and $\tilde{x}(w)$) is increasing in the wholesale price w, the above proposition suggests that the stipulated condition for contract renewal provides an incentive for the supplier to invest more capacity. This is also illustrated in Figure 2.2, where we show the supplier's optimal capacity both in case of an endogenous renewal probability ($\tilde{x}(w, \theta)$) and in case of an exogenous renewal probability ($\tilde{x}(w)$) for different values of wholesale price w, with $b = 1$, $\lambda = 1$, $\theta = 0.9$, $c = 1$ and $k = 0$. We observe that the optimal capacity under the contract with endogenous renewal probability is indeed considerably higher than in case of an exogenous renewal probability, for the same wholesale price.

2.4.3.1 Supply chain coordination

Recall from Section 2.4.2 that the wholesale price contract with exogenous renewal probability R can coordinate the supply chain (i.e., $\tilde{x}(w) = x^*$) only when the OEM sets $w = r$, which the OEM will not oblige because of zero profit. We now examine whether the use of endogenous renewal probability $R(x)$ would enable the OEM to coordinate the supply chain. By considering the supplier's optimal capacity $\tilde{x}(w, \theta)$ given in Equation (2.15) and x^* given in Proposition 1, obtain the result in Proposition 2.8.

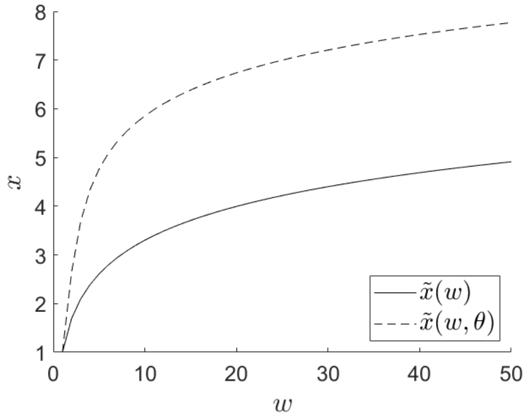

Figure 2.2: Supplier's optimal capacity for exogenous and endogenous renewal probabilities for different wholesale prices, with $b = 1$, $\lambda = 1$, $\theta = 0.9$, $c = 1$ and $k = 0$.

Proposition 2.8 *Suppose the OEM sets its wholesale price for each generation at w^θ, where:*

$$w^\theta = k + \frac{\theta c \left(1 + b\lambda + \log\left(\frac{r-k}{c}\right)\right) + (1-\theta)(r-k)}{1 + \theta b\lambda} < r.$$

Then corresponding contingent wholesale price contract with endogenous renewal probability $R(\tilde{x}(w^\theta, \theta))$ can coordinate the supply chain so that $\tilde{x}(w^\theta, \theta) = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right) = x^*$.

The above proposition reveals that, by imposing an explicit condition for contract renewal with the incumbent supplier, the OEM can leverage the indirect penalty associated with non-renewal to entice the supplier to select its capacity according to the first best solution x^* by offering the wholesale price w^θ. Since $w^\theta < r$, supply chain coordination is achieved at a strictly lower wholesale price compared to the wholesale price of r that is required for coordination when renewal is exogenous (cf. Section 2.4.2).

In Figure 2.3 we analyze numerically how the coordinating wholesale price changes with the margin on the end-product for different discount factors θ, where $c = 1$, $k = 0$, $b = 1$ and $\lambda = 1$. We observe that the coordinating wholesale price that the OEM pays to the supplier increases with $\frac{r-k}{c}$. When the margin on the end product is larger, the supplier will require a higher wholesale price to build the first-best capacity x^*. However, when the valuation of future profits by the supplier

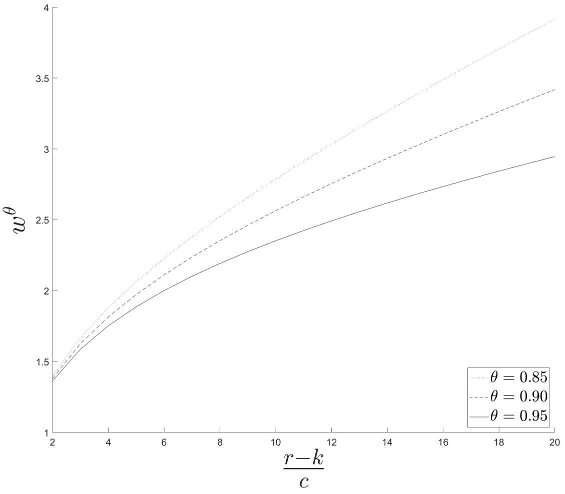

Figure 2.3: Coordinating wholesale price as a function of $\frac{r-k}{c}$ for different values of θ

(represented by θ) is larger, the increase in the coordinating wholesale price as $\frac{r-k}{c}$ increases becomes smaller as the supplier is more inclined to invest in future profits.

Let us now make a rough comparison between the single-generation wholesale price contract and the wholesale price contract with contingent renewals for such a high-tech end-product, based on a real-world case. The value of the end-product is around 60 million euros. The costs of capacity for the components are estimated at 4.1 million euros investment costs and 1 million euros production costs. Expected demand equals 20 units, of which 10 units are ordered in advance. In our model, this means we have $r = 60 \cdot 10^6$, $c = 4.1 \cdot 10^6$, $k = 1 \cdot 10^6$, $b = 10$ and $\lambda = \frac{1}{10}$, with $\theta = 0.96$. For the single-generation wholesale price contract, the OEM's optimal wholesale price equals roughly 12 million euro resulting in a capacity of 20 units, with the OEM's expected profit around 781 million euros. Under the renewal contract, the OEM offers a wholesale price of 11.6 million euros to entice the supplier to build the coordinating capacity of 37 units. The expected profit for the OEM in this case equals 989 million euros, which is 1.2 times as high as for the single-generation case. This illustrates that besides leading to higher total supply chain profits due to coordination, the benefit for the OEM of using contingent renewal is large. Next, we will further analyze the OEM's share of the profits.

2.4.3.2 Supplier Surplus Extraction

Proposition 2.8 shows that the contingent wholesale price contract can coordinate the supply chain by offering a wholesale price $w^\theta < r$. We now examine whether such coordinating contract can enable the OEM to extract the entire surplus from the supplier so that the corresponding contract is optimal. In preparation, let us first compute the NPV of the OEM's profit over all generations. For any given supplier capacity x and wholesale price w, the OEM will earn a profit $\beta^{t-1}\tilde{\pi}_m(x,w)$ for generation t, with $\tilde{\pi}_m(x,w) = \mathbb{E}[(r-w)\min\{b+A,x\}]$, regardless of which supplier its works with. In other words, even though the OEM may work with different suppliers upon contract non-renewals, the OEM's profit for each generation stays the same and it is independent of the contract renewal probability $R(x)$. Because of the assumptions that demand $D_t = b_t + A_t$ with $b_t = b$ and A_t i.i.d. for all t and all cost parameters scale with factor β, the NPV of the OEM's profit over all generations can be expressed as:

$$\tilde{\pi}_m^\theta(x,w) = \sum_{t=1}^{\infty} \theta^{t-1}\tilde{\pi}_m(x,w) = \frac{\tilde{\pi}_m(x,w)}{1-\theta} \tag{2.16}$$

Now, the OEM offers a wholesale price $w^\theta(< r)$ as stated in Proposition 2.8 (along with the contingent renewal condition) to entice the supplier to set $\tilde{x}(w^\theta,\theta) = x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$. Hence, we can use the fact that $A_t \sim Exp(\lambda)$ to determine the NPV of the OEM's profit as $\tilde{\pi}_m^\theta = \frac{1}{1-\theta}\left(r - w^\theta\right)\left(b + \frac{1}{\lambda}(1 - \frac{c}{r-k})\right) > 0$. By comparing $\tilde{\pi}_m^\theta$ against the optimal NPV of the centralized supply chain Π^θ as defined in Section 2.4.1, we get:

Proposition 2.9 *Under the coordinating contract (w^θ) with endogenous renewal probability, the OEM cannot extract the entire surplus from the supplier: the fraction of the NPV of the total supply chain profit captured by the OEM $\frac{\tilde{\pi}_m^\theta}{\Pi^\theta} < 1$.*

By considering different values of b and θ, for $\lambda = 1$, Figure 2.4 depicts the fraction of the NPV of the total supply chain profit captured by the OEM (given by $\frac{\tilde{\pi}_m^\theta}{\Pi^\theta}$) under the coordinating contract (w^θ) with endogenous renewal probability as a function of the margin on the end product. Figure 2.4 verifies that the fraction is strictly below 1. Also, from these figures, we notice that the OEM can capture a larger proportion of the NPV of the total supply chain profit when θ is large. This is due to the fact that a supplier that has a high valuation of future profits requires

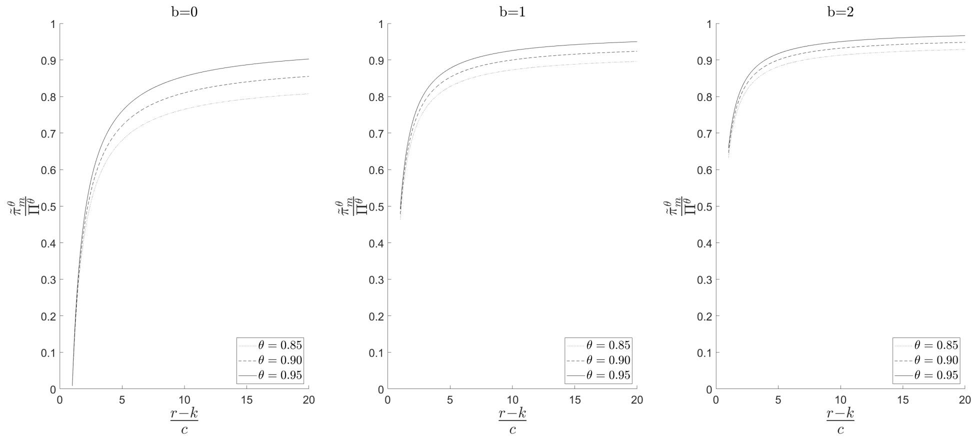

Figure 2.4: Fraction of the NPV captured by the OEM ($\tilde{\pi}_m^\theta/\Pi^\theta$) under the coordinating contingent wholesale price contract (w^θ), for different values of the base demand b.

less incentive to invest sufficient capacity. Similarly, the supplier is more willing to invest when the base demand, resulting in a certain profit, is higher, which results in a higher fraction of the profit for the OEM. Furthermore, the fraction of the NPV captured by the OEM is higher when his added value and thus $\frac{r-k}{c}$ is large.

Next, to gain analytic insights into whether the coordinating contract with endogenous renewal probability is suitable in the high-tech setting, we investigate the fraction of NPV captured by the OEM for the following special case of our general model that is very relevant in the high-tech setting, as explained in Section 2.1.

Special Case 1: When the ratio $\frac{r-k}{c}$ is very large (or the margin $r-k-c$ is very large). The following proposition characterizes the fraction of the optimal NPV captured by the OEM ($\tilde{\pi}_m^\theta/\Pi^\theta$) as defined in Proposition 2.9 when $\frac{r-k}{c} \to \infty$. This limiting case is of interest because it corresponds to the situation when the penalty under the augmented wholesale price contract is exorbitant to enforce as explained in Section 2.3.3.[13]

[13] The extreme case reflects situations where the supplier delivers a crucial part of the system developed by the OEM, with a value that is much lower than the selling price of the product, as is typical in high-tech manufacturing: $r-k$ corresponds to the gross margin when selling the product (e.g. an aircraft, a wafer-stepper), while c corresponds to the costs of capacity for producing a component of that product (e.g. a wing section, a wafer handler).

Proposition 2.10 *Suppose the OEM offers the coordinating contract (w^θ) with endogenous renewal probability. Then, when $\frac{r-k}{c} \to \infty$, the fraction of the NPV of the total supply chain profit captured by the OEM $\tilde{\pi}_m^\theta / \Pi^\theta \to \frac{\theta b\lambda+\theta}{\theta b\lambda+1}$.*

The above proposition has the following implications. First, when the base demand $b = 0$, the limit of the fraction $\tilde{\pi}_m^\theta / \Pi^\theta$ is equal to θ. Hence, when the base demand is low and when the combined discount factor is high, the OEM can extract the bulk of the surplus from the supplier by adopting the coordinating contingent wholesale price contract with endogenous renewal probability. Second, when the base demand b is substantial in comparison to $\mathbb{E}[A] = 1/\lambda$, the limit of the fraction $\tilde{\pi}_m^\theta / \Pi^\theta$ will approach 1. Hence, when a large portion of total demand is obtained through advance orders, the OEM can extract almost the entire surplus from the supplier. This means that having a strong market position, demonstrated by many advance orders, also gives the OEM a strong position vis-à-vis her supplier. These two observations enable us to characterize the business environment (i.e., when $\frac{r-k}{c}$ is high, base demand b is substantial compared to additional demand $\mathbb{E}[A]$, or the combined discount factor θ is high) in which the coordinating wholesale price contract with contingent renewal can enable the OEM to extract the bulk of the surplus from the supplier so that this contract is close to optimal.

Special Case 2: When the ratio $\frac{r-k}{c}$ is close to but strictly greater than 1 (or the margin $r-k-c$ is very small). This situation occurs when the supplier's capacity cost k and unit cost c are high, such that $\frac{r-k}{c} \to 1^+$.[14] By considering the case when $\theta \to 1$ we can compare the coordinating wholesale price w^θ given in Proposition 2.8 for the multi-generation case associated with the contingent renewal contract and the coordinating contingent penalty $\hat{p}$ given in Proposition 2.4 for the single generation case as examined in Section 2.3.3.

Proposition 2.11 *When $\frac{r-k}{c} \to 1^+$ and $\theta \to 1$, the coordinating wholesale price w^θ given in Proposition 2.8 for the multi-generation case satisfies: $w^\theta \approx r$. However, the coordinating contingent penalty $\hat{p}$ given in Proposition 2.4 for the single generation case satisfies: $\hat{p} \approx 0$.*

The above proposition has the following implications: when the margin $r-k-c$ is very small and θ is close to 1, the contingent penalty $\hat{p}$ is very small so that the augmented wholesale price contract is easily enforceable. Also, as revealed

[14] $\frac{r-k}{c} \to 1^+$ denotes $\frac{r-k}{c}$ approaches 1 from the right so that $\frac{r-k}{c} > 1$.

in Proposition 2.4, it allows the OEM to capture the entire supply chain profit. However, the wholesale price w^θ is close to r so that the contingent renewal contract renders the OEM essentially profitless. Therefore, when the margin $r - k - c$ is very small and θ is close to 1, the OEM is better off to treat each generation separately by adopting the contingent penalty contract instead of the contingent renewal contract for multiple generations.

Based on the analysis of these two special cases, we can make the following conclusions. First, the long-term supply contract with contingent renewals is effective for the OEM when the margin $r - k - c$ is very large. This is because, when the margin $r - k - c$ is very large, the contingent penalty $\hat{\rho}$ is exorbitant so that the augmented contract with contingent penalty is deemed impractical. However, the contingent renewal contract performs well because it enables the OEM to obtain almost the entire supply chain profit. Second, the short-term contingent penalty contract is more efficient for the OEM when the margin $r - k - c$ is very small and θ is close to 1. This is because, in this case, the coordinating contingent wholesale price $w^\theta \approx r$, leaving very little profit for the OEM under the contingent renewal contract. However, the contingent penalty $\hat{\rho}$ is very small so that the augmented contract with contingent penalty can be easily enforced and yet it enables the OEM to obtain the entire supply chain profit.

2.4.3.3 Duration of collaboration

Now that we have established the importance of long-term collaborations in high-tech supply chains, the question can be asked how long these collaborations will last. Since the duration of the collaboration Y is distributed geometrically with parameter $1 - R(x)$, where $R(x)$ is the renewal probability, the expected duration of the collaboration equals $\mathbb{E}[Y(x)] = \frac{1}{1-R(x)} = \frac{1}{e^{-\lambda(x-b)}}$. Proposition 2.12 shows that the duration of the collaboration when the OEM sets coordinating wholesale price w^θ, to induce the supplier to set capacity $x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$, is equal to $\frac{r-k}{c}$. This means that the higher the value of the end-product, the longer the collaboration lasts. Also, due to our assumption that $r - k - c > c$, we find:

Proposition 2.12 *Under the coordinating wholesale price w^θ, the expected duration of the collaboration is $\mathbb{E}[Y(x)] = \frac{1}{1-R(x)} = \frac{r-k}{c} > 2$.*

2.4.3.4 Optimal wholesale price renewal contract

Now that we have shown that a renewal contract with endogenous renewal probability can coordinate the supply chain while yielding a positive profit to both parties, the question remains whether the manufacturer has incentive to set this coordinating wholesale price. Since optimization of the wholesale price under an endogenous renewal probability is intractable, we answer this question based on numerical experiments. Since the OEM's profit function is concave, the optimal wholesale price can be determined numerically using Golden-section search.

We consider several instances with $c = 1$ and take $r - k \in \{10, 50, 100\}$, $B \in \{0, 1\}$, $\lambda \in \{0.1, 0.5, 1\}$ and $\theta \in \{0.85, 0.9, 0.95\}$. This gives in total $3^3 \cdot 2 = 54$ instances. For every instance we calculate the profit for the OEM under the contingent renewal contract with endogenous renewal probability for both the coordinating wholesale price and the OEM's optimal wholesale price and determine the percentage difference. In addition, we determine the expected number of generations that the collaboration will last, both under the optimal and the coordinating wholesale price. The summarizing statistics of the full factorial experiment are given in Table 2.1.

We can observe that in all instances the OEM looses some money by setting the coordinating wholesale price instead of optimizing the wholesale price. When we first consider the effect of $r - k$, we observe that the profit lost by coordinating the supply chain reduces as the value of the end product increases. Furthermore, we observe that θ has a large effect on the difference in profit. When θ increases, the difference in profit reduces considerably. The same holds for base demand b. The parameter values for which the difference in profit between choosing the optimal and coordinating wholesale price is smallest thus correspond to the case for which the coordinating contract is most suitable, namely a high value end-product and high valuation of future profits. When we additionally consider the duration of the collaboration, we observe that under both the optimal and coordinating wholesale price the collaboration is expected to span multiple generations. Furthermore, we observe that collaboration lasts considerably longer under the coordinating wholesale price than under the optimal wholesale price, especially for high-valued end-products.

Table 2.1: Results full factorial experiment on the difference in OEM's profit per period between using the coordinating (coord.) and optimal (opt.) wholesale price

		Avg profit per period		Difference profit (%)			Avg duration	
		Opt.	Coord.	Average	Max	Min	Opt.	Coord.
$r-k$	5	9.98	9.27	6.81	11.89	1.71	3.21	5.00
	10	29.23	27.81	4.67	8.97	0.96	5.79	10.00
	20	71.91	68.55	4.46	8.78	0.90	9.88	20.00
θ	0.85	36.36	33.46	8.15	11.89	4.07	5.55	11.67
	0.9	36.99	35.21	5.16	8.40	2.34	6.16	11.67
	0.95	37.76	36.96	2.62	5.28	0.90	7.17	11.67
b	0	31.89	30.09	6.52	11.89	2.08	6.32	11.67
	1	42.19	40.33	4.11	10.37	0.90	6.27	11.67
λ	0.1	78.69	74.49	6.12	11.89	1.84	6.31	11.67
	0.5	19.87	19.01	5.19	11.89	1.26	6.29	11.67
	1	12.56	12.12	4.63	11.89	0.90	6.28	11.67
Overall		12.56	12.12	5.31	11.89	0.90	6.29	11.67

2.5. Extensions

This section analyzes the effect of (1) a supplier's reservation profit and (2) more general demand distributions on the effectiveness of the different supply contracts and the corresponding division of profit between the supplier and the OEM.

2.5.1 Supplier's reservation profit

Until now we have assumed that the supplier will engage as long as the expected profit is non-negative. However, it is likely that a supplier will request a positive expected profit to justify his efforts. Since the contingent renewal contract already guarantees a positive profit for both parties, we investigate in this section how including a reservation profit, denoted by Z, affects our analysis of the single-generation contracts.

2.5.1.1 Single-generation wholesale price contract

Since the supplier's reservation profit does not affect the policy parameters, the supplier's profit function as given in Equation (2.2) remains the same. In Lemma 2.1, we analyzed for which values of wholesale price w the supplier's expected profit is non-negative. Analogously, Lemma 2.2 gives the minimum wholesale price for which the supplier attains the reservation profit.

Lemma 2.2 *The supplier's profit* $\tilde{\pi}_s(w) = (w - c - k)\left(b + \frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{w-k}{c}\right) \geq Z$ *if and only if* $w \geq k - \frac{c}{b\lambda+1} W\left(-(b\lambda + 1)e^{-(b\lambda+1+Z\frac{\lambda}{c}}\right)$.

This means that the supplier will engage in the wholesale price contract proposed in Proposition 2.2 provided that the condition given in Lemma 2.2 is satisfied. If the OEM's optimal wholesale price does not satisfy this condition, the OEM will need to pay a higher than optimal wholesale price to the supplier, leaving the OEM with lower profits.

2.5.1.2 Augmented wholesale price contract with lump-sum penalty

Under the OEM's optimal augemented wholesale price contract with lump-sum penalty, which was proposed in Proposition 2.4 in Section 2.3.3, the OEM was able to capture the entire supply chain profit. When the supplier has a reservation profit $Z > 0$, this is no longer possible. In this case, it will be optimal for the OEM to determine the policy parameters w and ρ for which the supplier builds the first best capacity x^* and the supplier's expected profit is exactly equal to the reservation profit. The details of this optimal policy are given in Proposition 2.13.

Proposition 2.13 *Any augmented wholesale price contract* (w, ρ) *that satisfies* $w + \rho\lambda = r$ *enables the OEM to coordinate the decentralized supply chain so that* $\hat{x}(w, \rho) = x^*$. *However, given the supplier's reservation profit* Z, *it is optimal for the OEM to set the wholesale price* $\hat{w} = k + c + \frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right) - Z\frac{\lambda}{b\lambda+1}$ *and the contingent lump-sum penalty* $\hat{\rho} = \frac{1}{\lambda}\left(r - k - c - \frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right) - Z\frac{\lambda}{b\lambda+1}\right)$ *such that* $\hat{\pi}_s = Z$, $\hat{\pi}_m = \Pi^* - Z$, *and* $\hat{\pi}_s + \hat{\pi}_m = \Pi^*$. *Furthermore,* $\hat{\rho} \geq 0$ *for* $Z \leq \Pi^*$.

The optimal augmented wholesale price contract with unit penalty can be analyzed accordingly.

2.5.2 Other demand distributions

To obtain tractable analytical results and to capture the "long tail" demand characteristics of high-tech products, we have assumed that $D = b + A$, where the base demand b is deterministic and the uncertain demand A is exponentially distributed. We now investigate whether our structural results would continue to hold when demand follows a more general distribution.

We will consider demand distributions that preserve the characteristics of high-tech industries. Specifically, we need to focus on demand distributions that only allow for positive demands and have a long tail, because there is a small but positive probability of very large demand. (Considering demand distributions with a finite upper bound on the domain leads to a fundamentally different problem that is less interesting, as it can be guaranteed that the supplier has sufficient capacity when the capacity decision equals the maximum demand.) In view of the long tail demand characteristics, we start by considering the case when A follows an Erlang-n distribution. Thereafter, we will show that our results continue to hold for general demand distributions.

2.5.2.1 Erlang-n distributed additional demand

Consider the case when the demand $D = b + A$, where $A \sim Erlang(\lambda, n)$ so that the uncertain portion of the demand A follows an Erlang distribution for any given value of n, yielding different coefficients of variation than the Exponential distribution. Due to the subsequent optimization of capacity and wholesale price with the renewal probability endogenously determined by these values, analytical analysis is intractable and we resort to numerical analysis. To do so, we use the expected sales defined in Lemma 2.3, such that we can express the supplier's and OEM's expected profit functions that we analyze numerically. Then we use a Golden-section search procedure to find the optimal capacity for a given wholesale price and bisection search to find the coordinating wholesale price.

Lemma 2.3 *For $A \sim Erlang(\lambda, n)$ and capacity x, the expected sales is given by:*

$$\begin{aligned}\mathbb{E}\left[\min\{D, x\}\right] &= b + \mathbb{E}\left[\min\{A, x-b\}\right] \\ &= b + \frac{\gamma(k+1, \lambda(x-b))}{\lambda(n-1)!} + (x-b)\mathbb{P}(A > x-b) \qquad (2.17)\end{aligned}$$

where $\gamma(s, x) = \int_0^x t^{s-1}e^{-t}dt$ is the lower incomplete gamma function.

It can be verified numerically that also for Erlang-distributed demand, an augmented wholesale price contract with lump-sum penalty may not be enforceable in high-tech supply chains. For example, let us return to the numerical example introduced in Section 2.3.3, with $r = 10^7$, $c = 10^5$ and $k = 0$ and assume that $b = 50$ and $A \sim Erlang\left(\frac{3}{100}, 3\right)$ such that again $\mathbb{E}[D] = 150$. We find that under the OEM's optimal contract $\hat{w} \approx 0.14 \cdot 10^6$ and $\hat{\rho} \approx 25.2 \cdot 10^6$. The supplier's optimal

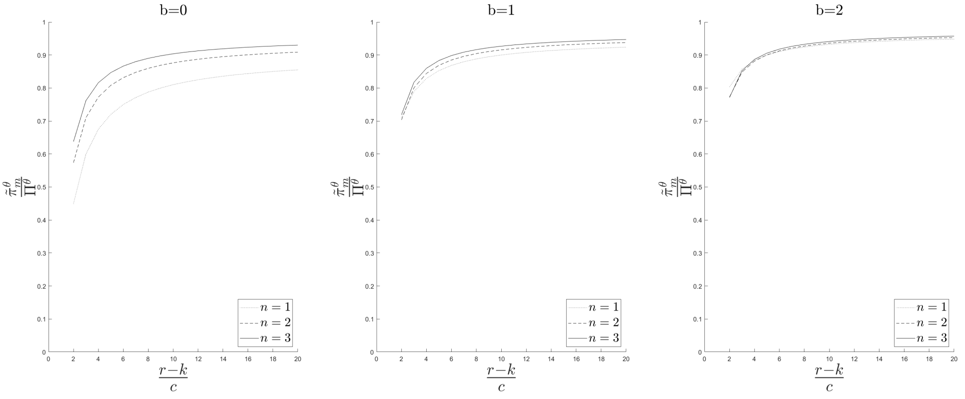

Figure 2.5: Fraction of the NPV captured by the OEM ($\tilde{\pi}_m^\theta/\Pi^\theta$) under the coordinating contingent wholesale price contract with Erlang-n demand.

capacity then equals $\hat{x} \approx 191$. When realized demand is exactly equal to $\hat{x}$, the supplier can use its entire capacity and earn $6.7 \cdot 10^6$, without being subjected to the penalty. However, when realized demand does exceed the available capacity, the supplier is subject to a penalty $\hat{\rho} \approx 25.2 \cdot 10^6$ that is 3.7 times as high as his best-case profit. Hence, also for $A \sim Erlang\left(\frac{3}{100}, 3\right)$, the optimal augmented wholesale price contract may not be enforceable when r is much larger than $c + k$.

Now that we have established that also for Erlang distributed demand an augmented wholesale price contract may not be suitable, we will examine whether a wholesale price contract with endogenous renewal probability again offers a suitable alternative. We analyze whether under the coordinating wholesale price both parties can earn a positive profit and how the total profits are divided. We can observe from Figure 2.5 that under the coordinating wholesale price, the division of profits between the supplier and the OEM for different Erlang distributions with $\lambda = 1$ for $\theta = 0.9$ has similar characteristics as the division that was found in Figure 2.4 (with Erlang-1 being equal to the exponential distribution). More specifically, when a substantial part of demand is fixed, the distribution of profits is similar for the different values of Erlang shape parameter n.

In Section 2.4.3.2, we found that the fraction of the profit captured by the OEM increases with θ. From Figure 2.6 we observe that also for $b = 1$ and $A \sim Erlang(1, 3)$, θ has a positive effect on the share of profit captured by the OEM. In summary, even when we extend our analysis to the case when the uncertain

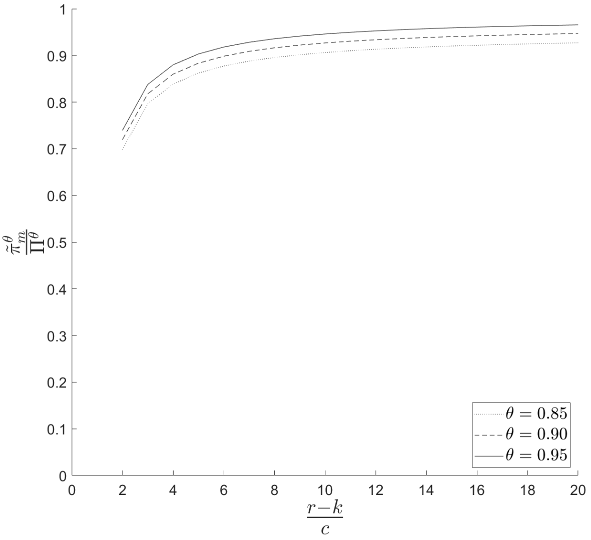

Figure 2.6: Fraction of the NPV captured by the OEM ($\tilde{\pi}_m^\theta / \Pi^\theta$) under the coordinating contingent wholesale price contract with Erlang-3 demand for different values of θ.

demand A follows the $Erlang(\lambda, n)$ distribution, the structural results presented in Section 2.4.3 as depicted in Figure 2.4 continue to hold.

2.5.2.2 General distributions

Instead of considering the case when the demand $D = b + A$, we now assume the demand D follows a general probability density function $f(d)$ so that the corresponding cumulative density function is $F(d)$. Let $R(x)$ again denote the renewal probability, which is increasing in capacity x. Additionally, we assume that $R(0) = 0$ and $R(x) > 0$. Akin to Proposition 2.7, Proposition 2.14 demonstrates that, when D follows a general demand distribution, the wholesale price contract with endogenous renewal probability motivates the supplier to build more capacity than under a single-generation wholesale price contract.

Proposition 2.14 *For any demand distribution that has support on the non-negative real numbers, the supplier's optimal capacity under the wholesale price contract with endogenous renewal probability $R(x)$, denoted by $\tilde{x}(w, \theta)$, is higher than the optimal capacity $\tilde{x}(w) = F^{-1}\left(\frac{w-k-c}{w-k}\right)$ under a single-generation wholesale price contract for any $w > c + k$.*

Using Proposition 2.14, we can also show that, under the wholesale price contract with endogenous renewal probability, the supply chain can be coordinated with a lower wholesale price than with a single-generation wholesale price contract. For any wholesale price w we have established in Proposition 2.14 that $\tilde{x}(w) < \tilde{x}(w, \theta)$. Since $\tilde{x}(w) = F^{-1}\left(\frac{w-k-c}{w-k}\right)$ is increasing in w, it follows that $\tilde{x}(w, \theta)$ is also increasing in w. Let $\tilde{w}$ be the wholesale price such that $\tilde{x}(\tilde{w}) = x^*$ and define w^θ such that $\tilde{x}(w^\theta, \theta) = x^*$. Then it follows that $\tilde{w} > w^\theta$. Thus, the required wholesale price to entice the supplier to build coordinating capacity x^* is lower for the renewal contract than for a single-generation wholesale price contract.

The question then remains whether we can extend our result that under the coordinating contract both parties are expected to earn positive profits for general demand distributions. Proposition 2.15, demonstrates that this result continues to holds.

Proposition 2.15 *For any demand distribution that has support on the non-negative real numbers, both parties are expected to earn positive profit for every generation.*

2.6. Discussion and conclusions

Motivated by our discussions with different European OEMs in the high-tech industry, we have examined different types of supply chain contracts between a high-tech OEM who designs and manufactures multiple generations of a state-of-the-art system and the supplier of a critical component of this system. Different from the existing literature, our model captures certain unique characteristics that are prevalent in the high-tech industry: demand consists of advance orders and a highly uncertain additional demand; components for each generation are single-sourced; capacity established by the supplier is not verifiable by the OEM; and the under-stock cost is very high, or, equivalently, the selling price is very high. Consequently, the cost of underinvestment by the supplier under a standard wholesale price contract is high and the OEM seeks for opportunities to entice the supplier to invest in more capacity.

Our work complements existing literature by examining two supply contracts that are of practical relevance to the OEMs in the high-tech industry: (a) augmented wholesale price contracts with contingent penalty for a single generation; and (b) expanded wholesale price contracts with contingent renewal for multiple generations. By examining the equilibrium outcomes, we have established the following results. First, the augmented wholesale price contract can coordinate the supply chain and it is optimal (in the sense that the OEM can capture all profit from the supply chain). Second, the contingent renewal contract can coordinate the supply chain but the OEM cannot capture the entire supply chain profit. Third, the augmented wholesale price contract is more efficient when the margin is very low, but the contingent renewal contract is more practical when the margin is very high. More specifically, in a high-tech setting, where the margin of the end product is usually large, an augmented wholesale price contract may not be enforceable in practice, while in this case the OEM can earn a large share of total profits when using a contingent renewal contract. Such a contingent renewal contract is especially attractive for the OEM when expected demand per period is large or when the supplier has a high valuation of future profits, since the supplier will be more willing to invest in capacity under these circumstances. Long-term collaborations are thus not only useful, but also essential for the functioning of high-tech supply chains.

Even though our problem setting differs from that of Taylor and Plambeck (2007b,a) in important ways, there are some similarities in the obtained results. Most importantly, both their studies and our study have shown that the gain from long-term collaboration is largest when the valuation of future profits is high. In this setting, Taylor and Plambeck (2007a) recommend a price-quantity contract. Our analytic result for $\frac{r-k}{c}$ large demonstrates that for θ large, even the price-only contract may perform well, which is important since such contracts are easier to adopt in the high-tech setting. We also found that the OEM can capture more profit when the amount of base demand increases.

Additionally, we analyzed the effect of including a positive reservation profit of the supplier and of using other demand distributions on our results. When considering the supplier's reservation profit we find that even though the supplier's expected profit is equal to the reservation profit, the augmented wholesale price contract with lump-sum penalty faces the same problems as without reservation profit. Hence, the main difficulties of the augmented wholesale price contract are not countered by including a positive reservation profit. Next, we showed by means of numerical analysis that the obtained results do not only hold when demand consists of a fixed base demand and an exponential part, but extend to Erlang-distributed demand.

On a higher level, we showed that our structural results even extend to general demand distributions.

Several key insights for high-tech manufacturers can be derived from our results. First, supply chain coordination in high-tech manufacturing may be vastly improved if OEMs make the sourcing of components for a new product generation *dependent* on the performance of suppliers for the previous product generation. For this, it is important that the sourcing department focuses not only on costs when selecting suppliers, but also on supply chain performance. Hence two-way communication between the sourcing and the operations departments at the OEM is important. Second, to reap actual benefits from making contract renewal dependent on supplier performance, suppliers must be made aware of this policy either through formal agreements or through clear communications. Finally, our insight that manufacturer profit depends on the supplier discount factor demonstrates that manufacturers in the high-tech industry may benefit from working with suppliers that focus on long term sustainable business rather than short term profit.

Even though the model presented in this study provides insights into collaborations in high-tech supply chains and shows the value of establishing long-term interactions, it has several limitations for further examination in the future. First, in our model we assume that the demand function is generation-independent. However, there may be trends in demand that are not captured by this model. Therefore, it is of interest to extend our model to generation-dependent demand functions, e.g. $b_t = b_0\alpha^t$, with $\alpha > 1$ the growth factor of demand. Another interesting aspect is when production equipment (i.e. capacity investment) can be used for more than 1 product generation.

2.A. Proofs

Proof of Lemma 2.1

Let $f(y) = (y-k-c)(b+\frac{1}{\lambda}) - \frac{c}{\lambda}\log\left(\frac{y-k}{c}\right)$. We now prove the claim that $f(y) \geq 0$ if $y \geq k+c$. To begin, let $z = \frac{y-k}{c} - 1$ so that $f(z) = cz(b+\frac{1}{\lambda}) - \frac{c}{\lambda}\log(1+z)$. By noting that $f(0) = 0$ and $f(z)$ is increasing and convex in $z \geq 0$ and by noting that $z \geq 0$ when $y \geq k+c$, we can conclude that $f(y) \geq 0$. □

Proof of Proposition 2.1

For any $x > b$ we have:

$$\begin{aligned}\Pi(x) &= -cx + \mathbb{E}[(r-k)\min\{b+A, x\}] \\ &= -cx + (r-k)b + \frac{r-k}{\lambda}\left(1 - e^{-\lambda(x-b)}\right)\end{aligned}$$

Taking the derivative w.r.t x gives

$$\frac{d}{dx}\Pi(x) = -c + (r-k)e^{-\lambda(x-b)}.$$

Thus, using that $r-k > 2c$, we find that $\frac{d}{dx}\Pi(x) = 0$ iff:

$$x = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right).$$

Hence, since $(r-k)/c > 1$, this implies that the x that maximizes $\Pi(x)$ must satisfy $x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$ Also, the corresponding supply chain profit is:

$$\begin{aligned}\Pi^* &= -c\left(b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)\right) + (r-k)b + \frac{r-k}{\lambda}\left(1 - \frac{c}{r-k}\right) \\ &= (r-k-c)\left(b + \frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right).\end{aligned}$$

Now, to prove that Π^* is positive, note that $\Pi^* = (r-k-c)(b+\frac{1}{\lambda}) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right) = f(r)$, with $f(\cdot)$ as in the proof of Lemma 2.1. From the lemma and the assumption $r > k+c$, it then follows that $\Pi^* > 0$. □

Proof of Proposition 2.2

Taking into account the supplier's capacity decision, the OEM's profit is:

$$\tilde{\pi}_m(w) = (r-w)b + \frac{r-w}{\lambda}\left(1 - \frac{c}{w-k}\right)$$

of which the derivative w.r.t w equals

$$\frac{d}{dw}\tilde{\pi}_m(w) = -\left(b + \frac{1}{\lambda}\right) + \frac{1}{\lambda}\frac{(r-k)c}{(w-k)^2}$$

Clearly, $\frac{d^2}{dw^2}\tilde{\pi}(w) < 0$ for $w > k$ and thus $\tilde{\pi}_m(w)$ is maximized by setting the first derivative equal to 0, yielding

$$\tilde{w} = k + \sqrt{\frac{(r-k)c}{b\lambda+1}}.$$

Since $r-k-c > c \to r-k > 2c$, and since $\frac{1}{\lambda} > b \to b\lambda+1 < 2$, we find $\tilde{w} = k + \sqrt{\frac{(r-k)c}{b\lambda+1}} > k + \sqrt{2c^2/2} = k+c$, thus the supplier participation constraint is satisfied. The resulting capacity developed by the supplier follows by substituting this in Equation (2.3)

$$\tilde{x} = b + \frac{1}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right).$$

□

Proof of Proposition 2.3

We first prove that the manufacturers and suppliers profit are strictly positive. Note that the manufacturer's profit is obtained by substituting $\tilde{w}$ into the manufacturer's profit function $\tilde{\pi}_m(w)$. Since $r-k-c > c \to r-k > 2c$, and since $\frac{1}{\lambda} > b \to b\lambda+1 < 2$, we find $\tilde{w} = k + \sqrt{\frac{(r-k)c}{b\lambda+1}} > k + \sqrt{2c^2/2} = k+c$. Since $r-k \leq r$, $c < r$, and $b\lambda+1 \geq 1$, we also have $\tilde{w} < r$. Thus $\tilde{\pi}_m = \tilde{\pi}_m(\tilde{w}) = (r-\tilde{w})(b + \frac{1}{\lambda}(1 - \frac{c}{\tilde{w}-k}) > 0$. The suppliers profit is obtained by substituting $\tilde{w}$ in $\tilde{\pi}_s(w)$. Note that $\tilde{\pi}_s(\tilde{w}) = f(\tilde{w})$, with $f(\cdot)$ as in the proof of Lemma 2.1. That $\tilde{\pi}_s > 0$ then follows from Lemma 2.1 and because $\tilde{w} > k+c$, as was shown above.

To prove that the total profit is lower than in the centralized case, we substitute the supplier's capacity and OEM's wholesale price decisions provided in Proposition 2.2 in the supplier's and OEM's profit functions given in Equations (2.4) and (2.5). This confirms our expressions in the Proposition:

$$\tilde{\pi}_s = \left(\sqrt{\frac{(r-k)c}{b\lambda+1}} - c\right) b + \frac{\sqrt{\frac{(r-k)c}{b\lambda+1}}}{\lambda} - \frac{c}{\lambda} - \frac{c}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right), \text{ and}$$

$$\tilde{\pi}_m = \left(r-k-\sqrt{\frac{(r-k)c}{b\lambda+1}}\right)\left(b+\frac{1}{\lambda}\left(1-\sqrt{\frac{(b\lambda+1)c}{r-k}}\right)\right).$$

Supply chain profit thus equals

$$\begin{aligned}
\tilde{\pi}_s + \tilde{\pi}_m = &\left(\sqrt{\frac{(r-k)c}{b\lambda+1}} - c\right) b + \frac{\sqrt{\frac{(r-k)c}{b\lambda+1}}}{\lambda} - \frac{c}{\lambda} - \frac{c}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right) \\
&+ \left(r-k-\sqrt{\frac{(r-k)c}{b\lambda+1}}\right)\left(b+\frac{1}{\lambda}\left(1-\sqrt{\frac{(b\lambda+1)c}{r-k}}\right)\right) \\
=&(r-k-c)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right) \\
&- \frac{r-k-\sqrt{\frac{(r-k)c}{b\lambda+1}}}{\lambda}\sqrt{\frac{(b\lambda+1)c}{r-k}} \\
=&(r-k-c)\left(b+\frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right) - \frac{1}{\lambda}\sqrt{(r-k)(b\lambda+1)c} + \frac{c}{\lambda}
\end{aligned}$$

Since

$$\frac{c}{\lambda}\log\left(\sqrt{\frac{r-k}{(b\lambda+1)c}}\right) + \frac{1}{\lambda}\sqrt{(r-k)(b\lambda+1)c} - \frac{c}{\lambda} > \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)$$

or equivalently

$$\sqrt{\frac{(r-k)(b\lambda+1)}{c}} > 1 + \log\left(\sqrt{\frac{(r-k)(b\lambda+1)}{c}}\right)$$

it follows that $\tilde{\pi}_s + \tilde{\pi}_m < \Pi^*$. □

Proof of Proposition 2.4

Suppose that the supplier participates in the penalty contract. Then his profit equals:

$$\begin{aligned}\hat{\pi}_s(x,w,\rho) &= -cx + \mathbb{E}[(w-k)\min(b+A,x) - \rho 1_{\{b+A>x\}}] \\ &= (w-c-k)b - c(x-b) + \frac{w-k}{\lambda}\left(1-e^{-\lambda(x-b)}\right) - \rho e^{-\lambda(x-b)}.\end{aligned}$$

By taking the derivative with respect to x, and setting it equal to zero, we find the capacity level at which the supplier maximizes the profit of participating in the contract:

$$\hat{x}(w,\rho) = b + \frac{1}{\lambda}\log\left(\frac{w-k+\rho\lambda}{c}\right).$$

Now note that by Proposition 2.1, the supply chain profit is maximized if $x = b + \frac{1}{\lambda}\log(\frac{r-k}{c})$. In view of the above, a necessary condition to achieve this is that $\frac{w-k+\rho\lambda}{c} = \frac{r-k}{c} \to \rho\lambda + w = r$. To arrive at a coordinating contract in which the OEM captures all the profit, we must in addition set ρ and w such that the profit of participating for the supplier equals 0. The supplier's profit equals:

$$\hat{\pi}_s(\hat{x}(w,\rho),w,\rho) = (w-c-k)b + \frac{w-k-c}{\lambda} - \frac{c}{\lambda}\log\left(\frac{w-k+\rho\lambda}{c}\right)$$

we impose $\rho\lambda + w = r$ by substituting $w = r - \rho\lambda$, which yields

$$\hat{\pi}_s(\rho) = (r-\rho\lambda-c-k)b + \frac{r-\rho\lambda-k-c}{\lambda} - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)$$

We in addition impose $\hat{\pi}_s(\rho) = 0$, which holds for:

$$\hat{\rho} = \frac{1}{\lambda}\left(r-k-c-\frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right)\right).$$

Note that $\hat{\rho} > 0$ whenever $r - k - c > \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$, or equivalently $\frac{r-k}{c} > 1 + \frac{1}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$. For $\frac{r-k}{c} = 1$ both sides would equal 1. For $\frac{r-k}{c} > 1$ the left-hand side increases linearly in $\frac{r-k}{c}$, while the right-hand side increases logarithmically. Also, since $b\lambda > 0$, the fraction $\frac{1}{b\lambda+1} < 1$. Therefore, using our earlier assumptions on r, b and $\frac{1}{\lambda}$ it holds that $\frac{r-k}{c} > 1 + \frac{1}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ and it thus follows that $\hat{\rho} > 0$. To satisfy $r = w + \rho\lambda$, the OEM sets

$$\hat{w} = k + c + \frac{c}{b\lambda + 1} \log\left(\frac{r-k}{c}\right).$$

By construction $\hat{\pi}_s(\hat{x}(\hat{w}, \hat{\rho}), \hat{w}, \hat{\rho}) = 0$ and it follows that

$$\begin{aligned} \hat{\pi}_m(\hat{w}, \hat{\rho}) &= \left(r - k - c - \frac{c}{b\lambda + 1} \log\left(\frac{r-k}{c}\right)\right)\left(b + \frac{1}{\lambda}\right) \\ &= \Pi^*. \end{aligned}$$

□

Proof of Proposition 2.5

The proof is along the same lines as for Proposition 2.4. Now the supplier's profit equals:

$$\begin{aligned} \hat{\pi}_s(x, w, \rho) &= -cx + \mathbb{E}[(w-k)\min(b + A, x) - \rho_1(b + A > x)^+] \\ &= (w - c - k)b - c(x - b) + \frac{w-k}{\lambda}\left(1 - e^{-\lambda(x-b)}\right) - \frac{\rho_1}{\lambda} e^{-\lambda(x-b)}. \end{aligned}$$

Since $\hat{x}(w, \rho_1) = b + \frac{1}{\lambda} \log\left(\frac{w-k+\rho_1}{c}\right)$, it follows that to achieve coordination it must hold that $w + \rho_1 = r$. Under the coordinating contract that allows the OEM to capture all profits, it must thus hold that:

$$(r - \rho_1 - c - k)b + \frac{r - \rho_1 - k - c}{\lambda} - \frac{c}{\lambda} \log\left(\frac{r-k}{c}\right) = 0.$$

Consequently, the augmented wholesale price contract with unit penalty $\hat{\rho}_1 = r - k - c - \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ and wholesale price $\hat{w} = k + c + \frac{c}{b\lambda+1} \log\left(\frac{r-k}{c}\right)$ coordinates the supply chain, while allowing the OEM to extract the entire surplus. For the same reasons as in Proposition 2.4, it holds that $\hat{\rho}_1 > 0$, $\hat{\pi}_s(\hat{x}(\hat{w}, \hat{\rho}_1), \hat{w}, \hat{\rho}_1) = 0$ and

$\hat{\pi}_m(\hat{w}, \hat{\rho}_1) = \Pi^*$. □

Proof of Proposition 2.6

Let $y = e^{-\lambda(x-b)}$. Then we can rewrite Equation (2.14) as

$$\tilde{\pi}_s^\theta(y) = \frac{1}{1-\theta(1-y)}\left(b(w-k-c) + \frac{c}{\lambda}\log(y) + \frac{w-k}{\lambda}(1-y)\right)$$

The first derivative with respect to y is given by

$$\frac{d}{dy}\tilde{\pi}_s^\theta(y) = \frac{(1-\theta(1-y))\left(\frac{c}{\lambda}\frac{1}{y} - \frac{w-k}{\lambda}\right) - \left(b(w-k-c) + \frac{c}{\lambda}\log(y) + \frac{w-k}{\lambda}(1-y)\right)\theta}{(1-\theta(1-y))^2}$$

Equating the derivative to 0 yields

$$(1-\theta(1-y))\left(\frac{c}{\lambda}\frac{1}{y} - \frac{w-k}{\lambda}\right) = \left(b(w-k-c) + \frac{c}{\lambda}\log(y) + \frac{w-k}{\lambda}(1-y)\right)\theta$$

which can be rewritten as

$$(w-k-\theta c + \theta b(w-k-c)\lambda)\, y + \theta c y\, \log(y) = (1-\theta)c$$

which is of the form

$$a_1 y + a_2 y \log(y) = a_3$$

with $a_1, a_2, a_3 > 0$ since $w \geq c+k$, $b \geq 0$, $\lambda > 0$ and $0 < \theta < 1$. Therefore, we can rewrite this as

$$\begin{aligned}
e^{a_1 y + a_2 y \log(y)} &= e^{a_3} \\
e^{a_2 y \log(y)} &= e^{a_3 - a_1 y} \\
e^{\log(y)} &= e^{\frac{a_3 - a_1 y}{a_2 y}} \\
y &= e^{\frac{a_3}{a_2 y} - \frac{a_1}{a_2}} \\
y e^{\frac{a_1}{a_2}} &= e^{\frac{a_3}{a_2 y}} \\
\frac{a_3}{a_2} e^{\frac{a_1}{a_2}} &= \frac{a_3}{a_2 y} e^{\frac{a_3}{a_2 y}}
\end{aligned}$$

$$W\left(\frac{a_3}{a_2}e^{\frac{a_1}{a_2}}\right) = \frac{a_3}{a_2 y}$$

to reach solution

$$y = \frac{a_3}{a_2 W\left(\frac{a_3}{a_2}e^{\frac{a_1}{a_2}}\right)}$$

where W is the Lambert's W function. Substituting $a_1 = w - k - \theta c + \theta b(w - k - c)\lambda$, $a_2 = \theta c$ and $a_3 = (1-\theta)c$ and simplifying the expression yields

$$y = \frac{1-\theta}{\theta W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}}\right)}$$

Since $y = e^{-\lambda(x-b)}$, we obtain

$$\tilde{x}(w,\theta) = b + \frac{1}{\lambda}\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}}\right)\right)$$

For the supplier participation constraint, note that Lemma 2.1 proved that setting capacity $\tilde{x}(w)$ when offered a wholesale price $w > k + c$ yields positive expected profits $\tilde{\pi}_s(\tilde{x}(w), w)$ per generation for the supplier whenever $w > k + c$. Hence, for any $w > k + c$, we must have that $\tilde{\pi}_s^{\theta}(\tilde{x}(w)) = \tilde{\pi}_s(\tilde{x}(w), w)/(1 - \theta R(\tilde{x}(w)))$. This shows that for any $w > k + c$, there exists a capacity level that yields positive profits for the supplier, hence the supplier will participate in any contingent wholesale contract with $w > k + c$. □

The main text claims the following, and we present a formal proof here for completeness.

Proposition 2.16 *For given w, $\tilde{x}(w,\theta)$ is increasing in $\theta \in (0,1)$.*

Proof of Proposition 2.16

Let $Z = \frac{\lambda b(w-k-c)}{c}$. Then

$$\tilde{x}(w,\theta) = b + \frac{1}{\lambda}\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)\right)$$

For $\theta \in (0,1)$ it is readily seen that both $\frac{1-\theta}{\theta}$ and $e^{\frac{w-k}{\theta c}-1+Z}$ are strictly decreasing in θ, thus so is $\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}$. Combined with the fact that $W(x)$ is increasing in $x \geq 0$ this shows that $W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)$ is decreasing in θ.

Considering $\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)$, this thus consists of a term $f(\theta) = \frac{\theta}{1-\theta}$ that is strictly increasing in θ and the term $g(\theta) = W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)$ that is strictly decreasing in θ, thus $f'(\theta)g'(\theta) < 0$. Therefore, the entire term is increasing in θ if and only if

$$(f(\theta)g(\theta))' = f'(\theta)g(\theta) + f(\theta)g'(\theta) > 0 \Leftrightarrow \frac{g(\theta)}{g'(\theta)} < -\frac{f(\theta)}{f'(\theta)}$$

Substituting

$$\begin{aligned} f'(\theta) &= \frac{-1}{(1-\theta)^2} \\ g'(\theta) &= -\frac{W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)}{W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)+1}\left(\frac{w-k}{\theta^2 c} + \frac{1}{\theta(1-\theta)}\right) \end{aligned}$$

yields the equivalent inequality

$$\frac{W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)+1}{\frac{w-k}{\theta^2 c} + \frac{1}{\theta(1-\theta)}} > \theta(1-\theta)$$

which can be rewritten as

$$\frac{w-k}{c} + Z > \log\left(\frac{w-k}{c}\right) + 1$$

Since $Z \geq 0$, this holds for all $w > c + k$. Therefore, $\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)$ is increasing in θ and so is $\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+Z}\right)\right)$. Thus we can conclude that $\tilde{x}(w,\theta)$ is increasing in $\theta \in (0,1)$. □

Proof of Proposition 2.7

Under the single-epoch wholesale price contract, the supplier's capacity decision equals

$$\tilde{x}(w) = b + \frac{1}{\lambda}\log\left(\frac{w-k}{c}\right)$$

Under the long-term contract with performance contingency, the supplier's capacity decision equals

$$\tilde{x}(w,\theta) = b + \frac{1}{\lambda}\log\left(\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}}\right)\right).$$

It follows that $\tilde{x}(w,\theta) > \tilde{x}(w)$ if and only if

$$\frac{\theta}{1-\theta}W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}}\right) > \frac{w-k}{c}$$

This relationship holds if and only if

$$\begin{aligned}
& W\left(\frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}}\right) > \frac{w-k}{c}\frac{1-\theta}{\theta} \\
\Leftrightarrow \quad & \frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}} > W^{-1}\left(\frac{w-k}{c}\frac{1-\theta}{\theta}\right) \\
\Leftrightarrow \quad & \frac{1-\theta}{\theta}e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}} > \frac{w-k}{c}\frac{1-\theta}{\theta}e^{\frac{w-k}{c}\frac{1-\theta}{\theta}} \\
\Leftrightarrow \quad & e^{\frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c}} > \frac{w-k}{c}e^{\frac{w-k}{c}\frac{1-\theta}{\theta}} \\
\Leftrightarrow \quad & \frac{w-k}{\theta c}-1+\frac{\lambda b(w-k-c)}{c} > \log\left(\frac{w-k}{c}\right)+\frac{w-k}{c}\frac{1-\theta}{\theta} \\
\Leftrightarrow \quad & \frac{\lambda b(w-k-c)}{c}-1 > \log\left(\frac{w-k}{c}\right)-\frac{w-k}{c} \\
\Leftrightarrow \quad & \frac{\lambda b(w-k-c)}{c}+\frac{w-k}{c}-1 > \log\left(\frac{w-k}{c}\right)
\end{aligned}$$

Here, in the first equivalence, we use that $W(\cdot)$ is strictly increasing, for the second equivalence, we used the definition of $W(\cdot)$, while the third equivalence is obtained by taking the logarithm on both sides. The remaining equivalences are simple manipulations. This holds for all $w-k>c$ and $b \geq 0$. □

Proof of Proposition 2.8

The supply chain is coordinated when the OEM sets w such that $\tilde{x}(w,\theta) = x^*$. This is the case when

$$\frac{\theta}{1-\theta} W\left(\frac{1-\theta}{\theta} e^{\frac{w-k}{\theta c} - 1 + \frac{\lambda b(w-k-c)}{c}}\right) = \frac{r-k}{c}$$

This equality holds if and only if

$$\begin{aligned}
&& \frac{1-\theta}{\theta} e^{\frac{w-k}{\theta c} - 1 + \frac{\lambda b(w-k-c)}{c}} &= W^{-1}\left(\frac{1-\theta}{\theta}\frac{r-k}{c}\right)\\
\Leftrightarrow && \frac{1-\theta}{\theta} e^{\frac{w-k}{\theta c} - 1 + \frac{\lambda b(w-k-c)}{c}} &= \frac{1-\theta}{\theta}\frac{r-k}{c} e^{\frac{1-\theta}{\theta}\frac{r-k}{c}}\\
\Leftrightarrow && e^{\frac{w-k}{\theta c} - 1 + \frac{\lambda b(w-k-c)}{c}} &= \frac{r-k}{c} e^{\frac{r-k}{\theta c} - \frac{r-k}{c}}\\
\Leftrightarrow && \frac{w-k}{\theta c} - 1 + \frac{\lambda b(w-k-c)}{c} &= \log\left(\frac{r-k}{c}\right) + \frac{r-k}{\theta c} - \frac{r-k}{c}\\
\Leftrightarrow && w - k - \theta c + \theta\lambda b(w-k-c) &= \theta c \log\left(\frac{r-k}{c}\right) + r - k - \theta(r-k)\\
\Leftrightarrow && (w-k)(1+\theta b\lambda) - \theta c(1+b\lambda) &= \theta c \log\left(\frac{r-k}{c}\right) + (1-\theta)(r-k)\\
\Leftrightarrow && w &= k + \frac{\theta c\left(1 + b\lambda + \log\left(\frac{r-k}{c}\right)\right) + (1-\theta)(r-k)}{1+\theta b\lambda}
\end{aligned}$$

Thus $w^\theta = k + \frac{\theta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(1-\theta)(r-k)}{1+\theta b\lambda}$. The resulting profits for the supplier and the OEM are:

$$\begin{aligned}
\tilde{\pi}_s =& (w^\theta - c - k)b + \frac{w^\theta - k}{\lambda}\left(1 - \frac{c}{r-k}\right) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)\\
=& \frac{(1-\theta)(r-k-c) + \theta c \log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda} b\\
&+ \frac{1}{\lambda}\frac{\theta c\left(1 + b\lambda + \log\left(\frac{r-k}{c}\right)\right) + (1-\theta)(r-k)}{1+\theta b\lambda}\left(1 - \frac{c}{r-k}\right) - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)
\end{aligned}$$

and

$$\tilde{\pi}_m = (r - w^\theta)b + \frac{r - w^\theta}{\lambda}\left(1 - \frac{c}{r-k}\right)$$

$$= \frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda} b$$
$$+ \frac{1}{\lambda}\frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda}\left(1-\frac{c}{r-k}\right)$$

Since $w-k>c$, the supplier's participation constraint is satisfied and $\tilde{\pi}_s > 0$.

We still need to prove $w^\theta < r$, which is equivalent to

$$\frac{\theta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(1-\theta)(r-k)}{1+\theta b\lambda} < r-k$$

This equality holds if and only if

$$\begin{aligned}
& \theta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(1-\theta)(r-k) < (1+\theta b\lambda)(r-k) \\
\Leftrightarrow \quad & \theta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)-\theta(r-k) < \theta b\lambda(r-k) \\
\Leftrightarrow \quad & 1+b\lambda+\log\left(\frac{r-k}{c}\right)-\frac{r-k}{c} < b\lambda\frac{r-k}{c} \\
\Leftrightarrow \quad & 1+b\lambda+\log\left(\frac{r-k}{c}\right) < (b\lambda+1)\frac{r-k}{c}
\end{aligned}$$

For $r-k>c$, since $b\lambda \geq 0$ the left-hand side increases logarithmically in $\frac{r-k}{c}$, while the right-hand side increases linearly in $\frac{r-k}{c}$. Therefore, we conclude that the inequality holds and the coordinating wholesale price when considering multiple product generations is lower than r. □

Proof of Proposition 2.9

If the fraction of the NPV of the total supply chain profit captured by the OEM is smaller than 1, this means that $\tilde{\pi}_m^\theta < \Pi^\theta$, or equivalently $\Pi^\theta - \tilde{\pi}_m^\theta > 0$. Inserting $\tilde{\pi}_m^\theta = \frac{1}{1-\theta}\left(r-w^\theta\right)\left(b+\frac{1}{\lambda}(1-\frac{c}{r-k})\right)$ with $w^\theta = k + \frac{\theta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(1-\theta)(r-k)}{1+\theta b\lambda}$ and $\Pi^\theta = \frac{1}{1-\theta}\left((r-k-c)\left(b+\frac{1}{\lambda}\right)-\frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)\right)$ and simplifying the resulting expression yields the condition:

$$\frac{(c\theta+(1-\theta)(r-k))\left(\left(b+\frac{1}{\lambda}\right)(r-c-k)-c\,log\left(\frac{r-k}{c}\right)\right)}{(1+\theta b\lambda)(r-k)} > 0.$$

Since $r > k$ and $\theta < 1$, this holds when $\left(b+\frac{1}{\lambda}\right)(r-c-k) > c\,log\left(\frac{r-k}{c}\right)$, or equivalently $\frac{r-k}{c} > 1 + \frac{1}{b\lambda+1} log\left(\frac{r-k}{c}\right)$. This holds for all $r-k-c > 0$, by the same reasoning as in the proof of Proposition 2.8. □

Proof of Proposition 2.10

The OEM's profit is denoted as in Proposition 2.8 by

$$\begin{aligned}\tilde{\pi}_m =& \frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda} b \\ &+ \frac{1}{\lambda}\frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda}\left(1-\frac{c}{r-k}\right)\end{aligned}$$

This means that the fraction of supply chain profit captured by the OEM equals

$$\frac{\tilde{\pi}_m}{\Pi^*} = \frac{\frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda} b + \frac{1}{\lambda}\frac{\theta b\lambda(r-k-c)+\theta(r-k-c)-\theta c\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda}\left(1-\frac{c}{r-k}\right)}{(r-k-c)\left(b+\frac{1}{\lambda}\right)-\frac{c}{\lambda}\log\left(\frac{r-k}{c}\right)}$$

or equivalently, by dividing both the numerator and denominator by $r-k$

$$= \frac{\frac{\theta b\lambda\frac{r-k-c}{r-k}+\theta\frac{r-k-c}{r-k}-\theta\frac{c}{r-k}\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda} b + \frac{1}{\lambda}\frac{\theta b\lambda\frac{r-k-c}{r-k}+\theta\frac{r-k-c}{r-k}-\theta\frac{c}{r-k}\,\log\left(\frac{r-k}{c}\right)}{1+\theta b\lambda}\frac{r-k-c}{r-k}}{\frac{r-k-c}{r-k}\left(b+\frac{1}{\lambda}\right)-\frac{c}{r-k}\frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)}$$

If $\frac{r-k}{c} \to \infty$:

$$\frac{c}{r-k} \to 0, \frac{r-k-c}{r-k} \to 1, \text{ and } \frac{c}{r-k}\log\left(\frac{r-k}{c}\right) \to 0$$

Therefore,

$$\frac{\tilde{\pi}_m}{\Pi^*} \to \frac{\frac{\theta b\lambda+\theta}{1+\theta b\lambda}b + \frac{1}{\lambda}\frac{\theta b\lambda+\theta}{1+\theta b\lambda}}{b+\frac{1}{\lambda}} = \frac{\theta b\lambda+\theta}{\theta b\lambda+1}$$

□

Proof of Proposition 2.11

For $\theta \to 1$ we obtain $w^\theta \approx k + \frac{c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)}{1+b\lambda}$. As $\frac{r-k}{c} \to 1^+$, $log\left(\frac{r-k}{c}\right) \to 0^+$. Therefore, $w^\theta \approx k + \frac{c(1+b\lambda)}{1+b\lambda} = k + c$. Since $\frac{r-k}{c} \to 1^+$ means that $r - (k+c) \to 0^+$ this means that $w^\theta \approx r$. Furthermore, in Proposition 2.4 it is defined that $\hat{\rho} = \frac{1}{\lambda}\left(r - k - c - \frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right)\right)$. Again using $\frac{r-k}{c} \to 1^+$ and thus $log\left(\frac{r-k}{c}\right) \approx 0$ we obtain $\hat{\rho} \approx \frac{1}{\lambda}(0 - \frac{c}{b\lambda+1} \cdot 0) = 0$. □

Proof of Proposition 2.12

From $Y \sim Geom(1 - R(x))$ it follows that $\mathbb{E}[Y(x)] = \frac{1}{1-R(x)}$. Since $R(x) = 1 - e^{-\lambda(x-b)}$, we have $\mathbb{E}[Y(x)] = \frac{1}{e^{-\lambda(x-b)}} = e^{\lambda(x-b)}$. Under the coordinating contract with capacity $\tilde{x}(w^\theta) = x^* = b + \frac{1}{\lambda}\log\left(\frac{r-k}{c}\right)$, it follows that $\mathbb{E}[Y(x)] = \frac{r-k}{c}$. □

Proof of Lemma 2.2

We want to determine the minimum wholesale price for which $\tilde{\pi}_s(w) = (w - c - k)\left(b + \frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{w-k}{c}\right) \geq Z$. This holds if and only if

$$\begin{aligned}
& \left(\frac{w-k}{c} - 1\right)(b\lambda + 1) - Z\frac{\lambda}{c} \geq \log\left(\frac{w-k}{c}\right) \\
\Leftrightarrow \quad & e^{\frac{w-k}{c}(b\lambda+1)-(b\lambda+1+Z\frac{\lambda}{c})} \geq \frac{w-k}{c} \\
\Leftrightarrow \quad & e^{-(b\lambda+1+Z\frac{\lambda}{c})} \geq \frac{w-k}{c}e^{-(b\lambda+1)\frac{w-k}{c}} \\
\Leftrightarrow \quad & -(b\lambda+1)e^{-(b\lambda+1+Z\frac{\lambda}{c})} \geq -(b\lambda+1)\frac{w-k}{c}e^{-(b\lambda+1)\frac{w-k}{c}} \\
\Leftrightarrow \quad & W\left(-(b\lambda+1)e^{-(b\lambda+1+Z\frac{\lambda}{c})}\right) \geq -(b\lambda+1)\frac{w-k}{c} \\
\Leftrightarrow \quad & \frac{w-k}{c} \geq -\frac{1}{b\lambda+1}W\left(-(b\lambda+1)e^{-(b\lambda+1+Z\frac{\lambda}{c})}\right) \\
\Leftrightarrow \quad & w \geq k - \frac{c}{b\lambda+1}W\left(-(b\lambda+1)e^{-(b\lambda+1+Z\frac{\lambda}{c})}\right)
\end{aligned}$$

So we conclude that $\tilde{\pi}_s(w) = (w - c - k)\left(b + \frac{1}{\lambda}\right) - \frac{c}{\lambda}\log\left(\frac{w-k}{c}\right) \geq Z$ if and only if $w \geq k - \frac{c}{b\lambda+1}W\left(-(b\lambda+1)e^{-(b\lambda+1+Z\frac{\lambda}{c}}\right)$. □

Proof of Proposition 2.13

Similar to the proof of Proposition 2.4, we have as necessary condition to achieve coordination that $\rho\lambda + w = r$. To arrive at a coordinating contract in which the supplier's profit is equal to the reservation profit Z and the OEM captures the remainder of the profit, we must in addition set ρ and w such that the profit of participating for the supplier equals Z.

The supplier's profit equals:

$$\hat{\pi}_s(\hat{x}(w,\rho),w,\rho) = (w-c-k)b + \frac{w-k-c}{\lambda} - \frac{c}{\lambda}\log\left(\frac{w-k+\rho\lambda}{c}\right)$$

we impose $\rho\lambda + w = r$ by substituting $w = r - \rho\lambda$, which yields

$$\hat{\pi}_s(\rho) = (r-\rho\lambda-c-k)b + \frac{r-\rho\lambda-k-c}{\lambda} - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right).$$

Solving $\hat{\pi}_s(\rho) = Z$ for ρ gives:

$$\hat{\rho} = \frac{1}{\lambda}\left(r-k-c-\frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right) - Z\frac{\lambda}{b\lambda+1}\right).$$

Note that $\hat{\rho} > 0$ whenever $r-k-c > \frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right) + Z\frac{\lambda}{b\lambda+1}$, or equivalently $Z < (r-k-c)\frac{b\lambda+1}{\lambda} - \frac{c}{\lambda}\log\left(\frac{r-k}{c}\right) = \Pi^*$. To satisfy $r = w + \rho\lambda$, the OEM sets

$$\hat{w} = k + c + \frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right) + Z\frac{\lambda}{b\lambda+1}.$$

By construction $\hat{\pi}_s(\hat{x}(\hat{w},\hat{\rho}),\hat{w},\hat{\rho}) = Z$ and it follows that

$$\begin{aligned}\hat{\pi}_m(\hat{w},\hat{\rho}) &= \left(r-k-c-\frac{c}{b\lambda+1}\log\left(\frac{r-k}{c}\right)\right)\left(b+\frac{1}{\lambda}\right) - Z \\ &= \Pi^* - Z.\end{aligned}$$

□

Proof of Lemma 2.3

Since,

$$\mathbb{E}\left[\min\{D,x\}\right] = \mathbb{E}\left[D|D<x\right]\mathbb{P}(D<x) + x\mathbb{P}(D>x)$$

we need to determine $\mathbb{E}\left[D|D<x\right]$.

$$\mathbb{E}\left[D|D<x\right] = \frac{\int_0^x y\frac{\lambda^n y^{n-1}}{(n-1)!}e^{-\lambda y}dy}{\mathbb{P}(D<x)}$$

The lower incomplete gamma function is defined as:

$$\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$$

such that

$$\gamma(n+1,\lambda x) = \int_0^x t^n e^{-t}dt$$

and

$$\frac{1}{\lambda}\gamma(n+1,\lambda x) = \int_0^x (\lambda y)^n e^{-\lambda y}dy.$$

Therefore,

$$\int_0^x y\frac{\lambda^n y^{n-1}}{(n-1)!}e^{-\lambda y}dy = \frac{1}{\lambda(n-1)!}\gamma(n+1,\lambda x)$$

and consequently,

$$\mathbb{E}\left[\min\{D,x\}\right] = \frac{\gamma(k+1,\lambda x)}{\lambda(n-1)!} + x\mathbb{P}(D>x).$$

□

Proof of Proposition 2.14

For random demand D with pdf $f(d)$ and CDF $F(d)$ we have:

$$\pi_s(x) = -cx + E[(w-k)\min\{D,x\}]$$

$$\pi_s^\theta(x) = \frac{1}{1-\theta R(x)}\pi_s(x)$$

If $d > x$ then the profit equals $(w-k-c)x$ and if $d \leq x$ then the profit equals $(w-k)d - cx$. Therefore, we can write $\pi_s(x)$ as:

$$\begin{aligned}
\pi_s(x) &= \int_o^x ((w-k)d - cx)\, f(d)dd + \int_x^\infty (w-k-c)xf(d)dd \\
&= \int_0^x (w-k-c)xf(d)dd - \int_0^x (w-k)(x-d)f(d)dd + \int_x^\infty (w-k-c)xf(d)dd \\
&= (w-k-c)x - (w-k)\int_0^x (x-d)f(d)dd
\end{aligned}$$

leading to the following derivatives:

$$\begin{aligned}
\frac{d}{dx}\pi_s(x) =&(w-k-c) - (w-k)F(x) \\
\frac{d}{dx}\pi_s^\theta(x) =&\frac{d}{dx}\left[\frac{1}{1-\theta R(x)}\right]\cdot \pi_s(x) + \frac{1}{1-\theta R(x)}\cdot\frac{d}{dx}\pi_s(x) \\
=&\frac{\theta R'(x)}{(1-\theta R(x))^2}\left((w-k-c)x - (w-k)\int_0^x (x-d)f(d)dd\right) \\
&+ \frac{1}{1-\theta R(x)}\left((w-k-c) - (w-k)F(x)\right).
\end{aligned}$$

When considering a single generation, the optimal capacity $\tilde{x}$ satisfies $\frac{d}{dx}\pi(x) = 0$, so $\tilde{x}(w) = F^{-1}\left(\frac{w-k-c}{w-k}\right)$. For $\tilde{x}(w,\theta) > \tilde{x}(w)$, with $\tilde{x}(w,\theta)$ the capacity under the wholesale price contract with endogenous renewal probability, we need to show that under the renewal contract higher expected profit can be attained by increasing capacity beyond $\tilde{x}(w)$, i.e., $\frac{d}{dx}\pi_s^\theta(x)|_{x=\tilde{x}(w)} > 0$. In other words, we need to show that

$$\begin{aligned}
\frac{d}{dx}\pi_s^\theta(x)|_{x=\tilde{x}(w)} = &\frac{\theta R'\left(F^{-1}\left(\frac{w-k-c}{w-k}\right)\right)}{\left(1-\theta R\left(F^{-1}\left(\frac{w-k-c}{w-k}\right)\right)\right)^2}\left((w-k-c)F^{-1}\left(\frac{w-k-c}{w-k}\right)\right. \\
&\left. -(w-k)\int_0^{F^{-1}\left(\frac{w-k-c}{w-k}\right)}\left(F^{-1}\left(\frac{w-k-c}{w-k}\right) - d\right) f(d)dd\right) > 0
\end{aligned}$$

Since $R(x)$ is increasing in x, it follows that $\frac{\theta R'\left(F^{-1}\left(\frac{w-k-c}{w-k}\right)\right)}{\left(1-\theta R\left(F^{-1}\left(\frac{w-k-c}{w-k}\right)\right)\right)^2} > 0$. So

$\frac{d}{dx}\pi_s^\theta(x)|_{x=\tilde{x}(w)} > 0$ if and only if

$$(w-k-c)F^{-1}\left(\frac{w-k-c}{w-k}\right) > \\ (w-k)\int_0^{F^{-1}\left(\frac{w-k-c}{w-k}\right)}\left(F^{-1}\left(\frac{w-k-c}{w-k}\right)-d\right)f(d)dd$$

or if we let $y = \frac{w-k-c}{w-k}$:

$$\begin{aligned}
& yF^{-1}(y) > \int_0^{F^{-1}(y)}\left(F^{-1}(y)-d\right)f(d)dd \\
\Leftrightarrow \quad & yF^{-1}(y) > F^{-1}(y)\int_0^{F^{-1}(y)} f(d)dd - \int_0^{F^{-1}(y)} df(d)dd \\
\Leftrightarrow \quad & yF^{-1}(y) > F^{-1}(y)F\left(F^{-1}(y)\right) - \int_0^{F^{-1}(y)} df(d)dd \\
\Leftrightarrow \quad & yF^{-1}(y) > yF^{-1}(y) - \int_0^{F^{-1}(y)} df(d)dd \\
\Leftrightarrow \quad & 0 > -\int_0^{F^{-1}(y)} df(d)dd
\end{aligned}$$

When $w > c+k$, this holds for any probability distribution that has support on the non-negative real numbers. □

Proof of Proposition 2.15

Since $w^\theta < \tilde{w} = r$, it follows that the manufacturer's expected profit under the renewal contract with coordinating wholesale price is positive. When the supplier decides not to invest in any capacity ($x = 0$), the supplier's expected profit is 0. Under the contingent renewal contract with endogenous renewal probability, we find that $\frac{d}{dx}\pi_s^\theta(x)|_{x=0} = w-k-c-(w-k)F(0) = w-k-c$ as $R(0) = 0$ and $F(0) = 0$. Consequently, the supplier can increase profit by choosing capacity $x > 0$, which then yields $\pi_s^\theta(x) = \frac{1}{1-\theta R(x)}\pi_s(x) > 0$. As $\frac{1}{1-\theta R(x)} > 0$, this means that the supplier's expected profit per generation, denoted by $\pi_s(x)$ is positive. Since there thus exists a positive capacity for which the supplier's expected profit per period is positive, this must also hold for the optimal capacity investment $\tilde{x}(w^\theta,\theta)$. □

3 Contingent Renewal Contracts in High-tech Manufacturing with Oligopolistic Suppliers

High-tech manufacturers (OEMs) often produce multiple generations of high-tech end-products. For each generation an OEM has to source complex components from a few oligopolistic suppliers. Due to the high shortage costs for missing components, resulting from costly delays in production of the end-product, it is important to align incentives between OEMs and suppliers of these components that are often single-sourced. We formulate an infinite horizon perfect information game with two possible suppliers where the payoffs in the current generation and transition probabilities for the next generation depend on the capacity investment of the current supplier. We express the suppliers' optimal capacity investment as a function of the wholesale price paid by the OEM and the capacity investment decision of the alternative supplier. We show that for every wholesale price there exists an equilibrium where neither supplier has incentive to adjust their capacity decision. Additionally, we show that the wholesale price for which in equilibrium the supply chain optimal capacity decision is made is lower than the coordinating

This chapter is based on Meijer et al. (2021b).

wholesale price when only a single supplier is present, but higher than in case there is an unlimited number of suppliers.

3.1. Introduction

High-tech supply chains involve original equipment manufacturers (OEMs), who design and produce high-tech end-products, and their suppliers. Production of critical components of these high-tech end-products requires specific technical knowledge and equipment. Due to this complex engineering, there often is a limited number of suppliers that is capable of supplying these high-tech components. This means that the OEM has to deal with oligopolistic suppliers. Additionally, critical components of high-tech systems are often sourced at a single supplier. Since suppliers need to invest in very specific production capacity and technical staff, they are often unwilling to engage unless they are the sole supplier of the component. Consequently, the OEM works closely with a single supplier to design and engineer such component. This single sourcing poses a risk for the OEM that the supplier invests insufficiently in production capacity. In this chapter, we focus on how to align incentives between an OEM and these oligopolistic suppliers of a critical component of the high-tech system, e.g., Yuasa who supplies the batteries for Boeing's 787 aircraft or VDL/ETG who makes the wafer handler for ASML's lithography machines.

In modeling this relationship between an OEM and a supplier, we need to take into account several other characteristics of high-tech supply chains. Like any other new product, demand is highly uncertain. To reduce this uncertainty, high-tech OEMs provide the possibility to place advance orders already before production has commenced. For example, for the Boeing 787 around 900 advance orders were placed, while many other orders only arrived after the aircraft was in production (Nolan, 2009). Therefore, demand consists of two parts: a fixed part representing *advance orders* and an uncertain additional demand. Furhtermore, *supply capacity is not verifiable* for the OEM: The OEM can audit the available equipment, but cannot verify how much production capacity is generated by this. Finally, an OEM continues to improve the end-product and introduces *multiple generations*. In this way, the OEM does not have to create a completely new product every time and avoids high development costs. The Boeing 747, which was launched in 1970 and had multiple upgrades before retiring in 2018, is an example of this.

We investigate how in a setting with oligopolistic suppliers this last characteristic of multiple product generations can be utilized to entice a supplier to perform well by offering the possibility of long-term collaboration. We analyze three situations with regard to the supply base: (1) there is only a single supplier available, (2) there are infinitely many suppliers available and (3) there are two oligopolistic suppliers. In case (1), the problem essentially reduces to a repetition of single-period problems, as there is no option to switch to another supplier and hence no credible threat. Case (2) refers to the contingent renewal contracts studied in Chapter 2. Case (3) is the main focus of this chapter. We model capacity investment of the suppliers as an infinite horizon game, where only a single supplier (the one currently working with the OEM) can invest in capacity at the same time. The capacity investment decision of this supplier determines the pay-offs in the current generation and determines the probability that the OEM continues working with this supplier or switches to a different supplier for the next product generation.

We show that the possibility of renewing the contract contingent upon performance works as a motivation for the supplier to perform well and invest in adequate production capacity. When we analyze the wholesale price required to entice the supplier to build the coordinating capacity, we find that this wholesale price in case (3) is lower than in case (1), meaning that the OEM earns a positive expected profit under this contract. However, when we compare this wholesale price to case (2), we find that the OEM needs to pay a higher wholesale price when the supplier pool is limited than in case the pool is unlimited, indicating that the oligopolistic supplier has more power. Furthermore, we find that the capacity investment of the incumbent supplier depends on the potential capacity investment of the alternative supplier. When the alternative supplier would invest in higher capacity, after switching to the alternative supplier the OEM is more likely to stay with that supplier. Hence, when the alternative supplier is willing to invest more capacity, so is the incumbent supplier, to avoid that the OEM switches. We show that there exists an equilibrium of investment decisions such that neither supplier has incentive to deviate.

This chapter is organized as follows. We review relevant literature in Section 3.2. In Section 3.3, we study the centralized supply chain, which will be used as a benchmark. Section 3.4 provides the analysis for the decentralized case, where we differentiate between the number of available suppliers (as described by cases (1), (2) and (3) above) and compare the wholesale prices required to entice the supplier to invest in coordinating capacity. The chapter concludes in Section 3.5. All proofs are provided in Appendix 3.A.

3.2. Literature review

In high-tech manufacturing, supplier selection and contracting are important decisions. An overview of possible sourcing strategies is given by Elmaghraby (2000), where the options of single- and multiple-sourcing are considered, as well as the possibility of having a single or multiple periods in which supplier selection can take place. Capacity reservations contracts are widely studied in literature (see e.g. Barnes-Schuster et al., 2002; Brown and Lee, 2003; Erkoc and Wu, 2005; Ren et al., 2010; Roels and Tang, 2017), but are often not practical in high-tech settings due to difficulties in verifying capacity investments. Furthermore, these studies often focus on contracts between a manufacturer and a single supplier and do not consider multiple suppliers or supplier switching.

Competition among suppliers is widely studied in literature. The vast majority of literature on supplier competition considers standard duopoly games with Cournot quantity competition or Bertrand price competition (cf. Sinha and Sarmah, 2010; Tsitsiklis and Xu, 2014; Wu et al., 2019). Wu et al. (2019) study a combination of these two types of competition, as the suppliers engage in price competition and the multiple retailers in quantity competition. Li and Gupta (2011) consider suppliers that may strategically invest in capabilities upfront to serve anticipated demand and to get an advantage over competitors. Hu et al. (2017) study competition among suppliers in an experimental laboratory study. They are interested in how decision makers actually behave in a capacity investment game and whether this behaviour is consistent with theoretical predictions. They conclude that suppliers invested in higher capacity than was predicted based on theory. The empirical study by Wu and Choi (2005) shows that the relationship between suppliers impacts the buyers strategy. They identify five archetypes of supplier-buyer relations through case studies. In many of these competitions it is assumed that the competing firms have market power. In our case, however, the OEM has most power, not the competing suppliers.

A key characteristic of our problem is the small supply base, where the number of suppliers having the required capabilities is limited. This is for example the case when there are substantial barriers to enter the market. Wilhite et al. (2014) study how to create incentives for suppliers to enter and stay in the market, to foster supplier competition. Li and Wan (2017) study how to stimulate supplier performance by fostering supplier competition and inducing supplier effort and what the consequences of this are for the supply base. A trade-off between a large number and small number of suppliers is studied by Li (2013): a small supply base may motivate supplier effort, whereas a large supply base encourages competition between suppliers. It is concluded that sole sourcing is preferred when supplier's cost uncertainty is low or when demand uncertainty is high. Nam et al. (2011) also study optimal sizing of the supply base. They argue that a limited number of suppliers may lead to opportunistic supplier behaviour. However, they also explain that the past decades many contractors have adopted long-lasting partnerships with fewer suppliers to increase cost-effectiveness and coordination. Our approach differs from these studies as we take the small supply base as given and analyse the effect thereof on supplier investments.

The main feature of the contracts considered in this chapter is the collaboration between an OEM and its suppliers that is continued over multiple periods or terminated based on supplier performance. A similar setting was considered by Meijer et al. (2021d) (on which Chapter 2 of this thesis is based) for an unlimited supply base, such that the OEM could always switch to a new supplier. Previous research also considered continued collaboration through relational contracts where an OEM and supplier make agreements that are not court-enforceable but are enforced by trigger strategies (Taylor and Plambeck, 2007b,a; Sun and Debo, 2014). Merckx and Chaturvedi (2020) study the trade-off between leveraging supplier competition in each period by offering short-term contracts and incentivizing investment by the incumbent supplier by offering a single long-term contract. Xia (2011) studies competition between two suppliers of substitutable products with multiple buyers choosing or switching between suppliers based on costs or their own preferences. Li and Debo (2009) consider a manufacturer that chooses between sole- and second-sourcing, where second-sourcing refers to keeping the option open to source from new suppliers in the future. They show that second-sourcing may lead to larger initial capacity investments. Pfeiffer (2010) and Wagner and Friedl (2007) both study conditions for a manufacturer to switch to a different supplier.

Pfeiffer (2010) considers a setting with asymmetric information, where complex contracts are employed to provide proper incentives for the incumbent supplier to share cost information truthfully. Wagner and Friedl (2007) also considers the option of partial switching, resulting in multi-sourcing. The main motivation for switching in the work of Pfeiffer (2010) and Wagner and Friedl (2007) is to source components at a lower cost. This is different from the setting we consider, where the threat of switching is used as an incentive to invest in higher capacity.

3.3. Centralized supply chain

We begin by analyzing a centralized supply chain, where the capacity decision x is made by a central body before demand is realized. Demand consists of two parts: a fixed part denoted by b representing advance orders and an uncertain part denoted by A. We assume that the additional uncertain demand is exponentially distributed, such that $D = b + A$ with $A \sim Exp(\lambda)$. Capacity x is established at a cost $c \geq 0$ per unit. Once demand is realized, up to x units can be produced at an additional cost k per unit. The final product is sold to the customer at retail price r. We consider a setting where multiple generations of a product are produced. To assure tractability, we assume that the cost structure is generation independent, such that in any generation t we have $c_t = c$, $k_t = k$ and $r_t = r$. Furthermore, we assume that demand is stationary, i.e. $D_t = b_t + A_t$ with $b_t = b$ and $A_t \sim Exp(\lambda)$. Future profits are discounted at rate $0 < \delta < 1$. Assuming that any demand exceeding the available capacity is lost, this allows us to write the following net present value (NPV) of the expected supply chain profit:

$$\Pi^{\delta}(x) = \mathbb{E}\left[\sum_{t=1}^{\infty} \delta^{t-1}\Pi(x)\right] = \frac{1}{1-\delta}\Pi(x), \tag{3.1}$$

with $\Pi(x)$ the expected profit function for each generation:

$$\Pi(x) = -cx + \mathbb{E}[(r-k)\min\{D, x\}] = -cx + \mathbb{E}[(r-k)\min\{b + A, x\}]. \tag{3.2}$$

By using the first-order conditions, we can find the optimal capacity and corresponding supply chain profits, as derived by Meijer et al. (2021d), to be:

$$x^* := b + \frac{1}{\lambda} \log \left(\frac{r-k}{c} \right) \tag{3.3}$$

and $\Pi^\delta = \frac{1}{1-\delta}\Pi^*$ with

$$\Pi^* := \Pi(x^*) = (r-k-c)\left(b + \frac{1}{\lambda}\right) - \frac{c}{\lambda} \log \left(\frac{r-k}{c} \right) > 0. \tag{3.4}$$

We can observe that the optimal capacity x^* covers the base demand b and has an additional part to satisfy the uncertain demand. Now that we have established the optimal capacity and corresponding profit as a benchmark, in the remainder of this chapter we will analyze the case where the supply chain is decentralized.

3.4. Decentralized supply chain

In a decentralized supply chain there no longer is a central decision maker. Instead, we study the interaction between a high-tech OEM and a supplier of a component for its end-product. The OEM determines a wholesale price w that it will pay to the supplier for each component and subsequently the supplier decides how much capacity x to build. Similar to the centralized supply chain, the supplier builds capacity at a cost c per unit and can produce up to x at a cost k per unit. The OEM pays the supplier wholesale price w and sells the end-product to its customers at retail price r.

Since the introduction of a new generation requires new technologies, production capacity is generation-specific meaning that capacity established for components of the current generation cannot be extended to the next generation. Consequently, every time a new generation of the product is introduced the OEM can make the decision to either stay with the incumbent supplier or switch to a different supplier when the performance of the current supplier is unsatisfactory. However, this requires the supply base to be sufficiently large for the OEM to choose another supplier. We will study the interaction between the OEM and its incumbent supplier in three cases: (1) the incumbent supplier is the only supplier that has the skills to produce the required components; (2) there is an unlimited base of suppliers that can produce the components; and (3) there are two suppliers having the required capabilities. The decisions corresponding to these cases are denoted by superscripts 1, ∞ and 2, respectively.

3.4.1 Single supplier

When there is only a single supplier that is able to produce the component required for the end-product, the OEM will work with this supplier for every generation. There is no alternative supplier, so the OEM has no possibility to switch. This means that the NPV of the supplier's profit function can be written as:

$$\pi_s^1(x,w) = \mathbb{E}\left[\sum_{t=1}^{\infty} \delta^{t-1} \pi_s(x,w)\right], \tag{3.5}$$

with

$$\pi_s(x,w) = -cx + \mathbb{E}[(w-k)\min\{b+A,x\}] \tag{3.6}$$

the single-generation expected profit function. The supplier maximizes its profit over all generations by maximizing its single-generation profit in each generation. The problem thus reduces to a repetition of single-period problems.

By considering the first-order condition and the fact that $A \sim Exp(\lambda)$, the supplier's profit maximizing capacity is:

$$x^1(w) = b + \frac{1}{\lambda}\log\left(\frac{w-k}{c}\right). \tag{3.7}$$

This confirms the well-known result that the supply chain can only be coordinated when the OEM sets wholesale price $w^1 = r$, shifting all profit to the supplier.

3.4.2 Unlimited suppliers

The threat of switching supplier becomes credible when the OEM has many alternative suppliers to work with. We assume that the OEM will continue working with the incumbent supplier for the next generation, unless this supplier is unable to satisfy all demand D for the current generation with the available capacity x. This means that the OEM will continue working with the incumbent supplier with probability $R(x) = Prob\{D \leq x\} = 1 - e^{-\lambda(x-b)}$ for $x \geq b$, and $R(x) = 0$ otherwise. For any given supplier capacity x, the number of generations $Y(x)$ that the incumbent supplier can work with the OEM follows a geometric distribution:

$Y \sim Geom(1 - R(x))$. The expected profit for the supplier in each generation t is $\delta^{t-1}\pi_s(x, w)$, with $\pi_s(x, w)$ as defined in Equation (3.6). We can then write the NPV of the supplier's expected profit function as:

$$\pi_s^\infty(x, w) = \mathbb{E}\left[\sum_{t=1}^{Y(x)} \delta^{t-1}\pi_s(x, w\right] = \sum_{t=1}^{\infty} P(Y(x) \geq t)\delta^{t-1}\pi_s(x, w)$$
$$= \sum_{t=1}^{\infty}(R(x)\delta)^{t-1}\pi_s(x, w) = \frac{\pi_s(x, w)}{1 - \delta R(x)}. \quad (3.8)$$

This case is analyzed in detail in Chapter 2, where it is shown that the supplier's optimal capacity as a function of the wholesale price in any generation equals

$$x^\infty(w) = b + \frac{1}{\lambda}\log\left(\frac{\delta}{1-\delta}\mathrm{W}\left(\frac{1-\delta}{\delta}e^{\frac{w-k}{\delta c} - 1 + \frac{b\lambda(w-k-c)}{c}}\right)\right). \quad (3.9)$$

Furthermore, it is shown that the coordinating wholesale price, as given in Equation (3.10), is smaller than the retail price r.

$$w^\infty = k + \frac{\delta c\left(1 + b\lambda + \log\left(\frac{r-k}{c}\right)\right) + (1-\delta)(r-k)}{1 + \delta b\lambda} < r. \quad (3.10)$$

3.4.3 Two suppliers

In reality, the OEM will likely find itself in a situation in between the two cases considered above, where there is a limited number of suppliers that it could work with. Unlike in the single supplier case, the OEM can use the threat of switching as an incentive for the supplier to build sufficient capacity. However, contrary to the case with an unlimited number of suppliers, the OEM will at some point return to the incumbent supplier.

We assume that the OEM has two potential suppliers to work with. We will refer to the supplier that the OEM is currently working with as the 'incumbent' supplier and the other supplier as the 'alternative' supplier. The interaction between the two suppliers can be modeled as an infinite horizon stochastic game. Stochastic games were introduced by Shapley (1953). In a stochastic game, the choices of the players at a point in time do not only determine their immediate pay-offs, but also the stochastic transitions to the game played in the next period. In our situation, only

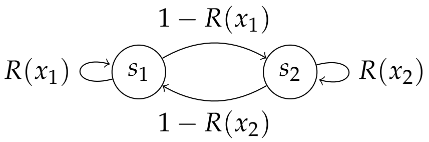

Figure 3.1: Illustration of supplier switching probabilities

a single player, namely the incumbent supplier, can make a decision at the current moment. This decision determines the expected pay-off in the current period. Next to that, it determines the probabilities that either the same game is played in the next period, i.e., the incumbent supplier continues to supply and again needs to decide on capacity, or the OEM switches to the alternative supplier which then becomes the decision maker.

The state space consists of two states: $S = \{s_1, s_2\}$ where s_1 denotes that supplier 1 is the incumbent supplier and s_2 denotes that supplier 2 is the incumbent supplier. When in state s_1, supplier 1 determines capacity x_1 and supplier 2 cannot choose anything. This results in expected pay-off $\pi_s(x_1, w)$, with $\pi_s(x, w)$ as defined in Equation (3.6), for supplier 1 and pay-off 0 for supplier 2. Similarly, when in state s_2, supplier 2 determines capacity x_2 and supplier 1 cannot choose anything. This results in expected pay-off $\pi_s(x_2, w)$ for supplier 2 and pay-off 0 for supplier 1. The transition probabilities are illustrated in Figure 3.1. The costs of supplier 1 are equal to c_1 and k_1, while the costs for supplier 2 are c_2 and k_2. The OEM will continue working with the incumbent supplier with probability $R(x) = Prob\{D \leq x\} = Prob\{A \leq (x - b)\} = 1 - e^{-\lambda(x-b)}$ for $x \geq b$, and $R(x) = 0$; otherwise. With probability $1 - R(x) = e^{-\lambda(x-b)}$ the OEM will switch to the alternative supplier.

Next, we will analyze optimal strategies in every period from the perspective of supplier 1, assuming that this is the supplier that is currently playing. Analysis for supplier 2, when supplier 2 is playing, is analogous. We will focus our attention to stationary strategies, where the same action, i.e., capacity investment, is chosen every time a specific state is encountered.

3.4.3.1 Incumbent supplier's stationary strategy

When the manufacturer leaves the incumbent supplier, every following period there is a probability $1 - R(x_2)$ that the manufacturer will return, with x_2 the decision of the alternative supplier. Therefore, the number of periods until the manufacturer returns to the incumbent supplier, denoted by Y_2, is *Geometric*$(1 - R(x_2))$. Therefore, we can write the expected profit function of the incumbent supplier, given in Equation (3.11), consisting of three parts: (i) the expected profit for the current generation; (ii) the expected profit of the next generation, when the OEM continues working with the incumbent supplier; (iii) the expected profit for the next generation the OEM and incumbent supplier will work together again, when the OEM now switches to the alternative supplier. The NPV of the incumbent supplier's profit therefore equals:

$$\begin{aligned}\pi_s^2(x_1) = -c_1 x_1 + (w - k_1)\mathbb{E}\left[\min\{x_1, D\}\right] + R(x_1)\delta\pi_s^2(x_1)\\ + (1 - R(x_1))\mathbb{E}\left[\delta^{Y_2}\right]\pi_s^2(x_1) \qquad (3.11)\end{aligned}$$

where $\mathbb{E}\left[\delta^{Y_2}\right]$ is the expected discount factor for the profit that the supplier earns when the OEM returns Y_2 periods from now, as given in Lemma 3.1.

Lemma 3.1 *Since $Y_2 \sim$ Geometric$(1 - R(x_2))$, it follows that $\mathbb{E}\left[\delta^{Y_2}\right] = \frac{\delta(1-R(x_2))}{1-\delta R(x_2)}$.*

Using the first-order conditions, we can derive the optimal capacity decision of the incumbent supplier in response to the capacity of the alternative supplier as a function of the wholesale price.

Proposition 3.1 *The optimal stationary capacity of supplier 1 (in response to the capacity of supplier 2) is given by*

$$\begin{aligned}x_1^2(w, x_2) = b + \frac{1}{\lambda}\log\left(\frac{\delta}{1-\delta}\left(1 - \frac{1 - R(x_2)}{1 - \delta R(x_2)}\right)\cdot\right.\\ \left. W\left(\left(\frac{1}{\delta R(x_2)} - 1\right)e^{\frac{w-k_1}{\delta c_1 R(x_2)} - 1 + b\lambda\frac{w-k_1-c_1}{c_1}}\right)\right), \qquad (3.12)\end{aligned}$$

where W is the Lambert-W function, which is the inverse function of $f(y) = ye^y$, and $R(x_2) = 1 - e^{-\lambda(x_2 - b)}$ is the probability that the OEM stays with the alternative supplier after working with the alternative supplier for a generation.

Analogous to the supplier's optimal capacity decision given in Equation (3.9) for the unlimited supplier case, the optimal capacity can cover the base demand b plus a certain amount of additional demand. The available capacity for covering additional demand in this case does not only depend on base demand b and discount factor δ, but also on the probability that the OEM will stay with the alternative supplier after switching, denoted by $R(x_2)$. When the probability that the OEM stays with the alternative supplier is higher, it will take longer for the OEM to return to the current supplier when he leaves. Hence, the current supplier is more willing to invest in additional capacity to prevent the OEM from switching to the alternative supplier. This is formalized in Proposition 3.2.

Proposition 3.2 *The optimal capacity of supplier 1 is monotonically increasing in the capacity investment of supplier 2.*

When $R(x_2) \to 1$ the probability of the OEM leaving the alternative supplier to return to the incumbent supplier goes to 0 ($1 - R(x_2) \to 0$). This means that after switching to the alternative supplier, the OEM will not return to the incumbent supplier. Consequently, the incumbent supplier essentially faces the same decision problem as in the unlimited supplier case. In Corollary 3.1, it is shown that the incumbent supplier's capacity decision in this case also converges to the capacity decision in the unlimited supplier case.

Corollary 3.1 *If $R(x_2) \to 1$, $x_1^2(w, x_2) \to x^\infty(w)$ for all wholesale prices $w > c_1 + k_1$.*

Next, when we compare the capacity investment by the supplier in the case with a limited number of suppliers to the capacity investment both when there is a single supplier and an unlimited number of suppliers, assuming that capacity and production costs are equal, i.e. $c_1 = c$ and $k_1 = k$, we get:

Proposition 3.3 *For all wholesale prices $w > c + k$ and investment decisions of the alternative supplier x_2 such that $0 < R(x_2) < 1$, it holds that $x^1(w) < x_1^2(w, x_2) < x^\infty(w)$.*

Similar to the case with unlimited suppliers, Proposition 3.3 suggests that the performance-based condition for switching to an alternative supplier poses an incentive to the supplier to invest in capacity. However, since the OEM will in this case, contrary to the unlimited supplier case, return the current supplier at some point, the incentive is weaker.

3.4.3.2 Equilibrium solution

Now that we have established the best response of the incumbent supplier to the capacity decision of the alternative supplier, and analogously the alternative supplier's capacity decision in response to the incumbent supplier's capacity decision, the question arises whether there exists an equilibrium solution where neither supplier has incentive to adjust their capacity decision. In Proposition 3.4 we show that for every wholesale price such that $w > c_1 + k_1$ and $w > c_2 + k_2$ such an equilibrium solution exists. Proposition 3.5 states that for pure exponential demand, when $b = 0$, this equilibrium is unique.

Proposition 3.4 *For every wholesale price w there exists at least one equilibrium $(\hat{x_1}, \hat{x_2})$ such that $x_1^2(w, \hat{x_2}) = \hat{x_1}$ and $x_2^2(w, \hat{x_1}) = \hat{x_2}$.*

Proposition 3.5 *When there is no base demand, i.e., $b = 0$, for every wholesale price w there exists a unique equilibrium $(\hat{x_1}, \hat{x_2})$.*

3.4.3.3 Coordination

In Proposition 3.4, we have just shown that for every wholesale price there exists an equilibrium solution. Next, we will investigate what wholesale price the OEM should offer to the suppliers such that there exists an equilibrium solution where both suppliers are enticed to build the coordinating capacity. To obtain a single wholesale price for which both parties are enticed to build the coordinating capacity, we will assume symmetric costs for both suppliers, such that $c = c_1 = c_2$ and $k = k_1 = k_2$. This means we want to find the wholesale price w for which we find $x_1^2(w, x^*) = x^*$ (and due to symmetry $x_2^2(w, x^*) = x^*$) as defined in Equation (3.3). This coordinating wholesale price for the limited supplier case is given in Proposition 3.6.

Proposition 3.6 *Suppose the OEM sets its wholesale price for each generation at w^2, where:*

$$w^2 = k + \frac{\delta c \left(1 - \frac{c}{r-k}\right)\left(1 + b\lambda + \log\left(\frac{r-k}{c}\right)\right) + \left(1 - \delta\left(1 - \frac{c}{r-k}\right)\right)(r-k)}{1 + \delta b \lambda \left(1 - \frac{c}{r-k}\right)}.$$

Then there exists an equilibrium solution where the suppliers' capacity decision equals the supply chain optimal capacity decision so that the supply chain is coordinated, i.e.,

$x_1^2(w, x^*) = b + \frac{1}{\lambda} \log\left(\frac{r-k}{c}\right) = x^*.$

In Chapter 2, we concluded that the wholesale price required to coordinate the supply chain in the unlimited supplier case is lower than the the coordinating wholesale price when only considering a single supplier. Proposition 3.7 shows that the coordinating wholesale price in the limited supplier case is in between these two values.

Proposition 3.7 *For the coordinating wholesale price, as given in Proposition 3.6, the following holds:* $w^1 > w^2 > w^\infty$.

The intuition behind this result is that the supplier requires more incentive, in terms of a higher wholesale price, than in the unlimited supplier case as the incumbent supplier knows the OEM will return to work with the incumbent supplier again at some point. However, there is some risk to the supplier of loosing the OEM temporarily to the alternative supplier, so that he is willing to invest in the coordinating capacity already for a lower wholesale price than in the single supplier case.

3.5. Conclusions

In this chapter, we have studied contracting between a high-tech OEM and suppliers of a critical component of the high-tech end-product. Supplier-buyer relationships are essential for the functioning of such high-tech supply chains. Since OEMs often produce multiple generations, this can be used to incentivize suppliers to perform well. We have investigated how a contingent renewal contract can entice the supplier to invest sufficient capacity to coordinate the decentralized supply chain by aligning the incentives of the OEM and the supplier when capacity is not verifiable. The main focus of this chapter is the case where there is only a limited number of suppliers that has the capacities to produce the required modules. Specifically, we have considered the case of two possible suppliers. We modeled their interaction as an infinite horizon stochastic game, where at any time only a single supplier has a non-trivial decision to make. This is the incumbent supplier that the OEM is currently working with. The capacity decision of the incumbent supplier determines not only the pay-off for the current generation, but also the probabability with which the OEM continues collaboration with the incumbent supplier.

We have formulated the optimal capacity investment decision of the incumbent supplier that anticipates on the capacity decision of the alternative supplier. When the alternative supplier is willing to invest more capacity, it is likely that the OEM will stay for a longer period with the alternative supplier after switching and therefore the incumbent supplier is also willing to invest in more capacity. We showed that for every wholesale price that covers the supplier's costs there exists a stationary equilibrium where neither supplier has incentive to adjust their capacity decision. Additionally, we have analyzed the coordinating wholesale price for which in this stationary equilibrium both suppliers will make the supply chain optimal capacity investment. This coordinating wholesale price is lower than in case there is only a single supplier with a standard wholesale price contract, meaning that both parties earn postive expected profits in every period. However, when the number of available supplier is limited, the suppliers have more power than in case there is an unlimited number of suppliers. Consequently, the coordinating wholesale price in the limited supplier case is higher than in the unlimited supplier case.

3.A. Proofs

Proof of Lemma 3.1

Since $Y_2 \sim Geometric(1 - R(x_2))$ is a discrete random variable taking on positive integer values and $\delta \in (0,1)$, $\mathbb{E}\left[\delta^{Y_2}\right]$ is the probability generating function of Y_2 evaluated at δ (Shaked and Shanthikumar, 2007). The probability generating function of a random variable X that is geometrically distributed with parameter p is given by $G_X(t) = \mathbb{E}\left[t^X\right] = \frac{pt}{1-(1-p)t}$. Therefore, it follows that $\mathbb{E}\left[\delta^{Y_2}\right] = \frac{\delta(1-R(x_2))}{1-\delta R(x_2)}$. □

Proof of Proposition 3.1

Let $y = e^{-\lambda(x_1-b)}$ and $z = e^{-\lambda(x_2-b)}$. Then we can rewrite Equation (3.11) as

$$\pi_s(y,z) = \frac{b(w-k_1-c_1) + \frac{c_1}{\lambda}\log(y) + \frac{w-k_1}{c_1}(1-y)}{1-\delta(1-y)-\delta y\frac{z}{1-\delta(1-z)}}$$

The first derivative with respect to y is given by

$$\begin{aligned}\frac{d}{dy}\pi_s(y,z) = &\frac{\left(1-\delta(1-y)-\delta y\frac{z}{1-\delta(1-z)}\right)\left(\frac{c_1}{\lambda}\frac{1}{y}-\frac{w-k_1}{\lambda}\right)}{(1-\delta(1-y)-\delta y\frac{z}{1-\delta(1-z)})^2} \\ &- \frac{\left(b(w-k_1-c_1)+\frac{c_1}{\lambda}\log(y)+\frac{w-k_1}{\lambda}(1-y)\right)\left(\delta-\delta\frac{z}{1-\delta(1-z)}\right)}{(1-\delta(1-y)-\delta y\frac{z}{1-\delta(1-z)})^2}\end{aligned}$$

such that the first-order condition yields

$$\begin{aligned}\left(\left(\delta-\delta\frac{z}{1-\delta(1-z)}\right)(b\lambda(w-k_1-c_1)-c_1)+\left(1-\delta\frac{z}{1-\delta(1-z)}\right)(w-k_1)\right)y \\ +\left(\delta-\delta\frac{z}{1-\delta(1-z)}\right)c_1 y\log(y) = (1-\delta)c_1\end{aligned}$$

which is of the form $a_1 y + a_2 y \log y = a_3$ with $a_1, a_2, a_3 > 0$ since $w \geq c_1 + k_1$, $b \geq 0$, $\lambda > 0$ and $0 < \delta < 1$.

The solution to this equation is given by

$$y = \frac{a_3}{a_2 W\left(\frac{a_3}{a_2} e^{\frac{a_1}{a_2}}\right)}$$

where W is the Lambert's W function. Substituting

$$\begin{aligned}
a_1 &= \left(\delta - \delta \frac{z}{1-\delta(1-z)}\right)\left(b\lambda(w-k_1-c_1)-c_1\right) + \left(1-\delta\frac{z}{1-\delta(1-z)}\right)(w-k_1)\\
a_2 &= \left(\delta - \delta \frac{z}{1-\delta(1-z)}\right) c_1\\
a_3 &= (1-\delta)c_1
\end{aligned}$$

and simplifying the expression yields

$$y = \frac{1-\delta}{\delta\left(1-\frac{z}{1-\delta(1-z)}\right)} \frac{1}{\mathrm{W}\left(\left(\frac{1}{\delta(1-z)}-1\right)e^{\frac{w-k_1}{\delta c_1(1-z)}-1+b\lambda\frac{w-k_1-c_1}{c_1}}\right)}.$$

Since $y = e^{-\lambda(x_1-b)}$ and $z = e^{-\lambda(x_2-b)} = 1-R(x_2)$, this gives

$$\begin{aligned}
x_1^*(w,x_2) = b + \frac{1}{\lambda}\log\Bigg(\frac{\delta}{1-\delta}\left(1-\frac{1-R(x_2)}{1-\delta R(x_2)}\right)\cdot \\
\mathrm{W}\left(\left(\frac{1}{\delta R(x_2)}-1\right)e^{\frac{w-k_1}{\delta c_1 R(x_2)}-1+b\lambda\frac{w-k_1-c_1}{c_1}}\right)\Bigg).
\end{aligned}$$

□

Proof of Proposition 3.2

$$\begin{aligned}
\frac{d}{dx_2}x_1(w,x_2) = -\frac{e^{-\lambda(x_2-b)}}{1-e^{-\lambda(x_2-b)}}\left[\frac{1}{W(\ldots)+1}\frac{1}{1-e^{-\lambda(x_2-b)}}\cdot\right.\\
\left.\left[\frac{1-e^{-\lambda(x_2-b)}}{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)}+\frac{w-k_1}{\delta c_1}\right]-\frac{1}{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)}\right]
\end{aligned}\tag{3.13}$$

Since $\frac{e^{-\lambda(x_2-b)}}{1-e^{-\lambda(x_2-b)}} > 0$, $\frac{d}{dx_2}x_1(w,x_2) > 0$ if and only if

$$\begin{aligned}
\frac{1}{W(\ldots)+1}\frac{1}{1-e^{-\lambda(x_2-b)}}\left[\frac{1-e^{-\lambda(x_2-b)}}{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)}+\frac{w-k_1}{\delta c_1}\right] &< \frac{1}{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)} \\
\frac{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)}{1-e^{-\lambda(x_2-b)}}\left[\frac{1-e^{-\lambda(x_2-b)}}{1-\delta\left(1-e^{-\lambda(x_2-b)}\right)}+\frac{w-k_1}{\delta c_1}\right] &< W(\ldots)+1 \\
1+\frac{w-k_1}{\delta c_1}\left(\frac{1}{1-e^{-\lambda(x_2-b)}}-\delta\right) &< W(\ldots)+1 \\
\frac{w-k_1}{\delta c_1}\left(\frac{1}{1-e^{-\lambda(x_2-b)}}-\delta\right) &< W(\ldots)
\end{aligned}$$

Using the fact that the Lambert-W function is the inverse function of $f(y) = ye^y$, we obtain

$$\begin{aligned}
&\frac{w-k_1}{c_1}\left(\frac{1}{\delta\left(1-e^{-\lambda(x_2-b)}\right)}-1\right)e^{\frac{w-k_1}{\delta c_1\left(1-e^{-\lambda(x_2-b)}\right)}-\frac{w-k_1}{c_1}} \\
&\qquad < \left(\frac{1}{\delta\left(1-e^{-\lambda(x_2-b)}\right)}-1\right)e^{\frac{w-k_1}{\delta c_1\left(1-e^{-\lambda(x_2-b)}\right)}-1+b\lambda\frac{w-k_1-c_1}{c_1}}
\end{aligned}$$

which holds if and only if

$$\begin{aligned}
\frac{w-k_1}{c_1}e^{-\frac{w-k_1}{c_1}} &< e^{-1+b\lambda\frac{w-k_1-c_1}{c_1}} \\
\log\left(\frac{w-k_1}{c_1}\right)-\frac{w-k_1}{c_1} &< -1+b\lambda\frac{w-k_1-c_1}{c_1} \\
1+\log\left(\frac{w-k_1}{c_1}\right) &< (1+b\lambda)\frac{w-k_1}{c_1}-b\lambda
\end{aligned}$$

This holds for all $\frac{w-k}{c} > 1$ since $b\lambda \geq 0$. □

Proof of Proposition 3.3

We will show that these inequalities hold for general demand distributions. Consequently, they also hold for the case of base demand plus exponential demand. First, we will show the first inequality holds: $x^1(w) < x_1^2(w,x_2)$.

For random demand D with pdf $f(d)$ and CDF $F(d)$ we have:

$$\pi_s(x) = -c_1 x + E[(w-k_1)\min\{D,x\}]$$
$$\pi_s^2(x_1) = \frac{1}{1-\delta R(x) - (1-R(x_1))\frac{\delta(1-R(x_2))}{1-\delta R(x_2)}}\pi_s(x_1)$$

If $d > x$ then the profit equals $(w-k_1-c_1)x$ and if $d \leq x$ then the profit equals $(w-k_1)d - c_1 x$. Therefore, we can write $\pi_s(x)$ as:

$$\begin{aligned}\pi_s(x) =& \int_0^x ((w-k_1)d - c_1 x) f(d)dd + \int_x^\infty (w-k_1-c_1)xf(d)dd\\ =& \int_0^x (w-k_1-c_1)xf(d)dd - \int_0^x (w-k_1)(x-d)f(d)dd\\ &+ \int_x^\infty (w-k_1-c_1)xf(d)dd\\ =&(w-k_1-c_1)x - (w-k_1)\int_0^x (x-d)f(d)dd\end{aligned}$$

leading to the following derivatives:

$$\begin{aligned}\frac{d}{dx}\pi_s(x) =&(w-k_1-c_1) - (w-k_1)F(x)\\ \frac{d}{dx_1}\pi_s^2(x_1) =&\frac{d}{dx_1}\left[\frac{1}{1-\delta R(x_1) - (1-R(x_1))\gamma}\right]\cdot \pi_s(x_1)\\ &+ \frac{1}{1-\delta R(x_1) - (1-R(x_1))\gamma}\cdot\frac{d}{dx_1}\pi_s(x_1)\end{aligned}$$

$$\begin{aligned}=&\frac{(\delta-\gamma)R'(x_1)}{(1-\gamma-(\delta-\gamma)R(x_1))^2}\left((w-k_1-c_1)x_1 - (w-k_1)\int_0^{x_1}(x_1-d)f(d)dd\right)\\ &+ \frac{1}{1-\gamma-(\delta-\gamma)R(x_1)}((w-k_1-c_1)-(w-k_1)F(x_1))\end{aligned}$$

where $\gamma = \frac{\delta(1-R(x_2))}{1-\delta R(x_2)}$.

When considering a single generation, the optimal capacity $x^1(w)$ satisfies $\frac{d}{dx}\pi(x) = 0$, so $x^1(w) = F^{-1}\left(\frac{(w-k_1-c_1)}{w-k_1}\right)$. For $x_1^2(w,x_2) > x^1(w)$, with $x_1^2(w,x_2)$ the capacity for the oligopolistic supplier, we need to show that higher expected profit can be attained by increasing capacity beyond $x^1(w)$, i.e., $\frac{d}{dx_1}\pi_s^2(x_1)|_{x_1=x^1(w)} > 0$. In other words, we need to show that

$$\begin{aligned}\frac{d}{dx_1}\pi_s^2(x_1)|_{x=x^1(w)} = &\frac{(\delta-\gamma)R'\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\right)}{\left(1-\gamma-(\delta-\gamma)R\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\right)\right)^2}\\ &\left((w-k_1-c_1)F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\right.\\ &\left.-(w-k_1)\int_0^{F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)}\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)-d\right)f(d)dd\right) > 0\end{aligned}$$

Since $R(x)$ is increasing in x, it follows that $\frac{(\delta-\gamma)R'\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\right)}{\left(1-\gamma-(\delta-\gamma)R\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\right)\right)^2} > 0$. So $\frac{d}{dx_1}\pi_s^2(x_1)|_{x_1=x^1(w)} > 0$ if and only if

$$\begin{aligned}&(w-k_1-c_1)F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)\\ &\qquad > (w-k_1)\int_0^{F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)}\left(F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)-d\right)f(d)dd\end{aligned}$$

or if we let $y = \frac{w-k_1-c_1}{w-k_1}$:

$$\begin{aligned} & yF^{-1}(y) > \int_0^{F^{-1}(y)}\left(F^{-1}(y)-d\right)f(d)dd\\ \Leftrightarrow\quad & yF^{-1}(y) > F^{-1}(y)\int_0^{F^{-1}(y)}f(d)dd - \int_0^{F^{-1}(y)}df(d)dd\\ \Leftrightarrow\quad & yF^{-1}(y) > F^{-1}(y)F\left(F^{-1}(y)\right) - \int_0^{F^{-1}(y)}df(d)dd\\ \Leftrightarrow\quad & yF^{-1}(y) > yF^{-1}(y) - \int_0^{F^{-1}(y)}df(d)dd\\ \Leftrightarrow\quad & 0 > -\int_0^{F^{-1}(y)}df(d)dd\end{aligned}$$

When $w > k_1 + c_1$, this holds for any probability distribution that has support on the non-negative real numbers. Therefore, the capacity built by the supplier in the oligopolistic supplier case is higher than in the single supplier case.

Next we will show that the second part of the inequality holds: $x_1^2(w, x_2) < x^\infty(w)$.

For this we will show that $\frac{d}{dx_1}\pi_s^2(x_1)|_{x_1=x^\infty(w)} < 0$, i.e., the oligopolistic supplier can attain higher profits by decreasing capacity compared to the unlimited supplier case. For $\frac{d}{dx_1}\pi_s^2(x_1) < 0$ it must hold that:

$$\frac{(\delta-\gamma)R'(x_1)}{(1-\gamma-(\delta-\gamma)R(x_1))^2}(w-k_1-c_1)x_1 + \frac{1}{1-\gamma-(\delta-\gamma)R(x_1)}(w-k_1-c_1) >$$

$$\frac{(\delta-\gamma)R'(x_1)}{(1-\gamma-(\delta-\gamma)R(x_1))^2}(w-k_1)\int_0^{x_1}(x_1-d)f(d)dd$$

$$+\frac{1}{1-\gamma-(\delta-\gamma)R(x_1)}(w-k_1)F(x_1)$$

or equivalently,

$$\frac{w-k_1-c_1}{w-k_1}\left(\frac{(\delta-\gamma)R'(x_1)}{1-\gamma-(\delta-\gamma)R(x_1)}+1\right)$$

$$< \frac{(\delta-\gamma)R'(x_1)}{1-\gamma-(\delta-\gamma)R(x_1)}\int_0^{x_1}(x_1-d)f(d)dd + F(x_1).$$

This can be rewritten as

$$\frac{w-k_1-c_1}{w-k_1}(1-\delta R(x_1)+\delta x_1 R'(x_1)) - \frac{w-k_1-c_1}{w-k_1}(\gamma-\gamma R(x_1)\gamma x_1 R'(x_1))$$

$$< (1-\delta R(x_1))F(x_1) + \delta R'(x_1)\int_0^{x_1}(x_1-d)f(d)dd$$

$$-(\gamma-\gamma R(x_1))F(x_1) - \gamma R'(x_1)\int_0^{x_1}(x_1-d)f(d)dd.$$

Since

$$\begin{aligned}\frac{d}{dx}\pi_s^\infty(x) =& \frac{d}{dx}\left[\frac{1}{1-\delta R(x)}\right]\cdot\pi_s(x) + \frac{1}{1-\delta R(x)}\cdot\frac{d}{dx}\pi_s(x)\\ =& \frac{\delta R'(x)}{(1-\delta R(x))^2}\left((w-k_1-c_1)x-(w-k_1)\int_0^x(x-d)f(d)dd\right)\\ &+\frac{1}{1-\delta R(x)}\left((w-k_1-c_1)-(w-k_1)F(x)\right),\end{aligned}$$

it follows that the optimal $x^\infty(w)$ for the unlimited supplier case satisfies $\frac{w-k_1-c_1}{w-k_1}(1-\delta R(x)+\delta x R'(x)) = (1-\delta R(x))F(x)+\delta R'(x)\int_0^x(x-d)f(d)dd$. This means that in order for $\frac{d}{dx_1}\pi_s^2(x_1)|_{x_1=x^\infty(w)} < 0$, it must hold that:

$$\begin{aligned}
&-\frac{w-k_1-c_1}{w-k_1}(\gamma - \gamma R(x^\infty(w)) + \gamma x^\infty(w) R'(x^\infty(w))) < \\
&\qquad -(\gamma - \gamma R(x^\infty(w)))F(x^\infty(w)) - \gamma R'(x^\infty(w)) \int_0^{x^\infty(w)} (x^\infty(w) - d)\, f(d) dd.
\end{aligned}$$

Rewriting this inequality, using again that the optimal $x^\infty(w)$ for the unlimited supplier case satisfies $\frac{w-k_1-c_1}{w-k_1}(1 - \delta R(x) + \delta x R'(x)) = (1-\delta R(x))F(x) + \delta R'(x) \int_0^x (x-d)\, f(d) dd$, gives

$$\begin{aligned}
&\frac{w-k_1-c_1}{w-k_1}(\delta - \delta R(x^\infty(w)) + \delta x^\infty(w) R'(x^\infty(w))) > \\
&\qquad (\delta - 1)F(x^\infty(w)) + \frac{w-k_1-c_1}{c_1}(1 - \delta R(x^\infty(w)) + \delta x^\infty(w) R'(x^\infty(w))).
\end{aligned}$$

This results in the condition that $F(x^\infty(w)) > \frac{w-k_1-c_1}{c_1}$. Since $x^1(w) = F^{-1}\left(\frac{w-k_1-c_1}{w-k_1}\right)$ and $x^1(w) < x^\infty(w)$ (cf. Proposition 7 Meijer et al. (2021d)), this condition holds and it follows that $\frac{d}{dx_1}\pi_s^2(x_1)|_{x_1=x^\infty(w)} < 0$. Therefore, we can conclude that the second inequality also holds. □

Proof of Proposition 3.4

Proposition 3.2 states that the optimal capacity investment of supplier 1 is increasing in the capacity investment of supplier 2. Furthermore, the capacity of supplier 1 is always larger than 0, also if supplier 2 does not invest, and the derivative of capacity supplier 1 w.r.t. capacity supplier 2, as given by Equation (3.13), converges to 0 as capacity supplier 2 becomes larger. The same holds for the capacity of supplier 2 in response to the capacity of supplier 1. Therefore, there must exist an equilibrium point $(\hat{x}_1, \hat{x}_2)$ such that $x_1^2(w, \hat{x}_2) = \hat{x}_1$ and $x_2^2(w, \hat{x}_1) = \hat{x}_2$. □

Proof of Proposition 3.5

When $b = 0$, it can be shown that $\frac{d}{dx_2}\left[x_1(w, x_2) - x_2\right] < 0$, so that $x_1(w, x_2) - x_2$ is strictly decreasing in x_2. In Proposition 3.4 we have established that an equilibrium solution, denoted by $(\hat{x}_1, \hat{x}_2)$, exists. This equilibrium is located on the line $x_1 = x_2 + C$ for some constant C. For all $\bar{x}_2 > \hat{x}_2$ we know that $x_1(w, \bar{x}_2) < \bar{x}_2 + C$ and $x_2(w, \bar{x}_1 = \bar{x}_2 + C) > \bar{x}_2 + C$. Therefore, $(\hat{x}_1, \hat{x}_2)$ is the only equilibrium solution. □

Proof of Proposition 3.6

The supply chain is coordinated when the OEM sets w such that

$$\frac{\delta}{1-\delta}\left(1-\frac{\frac{c}{r-k}}{1-\delta\left(1-\frac{c}{r-k}\right)}\right)\mathrm{W}\left(\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)e^{\frac{w-k}{\delta c\left(1-\frac{c}{r-k}\right)}-1+b\lambda\frac{w-k-c}{c}}\right) = \frac{r-k}{c}$$

This equality holds if and only if

$$\begin{aligned}
&\mathrm{W}\left(\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)e^{\frac{w-k}{\delta c\left(1-\frac{c}{r-k}\right)}-1+b\lambda\frac{w-k-c}{c}}\right) = \frac{r-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)\\
\Leftrightarrow\ &\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)e^{\frac{w-k}{\delta c\left(1-\frac{c}{r-k}\right)}-1+b\lambda\frac{w-k-c}{c}}\\
&\quad = \frac{r-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)e^{\frac{r-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)}\\
\Leftrightarrow\ &\frac{w-k}{\delta c\left(1-\frac{c}{r-k}\right)}-1+b\lambda\frac{w-k-c}{c} = \log\left(\frac{r-k}{c}\right)+\frac{r-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)\\
\Leftrightarrow\ &\frac{w-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}+b\lambda\right)\\
&\quad = 1+b\lambda+\log\left(\frac{r-k}{c}\right)+\log\left(\frac{r-k}{c}\right)+\frac{r-k}{c}\left(\frac{1}{\delta\left(1-\frac{c}{r-k}\right)}-1\right)\\
\Leftrightarrow\ &(w-k)\left(1+b\lambda\delta\left(1-\frac{c}{r-k}\right)\right)\\
&\quad = c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(r-k)\left(1-\delta\left(1-\frac{c}{r-k}\right)\right)\\
\Leftrightarrow\ &w = k+\frac{c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(r-k)\left(1-\delta\left(1-\frac{c}{r-k}\right)\right)}{1+b\lambda\delta\left(1-\frac{c}{r-k}\right)}
\end{aligned}$$

□

Proof of Proposition 3.7

1. The coordinating wholesale price in the limited supplier case is lower than in the single supplier case if and only if

$$\frac{c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(r-k)\left(1-\delta\left(1-\frac{c}{r-k}\right)\right)}{1+b\lambda\delta\left(1-\frac{c}{r-k}\right)} < r-k$$

This inequality holds if and only if

$$\begin{aligned} c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(r-k)\left(1-\delta\left(1-\frac{c}{r-k}\right)\right)& \\ < (r-k)\left(1+b\lambda\delta\left(1-\frac{c}{r-k}\right)\right)& \end{aligned}$$

$$\begin{aligned} \Leftrightarrow \qquad & c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)-\delta\left(1-\frac{c}{r-k}\right)(r-k) \\ & < b\lambda\delta\left(1-\frac{c}{r-k}\right)(r-k) \\ \Leftrightarrow \qquad & 1+b\lambda+\log\left(\frac{r-k}{c}\right)-\frac{r-k}{c} < b\lambda\frac{r-k}{c} \\ \Leftrightarrow \qquad & 1+b\lambda+\log\left(\frac{r-k}{c}\right) < (1+b\lambda)\frac{r-k}{c} \end{aligned}$$

which holds true since $b\lambda \geq 0$ and $r > k+c$.

2. The coordinating wholesale price in the limited supplier case is higher than in the unlimited supplier case if and only if

$$\begin{aligned} &\frac{c\delta\left(1-\frac{c}{r-k}\right)\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(r-k)\left(1-\delta\left(1-\frac{c}{r-k}\right)\right)}{1+b\lambda\delta\left(1-\frac{c}{r-k}\right)} \\ &\qquad > \frac{\delta c\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right)+(1-\delta)(r-k)}{1+\delta b\lambda} \end{aligned}$$

This inequality holds if and only if

$$c\delta\frac{c}{r-k}\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right) < \delta cb\lambda+\delta c$$

$$\begin{aligned}
\Leftrightarrow \qquad & \frac{c}{r-k}\left(1+b\lambda+\log\left(\frac{r-k}{c}\right)\right) < b\lambda+1 \\
\Leftrightarrow \qquad & 1+b\lambda+\log\left(\frac{r-k}{c}\right) < (1+b\lambda)\,\frac{r-k}{c}
\end{aligned}$$

which again holds true since $b\lambda \geq 0$ and $r > k + c$. □

4

Synchronization between Ordering a Fixed Lead-time Module and Capacitated Make-to-Order Production

A high-tech manufacturer often produces products that consist of many modules. These modules are either sourced from one of its suppliers or produced in-house. In this chapter, we study the common case of an assembly system in which one module is sourced from a supplier with a fixed lead-time, while the other module is produced by the manufacturer itself in a make-to-order production system. Since unavailability of one of the modules has costly consequences for the production of the end-product, it is important to coordinate between the ordering policy for one module and the production of the other. We propose an order policy for the lead-time module with base-stock levels depending on the number of outstanding orders in the production system of the in-house produced module. We prove monotonicity properties of this policy and show optimality. Furthermore, we conduct a computational experiment to evaluate how the costs of this policy compare to those of a policy with fixed base-stock levels and show that average savings of up to 17% are attained.

This chapter is based on Meijer et al. (2021c).

4.1. Introduction

High-tech original equipment manufacturers (OEMs) produce complex products composed of many different modules that are either produced by the OEM itself or sourced from one of its suppliers. To be able to assemble the final product and deliver it to the customer, the OEM needs to organize its production and ordering activities such that all modules are available at the point of assembly and inventory holding and waiting time costs are minimized. The key challenge herein is aligning deliveries from external suppliers with internal production, which is complicated since modules typically have long lead-times and considerable uncertainties are present in the in-house production process.

To study this problem, we consider an assembly system with an end-product consisting of two modules. One of the modules is produced by the OEM in a make-to-order (MTO) fashion, in order for it to be produced in line with specifications in the customer request, while the other module is sourced from a supplier with a given lead-time. The OEM must coordinate the deliveries of the supplier-sourced module with its internal production process. Such coordination challenges arise frequently at high-tech manufacturers and overcoming them is a crucial step towards controlling working capital while catering to ever more demanding customers. A noteworthy example arises at an OEM of wafer-steppers that we have extensively worked with. This OEM produces key modules of their product in its own production facility using highly skilled staff and specialized equipment. In addition to these modules, the final product contains the *lens*: a highly specialized component that has a long and predictable production lead-time and that is sourced from a specific supplier. Coordinating the deliveries of the lens with internal production is a challenging coordination problem.

In general, the importance of coordination in assembly processes is reflected in the amount of research on this general topic. Important results on coordination of ordering decisions for items with *deterministic* lead-times are obtained by Rosling (1989). Coordination of items with stochastic lead-times was first studied by Benjaafar and ElHafsi (2006). The problem studied in this paper falls in a third category that has to our knowledge not received attention in prior work: assembly systems with components with a fairly predictable lead-time as well as components with stochastic lead-times.

We consider two different variants of such assembly systems: a continuous-time model and a discrete-time model. In the continuous-time model, we assume a Poisson demand process and a single-server production system with exponential production times for the in-house produced module. Once a customer order arrives, production of the module can start as soon as there is available capacity. Until capacity becomes available, the order has to wait in the queue. We aim to synchronize the output process of this production system with the order policy of the other module, to avoid large inventories that give rise to holding costs as well as penalty costs incurred for waiting customers. Useful information on the expected production of the MTO module, and thus the required units of the second module, can be obtained from the number of customer orders waiting for production of the MTO module. Since the production capacity of the MTO module is fixed and assembly starts as soon as both modules are available, the inventory levels of both finished modules are influenced by the inventory position of the module sourced from the supplier. We propose a base-stock policy where the target inventory position of the supplier-sourced module is dependent on the number of customer orders waiting for production of the MTO module. We prove monotonicity properties of this policy and show optimality. Numerical results demonstrate that the proposed policy can generate considerable savings compared to a base-stock policy with fixed base-stock levels. We show that optimality of this state-dependent base-stock policy extends to the case of synchronizing the order policy of a lead-time module with the production of multiple MTO modules.

Next, we investigate whether the results we obtained for the continuous-time model can be extended to discrete time. We consider a discrete-time model with a production capacity per period that can either be random or fixed. This model would be more suitable when there is a fixed number of products that can be produced per period or when available equipment has a random yield per period. Also in this setting, we can prove optimality of the base-stock policy with target inventory positions depending on the number of outstanding orders. Numerical results again indicate that considerable savings can be attained by considering information on the number of waiting orders for the MTO module. However, we also observe some differences in the results compared to the continuous-time model. For example, in continuous time we observed an increase in the average percentage savings as lead-time increased, whereas in discrete time we observe a decrease.

This paper is organized as follows. We review relevant literature in Section 4.2. In Section 4.3 we explain our continuous-time model in detail. Using this model we derive the optimal base-stock policy for the module sourced from the supplier and show how to compute the policy parameters based on the state of the in-house production system. A computational experiment is provided in Section 4.3.4, which shows an example of what the inventory policy will look like and compares the expected costs of this state-dependent policy to those of a policy with a fixed base-stock level. In Section 4.4, we show optimality of this policy also for the discrete-time model and show numerically that also in the discrete-time case considerable savings can be attained. In Section 4.5, we reflect on some modeling assumptions. The chapter is concluded in Section 4.6.

4.2. Literature review

Assembly systems have been studied extensively, for example by Schmidt and Nahmias (1985) who provide optimal policies for assembly systems with two components. de Kok et al. (2018) review the extensive literature on multi-echelon inventory management over the past decades, covering convergent, divergent and more general structures and any combination of make-to-order, make-to-stock and assemble-to-order production. Atan et al. (2017) provide an overview of recent literature studying assemble-to-order systems. They state that the main challenge in continuous review models with a single end-product is the synchronization of component orders.

We can classify literature on assembly systems in two groups with respect to capacity constraints. Literature in the first group does not take into account capacity constraints. They assume stochasticity on the demand side, but deterministic lead-times and unlimited supply. In the second group we find literature concerning assembly systems with capacity constraints, resulting in stochastic lead-times.

First, we will focus on literature that studies coordination in uncapacitated assembly systems. Rosling (1989) shows that, under certain conditions, a multi-stage assembly system with fixed assembly times can be reduced to an equivalent serial system, for which optimal policies are derived by Clark and Scarf (1960). More recently, variations of assembly systems have been studied, such as systems with components that have both different lead-times and review periods (Karaarslan et al., 2018). Martínez-de Albéniz and Lago (2010) provide a closed-form formula to determine whether or not an order should be placed. Additionally, they provide conditions for optimality of such myopic policy. Lu et al. (2015) derive (asymptotically) optimal policies for both inventory replenishment and inventory allocation for assemble-to-order N- and W-systems.

Capacity constraints are often modeled using finite-capacity queueing systems. Song et al. (1999) consider a production system with exponential production times in single-server queues with a finite queue and show how to obtain performance measures. In the past decades, such assembly systems with finite production capacities are studied increasingly (e.g. Bollapragada et al., 2015; ElHafsi et al., 2010; Plambeck, 2008; Toktaş-Palut and Ülengin, 2011). Benjaafar and ElHafsi (2006) study an assembly system consisting of m components required to satisfy demand of n customer classes. A policy needs to specify when to produce each component and whether or not to satisfy incoming customer orders from on-hand inventory. They show a base-stock policy with dynamic base-stock levels is optimal. Benjaafar et al. (2011) extend this work to the case where production facilities do not only produce components, but also sub-assemblies. Huh and Janakiraman (2010) focus on base-stock policies, as these are often used in practice. They show convexity of the shortage costs with respect to the order-up-to levels and discuss algorithmic implications. Cheng et al. (2011) study a problem with unpredictable machine breakdowns and endogenous load-dependent lead-times. Song and Zipkin (1993) study an assembly system with stochastic lead-times and Markov-modulated demand, meaning that demand rates are dependent on the state of an underlying variable. Several variations and extensions of the work of Song and Zipkin (1993) have been considered, including the work of Chen and Song (2001). Gallego and Hu (2004) consider, besides a Markov-modulated demand process, also a Markov-modulated supply process that was driven by an independent Markov chain. Furthermore, they consider finite production capacities. Similarly, Mohebbi (2006) studies a situation where supply and demand are subject to independent random environmental conditions where production up to the storage capacity is initiated as soon as the inventory level drops below the limit. Muharremoglu and Tsitsiklis (2008) study a serial system with multiple stages and stochastic lead-times with Markov-modulated demand. They provide an approach for decomposing the serial inventory problem into decoupled subproblems each consisting of a single unit and a single customer. They show that state-dependent base-stock policies are optimal and provide an efficient algorithm to compute the base-stock levels.

Our work combines the two research streams discussed above by studying coordination in an assembly system that combines a module sourced from an unconstrained supplier with fixed lead-time with a module that is produced in a capacitated system with stochastic lead-time. Due to the prevalence of combinations of two such supply streams in practice, this is a relevant topic to study. However, besides its practical relevance, this problem is interesting from a theoretical perspective. Clearly, theory on uncapacitated systems with deterministic lead-times cannot be generalized to situations in which one supply stream has uncertain lead-time, since it is no longer possible to order items based on their lead-time. Furthermore, contrary to most studies on assembly with stochastic lead-times, we assume the capacity investment decision for the stochastic production system to be fixed and coordinate the availability of both modules through the inventory policy of the supplier-sourced module.

4.3. Continuous-time model

In this section we consider a continuous-time model of the assembly system. In Section 4.4 we will consider the case where the assembly process is modeled in discrete time.

Consider a high-tech end product that is composed of two modules. The first module, denoted by m_1, is customer-specific and made to order by the OEM itself, in order for it to be produced in line with specifications in the customer request. Production of m_1 starts as soon as capacity is available after arrival of the customer order; this assumption is discussed in Section 4.5. The second module, denoted by m_2, is sourced from a supplier with lead-time L. Demand of the end product is modeled as a Poisson process with customers arriving at rate λ. We assume that m_1 is produced in a single-server queue with exponential service times with rate μ. As a consequence, the production of module m_1 evolves as an $M/M/1$ queue. A departure from the queuing system then represents a finished m_1 module that can be used to assemble the end product.

When module m_1 is finished and module m_2 is available, the final product can be assembled. If module m_2 arrives before m_1 is available, it needs to be stored and holding costs are incurred. If module m_2 has not yet arrived when module m_1 is finished, m_1 needs to be stored. A sketch of this system is given in Figure 4.1. To formulate our model of the given assembly system, we introduce the additional notation given in Table 4.1.

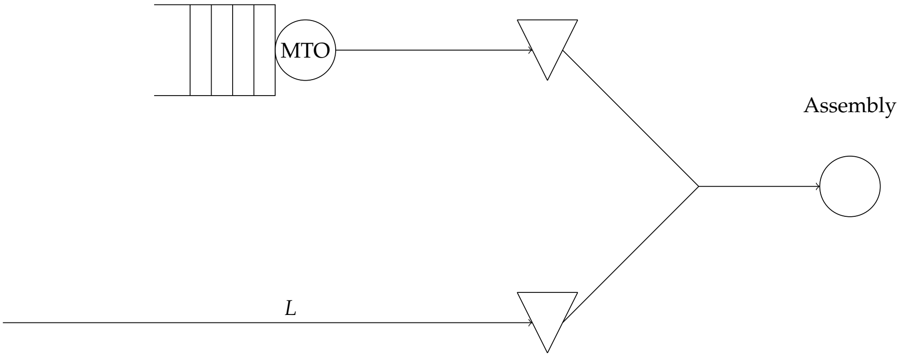

Figure 4.1: Sketch of the assembly system

Table 4.1: Notation

$I_1^P(t)$	number of unfinished orders for module m_1 in the MTO production system at time t
$I_1(t)$	inventory of finished m_1 at time t ready for assembly
$O_2(t)$	ordered m_2 modules that are in transit at time t
$I_2(t)$	inventory of m_2 at time t
$I^A(t)$	number of products in the final assembly step at time t
$IP_2(t)$	inventory position of module m_2 at time t
$M_1(t)$	cumulative production of module m_1 until time t
h_1	unit holding costs for m_1 per time unit
h_2	unit holding costs for m_2 per time unit
b	costs for customers waiting for their final product per time unit
L	lead-time for m_2

Customers want to receive the final product as soon as possible after placing their order, which is represented by waiting costs b. Since module m_1 is made to order and therefore there is a one-to-one correspondence between customer orders and m_1 modules, we can formulate the cost function at time t consisting of four parts:

1. m_1 orders that are waiting to be processed represent customers that are waiting, hence cost b is incurred for every item.

2. Similarly, m_1 modules that are finished, waiting to be assembled also represent customers that are waiting. Additionally, holding costs are incurred for the finished modules, leading to costs $b + h_1$ per unit.

3. m_2 modules that are delivered and are waiting to be merged with module m_1 give rise to holding costs h_2 per unit.

4. Final products that are being assembled consist of m_1 and m_2. Also, every final product is coupled to a customer order and thus represents a waiting customer. Therefore, costs $b + h_1 + h_2$ are incurred.

Combining these parts gives cost rate function

$$C(t) = bI_1^P(t) + (b + h_1)I_1(t) + h_2 I_2(t) + (b + h_1 + h_2)I^A(t). \tag{4.1}$$

Clearly, $I_2(t)$ depends on the ordering decision of module m_2. Since assembly can take place when both modules m_1 and m_2 are available and finished components m_1 are held on inventory when module m_2 is not available, $I_1(t)$ also depends on the ordering policy of module m_2. $I_1^P(t)$ is independent of this ordering policy, as it depends on the incoming orders of the end-product and whether there is capacity available to produce module m_1. The order policy of module m_2 can influence the timing of assembly of the end-product and shift corresponding costs over time. However, since the total number of assemblies that needs to take place is fixed (equal to the total demand of the end-product) and we assume ample assembly capacity, the total costs for assembly of the end-product are independent of the order policy. This means that the costs of interest that can actually be influenced by the order policy of module m_2 are those corresponding to the inventory of both modules:

$$\tilde{C}(t) = (b + h_1)I_1(t) + h_2 I_2(t).^{1} \tag{4.2}$$

Since production of the MTO module starts as soon as possible, we aim to synchronize the ordering of the lead-time module with the output of the MTO production system. We are interested in finding an order policy for module m_2, denoted by π, that minimizes the average costs of operating the assembly system

[1] Since $I_1^P(t)$ is independent of the order policy and the assembly costs can only be shifted in time, it holds that $\lim_{T\to\infty} \frac{1}{T}\int_0^T C(t) - \tilde{C}(t)dt = a$ for some constant a that is independent of the order policy.

over time. This minimization problem is formalized in Equation (4.3), where $\tilde{C}_\pi(t)$ denotes the value of $\tilde{C}(t)$ under order policy π for module m_2.

$$\min_{\pi \in \Pi} \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T \tilde{C}_\pi(t)dt\right] \tag{4.3}$$

4.3.1 Inventory policy for module m_2

Since minimizing the average costs over time is not straightforward, we will first try to solve a related problem of minimizing the expected costs at a certain point in time. After obtaining the solution that minimizes the expected costs at a fixed point in time, we will return to the problem of minimizing the average costs over time. Since decisions at time t affect costs at time $t+L$, we consider the expected cost rate $\tilde{C}(t+L)$ given the relevant information about the state of the system available at time t:

$$E[\tilde{C}(t+L)|I_1(t), I_2(t), O_2(t), I_1^P(t)] = E[\tilde{C}(t+L)|I_1(t), I_2(t)+O_2(t), I_1^P(t)]. \tag{4.4}$$

The equality holds because all m_2 in transit at time t will have been delivered at time $t+L$. Since $\tilde{C}(t+L)$ is determined by the values of $I_1(t+L)$ and $I_2(t+L)$, we are interested in $I_1(t+L)$ and $I_2(t+L)$ given the information available at time t ($I_1(t)$, $I_2(t)+O_2(t)$ and $I_1^P(t)$).

Since $M_1(t)$ denotes the cumulative production of module m_1 until time t, it follows that the production during time interval $[t, t+L]$ can be written as $M_1(t+L) - M_1(t)$. Since module m_1 and m_2 are combined into the final product, the inventory level of module m_1 can decrease by at most the availability of module m_2 during $[t, t+L]$, which is equal to $I_2(t)+O_2(t)$. This means that we can write the inventory of module m_1 at time $t+L$ as

$$I_1(t+L) = \max\left\{I_1(t) + M_1(t+L) - M_1(t) - I_2(t) - O_2(t), 0\right\}.$$

Similarly, we can write

$$I_2(t+L) = \max\left\{I_2(t) + O_2(t) - I_1(t) - (M_1(t+L) - M_1(t)), 0\right\}.$$

Normally, when analyzing an inventory system with a fixed lead-time, one knows that one lead-time from now all currently outstanding orders have been delivered.

This then forms the basis for the balance equations, in which information on demand and outstanding orders is incorporated. However, in this case we have a fixed lead-time L only for module m_2, while module m_1 is produced in a capacitated process. Therefore, it is unknown how many modules m_1 will be finished during the lead-time of module m_2. We thus introduce a different way of modeling this problem, where the demand process of the end-product is transformed by the production process of module m_1 such that the demand process for module m_2 is equal to the production process of m_1. This is reflected in the expressions of $I_1(t+L)$ and $I_2(t+L)$ given above. Since every demand occurrence initiates the production of an m_1 module, production of m_1 during the time interval $[t, t+L]$ can only take place if there was a queue of orders being processed or waiting to be processed at time t or new demand occurred in the time interval $[t, t+L]$. Therefore, demand is indirectly incorporated in these equations through the production of module m_1 which is denoted by $M_1(t+L) - M_1(t)$.

The inventory position of module m_2 increases when a new order for this module is placed. When production of a module m_1 is finished, the module is ready to be merged with module m_2, leading to a decrease in the inventory position of m_2. Therefore, the inventory position of module m_2 at time t is equal to $IP_2(t) = I_2(t) + O_2(t) - I_1(t)$. This allows us to rewrite the expression for expected cost rate given in Equation (4.4) as:

$$E[\tilde{C}(t+L)|IP_2(t), I_1^P(t)] = E[(b+h_1)(M_1(t+L) - M_1(t) - IP_2(t))^+ \\ + h_2(IP_2(t) - (M_1(t+L) - M_1(t)))^+ | I_1^P(t)]. \quad (4.5)$$

The expected cost is thus determined by the production of module m_1 and the inventory position of module m_2, which is controlled by the order policy.

Since the capacity for producing module m_1 is fixed and hence the production of m_1 cannot be influenced, the inventory control policy for module m_2 is the only decision that can influence the inventory levels of both modules and thus the expected costs. Therefore, we aim to find the inventory policy for module m_2 that minimizes the expected costs as given in Equation (4.5). We will consider a myopic inventory policy, where at any time t the target inventory position of module m_2 is determined that minimizes $E[\tilde{C}(t+L)|IP_2(t), I_1^P(t)]$. This yields the minimization problem given in Equation (4.6).

$$\min_{IP_2(t)} E[(b+h_1)(M_1(t+L)-M_1(t)-IP_2(t))^+ \\ + h_2(IP_2(t)-(M_1(t+L)-M_1(t)))^+|I_1^P(t)] \quad (4.6)$$

We denote the minimizing target inventory position by $\tilde{IP}_2(t)$. Observation 4.1 allows us to write $\tilde{IP}_2(t) = \tilde{IP}_2(I_1^P(t))$. This myopic inventory policy thus consists of a list of target inventory position levels for every current state of the system.

Observation 4.1 The target inventory position of module m_2 is a function of the number of waiting orders in the queue for module m_1.

Next, we want to assess whether this myopic policy is a good policy. For this purpose, we will further analyze how the target inventory position $\tilde{IP}_2(I_1^P(t))$ responds to changes in $I_1^P(t)$. First of all, in Theorem 4.1 we show that the target inventory position of module m_2 is non-decreasing in the number of customers in the queue. If this would not hold, there could be situations in which the target inventory position is lower than the current inventory position and one would like to place a negative order. Similarly, Theorem 4.2 shows that the increase of $\tilde{IP}_2(I_1^P(t))$ is at most one when $I_1^P(t)$ increases by one. All proofs are given in the Appendix.

Theorem 4.1 *$\tilde{IP}_2(I_1^P(t))$ is monotonically non-decreasing in $I_1^P(t)$.*

Theorem 4.2 *If an additional customer enters the system, the target inventory position of module m_2 increases by at most 1, i.e. $\tilde{IP}_2(I_1^P(t)+1) - \tilde{IP}_2(I_1^P(t)) \leq 1$*

Combining Theorem 4.1 and Theorem 4.2, we conclude that if $I_1^P(t)$ increases by one, the target IP level remains the same or increases by one. We give an illustration of this in Figure 4.2. On the horizontal axis we have the number of waiting order for production of module m_1 and on the vertical axis the inventory position of module m_2. The black dots indicate the target inventory position of m_2 for different values of $I_1^P(t)$. The dotted arrows correspond to an assembly step where both m_1 and m_2 are used, reducing both $I_1^P(t)$ and $IP_2(t)$ by one. The solid arrows correspond to a customer order arrival, leading to an increase in $I_1^P(t)$ of 1. The dashed arrows show the required orders of module m_2 to reach the target inventory position. We

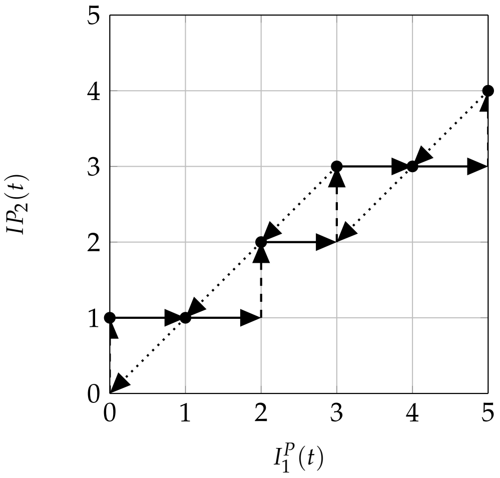

Figure 4.2: Illustration of order policy. Solid arrows correspond to customer order arrivals, dashed arrows to orders of m_2 and dotted arrows to assembly.

observe that both the solid and the dotted arrows never go to a point where the inventory position of module m_2 exceeds its target. Consequently, all dashed arrows correspond to an order of size 1. Hence, this policy behaves well in all situations, in the sense that it never prescribes negative orders.

The next question is how this policy performs when we return to the original problem in which we aim to minimize the average costs over time. When we consider the optimal inventory policy that minimizes the average costs there also is a corresponding inventory position $IP_2^*(t)$ at every time t. By definition, the myopically optimal target inventory position minimizes the costs at every point in time. Therefore, we can conclude that this myopic policy is also optimal for minimizing average costs over time and thus that $IP_2^*(t) = \tilde{IP}_2(I_1^P(t))$. This is formalized in Theorem 4.3.

Theorem 4.3 *The myopic inventory policy for module m_2 is optimal for minimizing average costs.*

4.3.2 Computing $\tilde{IP}_2(I_1^P(t))$

Now that we have characterized the myopic inventory policy, we are interested in how the target inventory level $\tilde{IP}_2(I_1^P(t))$ can be computed for a given value of

$I_1^P(t)$. This requires us to take a closer look at the rate at which module m_2 is used in the assembly process and the cost structure.

When module m_1 is finished and module m_2 is available, the final product can be assembled. However, if module m_2 arrives before m_1 is available it needs to be stored and holding costs are incurred and if module m_2 has not yet arrived when module m_1 is finished, m_1 needs to be stored and the customer is waiting. In other words, the minimization problem given in Equation (4.6) is a Newsvendor problem with shortage costs for module m_2 equal to $b + h_1$ and overage costs h_2.

The requirement for module m_2 in the assembly system during period $[t, t + L]$ is equal to the production of module m_1 during $[t, t + L]$, which is denoted by $M_1(t + L) - M_1(t)$. Since the production during $[t, t + L]$ depends on the number of waiting customers at time t, the distribution of $M_1(t + L) - M_1(t)$ depends on $I_1^P(t)$. Therefore, when synchronizing the order policy of module m_2 with the production of module m_1, we need to consider the distribution of $M_1(t + L) - M_1(t)$ given $I_1^P(t)$. We then want to determine $\tilde{IP}_2(I_1^P(t))$ such that $P(M_1(t + L) - M_1(t) < \tilde{IP}_2(I_1^P(t))) \geq CF$, where $CF = \frac{b+h_1}{b+h_1+h_2}$ is the critical fractile.

Every demand occurrence of the end-product triggers the production of the MTO module m_1. Since the end-product can only be assembled once both the MTO module and the lead-time module are available, we can use the output of the MTO production system during time $[t, t + L]$ as the demand process of the lead-time module. To derive the distribution of the production of this module during the lead-time of the second module, we model the state of the inventory system, consisting of the number of waiting customer orders and the number of finished modules m_1, as a Markov process. Let (i, j) denote the state of the system with i the number of m_1 jobs waiting or currently in process in the system and j the number of m_1 modules produced, with $i = 0, 1, 2, \ldots$ and $j = 0, 1, 2, \ldots$. When time is scaled such that $\lambda + \mu = 1$, a customer order arrival occurs with probability λ and generates a transition from (i, j) to $(i + 1, j)$. An exit corresponds to finished production of a module m_1. This occurs with probability μ. When there are customers in the system ($i > 0$), a completion of module m_1 means that the process moves to state $(i - 1, j + 1)$. When $i = 0$, meaning that there are no customer orders in the system, we do allow for exits, but the system then remains in the same state. This yields the following transition probabilities:

$$P_{(i,j),(i+1,j)} = \lambda,$$
$$P_{(i,j),(i,j)} = \mu \text{ if } i = 0, \text{ and}$$
$$P_{(i,j),(i-1,j+1)} = \mu \text{ if } i > 0.$$

We are interested in the number of modules m_1 produced by time $t + L$, given that there are $I_1^P(t)$ jobs waiting in the system at time t. In other words, we are interested in the value of j after L time units, starting from state $(I_1^P(t), 0)$.

In order to analyze the production during $[t, t + L]$, we condition on the total number of transitions, consisting of both customer arrivals and finished production of a module, during time period $[t, t + L]$, denoted by X. When we condition on X, both the number of customers in the queue and the number of modules m_1 that are produced during time interval $[t, t + L]$ are bounded. As a result the state space is bounded and we can define the transition probability matrix, which we denote by $\bar{P}$. Let s be a vector with length equal to the number of states with all zero entries, except for a one at the state corresponding to $(I_1^P(t), 0)$. The probability distribution over the state space after X= k events is then given by $s\bar{P}^k$. Since the production of m_1 is modeled as an $M/M/1$ queue, X has a Poisson distribution with parameter $(\lambda + \mu)L$, i.e.

$$P(X = k) = \frac{((\lambda + \mu)L)^k}{k!} e^{-(\lambda+\mu)L}.$$

For each realization of $X = k$ we can determine the probability distribution over the output of m_1 from $s\bar{P}^k$, from which we obtain the unconditional distribution over the state space after L time units starting from state $(I_1^P(t), 0)$. Then the distribution over the number of modules m_1 produced in the time interval $[t, t + L]$ can be obtained from Equation (4.7).

$$\begin{aligned} &P(M_1(t + L) - M_1(t) = j | I_1^P(t) = i) \\ &\qquad = \sum_{k=0}^{\infty} P(X = k) P(M_1(t + L) - M_1(t) = j | X = k, I_1^P(t) = i) \end{aligned} \tag{4.7}$$

Since the support of the Poisson distribution is all natural numbers starting from 0, i.e. $k \in \mathbb{N}_0$, we need to bound the states space in order to obtain closed-form expressions. Therefore, we introduce an upper bound on the number of transitions during $[t, t+L]$, denoted by X^U, such that $P(X \geq X^U)$ is negligible. We specify X^U using Cantelli's inequality, which is a one-sided Chebyshev inequality, as described by Ghosh (2002). Since $X \sim \mathit{Poisson}((\lambda+\mu)L)$, this gives the upper bound provided in Lemma 4.1.

Lemma 4.1 *For any* $0 < \epsilon < 1$, $P(X \geq X^U) \leq \epsilon$ *holds if* $X^U = (\lambda+\mu)L + \sqrt{\left(\frac{1}{\epsilon}-1\right)(\lambda+\mu)L}$.

To illustrate this procedure, we will now provide a small-scale example that shows the intuition behind this approach.

4.3.2.1 Small scale example production m_1

We consider a small example in which the parameters are such that the probabilities of having more than 2 customers in the queue or of producing more than 2 units of module m_1 during $[t, t+L]$ are negligible. This gives the following states and transition probabilities, where $(0,3)$ is added as an absorbing state such that the probabilities in each row add up to one:

$$\bar{P} = \begin{array}{c|cccccccccc|}
 & (0,0) & (1,0) & (2,0) & (0,1) & (1,1) & (2,1) & (0,2) & (1,2) & (2,2) & (0,3) \\
\hline
(0,0) & \mu & \lambda & & & & & & & & \\
(1,0) & & & \lambda & \mu & & & & & & \\
(2,0) & & & \lambda & & \mu & & & & & \\
(0,1) & & & & \mu & \lambda & & & & & \\
(1,1) & & & & & & \lambda & \mu & & & \\
(2,1) & & & & & & \lambda & & \mu & & \\
(0,2) & & & & & & & \mu & \lambda & & \\
(1,2) & & & & & & & & & \lambda & \mu \\
(2,2) & & & & & & & & & \lambda & \mu \\
(0,3) & & & & & & & & & & \lambda+\mu
\end{array}$$

Assuming we start in state $(0,0)$, the starting vector is $s = [1,0,0,0,0,0,0,0,0,0]$. By conditioning on $X = 2$, we obtain the distribution over the state space given by:

$$s\bar{P}^2 = [\mu^2, \lambda\mu, \lambda^2, \lambda\mu, 0,0,0,0,0,0]$$

Since the first three states ($\{(0,0),(1,0),(2,0)\}$) correspond to $M_1(t+L) - M_1(t) = 0$, the following three states ($\{(0,1),(1,1),(2,1)\}$) to $M_1(t+L) - M_1(t) = 1$, etc. we obtain the following conditional probabilities:

$$\begin{aligned}
&P(M_1(t+L) - M_1(t) = 0 | I_1^P(t) = 0, X = 2) = \mu^2 + \lambda\mu + \lambda^2 \\
&P(M_1(t+L) - M_1(t) = 1 | I_1^P(t) = 0, X = 2) = \lambda\mu \\
&P(M_1(t+L) - M_1(t) = 2 | I_1^P(t) = 0, X = 2) = 0
\end{aligned}$$

Similarly, we can obtain the conditional probability distribution over the production of m_1 for different values of X. Using Equation (4.7) we can then determine the distribution of the production given that there are $I_1^P(t) = 0$ customer orders waiting in the queue.

When starting with $i = 1$ and starting with $i = 2$ customer orders in the queue, we can repeat the calculations by using a different vector s (i.e., $s = [0,1,0,0,0,0,0,0,0,0]$ and $s = [0,0,1,0,0,0,0,0,0,0]$, respectively).

4.3.3 Cost calculations

Now that we have formulated the inventory policy with target inventory position $\tilde{IP}_2(i)$ when $I_1^P(t) = i$ and modeled the production of module m_1, we want to compare the costs of this policy to the costs of a policy with a fixed target inventory position. First, we formulate the expected costs of the proposed policy in Equation (4.8). By conditioning on the value of $I_1^P(t) = i$ we determine the expected costs for every i. Since the stationary probability of having i customers waiting in an M/M/1 queue is equal to $P(I_1^P(t) = i) = (1-\rho)\rho^i$, we can then calculate the overall expected costs.

$$\begin{aligned}
E[\tilde{C}(t+L) | \tilde{IP}_2(t)] &= \sum_{i=0}^{\infty} P(I_1^P(t) = i) E[\tilde{C}(t+L) | \tilde{IP}_2(t), I_1^P(t) = i] \\
&= \sum_{i=0}^{\infty} (1-\rho)\rho^i \sum_{j=0}^{X^U} P(M_1(t+L) - M_1(t) = j | I_1^P(t) = i) \\
&\qquad \left((b + h_1)(j - \tilde{IP}_2(i))^+ + h_2(\tilde{IP}_2(i) - j)^+\right) \qquad (4.8)
\end{aligned}$$

For an inventory policy with fixed target inventory position IP_2, the expected costs are given in Equation (4.9). Without taking into account the current state of the queue, the output process of the M/M/1 queue during $[t, t+L]$ in steady state is a Poisson process with rate λL. Therefore, $P(M_1(t+L) - M_1(t) = j) = \frac{(\lambda L)^j}{j!}e^{-\lambda L}$.

$$E[\tilde{C}(t+L)|IP_2(t)] = \sum_{j=0}^{X^U} \frac{(\lambda L)^j}{j!} e^{-\lambda L} \left((b+h_1)(j-IP_2)^+ + h_2(IP_2-j)^+\right) \qquad (4.9)$$

4.3.4 Computational analysis

In this section, we evaluate the effectiveness of the inventory policy that takes into account the customers currently waiting at the in-house production facility of module m_1 by means of a computational experiment. The results of this analysis are provided in Section 4.3.4.2. First, in Section 4.3.4.1 we illustrate how the procedure described in Section 4.3 results in a table of target inventory positions for different numbers of waiting customer orders. The complexity of the algorithm results from determining the state distribution after a certain number of events for each starting state. Therefore, the complexity is $O(S^3)$, where S denotes the size of the state space. Note that in case the number of events that can occur during the lead-time of module m_2 is large and hence the output size can be large, there are a lot of states that need to be considered. For example, when we increase lead-time such that the maximum number of events grows from 81 to 104, the computation time for the probability distribution of the output and the corresponding optimal policy increases from 0.64 to 0.92 seconds, where the time required for finding the optimal policy is negligible. We can thus observe that this maximum number of events has a large effect on the size of the state space and thus on the computation time.

4.3.4.1 Illustrative example of inventory policy

Consider the following numerical example: $\lambda = 0.8$, $\mu = 1$, $L = 4$, $h_1 = 4$, $h_2 = 1$, $b = 5$, giving critical fractile $CF = 0.9$. In Table 4.2 we tabulate the target IP for module m_2 for different values of $I_1^P(t)$.

Without taking into account the current state of the queue, the output process of the M/M/1 queue in steady-state during $[t, t+L]$ is a Poisson process with rate λL. Using this information, we can determine the target inventory position for which the cumulative probability reaches the critical fractile. In the given example, this leads to a target inventory position of 6.

Table 4.2: Target inventory position for different values of $I_1^P(t)$

$I_1^P(t)$	Target IP
0	4
1	4
2	5
3	6
4	6
5	6
≥ 6	7

4.3.4.2 Policy evaluation

In order to assess the value of taking into account information on outstanding orders in the order policy and its sensitivity to various model parameters, we perform a full factorial experiment. In our experiment, we vary the rate of production relative to the rate of incoming orders, the lead-time of module m_2, the holding costs of module m_1 relative to those of module m_2 and the waiting time costs for customers. The setup of the experiment is given in Table 4.3. We set $\mu = 1$ and $h_2 = 1$. In total we have $2 \cdot 3^3 = 54$ instances.

Table 4.3: Parameter settings for experiments

Parameter	Values
λ	0.8, 0.9
L	2, 3, 4
h_1	1, 2, 4
b	1, 5, 10

For each instance we calculate the expected costs at time $t + L$ under the reorder policy with fixed target inventory position and the policy with state-dependent target inventory position and the cost savings that can be achieved by taking into account information on customer orders. The parameters given in Table 4.3 are fixed one at a time and the remaining parameters are varied. The summary statistics are provided in Table 4.4. The last row provides the summary over all 54 instances.

According to the results provided in Table 4.4, on average 17.00% cost reduction can be achieved by letting the target inventory position depend on the number of waiting customers. The minimum and maximum cost reduction are 8.72% and 22.07%, respectively. The benefit of using the available information on the number

Table 4.4: Results of full factorial experiment continuous-time model

		Avg cost		Savings (%)		
		Fixed IP	State-dep. IP	Average	Max	Min
λ	0.8	2.71	2.30	14.65	18.27	8.72
	0.9	2.87	2.31	19.34	22.07	15.02
L	2	2.34	1.95	16.05	20.90	8.72
	3	2.82	2.33	16.95	20.77	10.68
	4	3.22	2.63	17.98	22.07	13.33
h_1	1	2.57	2.14	16.08	21.23	8.72
	2	2.77	2.28	17.45	22.07	11.62
	4	3.03	2.49	17.46	21.85	12.57
b	1	2.14	1.79	15.68	20.85	8.72
	5	2.88	2.39	16.95	20.90	10.51
	10	3.35	2.74	18.36	22.07	15.80
All		2.79	2.31	17.00	22.07	8.72

of waiting customer orders increases considerably in the arrival rate. When the arrival rate is low compared to the service rate, the number of waiting customers is likely to be small and the Poisson process with rate λL will be a good approximation of the production of module m_1, hence there is less value in using this information. However, when λ is close to μ, the queue of waiting customers may be more substantial and the information on the number of outstanding customer orders becomes more useful. Furthermore, the value of this information is higher when the holding cost of module m_1 and/or the customer waiting costs are high relative to the holding costs of module m_2 or when the lead-time of module m_2 is high. Overall, Table 4.4 shows that including the information on waiting customer orders in the order policy leads to substantial savings.

4.3.5 Extension: Multiple MTO modules

Until now, we considered the synchronization of a lead-time module with a single MTO module for which the in-house production evolves as an $M/M/1$ queue. However, in reality end-products may consist of multiple modules that need to be assembled, rather than just two. Consider for example the production of wafer-steppers that was mentioned in Section 4.1, where multiple key modules are produced at the OEM's own production facility, but the lens is sourced from an outside supplier with long but stable lead-times. Therefore, in this section we will extend this analysis to the synchronization of a lead-time module with N MTO modules, where the production of each of these modules evolves as an $M/M/1$ queue with arrival rates equal to the demand rate of the end-product. We again assume that production of each of the MTO modules commences as soon as there is capacity available in the corresponding production system. Whenever all MTO modules and the lead-time module, denoted by m_{N+1}, are available, assembly of the final product can take place. The requirement for the lead-time module in the assembly system during period $[t, t+L]$ is thus equal to the minimum of the production of all MTO components during this time period.

Similar to the original model, we incur holding costs and customer waiting costs when all MTO modules are ready, but assembly cannot take place as the lead-time module is missing. Holding costs for the lead-time module are incurred when this module is available while (one of the) MTO modules are (is) missing. Additionally, when one or more of the MTO modules are available while others are not, holding costs for the available module(s) and customer waiting costs are incurred. However, since the inventory policy of the lead-time module has no influence on these costs, these costs are considered to be exogenous. Therefore, we can determine the myopically optimal inventory position of the lead-time component as a function of the current state of the production system.

To analyze the joint production of the MTO modules, we model the state of the production system as a Markov process with $N+1$-dimensional state space. For the first MTO module, we include information on the number of outstanding orders and the number of finished modules. For the other MTO modules it suffices to keep track of the number of finished modules. Let $(i_1, j_1, \ldots, j_N)$ denote the state of the system with i_1 the number of jobs waiting for production of module m_1 and j_k, $k \in \{1, \ldots, N\}$ the number of modules m_k that are finished, with $i_1 = 0, 1, 2, \ldots$ and $j_k = 0, 1, 2, \ldots$ for $k \in \{1, \ldots, N\}$. Since production of all MTO modules m_k is triggered by a customer order and all modules are combined in the assembly of the end-product, for each state the corresponding number of jobs waiting for production of module m_k is equal to $i_k = i_1 + j_1 - j_k$. We can then find the distribution of the requirement for the lead-time module in the assembly system during period $[t, t+L]$ in a similar way as for the initial case with a single MTO module. Based on this, we determine the target inventory position of the lead-time module such that $P(\min\{M_1(t+L) - M_1(t), \ldots, M_N(t+L) - M_N(t)\} < \tilde{IP}_{N+1}(i_1, j_1, \ldots, j_N)) \geq CF$, where $CF = \frac{b + \sum_{k=1}^{N} h_k}{b + \sum_{k=1}^{N+1} h_k}$ is the critical fractile.

Similar to the case with a single MTO module, Theorems 4.4 and 4.5 show that the arrival of an additional customer order will never decrease the target inventory position of the lead-time module and increases it by at most one. Therefore, this myopic policy behaves well when considering multiple MTO modules in the sense that it will never prescribe negative orders.

Theorem 4.4 *The target inventory position of the lead-time module is monotonically non-decreasing in the number of outstanding customer orders.*

The reasoning remains the same as in the single MTO module case, namely that we use the output process of the production system of the MTO modules as the demand process for the lead-time module. Therefore, it is not surprising that this monotonicity result extends to the case with N MTO modules. Similar to the single MTO module case, the arrival of an additional customer order will never reduce the output of the in-house production process. Therefore, intuitively, it also seems reasonable that the target inventory position of the lead-time module does not decrease.

Theorem 4.5 *If an additional customer enters the system, the target inventory position of the lead-time module increases by at most 1.*

Similar reasoning can be used to explain why the target inventory position increases by at most one, as is shown in Theorem 4.5. In Theorem 4.2, this result was proven for the case with a single MTO module. It was shown that an additional customer order can increase the output of the production process of the MTO module by at most one. Now we consider the case with N MTO modules. If the output of the combined MTO modules were to increase by more than one, then for at least one of the MTO modules the output would need to increase by more than one. In Theorem 4.2 we showed that this does not happen in case of a single MTO module and therefore intuitively Theorem 4.5 should also hold.

Now that we have established that this myopic policy behaves well also for the case of multiple MTO components and thus never prescribes negative orders, we can again show that this myopic policy is optimal for synchronizing the ordering of the lead-time module m_{N+1} with the in-house production of the N MTO modules. This is formalized in Theorem 4.6.

Theorem 4.6 *The myopic inventory policy for module m_{N+1} is optimal for minimizing average costs.*

The structure of the optimal policy does not depend on the number of MTO modules in the in-house production system. Both in case of a single MTO module and in case of multiple MTO modules we have shown that the myopic state-dependent base-stock policy that takes into account the number of outstanding orders in the in-house production system is optimal. The optimal base stock levels for the case with multiple MTO modules can be determined in the same way as for the case with a single MTO module, with complexity $O(S^3)$. However, since the dimensionality of the state space is now $N+1$, the size of the state space will be much larger than in the case with a single MTO module. Consequently, the computation time for finding the optimal base-stock levels in case of multiple MTO modules will also be larger than in case of a single MTO module.

4.4. Discrete-time model

In the previous sections we have shown that an inventory policy with a state-dependent target inventory position is optimal and can generate considerable savings compared to a policy with a fixed target inventory position in a continuous-time model. Such a model may not be suitable for all assembly systems, for example when using periodic review. Therefore, we will now consider a model in discrete time. We will again consider the assembly system shown in Figure 4.1, consisting of two modules that need to be merged in a single end-product. Module m_2 is sourced from a supplier with a lead-time of L periods. For the other module there is in every period available production capacity that may either be fixed or random. We denote the (random) number of units that can be produced per period by C and demand per period by D.

The production of module m_1 can still be modeled as a Markov process with states (i, j), where i denotes the number of outstanding customer orders and j the number of units produced. Every period production is equal to the minimum of available capacity (C) and available customer orders consisting of both new demand and outstanding orders ($i + D$). If the total number of orders exceeds available capacity, the remaining orders will still be outstanding orders at the beginning of the next period. This means that we have the following transition:

$$(i, j) \to \left((i + D - C)^+, j + \min\{i + D, C\}\right).$$

The transition probabilities can be determined based on the distributions of demand and capacity.

The cost structure remains the same as in the continuous-time model. In every period holding costs h_1 are incurred for finished modules m_1 that need to be stored and similarly we have holding costs h_2 for m_2. Additionally, there is a per-period back-order cost b. Costs per period are as given in Equation (4.1). Using similar reasoning as for the continuous-time model, we can write the myopic inventory policy at time t as:

$$\min_{IP_2(t)} E[(b + h_1)(M_1(t + L) - M_1(t) - IP_2(t))^+ \\ + h_2(IP_2(t) - (M_1(t + L) - M_1(t)))^+ | I_1^P(t)]. \quad (4.10)$$

We can again show that the myopic inventory policy, in which the target inventory position is determined for every period separately, is optimal. For this we prove Theorems 4.7 and 4.8, which are similar to Theorems 4.1 and 4.2 in the continuous-time model. These theorems show that the target inventory position of module m_2 is non-decreasing in the number of outstanding customer orders and increases by at most 1 as the number of outstanding orders increases by 1.

Theorem 4.7 *For every period t, $\tilde{IP}_2(I_1^P(t))$ is monotonically non-decreasing in $I_1^P(t)$.*

Theorem 4.8 *For every period t, an additional customer order in the system increases the target inventory position of module m_2 by at most 1.*

Similar to the continuous-time model, we can use Theorems 4.7 and 4.8 to conclude that the proposed policy behaves well also for the discrete-time model. Furthermore, we can use the same reasoning as in Section 4.3.1 to conclude that the myopic policy for module m_2 is optimal in discrete time in Theorem 4.9.

Theorem 4.9 *In a discrete-time setting with demand D and capacity C per period, the myopic inventory policy for module m_2 is optimal.*

In summary, the proposed inventory policy is not only optimal in the continuous-time $M/M/1$ model, but all analytical results continue to hold when we consider a setting with discrete time periods and a capacity per period that is either fixed or random.

4.4.1 Computational analysis

Now that we have established that the proposed myopic policy is optimal also in the discrete-time case, we will evaluate the effectiveness of this policy compared to a base-stock policy with a fixed base-stock level. In our experiment, we vary the lead-time of module m_2, the holding costs of module m_1 relative to those of module m_2 and the waiting time costs for customers. We set $h_2 = 1$ and use the same parameter values for L, h_1 and b as given in Table 4.3. Furthermore, we consider different cases with respect to the distributions for demand and capacity:

$$\text{Case 1:} \quad P(D=d) = \begin{cases} 0.4 & \text{if } d=1 \\ 0.4 & \text{if } d=2, \\ 0.2 & \text{if } d=3 \end{cases} \quad P(C=c) = \begin{cases} 0.7 & \text{if } c=2 \\ 0.3 & \text{if } c=3 \end{cases}$$

$$\text{Case 2:} \quad P(D=d) = \begin{cases} 0.1 & \text{if } d=0 \\ 0.2 & \text{if } d=1 \\ 0.2 & \text{if } d=2 \\ 0.25 & \text{if } d=3' \\ 0.15 & \text{if } d=4 \\ 0.1 & \text{if } d=5 \end{cases} \quad P(C=c) = \begin{cases} 0.2 & \text{if } c=2 \\ 0.45 & \text{if } c=3 \\ 0.35 & \text{if } c=4 \end{cases}$$

In total we thus have again $2 \cdot 3^3 = 54$ instances. For each instance we calculate the expected costs at time $t+L$ under both the reorder policy with fixed target inventory position and the policy with state-dependent target inventory position. Subsequently, we determine the cost savings achieved by using the state-dependent policy. We again perform a full factorial experiment, where we vary the parameter values or demand and capacity distribution cases one by one. We determine the distribution of the number of outstanding orders for module m_1 by starting with an empty system and taking the distribution over the number of outstanding orders after n periods, where n is selected such that the probabilities have stabilized. The results of the experiment are given in Table 4.5.

Similar to the continuous-time model, we observe an increase in the average savings when considering the information on outstanding orders of module m_1 in the inventory policy for module m_2 as the holding costs h_1 increase. There are also some differences compared to the results for the continuous-time model. In the continuous-time case, the average savings percentage increased as the lead-time increased, whereas in the discrete-time case we observe the opposite effect. When $L = 2$ the average savings from incorporating the information on outstanding orders for module m_1 are 11.70% and for $L = 4$ this has decreased to 9.05%. This is due to the fact that we now have discrete distributions for demand and capacity with a finite number of outcomes. Therefore, there is no possibility of high peaks in the number of demands occurring during a certain time period and, hence, in the discrete case, the value of synchronization reduces as the lead-time grows large. When we consider the two cases with respect to the distributions of demand and capacity, we observe that the expected demand relative to the expected capacity is comparable in both cases. In the second case the variation in both capacity and demand per period is larger. This has a large effect on the average savings, as in case 1 the average savings are equal to 6.84% and in case 2 the average savings are 13.47%.

Table 4.5: Results of full factorial experiment discrete-time model

		Avg cost		Savings (%)		
		Fixed IP	State-dep. IP	Average	Max	Min
Case	1	1.55	1.45	6.84	25.02	0.02
	2	2.85	2.45	13.47	26.28	4.24
L	2	1.78	1.53	11.70	26.28	0.02
	3	2.21	1.96	9.71	21.28	1.21
	4	2.62	2.35	9.05	18.79	1.77
h_1	1	2.06	1.83	9.50	21.94	0.02
	2	2.14	1.93	8.53	20.78	0.04
	4	2.40	2.08	12.42	26.28	0.19
b	1	1.76	1.59	9.44	25.02	2.60
	5	2.28	2.00	11.71	26.28	1.36
	10	2.56	2.26	9.31	21.94	0.02
All		2.20	1.95	10.15	26.28	0.02

4.5. Discussion

Throughout this paper we have assumed that the production of module m_1 starts as soon as production capacity is available. Sometimes one may want to consider the possibility of postponing this production start time, to avoid inventory holding costs for module m_1 while m_2 is not yet available. However, this would generate high risks. If you decide right now to postpone the production of the MTO module and after that many customer orders arrive, you may run into long customer waiting times that could have been reduced by using this idle capacity. The long-term costs resulting from this postponement decision may thus be much higher than the short-term savings in inventory holding costs.

In Lemma 4.2 we formalize this argument. We consider the continuous-time model with a single MTO module and assume that the inventory of module m_2 is controlled using a base-stock policy with positive base-stock level. Under that assumption, we show that postponing production is not attractive when the ratio of customer waiting costs over holding costs for module m_1 is sufficiently large, as is usually the case in high-tech manufacturing.

Lemma 4.2 *When $\frac{b+h_2}{h_1} > e^{\mu L}$ it is optimal to start production of module m_1 as soon as there is available production capacity rather than postponing production.*

When the number of MTO modules is larger than one, it could be beneficial to coordinate among the production processes of the MTO modules to avoid large inventories of some of these MTO modules while others are not yet finished. However, combined with determining the order policy of the lead-time module this results in a complex problem that is outside the scope of this paper.

Additionally, it would be interesting to consider the case with a single MTO module and multiple modules that are sourced at a supplier with fixed lead-time. This would resemble practical cases where a manufacturer produces a single module in-house and sources all other modules at external suppliers. This system can be modeled in a way that is similar to the system studied by Rosling (1989), but with demand for the lead-time modules equal to the output of the MTO production system instead of the end-product demand. This is a crucial difference, as this means that demand for the lead-time modules at different times is dependent and not i.i.d. as in Rosling's work. Therefore, we cannot use the same approach to transform this system to an equivalent serial system. Hence, the model with multiple lead-time components alongside a single MTO component is intractable.

4.6. Conclusions

We have examined an assembly problem where an OEM builds a final product consisting of two modules, one of which is made-to-order by the OEM itself. Since production of the make-to-order module commences as soon as a customer order arrives and the final product can be assembled as soon as both modules are available, intermediate stocks and thus costs are controlled by the order policy of the module sourced from the supplier. Since the number of orders waiting for production of the in-house produced module gives an indication of the demand for the other module one lead-time from now, this information can be useful in determining the inventory policy for this module. Therefore, we consider an inventory policy where the target inventory position depends on the state of the production system for the in-house produced module. When there is a large number of orders waiting for production of the module, production during the lead-time of the other module is likely to be higher than in case none or very few orders are waiting. Consequently, the target inventory position of the supplier-sourced module will be higher when there are many customer orders waiting in the in-house production system.

We show that under this policy the target inventory position is monotonically non-decreasing in the number of customer orders in the queue. Additionally, we show optimality of this policy both in continuous and discrete time. We show that this approach extends to the case of synchronizing the order policy of a lead-time module with the production of multiple MTO modules. In support of our analytical results and to illustrate the proposed policy, we conducted a computational analysis. In this analysis we performed a full factorial experiment to evaluate the benefits of taking information on outstanding customer orders into account. We show that using this information can lead to considerable savings. Furthermore, we assess sensitivity to various model parameters and show that especially the arrival rate of customer orders relative to the production rate has a large influence on the savings. Furthermore, we show that there are differences in this sensitivity between the continuous-time and the discrete-time model. In the continuous-time case we observed that a longer lead-time clearly leads to higher savings from using the state-dependent base-stock policy, while in the discrete time case the effect is opposite.

4.A. Proofs

Proof of Theorem 4.1

We aim to show that $\tilde{IP}_2(I_1^P(t))$ is monotonically non-decreasing in $I_1^P(t)$. Note that $\tilde{IP}_2(I_1^P(t))$ is the minimizer of Equation (4.6). This is a typical Newsvendor equation with underage costs $b + h_1$ and overage costs h_2. Hence, the target inventory position of module m_2 that minimizes the expected costs, denoted by $\tilde{IP}_2(I_1^P(t))$, satisfies $P\left(M_1(t+L) - M_1(t) \leq IP_2(t)|I_1^P(t)\right) \geq \frac{b+h_1}{b+h_1+h_2}$.

To prove our result, we show that for every possible sample path (sequence of customer arrivals and production of m_1) $M_1(t+L) - M_1(t)$ will either stay the same or will increase if $I_1^P(t)$ increases. This implies that $M_1(t+L) - M_1(t)|I_1^P(t)$ is stochastically non-decreasing in $I_1^P(t)$ (see Shaked and Shanthikumar, 2007, Chapter 1), which implies that $P(M_1(t+L) - M_1(t) < \alpha|I_1^P(t) = i)$ is non-increasing in i for every value of α. This immediately yields the claim of Theorem 4.1. Indeed, for some i, define α_i and α_{i+1} as the smallest values for α that satisfy $P\left(M_1(t+L) - M_1(t) \leq \alpha|I_1^P(t) = i\right) \geq \frac{b+h_1}{b+h_1+h_2}$ and $P\left(M_1(t+L) - M_1(t) \leq \alpha|I_1^P(t) = i+1\right) \geq \frac{b+h_1}{b+h_1+h_2}$, respectively. It then follows that $\alpha_i \leq \alpha_{i+1}$. Consequently, we can conclude that $\tilde{IP}_2(I_1^P(t) = i) \leq \tilde{IP}_2(I_1^P(t) = i+1)$ for every i, proving the main claim.

We now turn to proving that for every possible sample path $M_1(t+L) - M_1(t)$ will either stay the same or will increase if $I_1^P(t)$ increases. We consider two cases: (i) we start with i customer orders waiting for production of module m_1 and (ii) we start with $i+1$ customer orders waiting for production of module m_1. In each case, we start of with no finished module m_1 available. Conditioning on a specific sequence of events, we show that after every event the number of finished modules m_1 in case (ii) will either stay the same as in case (i) or stay one ahead.

To formalize this, denote the following:

$$J(i) = \text{\# finished modules } m_1 \text{ when starting at } (i,0)$$
$$J(i+1) = \text{\# finished modules } m_1 \text{ when starting at } (i+1,0)$$

This means that $J(i) = M_1(t+L) - M_1(t)|I_1^P(t) = i$. We compare $J(i)$ and $J(i+1)$, so we consider starting at $(i,0)$ and $(i+1,0)$. There are two transitions, namely arrival of a new customer order to the queue and finished production of a module. This gives the following cases to consider.

1. **Arrival of customer order** gives transitions $(i, 0) \rightarrow (i+1, 0)$ and $(i+1, 0) \rightarrow (i+2, 0)$.
2. **Finished production &** $i > 0$ gives transitions $(i, 0) \rightarrow (i-1, 1)$ and $(i+1, 0) \rightarrow (i, 1)$.
3. **Finished production &** $i = 0$ gives transitions $(i, 0) = (0, 0) \rightarrow (0, 0)$ and $(i+1, 0) = (1, 0) \rightarrow (0, 1)$.

For cases 1 and 2, the difference in number of customers in the queue remains 1 and the difference between the number of finished modules remains 0. Thus, $J(i)$ and $J(i+1)$ stay equal in these two cases. For case 3, the number of customers in the queue becomes equal, but the difference in the number of finished modules increases by one and will thus always stay one ahead from now on. After these transitions, the next transition also falls within one of the three cases discussed above.

Since $J(i+1)$ either remains equal to $J(i)$ or stays one finished module ahead of $J(i)$, it follows that $J(i) \leq J(i+1)$. This means that for every number of finished modules j, $P(J(i+1) > j) \geq P(J(i) > j)$. From this we can conclude that $J(i+1)$ is stochastically larger than $J(i)$. Based on this, we know that the number of modules m_1 produced during the lead-time of module m_2 will be at least as high with $i+1$ waiting customer orders as with i waiting customer orders. This leads to our statement that $P(M_1(t+L) - M_1(t) < \alpha | I_1^P(t) = i)$ is non-increasing in i for every value of α. Therefore, the requirement for module m_2 will not decrease when an additional customer order is available and hence the target inventory position of module m_2 is non-decreasing in the number of customer orders waiting for production of module m_1, meaning that $\tilde{IP}_2(I_1^P(t) = i) \leq \tilde{IP}_2(I_1^P(t) = i+1)$. Since this holds for any sequence of events, it thus also holds for the expectation over all possible event sequences. □

Proof of Theorem 4.2

We aim to show that an additional waiting customer order will increase the target inventory position by at most 1, i.e. $\tilde{IP}_2(I_1^P(t) = i+1) - \tilde{IP}_2(I_1^P(t) = i) \leq 1$ for all $i \geq 0$. This can be proven along the same lines as Theorem 4.1, but this time we will show that $M_1(t+L) - M_1(t)|I_1^P(t) = i+1$ is stochastically smaller than $M_1(t+L) - M_1(t) + 1|I_1^P(t) = i$. To do this, we show that for every possible sample path that, for all values of i, $M_1(t+L) - M_1(t)|I_1^P(t) = i+1$ will either be smaller than or equal to $M_1(t+L) - M_1(t) + 1|I_1^P(t) = i$. Using the definitions of $J(i)$ and $J(i+1)$, it follows that $J(i) + 1$ equals the # finished modules m_1 when starting at $(i, 1)$. We thus want to show that $J(i) + 1$ is stochastically larger than $J(i+1)$.

We will consider the case of i orders in the queue and already one finished module and the case with $i+1$ customers in the queue and no finished modules. Again, there are two transitions, namely arrival of a new customer to the queue and departure of a customer from the system, corresponding to finished production of a module. This gives the following cases to consider.

1. **Arrival** gives transitions $(i, 1) \to (i+1, 1)$ and $(i+1, 0) \to (i+2, 0)$.
2. **Departure &** $i > 0$ gives transitions $(i, 1) \to (i-1, 2)$ and $(i+1, 0) \to (i, 1)$.
3. **Departure &** $i = 0$ gives transitions $(i, 1) = (0, 1) \to (0, 1)$ and $(i+1, 0) = (1, 0) \to (0, 1)$.

For cases 1 and 2, the difference in number of customers in the queue remains 1 and the difference between the number of exits remains 1. Thus, $J(i) + 1$ stays one finished module ahead of $J(i+1)$ in these two cases. In case 3, both the number of customers in the queue and the number of finished modules after the transition are equal, meaning that $J(i) + 1$ and $J(i+1)$ are equal and will follow the same trajectory when facing new arrivals or departures. After these transitions, the next transition also falls within one of the three cases discussed above.

Since $J(i) + 1$ either stays one finished module ahead of $J(i+1)$ or they become equal and follow the same trajectory, it follows that $J(i+1) \leq J(i) + 1$. This means that for every number of exits j, $P(J(i+1) > j) \leq P(J(i) + 1 > j)$ and thus that $J(i+1)$ is stochastically smaller than $J(i) + 1$. This means that the number of modules m_1 produced during the lead-time of module m_2 with $i+1$ waiting customer orders will be at most 1 larger than with i waiting customer orders.

In other words,

$$P\left(M_1(t+L) - M_1(t) + 1 \leq \alpha | I_1^P(t) = i\right)$$

$$\leq P\left(M_1(t+L) - M_1(t) \leq \alpha | I_1^P(t) = i+1\right) \quad (4.11)$$

for all α and all i. From this we can conclude that the target inventory position of module m_2 increases by at most one if the number of waiting customer order increases by one. Since this holds for any sequence of events, it thus also holds for the expectation over all possible event sequences. □

Proof of Theorem 4.3

The optimal inventory policy, denoted by π, for module m_2 is the one that minimizes average costs:

$$\min_{\pi \in \Pi} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\int_0^T \tilde{C}_\pi(t)dt\right].$$

Given the lead-time L, the inventory levels and thus costs at time t are affected by the decisions made at time $t-L$ and corresponding inventory position. Since every inventory policy for module m_2 has a corresponding inventory position, it holds that

$$\min_{\pi \in \Pi} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\int_0^T \tilde{C}_\pi(t)dt\right] \geq \lim_{T \to \infty} \frac{1}{T}\int_0^T \min_{IP_{t-L}} \mathbb{E}[\tilde{C}(t)|IP_{t-L}]dt$$

Hence, a lower bound on the costs is obtained for the policy in which the optimal inventory position is selected at every time t. This lower bound is attained by the proposed myopic inventory policy for module m_2. Therefore, it can be concluded that the proposed policy is optimal. □

Proof of Lemma 4.1

Cantelli's inequality states that for X with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, $P(X \geq r) \leq \frac{\hat{\sigma}^2}{\hat{\sigma}^2+(r-\hat{\mu})^2}$ for $r > \hat{\mu}$. Since we have $X \sim Poisson((\lambda+\mu)L)$, $\hat{\mu} = (\lambda+\mu)L$ and $\hat{\sigma}^2 = (\lambda+\mu)L$, this gives $P(X \geq r) \leq \frac{(\lambda+\mu)L}{(\lambda+\mu)L+(r-(\lambda+\mu)L)^2}$. To find r such that $P(X \geq r) \leq \epsilon$, we need to solve $\frac{(\lambda+\mu)L}{(\lambda+\mu)L+(r-(\lambda+\mu)L)^2} = \epsilon$. This yields $r = (\lambda+\mu)L + \sqrt{\left(1-\frac{1}{\epsilon}\right)(\lambda+\mu)L}$. Therefore, $P(X \geq X^U) \leq \epsilon$ holds if $X^U = (\lambda+\mu)L + \sqrt{\left(1-\frac{1}{\epsilon}\right)(\lambda+\mu)L}$. □

Proof of Theorem 4.4

Also for the case of N MTO modules, by conditioning on a specific sequence of events, we show that monotonicity holds for any possible event sequence and thus also for the expectation over all possible event sequences. This goes along the same lines as in the proof of Theorem 4.1.

Denote the following:

$$\begin{aligned} J(i_1) &= \text{\# finished modules of all modules } m_1, \ldots, m_N \\ &\quad \text{when starting at } (i_1, j_1, \ldots, j_N) \\ J(i_1+1) &= \text{\# finished modules of all modules } m_1, \ldots, m_N \\ &\quad \text{when starting at } (i_1+1, j_1, \ldots, j_N) \end{aligned}$$

This means that $J(i_1) = \min\{M_1(t+L) - M_1(t), \ldots, M_N(t+L) - M_N(t) | i_1 + 1, j_1, \ldots, j_N\}$.

We compare $J(i_1)$ and $J(i_1+1)$, so we consider starting at $(i_1, j_1, \ldots, j_N)$ and $(i_1 + 1, j_1, \ldots, j_N)$. There are three transitions, namely arrival of a new customer order to the queue, finished production of a module m_1 and finished production of a module m_k with $k \in \{2, \ldots, N\}$. This gives the following cases to consider.

1. **Arrival of customer order** gives transitions $(i_1, j_1, \ldots, j_N) \to (i_1+1, j_1, \ldots, j_N)$ and $(i_1+1, j_1, \ldots, j_N) \to (i_1+2, j_1, \ldots, j_N)$.

2. **Finished production of module** m_1 **&** $i_1 > 0$ gives transitions $(i_1, j_1, \ldots, j_N) \to (i_1 - 1, j_1 + 1, \ldots, j_N)$ and $(i_1+1, j_1, \ldots, j_N) \to (i_1, j_1+1, \ldots, j_N)$.

3. **Finished production of module** m_1 **&** $i_1 = 0$ gives transitions $(i_1, j_1, \ldots, j_N) \to (i_1, j_1, \ldots, j_N)$ and $(i_1+1, j_1, \ldots, j_N) \to (i_1, j_1+1, \ldots, j_N)$.

4. **Finished production of module** m_k **for** $k \in \{2, \ldots, N\}$ **&** $i_k > 0$ gives transitions $(i_1, j_1, \ldots, j_N) \to (i_1, j_1, \ldots, j_k + 1, \ldots, j_N)$ and $(i_1+1, j_1, \ldots, j_N) \to (i_1+1, j_1, \ldots, j_k+1, \ldots, j_N)$.

5. **Finished production of module** m_k **for** $k \in \{2, \ldots, N\}$ **&** $i_k = 0$ gives transitions $(i_1, j_1, \ldots, j_N) \to (i_1, j_1, \ldots, j_k, \ldots, j_N)$ and $(i_1+1, j_1, \ldots, j_N) \to (i_1+1, j_1, \ldots, j_k+1, \ldots, j_N)$.

For cases 1, 2 and 4, the difference in number of customers in the queues remains 1 and the difference between the number of finished modules remains 0. Thus, $J(i_1)$ and $J(i_1+1)$ stay equal in these two cases. For case 3, the number of customers in the queue of module m_1 becomes equal, but the difference in the number of finished modules of m_1 increases by one and will thus always stay one ahead from now on. For case 5, the number of customers in the queue of module m_k becomes equal, but the difference in the number of finished modules of m_k increases by one and will thus always stay one ahead from now on. After these transitions, the next transition also falls within one of the five cases discussed above.

Since $J(i_1+1)$ either remains equal to $J(i_1)$ or stays one finished module ahead of $J(i_1)$, it follows that $J(i_1) \leq J(i_1+1)$. This means that for every number of finished modules j, $P(J(i_1+1) > j) \geq P(J(i_1) > j)$. From this we can conclude that $J(i_1+1)$ is stochastically larger than $J(i_1)$. The result follows by the same reasoning as in the proof of Theorem 4.1. □

Proof of Theorem 4.5

Along the same lines as for Theorem 4.2, we show that this maximum increase in the target inventory position of the lead-time module holds for any possible event sequence and thus also for the expectation over all possible event sequences. Using the definitions of $J(i_1)$ and $J(i_1+1)$, it follows that $J(i_1)+1$ equals the # finished modules of all modules $m_1, \ldots, m_N$ when starting with one finished module of each of them. We will compare $J(i_1+1)$ and $J(i_1)+1$, so we consider starting at $(i_1+1, j_1, \ldots, j_N)$ and $(i_1, j_1+1, \ldots, j_N+1)$. There are again three transitions, namely arrival of a new customer order to the queue, finished production of a module m_1 and finished production of a module m_k for $k \in \{2, \ldots, N\}$. This gives the following cases to consider.

1. **Arrival of customer order** gives transitions $(i_1+1, j_1, \ldots, j_N) \rightarrow (i_1+2, j_1, \ldots, j_N)$ and $(i_1, j_1+1, \ldots, j_N+1) \rightarrow (i_1+1, j_1+1, \ldots, j_N+1)$.

2. **Finished production of module** m_1 **&** $i > 0$ gives transitions $(i_1+1, j_1, \ldots, j_N) \rightarrow (i_1, j_1+1, \ldots, j_N)$ and $(i_1, j_1+1, \ldots, j_N+1) \rightarrow (i_1-1, j_1+2, \ldots, j_N+1)$.

3. **Finished production of module** m_1 **&** $i = 0$ gives transitions $(i_1+1, j_1, \ldots, j_N) \rightarrow (i_1, j_1+1, \ldots, j_N)$ and $(i_1, j_1+1, \ldots, j_N+1) \rightarrow (i_1, j_1+1, \ldots, j_N+1)$.

4. **Finished production of module m_k for $k \in \{2, \ldots, N\}$ & $i > 0$** gives transitions $(i_1 + 1, j_1, \ldots, j_N) \rightarrow (i_1 + 1, j_1, \ldots, j_k + 1 \ldots, j_N)$ and $(i_1, j_1 + 1, \ldots, j_N + 1) \rightarrow (i_1, j_1 + 1, \ldots, j_k + 2, \ldots, j_N + 1)$.
5. **Finished production of module m_k for $k \in \{2, \ldots, N\}$ & $i = 0$** gives transitions $(i_1 + 1, j_1, \ldots, j_N) \rightarrow (i_1 + 1, j_1, \ldots, j_k + 1 \ldots, j_N)$ and $(i_1, j_1 + 1, \ldots, j_N + 1) \rightarrow (i_1, j_1 + 1, \ldots, j_N + 1)$.

For cases 1, 2 and 4, the difference in number of customers in the queues remains 1 and the difference between the number of finished modules remains 0. Thus, $J(i_1 + 1)$ and $J(i_1) + 1$ stay equal in these two cases. For case 3 (5), the number of customers in the queue of module m_1 (m_k for $k \in \{2, \ldots, N\}$) and the number of finished modules of m_1 (m_k for $k \in \{2, \ldots, N\}$) become equal and will remain the same from now on, so $J(i_1 + 1)$ and $J(i_1) + 1$ become equal. For module m_k for $k \in \{2, \ldots, N\}$ (m_1), the number of finished components will stay one ahead, meaning that there is already one order less in the queue, meaning that total output can increase by at most 1 when starting with one more finished combination. After these transitions, the next transition also falls within one of the five cases discussed above.

Since $J(i_1) + 1$ either remains equal to $J(i_1 + 1)$ or stays one finished combination of modules ahead of $J(i_1 + 1)$, it follows that $J(i_1 + 1) \leq J(i_1) + 1$. This means that for every number of finished modules j, $P(J(i_1) + 1 > j) \geq P(J(i_1 + 1) > j)$. From this we can conclude that $J(i_1 + 1)$ is stochastically smaller than $J(i_1) + 1$ and the result follows from the same reasoning as in the proof of Theorem 4.2. □

Proof of Theorem 4.6

The proof is analogous to that of Theorem 4.3.

Proof of Theorem 4.7

The proof is along the same lines as for Theorem 4.1. We again compare $J(i)$ and $J(i + 1)$. When we are in state (i, j), meaning that there are i product orders in the system and j units produced so far, and demand D occurs, there are $i + D$ units to be produced. Since the available capacity is C units, production is equal to $\min\{i + D, C\}$ and the remaining number of back-orders equals $(i + D - C)^+$.

By conditioning on demand $D = d$ and capacity $C = c$, the following transition occurs:

$$(i,j) \to \big((i+d-c)^+, j+\min\{i+d,c\}\big) \, .$$

We consider the following cases:

1. $i+d<c$ (so $i+1+d \leq c$): in this case we have transitions

   $$\begin{aligned}(i,j) &\to (0, j+i+d) \text{ and}\\ (i+1,j) &\to (0, j+i+d+1)\end{aligned}$$

   Since all product orders can be satisfied, the number of outstanding product orders reduces to zero for both (i,j) and $(i+1,j)$. Therefore, from now on production in any period will be the same and $J(i+1)$ will stay one unit ahead of $J(i)$.

2. $i+d \geq c$: gives transitions

   $$\begin{aligned}(i,j) &\to (i+d-c, j+c) \text{ and}\\ (i+1,j) &\to (i+1+d-c, j+c)\end{aligned}$$

   In this case the difference in the number of outstanding order stays 1 and the number of produced items remains equal. Hence, the following period starts of with the same situation as the current period.

Since $J(i+1)$ either stays equal to $J(i)$ or stays one finished module ahead of $J(i)$, it follows that $J(i) \leq J(i+1)$. This means that for every number of finished modules j, $P(J(i+1) > j) \geq P(J(i) > j)$. From this we can conclude that $J(i+1)$ is stochastically larger than $J(i)$ and thus that the target inventory position of module m_2 when $i+1$ customer orders are in the system is equal to or larger than in case i customer orders are in the system.

By conditioning on a specific occurrences of D and C, we have shown, following the same reasoning as in the proof of Theorem 4.1, that for any possible event sequence the target inventory position is monotonically non-decreasing in the number of orders. Therefore, this also holds for the expectation over all possible event sequences. □

Proof of Theorem 4.8

The proof is along the same lines as for Theorem 4.2. We again compare $J(i+1)$ and $J(i)+1$.

We again condition on demand $D = d$ and capacity $C = c$ and consider the following cases:

1. $i + d < c$ (so $i + 1 + d \leq c$): in this case we have transitions

$$\begin{aligned} (i, j+1) &\rightarrow (0, j+1+i+d) \text{ and} \\ (i+1, j) &\rightarrow (0, j+i+d+1) \end{aligned}$$

   The number of outstanding product orders again reduces to zero for both $(i, j+1)$ and $(i+1, j)$. Also, the number of produced items becomes equal. Therefore, from now on production in any period will be the same and $J(i+1)$ will remain the same as $J(i)$.

2. $i + d \geq c$: gives transitions

$$\begin{aligned} (i, j+1) &\rightarrow (i+d-c, j+1+c) \text{ and} \\ (i+1, j) &\rightarrow (i+1+d-c, j+c) \end{aligned}$$

   In this case the difference in the number of outstanding order and the difference in the number of produced items both remain equal. Hence, the following period starts of with the same situation as the current period.

After these transitions, the next transition also falls within one of the two cases discussed above. Since $J(i)+1$ either stays one finished module ahead of $J(i+1)$ or they become equal and follow the same trajectory, it follows that $J(i+1) \leq J(i)+1$. This means that for every number of exits j, $P(J(i+1) > j) \leq P(J(i)+1 > j)$ and thus that $J(i+1)$ is stochastically smaller than $J(i)+1$.

From this we can conclude, by the same reasoning as in the proof of Theorem 4.2, that the number of finished products when starting with $i+1$ orders in the queue is at most one larger than in case there are i orders in the queue. Therefore, the target inventory position of module m_2 is also at most one higher. □

Proof of Theorem 4.9

The optimal inventory policy, denoted by π, for module m_2 is the one that minimizes total costs:

$$\min_{\pi \in \Pi} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{0}^{T} \tilde{C}_{\pi}(t)\right]$$

where T denotes the planning horizon.

Given the lead-time of L periods, the inventory levels and thus costs in period t are affected by the decisions made in period $t - L$ and corresponding inventory position. Since every inventory policy for module m_2 has a corresponding inventory position, it holds that

$$\min_{\pi \in \Pi} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{0}^{T} \tilde{C}_{\pi}(t)\right] \geq \lim_{T \to \infty} \frac{1}{T}\sum_{0}^{T} \min_{IP_{t-L}} \mathbb{E}[\tilde{C}(t)|IP_{t-L}].$$

Hence, a lower bound on the costs is obtained for the policy in which the optimal inventory position is selected in every period t. This lower bound is attained by the proposed myopic inventory policy for module m_2. Therefore, it can be concluded that the proposed policy is optimal. □

Proof of Lemma 4.2

Consider a customer order. Suppose that at time t it is possible to start producing the MTO module for this customer order, and suppose that instead of starting at t, we postpone the production by z and start production at $t + z$. As a consequence, production will be in expectation finished z time units later, resulting in maximum savings of $h_1 \cdot z$. Since the order arrived before t, and since we assume a base-stock policy with positive base-stock level, the lead-time module that will be used to satisfy the order will arrive at or before $t + L$. Any delay in the completion of the MTO module that occurs after $t + L$ is thus guaranteed to cause additional waiting time for the final product. Such a delay happens with probability $\mathbb{P}(\text{production time} > L - z) = e^{-\mu(L-z)}$, in which case the expected time until production is finished equals $\frac{1}{\mu}$. When this happens, holding costs h_2 for module m_2 and customer waiting costs b are incurred. This means that an upper bound on the expected savings is given by $h_1 z - \frac{b+h2}{\mu} e^{-\mu(L-z)}$. Using the first-order condition, the optimal postponement equals $z = L - \frac{1}{\mu} \ln\left(\frac{b+h_2}{h_1}\right)$. From this it follows that it is optimal to not postpone production of the MTO module when $e^{\mu L} < \frac{b+h_2}{h_1}$. □

5

Extreme-value Theory for Large Fork-join Queues, with Applications to High-tech Supply Chains

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

We explore repercussions of this result for original equipment manufacturers (OEMs) that assemble a large number of components, each produced using specialized equipment, into complex systems. Component production capacity is subject to fluctuations, causing high risk of shortages of at least one component, which results in costly system production delays. OEMs hedge this risk by investing in a combination of excess production capacity and component inventories. We formulate a stylized model of the OEM that enables us to study the resulting trade-

This chapter is based on Meijer et al. (2021a).

off between shortage risk, inventory costs, and capacity costs. Our asymptotic extreme value results translate into asymptotically exact methods for cost-optimal inventory and capacity decisions, some of which are in closed form. We validate our asymptotic results with a set of detailed numerical experiments. These experiments indicate that our results are asymptotically exact, while for transient times the accuracy of the asymptotic approximations depends on model parameters.

5.1. Introduction

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

To this end, we consider a fork-join network of N statistically identical queues driven by a common arrival process and having independent service processes. All arrival and service processes consist of a deterministic term, perturbed by (independent) Brownian motions. We are interested in the behavior of the maximal queue length in steady state as the number of queues grows large. We examine separately the cases of purely deterministic arrivals and of perturbed arrivals. Our asymptotic results provide insight into the performance of large fork-join networks. The proof techniques we use are quite generic. For deterministic arrivals, we use standard extreme value theory, while for correlated arrivals, we rely on sample path analysis and conditional limit theorems for large suprema of Brownian motions.

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

When the arrival process is deterministic plus a random perturbation following a Brownian motion, the stationary queue lengths are still exponentially distributed, but no longer independent. The question is now how this affects the maximum queue length as the number of queues $N \to \infty$. Most of the work in extreme value theory has been done for independent random variables; cf. De Haan and Ferreira (2006); Resnick (1987). It turns out that suitable results from extreme value theory are absent for our setting. Thus, deriving a convergence result for the maximum queue length for perturbed arrivals as $N \to \infty$ is one of the key technical challenges underlying this chapter. Our answer to this challenge is somewhat surprising: the dependence structure causes the scaled maximum queue length to converge to a normally distributed random variable as $N \to \infty$. That this scaled maximum is in the domain of attraction of the normal distribution is remarkable since for independent random variables, such a scaled maximum can only converge to a Gumbel, a Weibull or a Fréchet distributed random variable. Thus, our result shows that the normal distribution has a non-empty domain of attraction in an extreme-value theory context. An intuitive explanation of this fact, based on asymptotic independence of hitting times, is provided in Section 5.5.

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

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

To appropriately model fluctuations in production capacity in continuous time, the cumulative production of each component is modelled as a Brownian motion with drift. This is a natural extension of normally distributed production capacity in discrete time, which is a common choice in the literature (e.g. Bradley and Glynn, 2002; Wu and Chao, 2014). OEMs typically *level* the demand to smooth the production process. Accordingly, in our base model we assume that demand is completely levelled/deterministic. For this base model, in Section 5.4 we derive easy to calculate expressions for capacity and base-stock levels that are asymptotically optimal as the number of components grows large. We provide order bounds between the costs under optimal and approximate base-stock level and capacity.

In particular, inspired by the literature on call centers: Borst et al. (2004); Gans et al. (2003) and van Leeuwaarden et al. (2019), we distinguish three regimes, which depend on the growth rates of cost parameters and are determined by the probability γ_N of not having enough inventory. Given that $\gamma_N \to \gamma$, 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.

Despite efforts to level demand, typically some demand variation remains. Therefore in Section 5.5, we assume that the stochastic demand for systems is modelled by a Brownian motion. This implies that the demand over any finite time period is a normal variable, which is a standard assumption in literature (e.g. Klosterhalfen et al., 2014; Atan and Rousseau, 2016). As a consequence, component delays become *dependent*, since they face the same stochastic demands from system assembly. Our main technical result for dependent Brownian motions implies that, with proper scaling of holding and backorder costs, the optimal base-stock level for stochastic demand converges to a scaled version of the quantile function of the normal distribution, while this quantile function also appears in the limit of the optimal capacity. Numerical experiments show that we typically are around 10% off the optimum (e.g. when N is in the range from 10 to 100); cf. Tables 5.4 and 5.5. Naturally, the difference goes to 0 as $N \to \infty$; cf. Theorem 5.3.

We give an improvement of this approximation by combining our results for deterministic demand and stochastic demand. Based on this approximation, we optimize the capacity and base-stock levels 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 typically result in costs that are less than 2% off the optimum.

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

The remainder of this chapter is organized as follows. In Section 5.2, we provide an overview of relevant literature. The content of the chapter is then structured around the application to high-tech assembly systems, with theoretical results appearing as we need them. In particular, we introduce the general mathematical model in Section 5.3 and subsequently present the optimization problem where we need to decide on capacity and base-stock levels to minimize costs. We study the assembly system with deterministic demand in Section 5.4. We provide explicit expressions and approximations for optimal base-stock levels and capacity. The stochastic demand case, with solutions to the minimization problem and convergence results, is studied in more detail in Section 5.5. That section also includes our key result on the extremal behavior of dependent Brownian fork-join queues, given in Theorem 5.2. A refinement of the approximations from Section 5.5 is provided in Section 5.6, where we combine the lessons learnt in Sections 5.4 and 5.5 to obtain better approximations for optimal capacity and base-stock levels. We give a summary and conclusions in Section 5.7 and provide most of the proofs in Appendix 5.A.

5.2. Literature review

In this chapter, we examine fork-join queueing networks with N servers where the arrival and service streams are almost deterministic with a Brownian component. Our goal is to find and investigate the maximum queue length as N goes to infinity. The queue lengths are dependent random variables due to the joint interarrivals. Thus, our work 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 only look at the largest of the N all-time suprema, where $N \to \infty$.

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 et al. (2016) give the solution of the Laplace transform of joint first passage times in terms of the solution of a partial differential equation, where the Brownian motions are dependent. Debicki et al. (2020) analyze the tail asymptotics of the all-time suprema of two dependent Brownian motions. The joint suprema of a finite number of Brownian motions is also studied; cf. Debicki et al. (2015), where the authors give tail asymptotics of the joint suprema of independent Gaussian processes over a finite time interval. These are just three examples, but the literature is rich with variations around assumptions on independence and dependence or around whether or not drift terms are linear, with joint suprema of two or more than two processes, with suprema over finite and infinite time intervals, and with extensions to other Gaussian processes. In this chapter, we specifically examine the maximum of N all-time suprema of dependent Brownian motions. In this respect, the work of Brown and Resnick (1977) comes the closest to our work. In that paper, the authors study process convergence of the scaled maximum of N independent Brownian motions to a stationary limiting process whose marginals are Gumbel distributed. However, we add to this by considering the maximum of the all-time suprema of N dependent Brownian motions.

Our work also relates to the literature on fork-join queues. Specifically, we study asymptotic results for a fork-join queueing system with N servers. Most exact results on fork-join queues are limited to systems with two service stations; cf. Flatto and Hahn (1984), Wright (1992), Baccelli (1985) and Klein (1988). For fork-join queues with more than two servers only approximations of performance measures are given; cf. Ko and Serfozo (2004), Baccelli and Makowski (1989) and Nelson and Tantawi (1988). Most of these papers focus on fork-join queueing systems where the number of servers is finite, while we investigate a fork-join queue where N goes to infinity. Furthermore, in these papers, the focus lies on steady-state distributions and other one-dimensional performance measures. Work on the heavy-traffic process limit has also been done. For example, Varma (1990) derives a heavy-traffic analysis for fork-join queues, and shows weak convergence of several processes, such as the joint queue lengths in front of each server. Furthermore, Nguyen (1993) proves that various appearing limiting processes are in fact multi-dimensional reflected Brownian motions. Nguyen (1994) extends this result to a fork-join queue with multiple job types. Lu and Pang study fork-join networks in Lu and Pang (2015, 2017a,b). In Lu and Pang (2015), they investigate a fork-join network where each service station has multiple servers under nonexchangeable synchronization and operates in the quality-driven regime. They derive functional central limit theorems for the number of tasks waiting in the waiting buffers for synchronization and for the number of synchronized jobs. In Lu and Pang (2017a), they extend this analysis to a fork-join network with a fixed number of service stations, each having many servers, where the system operates in the Halfin-Whitt regime. In Lu and Pang (2017b), the authors investigate these heavy-traffic limits for a fixed number of infinite-server stations, where services are dependent and could be disrupted. Finally, we mention Atar et al. (2012), who investigate the control of a fork-join queue in heavy traffic by using feedback procedures.

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

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

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

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

5.3. Problem formulation

We consider a manufacturing system in which a manufacturer assembles a final product from N components, each of which is produced on a single production line, where N is a large number. Random delays may occur in the production process for each of the components. To efficiently satisfy demand of the end-product, which may either be deterministic or stochastic, we need to decide how much capacity to establish for each component and how many finished components to keep on inventory as a buffer. Even though it is costly to establish capacity and to hold inventory, not being able to satisfy demand gives rise to backorder costs. Therefore, we need to find capacity and inventory levels that minimize total expected costs.

To formulate the cost-minimization problem, we model this assembly system by a fork-join queue. Demand is represented by the arrival stream of jobs going to each server and each server represents a component production line. The backlog of each component is represented by a queue of jobs that have not been served yet. After completion of a job, the finished component is stored in a warehouse. When all servers have a finished component in their warehouse, the end-product can be assembled. This system is visualized in Figure 5.1.

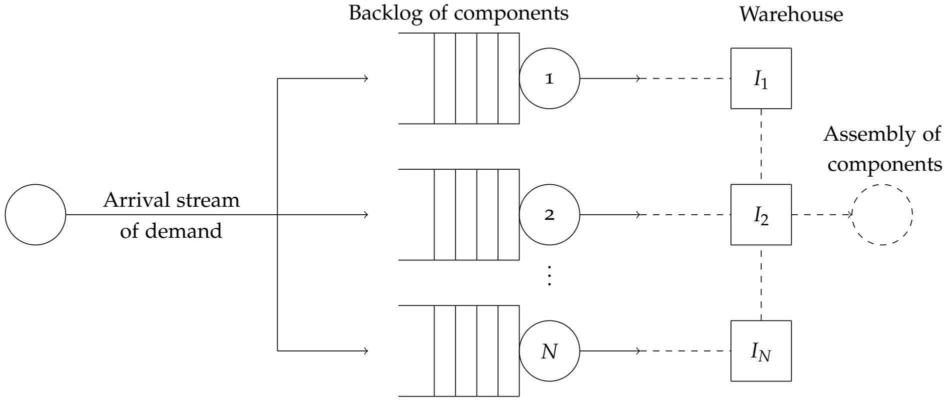

Figure 5.1: Fork-join queue

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

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

$$\begin{aligned}(I_i - Q_i(\beta_i))^+ + \max_{i\le N}\left(Q_i(\beta_i)-I_i\right)^+ - (Q_i(\beta_i)-I_i)^+ \\ = I_i - Q_i(\beta_i) + \max_{i\le N}\left(Q_i(\beta_i)-I_i\right)^+ .\end{aligned}$$

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

$$\begin{aligned}C_N(I,\beta) =& \mathbb{E}\left[\sum_{i\le N}\left[h^{(N)}\left(I - Q_i(\beta) + \left(\max_{i\le N} Q_i(\beta) - I\right)^+\right)\right]\right. \\ &\left. + b^{(N)}\left(\max_{i\le N} Q_i(\beta) - I\right)^+\right] \\ =& \mathbb{E}\left[Nh^{(N)}(I - Q_i(\beta)) + (Nh^{(N)} + b^{(N)})\left(\max_{i\le N} Q_i(\beta) - I\right)^+\right], \qquad (5.1)\end{aligned}$$

then, if $\beta_i = \beta$ and $I_i = I$ for given I and β, the expected total costs in the system are equal to $C_N(I,\beta) + \beta N$. In the centralized optimization problem, this expression is minimized with respect to I and β. In Appendix 5.A.1, we show that it suffices to consider symmetric solutions where both I_i and β_i are constant in i when $(Q_i(\beta_i))_{i\le N}$ are independent random variables or when we minimize over one drift parameter. For these two cases, we utilize the self-similarity property of Brownian motions, which allows us to simplify $C_N(I,\beta)$. Due to the self-similarity of Brownian motion, we can write

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

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

$$\min_{(I,\beta)}\left(C_N(I,\beta)+\beta N\right)=\min_{(I,\beta)}\left(\frac{1}{\beta}C_N(I\beta,1)+\beta N\right)$$
$$=\min_{(I,\beta)}\left(\frac{1}{\beta}C_N(I,1)+\beta N\right). \quad (5.2)$$

In the last part of Equation (5.2), I has the interpretation of the base-stock level where the net capacity $\beta = 1$. Therefore, from now on, the actual number of products on stock at time 0 equals I/β. 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 5.1.

Definition 5.1 We define

$$F_N(I,\beta) := \frac{1}{\beta}C_N(I)+\beta N, \quad (5.3)$$

with $C_N(I) := C_N(I,1)$ and $C_N(I,\beta)$ given in Equation (5.1).

Our goal is to solve $\min_{(I,\beta)} F_N(I,\beta)$, focusing on the case where N is large. Before we focus on this regime, we first derive some additional properties of this problem, which are valid for each N. In the next lemma, we show that we can write this minimization problem as two separate minimization problems. The proofs of this section can be found in Appendix 5.A.1.

Lemma 5.1 *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 capacity 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 5.1, we can characterize the optimal net capacity and base-stock level. In Lemma 5.2 we provide expressions for the optimal net capacity and costs in terms of the optimal base-stock level, which is given in Lemma 5.3.

Lemma 5.2 *Given $I_N^* = \arg\min_I C_N(I)$, minimizing $F_N(I,\beta)$ with respect to β yields $\beta_N^* = \sqrt{\frac{C_N(I_N^*)}{N}}$. Furthermore, the corresponding costs are $F_N(I_N^*,\beta_N^*) = 2N\beta_N^* = 2\sqrt{C_N(I_N^*)N}$.*

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

Lemma 5.3 *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 5.4 *$\max_{i\leq N} Q_i$ 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 chapter.

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

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

The optimal value $\bar{I}_N$ for the associated first-order minimization problem $\min_I \bar{C}_N(I)$

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

- We are in the *balanced regime* if $\gamma \in (0,1)$.
- If $\gamma = 0$, for large systems, the base-stock level is always sufficiently high to ensure that the manufacturer can assemble the end-product. We call this the *quality-driven regime.*
- Finally, if $\gamma = 1$, base-stock levels 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 5.5:

Lemma 5.5 *Assume* $\gamma_N = Nh^{(N)}/(Nh^{(N)}+b^{(N)})$, *with* $\gamma_N \overset{N\to\infty}{\longrightarrow} \gamma \in (0,1)$ *or* $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$. *Then*

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

In the next two sections, we carry out a more elaborate program using more refined extreme value estimates of $\max_{i\le N} Q_i$. This analysis gives sharper order bounds than those given in Lemma 5.5. In particular, in the following sections we consider the minimization in two distinct cases. First, in Section 5.4, we look at the case where demand is assumed to be deterministic, such that $W_A = 0$. Thereafter, in Section 5.5, we consider the stochastic demand case. In the former case, we utilize existing results in extreme value theory, while the latter case requires the development of a novel limit theorem. Furthermore, we use the result given in Corollary 5.1; 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 a solution to a minimization problem as well. This corollary follows trivially from Lemma 5.2.

Corollary 5.1 *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, then

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

5.4. The basic model: deterministic arrival stream

5.4.1 Solution and convergence of the minimization problem

We now analyze the minimization of the cost function described in Definition 5.1 for the special case with $W_A = 0$ representing deterministic demand. Although we can simplify the minimization problem significantly, by using the self-similarity of Brownian motions and by writing the minimization problem as two separate minimization problems as shown in Lemma 5.1, the function F_N still has a difficult form, since we have the expression $\max_{i \leq N} Q_i$ in this function. In Lemma 5.6 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 $N > 0$. Secondly, we consider the quality driven regime, where $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$. Finally, we investigate the efficiency driven regime, where $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$. All proofs for this section can be found in Appendix 5.A.2. We present numerical results for the three regimes in Section 5.4.2.

Lemma 5.6 *Let $Q_i = \sup_{s>0}(W_i(s) - s)$, with $(W_i, 1 \leq i \leq N)$ independent Brownian motions with mean 0 and variance σ^2. Let $h^{(N)}$ and $b^{(N)}$ be positive sequences. In order to minimize $F_N(I, \beta)$, the optimal base-stock level I^*_N satisfies,*

$$I^*_N = P_N^{-1}(1 - \gamma_N) = \frac{\sigma^2}{2} \log\left(\frac{1}{1 - (1 - \gamma_N)^{\frac{1}{N}}}\right), \tag{5.4}$$

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

To get a better understanding of the limiting behavior of the solution to $\min_{(I,\beta)} F_N(I, \beta)$, we would like to approximate the function F_N. Since $(Q_i, i \leq N)$ are independent and exponentially distributed, we know by standard extreme value theory (cf. De Haan and Ferreira (2006)) that $\frac{2}{\sigma^2} \max_{i \leq N} Q_i - \log N \xrightarrow{d} G$, 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 5.2 we give the resulting function $\hat{F}_N(I, \beta)$ that is to be minimized.

Definition 5.2

$$\hat{C}_N(I) := \mathbb{E}\left[Nh^{(N)}(I - Q_i) + \left(Nh^{(N)} + b^{(N)}\right)\left(\frac{\sigma^2}{2} G + \frac{\sigma^2}{2} \log N - I\right)^+\right], \tag{5.5}$$

and

$$\hat{F}_N(I, \beta) := \frac{1}{\beta} \hat{C}_N(I) + \beta N. \tag{5.6}$$

In the remainder of this section, we investigate whether the capacity and base-stock level minimizing $\hat{F}_N(I, \beta)$ result in costs that are close to those when we minimize $F_N(I, \beta)$. Note that we write (I^*_N, β^*_N) for the minimizers for the cost function F_N defined in Definition 5.1, and we write $(\hat{I}_N, \hat{\beta}_N)$ for the minimizers for the cost function $\hat{F}_N$ defined in Definition 5.2. Thus, throughout this chapter, we indicate second-order approximations by the $\wedge$-symbol.

In Proposition 5.1, 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 5.1 (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*

$$\hat{I}_N = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log\left(-\log\left(1-\gamma_N\right)\right), \tag{5.7}$$

and

$$\hat{C}_N(\hat{I}_N) = Nh^{(N)}\left(\hat{I}_N - \frac{\sigma^2}{2}\right) + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2} \\ \left(\int_{-\log(1-\gamma_N)}^{\infty} \frac{e^{-t}}{t}dt + \Gamma + \log\left(-\log\left(1-\gamma_N\right)\right)\right), \tag{5.8}$$

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

Combining Equations (5.7) and (5.8) with the results in Lemma 5.2 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 regime, quality driven regime and efficiency driven regime. We first present two lemmas that are needed to prove order bounds between the costs under the optimal base-stock level and net capacity, and the costs under the approximate base-stock level and net capacity. In Lemma 5.7 we show that we can define a random variable that follows a Gumbel distribution, and is on the same probability space as $\max_{i\leq N} Q_i$. 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 Lemma 5.8, which leads to sharp results in Theorem 5.1 and Lemma 5.9.

Lemma 5.7 *Define*

$$G_N := -\log\left(-\log\left(\left(1-\exp\left(-\frac{2}{\sigma^2}\max_{i\leq N} Q_i\right)\right)^N\right)\right), \tag{5.9}$$

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

$$\max_{i\leq N} Q_i > \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N, \tag{5.10}$$

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

Lemma 5.8 *Let* $\gamma_N = Nh^{(N)}/(Nh^{(N)}+b^{(N)})$*, then*

$$\left|C_N(I_N^*) - C_N(\hat{I}_N)\right| \leq \\ (I_N^* - \hat{I}_N)(Nh^{(N)}+b^{(N)})\left(1-\gamma_N-\left(1+\frac{\log(1-\gamma_N)}{N}\right)^N\right), \quad (5.11)$$

$$\left|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)\right| \leq (I_N^* - \hat{I}_N)Nh^{(N)}\left(1-\left(1+\frac{\log(1-\gamma_N)}{N}\right)^N\right). \quad (5.12)$$

Finally, by using the results from Lemmas 5.7 and 5.8, we prove the order bounds in the balanced, quality driven and efficiency driven regime in Theorem 5.1. In the efficiency driven regime, we impose the additional condition $\gamma_N < 1-\exp(-N)$ needed to make sure that $\hat{I}_N > 0$.

Theorem 5.1 (Order bounds) *Assume* $\gamma_N = Nh^{(N)}/(Nh^{(N)}+b^{(N)})$*, if* $\gamma_N = \gamma \in (0,1)$*, in the balanced regime, then*

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(1/(N\log N)), \quad (5.13)$$

if $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$*, in the quality driven regime, then*

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(\gamma_N/(N\log(N/\gamma_N))), \quad (5.14)$$

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

$$\frac{F_N(I_N^*,\beta_N^*)}{F_N(\hat{I}_N,\hat{\beta}_N)} = 1 - O(1/\log N). \quad (5.15)$$

Using the order bounds given in Theorem 5.1, we can establish for the three different regimes how $F_N(I_N^*,\beta_N^*)$ scales with N as N becomes large.

Lemma 5.9 *Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, if $\gamma_N = \gamma \in (0,1)$ in the balanced regime, then*

$$\begin{aligned} F_N(I_N^*, \beta_N^*) =& 2\sqrt{N}\left(Nh^{(N)}\frac{\sigma^2}{2}(\log N - \log(-\log(1-\gamma)) - 1)\right. \\ & \left. +(Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\mathbb{E}\Big[(G + \log(-\log(1-\gamma)))^+\Big]\right)^{\frac{1}{2}} \\ & + O(\sqrt{h^{(N)}}/\sqrt{\log N}), \end{aligned} \tag{5.16}$$

if $\gamma_N \overset{N\to\infty}{\longrightarrow} 0$ in the quality driven regime, then

$$\begin{aligned} F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh^{(N)}\frac{\sigma^2}{2}(\log(N/\gamma_N) - 1) + (Nh^{(N)} + b^{(N)})\frac{\sigma^2}{2}\gamma_N} \\ + O(\gamma_N\sqrt{h^{(N)}}/\sqrt{\log(N/\gamma_N)}), \end{aligned} \tag{5.17}$$

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

$$\begin{aligned} F_N(I_N^*, \beta_N^*) = 2\sqrt{N}\sqrt{Nh^{(N)}\frac{\sigma^2}{2}(\log N - 1) + b^{(N)}\frac{\sigma^2}{2}\log(-\log(1-\gamma_N))} \\ + O(N\sqrt{h^{(N)}}/\sqrt{\log N}). \end{aligned} \tag{5.18}$$

5.4.2 Numerical experiments

We now provide some numerical results to illustrate the solutions to the minimization problem and their characteristics discussed in Section 5.4.1. In all experiments, we let $\sigma = 1$ and let N vary from 10 to 1000. The results for the balanced regime, quality driven regime and efficiency driven regime are given in Tables 5.1, 5.2 and 5.3, respectively. We can observe that in all regimes the approximate solutions are close to the optimal solutions. Most importantly, already for small N, the fraction of the costs corresponding to the optimal solution over the costs corresponding to the approximate solution nearly equals 1.

Table 5.1: Balanced Regime, $h^{(N)} = 1$, $b^{(N)} = N$ such that $\gamma_N = \frac{1}{2}$.

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

Table 5.2: Quality Driven Regime, $h^{(N)} = 1$, $b^{(N)} = N^2$ such that $\gamma_N = \frac{1}{1+N}$.

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

Table 5.3: Efficiency Driven Regime, $h^{(N)} = N$, $b^{(N)} = 1$ such that $\gamma_N = \frac{N^2}{N^2+1}$.

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

5.5. Stochastic demand

We now extend our framework to the case where demand is stochastic. This means that stochasticity not only arises from the production process of the individual components, but also results from uncertain demands. Consequently, delays may no longer only be caused by low production of a specific component, but may also occur when there is a sudden peak in demand. Since all components need to be available to assemble the end-product and satisfy demand, delays of the different components are now correlated. We use the same strategy when demand is stochastic as in the basic model with deterministic demand. However, we can no longer approximate the maximum queue length distribution with the Gumbel distribution. In Section 5.5.1 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 5.3 with respect to I and β.

Definition 5.3

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

and

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

In Section 5.5.2 we elaborate on the solution and convergence of the minimization problem.

5.5.1 Extreme value limit

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

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

$$\frac{\max_{i\le N}\sup_{s>0}\left(W_i(s)+W_A(s)-\beta s\right)-\frac{\sigma^2}{2\beta}\log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma\sigma_A}{\sqrt{2}\beta}X, \tag{5.19}$$

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

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

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

A heuristic explanation of the result in Theorem 5.2 is as follows: though $(Q_i, i \le N)$ are dependent random variables, since we are adding the same Brownian motion W_A, $\max_{i\le N} W_i(s)$ will dominate more and more over W_A as N becomes larger. Consequently, W_A does not affect the time at which the supremum of $\max_{i\le N} W_i(s) + W_A(s) - \beta s$ is attained. Hence, for N large $\max_{i\le N} Q_i(\beta) \approx \max_{i\le N}\sup_{s>0}(W_i(s)-\beta s) + W_A(\tau)$, with τ the hitting time of the supremum of $\max_{i\le N}(W_i(s)-\beta s)$. Based on theory on conditional expectations of Lévy processes we know that the conditional expectation of the hitting time $\tau(x)$ to reach a point x is linear with x, to be precise, for $N=1$ it is known that $\mathbb{E}[\tau(x) \mid \tau(x) < \infty] = x/\beta$. Combining this with the fact that $\max_{i\le N}\sup_{s>0}(W_i(s)-\beta s) \sim \frac{\sigma^2}{2\beta}\log N$, we expect that the supremum of $\max_{i\le N}(W_i(s)-\beta s)$ is reached at $\tau \approx \frac{1}{\beta}\frac{\sigma^2}{2\beta}\log N = \frac{\sigma^2}{2\beta^2}\log N$. 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 (5.19).

The proof of Theorem 5.2 consists of four parts, which are stated in Lemmas 5.10, 5.11, 5.12 and 5.13 for which the proofs are provided in Appendix 5.A.3. For a process X we have for all $t > 0$ that

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

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

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

We prove that these lower and upper bounds are tight for the process given in Theorem 5.2 for appropriately chosen t, t_1, t_2. More specifically, in Lemma 5.10 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 5.11 and 5.12, respectively, leading up to Lemma 5.13 that shows the concentration around the critical time $d \log N$, proving a tight upper bound.

Lemma 5.10 *For $d = \frac{\sigma^2}{2\beta^2}$,*

$$\frac{\max_{i \le N}(W_i(d \log N) + W_A(d \log N)) - \beta d \log N - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma \sigma_A}{\sqrt{2\beta}} X, \tag{5.20}$$

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

Lemma 5.11 *For $d = \frac{\sigma^2}{2\beta^2}$ and $0 < \epsilon < d$, and for all x,*

$$\mathbb{P}\left(\frac{\max_{i \le N} \sup_{0<s<(d-\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \ge x\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{5.21}$$

Lemma 5.12 *For $d = \frac{\sigma^2}{2\beta^2}$ and all $\epsilon > 0$, and $x \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{\max_{i \le N} \sup_{s \ge (d+\epsilon)\log N} (W_i(s) + W_A(s) - \beta s) - \frac{\sigma^2}{2\beta} \log N}{\sqrt{\log N}} \ge x\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{5.22}$$

Lemma 5.13 *For $d = \frac{\sigma^2}{2\beta^2}$ and $\epsilon > 0$ and for all x,*

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

$$\leq \mathbb{P}\left(\sigma_A\sqrt{\frac{\sigma^2}{2\beta^2}-\epsilon}X_1+\sqrt{2\epsilon}\sigma_A|X_2|>x\right), \tag{5.23}$$

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

In Appendix 5.A.3 we show how these lemmas can be used to prove Theorem 5.2. In Lemma 5.14, we prove that convergence holds even in L_1, when X is chosen approprately.

Lemma 5.14 *Define $X_N := \frac{\sqrt{2}\beta}{\sigma\sigma_A}\frac{W_A\left(\frac{\sigma^2}{2\beta^2}\log N\right)}{\sqrt{\log N}}$. Then,*

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

The proof of Lemma 5.14 is also given in Appendix 5.A.3. In the next section, we apply Theorem 5.2 and Lemma 5.14 to solve and approximate the minimization problem. Specifically, Lemma 5.14 gives us an order bound between the optimal base-stock level and the approximate base-stock level.

5.5.2 Solution and convergence of the minimization problem

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

Using $P_N^A(z)$ as described in Definition 5.4, we can solve the minimization problem, which yields the optimal base-stock level and net capacity given in Lemma 5.15. The proofs concerning the solution and subsequent convergence results are provided in Appendix 5.A.4.

Definition 5.4 Let

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

Lemma 5.15 *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*

$$I_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}} {P_N^A}^{-1} (1-\gamma_N) \sqrt{\log N}. \tag{5.24}$$

When we are in the balanced regime, we can approximate the minimization problem given in Definition 5.3, using the convergence result in Theorem 5.2, and prove how far the approximate solution is from the optimal solution in terms of costs. This is done in Proposition 5.2 and Theorem 5.3. In Lemma 5.16 we show how the optimal costs scale with N when we are in the balanced regime. The proofs are given in Appendix 5.A.4.

Proposition 5.2 *For $(b^{(N)})_{N\geq 1}, (h^{(N)})_{N\geq 1}$ and $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$,*

$$\hat{I}_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}} \sqrt{\log N}\Phi^{-1}(1-\gamma_N), \tag{5.25}$$

and

$$\hat{C}_N^A(\hat{I}_N^A) = Nh^{(N)}\left(\frac{\sigma^2}{2}\log N - \frac{\sigma^2+\sigma_A^2}{2}\right) + \\ (Nh^{(N)} + b^{(N)})\frac{\sigma\sigma_A\sqrt{\log N}e^{-\frac{1}{2}\Phi^{-1}(1-\gamma_N)^2}}{2\sqrt{\pi}}. \tag{5.26}$$

Theorem 5.3 (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 5.16 (Balanced regime) *Assume $\gamma_N = Nh^{(N)}/(Nh^{(N)} + b^{(N)})$, with $\gamma_N =$*

$\gamma \in (0,1)$. *Then*

$$I_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\Phi^{-1}(1-\gamma) + o(\sqrt{\log N}), \tag{5.27}$$

and

$$F_N(I_N^A, \beta_N^A) = 2\sqrt{N}\sqrt{\hat{C}_N^A(\hat{I}_N^A)} + o(N\sqrt{h^{(N)}}). \tag{5.28}$$

The result in Lemma 5.16 only holds for the balanced regime, so a natural question is what we can say about the efficiency and the quality driven regime. As is shown in Lemma 5.5, in the efficiency driven regime, the first order approximation $\bar{I}_N = \frac{\sigma^2}{2}\log N$ gives that the ratio of the approximate costs and the optimal costs converge to 1. Thus we expect the approximation given in Equation (5.25) will also satisfy this convergence result. In order to determine whether this approximation also satisfies the order bound given in Theorem 5.3, a further analysis is needed. The analysis we provide for the balanced regime heavily relies on van der Vaart (1998, p. 305, Lem. 21.2), which says that if $Y_N \xrightarrow{d} Y$, then for $\gamma \in (0,1)$, $P_{Y_N}^{-1}(\gamma) \overset{N\to\infty}{\longrightarrow} P_Y^{-1}(\gamma)$. This gives us the convergence result in Equation (5.27) of the base-stock level in the balanced regime. In order to be able to prove a similar result for the efficiency driven regime, we should need an improvement of van der Vaart (1998, p. 305, Lem. 21.2) which also holds when $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$.

However, for the quality driven regime, this convergence result does not hold, because we see in Lemma 5.9 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 5.3 we should resort to the analysis of tail asymptotics, which is beyond the scope of this study.

5.5.3 Numerical experiments

In Section 5.5.2, 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. Since it is not straightforward to calculate $\mathbb{E}\left[(\max_{i\le N} Q_i - I)^+\right]$ for dependent Q_i, to evaluate the cost function given in Definition 5.1 we resort to simulation. First, we explain the details of our simulation experiment, after which we discuss the numerical results.

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

Using the above method to simulate $\max_{i \leq N} Q_i$, we can estimate ${P_N^A}^{-1}(1-\gamma_N)$ with $P_N^A(z)$ as described in Definition 5.4. To obtain a median-unbiased estimate of the quantile, we use the approach suggested by Zieliński (2009, p. 982-983). For this, we sample $\max_{i \leq N} Q_i$ 100 times and randomly choose between the observations $(1-\gamma_N) \cdot 100$ and $(1-\gamma_N) \cdot 100 + 1$, with weights depending on the value of the fractile. Our estimate is equal to the median over 100 iterations. Once we have our estimate of ${P_N^A}^{-1}(1-\gamma_N)$, we determine the value of the optimal base-stock level as given in Equation (5.24). Using the optimal base-stock level we determine the optimal net capacity given in Lemma 5.2. Since this also requires the expectation of $(\max_{i \leq N} Q_i - I)^+$, we determine this value by taking the average based on 10,000 simulations.

Next, we compare the costs under our asymptotic approximations of the net capacity and base-stock level (provided in Proposition 5.2) 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 $F_N(I, \beta)$.

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 5.6. We set $h^{(N)} = 1$ and $\sigma = 1$ in all experiments. In total we have 24 instances. The results are given in Tables 5.4 and 5.5 for $b^{(N)} = N$ and $b^{(N)} = 3N$, respectively.

Table 5.4: 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

Table 5.5: 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 5.6: Parameter settings for experiments

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

There are several important observations to be made from Table 5.4. 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, except for the case $N = 10$ and $\sigma_A = 1$ where the difference is around 15%. As N increases to 50, the difference decreases. Furthermore, the difference becomes larger when σ_A increases. In the last column, we verify the convergence result from Theorem 5.3. 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 5.5, we observe that the difference between the asymptotic and optimal costs is considerably higher than for $b^{(N)} = N$. Especially for $N = 10$, the difference is around 15% of the optimum, except for $N = 10$ and $\sigma_A = 0.1$, where the difference is around 20%. However, for a larger number of components, the difference is around 10% of the optimum. Interestingly, for the case $\sigma_A = 1$, the difference between $b^{(N)} = N$ and $b^{(N)} = 3N$ is relatively small.

Overall, in most of our experiments the difference between the costs under the optimal 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 indeed improves when N increases. Finally, the performance of the approximations highly depends on the backorder costs relative to the holding costs.

5.6. Mixed-behavior approximations

The numerical results in Section 5.5.3 show that the approximations are in most of the cases around 10-15% off the optimal value. In this section, we show how we can further improve the approximations.

Under deterministic demand and stochastic demand, the approximate problems are given in Definition 5.2 and Definition 5.3, respectively. If σ_A is small, then we know that on the one hand,

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

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

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

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

$$\max_{i\leq N} Q_i \approx \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G. \tag{5.29}$$

When we replace $\max_{i\leq N} Q_i$ with Equation (5.29) in the minimization problem, we get

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

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,

$$\int_{-\infty}^{\infty}\exp\left(-\exp\left(-\frac{2}{\sigma^2}\left(I_N^M - \frac{\sigma^2}{2}\log N - \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}x\right)\right)\right)\phi(x)dx = 1-\gamma_N. \tag{5.30}$$

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

$$\beta_N^M =$$

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

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

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

We can write

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

Now, we use the substitution $z = -\log s$. Then,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N - z\right)\exp(-\exp(-z) - z)dz\\
&= \int_{0}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)ds.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right]\\
&= \int_{I_N^M}^{\infty}\int_{0}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dsdx\\
&= \int_{0}^{\infty}\int_{I_N^M}^{\infty} \mathbb{P}\left(\frac{\sigma_A\sqrt{2}}{\sigma}\sqrt{\log N}X > \frac{2}{\sigma^2}x - \log N + \log s\right)\exp(-s)dxds.
\end{aligned}$$

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, since I_N^M can be numerically found by solving Equation (5.30), $\mathbb{E}\left[\left(\frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}X + \frac{\sigma^2}{2}G - I_N^M\right)^+\right]$ can be computed numerically as well. Observe that the running time of the procedure to obtain I_N^M and β_N^M is independent of the system size N and is efficient.

5.6.1 Numerical results mixed-behavior approximations

Using the same simulation procedure as described in Section 5.5.3, 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 5.7 and 5.8, respectively.

From the simulation results we can conclude that these adjusted approximations result in costs that are much closer to the optimal costs, already for small N. When comparing the last two columns, where the last column repeats the results from Section 5.5.3, we observe that the mixed-behavior approximations show better convergence, also when σ_A is larger. Furthermore, where we saw in Section 5.5.3 that the cost difference increased considerably with the change in $b^{(N)}$, we now do see an increase, but the difference is still small for a larger value of $b^{(N)}$. Therefore, we can conclude that these mixed-behavior approximations perform well especially when demand variations are no more than 75% of the variations in component production, even with a small number of components.

5.7. Conclusions

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

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

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

Table 5.7: 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

Table 5.8: 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

5.A. Proofs

5.A.1 Proofs of Section 5.3

Lemma 5.17 *(i) In the independent case $\sigma_A = 0$:*

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

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

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

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

Proof of Lemma 5.17

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

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

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

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

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

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

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

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

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

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

Remark 5.1 In the dependent case where all servers choose a different drift parameter, we have that $\sup_{s>0}(W_i(s) + W_A(s) - \beta_i s) = \sup_{s>0}(\hat{W}_i(s) + \hat{W}_A(s) - s)/\beta_i$ where $\hat{W}_i(s) = W_i(s/\beta_i^2)\beta_i$ and $\hat{W}_A(s) = W_A(s/\beta_i^2)\beta_i$. However, $\mathbb{E}\left[W_A(s/\beta_i^2)\beta_i W_A(s/\beta_j^2)\beta_j\right] = \sigma_A^2 \beta_i\beta_j s/\max(\beta_i, \beta_j)^2 \neq \sigma_A^2 s$ when $\beta_i \neq \beta_j$. Thus, when we have different drift parameters β_i and β_j, the joint distribution of $\sup_{s>0}(W_i(s) + W_A(s) - \beta_i s)$ and $\sup_{s>0}(W_j(s) + W_A(s) - \beta_j s)$ is not the same as the joint distribution of $\sup_{s>0}(W_i(s) + W_A(s) - s)/\beta_i$ and $\sup_{s>0}(W_j(s) + W_A(s) -$

$s)/\beta_j$. So to prove Lemma 5.17 when the drifts are different, other techniques are needed.

Proof of Lemma 5.1

$F_N(I,\beta) > 0$, hence F_N has a global infimum, and since $\lim_{\beta\downarrow 0} F_N(I,\beta) = \infty$, $\lim_{\beta\to\infty} F_N(I,\beta) = \infty$ 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

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

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

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

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

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

Proof of Lemma 5.2

$F_N(I,\beta)$ 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 5.1 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 $F_N(I_N^*,\beta_N^*) = 2\sqrt{NC_N(I_N^*)} = 2N\beta_N^*$. □

Proof of Lemma 5.3

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

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

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

Proof of Lemma 5.5

Following Corollary 5.1, we have

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

Furthermore, observe that

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

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

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

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

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

In the case that $\gamma_N = \gamma \in (0,1)$, we have by applying Lemma 5.4 that $|\bar{C}_N(\bar{I}_N) - C_N(\bar{I}_N)| = o((Nh^{(N)} + b^{(N)})\log N)$. Furthermore, $C_N(\bar{I}_N) \sim Nh^{(N)}\frac{\sigma^2}{2}\log N$, and since $\max_{i\leq N} Q_i/\log N \xrightarrow{\mathbb{P}} \sigma^2/2$, as $N \to \infty$, we also have that $I_N^*/\log N \overset{N\to\infty}{\longrightarrow}$

$\sigma^2/2$. 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 $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, we first observe that

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

Furthermore,

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

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

Thus,

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

By Lemma 5.4, we know that $\mathbb{E}\left[\left| \max_{i \le N} Q_i - \frac{\sigma^2}{2} \log N \right| \right] / \log N \overset{N\to\infty}{\longrightarrow} 0$. Thus

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

Finally,

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

$I_N^* = O(\log N)$, and $b^{(N)}/(Nh^{(N)}) \overset{N\to\infty}{\longrightarrow} 0$, therefore, $\liminf_{N\to\infty} \frac{C_N(I_N^*)}{Nh^{(N)} \log N} \ge \sigma^2/2$.

Combining these results gives

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

□

5.A.2 Proofs of Section 5.4

Proof of Lemma 5.6

In Lemma 5.3, it is shown that $I_N^* = P_N^{-1}(1-\gamma_N)$, with P_N^{-1} the quantile function of $\max_{i\leq N} Q_i$. Because $(Q_i, i \leq N)$ are independent and exponentially distributed,

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

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

Proof of Proposition 5.1

Minimizing $\hat{F}_N(\hat{I}_N, \hat{\beta}_N)$ goes analogously as minimizing $F_N(I_N, \beta_N)$ in Lemma 5.6. Hence $\hat{I}_N = \hat{P}_N^{-1}(1-\gamma_N)$. Thus, we have to solve

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

Therefore, $\hat{P}_N^{-1}(x) = \frac{\sigma^2}{2}\log N - \frac{\sigma^2}{2}\log(-\log x)$. Hence, the optimal inventory is given in Equation (5.7). Furthermore,

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

By using partial integration and substitution we can write

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

Hence, this gives us the expression of $\hat{C}_N(\hat{I}_N)$ in (5.8). □

Proof of Lemma 5.7

To prove that G_N follows a Gumbel distribution, we first observe that

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

Therefore, $\left(1-\exp\left(-\frac{2}{\sigma^2}\max_{i\le N} Q_i\right)\right)^N \sim \mathrm{Unif}[0,1]$. Then,

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

To prove (5.10), 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 $x>-\frac{\sigma^2}{2}\log\left(-\log\left(1-\exp\left(-\frac{2}{\sigma^2}x\right)\right)\right)$, 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 $\max_{i\le N} Q_i$ 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

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

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

Proof of Lemma 5.8

Due to the inequality in (5.10), $I_N^* > \hat{I}_N$, then, we have

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

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

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

Furthermore,

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

Equation (5.11) follows. To prove Equation (5.12), we observe that

$$\begin{aligned}
&|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)| \\
=&(Nh^{(N)} + b^{(N)})\,\mathbb{E}\left[\left(\max_{i\leq N} Q_i - \hat{I}_N\right)^+ - \left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N - \hat{I}_N\right)^+\right] \\
=&(Nh^{(N)} + b^{(N)})\,\mathbb{E}\left[\left(\max_{i\leq N} Q_i - \frac{\sigma^2}{2}G_N - \frac{\sigma^2}{2}\log N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\right)\right] \qquad (5.31)\\
+&(Nh^{(N)} + b^{(N)})\,\mathbb{E}\left[\left(\max_{i\leq N} Q_i - \hat{I}_N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N < \max_{i\leq N} Q_i\right)\right]. \qquad (5.32)
\end{aligned}$$

Because G_N and $\max_{i\leq N} Q_i$ are on the same probability space, we have

$$\mathbb{P}\Big(\max_{i\leq N} Q_i = I_N^* \Big| \frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N = \hat{I}_N\Big) = 1.$$

Furthermore, $x + \frac{\sigma^2}{2}\log(-\log(1 - e^{-\frac{2}{\sigma^2}x}))$ is decreasing in x. Thus, we can bound

$$\begin{aligned}
&\mathbb{E}\left[\left(\max_{i\leq N} Q_i - \frac{\sigma^2}{2}G_N - \frac{\sigma^2}{2}\log N\right)\mathbb{1}\left(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\right)\right] \\
\leq&(I_N^* - \hat{I}_N)\,\mathbb{P}\Big(\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N > \hat{I}_N\Big) \\
=&(I_N^* - \hat{I}_N)\gamma_N. \qquad (5.33)
\end{aligned}$$

Similarly, for (5.32), we observe that if $\frac{\sigma^2}{2}G_N + \frac{\sigma^2}{2}\log N < \hat{I}_N$, then $\max_{i\leq N} Q_i < I_N^*$, thus,

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

$$\leq (I_N^* - \hat{I}_N)\left(1 - \left(1 + \frac{\log(1-\gamma_N)}{N}\right)^N - \gamma_N\right). \tag{5.34}$$

Adding the bounds in (5.33) and (5.34) gives the result. □

Proof of Theorem 5.1

First of all, we assume that $\gamma_N = \gamma \in (0,1)$. Using Corollary 5.1, 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 (5.10), we have for all I that $C_N(I) > \hat{C}_N(I)$, thus

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

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

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

Thus, $(I_N^* - \hat{I}_N) \sim -\sigma^2\log(1-\gamma)/(4N)$. Following (5.12), we can conclude that $\frac{|\hat{C}_N(\hat{I}_N) - C_N(\hat{I}_N)|}{Nh^{(N)}} = O(1/N)$. 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 (5.11), we get $|C_N(I_N^*) - C_N(\hat{I}_N)|/(Nh^{(N)} + b^{(N)}) = O(1/N^2)$. We have

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

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

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

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

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

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

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

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

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

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

Earlier, we showed that when $\gamma_N = \gamma$, $(I_N^* - \hat{I}_N) = O(1/N)$, now I_N^* is larger, because $\mathbb{P}(\max_{i\leq N} Q_i < I_N^*) = 1 - \gamma_N \overset{N\to\infty}{\longrightarrow} 1$. Following the statement in Lemma 5.7 that the difference between $\max_{i\leq N} Q_i$ and $\frac{\sigma^2}{2} G_N + \frac{\sigma^2}{2}\log N$ decreases as $\max_{i\leq N} Q_i$ increases, we can conclude that $(I_N^* - \hat{I}_N) = O(1/N)$. Following the proof before, and by using the order bounds in (5.11) and (5.12), 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 $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$ and $\gamma_N \leq 1 - \exp(-N)$. Then, $\hat{I}_N \geq 0$. Furthermore, when $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, we have $\log(-\log(1-\gamma_N)) \overset{N\to\infty}{\longrightarrow} \infty$, from this it follows that

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

Thus

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

Since we consider the efficiency driven regime, we have $b^{(N)}/(Nh^{(N)}) \overset{N\to\infty}{\longrightarrow} 0$. Also, it is easy to deduce that when $\gamma_N < 1 - \exp(-N)$, we have $\log(-\log(1-\gamma_N)) < \log N$. Thus $\hat{C}_N(\hat{I}_N) \sim \frac{\sigma^2}{2} N h^{(N)} \log N$. Furthermore, $I_N^* - \hat{I}_N = O(1)$, thus the bounds in (5.11) and (5.12) are of $O(Nh^{(N)})$. By using the same argument as in the proof for the balanced regime,

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

□

Proof of Lemma 5.9

Following Equations (5.11) and (5.12) and using the same arguments as in the proof of Theorem 5.1, 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

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

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

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

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

From this it follows that

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

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 $\gamma_N \overset{N\to\infty}{\longrightarrow} 1$, we have

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

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

5.A.3 Proofs of Section 5.5.1

Proof of Lemma 5.10

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, p. 11, Ex. 1.1.7) for a proof. Observe that

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

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

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

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. □

Proof of Lemma 5.11

To prove Lemma 5.11, we first observe that

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

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

Furthermore, we can write

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

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

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

It turns out that

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

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

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

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

$$\begin{aligned}\frac{\max_{i\leq N}\sup_{s<(d-\epsilon)\log N}(W_i(s)-\beta s)}{\log N}&\xrightarrow{\mathbb{P}}\frac{\sigma\sqrt{\sigma^2-2\beta^2\epsilon}}{\beta}-\frac{\sigma^2}{2\beta}+\beta\epsilon\\ &=\frac{\sigma^2}{2\beta}+\left(\frac{\sigma\sqrt{\sigma^2-2\beta^2\epsilon}}{\beta}-\frac{\sigma^2}{\beta}+\beta\epsilon\right)\leq\frac{\sigma^2}{2\beta}-C\epsilon^2,\quad(5.35)\end{aligned}$$

with $C>0$. Hence,

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

Thus,

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

for N large. Therefore,

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

Furthermore, $\sup_{s>0}\left(W_A(s)-\frac{s}{\log\log N}\right)\sim\text{Exp}\left(\frac{2}{\sigma_A^2\log\log N}\right)$. Therefore,

$$\frac{\sup_{s>0}\left(W_A(s)-\frac{s}{\log\log N}\right)}{\sqrt{\log N}}\xrightarrow{\mathbb{P}}0,$$

as $N\to\infty$. The statement follows. □

Proof of Lemma 5.12

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

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

We have

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{P}\left(\max_{i\leq N}\left(W_i((d+\epsilon)\log N)+E_i\right) > x\sqrt{\log N} + \left(\frac{\sigma^2}{2\beta} + (\beta-\delta)(d+\epsilon)\right)\log N\right)\\ &\leq N^{1+\frac{\sigma^2(d+\epsilon)s^2}{2} - s\left(\frac{\sigma^2}{2\beta}+(\beta-\delta)(d+\epsilon)\right)} e^{-sx\sqrt{\log N}}\frac{\frac{2(\beta-\delta)}{\sigma^2}}{\frac{2(\beta-\delta)}{\sigma^2} - s}.\end{aligned} \tag{5.36}$$

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

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

Moreover, $\sup_{s>0}(W_A(s) - \delta s) \sim \text{Exp}\left(\frac{2\delta}{\sigma_A^2}\right)$. Therefore, $\frac{\sup_{s>0}(W_A(s)-\delta s)}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0$. The limit in (5.22) follows. □

Proof of Lemma 5.13

First of all, we bound

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

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

We can write

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

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

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

as $N \to \infty$, 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. □

Proof of Theorem 5.2

We have the following lower bound:

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

From this and Lemma 5.10, we know that

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

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

By using the union bound, we get

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

Combining this with the results from Lemmas 5.11, 5.12 and 5.13 gives

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

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

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

Hence, the statement follows. □

Proof of Lemma 5.14

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

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

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

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

and

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

We want to prove that

$$\mathbb{E}\left[\left|\frac{\max_{i\le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right] \overset{N\to\infty}{\longrightarrow} 0. \tag{5.37}$$

First observe that

$$\mathbb{E}\left[\left|\frac{\max_{i\le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right] \tag{5.38}$$

$$\le \mathbb{E}\left[\left|\frac{\max_{i\le N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\le N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right] \tag{5.39}$$

$$+\,\mathbb{E}\left[\left|\frac{\max_{i\le N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right]. \tag{5.40}$$

Because $Q_i > W_i(d\log N) + W_A(d\log N) - d\log N$, we can rewrite (5.39):

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{\max_{i\leq N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right]\\ =&\mathbb{E}\left[\left(\frac{\max_{i\leq N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right)^+\right].\end{aligned} \tag{5.41}$$

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

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}} - \frac{\sigma\sigma_A}{\sqrt{2}}X_N\right|\right]\\ =&\mathbb{E}\left[\left|\frac{\max_{i\leq N} W_i(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right|\right] \overset{N\to\infty}{\longrightarrow} 0.\end{aligned} \tag{5.42}$$

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

$$\mathbb{E}\left[\left(\frac{\max_{i\leq N} Q_i - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right)^+\right] \tag{5.43}$$

$$\leq \mathbb{E}\left[\left(\frac{\max_{i\leq N} Q_i^{(1,N)} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right)^+\right] \tag{5.44}$$

$$+ \mathbb{E}\left[\left(\frac{\max_{i\leq N} Q_i^{(2,N)} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right)^+\right] \tag{5.45}$$

$$+ \mathbb{E}\left[\left(\frac{\max_{i\leq N} Q_i^{(3,N)} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{\max_{i\leq N} W_i(d\log N) + W_A(d\log N) - \sigma^2\log N}{\sqrt{\log N}}\right)^+\right]. \tag{5.46}$$

For (5.44), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\max_{i\le N} Q_i^{(1,N)}-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}-\frac{\max_{i\le N} W_i(d\log N)+W_A(d\log N)-\sigma^2\log N}{\sqrt{\log N}}\right)^+\right]\\
\le&\,\mathbb{E}\left[\left(\frac{\max_{i\le N} Q_i^{(1,N)}-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}-\frac{W_A(d\log N)}{\sqrt{\log N}}\right)^+\right]\\
&+\mathbb{E}\left[\left(-\frac{\max_{i\le N} W_i(d\log N)-\sigma^2\log N}{\sqrt{\log N}}\right)^+\right]. \qquad (5.47)
\end{aligned}$$

By (5.42), the second term converges to 0. For the first term, following Lemma 5.11, we observe that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\max_{i\le N} Q_i^{(1,N)}-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}-\frac{W_A(d\log N)}{\sqrt{\log N}}\right)^+\right]\\
\le&\mathbb{E}\Bigg[\Bigg(\frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s)-(1-1/\log\log N)s))-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\\
&+\frac{\sup_{s>0}(W_A(s)-s/\log\log N)}{\sqrt{\log N}}-\frac{W_A(d\log N)}{\sqrt{\log N}}\Bigg)^+\Bigg]\\
\le&\mathbb{E}\Bigg[\Bigg(\frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s)-(1-1/\log\log N)s))-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\\
&-\frac{W_A(d\log N)}{\sqrt{\log N}}\Bigg)^+\Bigg]+\mathbb{E}\left[\left(\frac{\sup_{s>0}\left(W_A(s)-s/\log\log N\right)}{\sqrt{\log N}}\right)^+\right]. \qquad (5.48)
\end{aligned}$$

The term in (5.48) converges to 0. Furthermore,

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s)-(1-1/\log\log N)s))-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\\
&-\frac{W_A(d\log N)}{\sqrt{\log N}}\Bigg)^+\Bigg]\\
=&\int_0^\infty \mathbb{P}\Bigg(\frac{\max_{i\le N}(\sup_{0<s<(d-\epsilon)\log N}(W_i(s)-(1-1/\log\log N)s))-\frac{\sigma^2}{2}\log N}{\sqrt{\log N}}
\end{aligned}$$

$$
\begin{aligned}
&-\frac{W_A(d\log N)}{\sqrt{\log N}} > x\Bigg) dx\\
&= \int_0^\infty \mathbb{P}\left(\max_{i\le N} \sup_{0<s<(d-\epsilon)\log N} (W_i(s) - (1-1/\log\log N)s) - W_A(d\log N) > \right.\\
&\quad \left. x\sqrt{\log N} + \frac{\sigma^2}{2}\log N\right) dx\\
&\le \int_0^\infty N\mathbb{P}\Big(\sup_{0<s<(d-\epsilon)\log N} (W_i(s) - (1-1/\log\log N)s)\\
&> x/2\sqrt{\log N} + \Big(\frac{\sigma^2}{2} - \frac{1}{2}(\sigma^2 - \sigma\sqrt{\sigma^2-2\epsilon} - \epsilon)\Big)\log N\Big)dx\\
&+ \int_0^\infty \mathbb{P}\left(-W_A(d\log N) > x/2\sqrt{\log N} + \left(\frac{1}{2}\left(\sigma^2 - \sigma\sqrt{\sigma^2-2\epsilon} - \epsilon\right)\right)\log N\right)dx\\
&\overset{N\to\infty}{\longrightarrow} 0;
\end{aligned}
$$

see the inequality (5.35) in the proof of Lemma 5.11 for details. For the term in (5.45), we have that

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

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

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

Furthermore, (here we use the same bounds as in Lemma 5.13)

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\max_{i\le N} Q_i^{(2,N)} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right]\\
&\le \mathbb{E}\left[\frac{\max_{i\le N} \sup_{s>0}(W_i(s) - s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right]
\end{aligned}
$$

$$+\mathbb{E}\left[\frac{W_A((d-\epsilon)\log N)) + \sup_{0<2\epsilon\log N}\tilde{W}_A(s)}{\sqrt{\log N}}\right]$$

$$\overset{N\to\infty}{\longrightarrow} 0 + \sqrt{2\epsilon}\sigma_A\mathbb{E}[|X_N|] = \sqrt{2\epsilon}\sigma_A\sqrt{\frac{2}{\pi}}. \tag{5.49}$$

Similar as in (5.47), we have

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

The second term goes to 0, following the proof of Lemma 5.12, for the first term we have

$$\mathbb{E}\left[\left(\frac{\max_{i\le N} Q_i^{(3,N)} - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}} - \frac{W_A(d\log N)}{\sqrt{\log N}}\right)^+\right]$$
$$\le \mathbb{E}\Bigg[\Bigg(\frac{\max_{i\le N}\sup_{s>(d+\epsilon)\log N}(W_i(s)-(1-\delta)s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}$$
$$+\frac{\sup_{s>(d+\epsilon)\log N}(W_A(s)-\delta s) - W_A(d\log N)}{\sqrt{\log N}}\Bigg)^+\Bigg]$$
$$\le \mathbb{E}\left[\left(\frac{\max_{i\le N}\sup_{s>(d+\epsilon)\log N}(W_i(s)-(1-\delta)s) - \frac{\sigma^2}{2}\log N}{\sqrt{\log N}}\right)^+\right] \tag{5.50}$$
$$+\mathbb{E}\left[\left(\frac{\sup_{s>(d+\epsilon)\log N}(W_A(s)-\delta s) - W_A(d\log N)}{\sqrt{\log N}}\right)^+\right]. \tag{5.51}$$

For (5.50), we have, by using the inequality from Equation (5.36) with $s^\star = 1/(2\epsilon + \sigma^2) + (1-\delta)/\sigma^2$, that

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

For (5.51), we observe that

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{E}\left[\left(\frac{\sup_{s>(d+\epsilon)\log N}(W_A(s)-\delta s)-W_A(d\log N)}{\sqrt{\log N}}\right)^+\right]\\
&=\limsup_{N\to\infty}\mathbb{E}\left[\left(\frac{\tilde{W}_A(\epsilon\log N)-\delta(d+\epsilon)\log N+\sup_{s>0}(W_A(s)-\delta s)}{\sqrt{\log N}}\right)^+\right]\\
&\le\mathbb{E}\left[\left(\frac{\tilde{W}_A(\epsilon\log N)}{\sqrt{\log N}}\right)^+\right]+\mathbb{E}\left[\left(\frac{-\delta(d+\epsilon)\log N}{\sqrt{\log N}}\right)^+\right]\\
&\quad+\mathbb{E}\left[\left(\frac{\sup_{s>0}(W_A(s)-\delta s)}{\sqrt{\log N}}\right)^+\right]\\
&\overset{N\to\infty}{\longrightarrow}\sigma_A\sqrt{\epsilon}\sqrt{\frac{1}{2\pi}}. \qquad (5.52)
\end{aligned}$$

Concluding, after we collect the non-zero answers which are given in (5.49) and (5.52) we get

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

□

5.A.4 Proofs of Section 5.5.2

Proof of Lemma 5.15

From Lemma 5.1, we know that the optimal inventory I_N^A satisfies

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

We have

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

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

Proof of Proposition 5.2

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 (5.25) minimizes $\hat{C}_N^A(I)$. We know that

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

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

The expression in Equation (5.26) follows. □

Proof of Theorem 5.3

Using Corollary 5.1, 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

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

We know by van der Vaart (1998, p. 305, Lem. 21.2), that $(I_N^A - \hat{I}_N^A)/\sqrt{\log N} \overset{N\to\infty}{\longrightarrow} 0$. Furthermore, we prove in Lemma 5.14 that

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

From this it follows that $|\hat{C}_N^A(\hat{I}_N^A) - C_N(I_N^A)| = o((Nh^{(N)} + b^{(N)})\sqrt{\log N})$. Since $\hat{C}_N^A(\hat{I}_N^A) \sim \frac{\sigma^2}{2}Nh^{(N)}\log N$, we have

$$
\frac{\sqrt{C_N(I_N^A)}}{\sqrt{\hat{C}_N^A(\hat{I}_N^A)}} = 1 - o\left((Nh^{(N)} + b^{(N)})\sqrt{\log N}/(Nh^{(N)}\log N)\right) = 1 - o\left(1/\sqrt{\log N}\right).
$$

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

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

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

Proof of Lemma 5.16

We have $\hat{I}_N^A = \frac{\sigma^2}{2}\log N + \frac{\sigma\sigma_A}{\sqrt{2}}\sqrt{\log N}\Phi^{-1}(1-\gamma)$. Furthermore, $|I_N^A - \hat{I}_N^A| = o(\sqrt{\log N})$, thus (5.27) follows. Furthermore, by using the same argument as in Lemma 5.9, (5.28) follows. □

6

Conclusions and Future Work

In this thesis, we studied decisions regarding control of high-tech supply chains. We considered three different research problems, as discussed in Chapter 1. In Chapter 2, we focused on contracting between a high-tech OEM and a supplier of a critical component of the end-product, in particular, how long-term collaborations can help to achieve coordination. In Chapter 3, we further built on this analysis for the case where only a limited number of suppliers has the capabilities to supply the required component. Chapter 4 considered an assembly system in which an OEM combines components produced in-house with components sourced at an outside supplier into an end-product. We designed an order policy for the supplier-sourced component that minimizes overall costs. Finally, in Chapter 5, we investigated how the large scale of high-tech assembly systems can be used in optimizing both capacity and inventory for the required components.

In Section 6.1, we discuss the main findings of this thesis followed by their practical implications in Section 6.2. We conclude this thesis in Section 6.3 by providing potential directions for future research.

6.1. Main results

In Chapter 2, our research objective is to develop a supply chain contract that improves coordination in high-tech supply chains by offering the prospect of contract renewal. First, we study contracts that focus on a single generation. We confirm the well-known result that a wholesale price contract leads to double marginalization and hence a loss of total supply chain profit. A wholesale price contract augmented with a penalty for failure to satisfy demand theoretically performs well, as it coordinates the supply chain and allows the OEM to take all profits. However, in a high-tech supply chain, with a very high-valued end-product, such contract will likely not be enforceable as the required penalty is not achievable in practice. For this reason, we extend a standard wholesale price contract to the case where the OEM will source components for multiple generations of its end-product. We show that the promise of contract renewal, or the indirect penalty of non-renewal, contingent upon supplier performance can motivate the supplier to invest in sufficient capacity. Even though under this contract the OEM cannot capture the entire supply chain profits, we show that the OEM can capture a large share of the profits and has considerable benefits compared to a single-generation wholesale price contract.

Research objective 2 concerns the effectiveness of the renewal contract in case only a limited number of suppliers has the required capabilities. To this end, in Chapter 3, we extend the wholesale price contract with contingent renewal to the case where the OEM has two supplier options: the incumbent supplier and an alternative supplier. Once the OEM switches to the alternative supplier, the incumbent supplier knows that the OEM will return at some point, when switching back from the alternative supplier. Since the supplier knows that the OEM will not be lost as a customer forever, the supplier has a stronger position than in case there are infinitely many possible suppliers. Despite the stronger position of the supplier, we show that our structural results from Chapter 2 continue to hold and contract renewal remains an effective incentive for the supplier to invest in capacity.

In Chapter 4, we switch our focus from contracting for a single component to synchronizing between multiple components that are required in the production process. Specifically, we study synchronization between a component that is produced in-house by the OEM in a make-to-order production system and a component that is sourced at a supplier with a specific lead-time. Since both components need to be available to assemble and deliver the end-product, unavailability of one of the components gives rise to high holding costs and customer waiting costs. As the in-house produced component is made-to-order, every component is coupled to a customer order. Therefore, the number of orders waiting for the capacitated in-house production process gives an indication of the demand for the supplier-sourced module during its lead-time and can be used for optimizing the order policy. We prove optimality of the myopic state-dependent base-stock policy both in continuous time and discrete time. These results indicate that synchronization between a capacitated process and a lead-time component can be achieved by adjusting the order policy of the lead-time component to the expected output of the production process.

Our final research objective, which is the topic of Chapter 5, concerns simultaneous optimization of capacity investment decisions and inventory policies in large assembly systems. As high-tech end-products often consist of many modules, we investigate how the scale of these systems can be utilized to optimize decisions. Since demand of the end-product is stochastic and delays may occur in the production of individual components and shortage of a single components leads to costly delays in assembly of the end-product, a trade-off arises between shortage risk, capacity investments and inventory holding costs. The delay in assembly of the end-product equals the delay of the component with maximum delay. Therefore, the first step in deriving asymptotically optimal capacity and inventory decisions is expressing the expected maximum delay over all components. When demand of the end-product is deterministic, we use a well-known extreme value limit to obtain approximations for capacity and inventory that result in costs that are close to the optimal costs already for a limited number of components. For the case where demand is stochastic, delays of the individual components are correlated. We develop a novel limit theorem for the maximum delay as the number of components grows large, which we use to obtain approximate solutions. Since demand in high-tech supply chains is often predictable, resulting in a relatively low standard deviation, we further improve these approximations by considering a mixed-behavior regime where we use a combination of the approximations for the deterministic and stochastic demand scenarios. We show numerically that these approximations perform well already when the number of components in the end-product is limited. Our results thus indicate that the size of high-tech assembly systems allows for application of extreme-value theory for optimizing decisions.

6.2. Practical implications

From these results, we obtain several key insights for high-tech manufacturers. When considering the sourcing decision of a single component, there are some important factors to consider. First, when the sourcing decisions of the OEM for a new product generation depend on supplier performance for the previous product generation, coordination in high-tech supply chains may be improved tremendously. This means that the OEM's sourcing department should not make supplier selection solely dependent on costs, but also on past performance. This requires communication and information sharing between sourcing and operations departments at the OEM. Second, to entice the supplier to invest sufficiently by offering contract renewal dependent on supplier performance, it is important that the suppliers are aware of this policy. This requires either formal contracting with the suppliers regarding contract renewal or clear communication of the OEM's intentions. Finally, we have shown that the manufacturer's share of total profits is dependent on the supplier's valuation of future profits. A supplier that has a high valuation of future profits is more likely to increase capacity investments to increase the renewal probability. Consequently, it is beneficial for high-tech OEMs to work with suppliers that focus on long-term sustainable business rather than short-term profit.

When considering the assembly process of high-tech products, for which multiple components or modules are required, we show that synchronization is important to avoid costly delays in production of the end-product caused by shortages of a single component. We show that considerable savings can be achieved by using all available information when making ordering decisions for lead-time components, including the number of outstanding orders for in-house produced make-to-order components. Base-stock policies with varying base-stock levels that take into account this information can generate considerable savings compared to fixed base-stock levels. Also, when the end-product contains many components that are produced in capacitated systems, it is important to align inventories and capacities of the different components. In this way, one can avoid large backlog of a single component resulting in costly delays of production of the end-product and high inventory holding costs due to large inventories of all other components. Our asymptotic analysis provides easy to implement capacity and inventory decisions that result in near-optimal costs.

6.3. Future research

In this thesis we have studied specific problems in coordination and optimization in high-tech supply chains. To complete this thesis, we discuss some potential future research directions.

In Chapter 3 we extended our study of renewal contracts form Chapter 2 to the case where the supply base is limited. Specifically, we considered the case where there are only two oligopolistic suppliers. Further research could consider the case where there are N possible suppliers, with $2 < N < \infty$. This would allow us to analyze how fast our results converge to the unlimited supplier case as N grows large. In this case, it will be of importance how the OEM selects the next supplier to work with. When the supplier rotates through a fixed list of possible suppliers, the same solution approach may be used. When the supplier selection rules are more complex, we can no longer express $\mathbb{E}\left[\delta^{Y_2}\right]$ in the same way.

Chapter 4 concerns synchronization between an in-house produced component and a supplier-sourced component with lead-time. In practice, end-products often consist of multiple components, as discussed in Chapter 5. In Chapter 4, we considered the extension of synchronizing the order policy of a supplier-sourced component with lead-time with the output of an in-house production system in which multiple make-to-order components are produced and showed that the myopic state-dependent base-stock policy remains optimal. To further optimize the entire process, it would be beneficial to coordinate among the production processes of the make-to-order components. However, combined with determining the order policy of the lead-time component, this results in a complex problem.

In Chapter 5 we show that our approximations for capacity and inventory perform well for practically sized supply chains. In this chapter, we assumed that the assembly system is centrally controlled. In practice, many manufacturers source components that are incorporated in their end-product from suppliers. In such decentralized system, all suppliers make their own decisions regarding capacity investments. This results in different dynamics. Also, agreements need to be made about division of overall profits and who is responsible for back-order costs. An intuitive idea would be to transfer all penalty costs to the slowest supplier, who ultimately delays production of the end-product, but as discussed in Chapter 2 this may not always be feasible. To optimize this problem given all these dynamics, we suggest to look at the problem from a game theoretical perspective.

Bibliography

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

ASML Holding N.V. ASML annual report 2020. `https://www.asml.com/en/investors/annual-report/2020`, 2021.

S. Asmussen, P. W. Glynn, and J. Pitman. Discretization error in simulation of one-dimensional reflecting Brownian motion. *The Annals of Applied Probability*, pages 875–896, 1995.

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

Z. Atan, T. Ahmadi, C. Stegehuis, T. de Kok, and I. Adan. Assemble-to-order systems: A review. *European Journal of Operational Research*, 261(3):866–879, 2017.

R. Atar, A. Mandelbaum, and A. Zviran. Control of fork-join networks in heavy traffic. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 823–830. IEEE, 2012.

F. Baccelli. Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. *RR-0426, INRIA. ffinria-00076130*, 1985.

F. Baccelli and A. M. Makowski. Queueing models for systems with synchronization constraints. *Proceedings of the IEEE*, 77(1):138–161, 1989.

D. Barnes-Schuster, Y. Bassok, and R. Anupindi. Coordination and flexibility in supply contracts with options. *Manufacturing & Service Operations Management*, 4 (3):171–207, 2002.

S. Benjaafar and M. ElHafsi. Production and inventory control of a single product assemble-to-order system with multiple customer classes. *Management Science*, 52 (12):1896–1912, 2006.

S. Benjaafar, M. ElHafsi, C.-Y. Lee, and W. Zhou. Optimal control of an assembly system with multiple stages and multiple demand classes. *Operations Research*, 59 (2):522–529, 2011.

F. Bernstein and G. A. DeCroix. Inventory policies in a decentralized assembly system. *Operations Research*, 54(2):324–336, 2006.

R. Bollapragada, U. S. Rao, and J. Zhang. Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Science*, 50(12):1729–1743, 2004.

R. Bollapragada, S. Kuppusamy, and U. S. Rao. Component procurement and end product assembly in an uncertain supply and demand environment. *International Journal of Production Research*, 53(3):969–982, 2015.

S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

J. R. Bradley and P. W. Glynn. Managing capacity and inventory jointly in manufacturing systems. *Management Science*, 48(2):273–288, 2002.

A. O. Brown and H. L. Lee. The impact of demand signal quality on optimal decisions in supply contracts. In J. Shanthikumar, D. Yao, and W. Zijm, editors, *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, International Series in Operations Research & Management Science, pages 299–328. Springer, 2003.

B. M. Brown and S. I. Resnick. Extreme values of independent stochastic processes. *Journal of Applied Probability*, pages 732–739, 1977.

G. P. Cachon. Supply chain coordination with contracts. In S. Graves and A. de Kok, editors, *Handbooks in Operations Research and Management Science: Supply Chain Management*, volume 11, pages 227–339. Elsevier, 2003.

F. Chen and J.-S. Song. Optimal policies for multiechelon inventory problems with markov-modulated demand. *Operations Research*, 49(2):226–234, 2001.

T. Cheng, C. Gao, and H. Shen. Production planning and inventory allocation of a single-product assemble-to-order system with failure-prone machines. *International Journal of Production Economics*, 131(2):604–617, 2011.

P. Chintapalli, S. M. Disney, and C. S. Tang. Coordinating supply chains via advance-order discounts, minimum order quantities, and delegations. *Production and Operations Management*, 26(12):2175–2186, 2017.

A. J. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490, 1960.

C. J. Corbett. Stochastic inventory systems in a supply chain with asymmetric information: Cycle stocks, safety stocks, and consignment stock. *Operations Research*, 49(4):487–500, 2001.

J. Dai and J. M. Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, pages 65–86, 1992.

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2006.

T. de Kok, C. Grob, M. Laumanns, S. Minner, J. Rambau, and K. Schade. A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research*, 269(3):955–983, 2018.

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

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

M. ElHafsi, H. Camus, and E. Craye. Managing an integrated production inventory system with information on the production and demand status and multiple non-unitary demand classes. *European Journal of Operational Research*, 207(2):986–1001, 2010.

W. J. Elmaghraby. Supply contract competition and sourcing policies. *Manufacturing & Service Operations Management*, 2(4):350–371, 2000.

M. Erkoc and S. D. Wu. Managing high-tech capacity expansion via reservation contracts. *Production and Operations Management*, 14(2):232–251, 2005.

L. Flatto and S. Hahn. Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

M. R. Frascatore and F. Mahmoodi. Long-term and penalty contracts in a two-stage supply chain with stochastic demand. *European Journal of Operational Research*, 184 (1):147–156, 2008.

M. J. Fry, R. Kapuscinski, and T. L. Olsen. Coordinating production and delivery under a (z,Z)-type vendor-managed inventory contract. *Manufacturing & Service Operations Management*, 3(2):151–173, 2001.

G. Gallego and H. Hu. Optimal policies for production/inventory systems with finite capacity and markov-modulated demand and supply processes. *Annals of Operations Research*, 126(1-4):21–41, 2004.

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

B. Ghosh. Probability inequalities related to markov's theorem. *The American Statistician*, 56(3):186–190, 2002.

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

R. Guan and X. Zhao. On contracts for VMI program with continuous review (r,Q) policy. *European Journal of Operational Research*, 207(2):656–667, 2010.

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

J. M. Harrison. *Brownian Models of Performance and Control*. Cambridge University Press, 2013. doi: 10.1017/CBO9781139087698.

B. Hu, C. Meng, D. Xu, and Y.-J. Son. Supply chain coordination under vendor managed inventory-consignment stocking contracts with wholesale price constraint and fairness. *International Journal of Production Economics*, 202:21–31, 2018.

S. Hu, Z. Wan, Q. Ye, and W. Chi. Supplier behavior in capacity investment competition: An experimental study. *Production and Operations Management*, 26 (2):273–291, 2017.

W. T. Huh and G. Janakiraman. Base-stock policies in capacitated assembly systems: Convexity properties. *Naval Research Logistics*, 57(2):109–118, 2010.

M. Jin and S. D. Wu. Capacity reservation contracts for high-tech industry. *European Journal of Operational Research*, 176(3):1659–1677, 2007.

G. A. Karaarslan, Z. Atan, T. de Kok, and G. P. Kiesmüller. Optimal and heuristic policies for assemble-to-order systems with different review periods. *European Journal of Operational Research*, 271(1):80–96, 2018.

S. J. d. Klein. *Fredholm integral equations in queueing analysis.* PhD thesis, Rijksuniversiteit Utrecht, 1988.

S. T. Klosterhalfen, S. Minner, and S. P. Willems. Strategic safety stock placement in supply networks with static dual supply. *Manufacturing & Service Operations Management*, 16(2):204–219, 2014.

S.-S. Ko and R. F. Serfozo. Response times in M/M/s fork-join networks. *Advances in Applied Probability*, 36(3):854–871, 2004.

S. Kou, H. Zhong, et al. First-passage times of two-dimensional Brownian motion. *Advances in Applied Probability*, 48(4):1045–1060, 2016.

S. Kumar and R. S. Randhawa. Exploiting market size in service systems. *Manufacturing & Service Operations Management*, 12(3):511–526, 2010.

M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes.* Springer Science & Business Media, 1983.

J.-Y. Lee and R. K. Cho. Optimal (z, Z)-type contracts for vendor-managed inventory. *International Journal of Production Economics*, 202:32–44, 2018.

C. Li. Sourcing for supplier effort and competition: Design of the supply base and pricing mechanism. *Management Science*, 59(6):1389–1406, 2013.

C. Li and L. G. Debo. Second sourcing vs. sole sourcing with capacity investment and asymmetric information. *Manufacturing & Service Operations Management*, 11 (3):448–470, 2009.

C. Li and Z. Wan. Supplier competition and cost improvement. *Management Science*, 63(8):2460–2477, 2017.

Y. Li and S. Gupta. Strategic capability investments and competition for supply contracts. *European Journal of Operational Research*, 214(2):273–283, 2011.

H. Lu and G. Pang. Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Mathematics of Operations Research*, 41(2):560–595, 2015.

H. Lu and G. Pang. Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stochastic Systems*, 6(2):519–600, 2017a.

H. Lu and G. Pang. Heavy-traffic limits for an infinite-server fork–join queueing system with dependent and disruptive services. *Queueing Systems*, 85(1-2):67–115, 2017b.

L. Lu, J.-S. Song, and H. Zhang. Optimal and asymptotically optimal policies for assemble-to-order n-and w-systems. *Naval Research Logistics*, 62(8):617–645, 2015.

V. Martínez-de Albéniz and A. Lago. Myopic inventory policies using individual customer arrival information. *Manufacturing & Service Operations Management*, 12 (4):663–672, 2010.

B. Masih-Tehrani, S. H. Xu, S. Kumara, and H. Li. A single-period analysis of a two-echelon inventory system with dependent supply uncertainty. *Transportation Research Part B: Methodological*, 45(8):1128–1151, 2011.

M. E. Mayorga and H.-S. Ahn. Joint management of capacity and inventory in make-to-stock production systems with multi-class demand. *European Journal of Operational Research*, 212(2):312–324, 2011.

M. S. Meijer, D. Schol, W. van Jaarsveld, M. Vlasiou, and B. Zwart. Extreme-value theory for large fork-join queues, with an application to high-tech supply chains. *arXiv preprint arXiv:2105.09189v2*, 2021a.

M. S. Meijer, W. van Jaarsveld, and T. de Kok. Contingent renewal contracts in high-tech manufacturing with oligopolistic suppliers. *Working paper*, 2021b.

M. S. Meijer, W. van Jaarsveld, and T. de Kok. Synchronization in a two-supplier assembly system: Combining a fixed lead-time module with capacitated make-to-order production. *arXiv preprint arXiv:2105.08991*, 2021c.

M. S. Meijer, W. van Jaarsveld, T. de Kok, and C. S. Tang. Direct versus indirect penalties for supply contracts in high-tech industry. *European Journal of Operational Research*, 2021d.

G. Merckx and A. Chaturvedi. Short vs. long-term procurement contracts when supplier can invest in cost reduction. *International Journal of Production Economics*, 227:107652, 2020.

E. Mohebbi. A production-inventory model with randomly changing environmental conditions. *European Journal of Operational Research*, 174(1):539–552, 2006.

A. Muharremoglu and J. N. Tsitsiklis. A single-unit decomposition approach to multiechelon inventory systems. *Operations Research*, 56(5):1089–1103, 2008.

J. Nair, A. Wierman, and B. Zwart. Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base. *Management Science*, 62(6):1830–1841, 2016.

S.-H. Nam, J. Vitton, and H. Kurata. Robust supply base management: Determining the optimal number of suppliers utilized by contractors. *International Journal of Production Economics*, 134(2):333–343, 2011.

R. Nelson and A. N. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37(6):739–743, 1988.

V. Nguyen. Processing networks with parallel and sequential tasks: Heavy traffic analysis and brownian limits. *The Annals of Applied Probability*, pages 28–55, 1993.

V. Nguyen. The trouble with diversity: Fork-join networks with heterogeneous customer population. *The Annals of Applied Probability*, pages 1–25, 1994.

D. Nolan. Is Boeing's 787 Dreamliner a Triumph or a Folly? `https://hbr.org/2009/12/is-boeings-787-dreamliner-a-tr`, 12 2009. Accessed: 2020-02-14.

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

T. Pfeiffer. A dynamic model of supplier switching. *European Journal of Operational*

Research, 207(2):697–710, 2010.

J. Pickands III. Moment convergence of sample extremes. *The Annals of Mathematical Statistics*, 39(3):881–889, 1968.

E. L. Plambeck. Asymptotically optimal control for an assemble-to-order system with capacitated component production and fixed transport costs. *Operations Research*, 56(5):1158–1171, 2008.

K. N. Reddy and A. Kumar. Capacity investment and inventory planning for a hybrid manufacturing-remanufacturing system in the circular economy. *International Journal of Production Research*, pages 1–29, 2020.

J. Reed and B. Zhang. Managing capacity and inventory jointly for multi-server make-to-stock queues. *Queueing Systems*, 86(1-2):61–94, 2017.

Z. J. Ren, M. A. Cohen, T. H. Ho, and C. Terwiesch. Information sharing in a long-term supply chain relationship: The role of customer review strategy. *Operations Research*, 58(1):81–93, 2010.

S. I. Resnick. *Extreme values, regular variation and point processes*. Springer, 1987.

Reuters. Boeing CEO blames industry for 787 bolt shortage. `https://www.reuters.com/article/us-boeing-alcoa/boeing-ceo-blames-industry-for-787-bolt-shortage-idUSN1141086620070911`, 2007. Accessed: 2020-02-17.

G. Roels and C. S. Tang. Win-win capacity allocation contracts in coproduction and codistribution alliances. *Management Science*, 63(3):861–881, 2017.

K. Rosling. Optimal inventory policies for assembly systems under random demands. *Operations Research*, 37(4):565–579, 1989.

C. P. Schmidt and S. Nahmias. Optimal policy for a two-stage assembly system under random demand. *Operations Research*, 33(5):1130–1145, 1985.

D. A. Serel. Capacity reservation under supply uncertainty. *Computers & Operations Research*, 34(4):1192–1220, 2007.

M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):

1095–1100, 1953.

M. A. Sieke, R. W. Seifert, and U. W. Thonemann. Designing service level contracts for supply chain coordination. *Production and Operations Management*, 21(4):698–714, 2012.

S. Sinha and S. P. Sarmah. Coordination and price competition in a duopoly common retailer supply chain. *Computers & Industrial Engineering*, 59(2):280–295, 2010.

A. Sleptchenko, M. C. van der Heijden, and A. van Harten. Trade-off between inventory and repair capacity in spare part networks. *Journal of the Operational Research Society*, 54(3):263–272, 2003.

J.-S. Song and P. Zipkin. Inventory control in a fluctuating demand environment. *Operations Research*, 41(2):351–370, 1993.

J.-S. Song, S. H. Xu, and B. Liu. Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes. *Operations Research*, 47(1): 131–149, 1999.

J. Sun and L. Debo. Sustaining long-term supply chain partnerships using price-only contracts. *European Journal of Operational Research*, 233(3):557–565, 2014.

J. Svoboda, S. Minner, and M. Yao. Typology and literature review on multiple supplier inventory control models. *European Journal of Operational Research*, 2020.

C. S. Tang and J. D. Zimmerman. Managing new product development and supply chain risks: The boeing 787 case. In *Supply Chain Forum: An International Journal*, volume 10, pages 74–86. Taylor & Francis, 2009.

T. A. Taylor and E. L. Plambeck. Simple relational contracts to motivate capacity investment: Price only vs. price and quantity. *Manufacturing & Service Operations Management*, 9(1):94–113, 2007a.

T. A. Taylor and E. L. Plambeck. Supply chain relationships and contracts: The impact of repeated interaction on capacity investment and procurement. *Management Science*, 53(10):1577–1593, 2007b.

P. Toktaş-Palut and F. Ülengin. Coordination in a two-stage capacitated supply chain with multiple suppliers. *European Journal of Operational Research*, 212(1): 43–53, 2011.

S. Transchel, S. Bansal, and M. Deb. Managing production of high-tech products with high production quality variability. *International Journal of Production Research*, 54(6):1689–1707, 2016.

J. N. Tsitsiklis and Y. Xu. Efficiency loss in a cournot oligopoly with convex market demand. *Journal of Mathematical Economics*, 53:46–58, 2014.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

J. S. van Leeuwaarden, B. W. Mathijsen, and B. Zwart. Economies-of-scale in many-server queueing systems: Tutorial and partial review of the qed halfin–whitt heavy-traffic regime. *SIAM Review*, 61(3):403–440, 2019.

S. Varma. *Heavy and light traffic approximations for queues with synchronization constraints*. PhD thesis, University of Maryland, 1990.

S. M. Wagner and G. Friedl. Supplier switching decisions. *European Journal of Operational Research*, 183(2):700–717, 2007.

A. Wilhite, L. Burns, R. Patnayakuni, and F. Tseng. Military supply chains and closed-loop systems: resource allocation and incentives in supply sourcing and supply chain design. *International Journal of Production Research*, 52(7):1926–1939, 2014.

P. E. Wright. Two parallel processors with coupled inputs. *Advances in Applied Probability*, 24(4):986–1007, 1992.

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

J. Wu, H. Wang, and J. Shang. Multi-sourcing and information sharing under competition and supply uncertainty. *European Journal of Operational Research*, 278 (2):658–671, 2019.

Z. Wu and T. Y. Choi. Supplier–supplier relationships in the buyer–supplier triad: Building theories from eight case studies. *Journal of Operations Management*, 24(1): 27–52, 2005.

Y. Xia. Competitive strategies and market segmentation for suppliers with substitutable products. *European Journal of Operational Research*, 210(2):194–203, 2011.

R. Zieliński. Optimal nonparametric quantile estimators. Towards a general theory. A survey. *Communications in Statistics-Theory and Methods*, 38(7):980–992, 2009.

X. Zou, S. Pokharel, and R. Piplani. Channel coordination in an assembly system facing uncertain demand with synchronized processing time and delivery quantity. *International Journal of Production Research*, 42(22):4673–4689, 2004.

Summary

Optimization and Coordination in High-tech Supply Chains

High-tech original equipment manufacturers (OEMs) produce and assemble state-of-the art products that consist of many complex components sourced at dozens of suppliers. Examples are the production of lithography machines by ASML or the production of aircrafts by Boeing or Airbus. To assemble these high-tech end-products and deliver them to the customers, it is important that all required components are available as shortage of a single component may lead to costly delivery delays of the end-product. This requires sufficient capacity at the suppliers, who face their own trade-offs and may be hesitant to invest in too much capacity. This thesis investigates multiple optimization and coordination problems in these high-tech supply chains.

High-tech supply chains have characteristics that are not included in classical supply chain models, e.g. long-term collaboration between a manufacturer and supplier that may be terminated based on performance, large-scale assembly systems with many components sourced at specialized suppliers, or single-sourced components that are produced only by a single supplier. We introduce and analyze models that incorporate and utilize these characteristics to derive results specifically for high-tech supply chains. We extend classical single-period supply chain models

to multi-period models where the collaboration between a manufacturer and a supplier continues depending on the supplier's performance. Next, we combine ideas from queuing theory and Markov processes to develop an optimal inventory policy for a module sourced at a supplier that is combined with an in-house produced make-to-order module. Furthermore, we use the scale of large high-tech assembly systems to find asymptotically optimal decisions in the joint inventory and capacity optimization problem.

Most existing supply chain models focus on a single interaction between a manufacturer and its supplier. In Chapter 2, we propose a model for the case where a manufacturer and supplier work together for longer periods, as is often the case in the high-tech industry. This research is inspired by discussions with supply chain experts from ASML and Philips and several of their suppliers. We examine supply contracts that are intended to align the incentives between a high-tech manufacturer and a supplier so that the supplier will invest adequate and yet non-verifiable capacity to meet the manufacturer's uncertain demand. By using non-renewal as an implicit penalty, we show that a contingent renewal contract encourages a supplier to create more capacity and that the supply chain optimal capacity investment can be reached with positive expected profits for both parties. Subsequently, in Chapter 3, we consider a setting where only a limited number of suppliers has the required capabilities to supply the manufacturer. Since high-tech manufacturers often source specialized components at their suppliers, this is very relevant in high-tech industries. We show that long-term relations contingent on supplier performance continue to be valuable, but the supplier has a stronger position when the number of suppliers with these capabilities is limited.

In Chapter 4, we study an assembly system with a combination of a fixed lead-time component sourced at an outside supplier and an in-house produced component that is produced once a customer order is placed. Since unavailability of one of the components has costly consequences for the production of the end-product, it is important to synchronize between the ordering policy for one component and the production of the other. We propose a state-dependent base-stock policy for ordering the fixed lead-time component from the supplier while taking into account the number of outstanding orders for the in-house produced component. We show optimality of this policy and verify numerically that it generates considerable savings compared to a static policy that disregards this information. These results hold both in continuous time and discrete time, allowing for application in many practical settings.

Since high-tech end-products often consist of many components that need to be available at the time of assembly, we finally study a large-scale assembly system in Chapter 5. When one component is missing, this leads to costly delays in production of the end-product. Shortages can occur for example when there are disruptions in the component production process or when a peak in demand occurs. Since the demand for all N components results from stochastic demand of the end-product, delays of the different components are correlated. We model this as N identical queues with independent service processes and a common arrival process. Random perturbations in the arrival and service processes are modeled with Brownian motions. We prove that as N goes to infinity the scaled maximum of N steady-state queue lengths converges in distribution to a normally distributed random variable. We explore repercussions of this result for high-tech manufacturers. The probability of delays occurring can be reduced by increasing capacity or keeping inventory, which both also have associated costs. We formulate a stylized model that enables us to study the resulting trade-off between shortage risk, inventory costs, and capacity costs. Our asymptotic extreme value results translate into various asymptotically exact methods for cost-optimal inventory and capacity decisions, some of which are in closed-form. Numerical results indicate that our methods are asymptotically exact, while for transient times their performance depends on model parameters. Based on a simulation study we conclude that, when the variation in demand is limited relative to the variation in the component production processes, the expected costs using our approximations are very close to the optimal costs for any realistic number of components.

Acknowledgments

Already during my bachelor studies I became interested in doing research and pursuing a PhD. Now, looking back at the past 4 years, I can definitely say that this was a great decision. During these years I have learnt a lot and met many great people. I am very thankful for the stimulating research environment at OPAC and everyone I could work with to realize this thesis. The last 1.5 years of my PhD were completely different from what I envisioned. Instead of working in the office and traveling abroad for presenting research at international conferences, we were all forced to work mostly from home and have meetings online. This definitely needed some getting used to, and I missed standing in front of a whiteboard to discuss a problem together and the coffee breaks and casual chats with colleagues. Nonetheless, I was happy I could continue my research and I enjoyed writing this thesis. I would like to express my thanks to the people who have contributed to this or have supported me otherwise.

First of all, I want to thank my supervisors Willem van Jaarsveld and Ton de Kok. My weekly meetings with Willem varied from short progress updates, to inspiring brainstorm sessions, to thoughtful discussion about how to finalize a proof. If I got stuck somewhere in the middle of a long derivation, your usual advise would be not to make things too complicated, which repeatedly turned out to be great advice. I really enjoyed our meetings and discussions about my research, the writing of our papers or anything else. Although my meetings with Ton were sometimes a little less frequent, they always generated a lot of ideas for framing and positioning our research or even completely new research directions. I really enjoyed working together with both of you and have learnt a lot during the past years.

Another big thank you goes to Chris Tang. It was a pleasure to work with you on our joint paper and Chapter 2 of this thesis. I learnt a lot from your ideas and experience during our meetings and discussions and am very happy to have you on my committee. It would have been a great opportunity to visit you at UCLA to continue our collaboration, but unfortunately Covid-19 had other plans. Hopefully, we will be able to work together again in the future.

Next, I would like to express my gratitude to Dennis Schol, Bert Zwart and Maria Vlasiou. Dennis, I really enjoyed collaborating with you on our joint paper, which resulted in Chapter 5 of this thesis. By combining our expertise we were able to obtain very nice results. Bert and Maria, the discussions we had during our meetings were very valuable to me, not only for improving our paper, but also for developing myself as a researcher. I really appreciated your feedback and guidance. I am very thankful that Bert agreed to be part of my committee.

I would also like to thank Sandra Transchel, Joachim Arts and Zümbül Atan for being on my thesis committee, for reading my thesis and providing valuable feedback. Sandra and I first met during the ISIR symposium in Budapest in August 2018. For me this was a great introduction to presenting my research to the international academic community and I really appreciated your interest in and our discussions about my research. I am very thankful that we met each other again at the ISIR summer school in Luxembourg in 2021, where we raised the idea of a visit to KLU. I am very pleased to have had, after finising my thesis, the opportunity to be part of the KLU community and to work together with you and Nina. I look forward to continuing this collaboration. At the same ISIR symposium in 2018, I was introduced to Joachim. Your work showed me once again that practical relevance and solving theoretically interesting challenges can go hand-in-hand and inspired me to strive for achieving that as well. Already when we were still working at the Paviljoen building, I enjoyed the morning coffee breaks with Zümbül. Thank you for always being available to give advice and answer questions and for your genuine interest not only in my research, but also in general. I am very happy to have you all on my committee and hope to collaborate with each of you in the future.

The consortium meetings of our project “Complexity in high-tech manufacturing” gave me important insights into the operations and problems faced in high-tech supply chains, which formed the basis for the research in this thesis. I would like to thank everyone involved for the insightful presentations and discussions. Especially, I would like to thank Celine Netten and Michel Vissers for hosting me at Philips in Best and for sharing their knowledge. Also, I would like to thank Fred Janssen and Paul Enders, for sharing insights on the production process of ASML.

Of course, I would like to thank all my other colleagues in the OPAC group. Claudine, Christel, Jolanda and José, thank you for helping out with any practical issue. Claudine, thank you for the great conversations we had during our coffee and tea breaks. It was always a pleasure to have a chat with you.

I would like to thank Loe for the thoughtful discussions about performing research, publishing papers and building an academic career, not only in the office, but also during our regular walks in the woods of Lunteren. When the LNMB conference in 2021 went online due to the pandemic, I really missed our morning or lunch walks. These moments were very useful for exchanging thoughts and in addition also much fun. Hopefully, we can continue this tradition in the future.

Without the company of my fellow PhD colleagues, the past years would have been a completely different experience. Thank you Yeşim for being a great office mate and friend. We started our journey around the same time in the Paviljoen, but shared much more than just an office. Luckily, we did not move too far apart once we moved to Atlas. Unfortunately, the Covid-19 pandemic separated us a bit further, but I really enjoyed the tea breaks that we did have together every now and then.

Simon, thank you for the endless coffee and tea breaks and for working together on the many LNMB assignments. It is always nice to discuss research and other fun stuff with you. Hopefully, we will still be able to have a coffee/tea break together every once in a while.

I would like to thank Joni and Sjors for introducing me to the habit of regular coffee breaks and for all the useful advice on doing a PhD and answering all my questions. Joni, your excellent drawing of the map of Eindhoven on your whiteboard to point out the nice places to live really helped me out in my search for housing.

Thank you Natasja for joining me in Eindhoven. After doing our bachelor together it was quite a shock to not see each other on a daily basis anymore, but luckily that was not for long. I really enjoyed our trip through Oregon together with Lotte after the INFORMS conference in Seattle in 2019. Hopefully, we can do this again soon. Until then, let us continue our regular dinners together with Wendy.

Of course, there are many more people that I would like to thank for our (sometimes a bit too) regular coffee breaks. The composition of our group changed over the years, but it was always a nice moment to relax and have a chat.

I would like to thank Karel and Tarkan for their great support during the teaching of the Stochastic Operations Management course over the past years. It was a pleasure to work with you and I learnt a lot from both of you. Also, a thank you to Volkan and Ragnar for the nice collaboration in preparing the instruction sessions.

Next to my colleagues, I would like to thank my friends. Aniek, Tom and again Natasja, we spread out over the country and sometimes we do not see each other for a while, but I always have a lot of fun when we do get together. I would also like to thank my rowing teammates, especially Emily. When I returned from studying together in the US, I definitely did not think we would ever be living in the same city again, let alone row our doubles together. I am very glad that you decided to move to Eindhoven and we could take out our boat to enjoy the calmness of the Eindhovens kanaal.

Finally, I would like to thank my family for all their love and support. Mom and dad, thank you for all your support and interest in my research. You have encouraged me to work hard and enjoy it, which is exactly what I did during the past years of my PhD research. Titia, as my big sister you have always been an example for me. Thank you for suggesting that I should study Econometrics in the first place, who knows where I would be without your advice. Finally, a thank you to my brother-in-law Tim, and my little nephew Krijn, who is always able to offer distraction and bring back my inner child while digging tunnels in the sand.

About the author

Mirjam Meijer was born in Zwolle, the Netherlands, on February 18, 1994. She finished her pre-university education at Gymnasium Celeanum in Zwolle in 2012. During the academic year 2012-2013, she studied at Augustana College, Rock Island, Illinois (USA), through the Campus Scholarship Program of the Fulbright Center. In 2016, she obtained a BSc in Econometrics and Operations Research from Erasmus University Rotterdam, the Netherlands, after spending a semester abroad at the Norwegian University of Science and Technology (NTNU) in the fall of 2015. In 2017, she obtained an MSc in Econometrics and Management Science (cum laude), with a specialization in Operations Research and Quantitative Logistics, also from Erasmus University Rotterdam, the Netherlands.

In October 2017, she started her PhD research within the Operations, Planning, Accounting and Control group at the School of Industrial Engineering at Eindhoven University of Technology. Under supervision of dr. Willem van Jaarsveld and prof. dr. Ton de Kok, she worked on several optimization and coordination problems in high-tech supply chains. During this research, she collaborated with prof. dr. Christopher Tang from UCLA and with Dennis Schol, prof. dr. Bert Zwart and prof. dr. Maria Vlasiou from the department of Mathematics at Eindhoven University of Technology. She was awarded a Fulbright Promovendi Scholarship for visiting prof. Tang at UCLA, but this visit was cancelled due to Covid-19.