



Robust Spare Parts Inventory Management

Zhao Kang

Robust Spare Parts Inventory Management

This thesis is part of the PhD thesis series of the Beta Research School for Operations Management and Logistics in which research groups participate of CWI, Eindhoven University of Technology, Ghent University, Hasselt University, KU Leuven, Maastricht University, Tilburg University, University of Antwerp, University of Twente, VU Amsterdam, VU Brussels, and Wageningen University and Research.

This work is part of the research program PrimaVera: Predictive maintenance for Very effective asset management with project number NWA.1160.18.238, which is supported by the Dutch Research Council (NWO).

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

ISBN: 978-94-6510-653-3

Printed by ProefschriftMaken || www.proefschriftmaken.nl

Cover design by Maaïke Disco

No part of this publication may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the author.



Robust Spare Parts Inventory Management

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr. S.K. Lenaerts, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op *vrijdag 23 mei 2025 om 11:00 uur*

door

Zhao Kang

geboren te *Shangluo, People's Republic of China*

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.ir. R.J.I. Basten
2^e promotor: prof.dr. A.G. de Kok
co-promotor: dr. A. Marandi
leden: prof.dr.ir. G.J.J.A.N. van Houtum
prof.dr.ir. J.J. Arts (Université du Luxembourg)
prof.dr. W. Wiesemann (Imperial College London)
prof.dr. J. Zhang (New York University)

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

Contents

1	Introduction	1
1.1	Background	2
1.2	Methodology	4
1.2.1	Current Practice in Spare Parts Inventory Control	4
1.2.2	Robust Optimization as an Alternative Approach	6
1.3	Research topics and contributions	12
1.4	Notation	15
1.5	Outline of the thesis	15
2	Robust spare parts inventory control with lost sales	17
2.1	Introduction	18
2.2	Problem Formulation	19
2.2.1	Problem Setting	19
2.2.2	Stochastic Optimization Model	20
2.2.3	Adaptive Robust Optimization Model	22
2.3	Solution Method	23
2.3.1	Existing Approximation Methods	24
2.3.2	Exact Method	25
2.3.3	New Approximation Algorithms	26
2.4	Numerical Experiments	32
2.4.1	Performance of Three Different Uncertainty Sets	33
2.4.2	Comparison of Different Solution Methods	36
2.4.3	Comparison of Stochastic and Robust Models	39

2.5	ASML Case Study	44
2.6	Conclusion	47
2.A	Appendix	49
2.A.1	Proof of Infeasibility of SA to Problem (2.3)	49
2.A.2	Proof of Theorem 2.1	50
2.A.3	Fourier-Motzkin Elimination in Robust Optimization	51
2.A.4	Performance of ConGA	53
2.A.5	The Branch-and-cut (B&C) Method	54
2.A.6	An Algorithm for the Extended Budget Uncertainty Set	57
2.A.7	The Data Filtering Process, Data Decomposing, and Data Processing for ASML Case Study	57
2.A.8	Analysis of ConGA and LES at ASML	58
2.A.9	Analysis of Hybrid Method at ASML	59
3	Robust spare parts inventory control with emergency shipments	65
3.1	Introduction	66
3.2	Problem Description	67
3.2.1	Deterministic Model	67
3.2.2	Stochastic Optimization Model	68
3.2.3	Adaptive Robust Optimization Models	69
3.3	Solution method for Problem (3.5)	71
3.3.1	Equivalent Reformulations	72
3.3.2	Approximation Method	78
3.4	Solution method for Problem (3.4)	80
3.5	Uncertainty Sets	81
3.5.1	Classical Uncertainty sets	81
3.5.2	Incorporating Initial Failure Rate (IFR)	82
3.6	ASML Case study	83
3.6.1	Data Preparation and Model Setup	84
3.6.2	Incorporating IFR into Uncertainty Set Construction	85
3.6.3	Comparison of Uncertainty Sets	87
3.6.4	Comparison of RO and SO models	88
3.6.5	Sensitivity Analysis	89
3.7	Conclusion	93
3.A	Appendix	95

3.A.1	Proof of Theorem 3.1	95
3.A.2	Illustrative Example of Problem-solving Strategy	96
3.A.3	Proof of Optimality for Problem (3.11)	97
3.A.4	Proof of Theorem 3.3	100
3.A.5	Historical Demand Data-based Uncertainty Set Construction .	101
3.A.6	Benefits of Incorporating the IFR	102
3.A.7	Sensitivity Analysis of t_i^{em} and c_i^{em} on Stock Levels	104
4	Robust spare parts inventory control with backorders	107
4.1	Introduction	108
4.2	Problem Formulation	109
4.2.1	Stochastic Optimization Model	109
4.2.2	Robust Optimization Model	110
4.3	Solution Method	112
4.4	Uncertainty Set	116
4.4.1	Classical Uncertainty Sets	116
4.4.2	Lead Time Shift Method	117
4.5	ASML Case study	121
4.6	Conclusion	124
4.A	Appendix	126
4.A.1	Proof of Theorem 4.1	126
4.A.2	Alternative Robust Optimization Models with Backorders . .	129
4.A.3	Comparing Problems (4.2) and (4.14): A Numerical Example .	131
4.A.4	An Algorithm for the Budget Uncertainty Set	131
5	Conclusions	133
5.1	Research Topics Revisited	133
5.2	Future Research Directions	135
	Bibliography	141
	Summary	149
	Acknowledgments	153
	About the author	155

Chapter 1

Introduction

In service industries that rely heavily on capital equipment, “time is money” is not only an aphorism but a reality. Even brief operational disruptions can cause substantial financial losses. For example, a six-hour production line shutdown at TSMC’s semiconductor foundries resulted in losses of up to \$60 million (Asia Financial, 2024), and unexpected downtimes cost Fortune Global 500 companies around 11% of their revenue, totaling nearly \$1.5 trillion (SIEMENS, 2023). Therefore, companies perform maintenance activities to prevent disruptions caused by equipment deterioration and to quickly mitigate the impact when such disruptions occur despite preventive measures.

Various parties can be responsible for maintenance activities and associated spare parts inventory management. Equipment users like Air Canada and KLM manage their own maintenance operations and hold \$168 million and \$314 million in maintenance inventories, respectively (Air Canada, 2024; KLM, 2024). Original Equipment Manufacturers (OEMs) like ASML provide maintenance services to their customers and maintain substantial spare parts inventories. In some cases, third-party service providers specialize in equipment maintenance across different manufacturers. Regardless of who performs the maintenance, these activities are highly dependent on the availability of spare parts. Thus, effective spare parts inventory control is crucial and is the focus of this thesis.

The remainder of this chapter is structured as follows. Section 1.1 introduces the background to spare parts inventory control and how it is linked to the practice at ASML. Section 1.2 discusses the methodologies used in this thesis. Section 1.3

describes the problems studied in this thesis and the scientific contribution. Section 1.4 introduces the notation used throughout the thesis. Section 1.5 outlines the thesis.

1.1. Background

Our research on robust *Spare parts inventory control* originates from our knowledge about the situation at ASML, a world-leading original equipment manufacturer that produces lithography systems. These systems are critical for the production of integrated circuits in the semiconductor industry, and their breakdown can result in losses of up to 72,000 Euros per hour (ASML, 2014). In practice, ASML is required to determine stock levels for over 2,000 components at the new product introduction (NPI) stage. The state-of-the-art spare parts inventory control at ASML closely resembles the stochastic optimization (SO) problem discussed by (Van Houtum and Kranenburg, 2015), which is implemented by assuming a Poisson demand process with estimated demand rates.

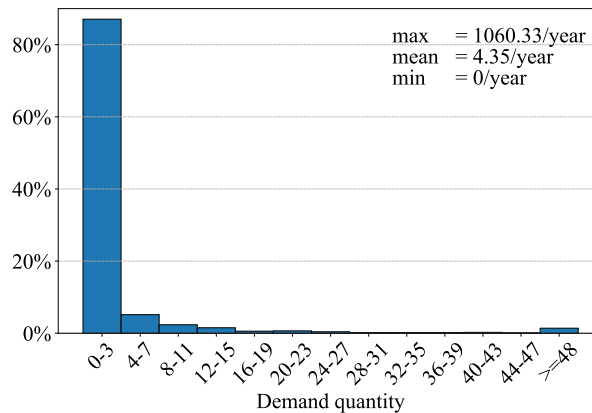


Figure 1.1: Distribution of average annual demand quantity over the initial three-year period for a specific machine type at ASML.

A primary challenge in spare parts inventory control is high demand uncertainty. Such uncertainty, for instance, arises during the early phases of the product life cycle, largely due to the scarcity of available data. Our analysis of ASML's data reveals that for a given machine type, over 80% of the components have an average annual demand of less than three units in the first three years of operation, as

shown in Figure 1.1. The lack of historical demand data is largely due to the long-lasting and reliable nature of the components, which translates into infrequent failures. Moreover, engineers tend to redesign components frequently in response to repeated failures, further aggravating the scarcity of consistent historical data. While the conventional SO approach can theoretically handle low demand patterns, obtaining reliable estimates for such low demand rates is particularly challenging. Additionally, demand uncertainty is complicated by the interdependence between different components. A failure in one component might increase the likelihood of failures in related components, creating complex demand correlations that are difficult to model using conventional approaches.

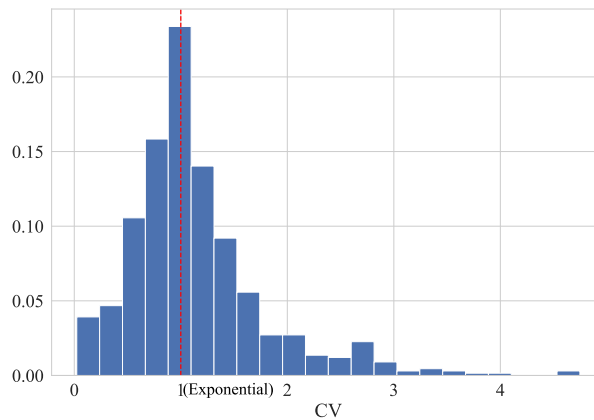


Figure 1.2: Distribution of CV values for components' demand inter-arrival times at ASML.

Also, the high variability in demand for components may make the SO approach less reliable. This variability is demonstrated in Figure 1.2 by showing the distribution of the Coefficient of Variation (CV) for demand inter-arrival times for all components of a type of ASML's machine. The CV is the ratio of the standard deviation to the mean of the demand inter-arrival times. For a Poisson arrival process, inter-arrival times follow an exponential distribution, where $CV = 1$ because the mean and standard deviation of exponentially distributed variables are identical. Our results show that 41.33% of components have $CV \geq 1.1$, indicating higher variability than a Poisson process, while 36.35% have $CV \leq 0.9$, indicating that actual demand is more concentrated around the mean. Notably, the maximum CV value observed is 4.75, indicating a substantial deviation from a Poisson process. In

the case of spare parts with a CV value higher than 1, the inventory management model employed by the company, which relies on a Poisson demand process, may not be appropriate for effectively managing the inventory of these spare parts.

Despite the uncertainty of demand, companies must maintain exceptionally high service performance due to the enormous costs of equipment downtime. High spare parts availability is crucial across industries, as evidenced in automotive (do Rego and De Mesquita, 2015) and textile manufacturing (Teunter et al., 2017). At ASML, 99% of the demands must be fulfilled in 24 hours due to the potential loss of millions in semiconductor production (ASML, 2025). This high service level requirement, combined with the high cost of components and the severe impact of stockouts, makes ASML an ideal case study for developing and testing new methods for spare parts inventory management.

Many organizations that provide worldwide maintenance services, such as ASML and other global equipment manufacturers, operate their spare parts inventory control through a multi-echelon network structure. At the top of the network, a central warehouse serves as the primary supply hub, replenishing inventory at local warehouses through regular shipments. These local warehouses, located closer to customer facilities, directly serve customer demand for spare parts. Customer facilities, such as semiconductor fabrication plants, aircraft maintenance bases, or medical centers, rely on rapid spare parts delivery to minimize equipment downtime. Understanding how these different warehouse types handle unfulfilled demand is crucial for developing effective inventory control, which we discuss in detail in Section 1.2.2.

1.2. Methodology

This section presents our methodological framework for robust spare parts inventory management. We begin by examining state-of-the-art academic approaches and current industry practices in Section 1.2.1, then introduce robust optimization as an alternative approach in Section 1.2.2.

1.2.1 Current Practice in Spare Parts Inventory Control

Spare parts inventory control has been a topic of interest in the literature for several decades. We refer the interested reader to Basten and van Houtum (2023) and Hu et al. (2018) for an overview of the methodologies developed in this area. In

his seminal work, Sherbrooke (1968) introduces the METRIC model for spare parts inventory control, which assumes that unmet demand is backordered. After that, many studies have recognized the importance of incorporating emergency shipments in spare parts inventory control to reduce stockout risks and improve system availability (e.g., Axsäter, 1990; Wong et al., 2006). These spare parts inventory control models typically employ an $(S - 1, S)$ policy, also known as the one-for-one replenishment policy or base stock policy (Feeney and Sherbrooke, 1966; Van Houtum and Kranenburg, 2015). This policy is preferred in the context of spare parts, especially for more expensive spare parts, for two reasons. First, it maintains minimal stock levels while ensuring immediate availability of spare parts, which is crucial for expensive spare parts where both holding and stockout costs are high. Second, since expensive spare parts typically have low demand patterns, the policy's one-for-one replenishment approach is more economical than batch ordering policies, where stockouts far exceed a single unit's holding cost.

Conventional approaches to handling the demand uncertainty rely on distribution fitting in stochastic inventory models, also known as stochastic optimization (SO) models, typically assuming demand follows a Poisson process (Sherbrooke, 1968; Basten and Van Houtum, 2014; Drent and Arts, 2021; Özkan and van Houtum, 2023). However, empirical evidence (Costantino et al., 2018; Turrini and Meissner, 2019) and our analysis of ASML data indicate that spare parts demand in the new product introduction phase does not always follow a Poisson process. Some researchers propose a Bayesian method for spare parts demand forecasting. However, this method is also typically based on the assumption of a (compound) Poisson process (Aronis et al., 2004; Babai et al., 2021). Furthermore, our analysis of ASML data demonstrates that in situations with very limited historical demand data, the Bayesian method is sensitive to the specification of the prior distribution, potentially leading to unreliable outcomes. Gelman et al. (2013, Chap. 7) similarly emphasize that with sparse data, poor choice of prior distribution can lead to weak inferences and poor predictions.

A greedy algorithm is commonly used in stochastic spare parts inventory control (Sherbrooke, 2006; Van Houtum and Kranenburg, 2015, Chap. 2). The algorithm works by iteratively selecting the alternative that offers the highest ratio of performance improvement to cost increase until reaching a feasible solution, i.e., one that achieves the target service level. In spare parts inventory control, this improvement is typically measured in terms of reduced waiting time or expected number of

backorders. Using numerical experiments, Wong et al. (2007) find that the greedy algorithm is more efficient in terms of computation time than the Lagrangian heuristic and Dantzig-Wolfe decomposition.

1.2.2 Robust Optimization as an Alternative Approach

This thesis proposes a robust optimization (RO) approach to handle demand uncertainty for spare parts inventory control. We choose RO for this thesis for three reasons.

- The RO approach does not require knowledge about the probability distribution of the uncertain parameters. As a non-probabilistic method for decision-making under uncertainty, the RO approach is more suitable for the initial stage of the product life cycle with limited demand data.
- The RO approach proves to be implementable in practice through appropriate reformulation techniques. Our research demonstrates that we can develop efficient solution methods for large-scale spare parts inventory problems using RO.
- The RO approach can capture demand correlations between components, accounting for the interdependence of failures, which the SO approach struggles to address.
- The decision-maker has the flexibility of choosing the robustness level to better manage the trade-off between solution robustness and performance.

Instead of using probability distributions, the RO solution safeguards against any realization of the uncertain demand in a given set, called the *uncertainty set*. RO is an effective approach for modeling uncertain demand in inventory control, though it has not yet been applied to spare parts settings. Early work by Bertsimas and Thiele (2006) applied RO to develop a periodic review inventory control model for a single warehouse as well as a tree-type divergent inventory network with backorders. Later, Bienstock and Özbay (2008) extend this work by considering the same single warehouse model while incorporating nonstationary costs for the stock. Ardestani-Jaafari and Delage (2016) further show that the approach of Bertsimas and Thiele (2006) is a conservative approximation and propose new approximations for robust optimization of sums of piecewise linear functions. In recent work, Chen

et al. (2023) consider both backlogging and lost sales for unfulfilled demand and propose a cycle-based single-item, single-warehouse inventory model. They use RO approaches and propose algorithms to approximate the optimal inventory. In addition to demand, Thorsen and Yao (2017) incorporate uncertain lead time in a periodic review model with backorders and introduce a central limit theorem-based polyhedral uncertainty set.

Our work differs from these existing robust optimization models. First, we focus specifically on spare parts inventory where stock levels are determined by both service level requirements (e.g., spare parts availability) and costs, rather than solely by costs. Second, while previous work commonly uses periodic review policies, we use a continuous review base stock policy, which is more appropriate for spare parts inventory control. These differences in both the operational context and control policy necessitate our novel modeling approach.

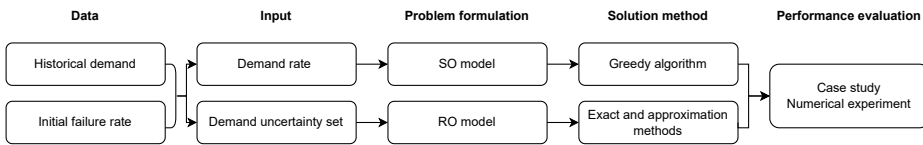


Figure 1.3: Overview of methodological framework comparing SO and RO approaches.

Figure 1.3 presents our methodological framework for applying robust optimization to spare parts inventory control. We use the conventional SO approach as a benchmark. To be able to apply this method to our setting, we first assume Poisson demand processes for all spare parts, then estimate the Poisson demand rates for different spare parts, and finally apply the SO model as described by (Van Houtum and Kranenburg, 2015, Chap. 2) to find optimal base stock levels. Therefore, we estimate the base stock levels by fitting an exponential distribution to the demand inter-arrival times of different components. As we demonstrate throughout this thesis, the conventional SO approach may face limitations when actual demand deviates from a Poisson process or when demand rates cannot be reliably estimated due to data scarcity, particularly during new product introductions. Furthermore, the fill rate calculation in the conventional SO approach is based on the steady state of an infinite horizon. However, we are interested in the fill rate calculation in a finite horizon. Therefore, the conventional SO approach approximates our studied

problem in this regard as well.

Our methodological framework consists of four main components: input, problem formulation, solution method, and performance evaluation. The remainder of this section details each component of this methodological framework.

Input

Unlike the SO model, which primarily relies on the estimated mean value of the demand rate as its key input, the quality of the RO solution heavily depends on the structure of the demand uncertainty set. In Chapter 2, we begin uncertainty set construction by using only historical demand data. We first consider a box uncertainty set that assumes independence between components. To capture the demand interaction between components, we then introduce a budget uncertainty set that adds constraints on the total demand across components. We extend it to an extended budget uncertainty set that captures demand relationships among all possible subsets of components.

When historical data is limited, particularly during the new product introduction stage, we incorporate *initial failure rate* (IFR) estimates from reliability engineers into our uncertainty sets in Chapter 3. The incorporation of IFR follows a dynamic weighting scheme that transitions from expert-based to data-driven estimates as more historical data becomes available. We then develop a price-based budget uncertainty set that focuses on demand interdependency between relatively expensive components while maintaining computational tractability.

To calculate the total demand within a given period for components with different lead times, we introduce a lead time shift method in Chapter 4. This method segments historical demand data into standardized time periods and provides a way to compute aggregate bounds for components with different lead times.

Problem formulation

Research on RO initially focused on static decision-making, which assumes that all variables, so-called *here-and-now*, are *independent* of the uncertain parameter and are chosen before the realization of the uncertain parameter. In other words, variables regarding stock levels and fill rates, i.e., the fraction of immediately fulfilled demand, are decided such that they are safeguarded against the demands in the uncertainty set, and they are independent of the realization of the demand for spare

parts. This approach, known as static robust optimization, can result in conservative solutions, as some variables, e.g., the fill rates, are *dependent* and decided after the realization of uncertain parameters in real-life situations. To address this limitation, Ben-Tal et al. (2004) introduce the adaptive robust optimization (ARO) method, where part of the variables are here-and-now, and the rest are *wait-and-see*, whose actual values depend on the realization of the uncertain parameter.

We develop spare parts inventory control models using ARO. Our ARO models follow the same setting as the conventional SO models to ensure fair comparison. We employ a two-stage decision-making process. In the first stage, prior to demand realization, we determine the stock levels as here-and-now variables. Because fill rates, i.e., the fraction of immediately fulfilled demand, depend on the realization of demand, we consider them as wait-and-see variables, which are decided in the second stage.

As introduced in Section 1.1, spare parts inventory control operates across both local and central warehouses. What mainly distinguishes these warehouse types is how they handle unfulfilled demand. At local warehouses, when the demand cannot be fulfilled immediately from stock, the unfulfilled demand is lost for the warehouse as service providers seek alternative solutions due to the urgency of their needs. We model this as a lost sales inventory control model in Chapter 2. In Chapter 3, we explicitly incorporate that if demand is lost, high costs are incurred as a costly emergency shipment from another local warehouse or the central warehouse is performed. At the central warehouse, investigated in Chapter 4, the dynamics change fundamentally. When central warehouses cannot fulfill demand, it typically results in backorders as there are no alternative quick-supply sources.

Solution method

This thesis focuses on how to solve the ARO problem. Solving ARO problems is generally computationally challenging. Reformulation and cutting-plane methods are two primary approaches for solving RO problems. Bertsimas et al. (2016) show that the effectiveness of these methods varies depending on the problem type and uncertainty set. In the realm of cutting-plane methods, researchers have applied various techniques to specific problems. For instance, in the context of robust vehicle routing problems, Chen et al. (2016) use a branch-and-cut method, while Pessoa et al. (2021) use a branch-cut-and-price method. Turning to reformulation

techniques, for general linear ARO problems involving continuous wait-and-see decisions, Bertsimas and De Ruiter (2016) propose a dual reformulation method. Moreover, Zhen et al. (2018) introduce Fourier-Motzkin elimination (FME), which can reformulate a class of ARO problems into equivalent counterparts with fewer wait-and-see variables at the expense of an increase in the number of constraints. Nevertheless, reformulating an ARO problem does not always lead to a tractable optimization problem (Bertsimas and De Ruiter, 2016).

To tackle the computational intractability of ARO problems, various efficient methods have been proposed in the literature to approximate solutions. The static RO solution is often used to approximate the ARO solution since it requires fewer computational resources (Lim and Wang, 2017). Previous studies have shown that under some conditions, the optimal objective values of the static RO problem and the adjustable variant are equivalent (Marandi and Den Hertog, 2018), while the static solution can be far from optimal for general problems (Bertsimas et al., 2011). An extension to the static RO approximation is to consider the wait-and-see decisions to be affine instead of constant, leading to affinely adaptive RO (Ben-Tal et al., 2004). According to El Housni and Goyal (2021), this approach provides a tight approximation for problems with fixed recourse and right-hand-side uncertainty using a specific type of uncertainty set. Another extension that can improve the quality of the affine policy is the use of piece-wise constant decision rules, known as the finitely adaptive method, where the uncertainty set is partitioned into smaller subsets (Bertsimas and Caramanis, 2010).

In this thesis, we employ two main approaches to find the exact solution to the ARO problem. The first approach involves reformulation techniques where we eliminate wait-and-see variables using Fourier-Motzkin elimination (Zhen et al., 2018), transforming the ARO problem into a deterministic mixed-integer optimization problem. While this reformulated problem contains an exponential number of constraints, it reveals important structural properties of the optimal solution. The second approach leverages decomposition. Specifically, in Chapter 3, we show that the deterministic counterpart can be decomposed into two independent mixed-integer optimization problems, which significantly reduces the computational complexity.

For large-scale ARO problems, we consider both existing approximation methods and develop new algorithms. Among the existing approaches, two main methods

are considered: static approximation (SA), which treats wait-and-see variables as here-and-now variables, and affine decision rule (ADR), which restricts wait-and-see variables to be affine functions of uncertain parameters. However, in Chapter 2, we prove that SA often leads to infeasible solutions for our problem with lost sales and show that ADR faces computational limitations for large instances. To address these limitations, we propose new algorithms that show superior performance in both solution quality and computational efficiency compared to existing methods, especially for problems involving hundreds or thousands of components.

Performance evaluation

We evaluate our proposed approaches through comprehensive numerical experiments in Chapter 2 and real-world case studies at ASML in Chapters 2, 3, and 4.

In Chapter 2, we first conduct numerical experiments to understand how different uncertainty sets affect the quality of the solution. We then evaluate our lost sales model and solution methods using various uncertainty sets with both Poissonian and non-Poissonian demand. The ASML case study with 710 components validates the effectiveness of our method in handling real-world complexity, particularly by comparing the performance of different uncertainty sets and solution methods.

In Chapter 3, we focus on evaluating the RO method during the critical new product introduction phase through an extensive ASML case study with more than 2,400 components. We examine how incorporating initial failure rate estimates affects model performance and investigate the inventory model's sensitivity to key operational parameters such as emergency shipment times and costs. This case study demonstrates the practical applicability of the RO model in high-stakes industrial settings where emergency shipments are crucial.

In Chapter 4, we evaluate the performance of the models to optimize the central warehouse inventory through an ASML case study. We consider nearly 1,600 components and examine how our lead time shift method performs under different historical data availability scenarios. We evaluate model performance across multiple service level targets, analyzing the trade-offs between stimulated fill rates and costs.

1.3. Research topics and contributions

This thesis addresses three research topics in robust spare parts inventory management, each focusing on different warehouse settings and demand fulfillment strategies when stockouts occur. We also discuss the contribution of this thesis by summarizing the main insights of each core chapter.

Research topic 1: Robust spare parts inventory control with lost sales at the local warehouse.

We focus on this research topic in Chapter 2. When immediate fulfillment is impossible due to long lead times, demand is considered lost from that warehouse. We propose an ARO model for multi-component, single-location spare parts inventory control. Since ARO problems are generally computationally difficult to solve, we show how the ARO problem can be reformulated into a deterministic counterpart. Despite the fact that the deterministic counterpart has an exponential number of constraints, it interestingly allows us to prove the structure of the optimal solution.

Leveraging the structure of the optimal solution, we develop an algorithm called *iterative projection in descending order* (IPDO). We show that IPDO can obtain an optimal solution under some conditions, largely dependent on the uncertainty set. To enhance IPDO's efficiency for large-scale problems, we develop a preprocessing step that provides a high-quality initial estimation of stock levels for certain products, enabling faster convergence to near-optimal solutions.

While IPDO is much faster than most existing methods for solving the ARO model, it still requires extensive computational resources for large-scale inventory problems. To address this, we design two heuristic algorithms based on IPDO's foundation. The first, *constraint generation* (ConGA), efficiently finds near-optimal solutions. The second, *linear equation system* (LES), is a highly efficient heuristic capable of handling hundreds of products within seconds. However, its accuracy depends on the type of uncertainty sets used. Therefore, we develop a hybrid method that dynamically combines ConGA and LES. This hybrid approach offers the flexibility to balance solution quality and computational efficiency by selectively applying ConGA and LES to different uncertainty sets, making it particularly valuable for large-scale inventory problems.

Our numerical experiments reveal that our algorithms surpass our branch-and-cut method and existing approximation methods in terms of both accuracy and

computational efficiency, demonstrating their superiority using various uncertainty sets.

We perform simulation-based experiments with both Poissonian and non-Poissonian demand, along with a case study at ASML with 710 components, to compare the performance of our ARO model with the conventional Poisson-based model. The experimental results demonstrate that our method is more reliable in achieving the target fill rate than the conventional stochastic optimization model when dealing with non-Poissonian demand.

Research topic 2: Robust spare parts inventory control with emergency shipments at the local warehouse.

We focus on this research topic in Chapter 3, which extends our work in Chapter 2 by incorporating emergency shipments as a costly but necessary backup option to achieve the target service performance. This research topic places particular emphasis on handling extreme demand uncertainty during the *new product introduction* (NPI) stage when historical demand data is scarce.

To ensure computational tractability, we reformulate the ARO model as a deterministic counterpart and prove that it can be approximated and decomposed into two mixed-integer optimization problems, drastically reducing the computational complexity. We then develop an efficient algorithm to obtain near-optimal solutions for large-scale problems with thousands of components.

When historical demand data are limited, we propose a phased approach that incorporates the IFR estimated by reliability engineers into the construction of uncertainty sets. In the initial phase, when demand data is extremely scarce, the phased approach constructs the uncertainty set solely based on the IFR. As more historical demand data becomes available, the weight of the IFR is gradually reduced. This approach addresses the challenge of constructing uncertainty sets for spare parts demand when historical data are limited and provides a structured framework throughout the product lifecycle. Our study shows that the phased approach improves model performance during the NPI stage compared to relying solely on historical demand.

The ASML case study demonstrates that using the ARO model consistently results in shorter simulated waiting times and lower costs than ASML's state-of-the-art stochastic optimization model. For the same simulated total cost, the robust solution achieves a simulated mean waiting time up to 3.5 hours shorter than the

stochastic model, potentially saving over €250,000 in lost production per breakdown of an expensive lithography system. Moreover, the sensitivity analysis demonstrates the strong adaptability of the ARO model to variations in key parameters, such as emergency shipment times, costs, and the exact way in which the IFR is incorporated into the uncertainty set. The ARO model can provide robust and economically viable solutions, demonstrating its superiority in dealing with uncertainty.

Research topic 3: Robust spare parts inventory control with backorders at the central warehouse.

We focus on this research topic in Chapter 4, addressing the fundamentally different dynamics of central warehouses where emergency shipments are impossible and stockouts result in backorders. The central warehouse sources directly from suppliers and serves as the emergency shipment source for local warehouses.

We propose an ARO model for the central warehouse. To the best of our knowledge, we are the first to use robust optimization to formulate a continuous review inventory model with backorders. To solve the ARO problem, we first reformulate it as its deterministic counterpart. We then develop a three-step solution approach. The first step establishes solution bounds by deriving an approximate lower bound through a lost sales problem (Chapter 2) and an upper bound through conservative estimation based on worst-case demand scenarios. This preprocessing step proves highly efficient, allowing us to immediately identify near-optimal stock levels for approximately 90% of components where the bounds coincide in our case study at ASML. We proceed to the second step for components where the bounds differ, introducing a tighter upper bound through a relaxed version of the ARO problem. This relaxed problem shares structural similarities with the lost sales problem from Chapter 2, allowing us to apply the IPDO and ConGA algorithms developed earlier in this thesis. Finally, in the third step, we employ additional approximation methods to determine near-optimal solutions for the remaining components where the bounds differ after the second step.

Unlike research topics 1 and 2, where we assume identical repair lead times for all SKUs, in this research topic, we extend our analysis to a more realistic case where each SKU may have a different repair lead time. This extension requires modifications to our modeling approach, particularly in the aggregate fill rate constraint. To address varying lead times for different components in the construction of the uncertainty set, we introduce a *lead time shift* method. This method segments

historical demand data into standardized time periods that account for the lead time differences, enabling us to compute meaningful aggregate demand bounds across components.

To demonstrate the applicability of our model, we conduct a case study at ASML. The results show that our robust model demonstrates superior cost efficiency compared to the SO model, particularly at very high service level targets.

This thesis advances both theoretical and practical aspects of spare parts inventory management through these three topics. From a theoretical perspective, we are the first to propose a robust optimization approach in spare parts inventory control. We develop ARO models that progress from lost sales to complex emergency shipment and backorder settings. We contribute new solution methods that make these models computationally tractable for real-world applications involving thousands of components. On the practical side, extensive case studies at ASML demonstrate substantial improvements over conventional methods, with potential savings of hundreds of thousands of euros per breakdown.

1.4. Notation

Throughout this thesis, we use \mathbb{Z} to denote the set of integers, \mathbb{N}_0 to denote the set of non-negative integers, and \mathbb{R} to denote the set of real numbers. Vectors and matrices are represented by boldfaced characters. For a vector $\mathbf{a} \in \mathbb{R}^n$, we use $[a_i]_{i=1,\dots,n}$ to denote its elements. We use \emptyset to denote the empty set and $\lceil \cdot \rceil$ to denote the ceiling function.

1.5. Outline of the thesis

The remainder of this thesis is organized as follows. Chapters 2 and 3 consider spare parts inventory control at a local warehouse, with Chapter 2 focusing on lost sales and Chapter 3 on emergency shipments. Chapter 4 considers spare parts inventory control at a central warehouse with backorders. The thesis concludes with a summary of our main findings and directions for future research in Chapter 5.

Chapter 2

Robust spare parts inventory control with lost sales

In this chapter, we propose an adaptive robust optimization (ARO) approach for multi-item, single-location spare parts inventory control with lost sales. We reformulate the ARO problem as a deterministic integer optimization problem with an exponential number of constraints. To solve this problem, we first introduce the iterative projection in descending order algorithm, which efficiently provides optimal solutions under certain conditions. Recognizing that only a few constraints are active in an optimal solution, we introduce the more time-efficient constraint generation algorithm (ConGA) and the linear equation system (LES) heuristic. To leverage the strengths of both methods, we propose a hybrid approach that combines ConGA and LES, enhancing performance across various uncertainty sets for large-scale problems involving hundreds of products.

Comprehensive simulation-based experiments with non-Poissonian demand demonstrate that our ARO model outperforms the implementation of the conventional SO model, achieving higher fill rates at lower holding costs. We validate our model through a case study of 710 products at ASML, a world-leading lithography machine manufacturer. The results confirm our model's cost-effectiveness in meeting target service performance, making it attractive for ASML and other expensive equipment maintenance service providers.

2.1. Introduction

Spare parts inventory control is not merely a distant industrial concern. Its impact resonates throughout our daily lives. For instance, recent spare parts shortages have forced major airlines like Lufthansa, Qatar, and Silver to ground planes and urgently request increased production from suppliers (Business Insider, 2022). Effective spare parts inventory control can mitigate such losses significantly, as evidenced by TSMC's swift 6-hour production line recovery in 2024 (Asia Financial, 2024).

In this chapter, we focus on local warehouses where unfulfilled demand results in lost sales, as service providers must seek alternative solutions due to urgency. We propose an ARO model that employs a two-stage decision-making process to handle demand uncertainty. We consider stock levels as here-and-now variables and fill rates as wait-and-see variables. As the quality of the solution highly depends on the selection of an uncertainty set, we consider three types of uncertainty sets to manage the trade-off between solution robustness and costs.

The key focus of this chapter is on developing computationally efficient methods for solving the ARO model in practical settings. While ARO problems are generally difficult to solve, we show how our problem can be reformulated into a deterministic counterpart despite having an exponential number of constraints. This reformulation reveals important structural properties of the optimal solution. Based on these properties, we develop the IPDO algorithm that can find optimal solutions under certain conditions. For large-scale problems where IPDO may be computationally intensive, we introduce two efficient heuristic algorithms. We introduce ConGA to find near-optimal solutions and LES to rapidly process hundreds of products. We then combine these approaches in a hybrid method that balances solution quality with computational speed. Through extensive numerical experiments and an ASML case study, we demonstrate that our approach consistently outperforms the conventional SO approach, which relies heavily on estimated demand rates and the assumption that demand for each spare part follows a Poisson process. The superiority of our approach in both computational efficiency and solution quality is particularly evident when demand processes are non-Poissonian or when the demand rate is uncertain.

The rest of this chapter is organized as follows. In Section 2.2, we describe a spare part inventory problem, show the existing model with which we compare later,

and develop the ARO model. In Section 2.3, we present the solution methods for the proposed ARO model. In Section 2.4, we conduct numerical experiments to show the applicability of our proposed model and algorithms. More specifically, in Section 2.4.1, we show how the choice of the uncertainty set affects the conservativeness of the solution. Section 2.4.2 contains the numerical experiments that show the efficiency of our algorithms in comparison with existing approaches. Section 2.4.3 contains the simulation-based experiments that identify under which conditions the use of RO is beneficial compared to the stochastic model. Section 2.5 contains a case study at ASML, where we show a significant cost saving achieved by our model and algorithms compared to ASML's current stochastic optimization approach. Finally, we conclude the chapter in Section 2.6.

2.2. Problem Formulation

In this section, we first introduce the general problem setting, and then we present two mathematical formulations of a spare parts inventory control problem: a stochastic optimization model and an adaptive robust optimization model. These models differ in their assumptions about the demand process.

2.2.1 Problem Setting

Consider a single warehouse that stocks spare parts for multiple types of critical components, thus servicing an installed base of machines of one type. If a critical component in a machine fails, the machine goes down. We refer to each distinct type of critical component as a stock keeping unit (SKU).

Let us denote the set of SKUs by $\mathcal{I} = \{1, \dots, n\}$. Demand for an SKU arises due to a component's failure. The defective component is sent for repair immediately and returns to stock in an as-good-as-new state after a constant repair lead time $t (> 0)$. Although we use repair terminology throughout the chapter, everything remains analogous in the model if spare parts are not repairable, i.e., if they are consumable. In that case, a defective spare part is discarded, and a new one is acquired. The repair lead time is then replaced by the order-and-ship time for the new part. Due to the relatively long lead times in practical situations, we assume that when the demand cannot be fulfilled from stock immediately, it is lost from the normal replenishment system of the local warehouse. This mirrors the operational strategies employed by companies like ASML, where high machine downtime costs

mean that unfulfilled demand is satisfied by other warehouses, resulting in lost demand at the considered warehouse (Van Wingerden et al., 2019; Lamghari-Idrissi et al., 2022).

2

For each SKU, the stock is controlled by a continuous review basestock policy. This implies that the inventory position of a given SKU $i \in \mathcal{I}$ remains constant at the basestock level $S_i (\geq 0)$. We use \mathbf{S} to denote the vector containing the basestock levels for all SKUs. The total holding cost per time unit for spare parts is calculated as $\sum_{i \in \mathcal{I}} c_i^h S_i$, where $c_i^h (> 0)$ represents the inventory holding cost per time unit for SKU i . In practice, the annual holding cost per SKU is typically set at about 20% of the price of a new spare part. The fill rate $\beta_i (\geq 0)$ is the fraction of demand satisfied immediately from stock. The calculation of β_i varies depending on the specific model employed. We consider a finite horizon and are interested in β_i calculated by the ratio of fulfilled demand to total demand over a specific time interval (e.g., a planning period). The aggregate fill rate, defined as the proportion of demand for all SKUs immediately filled from stock in a finite horizon should be at least equal to the target fill rate β^{obj} , which is typically high (≥ 0.9) in practice. We focus on the initial supply and inventory planning during the exploitation phase. To ensure service performance, companies typically specify service level agreements in customer contracts. Consequently, inventory planners use β^{obj} to achieve the target performance levels. To be consistent with industry practices, we use the constraint containing β^{obj} to achieve service performance.

2.2.2 Stochastic Optimization Model

We use the stochastic optimization (SO) method as a benchmark. This conventional spare parts inventory control model, as in Chapter 2 of the book by Van Houtum and Kranenburg (2015), is widely used in the literature (Thonemann et al., 2002; Caglar et al., 2004; Drent and Arts, 2021).

Despite our setting being a finite horizon, the SO approach considers an infinite horizon. The demand of SKU i per lead time follows a Poisson distribution with a constant rate of $m_i (> 0)$ per time unit in an infinite horizon. The behavior of spare parts in repair for SKU i is modeled as the number of customers in an $M|G|c|c$ queue with $c = S_i$ parallel servers, arrival rate m_i , and lead time t , which is also known as the *Erlang loss system*. The achieved fill rate for SKU i based on the Erlang

loss function is defined as:

$$\beta_i = 1 - \frac{\frac{1}{S_i!} (m_i t)^{S_i}}{\sum_{j=0}^{S_i} \frac{1}{j!} (m_i t)^j},$$

which is often referred to as the long run fill rate. The total demand rate for all SKUs together is denoted by $M = \sum_{i \in \mathcal{I}} m_i$. The aggregate fill rate $\beta (> 0)$ is given by:

$$\beta = \sum_{i \in \mathcal{I}} \frac{m_i}{M} \beta_i.$$

We consider S_i to be the main decision variable and β_i to be an auxiliary variable (since its value depends on S_i and m_i). To find the optimal stock level, we solve the following Problem (2.1):

$$\min_{\substack{S_i \in \mathbb{N}_0^n \\ \beta_i \in \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i \quad (2.1a)$$

$$\text{s.t. } \beta_i = 1 - \frac{\frac{1}{S_i!} (m_i t)^{S_i}}{\sum_{j=0}^{S_i} \frac{1}{j!} (m_i t)^j}, \quad \forall i \in \mathcal{I}, \quad (2.1b)$$

$$\sum_{i \in \mathcal{I}} \frac{m_i}{M} \beta_i \geq \beta^{\text{obj}}. \quad (2.1c)$$

The objective function (2.1a) minimizes the total holding costs per time unit, which is equivalent to minimizing the total cost of initial investment, assuming that both types of costs are linear in the number of spare parts. Constraints (2.1b) show the calculation of item fill rates where the demand for spare parts follows a Poisson process. Constraint (2.1c) ensures that the aggregate fill rate for all SKUs is higher than the target service level β^{obj} , which restricts the items, and thus leads to joint control, a key setting in a large stream of literature on spare parts inventory control. Using this joint control setting, which is called a *system approach*, to stock spare parts, leads to much lower costs than using an item approach (Thonemann et al., 2002).

The implementation of the SO model in practice faces fundamental limitations when actual demand deviates from a Poisson process or when the demand rate cannot be reliably estimated due to data scarcity.

2.2.3 Adaptive Robust Optimization Model

Since the demand $\zeta_i (\geq 0)$ during the lead time for SKU $i \in \mathcal{I}$ is uncertain, in ARO, we assume that the only available information about demand vector ζ is that it lies in the set $\mathcal{D} \subset \mathbb{R}^n$. In the first stage, the decision about stock level S_i for SKU $i \in \mathcal{I}$ should be made, which is a here-and-now variable. Because the auxiliary variable β_i depends on the realization of demand ζ_i during the lead time t , we reformulate it as a wait-and-see variable, which is decided in the second stage. Unlike the SO model, which considers an infinite horizon, in the RO model we calculate β_i only focusing on the lead time with the worst-case demand.

So, the ARO model for this Problem (2.2) reads:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i \quad (2.2a)$$

$$\text{s.t. } \beta_i(\zeta) \zeta_i \leq S_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (2.2b)$$

$$\sum_{i \in \mathcal{I}} \beta_i(\zeta) \zeta_i \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \quad (2.2c)$$

$$1 \geq \beta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}. \quad (2.2d)$$

Constraint (2.2b) guarantees that for all possible demand arrival sequences during the lead time, the actual fill rate is always greater than or equal to β_i . Constraint (2.2c) ensures that the aggregate fill rate for all SKUs is higher than the target service level β^{obj} . Constraint (2.2d) limits the range interval of an item fill rate.

Demand Uncertainty Sets

As known from the literature on RO (see, e.g., Bertsimas and den Hertog, 2022), the choice of uncertainty set is crucial in obtaining a practical inventory solution for the ARO model. We study three types of uncertainty sets, each reflecting different assumptions about demand interdependence. Let $\bar{\mathbf{d}}, \underline{\mathbf{d}} \in \mathbb{R}^n$, $\bar{\boldsymbol{\Theta}}, \underline{\boldsymbol{\Theta}} \in \mathbb{R}$, and $\bar{\boldsymbol{\Gamma}}, \underline{\boldsymbol{\Gamma}} \in \mathbb{R}^{2^n - 1}$, such that $\bar{\mathbf{d}} \geq \underline{\mathbf{d}}$, $\bar{\boldsymbol{\Theta}} \geq \underline{\boldsymbol{\Theta}}$, and $\bar{\boldsymbol{\Gamma}} \geq \underline{\boldsymbol{\Gamma}}$.

A basic uncertainty set, where we assume no interactions exist between demands for different SKUs, can be formulated as a box uncertainty set:

$$\mathcal{D}^{\text{box}} = \{\zeta \in \mathbb{R}^n : \underline{d}_i \leq \zeta_i \leq \bar{d}_i, \forall i \in \mathcal{I}\}.$$

Box uncertainty sets lead to problems that are computationally less challenging (Marandi and Den Hertog, 2018; Bertsimas and den Hertog, 2022). This uncertainty

set is used when spare parts serve multiple unrelated systems or when demand correlations are minimal. However, it is rare in practice that all SKUs have independent demand. To address the limitations of the box uncertainty set, we consider a budget uncertainty set:

$$\mathcal{D}^{\text{bud}} = \{\zeta \in \mathbb{R}^n : \underline{d}_i \leq \zeta_i \leq \bar{d}_i, \forall i \in \mathcal{I}, \underline{\Theta} \leq \sum_{i \in \mathcal{I}} \zeta_i \leq \bar{\Theta}\}.$$

Notice that $\mathcal{D}^{\text{bud}} \subseteq \mathcal{D}^{\text{box}}$. In \mathcal{D}^{bud} , we exclude scenarios where all SKUs reach their upper or lower bounds simultaneously, in order to reduce the conservativeness of the robust solution. However, demand interactions often exist among specific subsets of SKUs, rather than across all SKUs. The budget uncertainty set cannot capture these nuanced relationships. Therefore, we consider an extended budget uncertainty set:

$$\mathcal{D}^{\text{ext}} = \{\zeta \in \mathbb{R}^n : \underline{\Gamma}_\alpha \leq \sum_{i \in \alpha} \zeta_i \leq \bar{\Gamma}_\alpha, \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset\}.$$

Here, for any $i \in \mathcal{I}$, we have $\underline{\Gamma}_{\{i\}} = \underline{d}_i$, $\bar{\Gamma}_{\{i\}} = \bar{d}_i$, and $\underline{\Gamma}_{\mathcal{I}} = \underline{\Theta}$, $\bar{\Gamma}_{\mathcal{I}} = \bar{\Theta}$. So, $\mathcal{D}^{\text{ext}} \subseteq \mathcal{D}^{\text{bud}}$. In this uncertainty set, we add a constraint for every combination of SKUs. If there are no interactions between demands for different SKUs, these additional constraints become redundant. However, when component failures exhibit correlations, constraints can exclude unrealistic scenarios. This set is valuable for industries with complex interrelationships between component failures.

2.3. Solution Method

For the SO Problem (2.1), a greedy algorithm is commonly used in spare parts inventory control (Sherbrooke, 2006; Van Houtum and Kranenburg, 2015, Chap. 2).

Our main focus in this section is on how to solve the ARO Problem (2.2). Since most of the algorithms to solve an ARO problem are developed for fixed-recourse problems, we first reformulate Problem (2.2) into a fixed-recourse ARO problem. For a given $i \in \mathcal{I}$ and $\zeta \in \mathcal{D}$, we define $\epsilon_i(\zeta) := \beta_i(\zeta)\zeta_i$. Now, Problem (2.2) can be

reformulated to become Problem (2.3):

$$\begin{aligned}
 \min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^n \rightarrow \mathbb{R}}} & \sum_{i \in \mathcal{I}} c_i^h S_i \\
 \text{s.t.} & \epsilon_i(\zeta) \leq S_i, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
 & \sum_{i \in \mathcal{I}} \epsilon_i(\zeta) \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, & \forall \zeta \in \mathcal{D}, \\
 & \zeta_i \geq \epsilon_i(\zeta) \geq 0, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}.
 \end{aligned} \tag{2.3}$$

To solve Problem (2.3), we begin by discussing two existing approximation methods. Subsequently, we propose a reformulation of Problem (2.3) to obtain an exact solution. To enhance the computational efficiency of solving the reformulated problem beyond the capabilities of conventional solvers, we develop two dedicated algorithms.

2.3.1 Existing Approximation Methods

In this section, we discuss two existing approximation methods for solving Problem (2.3): the static approximation approach (SA) and the affine decision rule approach (ADR).

Static approximation: By treating $\epsilon_i(\zeta)$ as a here-and-now variable, we can approximate the reformulated Problem (2.3) as a static RO model. However, applying SA to the reformulated Problem (2.3) most likely results in an optimization problem that is infeasible (see Appendix 2.A.1 for the proof). More generally, given a partitioning $\mathcal{D} = \cup_{\ell=1}^L \mathcal{D}^\ell$, applying any piece-wise constant decision rule to Problem (2.3), where $\epsilon_i(\zeta)$ is approximated by a piece-wise constant function

$$\epsilon_i(\zeta) \approx \epsilon_i^\ell, \quad \text{if } \zeta \in \mathcal{D}^\ell,$$

we need to have very small partitions to have a feasible approximation. This implies the impracticality of using piece-wise constant decision rules to approximate Problem (2.3). Therefore, in the numerical experiment, we only evaluate the performance of SA when applied to approximate the original Problem (2.2).

Affine decision rule: The ADR is a popular approximation approach for ARO problems with fixed recourses. To apply ADR, we restrict $\epsilon_i(\zeta)$ to be affine: i.e., $\epsilon_i(\zeta) = V^i \cdot \zeta + u^i$, where $V^i \in \mathbb{R}^n$ and $u^i \in \mathbb{R}$, for any $i \in \mathcal{I}$.

In the numerical experiment, we compare the performance of these existing approximations with the newly developed ones presented in Section 2.3.3.

2.3.2 Exact Method

To solve Problem (2.3), we can eliminate the wait-and-see variables and cast the problem as a static RO problem. Theorem 2.1 shows the robust reformulation of Problem (2.3) after eliminating all wait-and-see variables.

Theorem 2.1 *The ARO Problem (2.3) is equivalent to the following static RO problem:*

$$\begin{aligned} \min_{S \in \mathbb{N}_0^n} \quad & \sum_{i \in \mathcal{I}} c_i^h S_i \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I} \setminus \alpha} S_i \geq \beta^{obj} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \alpha} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}. \end{aligned} \quad (2.4)$$

Proof. We present the proof of Theorem 2.1 in Appendix 2.A.2. \square

Theorem (2.1) presents a static reformulation of the ARO Problem (2.3), where all wait-and-see variables are eliminated. The constraints show that the optimal inventory policy is determined by considering the worst-case demand scenario across all possible subsets of SKUs. It guarantees that if demand spikes for some parts, there is enough stock across the system to maintain the overall service level. These constraints make the model suitable for situations with spare parts where system-wide stockouts have severe consequences, but holding excessive inventory for each individual spare part based on its worst-case demand would be prohibitively expensive. The static RO Problem (2.4) provides insight into the situation when we have a target fill rate of 100%, which is captured in the following corollary. More specifically, for $\beta^{obj} = 1$, most constraints in Problem (2.4) are redundant.

Corollary 2.1 *Given $\beta^{obj} = 1$, Problem (2.4) is equivalent to:*

$$\begin{aligned} \min_{S \in \mathbb{N}_0^n} \quad & \sum_{i \in \mathcal{I}} c_i^h S_i \\ \text{s.t.} \quad & S_i \geq \zeta_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}. \end{aligned} \quad (2.5)$$

Corollary 2.1 states that for $\beta^{obj} = 1$, the multi-item optimization problem (system approach) decomposes into separate single-item problems (item approach).

We notice that after eliminating all the wait-and-see variables, both Problems (2.4) and (2.5) are static linear robust optimization problems under right-hand side uncertainty. An optimal solution to Problem (2.5) is achieved when the stock levels of all SKUs reach their lower bounds. Intuitively, we see the same structure for the optimal solution of Problem (2.4). More specifically, to find the optimal solution to Problem (2.4), we can prioritize approaching the lower bounds of stock levels for higher-priced SKUs first.

While we can easily reformulate Problem (2.4) into a linear optimization problem, the obvious drawback is that Problem (2.4) has $2^n - 1$ constraints (the constraint with $\alpha = \mathcal{I}$ is redundant), which is computationally challenging. To address this, one can employ the branch-and-cut (B&C) method to obtain an exact solution (detailed in Appendix 2.A.5), which is typically effective for mixed integer optimization problems with numerous constraints. Our numerical results show that the B&C method is still computationally expensive. Therefore, in Section 2.3.3, we propose several algorithms that provide close-to-optimal solutions for Problem (2.4).

2.3.3 New Approximation Algorithms

The structure of the solution for Problem (2.4) served as our inspiration for designing two algorithms capable of finding solutions when n is large. Without loss of generality, we assume that the indices of SKUs are based on their prices in descending order. In other words, $c_1^h \geq c_2^h \geq \dots \geq c_n^h$. Let us consider the following equivalent formulation of Problem (2.4):

$$\begin{aligned} \min_{S \in \mathbb{N}_0^n} \quad & \sum_{i \in \mathcal{I}} c_i^h S_i \\ \text{s.t.} \quad & \sum_{i \in \alpha} S_i \geq \lceil \max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i \rceil, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset. \end{aligned} \quad (2.6)$$

Now, we write the constraints of Problem (2.6) in matrix form as follows:

$$AS \geq \mathbf{b}, \quad S \geq \mathbf{0}, \quad (2.7)$$

where $S \in \mathbb{N}_0^n$ is the vector of decision variables, $A \in \mathbb{R}^{2^n - 1 \times n}$ is the constraint coefficient matrix, and $\mathbf{b} \in \mathbb{R}^{2^n - 1}$ is the right-hand side parameter vector. We can

express A and \mathbf{b} as

$$a_{\alpha,j} = \begin{cases} 0 & \text{if } j \in \mathcal{I} \setminus \alpha \\ 1 & \text{if } j \in \alpha \end{cases}, \quad b_{\alpha} = \lceil \max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i \rceil, \quad (2.8)$$

where $a_{\alpha,j}$ represents the component in the α^{th} row and the j^{th} column of the matrix A , b_{α} is the component in the α^{th} row of the vector \mathbf{b} for any non-empty set $\alpha \subseteq \mathcal{I}$. Note that $b_{\alpha} \in \mathbb{N}_0$, for any $\alpha \subseteq \mathcal{I}, \alpha \neq \emptyset$.

Iterative Projection in Descending Order (IPDO) Algorithm:

We now introduce an algorithm that we call the iterative projection in descending order algorithm (IPDO). In this algorithm, we first restrict ourselves to SKU 1 and find its optimal stock level: $S_1 = b_{\{1\}}$. We next go to SKU 2. For this SKU, we have two lower bounds: $S_2 \geq b_{\{2\}}$ and $S_2 \geq b_{\{1,2\}} - S_1$. Using these lower bounds, we can find the best stock level for SKU 2. We continue this process until we have the stock levels for all SKUs.

Algorithm IPDO *An algorithm to solve Problem (2.6).*

- 1: **for** $k = 1, \dots, n$:
 - 2: $\Omega_k := \{\alpha \subseteq \{1, \dots, k\} : \alpha \neq \emptyset, k \in \alpha\}$
 - 3: $S_k := \max\{\max_{\alpha \in \Omega_k} \{b_{\alpha} - \sum_{i \in \alpha \setminus \{k\}} S_i\}, 0\}$
 - 4: **endfor**
 - 5: **Output:** \mathbf{S}
-

Remark 2.1 *Compared to the greedy algorithm often used to solve the stochastic Problem (2.1), see, e.g., Basten and Van Houtum (2014), our algorithm has an entirely different approach to achieving the solution. To reach the desired aggregate fill rate, the greedy algorithm first increases the stock level of the SKU that gives the largest ratio of increase in fill rate divided by the cost of stocking one more unit of that SKU. This often implies an increase in the stock level of a cheaper SKU. In IPDO, however, we first determine the stock level of the most expensive SKU. Then, we determine the base stock levels of the SKUs sequentially in descending order of their prices. In our numerical experiments, we delve into what these differences imply for the resulting solutions.*

In the following theorem, we show that IPDO provides us with an optimal solution to Problem (2.6) under some conditions. To show this, for any non-empty set $\alpha \subseteq \mathcal{I}$,

let Π^α be a partition of α . In other words, there exists a $p \in \mathbb{N}$ such that $\Pi^\alpha = \{\Pi_1^\alpha, \dots, \Pi_p^\alpha\}$ and

$$\Pi_1^\alpha \cup \Pi_2^\alpha \cup \Pi_3^\alpha \cup \dots \cup \Pi_p^\alpha = \alpha, \quad \Pi_i^\alpha \cap \Pi_j^\alpha = \emptyset, \quad \forall i \neq j. \quad (2.9)$$

Theorem 2.2 *The solution generated by IPDO is optimal when for any non-empty subset $\alpha \subseteq \mathcal{I}$ and any partition Π^α , we have $b_\alpha \geq \sum_{B \in \Pi^\alpha} b_B$, where b_B , defined in Equation (2.8), is the right-hand side of the constraint associated with the set B .*

Proof. We first show that if we remove the integrality restriction in Problem (2.6), leading to what we call the relaxed problem, then $S^* = (S_1^*, \dots, S_n^*)$ obtained by IPDO is a basic feasible solution that is optimal. According to the assumption and procedures in IPDO, S^* can be expressed as $S_k^* = b_{\{1,2,\dots,k\}} - \sum_{i=1}^{k-1} S_i^*$, i.e.

$$\begin{aligned} S_1^* &= b_{\{1\}}, \\ S_1^* + S_2^* &= b_{\{1,2\}}, \\ S_1^* + S_2^* + S_3^* &= b_{\{1,2,3\}}, \\ &\vdots \\ \sum_{i \in \mathcal{I}} S_i^* &= b_{\mathcal{I}}. \end{aligned} \quad (2.10)$$

Now we use the Simplex algorithm (Bazaraa et al., 2011) to show that S^* is an optimal solution to the relaxed problem. Let us set A_1 to be the lower triangular matrix of ones with n rows and n columns. Also, let $\bar{b} := [b_{\{1\}}, b_{\{1,2\}}, b_{\{1,2,3\}}, \dots, b_{\mathcal{I}}]$.

Now let us denote by A_2 the rows of A excluding A_1 , $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$. Set $\mathcal{M} := \{\{1\}, \{1,2\}, \{1,2,3\}, \dots, \mathcal{I}\}$ and let $2^{\mathcal{I}}$ be the power set of \mathcal{I} . Introducing slack variables x_α , for any non-empty set $\alpha \subseteq \mathcal{I}$, we can rewrite Constraints (2.7) in a standard form as

$$\begin{bmatrix} A_1 & \mathbf{0}_{n,2^n-1-n} & -I^n \\ A_2 & -I^{2^n-1-n} & \mathbf{0}_{2^n-1-n,n} \end{bmatrix} \begin{bmatrix} S \\ [x_\alpha]_{\alpha \in 2^{\mathcal{I}} \setminus (\mathcal{M} \cup \emptyset)} \\ [x_\alpha]_{\alpha \in \mathcal{M}} \end{bmatrix} = \bar{b}, \quad S \geq 0, \quad x_\alpha \geq 0,$$

where I^{2^n-1-n} and I^n are the identity matrices in $\mathbb{R}^{(2^n-1-n) \times (2^n-1-n)}$ and $\mathbb{R}^{n \times n}$, respectively, $\mathbf{0}_{n,2^n-1-n}$ is the $n \times (2^n - 1 - n)$ zero matrix, and $\mathbf{0}_{2^n-1-n,n}$ is the $(2^n - 1 - n) \times n$ zero matrix.

Consider the basic feasible solution corresponding to $B = \begin{bmatrix} A_1 & \mathbf{0}_{n,2^n-1-n} \\ A_2 & -I^{2^n-1-n} \end{bmatrix}$. For

the basic solution, S and $[x_\alpha]_{\alpha \in 2^{\mathcal{I}} \setminus (\mathcal{M} \cup \emptyset)}$ are the basic variables. Since B is a block matrix with invertible matrices in the diagonal, B is invertible (Bernstein, 2009), and its inverse is

$$B^{-1} = \begin{bmatrix} A_1^{-1} & \mathbf{0} \\ A_2 A_1^{-1} & -I^{2^n - 1 - n} \end{bmatrix}$$

where A_1^{-1} is

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Therefore, B is a basis. Since its corresponding solution is feasible, it is a feasible basis. Now, we check the optimality criteria of the Simplex algorithm. To this end, set $c_B^a = (c_1^a, c_2^a, \dots, c_n^a, 0, 0, \dots, 0)$ with n zeros. So, we have

$$c_B^a B^{-1} = (c_1^a - c_2^a, c_2^a - c_3^a, \dots, c_{n-1}^a - c_n^a, c_n^a, c_n^a, 0, 0, \dots, 0).$$

Let $e_i = [0, 0, 0, \dots, 1, \dots, 0]^T$ be the n -tuple with all components equal to 0, except the i th one being 1. And let z_α denote the objective function value for any $\alpha \in \mathcal{M}$. Then, we have

$$\begin{aligned} z_{\{1\}} - c_{x_{\{1\}}}^a &= c_B^a B^{-1}(-e_1) - c_{x_{\{1\}}}^a = c_2^a - c_1^a < 0, \\ z_{\{1,2\}} - c_{x_{\{1,2\}}}^a &= c_B^a B^{-1}(-e_2) - c_{x_{\{1,2\}}}^a = c_3^a - c_2^a < 0, \\ &\vdots \\ z_{\mathcal{I}} - c_{x_{\mathcal{I}}}^a &= c_B^a B^{-1}(-e_n) - c_{x_{\mathcal{I}}}^a = -c_n^a < 0, \end{aligned}$$

where the negativity comes from the fact that the initial investment is in descending order. It is clear that all the reduced cost coefficients of the non-basic variables are non-positive. Therefore, the solution to the relaxed problem is optimal. Because $b_\alpha \in \mathbb{N}_0$, S^* is integer, hence S^* is optimal for Problem (2.6). \square

The assumptions in Theorem 2.2 are linked to the absence of a redundant constraint in Problem (2.6). The validity of this assumption depends heavily on the choice of the uncertainty set and the value of β^{obj} (see Corollary 2.1). However, the presence of redundant constraints does not necessarily imply that our algorithm cannot obtain an optimal solution. In the following, we provide an example in which most constraints are redundant, yet our algorithm still provides an optimal solution.

Example 2.1 Given an uncertainty set:

$$\mathcal{D} = \{\zeta \in \mathbb{R}^n : \underline{d}_i \leq \zeta_i \leq \bar{d}_i, \forall i \in \mathcal{I}\},$$

where $\bar{d}, \underline{d} \in \mathbb{Z}^n$, and $\bar{d} \geq \underline{d}$. We obtain \mathbf{b} in Constraints (2.7) from the following equation:

$$b_\alpha = \max_{\zeta \in \mathcal{D}} \beta^{obj} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i = \beta^{obj} \sum_{i \in \alpha} \bar{d}_i + (\beta^{obj} - 1) \sum_{i \in \mathcal{I} \setminus \alpha} \underline{d}_i, \quad \forall \alpha \neq \emptyset, \alpha \subseteq \mathcal{I}. \quad (2.11)$$

When $\underline{d}_i = 0$ for any $i \in \mathcal{I}$, we have $b_\alpha = \sum_{B \in \Pi^\alpha} b_B$ for any $\alpha \subseteq \mathcal{I}$, $\alpha \neq \emptyset$. Therefore for any α containing more than one element, the constraint corresponding to b_α is redundant. This leads us to only n active constraints, while IPDO still finds an optimal solution. \square

The main disadvantage of IPDO is its computational effort. For the k -th iteration, we need to calculate S_k based on 2^{k-1} values. Therefore, the computation time still grows dramatically, which is not feasible in practice when we have a large number of SKUs. To address this limitation, we propose two alternative approaches and introduce a preprocessing step.

Linear Equation System Heuristic (LES) and Constraint Generation Algorithm (ConGA):

We first use the proof of Theorem 2.2 to introduce a heuristic based on the linear equation system (2.10), which we call LES. This method allocates stock levels sequentially, starting with $S_1 = b_{\{1\}}$ for SKU 1. For each subsequent SKU i , we set $S_i = b_{\{1,2,\dots,i\}} - \sum_{j=1}^{i-1} S_j$, which is the difference between the cumulative lower bound up to and including SKU i and the sum of previously allocated stock levels. This process continues until the final SKU n , where $S_n = b_{\mathcal{I}} - \sum_{i \in \mathcal{I} \setminus n} S_i$. The LES heuristic provides an optimal solution when, for any non-empty subset $\alpha \subseteq \mathcal{I}$ and any partition Π^α , the inequality $b_\alpha \geq \sum_{B \in \Pi^\alpha} b_B$ holds.

The other heuristic is a simpler version of IPDO for cases with many SKUs, which we call the constraint generation algorithm (ConGA). The idea of this algorithm is similar to IPDO. However, we only consider restricting ourselves to a subset of constraints. Specifically, for a given value $j \leq n$ (which we call the layer of ConGA), we consider $\sum_{l=1}^j \binom{n}{l}$ constraints of the problem, corresponding to all non-empty subsets of SKUs with at most j members. For example, when $j = 2$ and $n = 4$, we would include 10 constraints with right-hand side values of $b_{\{1\}}$, $b_{\{2\}}$, $b_{\{3\}}$, $b_{\{4\}}$, $b_{\{1,2\}}$, $b_{\{1,3\}}$, $b_{\{1,4\}}$, $b_{\{2,3\}}$, $b_{\{2,4\}}$, and $b_{\{3,4\}}$. When $j = n$, ConGA and IPDO coincide.

Algorithm ConGA *An algorithm to solve Problem (2.6) with $j(\leq n)$ layers.*

```

1: Given:  $j$ ;
2: for  $k = 1, \dots, n$ :
3:    $\Omega_k := \{\alpha \subseteq \{1, \dots, k\} : \alpha \neq \emptyset, k \in \alpha, |\alpha| \leq j\}$ 
4:    $S_k := \max\{\max_{\alpha \in \Omega_k} \{b_\alpha - \sum_{i \in \alpha \setminus \{k\}} S_i\}, 0\}$ 
5: endfor
6: Output:  $S$ 

```

To enhance the computational efficiency of IPDO and ConGA, we introduce a preprocessing step as a heuristic add-on. The idea comes from the fact that the solution obtained from LES with box uncertainty sets provides an upper bound on the total cost since we only consider a limited number of constraints. While analytically not proven, we observe that this solution gives, for almost all SKUs, an upper bound on the optimal stock level. Therefore, in this preprocessing heuristic, we set the solution obtained by LES as an upper bound on the individual stock level. For more clarification, let us consider the box uncertainty set, and denote by S'_i the solution obtained by applying the LES heuristic. So, we assume that $S'_i \geq S_i$. We then calculate $\lceil b_i \rceil$ using the first layer of ConGA. For each SKU i , if $S'_i = \lceil b_i \rceil$, then we report S'_i as the desired stock level for SKU i . For the remaining SKUs, we proceed with IPDO and ConGA, where ConGA uses a higher number of layers to refine the solution.

Hybrid Method

As it is known and our analysis later in Section 2.4.2 shows, Problem (2.3) with box uncertainty sets, particularly when solved using the LES heuristic, provides computationally efficient solutions for large-scale problems but may be overly conservative. Conversely, an extended budget uncertainty set yields less conservative solutions but faces computational challenges as the problem size increases. Therefore, we propose a hybrid method that combines these approaches to balance solution quality and computational efficiency, which consists of three key steps.

- **Initial Solution Generation:** We first apply the computationally efficient LES heuristic with box uncertainty set to obtain initial stock levels S'_i for all SKUs. This provides a feasible baseline solution that can be computed quickly even for large-scale problems.

- **Improvement Potential Assessment:** For each SKU i , we compute $\lceil b_{\{i\}} \rceil$ using RO-ext, representing a lower bound on the required stock level. The gap between S'_i and $\lceil b_{\{i\}} \rceil$ indicates the potential for stock level reduction.
- **Selective Refinement:** We introduce a threshold parameter τ , meaning the upper and lower bounds are close enough. If $S'_i - \lceil b_{\{i\}} \rceil \leq \tau$, we retain S'_i as the final stock level, since the potential reduction in stock levels when applying ConGA with extended budget uncertainty set is limited, while the computational time remains high. If $S'_i - \lceil b_{\{i\}} \rceil > \tau$, we keep the SKU in the pool. For all SKUs in the pool, we apply ConGA with an extended budget uncertainty set to determine the stock level.

Section 2.5 demonstrates the practical effectiveness of this hybrid approach through a real-world case study at ASML with 710 SKUs.

2.4. Numerical Experiments

We perform numerical experiments to evaluate both the effectiveness and computational efficiency of the robust solutions developed in this chapter. First, in Section 2.4.1, we investigate how the choice of the uncertainty set affects the stock levels and objective function values using only 9 SKUs, considering the very long computational time for exact \mathbf{b} calculations. Section 2.4.2 evaluates the performance of our heuristic algorithms and solution methods regarding the accuracy and computation time, scaling up to 400 SKUs for the box uncertainty set and 36 SKUs for the extended budget uncertainty set, with limits imposed by the exact solution method. Section 2.4.3 compares the approximated solution of the RO problem with that of the SO problem for 40 SKUs.

In the literature on spare part inventory control, it is typically assumed that demand follows a Poisson process. Therefore, for Sections 2.4.1 and 2.4.2, we generate such demand data based on the predicted demand rate \bar{m}_i . We construct the uncertainty set using the 95% bootstrap confidence level (Wood, 2005) of the generated demand. For Section 2.4.3, we generate demand data considering different distributions within a simulation framework.

We implement the experiment in Python version 3.11 on a 12-core CPU MacBook M3 Pro Chip with 18GB RAM. We solve deterministic mixed-integer linear opti-

mization models using Gurobi 11.0.2 (Gurobi Optimization, 2018). For the ADR and SA methods, we utilize the RSOME package (Chen et al., 2020).

2.4.1 Performance of Three Different Uncertainty Sets

In this section, we investigate the performance of solutions considering the three uncertainty sets. We focus on a small-scale problem with $n = 9$ SKUs in this section and expand to larger instances with up to 400 SKUs in later sections.

To facilitate a clear understanding of the differences in solutions resulting from the three uncertainty sets, we first present an illustrative example. Table 2.1 displays the input parameters for each SKU $i \in \mathcal{I}$: the predicted demand rate \bar{m}_i , price c_i^a , and lead time t . The nine SKUs are formed by all combinations of Low, Medium, and High demand rates and Low, Medium, and High prices. The parameter settings are exactly the same as the examples discussed in Chapter 2 of Van Houtum and Kranenburg (2015).

Figure 2.1 depicts the stock levels for each SKU using three uncertainty sets in RO problem (2.6). We see that in this figure, the stock levels are the same in most cases when using the box uncertainty set (RO-box) and the budget uncertainty set (RO-bud), with only slight differences observed for expensive SKUs at certain β^{obj} . As expected, the extended budget uncertainty set (RO-ext) results in slightly lower stock levels in most cases, which consequently increases the occurrence of zero-stock situations for low-demand SKUs.

Table 2.1: Inputs for the illustrative example.

Parameter	Value by Category
Mean demand (\bar{m}_i) (units/year)	Low: 1 Medium: 5 High: 15
Acquisition cost* (c_i^a) ($\times 1000$ EUR per time unit)	Low: 0.2 Medium: 0.6 High: 4
Lead time (t)	2 months

* The annual holding cost c_i^h per SKU is set at 20% of c_i^a .

We then extend our analysis to the general case. We generate 100 problem instances for each β^{obj} value, where each problem instance is created by randomly drawing values for \bar{m}_i for all $i \in \mathcal{I}$ from a discrete uniform distribution over the interval

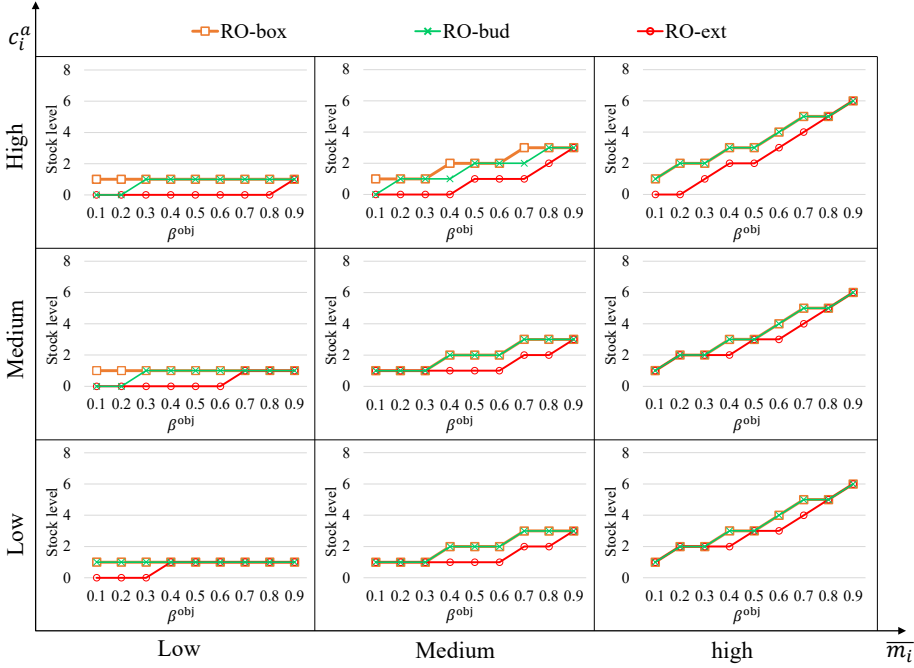


Figure 2.1: Solutions achieved by three uncertainty sets for the illustrative example.

[1,30]. To obtain a comprehensive representation of demand uncertainty, we generate 1,000 demand data points for each SKU from a Poisson process with a mean of \bar{m}_i . We randomly generate the values of c_i^h from a discrete uniform distribution over the interval $[0,2,4]$. These parameter settings are motivated by the real-life value of spare parts at ASML. In practice, β^{obj} typically takes high values (usually ≥ 0.9) to ensure customer satisfaction (Thonemann et al., 2002; Tan et al., 2017). We also include lower values to better demonstrate how performance changes across different β^{obj} . Therefore, we examine $\beta^{obj} \in \{0.8, 0.85, 0.9, 0.95\}$. We define

$$\Delta C^{\text{box-bud}} = \frac{C^{\text{box}} - C^{\text{bud}}}{C^{\text{box}}}, \quad \Delta C^{\text{box-ext}} = \frac{C^{\text{box}} - C^{\text{ext}}}{C^{\text{box}}},$$

where C^{box} , C^{bud} , and C^{ext} are the total holding costs per time unit using the box, budget, and extended budget uncertainty set. Therefore, $\Delta C^{\text{box-bud}}$ and $\Delta C^{\text{box-ext}}$ represent the relative difference in the total holding cost per time unit between the box and budget uncertainty set, and between the box and extended budget uncertainty set, respectively. Since we have $\mathcal{D}^{\text{ext}} \subset \mathcal{D}^{\text{bud}} \subset \mathcal{D}^{\text{box}}$, we know that $\Delta C^{\text{box-bud}}, \Delta C^{\text{box-ext}} \geq 0$.

Figure 2.2 (a) shows the histogram of the values of $\Delta C^{\text{box-bud}}$ across all test instances, revealing that while the total holding costs per time unit using the box uncertainty set (RO-box) are consistently higher than when using the budget uncertainty set (RO-bud), this difference is small (typically less than 5%). In Figure 2.2 (b), the histogram of the values of $\Delta C^{\text{box-ext}}$ shows a different pattern, with a narrowing difference of the total holding costs per time unit as β^{obj} increases. However, 55% of the instances still have strictly lower total holding costs per time unit using the extended budget uncertainty set (RO-ext). As expected, using the RO-box yields the highest total holding costs per time unit across all instances, indicating the most conservative approach.

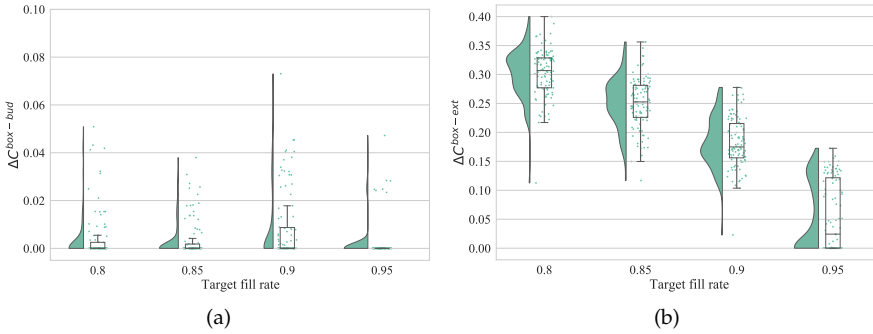


Figure 2.2: Comparison of total holding costs per time unit for different uncertainty sets.

Note: The y-axis in (a) has a different scale than in (b).

From the above analysis, we observe that the additional constraint on the sum of demand for the budget uncertainty set has little effect on solutions, resulting in similar behavior to the box uncertainty set. From a mathematical perspective, the additional constraints on $\sum_{i \in \mathcal{I}} \zeta_i$ only make slight differences to the values of $\max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i$, for any non-empty set $\alpha \subseteq \mathcal{I}$. However, the optimal stock is sensitive to $\lceil \max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i \rceil$ for any non-empty set $\alpha \subseteq \mathcal{I}$, which is not different in most of the instances considering the box or budget uncertainty set. Using the extended budget uncertainty set, however, significantly affects the values of $\lceil \max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \alpha} \zeta_i \rceil$, for any non-empty set $\alpha \subseteq \mathcal{I}$, resulting in less conservative solutions. Because RO-box and RO-bud behave similarly, we do not explore the latter further in the rest of the

numerical experiment.

2.4.2 Comparison of Different Solution Methods

In the previous section, we compared the solutions obtained using different uncertainty sets by solving Problem (2.6) exactly. In this section, we evaluate two exact methods: Gurobi and branch-and-cut (B&C). We evaluate the IPDO, ConGA, LES heuristic, and ADR for approximation methods, all applied to Problem (2.3). We explore the performance of ConGA with one or two layers, referred to as ConGA 1 ($j = 1$) and ConGA 2 ($j = 2$), respectively. Additionally, we evaluate solutions obtained by applying SA to Problem (2.2). We emphasize that based on Proposition 2.1 in Appendix 2.A.1, applying SA to Problem (2.3) likely gives an infeasible solution. Appendix 2.A.4 contains a more detailed comparison of ConGA with different numbers of layers, and Appendix 2.A.5 provides details on the B&C method.

We set $\beta^{\text{obj}} = 0.9$ and vary n from 4 to 400 for RO-box and from 4 to 36 for RO-ext, limited by the computational time of the B&C method (used as a benchmark to evaluate accuracy). For each given n , we generate 10 random problem instances. The values of c_i^h for each problem instance are obtained following the same procedure as described in Section 2.4.1. Since most SKUs in the actual situation have low demand, and inspired by the demand patterns of ASML's SKUs, we categorize them into low (80% of all SKUs), medium (12% of all SKUs), and high (8% of all SKUs) demand groups. The corresponding \bar{m}_i values are drawn from discrete uniform distributions over $[1, 4]$, $[5, 10]$, and $[11, 30]$, respectively. To assess the accuracy of the solutions, we calculate the mean absolute percentage error (MAPE) of the stock levels of the approximation method. The MAPE is formulated as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\tilde{S}_i - S_i}{S_i} \right|,$$

where \tilde{S}_i and S_i are the approximated and exact stock levels for SKU i , respectively.

Box Uncertainty Set

In this section, we discuss the performance of algorithms on problems using the RO-box. We use Equation (2.11) to calculate the exact value of \mathbf{b} , which represents the right-hand side values of the constraints in Problem (2.6).

Figure 2.3 illustrates the performance of different methods by comparing the computation time and the MAPE value. It is clear that B&C substantially reduces com-

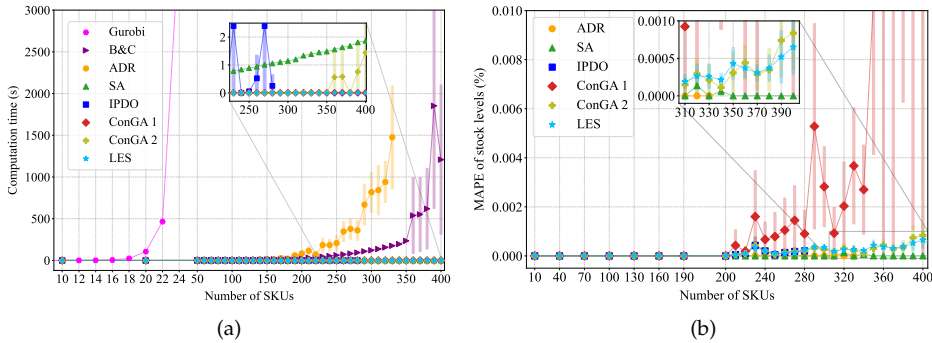


Figure 2.3: Mean (marker point) and standard deviation (shaded range) of computation time (a) and MAPE values (b) for different methods using RO-box. Note: The shaded range of computation time for some solutions is too narrow to discern. Gurobi and B&C give exact solutions and thus are not shown in (b).

putation time compared to Gurobi while still providing optimal solutions. However, it still experiences rapid increases in computation time when $n \geq 350$. IPDO and ADR face memory limitations for larger instances ($n > 280$ and 330 , respectively). ConGA ₁, while computationally efficient, shows declining solution quality for larger instances. ConGA ₂ addresses this limitation by maintaining low MAPE values even for larger problem sizes. SA maintains good performance across problem sizes, balancing reasonable computation times with high solution quality. The LES heuristic demonstrates remarkable efficacy, exhibiting the shortest computation time while consistently attaining very low MAPE values.

Extended Budget Uncertainty Set

The extended budget uncertainty set contains $2^{n+1} - 2$ constraints when there are n SKUs in the model. To efficiently derive the values of \mathbf{b} using this uncertainty set, we propose an approximation algorithm, which is explained in Appendix 2.A.6. The error in approximating \mathbf{b} using this algorithm is minimal, resulting in identical stock levels compared to using the exact \mathbf{b} while being notably faster than obtaining the exact \mathbf{b} . In the following comparison, we use the approximated value of \mathbf{b} for IPDO, B&C, ConGA, and LES.

In Figure 2.4 we compare the performances of different methods. Figure 2.4(a) displays the computation time, and Figure 2.4(b) shows the MAPE of stock levels

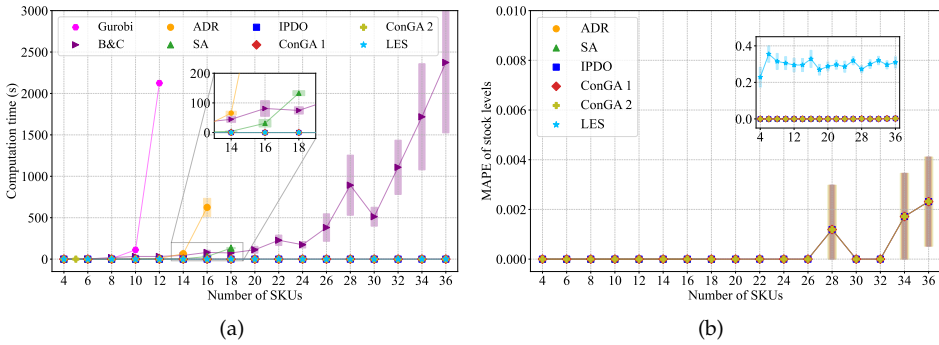


Figure 2.4: Mean (marker point) and standard deviation (shaded range) of computation time (a) and MAPE values (b) for different methods using the RO-ext. Note: The shaded range of computation time for some solutions is too narrow to discern. Gurobi and B&C give exact solutions and thus are not shown in (b).

for various methods as the number of SKUs increases. Exact methods (Gurobi and B&C) show rapidly increasing computation times, with Gurobi facing limitations beyond $n = 12$. ADR and SA also exhibit growing computation times and reach memory limitations at $n > 16$ and 18 , respectively. Therefore, these methods are excluded from Figure 2.4(b). In contrast, IPDO, ConGA (1 and 2), and the LES heuristic maintain low computation times across all problem sizes. For solution quality, the LES heuristic shows high MAPE (20-40%) as n increases. IPDO and ConGA maintain very low MAPE (0.1-0.2%) in a few instances while achieving optimal solutions in all other cases. Our further investigation in Appendix 2.A.4 shows that IPDO encounters memory limitations when $n > 70$, while ConGA performs effectively beyond this number.

Overall, the effectiveness of all existing methods is heavily dependent on the complexity of the uncertainty set, which can lead to errors and memory shortages, especially when using the extended budget uncertainty set. Our proposed ConGA performs consistently well for both types of uncertainty sets, while the LES heuristic demonstrates good performance for only the box uncertainty set. As n increases, ConGA's performance can be improved by increasing the number of layers with an increase in the computational time. Specifically, when the number of layers equals n , ConGA coincides with IPDO.

2.4.3 Comparison of Stochastic and Robust Models

This section compares the performance of implementing SO and ARO models. We first present an illustrative example with 9 SKUs, followed by a comprehensive simulation-based experiment with 40 SKUs. The simulation examines two cases: (i) uncertainty in the shape of the demand distribution and (ii) combined uncertainty in both the shape and size of the demand.

Illustrative example: We start with an illustrative example featuring nine SKUs, which is the same problem instance discussed in Section 2.4.1. For the ARO model, we construct the uncertainty set using the 95% confidence interval of the Poisson process given \bar{m}_i .

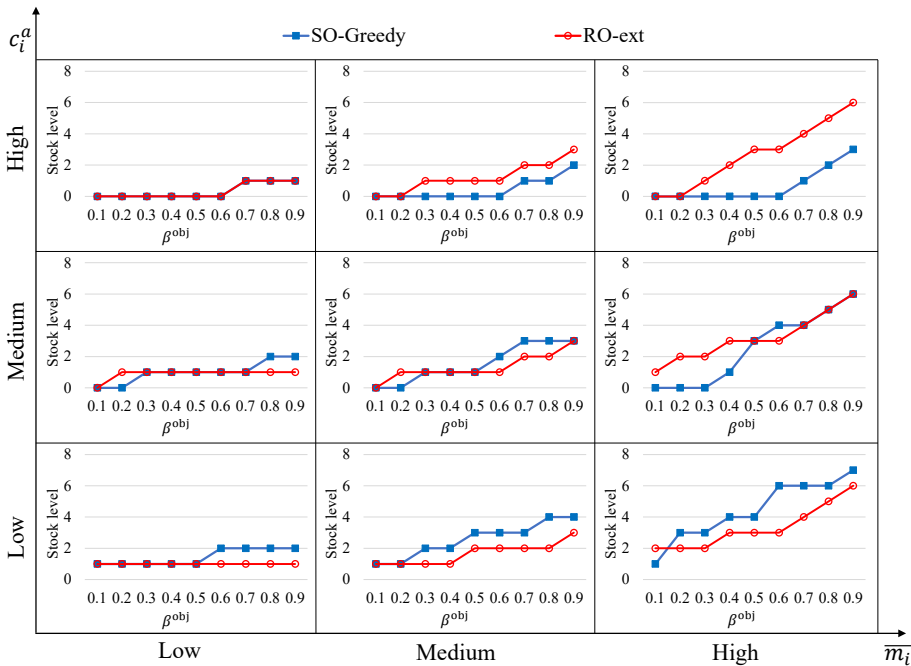


Figure 2.5: Solutions to RO-ext and SO-Greedy for the illustrative example.

Figure 2.5 shows the solution of the SO model obtained by the greedy algorithm (SO-Greedy) (Basten and Van Houtum, 2014) and the exact solution of the robust model considering the extended budget uncertainty set. The SO model tends to stock more spare parts with lower holding costs for SKUs with the same \bar{m}_i . As

β^{obj} increases, the low-cost spare parts are prioritized to reach a higher stock level. In contrast, the robust solution prioritizes high-demand spare parts, maintaining higher stock levels for high-cost spare parts compared to SO-Greedy. For SKUs with equal \bar{m}_i , the stock level obtained by the RO-ext model essentially follows a similar increasing trend when β^{obj} increases.

Simulation-based experiment: Now, we conduct a simulation-based experiment to compare the solutions obtained by the SO and ARO model, following the steps shown in Figure 2.6. The experiment involves $n = 40$ SKUs, each with a two-month lead time. The annual holding cost c_i^h and the predicted demand rate \bar{m}_i are obtained using the same range as described in Section 2.4.2. For an overview, we refer to Table 2.2.

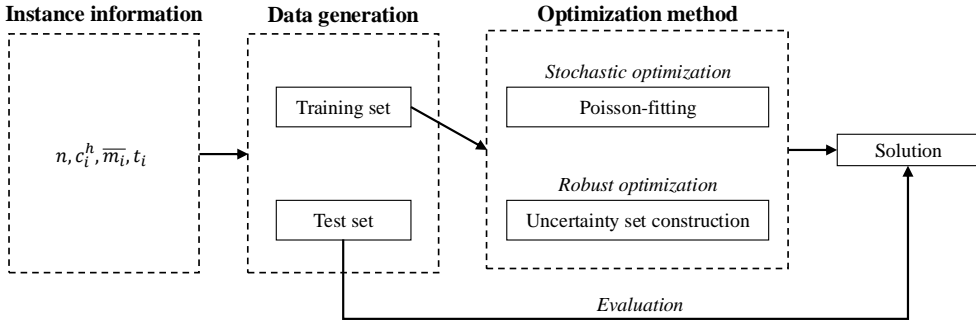


Figure 2.6: Steps for the simulation.

Table 2.2: Input values for the simulation experiments (\mathcal{U} denotes a discrete uniform distribution).

Parameter	Value(s)
Number of SKUs (n)	40
Annual holding cost (c_i^h)	$\mathcal{U}[0.2, 4]$ ($\times 1000$ EUR)
Mean demand (\bar{m}_i) (units/year)	Low: $\mathcal{U}[1, 4]$ Medium: $\mathcal{U}[5, 10]$ High: $\mathcal{U}[11, 30]$
SKU distribution by demand category	Low: 32, Medium: 5, High: 3
Lead time (t)	2 months
Demand inter-arrival distributions	Gamma, Weibull, Lognormal, Inverse Gaussian
Coefficient of variation (CV)	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8

This experiment assumes the availability of sufficient historical data for both approaches, which we acknowledge is an optimistic scenario that may not always hold in practice (we examine the implications of data limitations in Section 2.5). To construct the training dataset, we generate 1,000 demand data points from a given inter-arrival time distribution, each with a mean value of \bar{m}_i for the corresponding SKU. For example, we can use an exponential distribution for the inter-arrival times, which would result in a Poisson demand process. For the SO Problem (2.1), we fit a Poisson process to the training dataset of each SKU. For the ARO model, we construct the uncertainty set using the 95% confidence level based on the empirical distribution of the training dataset. We then obtain solutions for both models with $\beta^{\text{obj}} = 0.9$. To evaluate the performance of the obtained solution, we generate 500 instances. Each instance contains a demand arrival pattern over a two-year simulation period. The first year is considered a warm-up phase, during which fill rates are gradually decreasing as demand is processed due to the initial storage position being at its maximum. Therefore, we commence the calculation of simulated fill rates in the second year, enabling us to assess the solution's performance under stable conditions.

We examine two types of demand uncertainty in the experiment: uncertainty in the shape of the demand distribution and uncertainty in the size of demand. In Case (i), we consider situations where the predicted demand rate \bar{m}_i is accurate. However, the actual demand per lead time follows a distribution with some variation in shape, deviating from the Poisson process. Building on Case (i), we further investigate Case (ii), which includes both the occurrence of demand dispersion from the Poisson process and the misestimation of the predicted demand rate.

Case (i): In the first case, we examine the impact of demand variability on spare parts inventory control. We measure the demand variability by the *coefficient of variation (CV)*, which is the ratio of the standard deviation to the mean of the demand inter-arrival times. A high $CV(> 1)$ indicates a large dispersion of the actual demand rates in relation to the predicted demand rate \bar{m}_i , while a low $CV(< 1)$ implies that the actual demand rates are more concentrated around \bar{m}_i . In Case (i), we vary the standard deviation of the distribution of demand inter-arrival times while keeping the same \bar{m}_i , thus varying CV values.

We extend beyond the conventional exponential distribution for demand inter-arrival times, which yields a Poisson arrival process with a constant CV of 1. To

incorporate varying CV values, we employ four distributions widely recognized for representing positive and continuous demand inter-arrival times: gamma, Weibull, lognormal, and inverse Gaussian (Burgin, 1975; Ghodrati, 2006; Chaves and Gosavi, 2022). Notably, when CV equals 1, both gamma and Weibull inter-arrival times result in a Poisson arrival process. Table 2.2 provides an overview of the test bed. We generate 10 random problem instances for each parameter setting.

We evaluate the performance of the solutions based on their simulated fill rate and total holding cost per time unit. Let us define $\Delta C^{\text{box-greedy}}$ and $\Delta C^{\text{ext-greedy}}$ as the relative difference in the total holding costs per time unit between RO-box and SO-Greedy, and between RO-ext and SO-Greedy, respectively. Specifically, we have:

$$\Delta C^{\text{box-greedy}} = \frac{C^{\text{box}} - C^{\text{greedy}}}{C^{\text{greedy}}}, \quad \Delta C^{\text{ext-greedy}} = \frac{C^{\text{ext}} - C^{\text{greedy}}}{C^{\text{greedy}}},$$

where C^{greedy} represents the total holding cost per time unit using SO-Greedy.

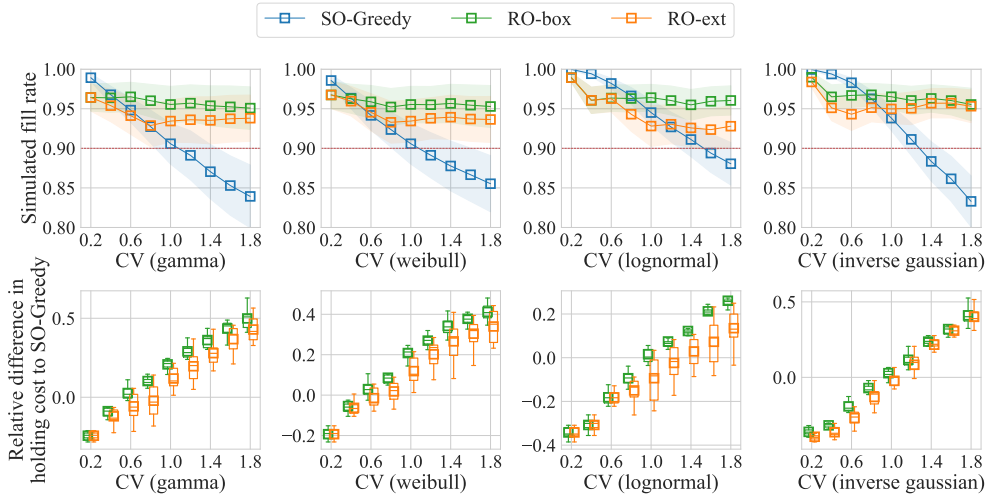


Figure 2.7: Simulated fill rates (top) with mean (marker point) and standard deviation (shaded range), and relative differences in holding cost per time unit to SO-Greedy (bottom) for different demand inter-arrival time distributions.

Figure 2.7 displays the simulated fill rates, along with $\Delta C^{\text{box-greedy}}$ and $\Delta C^{\text{ext-greedy}}$ for different demand inter-arrival time distributions: gamma, Weibull, lognormal, and inverse Gaussian. The top row shows that robust solutions maintain the simulated fill rates consistently above the target ($\beta^{\text{obj}} = 0.9$) across all distributions. The bottom row shows that RO-ext generally yields the target fill rate with lower

costs than RO-box. Notably, the robust solutions perform significantly better than the SO solution when the dispersion of demand is high ($CV > 1$). To keep such a high fill rate, we see that the total holding cost per time unit of robust solutions needs to be increased in such cases. Conversely, when demands concentrate around \bar{m}_i ($CV < 1$), all solutions achieve the target fill rate with closer costs. In some instances, the robust solutions are even more cost-effective than the SO solution. It is worth noting that at $CV = 0.2$, RO-box and RO-ext perform similarly. These patterns are consistent across all four distributions examined, demonstrating the robustness of our findings across different demand inter-arrival time characteristics.

Case (ii): After analyzing Case (i), we now consider the case where both types of demand uncertainty co-occur, i.e., when demand does not arrive according to the Poisson process, and the predicted demand rate \bar{m}_i is misestimated. We examine the extent of misestimation by randomly generating the ratio of the actual demand rate for the test dataset to the predicted demand rate. We draw the ratio from a discrete uniform distribution with the interval $\mathcal{U}[0.8, 1.2]$ for small misestimations and $\mathcal{U}[0.6, 1.4]$ for large misestimations.

We show the simulated fill rates of the solutions in Figure 2.8. As the extent of misestimations increases (Figure 2.8 (bottom)), the simulated fill rates are more affected, with wider ranges. The robust solutions yield a higher simulated fill rate than the SO solutions in most cases for gamma and Weibull distributions and when demand deviates from the mean ($CV > 1$) for lognormal and inverse Gaussian distributions. This indicates the superior performance of robust solutions in mitigating the effects of demand uncertainty in most cases.

Overall, the RO solutions perform well in achieving the target fill rate, especially when the actual demands deviate from a Poisson arrival process or in instances where the demand rate is misestimated. However, the simulated fill rate achieved by SO-Greedy only reaches the target under the assumption of actual demand following a Poisson arrival process and with a relatively accurate prediction of the demand rate. A critical limitation in implementing the SO model becomes evident when these assumptions are violated. In this case, the RO model considering either of the two uncertainty sets yields solutions that achieve the goal as effectively as SO-Greedy and incur similar or lower holding costs per time unit. Therefore, the RO solutions are more stable for coping with demand uncertainty than the SO solution in this investigation. This stability makes them appealing for practical

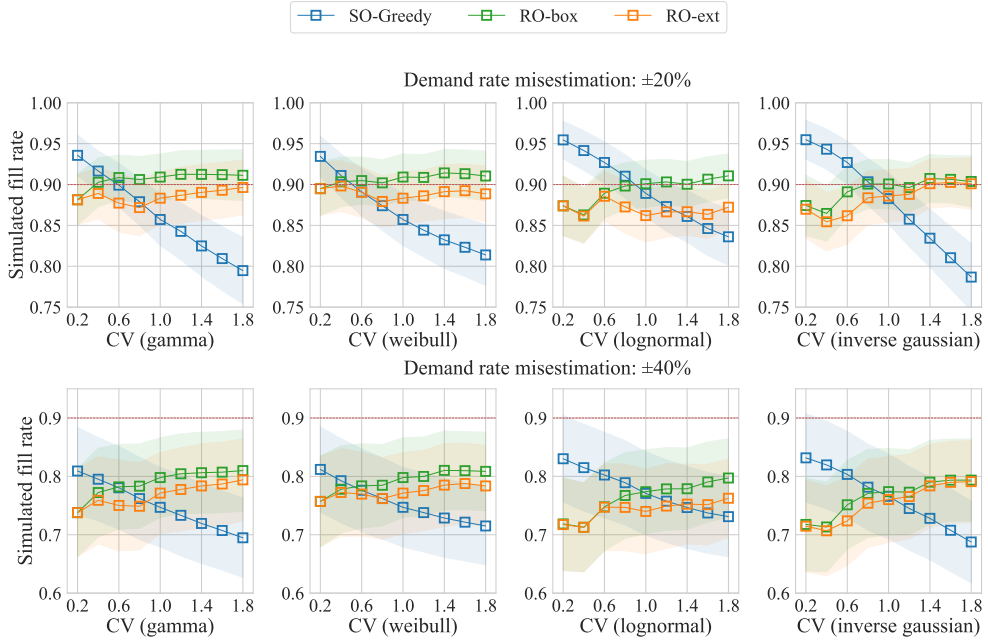


Figure 2.8: Mean (marker point) and standard deviation (shaded range) of simulated fill rates when the ratio of actual demand rate to predicted demand rate follows a discrete uniform distribution of $\mathcal{U}[0.8, 1.2]$ (top) and in $\mathcal{U}[0.6, 1.4]$ (bottom).

implementations, offering enhanced managerial certainty and reliability.

The numerical experiment conducted above demonstrates the effectiveness of the RO and SO models when sufficient data is available. However, this is not always the case in practical applications for spare parts inventory management, particularly during the early stages of the product life cycle. Thus, to gain insight into how data limitation impacts the solutions generated by various models, we conduct a case study in the next section.

2.5. ASML Case Study

As the largest supplier for the semiconductor industry (Tarasov, 2022), ASML manages spare parts to guarantee the availability of its machines. In practice, ASML is required to determine stock levels for over 2,000 SKUs at the new product introduction (NPI) stage. The state-of-the-art spare parts inventory control at ASML

closely resembles the SO Problem (2.1) (Lamghari-Idrissi et al., 2022), which is implemented by first assuming Poisson demand processes, then estimating demand rates, and finally employing a greedy algorithm for generating solutions. Based on the CV analysis of SKUs' demand at ASML (see Section 1.1), and the results from Section 2.4, the conventional SO solution may not be reliable. Therefore, in this section, we investigate the actual performance of the RO solution.

After filtering out SKUs with incomplete or inconsistent historical demand data (more details on data processing can be found in Appendix 2.A.7), we focus on a dataset comprising 710 SKUs out of 2,428 SKUs. This dataset covers the first three years of demand data for a specific generation of ASML machines during the early stages of their product lifecycle. The average annual demand is low, on average 4.22 per year per SKU, and varies between 0.34 and 97. We divide the demand data into two sets: the first two years for training and the third year for testing. We account for non-stationary demand by incorporating machine sales trends into our demand projections, as detailed in Appendix 2.A.7.

ASML categorizes its SKUs based on two primary factors: price and demand rate. Table 2.3 presents the distribution of SKUs across these categories, expressed as proportions of the total SKU count.

Table 2.3: Distribution of SKUs by Price and Demand Categories.

		\bar{m}_i (per year)			Total
		< 5	5 – 10	> 10	
c_i^a (\times 1000 Euros)	< 3	0.65	0.08	0.09	0.82
	3 – 10	0.12	0.02	0.01	0.15
	> 10	0.02	0.00	0.01	0.03
Total		0.79	0.11	0.10	1.00

To avoid costly downtime caused by unfulfilled demand, ASML is committed to achieving a very high fill rate. We set $\beta^{\text{obj}} \in \{0.80, 0.85, 0.90, 0.95, 0.99\}$ to explore the performance of the model across different service targets and to identify the highest fill rate that is practically achievable. ASML standardizes lead time parameters for all SKUs at the inventory planning stage. To reflect a realistic range of parameters while maintaining confidentiality, we set the lead time for each SKU to 3 months.

We employ SO-Greedy to capture the current practice in ASML. We use the LES heuristic for RO-box. For each SKU, we construct the uncertainty set using \underline{d}_i and

Table 2.4: Comparison of the simulated fill rates, the relative difference in total holding costs per time unit when SO-Greedy and RO models achieve the exact same simulated fill rate, and the total computation time using different inventory policies.

β^{obj}	Number of SKUs per (sub) dataset	SO-Greedy	RO-box by LES	RO-ext by ConGA			Hybrid method
		710	710	≈ 48	71	142	710
0.85	Simulated fill rate**	0.660	0.899	0.884	0.878	0.865	0.867
	ΔC^*	–	-8.18%	-11.68%	-9.38%	-9.32%	-5.53%
	Total computation time (s)	20.943	0.054	7.190	19.598	190.390	0.228
0.90	Simulated fill rate**	0.704	0.905	0.897	0.899	0.883	0.888
	ΔC^*	–	-7.22%	-12.21%	-14.62%	-10.12%	-8.78%
	Total computation time (s)	23.554	0.059	6.157	59.597	200.622	0.063
0.95	Simulated fill rate**	0.754	0.913	0.911	0.910	0.908	0.909
	ΔC^*	–	-11.84%	-14.60%	-14.86%	-15.65%	-17.27%
	Total computation time (s)	25.643	0.058	2.258	69.371	208.604	0.000
0.99	Simulated fill rate**	0.819	0.915	0.915	0.915	0.915	0.915
	ΔC^*	–	-10.20%	-10.40%	-10.40%	-10.90%	-10.20%
	Total computation time (s)	31.988	0.054	0.010	0.001	0.007	0.000

* ΔC is the same as $\Delta C^{box-greedy}$ for the comparison of SO-Greedy and RO-box, equivalent to $\Delta C^{ext-greedy}$ for the comparison of SO-Greedy and RO-ext, and equivalent to $\Delta C^{hybrid-greedy}$ for the comparison of SO-Greedy and the hybrid method.

** Simulated fill rate refers to the realized fill rate measured on the third-year data (test set), which is out of sample.

\bar{d}_i , calculated from the minimum and maximum historical demand per lead time, respectively.

For RO-ext, we additionally calculate $\underline{\Gamma}_\alpha$ and $\bar{\Gamma}_\alpha$, representing the minimum and maximum of the aggregated demand for any subset of SKUs α . We use ConGA 3 based on the discussion in Appendix 2.A.4. Due to memory capacity constraints for some algorithms, we decompose the 710 SKUs into 5, 10, or 15 sub-datasets of similar size and demand patterns, resulting in approximately 142, 71, and 48 SKUs per sub-dataset, respectively. Further details can be found in Appendix 2.A.7.

We also investigate the hybrid method outlined in Section 2.3.3, which combines RO-ext with RO-bud and does not require dataset decomposition. In addition, our computational results (detailed in Appendix 2.A.7) show that B&C achieves very similar solution quality to the LES heuristic and ConGA 3 but with significantly longer computational times. For a detailed analysis of solution performance using different methods, including ConGA, LES, and the hybrid method, we refer to Appendix 2.A.8 and 2.A.9.

Table 2.4 compares different inventory policies through simulated fill rates (realized on third-year test data), relative holding cost differences when robust solutions and SO-Greedy achieve the same simulated fill rate, and computation times. The robust solutions (RO-box, RO-ext, and hybrid method) consistently outperform SO-Greedy

in achieving higher simulated fill rates while offering cost savings of 5 to 17% for the same simulated fill rate. We notice that when $\beta^{\text{obj}} = 0.9$, the simulated fill rates of the robust solutions either meet or are very close to β^{obj} . However, the simulated fill rate fails to reach β^{obj} when $\beta^{\text{obj}} = 0.95$. This discrepancy is predominantly attributed to an unexpected surge in demand for certain SKUs in our dataset during the third year. This surge, which is even more than tenfold in the third year compared to the first two years, creates a challenge for further improvement beyond a 90% simulated fill rate.

Regarding computational efficiency, RO-box by LES requires less than 0.06 seconds in all scenarios, and the hybrid method requires at most 0.228 seconds, both notably faster than SO-Greedy (up to 32 seconds) and RO-ext with larger sub-datasets (up to 208 seconds). The hybrid method is particularly effective at $\beta^{\text{obj}} = 0.95$, achieving the highest cost savings of 17.27% while maintaining a simulated fill rate of 0.909, demonstrating its ability to balance solution quality and computational time.

This case study demonstrates that by implementing our robust spare parts inventory solution for the new generation of machines, ASML can achieve the target fill rate more efficiently and cost-effectively than the currently employed SO-greedy approach. While using RO-ext initially required decomposition into smaller sub-datasets, the hybrid method eliminates this need while maintaining high-quality robust solutions, making it particularly valuable for large-scale inventory problems.

2.6. Conclusion

In this chapter, we propose a robust inventory model for spare parts with lost sales. We reformulate the ARO problem into a deterministic counterpart. To solve the problem efficiently, we introduce the IPDO algorithm, which provides optimal solutions under certain conditions. To improve the problem-solving efficiency in large-scale problems, we develop the ConGA and the LES heuristic and propose a hybrid approach combining ConGA and LES.

The results of simulation experiments demonstrate that the solutions obtained by the RO method exhibit greater stability compared to those obtained from the SO model in terms of the simulated fill rate. This stability is evident when the demand deviates from a Poisson process or when there are inaccuracies in estimating the demand rate. The SO model's instability stems from its assumption of Poisson

demand, making the model unreliable when this assumption is violated in practice. The inherent stability of the RO model makes it an attractive option for practical implementations, providing increased certainty and reliability.

We validate our approach through a case study at ASML involving 710 SKUs for one type of machine. The results show that the robust solutions outperform the currently employed inventory policies, making the robust solutions more reliable and cost-effective for coping with demand uncertainty in the spare parts inventory at ASML, especially at the beginning of a product life cycle.

2.A. Appendix

This chapter includes nine appendices. Appendix 2.A.1 provides the proof that in practical cases, applying SA to Problem (2.3) results in an infeasible solution. Appendix 2.A.2 presents a proof of Theorem 2.2. Appendix 2.A.3 presents an explanation of the Fourier-Motzkin elimination procedure as applied to two-stage robust optimization problems. Appendix 2.A.4 investigates the performance of different layers for ConGA, which can help in selecting an appropriate layer for a given number of SKUs. Appendix 2.A.5 introduces the branch-and-cut method, providing pseudocode and discussing its performance in the ASML Case Study. Appendix 2.A.6 presents an algorithm to efficiently derive the values of \mathbf{b} in Constraints (2.7) using the extended budget uncertainty set. Appendix 2.A.7 describes the data preparation process for the ASML Case Study. Appendices 2.A.8 and 2.A.9 evaluate the performance of ConGA, LES, and the hybrid method at ASML.

2.A.1 Proof of Infeasibility of SA to Problem (2.3)

Proposition 2.1 *Applying SA to Problem (2.3) results in an optimization problem that is infeasible if*

$$\frac{\min_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i}{\max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i} < \beta^{\text{obj}}.$$

Proof. Applying SA to Problem (2.3) implies that ϵ_i is a here-and-now variable. Therefore, after applying SA, we have

$$\min_{\substack{S \in \mathbb{N}_0^{|I|} \\ \epsilon_i \in \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^a S_i \quad (2.12a)$$

$$\text{s.t. } \epsilon_i \leq S_i, \quad \forall i \in \mathcal{I}, \quad (2.12b)$$

$$\sum_{i \in \mathcal{I}} \epsilon_i \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \quad (2.12c)$$

$$\zeta_i \geq \epsilon_i \geq 0, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}. \quad (2.12d)$$

For any feasible solution (S, ϵ) in Problem (2.12), we have

$$\beta^{\text{obj}} \max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i \leq \sum_{i \in \mathcal{I}} \epsilon_i \leq \sum_{i \in \mathcal{I}} \min_{\zeta \in \mathcal{D}} \zeta_i \leq \min_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i, \quad (2.13)$$

where, from the left, the first inequality is due to constraint (2.12c), the second inequality is because of (2.12d), and the last inequality is because of the definition of min. Therefore, if Problem (2.12) is feasible, then $\frac{\min_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i}{\max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i} \geq \beta^{\text{obj}}$, which contradicts the assumption. Therefore, Problem (2.12) is infeasible. \square

We emphasize that ζ_i is the demand for SKU i ; hence nonnegative. Therefore, $\max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \zeta_i \neq 0$ as long as \mathcal{D} contains a non-zero element.

In practical situations, the value of β^{obj} is typically quite large (e.g., larger than 0.9), while the demand for spare parts tends to be relatively low, with the lower bound of the uncertainty set close to zero. Consequently, it is likely that the assumption of Proposition 2.1 holds in most situations, which implies that SA is not practical.

2

2.A.2 Proof of Theorem 2.1

We prove Theorem 2.1 using the Fourier-Motzkin elimination (FME) procedure (Zhen et al., 2018). A more detailed explanation of the FME procedure is presented in Appendix 2.A.3. We first show the robust reformulation of Problem (2.3) after eliminating $k(> 0)$ wait-and-see variables in the following proposition.

Proposition 2.2 Let $\mathcal{I}^{n-k+1} = \{n-k+1, \dots, n\}$ and α be a subset of \mathcal{I}^{n-k+1} . After eliminating $\epsilon_i(\zeta)$, for any $i \in \mathcal{I}^{n-k+1}$, Problem (2.3) is equivalent to:

$$\begin{aligned}
 & \min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^{n-k} \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i \\
 & \text{s.t.} \quad \epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
 & \quad \sum_{i=1}^{n-k} \epsilon_i(\zeta) + \sum_{i \in \alpha} \zeta_i + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} S_i \geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
 & \quad \zeta_i \geq \epsilon_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}.
 \end{aligned} \tag{2.14}$$

Proof. Let $k = 1$, we have:

$$\begin{aligned}
 & \epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \\
 & \epsilon_n(\zeta) \leq S_n, \quad \forall \zeta \in \mathcal{D}, \\
 & \sum_{i=1}^{n-1} \epsilon_i(\zeta) + \epsilon_n(\zeta) \geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \\
 & \zeta_i \geq \epsilon_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \\
 & \zeta_n \geq \epsilon_n(\zeta) \geq 0, \quad \forall \zeta \in \mathcal{D},
 \end{aligned}$$

where $S \in \mathbb{N}_0^n$. Eliminating $\epsilon_n(\zeta)$ using FME results in

$$\epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D},$$

$$\begin{aligned}
\sum_{i=1}^{n-1} \epsilon_i(\zeta) + S_n &\geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, & \forall \zeta \in \mathcal{D}, \\
\sum_{i=1}^{n-1} \epsilon_i(\zeta) + \zeta_n &\geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i & \forall \zeta \in \mathcal{D}, \\
\zeta_i \geq \epsilon_i(\zeta) &\geq 0, & \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}.
\end{aligned}$$

So, the formulation of Problem (2.14) is valid for $k = 1$. Let us assume for a given $k \in \mathbb{Z}_+$ that the formulation is equivalent. So, by eliminating $\epsilon_i(\zeta)$ for any $i \in \mathcal{I}^{n-k+1}$ from Problem (2.3), we have

$$\begin{aligned}
\min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^n \rightarrow \mathbb{R}}} & \sum_{i \in \mathcal{I}} c_i^a S_i \\
\text{s.t.} & \epsilon_i(\zeta) \leq S_i, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
& \sum_{i=1}^{n-k-1} \epsilon_i(\zeta) + \epsilon_{n-k}(\zeta) + \\
& \sum_{i \in \alpha} \zeta_i + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} S_i \geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, & \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
& \zeta_i \geq \epsilon_i(\zeta) \geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}.
\end{aligned} \tag{2.15}$$

We show that Problem (2.14) also holds for $k + 1$. Eliminating $\epsilon_{n-k}(\zeta_{n-k})$ from Problem (2.15) using FME results in

$$\begin{aligned}
\epsilon_i(\zeta) &\leq S_i, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D}, \\
\sum_{i=1}^{n-k-1} \epsilon_i(\zeta) + S_{n-k} + \sum_{i \in \alpha} \zeta_i + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} S_i &\geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, & \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
\sum_{i=1}^{n-k-1} \epsilon_i(\zeta) + \zeta_{n-k} + \sum_{i \in \alpha} \zeta_i + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} S_i &\geq \beta^{\text{obj}} \sum_{i=1}^n \zeta_i, & \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
\zeta_i \geq \epsilon_i(\zeta) &\geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D}.
\end{aligned} \tag{2.16}$$

Rearranging (2.16) concludes the proof of Proposition 2.2. \square

The proof of Theorem 2.1 follows from Proposition 2.2 by setting $k = n$.

2.A.3 Fourier-Motzkin Elimination in Robust Optimization

Fourier-Motzkin elimination (FME) is a mathematical technique to remove variables from systems of linear inequalities. In robust optimization, FME serves as an essential tool for reformulating adjustable RO problems into static RO problems.

Based on the work of Zhen et al. (2022), we demonstrate the elimination of a single wait-and-see variable $\epsilon(\zeta)$ by FME from a system of linear inequalities with multiple constraints. Consider the following robust optimization problem with an uncertain vector $\zeta = \{\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5\} \in \mathcal{D}$, where each ζ_i represents a component of the uncertainty, and here-and-now variables are S_1, S_2, S_3 :

$$\begin{aligned}
 & \min_{S \in \mathbb{R}^3} \sum_{i=1}^3 c_i S_i && \text{(Objective)} \\
 & \text{s.t. } \epsilon(\zeta) \leq S_1 + \zeta_1, && \forall \zeta \in \mathcal{D}, \quad \text{(UB1)} \\
 & \epsilon(\zeta) \leq S_2 + \zeta_2, && \forall \zeta \in \mathcal{D}, \quad \text{(UB2)} \\
 & \epsilon(\zeta) \geq \zeta_3 - S_3, && \forall \zeta \in \mathcal{D}, \quad \text{(LB1)} \\
 & \epsilon(\zeta) \geq \zeta_4, && \forall \zeta \in \mathcal{D}, \quad \text{(LB2)} \\
 & \epsilon(\zeta) + \epsilon_1(\zeta) \geq \zeta_5, && \forall \zeta \in \mathcal{D}, \quad \text{(KC)}
 \end{aligned}$$

where $c_i > 0$ are cost coefficients and $\epsilon_1(\zeta)$ is another wait-and-see variable. Now we show how to eliminate $\epsilon(\zeta)$ using FME.

Step 1: Identify Bounds on $\epsilon(\zeta)$

The variable $\epsilon(\zeta)$ is bounded by:

$$\underbrace{\max(\zeta_3 - S_3, \zeta_4, \zeta_5 - \epsilon_1(\zeta))}_{\text{Lower Bound}} \leq \epsilon(\zeta) \leq \underbrace{\min(S_1 + \zeta_1, S_2 + \zeta_2)}_{\text{Upper Bound}}, \quad \forall \zeta \in \mathcal{D}.$$

Step 2: Feasibility Condition

For $\epsilon(\zeta)$ to exist, the lower bound should be smaller than the upper bound, implying:

$$\max(\zeta_3 - S_3, \zeta_4, \zeta_5 - \epsilon_1(\zeta)) \leq \min(S_1 + \zeta_1, S_2 + \zeta_2), \quad \forall \zeta \in \mathcal{D}. \quad \text{(FC)}$$

Step 3: Eliminate $\epsilon(\zeta)$ via FME

Step 2 provides a sufficient and necessary condition to eliminate $\epsilon(\zeta)$. So, expanding Constraint (FC) by pairing all lower bounds with all upper bounds:

$$\zeta_3 - S_3 \leq S_1 + \zeta_1, \quad \forall \zeta \in \mathcal{D}, \quad \text{(C1)}$$

$$\zeta_3 - S_3 \leq S_2 + \zeta_2, \quad \forall \zeta \in \mathcal{D}, \quad \text{(C2)}$$

$$\zeta_4 \leq S_1 + \zeta_1, \quad \forall \zeta \in \mathcal{D}, \quad \text{(C3)}$$

$$\zeta_4 \leq S_2 + \zeta_2, \quad \forall \zeta \in \mathcal{D}, \quad \text{(C4)}$$

$$\zeta_5 - \epsilon_1(\zeta) \leq S_1 + \zeta_1, \quad \forall \zeta \in \mathcal{D}, \quad \text{(C5)}$$

$$\zeta_5 - \epsilon_1(\zeta) \leq S_2 + \zeta_2, \quad \forall \zeta \in \mathcal{D}. \quad (\text{C6})$$

Final Reformulated Problem

After eliminating $\epsilon(\zeta)$, the problem retains Constraints (C1) to (C6). The reformulated problem becomes:

$$\begin{aligned} \min_{S \in \mathbb{R}^3} \quad & \sum_{i=1}^3 c_i S_i \\ \text{s.t.} \quad & \zeta_3 - S_3 \leq S_1 + \zeta_1, & \forall \zeta \in \mathcal{D}, \\ & \zeta_3 - S_3 \leq S_2 + \zeta_2, & \forall \zeta \in \mathcal{D}, \\ & \zeta_4 \leq S_1 + \zeta_1, & \forall \zeta \in \mathcal{D}, \\ & \zeta_4 \leq S_2 + \zeta_2, & \forall \zeta \in \mathcal{D}, \\ & S_1 + \zeta_1 + \epsilon_1(\zeta) \geq \zeta_5, & \forall \zeta \in \mathcal{D}, \\ & S_2 + \zeta_2 + \epsilon_1(\zeta) \geq \zeta_5, & \forall \zeta \in \mathcal{D}. \end{aligned}$$

2.A.4 Performance of ConGA

In this section, we investigate the performance of different numbers of layers of ConGA. We conduct tests for the number of layers $j = 1, 2, 3, 4$, and 5, and vary n between 30 to 90. The parameter settings are the same as those employed in Section 2.4.2. Because the exact inventory levels are unavailable for a large number of SKUs (n), we compare the relative differences in the objective values obtained by different solution methods as follows:

$$\Delta C^{\ell-\ell'} = \frac{C^\ell - C^{\ell'}}{C^{\ell'}},$$

where $\Delta C^{\ell-\ell'}$ represents the relative difference in the total investment costs obtained from solution ℓ compared to solution ℓ' , and where $C^{\ell'}$ and C^ℓ denote the total investment costs of solution ℓ' and solution ℓ , respectively.

In our analysis of the RO model with a box uncertainty set in Section 2.4.2, the LES heuristic demonstrates superior performance, obviating the need for further exploration of ConGA in this context. Therefore, we investigate the performance of ConGA when using the extended budget uncertainty set.

ConGA 1, 2, and 3 exhibit remarkably short computation times across all values of n , as illustrated in 2.9 (a). Meanwhile, ConGA 5 has a significantly longer computation time than other ConGA layers and reaches computational limits at $n = 78$ due to memory constraints.

Given these resource limitations, ConGA emerges as the sole viable option for $n > 30$. Therefore, we establish ConGA 5 as the benchmark for total investment cost comparisons. Figure 2.9 (b) shows the relative difference in total investment costs using ConGA 1 to 4 compared to ConGA 5. ConGA 1 exhibits a more substantial increase as n increases. However, ConGA 2, 3, and 4 slightly differ in total investment costs from ConGA 5.

2

It is important to note that ConGA provides an approximate lower bound for the optimal solution when $j < n$. To quantify the precision of this approximation relative to the exact lower bound, we calculate the mean Mean Absolute Percentage Error (MAPE) of stock levels derived from ConGA, using the exact solution obtained by Gurobi under identical constraints as the reference. Figure 2.9 (c) shows the results. When $j = 1$, the constraint is expressed as $S_i \geq b_i$ for any $i \in \mathcal{I}$, so the mean MAPE value of ConGA is zero. For $j > 1$, the values of mean MAPE increase with n but remain relatively low at less than 8%. This suggests that ConGA solutions yield a comparatively higher lower bound, with higher ConGA layers providing a more elevated upper bound for the lower bound, ultimately resulting in more conservative solutions.

Based on the above analysis, we can select the appropriate ConGA layer for n SKUs by considering running time and accuracy.

2.A.5 The Branch-and-cut (B&C) Method

The B&C method has been utilized in previous research to solve MIP problems with a large number of constraints Côté et al. (2014, 2021). In our implementation, the B&C algorithm is based on iteratively solving Problem (2.6).

Algorithm B&C *Branch-and-cut algorithm for solving Problem (2.6).*

-
- 1: Initialize Problem (2.6) with partial constraints
 - 2: Solve the restricted problem and obtain the objective value \underline{obj} , set $LB = \underline{obj}, UB = +\infty$
 - 3: Solve Problem (2.6) in a B&C fashion:
 - 4: **for** each integer solution found in the B&C:
 - 5: **if** the objective value of the current solution $\geq UB$:
 - 6: This integer solution is non-promising; continue to verify the next integer solution
 - 7: **else** Determine the feasibility of the current solution
 - 8: **if** the solution of the subproblem is feasible:
 - 9: Denote the corresponding objective value \overline{obj} , update $UB = \min\{UB, \overline{obj}\}$
 - 10: Retrieve the best LB from the solver
 - 11: **if** $UB = \lceil LB \rceil$:
 - 12: UB is the optimal solution
 - 13: **return** UB
 - 14: **else** Identify at least a subset $\tilde{\alpha}$ such that the current solution is not satisfied, add the following lazy constraints to the problem

$$\sum_{i \in \tilde{\alpha}} S_i \geq \lceil \max_{\zeta \in \mathcal{D}} \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i - \sum_{i \in \mathcal{I} \setminus \tilde{\alpha}} \zeta_i \rceil, \forall \tilde{\alpha} \subseteq \mathcal{I}, \tilde{\alpha} \neq \emptyset$$

- 15: **endfor**
 - 16: **Output:** S
-

For a given $\alpha \in \mathcal{I}$, we solve Problem (2.6) with standard off-the-shelf commercial MIP solvers. When a new solution is found, we store it in a global variable that is accessible by the branch-and-cut algorithm (implemented in Gurobi 11.0.0). The branch-and-cut algorithm checks for improved upper bounds during its search, and whenever available it uses the solution provided by Problem (2.6) as the new incumbent solution.

Applicability to the ASML case study: Due to memory capacity and computational time constraints, we employ a decomposition strategy when applying the B&C method. For RO-box, we divide the 710 SKUs into 5 sub-datasets of similar size and demand patterns. Similarly, for RO-ext, we utilize 15 sub-datasets to accommodate the B&C method's computational requirements. To address the potential for excessively long running times, we implement a time limit of 1 hour per sub-dataset, resulting in a total limit of 5 hours for the 5 sub-datasets in our ASML case study. This means we consider the best bound and solution found within the 1-hour timeframe for each sub-dataset.

Figure 2.10 illustrates the simulated fill rate and investment cost of LES, ConGA, and B&C methods for both RO-box and RO-ext approaches. For RO-box, the results

from ConGA and B&C methods show close alignment across all β^{obj} . In the case of RO-ext, the differences become negligible when $\beta^{\text{obj}} \geq 0.95$. As β^{obj} decreases to 0.85, the gap widens slightly; however, the investment costs remain nearly identical, with the simulated fill rate differing by merely 0.5%.

2

While the solution quality is comparable, the computational efficiency varies clearly among the methods. Table 2.5 presents a comparison of computation times. Notably, the B&C method's computational time substantially exceeds that of LES for RO-box and ConGA for RO-ext across all β^{obj} .

Table 2.5: Comparison of computation times (s) for different RO methods.

Method	Target Fill Rate (β_{obj})			
	0.85	0.90	0.95	0.99
RO-box by LES	0.054	0.059	0.058	0.054
RO-box by B&C	18,094.082	3,732.563	53.483	42.324
RO-ext by ConGA	7.190	6.157	2.258	0.010
RO-ext by B&C	4,655.211	3,669.387	1,584.105	125.092

Discussion: The computational time of the B&C method is closely related to the constraints, specifically the value of b in Constraint (2.7). The value of b is not only related to the uncertainty set but also to β^{obj} and historical demand data. Consequently, the performance of the B&C method exhibits considerable variability.

To illustrate this variability, we compare results from different scenarios. In Section 2.4.2, we generate ten instances with Poisson demand and set $\beta^{\text{obj}} = 0.90$. For the box uncertainty set, the B&C method required only a few seconds for 150 SKUs. However, our ASML case study, involving 5 sub-datasets, each containing 147 SKUs, yielded different results. When $\beta^{\text{obj}} = 0.90$, four of the sub-datasets exhibited computational times similar to the numerical example in Section 2.4.2. Interestingly, one sub-dataset could not complete its computation even after several hours. The situation became more challenging when $\beta^{\text{obj}} = 0.85$, as all sub-datasets using the B&C method failed to complete within several hours.

These findings underscore the sensitivity of the B&C method's computational time to specific problem characteristics and parameter settings, highlighting the need for careful consideration when applying this method to large-scale inventory optimization problems.

2.A.6 An Algorithm for the Extended Budget Uncertainty Set

The extended budget uncertainty set comprises $2^{n+1} - 2$ constraints when the model includes n SKUs. Using this uncertainty set, we propose the following approximation algorithm to efficiently derive the \mathbf{b} values in Constraints (2.7).

$$\tilde{b}_\alpha := \lceil \beta^{\text{obj}} \bar{\Gamma}_\alpha + (\beta^{\text{obj}} - 1) \underline{\Gamma}_{\mathcal{I} \setminus \alpha} \rceil, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset. \quad (2.17)$$

We use Gurobi to solve the problem instances with approximated and exact \mathbf{b} , then compare their accuracy and computation time. We find that using either the approximated or exact values of \mathbf{b} produced the same results in all instances, which indicates that the error generated by the approximated \mathbf{b} has a minimal effect on the final result. In Figure 2.11, the computation time is significantly reduced when using the approximated \mathbf{b} . The exact \mathbf{b} cannot be obtained due to limited memory when $n > 11$, while the approximation method still performs well.

2.A.7 The Data Filtering Process, Data Decomposing, and Data Processing for ASML Case Study

In this section, we outline the methods utilized for data filtering, decomposition, and processing in the ASML case study, on which we report in Section 2.5.

Data filtering process: We implemented a data filtering process to refine the dataset for our analysis. Initially, the dataset consisted of 2,428 SKUs. We conduct the following steps for the data filtering:

- SKUs lacking historical demand data are excluded to ensure dataset integrity, reducing the count from 2,428 to 1,012 SKUs.
- SKUs demonstrating zero demand in the initial two years are removed, as they do not offer sufficient information for generating viable solutions. This further reduces the SKU count from 1,012 to 710.

Data decomposing: Due to the limited memory of the laptop, we adopt data decomposition using systematic sampling to create smaller sub-datasets with similar demand patterns. For instance, if we aim to decompose the data into m sub-datasets, each sub-dataset will have a sample size of approximately $710/m$. This process works as follows:

- We initiate the process by sorting all SKUs from the smallest to the largest based on their demand size.

- We choose 1 as the starting number and add the 1st SKU, $1 + m^{\text{th}}$ SKU, $1 + 2m^{\text{th}}$ SKU, and so on, to form the elements of the first sub-dataset.
- Next, we choose 2 as the starting number and add the 2nd SKU, $2 + m^{\text{th}}$ SKU, $2 + 2m^{\text{th}}$ SKU, and so forth, to create the second sub-dataset.
- We continue this procedure, incrementing the starting number up to m , to add the remaining SKUs to the subsequent sub-datasets.

Data processing: In this case study, we leverage data from the first two years to obtain solutions for the third year. An essential factor to consider is the continuous surge in demand for spare parts at ASML, which directly corresponds to the increasing sales of machines annually. Consequently, we analyze demand variations over the initial two years, and we assume that the observed trends in demand fluctuations during the first two years will remain consistent in the third year, but multiplied by a certain factor due to the steady annual growth in machine sales.

We estimate the predicted mean demand rate for the third year by multiplying the mean demand rate of the first two years by 1.5, which aligns with the observed sales increase (ASML, 2023). To obtain robust solutions for the third year, we conduct an analysis of the minimum and maximum lead time demands in the second year relative to the first year (Figure 2.12(a) and Figure 2.12(b) respectively). To ensure the robustness of our solutions, we incorporate a wide range of ratios derived from the analysis results presented in Figure 2.12. Specifically, we set 2.5 times the maximum lead time demand observed in the first two years as our predicted maximum lead time demand for the third year. Similarly, we choose 0.5 times the minimum lead time demand observed in the first two years as our predicted minimum lead time demand for the third year.

2.A.8 Analysis of ConGA and LES at ASML

We use a color spectrum from red to yellow to denote RO solutions with β^{obj} ranging from 0.85 to 0.99, and a spectrum from blue to green for SO solutions within the same β^{obj} range. Black lines illustrate the efficient frontier for SO-Greedy.

We begin by examining the performance of RO models using RO-box and RO-ext. Figure 2.13 (a) demonstrates that for RO-ext, the cost-efficiency trends remain similar across different sub-datasets. However, total computational times vary notably. As shown in Table 2.6, the time required for 5, 10, and 15 sub-datasets

Table 2.6: Computational time (s) for different methods.

Method	β^{obj}			
	0.85	0.90	0.95	0.99
RO-box	0.054	0.059	0.058	0.054
RO-ext, 5 sub-datasets	190.390	200.622	208.604	0.007
RO-ext, 10 sub-datasets	19.598	20.481	16.337	0.001
RO-ext, 15 sub-datasets	7.190	6.157	2.258	0.010
SO-Greedy	20.943	23.554	25.643	31.988

are around 200, 20, and 8 seconds, except for when $\beta^{\text{obj}} = 0.99$. When $\beta^{\text{obj}} = 0.99$, implying that most demand should be satisfied from stock, RO-ext and RO-box solutions coincide. However, for $\beta^{\text{obj}} \leq 0.95$, RO-ext solutions demonstrate markedly superior cost-efficiency than RO-box. We notice that the simulated fill rate cannot exceed the value of 0.95, even when we consider $\beta^{\text{obj}} \geq 0.95$. This discrepancy is predominantly attributed to an unexpected surge in demand for certain SKUs in our dataset during the third year. This surge, which is even more than tenfold in the third year compared to the first two years, creates a challenge for further improvement beyond a 91.6% simulated fill rate.

Figure 2.13 (b) shows SO and RO model solutions at various β^{obj} levels. RO solutions clearly outperform SO solutions in terms of cost-efficiency. For instance, at $\beta^{\text{obj}} = 0.95$, the RO-ext solution achieves approximately 10.4% savings in holding costs per time unit compared to SO-Greedy while maintaining the same simulated fill rate. Moreover, SO-Greedy solutions show a considerable gap between β^{obj} and simulated fill rates. When $\beta^{\text{obj}} = 0.99$, the simulated fill rate only reaches 82%. Additionally, SO-Greedy solutions require a longer computation time than the RO-box and the RO-ext for 10 and 15 sub-datasets.

2.A.9 Analysis of Hybrid Method at ASML

To evaluate the performance of the hybrid approach, we examine $\tau \in \{4, 6, 8, 10, 12, 14\}$ for $\beta^{\text{obj}} \in \{0.85, 0.90, 0.95\}$.

Figure 2.14 illustrates the efficient frontiers of solutions obtained by the hybrid method. The grey lines represent the efficient frontiers for RO-ext with five sub-datasets and RO-box with a single dataset. Red squares and green triangles mark the solutions for RO-box and RO-ext, respectively, at the corresponding β^{obj} values.

Pink circles denote the hybrid method solutions for different τ .

The hybrid method produces solutions that lie between those of RO-box and RO-ext in most cases. As τ increases, the hybrid solutions move closer to the RO-box solution. For $\tau \leq 8$, the solutions are closer to the RO-ext, suggesting that it can capture most of the cost-efficiency benefits of RO-ext. We examine the computation times of the hybrid method for different β^{obj} and τ , as presented in Table 2.7. As τ increases, computation time decreases dramatically, approaching the efficiency of the box uncertainty set. Even for smaller τ , computation times are reduced compared to the full RO-ext implementation while maintaining solution quality close to RO-ext.

Table 2.7: Computational time (s) for different β^{obj} and τ

β^{obj}	τ					
	4	6	8	10	12	14
0.85	9.310	0.936	0.228	0.063	0.000	0.000
0.90	9.861	0.956	0.263	0.063	0.000	0.000
0.95	12.231	1.701	0.349	0.088	0.004	0.000

This case study demonstrates that by implementing our robust spare parts inventory solution for the new generation of machines, ASML can achieve the target fill rate more efficiently and cost-effectively than the currently employed SO-greedy approach. While obtaining robust solutions for large datasets initially required decomposition into smaller sub-datasets due to memory constraints, the hybrid method eliminates this need while maintaining high-quality robust solutions. This method proves particularly valuable for large-scale inventory problems where the full implementation of the RO-ext may be computationally prohibitive while still capturing much of its cost-efficiency benefits.

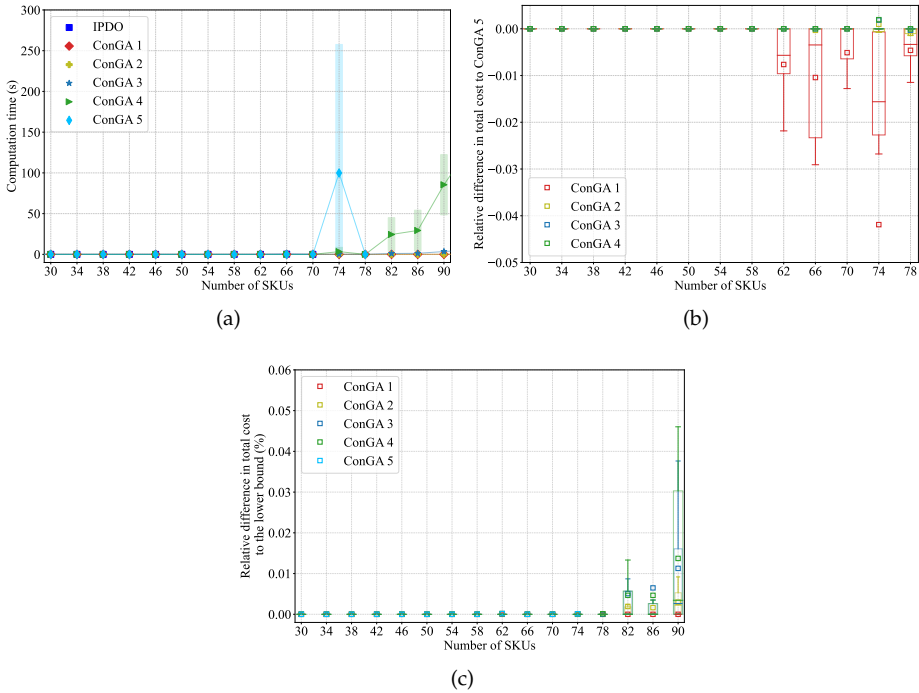


Figure 2.9: Mean (marker point) and standard deviation (shaded range) of the computation time (a), the relative difference in total cost to ConGA 5 (b), and the mean MAPE values of stock levels to the lower bound (c) considering the extended budget uncertainty set.

Note: In (a), some solutions have negligible time variance. IPDO and ConGA 5 terminate at $n > 70$ and $n > 78$, respectively, due to memory issues. In (c), the calculation of MAPE values for different layers of ConGA stops at different n due to the memory issues of obtaining the exact solution using Gurobi.

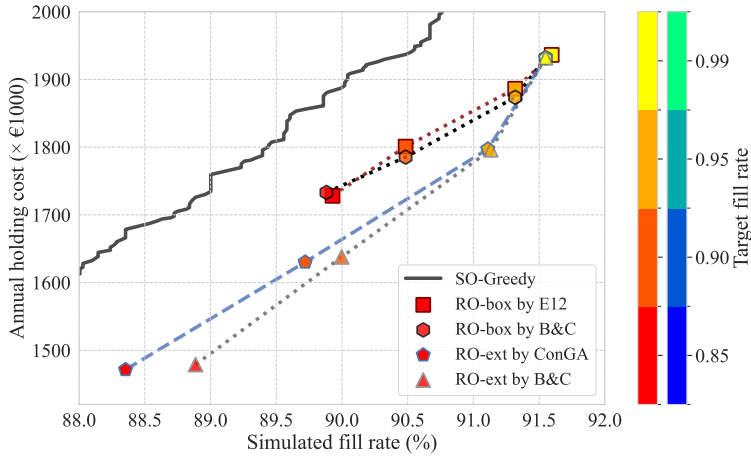


Figure 2.10: Efficient frontier of RO solutions using B&C, LES and ConGA.

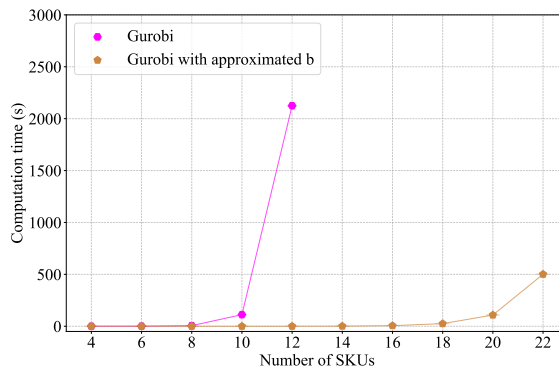


Figure 2.11: Mean (marker point) and standard deviation (shaded range) of computation time for obtaining b .

Note: The shaded range of computation time is too narrow to discern.

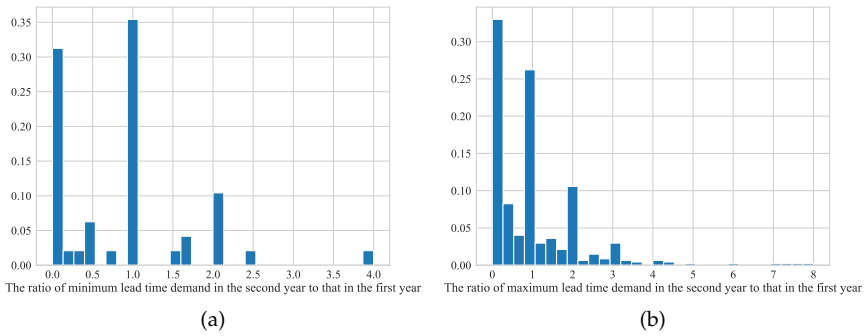


Figure 2.12: The ratio of minimum lead time demand (a) to maximum lead time demand (b) in the second year compared to the first year.

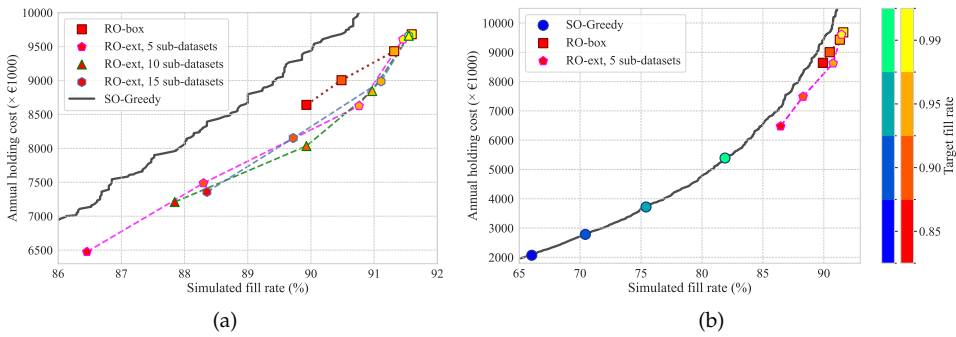


Figure 2.13: Efficient frontier of RO solutions (a) and the comparison with SO solutions (b).

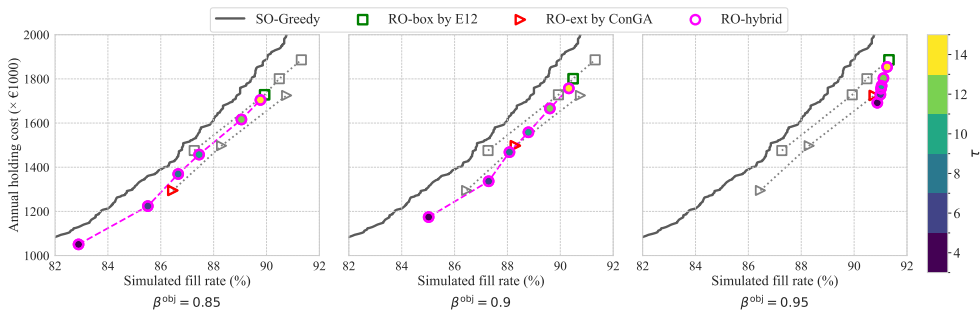


Figure 2.14: Efficient frontier of solutions using the hybrid method.

Chapter 3

Robust spare parts inventory control with emergency shipments

In this chapter, we study the spare parts inventory control problem under high demand uncertainty, particularly during the new product introduction (NPI) stage when historical demand data is scarce. To mitigate stockouts and achieve a low target waiting time, we satisfy unfulfilled demand through emergency shipments.

We formulate a multi-item spare parts inventory control problem as an ARO model. To ensure computational tractability, we reformulate the ARO model as a deterministic counterpart and prove that it can be approximated by decomposing it into two mixed-integer optimization problems. We then develop an efficient algorithm to obtain near-optimal solutions for large-scale problems with thousands of items. When limited demand data is available, we propose an approach that incorporates initial failure rate estimates from engineers into the uncertainty set construction, which enhances the model's performance in the NPI stage.

We demonstrate the practical value of our model through a comprehensive case study at ASML. The case study shows that our model achieves a simulated mean waiting time up to 3.5 hours shorter than the state-of-the-art stochastic optimization model employed at ASML at the same simulated total cost, potentially saving over €250,000 in lost production per breakdown of an expensive lithography system. The sensitivity analysis shows the ARO model's strong adaptability to variations in key parameters of the model. Our research contributes to both the theoretical advancement of robust optimization techniques and their practical application in spare parts inventory management, particularly for industries that sell or use capital equipment.

3.1. Introduction

Service providers of expensive equipment must maintain a sufficient spare parts inventory to ensure high equipment availability while managing inventory costs effectively. When stock is unavailable locally, providers often use emergency shipments from other warehouses instead of waiting for regular replenishment, especially when equipment downtime costs are high. The trade-off between inventory-related costs and the urgency of spare parts availability through emergency shipments forms the foundation of this chapter. We propose a robust optimization approach for spare parts inventory control to incorporate emergency shipments.

Extending to emergency shipments introduces new modeling complexities beyond the problem we considered in Chapter 2. We now need to consider not only the trade-off between holding costs and service levels but also the cost and time implications of emergency shipments across multiple components. This requires careful consideration of how emergency shipment decisions interact with base stock levels and target service requirements.

Our research makes several contributions to both theory and practice. We develop the first robust optimization model for spare parts inventory management with emergency shipments. We introduce an efficient decomposition method that makes the model computationally tractable for large-scale industrial applications. We propose a method for incorporating engineering knowledge into uncertainty set construction when historical data is limited. The practical value of these innovations is validated through a comprehensive ASML case study that shows significant improvements in both service levels and cost efficiency compared to current state-of-the-art approaches.

The remainder of this chapter is organized as follows. Section 3.2 formulates the spare parts inventory control problem with emergency shipments and develops our adaptive robust optimization model. Section 3.3 develops our solution methodology for a conservative formulation of the ARO problem, introducing decomposition techniques and algorithms. Section 3.4 extends to solve a less conservative formulation that considers dependencies between SKUs. Section 3.5 discusses the construction of uncertainty sets that incorporate both historical data and engineering estimates. Section 3.6 demonstrates the practical application through a comprehensive ASML case study. Finally, Section 3.7 concludes with a summary of the findings.

3.2. Problem Description

Consider a single warehouse stocking spare parts for various types of critical components. These components are known as Stock Keeping Units (SKUs) and are represented by the set $\mathcal{I} = \{1, \dots, n\}$, where n is the number of SKUs. If component i fails, it is immediately sent for repair, and it is added to the inventory in an as-good-as-new condition after a fixed repair lead time $t(> 0)$. It is important to note that the model also applies if spare parts are non-repairable, i.e., consumables. In this case, a defective part is discarded, and a new part is procured. Consequently, the repair lead time is substituted by the order-and-ship time of new parts.

Due to the long repair lead times experienced by companies selling expensive equipment, we assume that any demand that cannot be fulfilled immediately is lost from the normal replenishment system of the local warehouse. As a result, such unsatisfied demand is satisfied by emergency shipments sourced from other warehouses.

We consider a continuous review base stock policy for SKU $i \in \mathcal{I}$ with base stock level $S_i(\geq 0)$. The fill rate $\beta_i(\geq 0)$ indicates the fraction of demand that is immediately satisfied from stock, and its calculation depends on the demand assumptions used in the model, so we come back to it in Sections 3.2.1 - 3.2.3. If no stock is available, the unsatisfied demand of SKU i is satisfied by an emergency shipment, which has a cost of $c_i^{\text{em}}(> 0)$ and takes a fixed amount of time $t_i^{\text{em}}(> 0)$. We denote the inventory holding cost per time unit per part of SKU i as $c_i^{\text{h}}(> 0)$. To avoid long downtime, the aggregate mean waiting time should not exceed a target service level W^{obj} . We discuss the mathematical formulations under different demand assumptions in the next sections.

3.2.1 Deterministic Model

We propose a mixed-integer linear optimization model for the deterministic case, where the demand during lead time $\zeta_i \in \mathbb{N}_0$ for SKU $i \in \mathcal{I}$ is deterministic. In this model, S_i is the main decision variable, while β_i serves as an auxiliary variable due to its dependence on S_i and ζ_i . The optimal inventory policy can be obtained by solving Problem (3.1):

$$\min_{\substack{S_i \in \mathbb{N}_0^n \\ \beta_i \in \mathbb{R}}} \sum_{i \in \mathcal{I}} (c_i^{\text{h}} S_i + \frac{1}{t} \zeta_i (1 - \beta_i) c_i^{\text{em}}) \tag{3.1a}$$

$$\text{s.t. } \beta_i \zeta_i \leq S_i, \quad \forall i \in \mathcal{I}, \tag{3.1b}$$

$$\sum_{i \in \mathcal{I}} (1 - \beta_i) \zeta_i t_i^{\text{em}} \leq W^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad (3.1c)$$

$$0 \leq \beta_i \leq 1, \quad \forall i \in \mathcal{I}. \quad (3.1d)$$

The objective function (3.1a) aims to minimize the total cost per time unit, where the first term is the total holding cost per time unit, and the second is the total emergency shipment cost per time unit. Notice that the term $\frac{1}{t}$ scales the emergency shipment cost during the lead time into a cost per time unit. Constraint (3.1b) states that the stock level S_i should exceed the quantity of demand satisfied from the stock during the lead time for SKU i . Constraint (3.1c) ensures that the aggregate mean waiting time remains below the target service level. Constraint (3.1d) specifies the allowed range for the item fill rate.

3

3.2.2 Stochastic Optimization Model

The stochastic optimization (SO) model is a state-of-the-art model for spare parts inventory control. This model is detailed by Van Houtum and Kranenburg (2015, p. 40). The SO model assumes that the demand per lead time follows a Poisson process with a constant rate of $m_i (> 0)$ per time unit. The dynamics of spare parts in repair for SKU i are captured through an $M|G|c|c$ queue in an infinite horizon, i.e., an Erlang loss system, where $c = S_i$ represents parallel servers, m_i represents the arrival rate, and t represents the lead time. With these assumptions, the fill rate for SKU i can be calculated using the Erlang loss function:

$$\beta_i(S_i) = 1 - \frac{\frac{1}{S_i!} (m_i t)^{S_i}}{\sum_{j=0}^{S_i} \frac{1}{j!} (m_i t)^j}. \quad (3.2)$$

We denote the total demand rate for all SKUs by $M = \sum_{i \in \mathcal{I}} m_i$. To find the optimal stock level, we solve Problem (3.3):

$$\min_{S \in \mathbb{N}_0^n} \sum_{i \in \mathcal{I}} c_i^h S_i + m_i (1 - \beta_i(S_i)) c_i^{\text{em}} \quad (3.3a)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} \frac{m_i}{M} (1 - \beta_i(S_i)) t_i^{\text{em}} \leq W^{\text{obj}}. \quad (3.3b)$$

In practice, the calculated fill rate becomes less reliable when demand rates are uncertain, especially during new product introductions. When the actual demand pattern deviates from the assumed Poisson distribution, the fill rate calculation in Equation (3.2) no longer accurately reflects the system's behavior. Additionally, since Equation (3.2) is based on the steady state of an infinite horizon, it may not be directly applicable to finite horizon settings.

3.2.3 Adaptive Robust Optimization Models

Adaptive Robust Optimization (ARO) assumes that the uncertain demand vector $\zeta = [\zeta_i]_{i \in \mathcal{I}}$ lies within a set $\mathcal{D} \subset \mathbb{N}_0^n$, which encapsulates all possible realizations of the demand vector. The decision-making process takes place in two stages. In the first stage, we decide on the stock level S_i for each SKU $i \in \mathcal{I}$, which are here-and-now variables. The auxiliary variable β_i depends on the actual demand ζ during the lead time t , hence a wait-and-see variable, which is decided in the second stage. Unlike the deterministic and stochastic models where emergency shipment costs per time unit are explicit functions of S_i and β_i (or S_i alone), in ARO models, these costs depend on the uncertain demand vector ζ . Therefore, we introduce the auxiliary variables $\eta_i(\zeta)$ to represent the emergency shipment cost per time unit for each SKU i , and η to represent the total worst-case emergency shipment cost per time unit across all SKUs.

Now, we introduce Problem (3.4), which considers the dependency of emergency shipment costs across SKUs:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta) \quad (3.4a)$$

$$\text{s.t.} \quad \frac{1}{t} \zeta_i (1 - \beta_i(\zeta)) c_i^{\text{em}} \leq \eta_i(\zeta), \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.4b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \quad (3.4c)$$

$$\beta_i(\zeta) \zeta_i \leq S_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.4d)$$

$$\sum_{i \in \mathcal{I}} (1 - \beta_i(\zeta)) \zeta_i t_i^{\text{em}} \leq W^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \quad (3.4e)$$

$$0 \leq \beta_i(\zeta) \leq 1, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.4f)$$

$$\eta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.4g)$$

In this formulation, constraint (3.4c) ensures that the sum of individual emergency shipment costs per time unit does not exceed η , which moves the uncertainty from the objective function to a constraint (Bertsimas and den Hertog, 2022, p. 26). Constraint (3.4d) establishes the relationship between S_i and $\beta_i(\zeta)$, which guarantees that the actual fill rate per lead time is at least $\beta_i(\zeta)$ considering the worst-case demand.

To make Problem (3.4) more tractable, we approximate it with a more conservative

formulation given by Problem (3.5), where we consider the worst-case emergency shipment cost for each SKU independently:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta_i) \quad (3.5a)$$

$$\text{s.t. } \frac{1}{t} \zeta_i (1 - \beta_i(\zeta)) c_i^{\text{em}} \leq \eta_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.5b)$$

$$\beta_i(\zeta) \zeta_i \leq S_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.5c)$$

$$\sum_{i \in \mathcal{I}} (1 - \beta_i(\zeta)) \zeta_i t_i^{\text{em}} \leq W^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \quad (3.5d)$$

$$0 \leq \beta_i(\zeta) \leq 1, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.5e)$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.5f)$$

Problem (3.5) is more conservative than Problem (3.4) due to the way it aggregates worst-case emergency shipment costs per time unit. In Problem (3.4), η represents the worst-case total emergency shipment cost per time unit *across all SKUs for a single scenario* $\zeta \in \mathcal{D}$, i.e., $\eta = \max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \eta_i(\zeta)$. In contrast, Problem (3.5) enforces $\eta_i = \max_{\zeta \in \mathcal{D}} \eta_i(\zeta)$ for each SKU i , and the objective sums these *individual* worst-case emergency shipment costs per time unit. Mathematically, $\sum_{i \in \mathcal{I}} \max_{\zeta \in \mathcal{D}} \eta_i(\zeta) \geq \max_{\zeta \in \mathcal{D}} \sum_{i \in \mathcal{I}} \eta_i(\zeta)$, which means Problem (3.5) overestimates the true worst-case total emergency shipment cost per time unit. For example, if \mathcal{D} contains two scenarios where $\eta_1 = 10, \eta_2 = 5$ and $\eta_1 = 5, \eta_2 = 10$, Problem (3.4) yields $\eta = 15$ (the maximum total across scenarios), while Problem (3.5) yields $\eta_1 + \eta_2 = 20$ (the sum of individual maxima). Thus, Problem (3.5) protects against unrealistic scenarios where all SKUs simultaneously face their worst-case emergency shipment costs per time unit.

Problem (3.5) provides a highly robust solution. It is particularly suitable for contexts where the consequences of stockouts are severe, such as in critical industries like healthcare or defense, where the availability of spare parts is crucial for maintaining operational continuity. Moreover, we use Problem (3.5) in Section 3.4 as part of our solution method for Problem (3.4).

Remark 3.1 If t_i^{em} are equal for any $i \in \mathcal{I}$, i.e., $t_i^{\text{em}} = t^{\text{em}}$, Constraint (3.5d) is equivalent to

$$\sum_{i \in \mathcal{I}} \beta_i(\zeta) \zeta_i \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D},$$

where $\beta^{obj} = 1 - \frac{W^{obj}}{f^{em}}$. This equivalence implies that the aggregate waiting time constraints are convertible to aggregate fill rate constraints. Therefore, we can calculate waiting times using predetermined fill rates and vice versa. This simplification is practical because companies selling expensive equipment often use a standard value for t_i^{em} for all SKUs in emergency logistics. Additionally, this simplification allows managers to focus on achieving predetermined fill rates, which are often more intuitive and easier to communicate within an organization. By ensuring that the fill rate meets the target W^{obj} , the organization can be confident that its waiting times are also under control. This equivalence is also seen in the stochastic model discussed by Van Houtum and Kranenburg (2015, p. 54).

3.3. Solution method for Problem (3.5)

The deterministic Problem (3.1) is a mixed-integer linear optimization problem that can be solved using off-the-shelf solvers. For the SO Problem (3.3), a greedy algorithm is commonly used in spare parts inventory control literature (Sherbrooke, 2006; Van Houtum and Kranenburg, 2015, Chap. 2; Basten and van Houtum, 2023). In this section, we focus on solving the ARO Problem (3.5), which is computationally challenging.

In Section 3.3.1, we reformulate the ARO problem into a deterministic counterpart, which is a mixed integer linear optimization problem (MILP) with an exponential number of constraints. To deal with the computational complexity of the problem, we show that the reformulation can be decomposed into two smaller MILPs, one of which can be solved analytically. This decomposition helps reduce the computational complexity drastically. Finally, we develop an algorithm in Section 3.3.2.

Let us first reformulate Problem (3.5) into a fixed-recourse ARO problem, where the uncertain parameter is not multiplied by a wait-and-see variable. For a given $i \in \mathcal{I}$ and a vector $\zeta \in \mathcal{D}$, we define $\epsilon_i(\zeta) := \beta_i(\zeta)\zeta_i$, which can be interpreted as the demand quantity during the lead time being immediately satisfied from stock.

Now, Problem (3.5) can be reformulated as:

$$\begin{aligned}
& \min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta_i) \\
& \text{s.t. } \zeta_i - \epsilon_i(\zeta) \leq \frac{t\eta_i}{c_i^{\text{em}}}, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
& \epsilon_i(\zeta) \leq S_i, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
& \sum_{i \in \mathcal{I}} (\zeta_i - \epsilon_i(\zeta)) t_i^{\text{em}} \leq W^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, & \forall \zeta \in \mathcal{D}, \\
& 0 \leq \epsilon_i(\zeta) \leq \zeta_i, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
& \eta_i \geq 0, & \forall i \in \mathcal{I}.
\end{aligned} \tag{3.6}$$

In Section 3.3.1, we explain how to reformulate Problem (3.6) into different problems, helping us solve it efficiently.

3.3.1 Equivalent Reformulations

To solve Problem (3.6), we first show how we can reduce the number of wait-and-see variables in Theorem 3.1.

Theorem 3.1 *Given $k \in \mathcal{I}$, let $\mathcal{I}^{n-k+1} = \{n-k+1, \dots, n\}$. Problem (3.6) is equivalent to Problem (3.7):*

$$\begin{aligned}
& \min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^{n-k} \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta_i) \\
& \text{s.t. } S_i \geq \zeta_i - \frac{t\eta_i}{c_i^{\text{em}}}, & \forall i \in \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
& \zeta_i - \epsilon_i(\zeta) \leq \frac{t\eta_i}{c_i^{\text{em}}}, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
& \epsilon_i(\zeta) \leq S_i, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
& \sum_{i=1}^{n-k} (\zeta_i - \epsilon_i(\zeta)) t_i^{\text{em}} + \sum_{i \in \alpha} (\zeta_i - S_i) t_i^{\text{em}} \\
& \leq W^{\text{obj}} \sum_{i=1}^n \zeta_i, & \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
& \zeta_i \geq \epsilon_i(\zeta) \geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, \zeta \in \mathcal{D}, \\
& \eta_i \geq 0, & \forall i \in \mathcal{I}.
\end{aligned} \tag{3.7}$$

The proof has the same line of reasoning as for Theorem 1 in Chapter 2, so we postpone it to Appendix 3.A.1. Based on this theorem, we can reduce the number of wait-and-see variables in Problem (3.6). By setting $k = n$ in this theorem, we can eliminate all the wait-and-see variables, hence the following corollary.

Corollary 3.1 *Problem (3.6) is equivalent to Problem (3.8):*

$$\min_{\substack{S_i \in \mathbb{N}_0^n \\ \eta_i \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta_i) \quad (3.8a)$$

$$\text{s.t. } S_i \geq \zeta_i - \frac{t\eta_i}{c_i^{em}}, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.8b)$$

$$\sum_{i \in \alpha} S_i t_i^{em} \geq \sum_{i \in \alpha} \zeta_i t_i^{em} - W^{obj} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \quad (3.8c)$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.8d)$$

Proof. Use Theorem 3.1 and setting $k = n$. □

In Problem (3.8), as the number of SKUs increases, the number of constraints grows exponentially. This creates computational difficulties for larger instances. Therefore, we show for Problem (3.8) how to efficiently identify the optimal stock levels. Intuitively, if for an SKU $i \in \mathcal{I}$, $S_i \geq \bar{\zeta}_i$, there should no longer be an emergency shipment cost, and any increase in S_i incurs an additional holding cost but does not contribute to the decrease in W^{obj} . In other words, intuitively, we know that $S_i \in [0, \bar{\zeta}_i]$ for any SKU $i \in \mathcal{I}$. In the next lemma, we formally prove this statement.

Lemma 3.1 *For Problem (3.8), $S_i^* \in [0, \bar{\zeta}_i]$, for any $i \in \mathcal{I}$.*

Proof. Let $(\mathbf{S}^*, \boldsymbol{\eta}^*) = (S_1^*, \dots, S_n^*, \eta_1^*, \dots, \eta_n^*)$ be an optimal solution with an objective value obj^* . By contradiction, let us assume that there exists $j \in \mathcal{I}$, where $S_j^* > \bar{\zeta}_j$, which implies $S_j^* \geq \bar{\zeta}_j + 1$. Now, we construct a new feasible solution:

$$\tilde{S}_i = \begin{cases} S_i^* & \text{if } i \neq j, \\ \bar{\zeta}_j & \text{if } i = j, \end{cases} \quad \tilde{\eta}_i = \begin{cases} \eta_i^* & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

We check the feasibility of this solution. Constraints (3.8b) and (3.8d) of Problem (3.8) are clearly satisfied. For an arbitrary subset $\alpha \subseteq \mathcal{I}$, if $j \notin \alpha$, Constraint (3.8c)

holds. If $j \in \alpha$, and $\emptyset \neq \alpha \setminus \{j\}$, from feasibility of (S^*, η^*) we have:

$$\sum_{i \in \alpha \setminus \{j\}} S_i^* t_i^{\text{em}} \geq \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \setminus \{j\}} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\}.$$

So, by adding $\bar{\zeta}_j t^{\text{em}}$ to both sides, we have

$$\begin{aligned} \bar{\zeta}_j t^{\text{em}} + \sum_{i \in \alpha \setminus \{j\}} S_i^* t_i^{\text{em}} &\geq \bar{\zeta}_j t^{\text{em}} + \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \setminus \{j\}} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\} \\ &= \max_{\zeta \in \mathcal{D}} \zeta_j t^{\text{em}} + \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \setminus \{j\}} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\} \\ &\geq \max_{\zeta \in \mathcal{D}} \left\{ \zeta_j t^{\text{em}} + \sum_{i \in \alpha \setminus \{j\}} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\} \\ &= \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\}. \end{aligned}$$

Thus, the following holds:

$$\sum_{i \in \alpha} \tilde{S}_i t_i^{\text{em}} \geq \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\},$$

which implies Constraint (3.8c) also holds. Therefore, $(\tilde{S}, \tilde{\eta})$ is feasible. The objective function value of this solution is:

$$\sum_{i \in \mathcal{I}} c_i^h \tilde{S}_i + \tilde{\eta}_i = \text{obj}^* - c_j^h (S_j^* - \bar{\zeta}_j) - \eta_j^* < \text{obj}^*,$$

which contradicts the optimality of (S^*, η^*) . Therefore, we deduce that $S_j^* \leq \bar{\zeta}_j$ for any $j \in \mathcal{I}$. \square

Lemma 3.1 shows the upper bounds on optimal stock levels. We now further characterize the solution by examining the trade-off between holding costs and emergency shipment costs per time unit. Let us define \mathcal{I}_1 as

$$\mathcal{I}_1 := \left\{ i \in \mathcal{I} \mid c_i^{\text{em}} \geq c_i^h t \right\}.$$

By definition, \mathcal{I}_1 is the set of cheap SKUs whose holding costs during the lead time are lower than the unit emergency shipment cost. From a practical perspective, maintaining a large inventory for inexpensive SKUs, i.e., SKUs $i \in \mathcal{I}_1$, offers a financial advantage. Since holding stock is inexpensive, the practitioners consider the most conservative stock levels to buffer against demand uncertainty for these SKUs.

For a given $i \in \mathcal{I}$, let $\bar{\zeta}_i$ denote the maximum value of ζ_i over all $\zeta \in \mathcal{D}$, that is, $\bar{\zeta}_i = \max_{\zeta \in \mathcal{D}} \zeta_i, \forall i \in \mathcal{I}$. The following theorem proves that maintaining a large inventory is optimal for inexpensive SKUs.

Theorem 3.2 For any optimal solution $(\mathbf{S}^*, \boldsymbol{\eta}^*) = (S_1^*, \dots, S_n^*, \eta_1^*, \dots, \eta_n^*)$ of Problem (3.8), we have $S_i^* = \bar{\zeta}_i$ for any $i \in \mathcal{I}_1$.

Proof. Let $(\mathbf{S}^*, \boldsymbol{\eta}^*) = (S_1^*, \dots, S_n^*, \eta_1^*, \dots, \eta_n^*)$ be an optimal solution with an objective value obj^* . We emphasize that since $\mathcal{D} \subset \mathbb{N}_0^n$, we know $\bar{\zeta} \in \mathbb{N}_0^n$. From Lemma 3.1, we know that $S_i^* \leq \bar{\zeta}_i$ for all $i \in \mathcal{I}$. For contradiction, let us assume that there exists $j \in \mathcal{I}_1$ where $S_j^* < \bar{\zeta}_j$, which implies $S_j^* \leq \bar{\zeta}_j - 1$. Since the solution is feasible, Constraint (3.8b) holds, implying that $\eta_j^* \geq \frac{c_j^{em}}{t}$. Now, we construct another feasible solution $(\tilde{\mathbf{S}}, \tilde{\boldsymbol{\eta}})$, where

$$\tilde{S}_i = \begin{cases} S_i^* & \text{if } i \neq j, \\ S_j^* + 1 & \text{if } i = j, \end{cases} \quad \tilde{\eta}_i = \begin{cases} \eta_i^* & \text{if } i \neq j, \\ \eta_j^* - \frac{c_j^{em}}{t} & \text{if } i = j. \end{cases}$$

This solution is feasible for Problem (3.8) since $(\mathbf{S}^*, \boldsymbol{\eta}^*)$ is feasible. The value of the objective function becomes:

$$\sum_{i \in \mathcal{I}} (c_i^h \tilde{S}_i + \tilde{\eta}_i) = obj^* + c_j^h - \frac{c_j^{em}}{t} < obj^*,$$

where the inequality holds because $j \in \mathcal{I}_1$. This contradicts the optimality of $(\mathbf{S}^*, \boldsymbol{\eta}^*)$. Hence, we conclude that for any $i \in \mathcal{I}_1$, $S_i^* \geq \bar{\zeta}_i$, and hence $\eta_i^* = 0$. Therefore, we deduce that $S_j^* = \bar{\zeta}_j$ for any $j \in \mathcal{I}_1$. \square

Theorem 3.2 asserts that for the cheap SKUs, we need to stock up regardless of W^{obj} . Therefore, we can add Constraint (3.9b) to Problem (3.8), which implies that Problem (3.8) is equivalent to Problem (3.9).

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta_i) \quad (3.9a)$$

$$\text{s.t. } S_i = \bar{\zeta}_i, \quad \forall i \in \mathcal{I}_1, \quad (3.9b)$$

$$S_i \geq \zeta_i - \frac{t\eta_i}{c_i^{em}}, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.9c)$$

$$\sum_{i \in \alpha} S_i t_i^{em} \geq \sum_{i \in \alpha} \zeta_i t_i^{em} - W^{obj} \sum_{i=1}^n \zeta_i, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \zeta \in \mathcal{D}, \quad (3.9d)$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.9e)$$

Remark 3.2 (Worst-case assortment scenario) If $c_i^{em} \geq c_i^h t$ for all $i \in \mathcal{I}$, the optimal solution for Problem (3.9) is $S_i = \bar{\zeta}_i$, $\eta_i = 0$ for all $i \in \mathcal{I}$. Consequently, Constraints (3.9c) and (3.9d) are redundant.

The advantage of Theorem 3.2 is that we can reduce the number of variables, but we still have an exponential number of constraints on n due to Constraint (3.9d), if there exists even one SKU with $c_i^{em} < c_i^h t$. Let us consider SKU in \mathcal{I}_2 , i.e., $\mathcal{I}_2 = \mathcal{I} \setminus \mathcal{I}_1$. From Remark 3.2, we know the optimal stock levels if $\mathcal{I}_2 = \emptyset$. So, from now on, we assume $\mathcal{I}_2 \neq \emptyset$. We are interested in identifying how the stock levels of cheap SKUs ($i \in \mathcal{I}_1$) influence the stock levels of expensive SKUs ($i \in \mathcal{I}_2$). The following theorem shows that these two sets are fully independent despite their stock levels being intertwined via Constraint (3.9d).

Lemma 3.2 The optimal solution for any SKU $i \in \mathcal{I}_2$ in Problem (3.9) can be obtained by solving Problem (3.10):

$$\min_{\substack{s \in \mathbb{N}_0^{|\mathcal{I}_2|} \\ \eta \in \mathbb{R}^{|\mathcal{I}_2|}}} \sum_{i \in \mathcal{I}_2} (c_i^h S_i + \eta_i) \quad (3.10a)$$

$$\text{s.t. } S_i \geq \zeta_i - \frac{t\eta_i}{c_i^{em}}, \quad \forall i \in \mathcal{I}_2, \zeta \in \mathcal{D}, \quad (3.10b)$$

$$\sum_{i \in \alpha} S_i t_i^{em} \geq \sum_{i \in \alpha} \zeta_i t_i^{em} - W^{obj} \sum_{i \in \mathcal{I}_2} \zeta_i, \quad \forall \alpha \subseteq \mathcal{I}_2, \alpha \neq \emptyset, \zeta \in \mathcal{D}, \quad (3.10c)$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}_2. \quad (3.10d)$$

Proof. According to Theorem 3.2, for any optimal solution (S^*, η^*) , we have $S_i^* = \bar{\zeta}_i$ and $\eta_i^* = 0$ for all $i \in \mathcal{I}_1$. So, we only need to show that Constraint (3.9d) is redundant if $\alpha \cap \mathcal{I}_1 \neq \emptyset$. We know Constraint (3.9d) is redundant if $\alpha \subseteq \mathcal{I}_1$ since $S_i^* = \bar{\zeta}_i$ for any $i \in \mathcal{I}_1$. So, let $\alpha \subseteq \mathcal{I}$ such that $\alpha \cap \mathcal{I}_1 \neq \emptyset$ and $\alpha \cap \mathcal{I}_2 \neq \emptyset$. From Constraint (3.9d) we have:

$$\sum_{i \in \alpha} S_i t_i^{em} \geq \sum_{i \in \alpha} \zeta_i t_i^{em} - W^{obj} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D},$$

which is equivalent to

$$\sum_{i \in \alpha \cap \mathcal{I}_1} S_i t_i^{em} + \sum_{i \in \alpha \cap \mathcal{I}_2} S_i t_i^{em} \geq \sum_{i \in \alpha \cap \mathcal{I}_1} \zeta_i t_i^{em} + \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{em} - W^{obj} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D},$$

$$\begin{aligned} &\Leftrightarrow \sum_{i \in \alpha \cap \mathcal{I}_1} \bar{\zeta}_i t_i^{\text{em}} + \sum_{i \in \alpha \cap \mathcal{I}_2} S_i t_i^{\text{em}} \geq \sum_{i \in \alpha \cap \mathcal{I}_1} \zeta_i t_i^{\text{em}} + \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \\ &\Leftrightarrow \sum_{i \in \alpha \cap \mathcal{I}_2} S_i t_i^{\text{em}} \geq \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \cap \mathcal{I}_1} (\zeta_i - \bar{\zeta}_i) t_i^{\text{em}} + \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\}. \end{aligned}$$

Given that $\sum_{i \in \alpha \cap \mathcal{I}_1} (\zeta_i - \bar{\zeta}_i) t_i^{\text{em}} \leq 0$, for any $\zeta \in \mathcal{D}$, we obtain:

$$\begin{aligned} &\max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\} \geq \\ &\max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \cap \mathcal{I}_1} (\zeta_i - \bar{\zeta}_i) t_i^{\text{em}} + \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\} \end{aligned}$$

By defining $\alpha' = \alpha \cap \mathcal{I}_2$ and denoting $b_{\alpha'} = \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha'} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\}$, we then obtain:

$$\sum_{i \in \alpha'} S_i t_i^{\text{em}} \geq b_{\alpha'} \geq \max_{\zeta \in \mathcal{D}} \left\{ \sum_{i \in \alpha \cap \mathcal{I}_1} (\zeta_i - \bar{\zeta}_i) t_i^{\text{em}} + \sum_{i \in \alpha \cap \mathcal{I}_2} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i \right\}.$$

This implies that for any $\alpha \subseteq \mathcal{I}_2$ where $\alpha \cap \mathcal{I}_1 \neq \emptyset$ and $\alpha \cap \mathcal{I}_2 \neq \emptyset$, Constraint (3.9d) is redundant. \square

Lemma 3.2 shows that the optimal stock levels for SKUs $i \in \mathcal{I}_2$ can be determined independently of SKUs in \mathcal{I}_1 , which allows us to reduce the computational complexity drastically. In Appendix 3.A.2, we provide an illustrative example to further illustrate the results of Theorem 3.2 and Lemma 3.2.

From now on, we focus on finding characteristics of Problem (3.10) that can help us solve it more efficiently. From Problem (3.10), we know that $\eta_i^* = \max\{0, \frac{c_i^{\text{em}}(\bar{\zeta}_i - S_i)}{t}\}$ for any $i \in \mathcal{I}_2$. Due to Lemma 3.1, we see that $S_i \leq \bar{\zeta}_i$ for any $i \in \mathcal{I}_2$. Therefore, we have $\eta_i^* = \frac{c_i^{\text{em}}(\bar{\zeta}_i - S_i)}{t}$. By substituting this expression for η_i^* into the objective function of Problem (3.10), the optimal stock level of SKU i for any $i \in \mathcal{I}_2$ can be achieved by solving:

$$\min_{S \in \mathbb{N}_0^{|\mathcal{I}_2|}} \sum_{i \in \mathcal{I}_2} \left(c_i^{\text{h}} - \frac{c_i^{\text{em}}}{t} \right) S_i + \frac{c_i^{\text{em}}}{t} \bar{\zeta}_i \quad (3.11a)$$

$$\text{s.t.} \quad \sum_{i \in \alpha} S_i t_i^{\text{em}} \geq -W^{\text{obj}} \sum_{i \in \mathcal{I}_2} \zeta_i + \sum_{i \in \alpha} \zeta_i t_i^{\text{em}}, \quad \forall \alpha \subseteq \mathcal{I}_2, \alpha \neq \emptyset, \zeta \in \mathcal{D}. \quad (3.11b)$$

The size of set \mathcal{I}_2 determines the tractability of Problem (3.11). For small \mathcal{I}_2 , one can solve Problem (3.11) to optimality using off-the-shelf solvers, such as Gurobi. However, for large \mathcal{I}_2 , Problem (3.11) remains challenging due to Constraints (3.11b). To

address this, we introduce approximation approaches in Section 3.3.2 that provide near-optimal solutions with reduced complexity.

3.3.2 Approximation Method

Let us focus on determining the stock levels for SKUs in \mathcal{I}_2 , as Theorem 3.2 provides the optimal stock levels for SKUs in \mathcal{I}_1 . Leveraging the structure of the optimal policy for Problem (3.11) for SKUs in \mathcal{I}_2 and inspired by the algorithms proposed in Chapter 2, we design an algorithm to find approximate solutions when $|\mathcal{I}_2|$ is large. More specifically, we show how we can use a preprocessing step to significantly improve the efficiency of the algorithm. We then propose a fast heuristic based on the algorithm.

Let $\theta_i := (c_i^h - \frac{c_i^{\text{em}}}{t_i})/t_i^{\text{em}}$, for any $i \in \mathcal{I}_2$. To get an intuition, let us consider two cases.

- Case 1: SKUs with the same t_i^{em} , but different $c_i^h - \frac{c_i^{\text{em}}}{t_i}$. In this case, a larger θ_i indicates a higher cost of holding stock than performing an emergency shipment, which means that using emergency shipments is more cost-efficient.
- Case 2: SKUs with the same $c_i^h - \frac{c_i^{\text{em}}}{t_i}$, but different t_i^{em} . In this case, the SKU with a smaller t_i^{em} (and thus larger θ_i) poses less stockout risk when using emergency shipments compared to an SKU with a larger t_i^{em} , which means that emergency shipments are more favorable than holding stock.

Summarizing, a larger θ can be interpreted as more favoring emergency shipments.

Our algorithm starts by ordering the SKUs such that $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$, where $m = |\mathcal{I}_2|$. For simplicity, let us denote the right-hand side of Constraint (3.11b) as b_α , i.e., $b_\alpha := \max_{\zeta \in \mathcal{D}} \left\{ -W^{\text{obj}} \sum_{i \in \mathcal{I}_2} \zeta_i + \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} \right\}$. Leveraging the constraints outlined in Equation (3.11b), we introduce a preprocessing step to enhance the efficiency of the algorithm. We know from Lemma 3.1 that $S_i \leq \bar{\zeta}_i$, for any $i \in \mathcal{I}_2$. So, in the preprocessing step, we calculate $\lceil \frac{b_{\{i\}}}{t_i^{\text{em}}} \rceil$. Since we have the constraint $S_i \geq \lceil \frac{b_{\{i\}}}{t_i^{\text{em}}} \rceil$, if $\bar{\zeta}_i = \lceil \frac{b_{\{i\}}}{t_i^{\text{em}}} \rceil$, we know that $S_i^* = \lceil \frac{b_{\{i\}}}{t_i^{\text{em}}} \rceil$. It is worth noting that the number of SKUs satisfying $\bar{\zeta}_i = \lceil \frac{b_{\{i\}}}{t_i^{\text{em}}} \rceil$ is intrinsically linked to how the uncertainty set is constructed. We delve deeper into this aspect in Section 3.5.

For the remaining SKUs whose stock levels are not determined in the preprocessing step, we proceed as follows. For SKU 1, we consider the lowest possible stock level,

which is $\lceil \frac{b_{\{1\}}}{t_1^{\text{em}}} \rceil$. We then go to SKU 2 and have two lower bounds: $S_2 \geq \lceil \frac{b_{\{2\}}}{t_2^{\text{em}}} \rceil$ and $S_2 \geq \lceil \frac{b_{\{1,2\}} - S_1 t_1^{\text{em}}}{t_2^{\text{em}}} \rceil$. So, the stock level of SKU 2 is the maximum of these two bounds. We continue this process until we have the stock levels for all m SKUs in \mathcal{I}_2 .

Algorithm ISP *An algorithm to solve Problem (3.11) including preprocessing.*

```

1:  $\theta_i := (c_i^h - \frac{c_i^{\text{em}}}{t}) / t_i^{\text{em}}$ 
2: Sort SKUs  $i \in \mathcal{I}_2$  according to the descending order of  $\theta_i$ 
3:  $\mathcal{K} := \{k \in \{1, \dots, m\} \mid \bar{\zeta}_k := \lceil \frac{b_{\{k\}}}{t_k^{\text{em}}} \rceil\}$ 
4: for  $k = 1, \dots, m$ 
5:   if  $k \in \mathcal{K}$ 
6:      $S_k := \bar{\zeta}_k$ 
7:      $\eta_k := 0$ 
8:   else
9:      $\Omega_k := \{\alpha \subseteq \{1, \dots, k\} \setminus \mathcal{K} : \alpha \neq \emptyset, k \in \alpha\}$ 
10:     $S_k := \max\{\max_{\alpha \in \Omega_k} \{[(b_\alpha - \sum_{i \in \alpha \setminus \{k\}} S_i t_i^{\text{em}}) / t_k^{\text{em}}], 0\}\}$ 
11:     $\eta_k := \max\{c_i^{\text{em}}(\bar{\zeta}_k - S_k), 0\}$ 
12:  endfor
13: Output:  $S, \eta$ 

```

} Preprocessing

The pseudocode of the algorithm, called Iterative Stocking including Preprocessing (ISP), is provided in Algorithm ISP. In Appendix 3.A.3, we show that ISP provides an optimal solution to Problem (3.11) under some conditions.

Algorithm ConGAP *An algorithm to solve Problem (3.11) with $j(\leq m)$ layers.*

```

1:  $\theta_i := (c_i^h - \frac{c_i^{\text{em}}}{t}) / t_i^{\text{em}}$ 
2: Sort SKUs  $i \in \mathcal{I}_2$  according to the descending order of  $\theta_i$ 
3:  $\mathcal{K} := \{k \in \{1, \dots, m\} \mid \bar{\zeta}_k := \lceil \frac{b_{\{k\}}}{t_k^{\text{em}}} \rceil\}$ 
4: for  $k = 1, \dots, m$ 
5:   if  $k \in \mathcal{K}$ 
6:      $S_k := \bar{\zeta}_k$ 
7:      $\eta_k := 0$ 
8:   else
9:      $\Omega_k := \{\alpha \subseteq \{1, \dots, k\} : \alpha \neq \emptyset, k \in \alpha, |\alpha| \leq j\}$ 
10:     $S_k := \max\{\max_{\alpha \in \Omega_k} \{[(b_\alpha - \sum_{i \in \alpha \setminus \{k\}} S_i t_i^{\text{em}}) / t_k^{\text{em}}], 0\}\}$ 
11:     $\eta_k := \max\{c_i^{\text{em}}(\bar{\zeta}_k - S_k), 0\}$ 
12:  endfor
13: Output:  $S, \eta$ 

```

} Preprocessing

To further improve computational efficiency, we leverage the concept of ConGA,

introduced in Chapter 2, to find a heuristic solution for SKU $k \notin \mathcal{K}$ (steps 8 to 11 of ISP). We refer to this enhanced approach as ConGA including Preprocessing (ConGAP), which can achieve near-optimal solutions based on the results of Chapter 2. The pseudocode of the ConGAP algorithm is given above. For a given value $j \leq m$ (which we call the layer of ConGAP), we consider $\sum_{l=1}^j \binom{m}{l}$ constraints of Problem (3.11) and restrict ourselves to $\alpha \subseteq \mathcal{I}_2$ with at most j members. When $j = m$, ConGAP and ISP coincide.

3

3.4. Solution method for Problem (3.4)

In this section, we propose an algorithm to solve the ARO Problem (3.4). We first reformulate Problem (3.4) into a fixed-recourse ARO problem:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta) \quad (3.12a)$$

$$\text{s.t. } \zeta_i - \epsilon_i(\zeta) \leq \frac{t\eta_i(\zeta)}{c_i^{\text{em}}}, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.12b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \quad (3.12c)$$

$$\epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.12d)$$

$$\sum_{i \in \mathcal{I}} (\zeta_i - \epsilon_i(\zeta)) t_i^{\text{em}} \leq W^{\text{obj}} \sum_{i \in \mathcal{I}} \zeta_i, \quad \forall \zeta \in \mathcal{D}, \quad (3.12e)$$

$$0 \leq \epsilon_i(\zeta) \leq \zeta_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.12f)$$

$$\eta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}. \quad (3.12g)$$

After eliminating all wait-and-see variables, we have the following theorem:

Theorem 3.3 *Problem (3.12) is equivalent to Problem (3.13):*

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta) \quad (3.13a)$$

$$\text{s.t. } \sum_{i \in \alpha} \frac{c_i^{\text{em}}}{t} (\zeta_i - S_i) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \quad (3.13b)$$

$$\sum_{i \in \alpha} S_i t_i^{\text{em}} \geq \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \quad (3.13c)$$

$$\eta \geq 0. \quad (3.13d)$$

Proof. See Appendix 3.A.4. □

To solve this problem efficiently, we develop a preprocessing approach consisting of two key steps. In the first step, we establish bounds on optimal stock levels by calculating a lower bound from the lost sales Problem (2.2) in Chapter 2 and an upper bound from the conservative Problem (3.5). For SKUs where these bounds coincide, we can directly determine their near-optimal stock levels without further computation. In the second step, for the remaining SKUs, we employ existing approximation methods, including the static approximation approach and the affine decision rule approach, to determine near-optimal solutions.

This preprocessing approach has shown remarkable effectiveness in practice. In a case study with ASML, we determine near-optimal stock levels for 2,345 SKUs out of 2,428 SKUs through the first step alone. Only 83 SKUs required further calculations using the approximation methods, drastically reducing the computational burden of the original problem.

Given that Problems (3.5) and (3.4) yield identical solutions for about 97% of the SKUs, and solving Problem (3.5) provides sufficient insight while being computationally more tractable, we conduct our detailed case study in Section 3.6 using only Problem (3.5).

3.5. Uncertainty Sets

We first recall two types of uncertainty sets that are typically considered in the literature for this type of problem: the box and budget uncertainty sets (see, e.g., Bertsimas and Thiele, 2006; Ardestani-Jaafari and Delage, 2016; Lim and Wang, 2017), which we jointly refer to as the classical uncertainty sets. Then, we explain how to incorporate the initial failure rate estimated by reliability engineers into constructing the uncertainty set, particularly when historical demand data is limited. Let $\underline{d}, \bar{d} \in \mathbb{N}_0^n$, and $\underline{\Gamma}, \bar{\Gamma} \in \mathbb{N}_0^{2^m-1}$, such that $\underline{d} \leq \bar{d}$, and $\underline{\Gamma} \leq \bar{\Gamma}$.

3.5.1 Classical Uncertainty sets

Box uncertainty set: We use the box uncertainty set if the only knowledge about ζ_i is that $\underline{d}_i \leq \zeta_i \leq \bar{d}_i$, for any $i \in \mathcal{I}$. Here, \underline{d}_i and \bar{d}_i denote the i_{th} components of \underline{d} and \bar{d} , respectively. The main assumption behind the box uncertainty set is the

independence of the demands of different SKUs. We can write the uncertainty set as:

$$\mathcal{D}^{\text{box}} = \{\zeta \in \mathbb{N}_0^n : \underline{d}_i \leq \zeta_i \leq \bar{d}_i, \forall i \in \mathcal{I}\}.$$

Price-based budget uncertainty set: In practice, it is highly unlikely that all SKUs experience their highest or lowest demand simultaneously, resulting in the corner points of a box uncertainty set being highly unlikely. Therefore, similar to the approach taken in Bertsimas and Sim (2004), we avoid these points by cutting them out in a budget uncertainty set.

3

Since companies usually need to handle thousands of SKUs, it is critical to ensure computational efficiency when adding cuts. Typically, SKUs falling under category \mathcal{I}_1 account for a high proportion of total demand and are identified by their low holding cost. Although SKUs in \mathcal{I}_2 represent only a small proportion of total demand, they have significant holding costs that can greatly affect total expenditures. Therefore, in the price-based budget uncertainty set, we only exclude extreme demand scenarios for SKU $i \in \mathcal{I}_2$ by considering the following uncertainty set:

$$\mathcal{D}^{\text{bud}} = \{\zeta \in \mathbb{N}_0^n : \underline{d}_i \leq \zeta_i \leq \bar{d}_i, \forall i \in \mathcal{I}_1, \underline{\Gamma}_\alpha \leq \sum_{i \in \alpha} \zeta_i \leq \bar{\Gamma}_\alpha, \forall \alpha \subseteq \mathcal{I}_2, \alpha \neq \emptyset\}.$$

Here, for any $i \in \mathcal{I}_2$, we have $\underline{\Gamma}_{\{i\}} = \underline{d}_i$, $\bar{\Gamma}_{\{i\}} = \bar{d}_i$. So, $\mathcal{D}^{\text{bud}} \subseteq \mathcal{D}^{\text{box}}$. The values of Γ are determined based on correlations between SKU demands. We use the algorithm presented in Appendix 2.A.6 to approximate b_α containing expensive SKUs.

The idea behind the added cuts is to incorporate the well-known concept of failure interactions on multi-component systems (Murthy and Nguyen, 1985). The price-based budget uncertainty set accounts for the interdependencies among the demand behaviors of expensive SKUs while avoiding the computational complexity of adding cuts for all SKUs as in Chapter 2.

3.5.2 Incorporating Initial Failure Rate (IFR)

The IFR is used to quantify the expected failure rate of spare parts when historical demand data is limited. It relies on the reliability engineer's knowledge and experience with similar machines. Bayesian methods have been used in reliability engineering to update prior beliefs about failure rates with observed data (Hamada et al., 2008; Liu and Abeyratne, 2019). Our approach is inspired by the Bayesian idea of combining prior information (in this case, the IFR) with observed data (historical

demand). However, instead of using the combined information to obtain a point estimate of the expected demand rate based on assumed demand distributions, we use it to construct uncertainty sets that provide bounds on the plausible values of the demand rate without assuming any particular distribution. This distinguishes our approach by focusing specifically on the construction of uncertainty sets for spare parts demand when historical data is limited.

We propose a phased approach to incorporate the IFR into the construction of uncertainty sets for spare parts demand. In the first phase, when no historical demand data is available, we rely solely on the IFR to construct the uncertainty set. As more historical demand data is accumulated in later phases, we gradually reduce the weight given to the IFR. By incorporating the IFR in a phased manner, our method provides a tailored solution for constructing uncertainty sets in spare parts inventory management, starting at the NPI stage and continuing throughout the product lifecycle.

The conservativeness of a solution is related to the calculation of \underline{d} , \bar{d} , $\underline{\Gamma}$, and $\bar{\Gamma}$. We estimate \underline{d} and \bar{d} based on the use of historical demand data (HIS) and the IFR, while $\bar{\Gamma}$, $\underline{\Gamma}$ are estimated only based on the HIS. We show how these bounds are calculated and how the HIS and IFR data are integrated in Section 3.6.2. The IFR is typically overestimated for most SKUs in practice when few or no failures are observed. Despite this, it remains practical to use the IFR for SKUs belonging to \mathcal{I}_1 (considered inexpensive), as the overestimation has a limited impact on the holding cost for these SKUs. However, the IFR overestimation substantially affects the holding cost for SKUs belonging to \mathcal{I}_2 (considered expensive). To mitigate this impact, the price-based budget uncertainty set implies a relatively low utilization of the IFR, resulting in lower stock levels for SKUs in \mathcal{I}_2 . Our case study in Section 3.6 further illustrates the notable difference between using the box uncertainty set, which does not mitigate the impact of IFR overestimation for expensive SKUs, and the price-based budget uncertainty set.

3.6. ASML Case study

With this case study, we evaluate the efficacy of the ARO Problem in enhancing spare parts inventory control at ASML, the world's largest supplier of photolithography systems for the semiconductor industry and the sole provider of extreme ultraviolet (EUV) lithography machines (CNBC, 2022). As the most highly valued

European tech company as of May 2024 (Companies Market Cap, 2024), ASML must ensure a robust spare parts inventory to prevent costly downtimes in the fabrication process for its customers, such as TSMC, Intel, and Samsung. Currently, ASML uses a stochastic optimization (SO) approach for its spare parts inventory control system. This approach applies Problem (3.3) by first assuming Poisson demand processes, then estimating demand rates, and finally employing a greedy algorithm to solve the problem. This SO model is state-of-the-art in the industry, as discussed by Lamghari-Idrissi et al. (2022). However, the variability in demand for SKUs, as indicated by the Coefficient of Variation (CV) analysis presented in Section 1.1, suggests that the SO approach may not always be reliable during the new product introduction (NPI) phase. Therefore, we compare the performance of the RO method to ASML's current SO approach.

The case study consists of five main parts. Section 3.6.1 discusses data preparation and model setup. In Section 3.6.2, we show how the IFR estimated by reliability engineers is used in the construction of uncertainty sets at ASML. In Section 3.6.3, we compare the solutions obtained using the box uncertainty set (RO-box) against the price-based budget uncertainty set (RO-bud) to determine the most suitable uncertainty set for the RO model. In Section 3.6.4, we compare the solutions using the RO model with the SO model, which is used at ASML. In Section 3.6.5, we perform a sensitivity analysis on t_i^{em} , c_i^{em} , and the exact method of incorporating the IFR in the uncertainty set.

3.6.1 Data Preparation and Model Setup

We begin by filtering SKUs that do not have any historical demand or IFR data, as there is nothing we can do if we have no data at all. This filtering process reduces the total number of SKUs from 3,144 to 2,428. After filtering, we have a dataset of 2,428 SKUs covering the first three years of demand data for one type of machine at ASML. Figure 1.1 (in Section 1.1) depicts the annual historical demand distribution for these SKUs, which indicates that over 80% of the SKUs have an annual demand quantity of three or fewer.

ASML standardizes the emergency logistics parameters for all SKUs shipped from specific origins to specific destinations. To reflect a realistic range of parameters while maintaining confidentiality, we use the following representative values. The replenishment lead time t is 60 days, the unit emergency shipment cost c_i^{em} is 750 Euros, and the unit emergency shipment time t_i^{em} is 1 day for each SKU $i \in \mathcal{I}$.

The holding cost c_i^h per year is 18% of the price of a new spare part, denoted by c_i^a (> 0). As a result, approximately 93% of the SKUs belong to \mathcal{I}_1 (considered cheap), while 7% belong to \mathcal{I}_2 (considered expensive). ASML typically aims for high service performance ($W^{\text{obj}} \leq 0.1$ days if t_i^{em} is 1).

To make the stocking decisions for SKUs, we consider the historical demand data available over the first T quarters (where $T \geq 2$). We then evaluate the performance of this decision based on the available data in quarters $T + 1$ and $T + 2$. For example, when $T = 2$, we use the available data from the first two quarters to create an uncertainty set and determine the stock levels. We then evaluate the stock levels based on the realized demand in quarters 3 and 4. As a result, we begin making stock decisions from the third quarter onwards.

To obtain solutions for the RO model, we use ConGAP with 3 layers. ConGAP is computationally efficient and can generate solutions for 2,428 SKUs in just a few minutes. Notably, during the preprocessing step, we can find the optimal stock level for around 30% of the SKUs in \mathcal{I}_2 . For the remaining SKUs in \mathcal{I}_2 , we apply the subsequent steps of ConGAP. In addition to the robust optimization approach, we consider the SO Problem (3.3) to capture the current method used in ASML and apply the greedy algorithm (SO-Greedy) (Basten and Van Houtum, 2014) to solve the SO Problem (3.3).

3.6.2 Incorporating IFR into Uncertainty Set Construction

Figure 3.1 shows how to incorporate the IFR and HIS for the point forecasts of the demand rate at ASML. This approach is based on ASML's current way of working and uses two criteria: historical demand quantity and machine years. The latter is defined as a number of machines using a specific SKU times the number of years these machines have been in operation. The approach for each SKU starts with IFR-based demand estimation when there is limited historical demand data in the very early stages of the machine lifecycle. As the machines accumulate years in operation and demand increases, the approach gradually increases the weight of the historical demand. Therefore, we have an intermediate stage where IFR and HIS are weighted equally and a final stage where demand estimation relies entirely on HIS. The term *IFR+Review* refers to the periodic reassessment of the IFR as operational experience increases. Over time, the approach increasingly relies on HIS-based demand estimation.

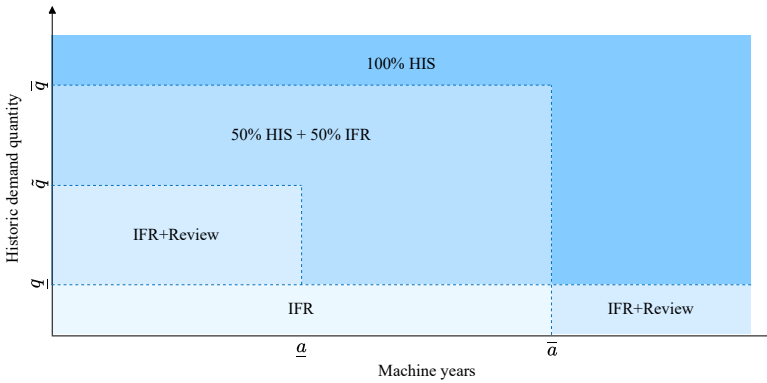


Figure 3.1: The IFR transition approach.

We adopt a similar approach to incorporate the IFR in our uncertainty set construction. First, let \underline{IFR} and \overline{IFR} denote the lower and upper bounds of demand estimated based on IFR. Similarly, let \underline{HIS} and \overline{HIS} denote the lower and upper bounds of demand estimated based on historical demand data. We show how to estimate demand using HIS in Appendix 3.A.5. We categorize SKUs into three main groups when incorporating the IFR and HIS in the construction of the uncertainty set.

- Category Full-IFR: The first category involves SKUs for which we only use IFR. In this case study, we set $\underline{d} = \underline{IFR} = 0.5 \times IFR$ and $\bar{d} = \overline{IFR} = 1.5 \times IFR$ to ensure that the mean remains the same as IFR, which is the input to the SO model. In Appendix 3.A.6, we perform sensitivity analysis, which shows that varying \underline{IFR} and \overline{IFR} has a minimal effect on the performance of the RO model when W^{obj} is small, and that the RO approach consistently outperforms the SO approach across the parameter ranges tested.
- Category MIX: The second category involves SKUs for which both IFR and HIS are used. We calculate the upper and lower bounds on demand as follows: $\underline{d} = 50\% \times \underline{IFR} + 50\% \times \underline{HIS}$ and $\bar{d} = 50\% \times \overline{IFR} + 50\% \times \overline{HIS}$.
- Category Full-HIS: The last category involves SKUs for which we only use HIS. In this case, we simply set \underline{d} and \bar{d} to \underline{HIS} and \overline{HIS} , respectively.

We set $q = 1$, $\tilde{q} = 3$, $\bar{q} = 6$, $a = 50$, and $\bar{a} = 100$ for Figure 3.1. We derive these

values from ASML's expert opinion. Based on this, we find that when $T \in \{2, 4\}$, more than 90% of the SKUs are in Categories Full-IFR and MIX, i.e., relying on the IFR to construct the uncertainty set. When $T \geq 6$, about 70% of the SKUs are in Category Full-IFR, and the remaining are in Category Full-HIS. Because the historical demand for SKUs in Category Full-IFR continues to be almost zero as the machine ages, inventory planners typically assume that future demand is unlikely as the SKUs proved reliable. Therefore, they set the IFR to zero and manually set a minimal stock level for these SKUs to account for any unexpected future demands. In our study, we set the minimal stock level to zero for these SKUs in both the RO and SO models. This means that when $T \geq 6$, we only need to determine stock levels for SKUs in Category Full-HIS, and stock levels for SKUs in Category Full-IFR will be 0 in our experiments. We refer to Appendix 3.A.6 for more information on evaluating the potential benefits of incorporating the IFR compared to relying solely on the HIS for all SKUs.

3.6.3 Comparison of Uncertainty Sets

We evaluate the effectiveness of solutions obtained from the two classical uncertainty sets that we discussed in Section 3.5.1: RO-box and RO-bud. We examine the robust solutions for values of $T \in \{2, 4, 6, 8, 10\}$. As companies like ASML aim for high service performance, we set $W^{\text{obj}} \in \{0.05, 0.1\}$ days to align with the goal of $W^{\text{obj}} \leq 0.1$ days. However, we also include $W^{\text{obj}} \in \{0.15, 0.2\}$ to observe the impact of higher waiting time targets on costs and performance.

Table 3.1 presents the realized performance metrics, including the simulated mean waiting time for all demands and the total simulated costs. These metrics are evaluated using the available data in quarters $T + 1$ and $T + 2$ (test set) for both RO-box and RO-bud, where T is the number of quarters used in the training set. The simulated mean waiting times for RO-box mostly remain below W^{obj} , except for $T = 2$ or when $W^{\text{obj}} = 0.05$ days. For instance, when $T = 4$, the simulated mean waiting time remains constant at 0.091 days for RO-box as W^{obj} varies from 0.05 to 0.2 days. Compared with the RO-box, the RO-bud achieves slightly higher simulated mean waiting times but offers substantial cost savings. For instance, when $T = 4$, the simulated mean waiting time only slightly increases from 0.096 to 0.112 days for RO-bud as W^{obj} varies from 0.05 to 0.2 days. This is because the demand for SKUs $i \in \mathcal{I}_2$ is a small fraction of the total demand, causing W^{obj} to have little influence on the simulated mean waiting time for all demands. The

Table 3.1: Comparison of simulated mean waiting time and simulated total cost using the RO-box and RO-bud.

		RO-box				RO-bud			
		0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2
Simulated mean waiting time (days)	T \ W^{obj} (days)								
	2	0.195	0.195	0.195	0.195	0.210	0.212	0.218	0.220
	4	0.091	0.091	0.091	0.091	0.096	0.100	0.104	0.112
	6	0.107	0.107	0.107	0.107	0.108	0.110	0.116	0.125
	8	0.064	0.064	0.064	0.064	0.065	0.068	0.070	0.073
	10	0.035	0.035	0.035	0.035	0.035	0.037	0.040	0.044
Simulated total cost ($\times 1,000$ Euros)	2	35,941	35,615	34,977	34,623	5,029	4,723	4,593	4,509
	4	41,989	41,682	40,994	40,614	7,682	6,641	5,788	5,445
	6	6,924	6,919	6,914	6,909	5,397	4,337	3,401	2,695
	8	8,350	8,345	8,340	12,173	6,685	5,539	4,591	3,731
	10	14,564	14,559	14,554	14,549	8,877	6,698	5,388	4,164

Note: For $T \in \{2, 4\}$, most SKUs are in Categories Full-IFR and MIX. When $T \geq 6$, we only need to consider SKUs in Category Full-HIS.

slight increase in simulated mean waiting time at $T = 6$ is due to extreme demand for some SKUs, which the uncertainty sets based on historical demand data and estimates can not capture. After incorporating this extreme demand at $T = 6$, the uncertainty set is adjusted accordingly and can better prepare for potential worst-case scenarios, resulting in shorter simulated mean waiting times for $T > 6$.

As mentioned at the end of Section 3.6.2, more than 90% of the SKUs rely on the IFR to construct the uncertainty set when $T \leq 4$. During this period, the simulated total costs for the RO-box are quite high due to the overestimation of the IFR values for most SKUs. In contrast, using RO-bud yields notably lower simulated total costs. This cost reduction is because the RO-bud adds cuts to the RO-box based solely on historical demand data, thereby mitigating demand overestimation risk for SKUs with low historical demand but high IFR values.

Overall, the RO-bud results in more cost-efficient solutions than the RO-box, particularly when incorporating the IFR at the beginning of the product life cycle. Therefore, in Sections 3.6.4 and 3.6.5, we further analyze the performance of the RO model by utilizing the RO-bud solution.

3.6.4 Comparison of RO and SO models

We compare the performance of the RO and SO models for the same pairs of quarters outlined earlier, where $T \in \{2, 4, 6, 8, 10\}$. Figure 3.2 presents the trade-off between the simulated mean waiting time and the simulated total cost of the

solutions to both models. A black line with dashed circles on the graph shows the SO model's performance, while a distinct blue line with solid circles represents the RO solution. We also show the impact of adjusting W^{obj} , which takes values from the set $W^{\text{obj}} \in \{0.05, 0.1, 0.15, 0.2\}$ days. The RO solutions under different W^{obj} are displayed using a blue-to-green gradient, while the SO solutions are displayed using a red-to-yellow gradient.

Figure 3.2 shows that robust solutions consistently outperform stochastic ones. The simulated mean waiting time for SO-Greedy can meet the service target when W^{obj} is high. However, the robust solutions can achieve the service target even when W^{obj} is low, which aligns with ASML's needs. The robust solution provides the most benefit when $T = 6$. For the same simulated total cost, the simulated mean waiting time of the robust solution is up to 0.15 days, so over 3.5 hours, shorter than that of SO-Greedy. Considering that a breakdown of lithography systems at ASML can result in losses of up to 72,000 Euros per hour (ASML, 2014), it can be estimated that the robust solution has the potential to save over $72,000 \times 3.5 = 252,000$ Euros in lost production per breakdown of an expensive lithography system compared to SO-Greedy. For $T \in \{4, 8, 10\}$, the simulated mean waiting time of the robust solution remains more than 0.10 days shorter than that of SO-Greedy, potentially saving over 170,000 Euros in lost production per breakdown of an expensive lithography system.

The performance of the SO solutions heavily depends on the accuracy of the predicted demand using HIS, which is particularly clear when SKUs experience unexpectedly high demand at $T = 6$. Furthermore, approximately half of the SKUs do not follow a Poisson demand process, resulting in longer simulated mean waiting times for SO solutions compared to previous quarters.

3.6.5 Sensitivity Analysis

In this section, we perform sensitivity analysis focusing on five key elements of our model: t_i^{em} , c_i^{em} , q , \bar{q} , and $\bar{\bar{q}}$. Emergency shipments at ASML are inherently variable in costs and timeframes, influenced by the geographical distance between warehouses and part suppliers. By examining t_i^{em} and c_i^{em} , we aim to understand the impact of emergency logistics parameters on the performance of both RO and SO models. q , \bar{q} , and $\bar{\bar{q}}$ shown in Figure 3.1 are crucial for the integration of HIS and IFR. We also investigate its applicability to the RO and SO models.

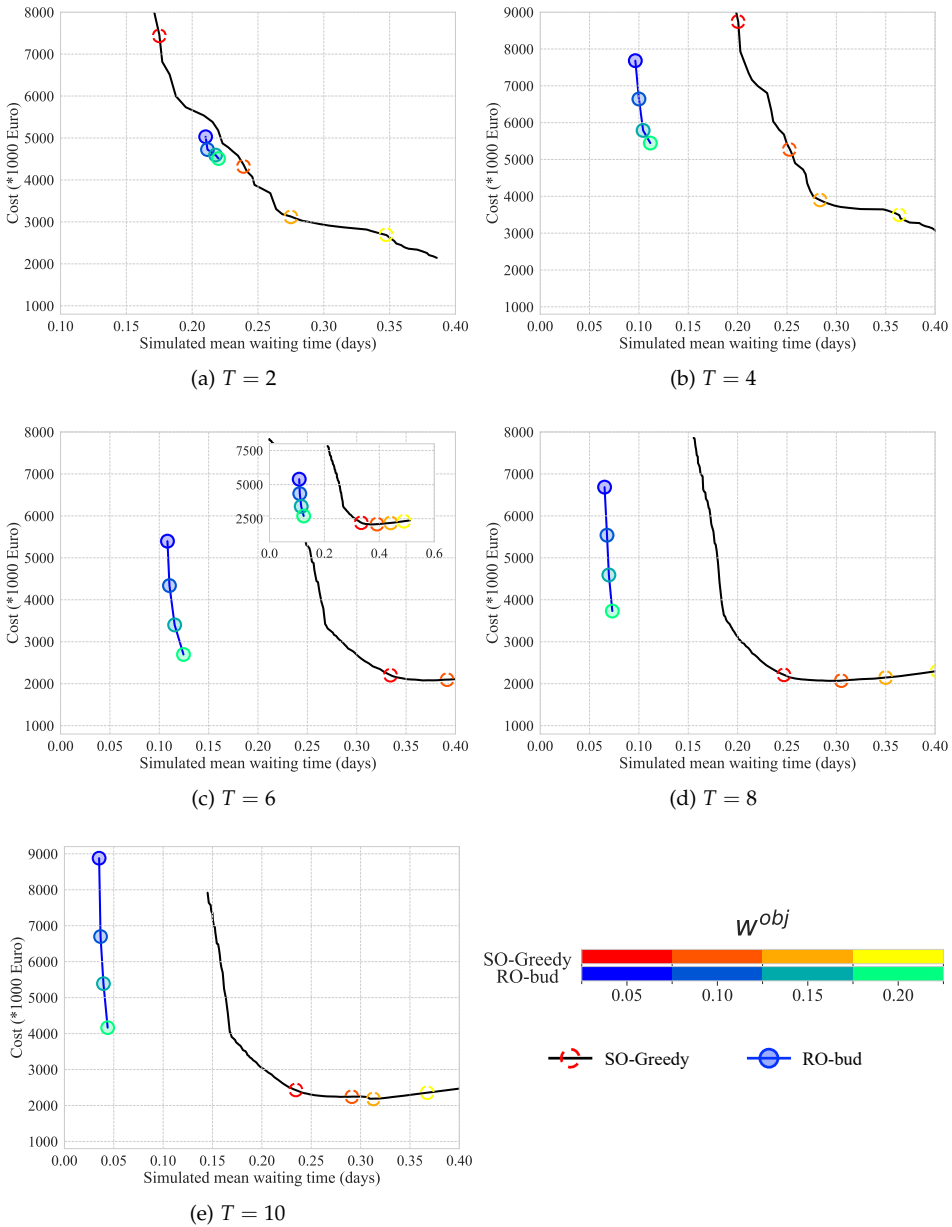


Figure 3.2: The trade-off between the simulated mean waiting time and the simulated total cost of solutions using RO-bud and SO-Greedy. Note: For $T \in \{2, 4\}$, most SKUs are in Categories Full-IFR and MIX. When $T \geq 6$, we only need to consider SKUs in Category Full-HIS.

Our sensitivity analysis examines the effect of changes in these elements on the simulated total cost and the simulated mean waiting time. In addition, Appendix 3.A.7 explores how changes in these elements can affect stock levels. We select two representative periods: $T = 2$, characterized by a predominant reliance on the IFR to obtain solutions, and $T = 10$, characterized by a predominant reliance on the HIS.

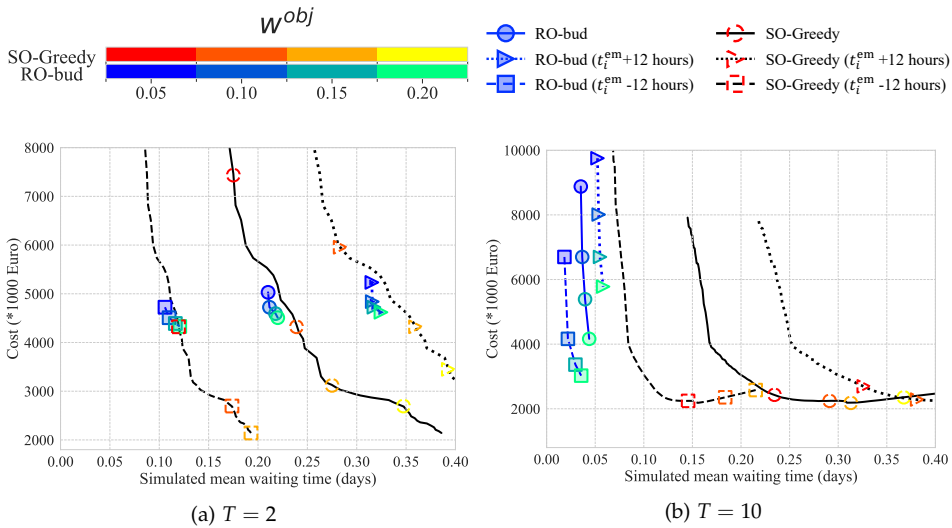


Figure 3.3: Sensitivity analysis of t_i^{em} for RO and SO solutions.

Sensitivity analysis of t_i^{em} : We conduct sensitivity analysis on t_i^{em} by changing its value by ± 12 hours from the baseline of 24 hours. Figure 3.3 shows how variations in t_i^{em} affect the simulated mean waiting time and simulated total cost of solutions. The RO solution consistently outperforms the SO solution as t_i^{em} varies. The RO model exhibits adaptability by adjusting the inventory levels of expensive SKUs following the prespecified Constraints (3.10c). Due to this, the stock levels of only a small proportion of SKUs are impacted. See Appendix 3.A.7 for more information.

Sensitivity analysis of c_i^{em} : In our sensitivity analysis of c_i^{em} , we vary it by ± 250 Euros from its baseline value of 750 Euros. Figure 3.4 shows that the RO solution is more sensitive to changes in c_i^{em} when using IFR for $T = 2$ and becomes more robust as more historical demand data is available for $T = 10$.

The sensitivity of the two models to c_i^{em} is related to their solution strategies. For the

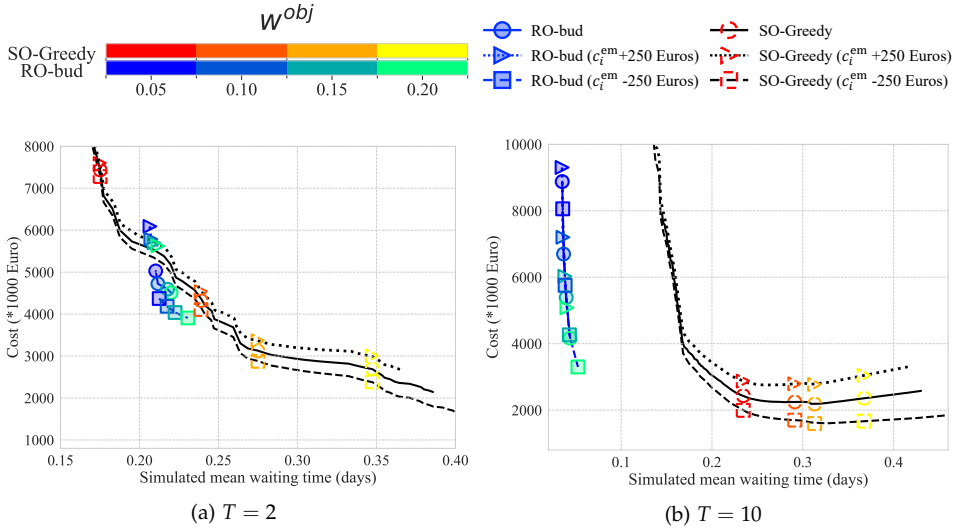


Figure 3.4: Sensitivity analysis of c_i^{em} for RO and SO solutions.

SO solution, c_i^{em} affects the starting point of the iteration where the simulated total cost for each SKU is minimized. However, as the number of iterations increases, the impact of changing c_i^{em} on achieving W^{obj} decreases, resulting in a similar simulated mean waiting time. The investigation in Appendix 3.A.7 confirms this finding. For the RO solutions, c_i^{em} affects the classification of SKUs into expensive and cheap categories. An increase in c_i^{em} leads to a higher number of SKUs labeled as cheap SKUs (\mathcal{I}_1), resulting in increased inventory holding costs but only a slight decrease in simulated mean waiting times.

Sensitivity analysis of \underline{q} , \tilde{q} , and \bar{q} : As shown in Figure 3.1, we classify SKUs based on the following thresholds: $\underline{q} = 1$, $\tilde{q} = 3$, and $\bar{q} = 6$. We perform sensitivity analysis by adjusting the thresholds by ± 2 units. Increasing the thresholds to $\underline{q} = 3$, $\tilde{q} = 5$, and $\bar{q} = 8$ results in the use of the IFR for a greater number of SKUs. Conversely, lowering the thresholds to $\underline{q} = 0$, $\tilde{q} = 1$, and $\bar{q} = 4$ leads to a preference for HIS.

Figure 3.5 shows how adjustments to the threshold can affect the solution performance. The RO and SO solutions are more cost-efficient when the threshold decreases by 2, i.e., as more SKUs use HIS-based demand estimation. This finding supports our conclusion in Appendix 3.A.6 that implementing the IFR can reduce

simulated mean waiting times, but it also leads to unnecessary holding costs due to the overestimation of the IFR for most low-demand SKUs.

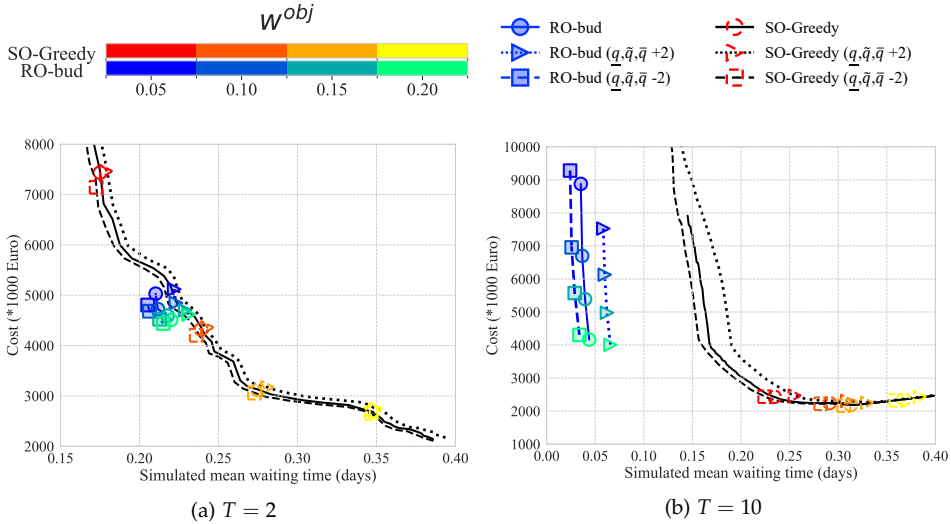


Figure 3.5: Sensitivity analysis of \underline{q} , \tilde{q} , and \bar{q} for RO and SO solutions.

Overall, the RO solution outperforms the SO solution in most cases, highlighting the adaptability of the robust model in spare parts inventory management. Besides, incorporating the IFR provides benefits in the early stages of the product life cycle in achieving a lower simulated mean waiting time, but may lead to higher holding costs. Therefore, in future research, it would be valuable to explore ways to incorporate the IFR more effectively for both RO and SO models.

3.7. Conclusion

In this chapter, we develop an adaptive robust optimization (ARO) model for managing spare parts inventory, which is especially useful in the early stages of a product's life cycle. This model takes into account emergency shipments. To obtain the exact solutions of the ARO model, we establish its equivalence to a deterministic counterpart. However, as the number of SKUs grows, the constraints in the deterministic counterpart increase exponentially, which can result in computational challenges. To overcome this, we prove that the deterministic counterpart can be decomposed into two mixed integer optimization problems. Based on this

decomposition, we develop an efficient algorithm to deal with integer uncertainty sets and to obtain near-optimal solutions for thousands of SKUs.

We conduct a case study at ASML, a leading supplier in the semiconductor industry, highlighting the practical relevance and effectiveness of our ARO model for inventory management in the semiconductor industry and other industries that sell or use capital equipment. While effective under Poisson demand, the SO model struggles with the demand variability and uncertainty inherent in new product introductions. The ARO model adapts better to diverse demand patterns without distributional assumptions, although it exhibits less flexibility in the sensitivity analysis, as its performance remains relatively stable across parameter changes.

3.A. Appendix

This chapter contains seven appendices. Appendix 3.A.1 presents the proof of Theorem 3.1. Appendix 3.A.2 provides an illustrative example comparing the optimal solutions derived from the SO-Greedy algorithm and the RO model. Appendix 3.A.3 gives proof of the optimality using the ISP algorithm for Problem (3.10) under some conditions. Appendix 3.A.4 presents the proof of Theorem 3.3. Appendix 3.A.5 through 3.A.7 are related to the ASML case study in Section 3.6.

3.A.1 Proof of Theorem 3.1

Proof. Let $k = 1$, then eliminating $\epsilon_n(\zeta)$ using Fourier–Motzkin elimination (FME) results in

$$\epsilon_i(\zeta) \geq \zeta_i - \frac{t\eta_i}{c_i^{\text{em}}}, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \quad (3.14a)$$

$$S_n \geq \zeta_n - \frac{t\eta_n}{c_n^{\text{em}}}, \quad \forall \zeta \in \mathcal{D}, \quad (3.14b)$$

$$\epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \quad (3.14c)$$

$$(S_n - \zeta_n)t_n^{\text{em}} \geq -W^{\text{obj}} \sum_{i=1}^n \zeta_i + \sum_{i=1}^{n-1} (\zeta_i - \epsilon_i(\zeta))t_i^{\text{em}}, \quad \forall \zeta \in \mathcal{D}, \quad (3.14d)$$

$$0 \geq -W^{\text{obj}} \sum_{i=1}^n \zeta_i + \sum_{i=1}^{n-1} (\zeta_i - \epsilon_i(\zeta))t_i^{\text{em}}, \quad \forall \zeta \in \mathcal{D}, \quad (3.14e)$$

$$0 \leq \epsilon_i(\zeta) \leq \zeta_i, \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \quad (3.14f)$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.14g)$$

This proves that Theorem 3.1 holds if $k = 1$.

Let us assume that Theorem 3.1 holds for a given k . We show that it holds for $k + 1$, too. Eliminating $\epsilon_{n-k}(\zeta)$ from Problem (3.6) using FME results in

$$S_i \geq \zeta_i - \frac{t\eta_i}{c_i^{\text{em}}}, \quad \forall i \in \mathcal{I}^{n-k}, \zeta \in \mathcal{D},$$

$$\epsilon_i(\zeta) \geq \zeta_i - \frac{t\eta_i}{c_i^{\text{em}}}, \quad \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D},$$

$$\epsilon_i(\zeta) \leq S_i, \quad \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D},$$

$$(S_{n-k} - \zeta_{n-k})t_{n-k}^{\text{em}} \geq -W^{\text{obj}} \sum_{i=1}^n \zeta_i + \sum_{i=1}^{n-k-1} (\zeta_i - \epsilon_i(\zeta))t_i^{\text{em}} + \sum_{i \in \alpha} (\zeta_i - S_i)t_i^{\text{em}}, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}$$

$$\begin{aligned}
0 &\geq -W^{\text{obj}} \sum_{i=1}^n \zeta_i + \sum_{i=1}^{n-k-1} (\zeta_i - \epsilon_i(\zeta)) t_i^{\text{em}} + \sum_{i \in \alpha} (\zeta_i - S_i) t_i^{\text{em}}, & \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1} \\
\zeta_i &\geq \epsilon_i(\zeta) \geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D}, \\
\eta_i &\geq 0, & \forall i \in \mathcal{I}.
\end{aligned} \tag{3.15}$$

Rearranging (3.15) concludes the proof. \square

3.A.2 Illustrative Example of Problem-solving Strategy

We present an illustrative example involving nine SKUs to compare the optimal solutions derived from the stochastic optimization model considering the Greedy algorithm (SO-Greedy) and the robust optimization model considering the price-based budget uncertainty set (RO-bud). Table 3.2 shows the values of all input parameters: predicted demand rate \bar{m}_i , price c_i^a , unit emergency shipment cost c_i^{em} , unit emergency shipment time t_i^{em} , and lead time t for each SKU, $i = 1, \dots, 9$. All combinations of Low, Medium, and High demand rates and Low, Medium, and High prices form the nine SKUs. These parameter settings are motivated by the real-life value of spare parts at ASML.

Table 3.2: Inputs for the illustrative example.

Parameter	Low	Medium	High
Mean demand \bar{m}_i (per year)	1	5	15
Acquisition cost c_i^a ($\times 1,000$ €)	1	3	30
Emergency shipment cost c_i^{em} ($\times 1,000$ €)		0.75	
Emergency shipment time t_i^{em} (days)		2	
Regular replenishment time t (days)		60	

In this example, SKUs priced at low and medium levels belong to \mathcal{I}_1 , while those with high prices are categorized under \mathcal{I}_2 . In the literature on spare part inventory control, it is typically assumed that demand follows a Poisson process. Therefore, we generate such demand data based on the predicted demand rate \bar{m}_i . We construct the price-based budget uncertainty set using the 95% bootstrap confidence level (Wood, 2005) of the generated demand.

Figure 3.6 provides a comparative analysis of the optimal solutions derived from the SO-Greedy algorithm and the optimal solutions of the robust model considering the RO-bud by solving it exactly. The SO-Greedy solution tends to stock a greater

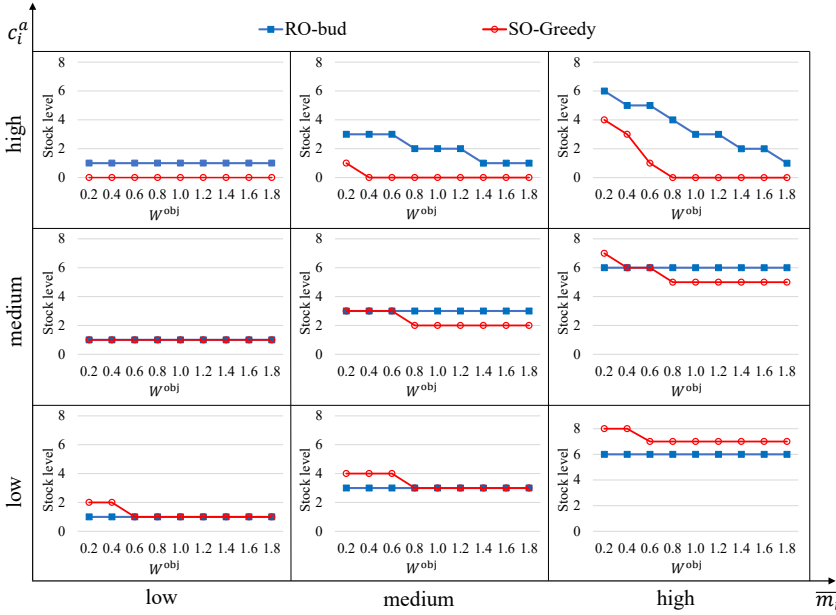


Figure 3.6: Solutions to RO-bud and SO-Greedy for the illustrative example.

quantity of low-priced spare parts for SKUs sharing identical \bar{m}_i values. As W^{obj} increases, the high-price spare parts are prioritized to reach a lower stock level. In contrast, the robust solution remains unaffected by changes in W^{obj} for medium-priced and low-priced SKUs. This observation confirms Theorem 3.2 that for SKUs within \mathcal{I}_1 , stock levels equals $\bar{\zeta}_i$, hence not affected by W^{obj} . Even though the stock level of the RO model is $\bar{\zeta}_i$, for these SKUS, the stock level using the SO model remains higher when W^{obj} is relatively small. The robust solution generally maintains higher stock levels for high-priced spare parts than the solution obtained by SO-Greedy.

3.A.3 Proof of Optimality for Problem (3.11)

We show that ISP provides an optimal solution to Problem (3.11) under some conditions. We first write the constraints of Problem (3.11) in the matrix form as follows:

$$AS \geq b, \quad S \geq 0, \tag{3.16}$$

where $S \in \mathbb{N}_0^m$ is the vector of decision variables, $A \in \mathbb{R}^{2^m-1 \times m}$ is the constraint coefficient matrix, and $b \in \mathbb{R}^{2^m-1}$ is the right-hand side parameter vector. We can

express A and b as

$$a_{\alpha,j} = \begin{cases} 0 & \text{if } j \in \mathcal{I} \setminus \alpha \\ t^{\text{em}} & \text{if } j \in \alpha \end{cases}, \quad b_{\alpha} = -W^{\text{obj}} \sum_{i \in \mathcal{I}_2} \zeta_i + \sum_{i \in \alpha} \zeta_i t_i^{\text{em}}, \quad (3.17)$$

where $a_{\alpha,j}$ represents the component in the α^{th} row and the j^{th} column of the matrix A , b_{α} is the component in the α^{th} row of the vector b for any non-empty set $\alpha \subseteq \mathcal{I}_2$.

For any non-empty set $\alpha \subseteq \mathcal{I}_2$, let Π^{α} be a partition of α . In other words, there exists a $p \in \mathbb{N}$ such that $\Pi^{\alpha} = \{\Pi_1^{\alpha}, \dots, \Pi_p^{\alpha}\}$ and

$$\Pi_1^{\alpha} \cup \Pi_2^{\alpha} \cup \Pi_3^{\alpha} \cup \dots \cup \Pi_p^{\alpha} = \alpha, \quad \Pi_i^{\alpha} \cap \Pi_j^{\alpha} = \emptyset, \quad \forall i \neq j. \quad (3.18)$$

Theorem 3.4 Let $b_{\alpha} \in \mathbb{N}_0$, for any $\alpha \subseteq \mathcal{I}_2, \alpha \neq \emptyset$. The solution generated by IS is optimal when for any non-empty subset $\alpha \subseteq \mathcal{I}_2$ and any partition Π^{α} , we have $b_{\alpha} \geq \sum_{B \in \Pi^{\alpha}} b_B$, where b_B , defined in Equation (3.17), is the right-hand side of the constraint associated with the set B .

Proof. We show that the solution generated by ISP is optimal for Problem (3.11).

We first show that if we remove the integrality restriction in Problem (3.11), leading to what we call the relaxed problem, then $S^* = (S_1^*, \dots, S_m^*)$ obtained by IS is a basic feasible solution that is optimal.

According to the assumption and procedures in ISP, S^* can be expressed as $S_k^* = \lceil (b_{\{1,2,\dots,k\}} - \sum_{i=1}^{k-1} S_i^* t_i^{\text{em}}) / t_k^{\text{em}} \rceil$, i.e. the solution of the following linear system:

$$\begin{aligned} t_1^{\text{em}} S_1^* &= b_{\{1\}}, \\ t_1^{\text{em}} S_1^* + t_2^{\text{em}} S_2^* &= b_{\{1,2\}}, \\ t_1^{\text{em}} S_1^* + t_2^{\text{em}} S_2^* + t_3^{\text{em}} S_3^* &= b_{\{1,2,3\}}, \\ &\vdots \\ \sum_{i \in \mathcal{I}} t_i^{\text{em}} S_i^* &= b_{\mathcal{I}}. \end{aligned}$$

Now, we use the Simplex algorithm (Bazaraa et al., 2011) to show that S^* is an

optimal solution to the relaxed problem. Let us set A_1 as:

$$\begin{bmatrix} t_1^{\text{em}} & 0 & 0 & \cdots & 0 \\ t_1^{\text{em}} & t_2^{\text{em}} & 0 & \cdots & 0 \\ t_1^{\text{em}} & t_2^{\text{em}} & t_3^{\text{em}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^{\text{em}} & t_2^{\text{em}} & t_3^{\text{em}} & \cdots & t_m^{\text{em}} \end{bmatrix}.$$

Note that A_1 is invertible since it is a lower triangular matrix with strictly positive diagonal entries t_i^{em} . And let us denote by A_2 the rows of A excluding A_1 . So $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$. Set $\mathcal{M} := \{\{1\}, \{1,2\}, \{1,2,3\}, \dots, \mathcal{I}_2\}$ and let $2^{\mathcal{I}_2}$ be the power set of \mathcal{I}_2 . Introducing slack variables x_α , for any $\emptyset \neq \alpha \subseteq \mathcal{I}_2$, we can rewrite Constraints (3.16) in a standard form as

$$\begin{bmatrix} A_1 & \mathbf{0}_{m,2^m-1-m} & -I^m \\ A_2 & -I^{2^m-1-m} & \mathbf{0}_{2^m-1-m,m} \end{bmatrix} \begin{bmatrix} S \\ [x_\alpha]_{\alpha \in 2^{\mathcal{I}_2} \setminus (\mathcal{M} \cup \emptyset)} \\ [x_\alpha]_{\alpha \in \mathcal{M}} \end{bmatrix} = \mathbf{b}, \quad S \geq 0, \quad x_\alpha \geq 0, \quad (3.19)$$

where I^{2^m-1-m} and I^m are identity matrices in \mathbb{R}^{2^m-1-m} and \mathbb{R}^m , respectively, $\mathbf{0}_{m,2^m-1-m}$ is the $m \times (2^m - 1 - m)$ zero matrix, and $\mathbf{0}_{2^m-1-m,m}$ is the $(2^m - 1 - m) \times m$ zero matrix.

Consider the basic feasible solution corresponding to $B = \begin{bmatrix} A_1 & \mathbf{0}_{m,2^m-1-m} \\ A_2 & -I^{2^m-1-m} \end{bmatrix}$. For the basic solution, S and $[x_\alpha]_{\alpha \in 2^{\mathcal{I}_2} \setminus (\mathcal{M} \cup \emptyset)}$ are the basic variables. Since B is a block matrix with invertible matrices in the diagonal, B is invertible (Bernstein, 2009), and its inverse is

$$B^{-1} = \begin{bmatrix} A_1^{-1} & \mathbf{0} \\ A_2 A_1^{-1} & -I^{2^m-1-m} \end{bmatrix}$$

where A_1^{-1} is

$$\begin{bmatrix} \frac{1}{t_1^{\text{em}}} & 0 & 0 & \cdots & 0 \\ -\frac{1}{t_2^{\text{em}}} & \frac{1}{t_2^{\text{em}}} & 0 & \cdots & 0 \\ 0 & -\frac{1}{t_3^{\text{em}}} & \frac{1}{t_3^{\text{em}}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{1}{t_m^{\text{em}}} \end{bmatrix}.$$

Therefore, B is a basis. Since its corresponding solution is feasible, it is a feasible

basis. Now, we check the optimality criteria of the Simplex algorithm. To this end, set $\mathbf{c}_B^a = (c_1^h - \frac{c_1^{em}}{t_1}, c_2^h - \frac{c_2^{em}}{t_2}, \dots, c_m^h - \frac{c_m^{em}}{t_m}, 0, 0, \dots, 0)$ containing m zeros. So, we have

$$\mathbf{c}_B^a \mathbf{B}^{-1} = (\theta_1 - \theta_2, \theta_2 - \theta_3, \dots, \theta_{m-1} - \theta_m, \theta_m, 0, 0, \dots, 0).$$

Let $\mathbf{e}_i = [0, 0, 0, \dots, 1, \dots, 0]^T$ be the m -tuple with all components equal to 0, except the i th one being 1. And let z_α denote the objective function value for any $\alpha \subseteq \mathcal{M}$. Then, we have the reduced cost coefficients:

$$\begin{aligned} z_{\{1\}} - c_{x_{\{1\}}}^a &= \mathbf{c}_B^a \mathbf{B}^{-1}(-\mathbf{e}_1) - c_{x_{\{1\}}}^a = \theta_2 - \theta_1 < 0, \\ z_{\{1,2\}} - c_{x_{\{1,2\}}}^a &= \mathbf{c}_B^a \mathbf{B}^{-1}(-\mathbf{e}_2) - c_{x_{\{1,2\}}}^a = \theta_3 - \theta_2 < 0, \\ &\vdots \\ z_{\mathcal{I}_2} - c_{x_{\mathcal{I}_2}}^a &= \mathbf{c}_B^a \mathbf{B}^{-1}(-\mathbf{e}_m) - c_{x_{\mathcal{I}_2}}^a = -\theta_m < 0, \end{aligned}$$

where the nonnegativity comes from the fact that the initial investment is in descending order. It is clear that all the reduced cost coefficients of the non-basic variables are non-positive. Therefore, the solution to the relaxed problem is optimal. \square

The assumptions outlined in Theorem 3.4 are related to the absence of redundant constraints in Problem (3.11). The validity of this assumption depends heavily on the choice of the uncertainty set and the value of W^{obj} .

3.A.4 Proof of Theorem 3.3

Proof. Given Corollary 3.1, after eliminating the wait-and-see variable $\epsilon_i(\zeta)$, Problem (3.12) is equivalent to Problem (3.20):

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta) \quad (3.20a)$$

$$\text{s.t. } S_i \geq \zeta_i - \frac{t\eta_i(\zeta)}{c_i^{em}}, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (3.20b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \quad (3.20c)$$

$$\sum_{i \in \alpha} S_i t_i^{em} \geq \sum_{i \in \alpha} \zeta_i t_i^{em} - W^{obj} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \quad (3.20d)$$

$$\eta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}. \quad (3.20e)$$

Compared to Problem (3.8), Problem (3.20) has an additional constraint (3.20c) and a wait-and-see variable $\eta_i(\zeta)$. We now eliminate this variable using the FME method

to prove Theorem 3.3. Let $k = 1$. Eliminating $\eta_n(\zeta)$ using FME results in

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta) \quad (3.21a)$$

$$\text{s.t. } \eta_i(\zeta) \geq \frac{c_i^{\text{em}}}{t} (\zeta_i - S_i) \quad \forall i \in \mathcal{I} \setminus \{n\}, \zeta \in \mathcal{D}, \quad (3.21b)$$

$$\sum_{i=1}^{n-1} \eta_i(\zeta) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \quad (3.21c)$$

$$\sum_{i=1}^{n-1} \eta_i(\zeta) + \frac{c_n^{\text{em}}}{t} (\zeta_n - S_n) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \quad (3.21d)$$

$$\sum_{i \in \alpha} S_i t_i^{\text{em}} \geq \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, \quad (3.21e)$$

$$\eta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I} \setminus \{n\}. \quad (3.21f)$$

This proves that Theorem 3.3 holds if $k = 1$. Assume that Theorem 3.3 holds for a given k . We now show that it holds for $k + 1$ as well. Eliminating $\epsilon_{n-k}(\zeta)$ from Problem (3.6) using FME results in:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} (c_i^h S_i + \eta)$$

$$\text{s.t. } \eta_i(\zeta) \geq \frac{c_i^{\text{em}}}{t} (\zeta_i - S_i) \quad \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, \zeta \in \mathcal{D},$$

$$\sum_{i=1}^{n-k-1} \eta_i(\zeta) + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} \frac{c_i^{\text{em}}}{t} (\zeta_i - S_i) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1},$$

$$\sum_{i=1}^{n-k-1} \eta_i(\zeta) + \frac{c_{n-k}^{\text{em}}}{t} (\zeta_{n-k} - S_{n-k}) + \quad (3.22)$$

$$\sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} \frac{c_i^{\text{em}}}{t} (\zeta_i - S_i) \leq \eta, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1},$$

$$\sum_{i \in \alpha} S_i t_i^{\text{em}} \geq \sum_{i \in \alpha} \zeta_i t_i^{\text{em}} - W^{\text{obj}} \sum_{i=1}^n \zeta_i, \quad \forall \zeta \in \mathcal{D}, \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset,$$

$$\eta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}.$$

Rearranging (3.22) and setting $k = n$ concludes the proof. \square

3.A.5 Historical Demand Data-based Uncertainty Set Construction

The following section details how to estimate \underline{HIS} and \overline{HIS} , which are the upper and lower bounds of the demand quantities per lead time.

In the ASML case study, we focus on a two-month lead time and use monthly demand data to construct the uncertainty set. We prefer to use monthly demand data because of the limited data availability at the beginning of the analysis. For example, when making stock decisions for the third and fourth quarters, we can only access two quarterly demand data points or six monthly demand points per SKU.

Another essential factor to consider is the continuous increase in demand for spare parts at ASML, which directly corresponds to the increase in machine sales. Therefore, based on the monthly demand data, we first calculate the maximum and minimum demand quantity per machine per time unit. We then multiply these by the expected number of machines and the predicted period length to get \underline{HIS} and \overline{HIS} .

3

3.A.6 Benefits of Incorporating the IFR

We evaluate the potential benefits of incorporating the IFR using the ASML dataset. Most SKUs rely on the IFR for $T = \{2, 4\}$. For $T > 4$, the proportion of SKUs using the IFR steadily declines. Because historical demand for these SKUs continues to be almost non-existent as the machine ages, inventory planners usually assume that their future demand is unlikely and set the IFR to zero.

We first examine the potential consequences of varying \underline{IFR} and \overline{IFR} on RO solutions. As shown in Figure 3.7, we compare various RO solutions with different \underline{IFR} and \overline{IFR} for the uncertainty set scaled by different multiples of the IFR values. Our results reveal that decreasing or increasing the multiplier of IFR values has a limited effect on the simulated mean waiting time and simulated total costs.

We then analyze three scenarios to assess whether the incorporation of IFR improves solution performance: a combination of IFR and HIS, exclusive use of HIS, and exclusive use of IFR. Figure 3.8(a) shows that for $T = 2$, when the simulated mean wait time exceeds 0.26 days, using only HIS outperforms incorporating IFR. However, when we aim to achieve a simulated average waiting time of less than 0.26 days, incorporating the IFR outperforms the other two scenarios. Since ASML typically aims for high service performance, targeting $W^{\text{obj}} \leq 0.1$ days when t_i^{em} is 1 day. Incorporating the IFR is beneficial to achieve a smaller W^{obj} at the very beginning of the product life cycle, aligning with ASML's service performance goals.

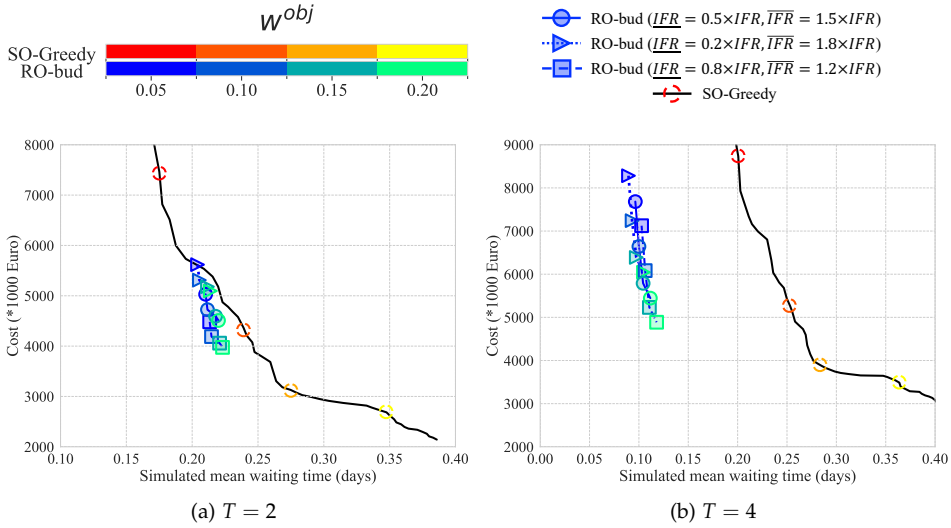


Figure 3.7: Sensitivity analysis of IFR and \overline{IFR} for RO and SO solutions.

The dynamic shifts for $T = 10$, as shown in figure 3.8(b). During this period, historical demand data accumulates, and using only HIS for all SKUs shows a marginal overperformance compared to the exclusive use of the HIS. This shift in performance is due to two primary factors. First, more historical demand data improves the accuracy of future demand forecasts. Second, at $T = 10$, inventory planners often set IFR values to zero for low-demand SKUs, even though potential future demand exists. Overall, the performance difference between these two scenarios is small.

The effectiveness of the solution incorporating the IFR depends on the accuracy of the IFR as estimated by reliability engineers. We find an overestimation of the IFR for most SKUs. While this overestimation may result in higher inventory holding costs, it is consistent with ASML's goal of achieving robust inventory performance for unexpected failures during the NPI stage.

Overall, incorporating the IFR can improve the performance of both RO and SO solutions in the early stages of a product's life cycle, especially when making stock decisions for SKUs that don't have any historical demand data but are expected to have demand in the future. Since IFR is usually overestimated, an appropriate underestimation of the worst-case scenario when constructing the uncertainty set

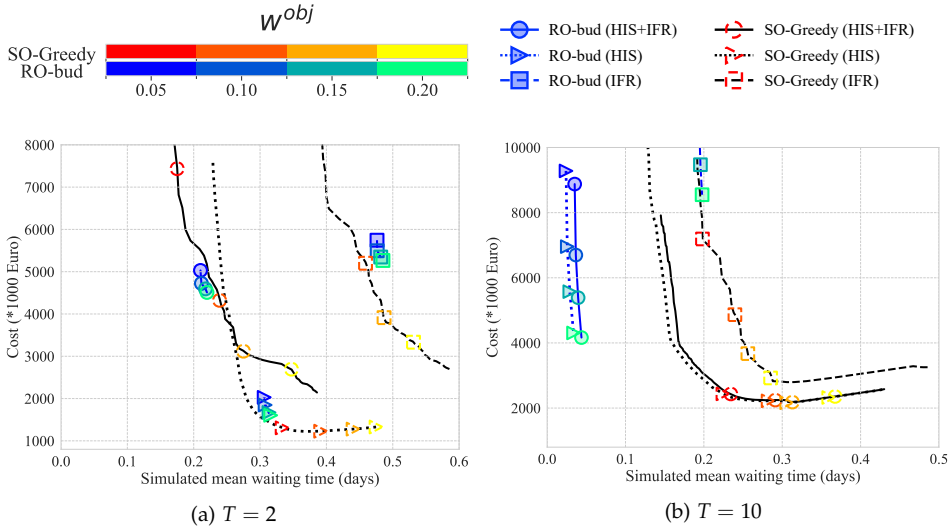


Figure 3.8: Comparison of RO and SO solutions under three scenarios: combined IFR and HIS, HIS-only, and IFR-only.

can reduce the conservatism of the RO solution.

3.A.7 Sensitivity Analysis of t_i^{em} and c_i^{em} on Stock Levels

In this section, we investigate how variations in t_i^{em} and c_i^{em} affect stock levels. We use the mean percentage error (MPE) of the stock levels to quantify the effects of these variations. The MPE is calculated by comparing the adjusted stock levels for SKU i (denoted by \tilde{S}_i) when either t_i^{em} or c_i^{em} is varied, to the original stock level for the same SKU (denoted by S_i). The formula for MPE is:

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{\tilde{S}_i - S_i}{S_i},$$

where n is the number of SKUs in consideration.

We show the MPE values derived from the sensitivity analysis of t_i^{em} and c_i^{em} in Table 3.3. Regarding the sensitivity analysis for t_i^{em} , the changes in the MPE values of the SO solution are more pronounced than those of the RO model. This indicates that the stock level of the SO solution is more sensitive to changes in t_i^{em} than the RO solution.

For the sensitivity analysis of c_i^{em} , the MPE value of the SO solution is zero, meaning that the stock level of the SO model is insensitive to changes in c_i^{em} in this case. For

Table 3.3: MPE for the sensitivity analysis of t_i^{em} and $c_i^{\text{em}}(\%)$.

W^{obj}		RO-bud				SO-Greedy			
		0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2
$t_i^{\text{em}} : +12$	T=2	0.02	0.10	0.24	0.16	3.99	9.17	10.16	16.00
	T=10	0.30	0.60	0.63	0.89	2.36	10.75	7.35	10.20
$t_i^{\text{em}} : -12$	T=2	-0.10	-0.37	-0.36	-0.35	-11.02	-19.20	-24.73	-15.45
	T=10	-0.87	-1.42	-2.04	-1.89	-14.87	-14.35	-16.93	-9.66
$c_i^{\text{em}} : +250$	T=2	2.92	2.98	3.13	3.17	0.00	0.00	0.00	0.00
	T=10	0.24	0.52	0.70	1.41	0.00	0.00	0.00	0.00
$c_i^{\text{em}} : -250$	T=2	-1.13	-1.29	-1.31	-1.43	0.00	0.00	0.00	0.00
	T=10	-0.30	-0.71	-0.90	-1.54	0.00	0.00	0.00	0.00

the RO solution, although the stock level changes due to c_i^{em} , the MPE values are small and do not exceed 3.17%. This indicates that the RO solution is stable when c_i^{em} changes.

Chapter 4

Robust spare parts inventory control with backorders

In this chapter, we propose an ARO approach for spare parts inventory control at a central warehouse. The central warehouse is sourced directly from suppliers and serves as the emergency shipment source for local warehouses, which means that stock shortages at the central warehouse result in backorders rather than lost sales.

To the best of our knowledge, we are the first to use robust optimization to formulate a continuous review inventory model with backorders. To efficiently solve our model, we develop a three-step approach. First, we establish bounds on optimal stock levels by calculating an approximate lower bound from a lost sales problem and an upper bound from a conservative estimation. For SKUs where these bounds coincide, we directly determine their near-optimal stock levels. For components where the bounds differ, our second step introduces a tighter upper bound through a relaxation of the original problem. Third, for the remaining SKUs, we employ existing approximation methods to determine near-optimal stock levels. Unlike Chapters 2 and 3, where we assume identical repair lead times for all SKUs, in this chapter, we consider the more general case where different SKUs have distinct repair lead times. This extension presents a key challenge in implementing robust optimization for spare parts inventory control, and thus, we introduce a lead time shift method to handle this complexity when constructing uncertainty sets.

To demonstrate the applicability of our model, we conduct a case study involving 1,597 SKUs at ASML. Our results show that the robust model achieves target service levels more cost-effectively than the SO model, particularly when service level requirements are stringent.

4.1. Introduction

The availability of spare parts at a central warehouse plays a critical role in maintaining high equipment uptime across a company's service network. Unlike local warehouses, where stock shortages can often be mitigated by emergency shipments, central warehouse stockouts result in backorders.

The robust optimization models developed in Chapters 2 and 3 demonstrate that appropriate uncertainty sets can effectively capture the relationships between spare parts demands. In those chapters, we introduce a (price-based) budget uncertainty set that accounts for demand interaction by constraining the joint occurrence of demands for multiple types of spare parts within specific time frames. However, both chapters assume identical lead times across spare parts for computational simplicity. In practice, spare parts lead times can vary dramatically, from days to months, which introduces considerable complexity in constructing uncertainty sets that accurately reflect demand patterns.

This chapter presents an ARO approach to the central warehouse inventory control problem. We discuss how to apply robust optimization to a continuous review inventory policy with backorders. We then propose a three-step approach to make our ARO model computationally tractable for large-scale industrial applications. To handle different lead times when constructing an uncertainty set, we develop a lead time shift method. Through a case study at ASML, we demonstrate that our model achieves higher service levels and is more cost-effective than the SO model, particularly when historical demand data are limited or service requirements are stringent.

The remainder of this chapter is organized as follows. In Section 4.2, we formulate both the stochastic optimization model and our proposed robust optimization model for central warehouse inventory control. Section 4.3 presents our solution method. Section 4.4 introduces our lead time shift method for constructing uncertainty sets. In Section 4.5, we validate our approach through a case study at ASML, comparing the performance of our robust model against their current stochastic optimization approach. Finally, Section 4.6 concludes with key findings.

4.2. Problem Formulation

We consider a central warehouse stocking a set of spare parts, denoted by $i \in \mathcal{I} = \{1, \dots, n\}$. These spare parts are used to service an installed base of machines of one type and can be used for multiple types of critical components in case of failure. We refer to these components as stock keeping units (SKUs). The inventory management system operates under a continuous review base-stock policy. Whenever demand for SKU i occurs, a new unit is ordered to bring the inventory position back up to the base-stock level S_i , and it arrives after a deterministic lead time $t_i (\geq 0)$.

When the central warehouse runs out of stock, no emergency supply option exists, and thus, the long lead times for new spare parts cause backorders. These backorders can lead to substantial disruptions in manufacturing operations and equipment uptime at all local warehouses that depend on the central warehouse. To measure the warehouse's service performance, we define the fill rate $\beta_i (\geq 0)$ for each SKU as the fraction of demand that can be fulfilled immediately from available stock. We establish an aggregate fill rate target β^{obj} (typically exceeding 90% or even 95%) across all SKUs to ensure overall service quality, representing the minimum required fraction of total demand that must be satisfied directly from stock.

Currently, ASML, like many other organizations, employs spare parts inventory control using a stochastic optimization model, as detailed in Section 4.2.1. We propose a robust optimization approach in Section 4.2.2. While the stochastic optimization model discussed by Van Houtum and Kranenburg (2015) focuses solely on minimizing holding costs for a given service level target, previous robust optimization models for periodic review inventory control (Bertsimas and Thiele, 2006; Ardestani-Jaafari and Delage, 2016; Chen et al., 2023) include both holding and backorder costs in their objective function. We follow the robust optimization literature and include backorder costs in our problem formulation.

4.2.1 Stochastic Optimization Model

The stochastic optimization (SO) model is a state-of-the-art model for spare parts inventory control at a central warehouse, allowing for the direct determination of stock levels for each SKU. The SO model assumes that the demand during the lead time follows a Poisson process with a constant rate of $m_i (> 0)$ per time unit for SKU $i \in \mathcal{I}$ (Van Houtum and Kranenburg, 2015). Under this assumption, the fill

rate for SKU i given base stock level S_i can be calculated as:

$$\beta_i(S_i) = \sum_{x=0}^{S_i-1} \frac{(m_i t_i)^x}{x!} e^{-m_i t_i}.$$

We denote the total demand rate for all SKUs by $M = \sum_{i \in \mathcal{I}} m_i$. For any SKU i in \mathcal{I} , let $c_i^h (> 0)$ denote the holding cost per time unit and $c_i^{\text{bo}} (> 0)$ denote the backorder cost per time unit. To find the optimal stock level, we solve Problem (4.1):

$$\min_{S \in \mathbb{N}_0} \sum_{i \in \mathcal{I}} c_i^h S_i + \sum_{i \in \mathcal{I}} c_i^{\text{bo}} EBO_i(S_i) \quad (4.1a)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} \frac{m_i}{M} \beta_i(S_i) \geq \beta^{\text{obj}}, \quad (4.1b)$$

where $EBO_i(S_i)$ represents the expected number of backorders for SKU $i \in \mathcal{I}$ given base stock level S_i , calculated as (see Section 2.3 of Van Houtum and Kranenburg (2015)):

$$EBO_i(S_i) = m_i t_i - S_i + \sum_{x=0}^{S_i} (S_i - x) P\{X_i = x\}, \quad S_i \in \mathbb{N}_0.$$

Problem (4.1) aims to minimize the total average cost per time unit, where the first term is the sum of mean holding costs per time unit, and the second is the sum of mean backorder costs per time unit. Constraint (4.1b) ensures that the aggregate fill rate remains above the target service level.

4.2.2 Robust Optimization Model

Unlike the SO model that assumes a known probability distribution for demand, the adaptive robust optimization (ARO) model considers a set of possible demand scenarios. Developing an ARO model with backorders under a continuous review policy presents unique challenges, particularly in capturing the dynamic nature of backorder tracking. After exploring several modeling approaches (detailed in Appendix 4.A.2), we propose the following model.

For each SKU i , we introduce two types of uncertain demand to keep track of backordered demand over consecutive lead time cycles. ζ_i^1 represents the demand in the previous lead time $[t - t_i, t]$ that may result in backorders that carried to the current lead time, and ζ_i^2 represents the new incoming demand in the current lead time $[t, t + t_i]$. Both demands are non-negative, and their joint vectors lie within the uncertainty set $(\zeta^1, \zeta^2) \in \mathcal{D}$, where $\zeta^1 = [\zeta_i^1]_{i \in \mathcal{I}}$ and $\zeta^2 = [\zeta_i^2]_{i \in \mathcal{I}}$. During lead time t_i , we first handle any backorders $(\zeta_i^1 - S_i)^+$ from the previous lead time using stock level S_i , then observe new demand ζ_i^2 and determine what fraction of

this new demand can be satisfied from the remaining stock. We introduce Problem (4.2), which considers the dependency of backorder costs across SKUs:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}, \eta_i: \mathbb{R}^n \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i + \eta \quad (4.2a)$$

$$\text{s.t. } \frac{1}{t_i} \zeta_i^2 (1 - \beta_i(\zeta^2)) c_i^{\text{bo}} \leq \eta_i(\zeta^2), \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.2b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) \leq \eta, \quad \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.2c)$$

$$\beta_i(\zeta^2) \zeta_i^2 \leq S_i - (\zeta_i^1 - S_i)^+, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.2d)$$

$$\sum_{i \in \mathcal{I}} \beta_i(\zeta^2) \frac{\zeta_i^2}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i}, \quad \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.2e)$$

$$0 \leq \beta_i(\zeta^2) \leq 1, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.2f)$$

$$\eta_i(\zeta^2) \geq 0, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.2g)$$

Here, η is the total amount of back ordered cost per time unit, which is introduced to move the uncertainty from the objective function to Constraint (4.2c). η_i is the amount of back ordered cost per time unit per SKU, which depends on the realization of the actual demand ζ^2 . Constraint (4.2d) ensures that the stock level is sufficient to fulfill part of the demand ζ^2 and backorders from the previous lead time $(\zeta_i^1 - S_i)^+$. Constraint (4.2e) ensures that the aggregate fill rate across all SKUs meets a target service level β^{obj} . Unlike Chapters 2 and 3, where we assume identical repair lead times for all SKUs (i.e., $t_i = t$ for all $i \in \mathcal{I}$), in this chapter, we extend our analysis to the more realistic case where each SKU may have a distinct repair lead time t_i . This extension necessitates the modification seen in Constraint (4.2e), where each term in the aggregate fill rate calculation is normalized by dividing by the corresponding lead time t_i . This normalization ensures that SKUs with different lead times are appropriately weighted in the service-level constraint, preventing SKUs with longer lead times from dominating the fill rate calculation. Compared with the SO Problem, Constraint (4.2f) ensures that β_i is a proper fraction between 0 and 1, and Constraint (4.2g) guarantees that the backorder amount is non-negative.

Problem (4.2) enables dynamic backorder tracking across lead times by coupling ζ^1 and ζ^2 in a joint uncertainty set. However, this formulation may lead to overestimation of backorders in certain scenarios. This conservatism stems primarily from

simplified assumptions about replenishment timing. Specifically, Constraint (4.2d) assumes that replenishment occurs only at lead time endpoints, while in practice, deliveries often arrive throughout the lead time. We illustrate this limitation in the example below. Despite the limitation, Problem (4.2) establishes a foundation for future research in this domain.

Example 4.1 Consider a single SKU with base stock level $S = 2$ and lead time $t = 10$ days. We observe demand in two consecutive lead times: $[0, 10]$ and $[10, 20]$.

Actual Continuous Review: The scenario begins identically. We start with 2 units on hand, receive 3 demands on day 1, fulfill 2 demands immediately, and backorder 1 unit. However, in a continuous review system, we immediately place a replenishment order for 3 units when the demands arrive on day 1. These replenishments arrive on day 11, allowing us to clear the backorder and restore our stock on hand to 2 units. When 3 new demands arrive on day 12, we can fulfill 2 demands immediately, leaving only 1 unit backordered.

Model's Perspective: We start with a base stock level and stock on hand of 2 units. When 3 demands arrive on day 1, we can fulfill 2 demands immediately, while 1 unit becomes backordered. At the beginning of the subsequent lead time $[10, 20]$, the scheduled replenishment arrives after a fixed 10-day cycle, and the stock level returns to the base stock level. The backordered demand is satisfied, so the stock on hand becomes $2 - 1 = 1$. When 3 more demands arrive on day 12, we can only fulfill 1 demand with our remaining stock, resulting in 2 more units being backordered.

This example illustrates how our model approximates backorders by not capturing the exact timing of demand fulfillment.

4.3. Solution Method

The SO Problem (4.1) can be solved using an extended version of the greedy algorithm (Algorithm 2.3) as discussed in Van Houtum and Kranenburg (2015). Our extension incorporates backorder costs in the objective function. The algorithm works by iteratively increasing the stock level of the SKU to provide the greatest marginal benefit, where the marginal benefit is defined as the ratio of the increase in fill rate and the increase in total cost per time unit. In this section, we show how to solve the ARO Problem (4.2). First, we reformulate it into a deterministic counterpart, which is a mixed integer linear optimization problem (MILP) with an exponential number

of constraints. Then, to deal with the computational complexity of the problem, we introduce a three-step approach to approximate the solution.

To solve Problem (4.2), we first show how we can reduce the number of wait-and-see variables in Theorem 4.1.

Theorem 4.1 *Problem (4.2) is equivalent to Problem (4.3):*

$$\min_{\substack{\mathbf{S} \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i + \eta \quad (4.3a)$$

$$\text{s.t.} \quad \sum_{i \in \alpha} \frac{c_i^{bo}}{t_i} (\zeta_i^2 - S_i + (\zeta_i^1 - S_i)^+) \leq \eta, \quad \forall \alpha \subseteq \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.3b)$$

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{(S_i - (\zeta_i^1 - S_i)^+)}{t_i} \geq \beta^{obj} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.3c)$$

$$S_i \geq (\zeta_i^1 - S_i)^+, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.3d)$$

Proof. We give the proof in Appendix 4.A.1. □

Problem (4.3) poses computational challenges. Specifically, the function $(\cdot)^+$ in both constraints makes them convex in uncertain parameters ζ^1 and ζ^2 . For robust optimization problems, we prefer constraints that are concave in uncertain parameters to derive a tractable reformulation. To address this, we eliminate the $(\cdot)^+$ terms by introducing additional subset variables γ . For constraint (4.3b), we introduce $\gamma \subseteq \alpha$ to represent the components where $(\zeta_i^1 - S_i)$ is positive. Similarly, for constraint (4.3c), we introduce $\gamma \subseteq \mathcal{I} \setminus \alpha$ to represent components where $(\zeta_i^1 - S_i)$ is positive. For Constraint (4.3d), we rewrite it more explicitly. When $\zeta_i^1 > S_i$, this constraint becomes $S_i \geq \zeta_i^1 - S_i$, which simplifies to $S_i \geq \frac{1}{2}\zeta_i^1$. When $\zeta_i^1 \leq S_i$, this constraint becomes $S_i \geq 0$, which is already satisfied by $\mathbf{S} \in \mathbb{N}_0^n$. Thus, Constraint (4.3d) can be expressed as $S_i \geq \frac{\zeta_i^1}{2}$ for any $i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}$. This yields the following equivalent reformulation of Problem (4.3):

$$\min_{\substack{\mathbf{S} \in \mathbb{N}_0^n \\ \eta \in \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i + \eta \quad (4.4a)$$

$$\text{s.t.} \quad \sum_{i \in \alpha} \frac{c_i^{bo}}{t_i} (\zeta_i^2 - S_i) + \sum_{i \in \gamma} \frac{c_i^{bo}}{t_i} (\zeta_i^1 - S_i) \leq \eta, \quad \forall \alpha \subseteq \mathcal{I}, \gamma \subseteq \alpha, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.4b)$$

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{(S_i - \sum_{i \in \gamma} (\zeta_i^1 - S_i))}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, \gamma \subseteq \mathcal{I} \setminus \alpha, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.4c)$$

$$S_i \geq \frac{\zeta_i^1}{2}, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.4d)$$

Problem (4.4) still has high computational complexity due to its exponential number of constraints. We propose a three-step approach.

Step 1: First, we obtain initial bounds. We construct an approximate lower bound of the optimal solution through a modified version of Problem (2.2) with lost sales. Note that in Chapter 2, we assume constant lead times for all SKUs, whereas in this chapter, we consider different lead times for each SKU. Therefore, we adapt Problem (2.2) as:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i \quad (4.5a)$$

$$\text{s.t. } \beta_i(\zeta) \zeta_i \leq S_i, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \quad (4.5b)$$

$$\sum_{i \in \mathcal{I}} \frac{\beta_i(\zeta) \zeta_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i}{t_i}, \quad \forall \zeta \in \mathcal{D}, \quad (4.5c)$$

$$1 \geq \beta_i(\zeta) \geq 0, \quad \forall i \in \mathcal{I}, \zeta \in \mathcal{D}. \quad (4.5d)$$

The key difference from Problem (2.2) is in Constraint (4.5c), where each term is now normalized by dividing by the corresponding lead time t_i . The lost sales problem provides a good lower bound because, in spare parts inventory management, cheaper SKUs are typically stocked at higher levels regardless of whether we use a lost sales or backorder model.

We construct an upper bound through conservative estimation. For conservative estimation, we consider the worst-case demand scenario for each SKU in paired lead times, setting $S_i = \max_{(\zeta^1, \zeta^2) \in \mathcal{D}} \{\zeta_i^1, \zeta_i^2\}$. This preprocessing step can immediately identify the near-optimal stock levels for SKUs where the lower and upper bounds meet.

Step 2: Second, for SKUs where the bounds differ, we construct a relaxation of Problem (4.4) to obtain a tighter upper bound. For Constraint (4.4c), observe that

for any feasible solution \mathbf{S}^* :

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{(S_i^* - \sum_{i \in \gamma} (\zeta_i^1 - S_i^*))}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, \gamma \subseteq \mathcal{I} \setminus \alpha, (\zeta^1, \zeta^2) \in \mathcal{D},$$

which rearranges to:

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{S_i^*}{t_i} + \sum_{i \in \gamma} \frac{S_i^*}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \gamma} \frac{\zeta_i^1}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, \gamma \subseteq \mathcal{I} \setminus \alpha, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.6)$$

Dropping $\sum_{i \in \gamma} S_i^* \geq 0$ (as $S_i^* \geq 0$), a valid relaxation is:

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{S_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \gamma} \frac{\zeta_i^1}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, \gamma \subseteq \mathcal{I} \setminus \alpha, (\zeta^1, \zeta^2) \in \mathcal{D}.$$

Any solution satisfying the above constraint would satisfy Constraint (4.6). Therefore, replacing (4.6) by the above constraint shrinks the feasible region, resulting in an upper bound on Problem (4.4). To avoid enumerating all γ , note the worst case occurs when $\sum_{i \in \gamma} \frac{\zeta_i^1}{t_i}$ is maximized. Since $\gamma \subseteq \mathcal{I} \setminus \alpha$, we have $\max_{(\zeta^1, \zeta^2) \in \mathcal{D}} \sum_{i \in \gamma} \frac{\zeta_i^1}{t_i} \leq \max_{(\zeta^1, \zeta^2) \in \mathcal{D}} \sum_{i \in \mathcal{I} \setminus \alpha} \frac{\zeta_i^1}{t_i}$. We expand γ to $\mathcal{I} \setminus \alpha$ for terms with $\frac{\zeta_i^1}{t_i}$, yielding:

$$\sum_{i \in \mathcal{I} \setminus \alpha} \frac{S_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \mathcal{I} \setminus \alpha} \frac{\zeta_i^1}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.7)$$

This ensures feasibility for all $\gamma \subseteq \mathcal{I} \setminus \alpha$. So, we can calculate an upper bound on the stock levels by solving the following optimization problem.

$$\min_{\mathbf{S} \in \mathbb{N}_0^n} \sum_{i \in \mathcal{I}} c_i^h S_i \quad (4.8a)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I} \setminus \alpha} \frac{S_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \mathcal{I} \setminus \alpha} \frac{\zeta_i^1}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.8b)$$

$$S_i \geq \frac{\zeta_i^1}{2}, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.8c)$$

This relaxed Problem (4.8) has a structure similar to the lost sales problem in Chapter 2, enabling us to solve it using established algorithms such as the iterative projection in descending order (IPDO) algorithm or the constraint generation (ConGA) algorithm. For implementing ConGA, we provide an estimation of the right-hand side value of Constraint (4.8b) in Appendix 4.A.4.

Step 3: Finally, for those SKUs whose bounds are not tight yet, we employ existing approximation methods to determine near-optimal solutions. For example, we can use the affine decision rule approach (ADR), which restricts decision variables to

affine functions of uncertain parameters.

4.4. Uncertainty Set

This section introduces two classical uncertainty sets in Section 4.4.1 and describes our method to construct the uncertainty set for SKUs with different lead times in Section 4.4.2.

4.4.1 Classical Uncertainty Sets

The box uncertainty set, as discussed in Sections 2.2.3 and 3.5.1, provides individual bounds for each SKU's demand in each lead time and serves as our starting point:

$$\mathcal{D}^{\text{box}} = \left\{ (\zeta^1, \zeta^2) \in \mathbb{R}^n \times \mathbb{R}^n : \begin{array}{l} d_i^1 \leq \zeta_i^1 \leq \bar{d}_i^1, \quad \forall i \in \mathcal{I} \\ d_i^2 \leq \zeta_i^2 \leq \bar{d}_i^2, \quad \forall i \in \mathcal{I} \end{array} \right\}.$$

Here, d_i^1 and \bar{d}_i^1 represent the lower and upper bounds for demand in the previous lead time, and d_i^2 and \bar{d}_i^2 represent those for the current lead time. Note that this uncertainty set assumes the independence between demands in two consecutive lead times, which can be overly conservative in practice. Therefore, we introduce a budget uncertainty set.

$$\mathcal{D}^{\text{bud}} = \left\{ (\zeta^1, \zeta^2) \in \mathbb{R}^n \times \mathbb{R}^n : \begin{array}{l} d_i^{12} \leq \zeta_i^1 + \zeta_i^2 \leq \bar{d}_i^{12}, \quad \forall i \in \mathcal{I}, \\ \underline{\Gamma}_\alpha^1 \leq \sum_{i \in \alpha} \frac{\zeta_i^1}{t_i} \leq \bar{\Gamma}_\alpha^1, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset \\ \underline{\Gamma}_\alpha^2 \leq \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} \leq \bar{\Gamma}_\alpha^2, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset \end{array} \right\}$$

The first constraint captures the aggregate demand bounds for each SKU in two consecutive lead times. For each SKU $i \in \mathcal{I}$, this constraint ensures that the sum of the demands in the previous lead time (ζ_i^1) and the current lead time (ζ_i^2) are bounded by d_i^{12} and \bar{d}_i^{12} . Subsequent constraints account for demand interdependencies between different SKUs by establishing upper and lower bounds of demand during the lead time for every subset α . Note that for any $i \in \mathcal{I}$, $\underline{\Gamma}_{\{i\}}^1 = \frac{d_i^1}{t_i}$ and $\bar{\Gamma}_{\{i\}}^1 = \frac{\bar{d}_i^1}{t_i}$, and similarly, $\underline{\Gamma}_{\{i\}}^2 = \frac{d_i^2}{t_i}$ and $\bar{\Gamma}_{\{i\}}^2 = \frac{\bar{d}_i^2}{t_i}$.

The construction of this set, specifically the estimation of d_i^1 , \bar{d}_i^1 , d_i^2 , \bar{d}_i^2 , $\underline{\Gamma}_\alpha^1$, $\bar{\Gamma}_\alpha^1$, $\underline{\Gamma}_\alpha^2$ and $\bar{\Gamma}_\alpha^2$, relies on two primary data sources: historical demand data and the initial

failure rate (IFR) estimated by engineers. The incorporation of these data types into the uncertainty set is detailed in Section 3.6.2.

4.4.2 Lead Time Shift Method

Different SKUs usually have different lead times. For notational simplicity, we omit superscripts 1 and 2 (which represent two consecutive lead times) in this section. Although estimating \underline{d}_i and \bar{d}_i from historical demand data is straightforward, determining $\underline{\Gamma}_\alpha$ and $\bar{\Gamma}_\alpha$ becomes complex when SKUs within α have different lead times. The analysis presented here extends the initial work done in the Master thesis of Pessers (2024). Figure 4.1 illustrates this complexity. The top three lines represent demand patterns for three SKUs, indicated by different colored points (orange, blue, and pink). Vertical blue lines on each SKU’s timeline denote the start and end of its lead time, which varies among SKUs. The bottom line shows their combined demand over time. The noncoinciding lead times complicate the determination of combined upper and lower demand bounds for multiple SKUs.

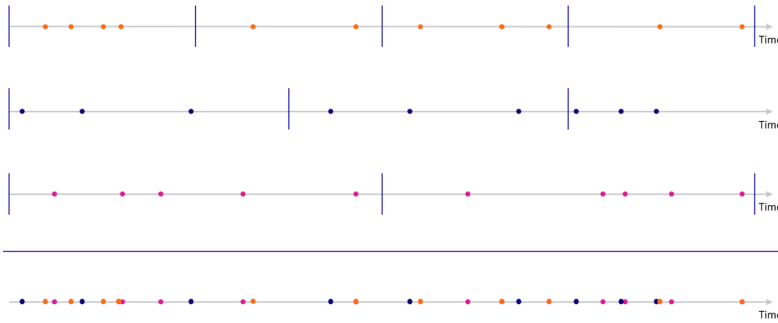


Figure 4.1: Illustration of observed demand over time for different SKUs.

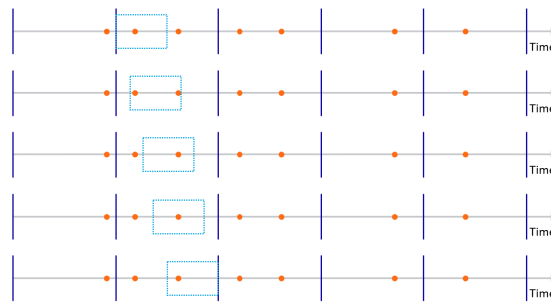
We develop the *lead time shift* method to compute aggregate bounds $\underline{\Gamma}_\alpha$ and $\bar{\Gamma}_\alpha$ for SKUs with different lead times. Let us start with an illustrative example to understand the method.

Example 4.2 Consider three years of historical demand data divided into 6-month periods ($P = 182$ days), giving us six periods $Q = \{1, 2, 3, 4, 5, 6\}$. For an SKU $i \in \mathcal{I}$ with a lead time $t_i = 100$ days, we perform the following process for each period $q \in Q$.

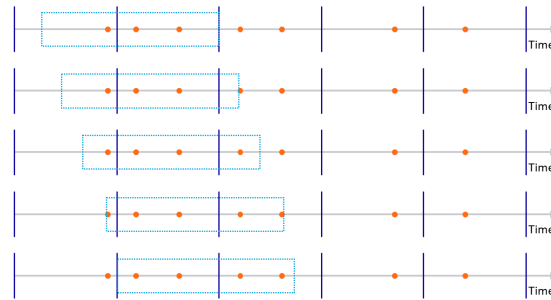
For the period $q = 1$ (first 6 months), we start with the lead time window (100 days) aligned to the beginning of the period. We count demands within this 100-day window, then shift

the window by one day and count again. This shifting and counting continue for all possible $|100 - 182| + 1 = 83$ positions within the period. All 83 counts are recorded in vector $\vec{e}_{i,1}$.

We repeat this same process for each subsequent period ($q = 2, 3, 4, 5, 6$), creating vectors $\vec{e}_{i,2}$ through $\vec{e}_{i,6}$, each with 82 elements representing different possible demand counts based on where the 100-day lead time window falls within that 6-month period. For each SKU $i \in \mathcal{I}$ and period $q \in Q$, we then calculate the lower and upper bounds: $LB_{i,q} = \min_{1 \leq l \leq 83} (e_{i,q})_l$ and $UB_{i,q} = \max_{1 \leq l \leq 83} (e_{i,q})_l$.



(a) Lead time shifts when $t_i < P$



(b) Lead time shifts when $t_i > P$

Figure 4.2: Visualization of the lead time shifts. Solid vertical lines indicate the length of the pre-specified time period (P). The blue rectangle represents the lead time window in different positions as it shifts across the timeline.

Figure 4.2 illustrates the lead time shift method in $q = 2$, with (a) showing $t_i < P$ and (b) $t_i > P$. In Figure 4.2 (a), we have $UB_{i,2} = 2$ and $LB_{i,2} = 1$. For a longer lead time with an identical demand pattern, Figure 4.2 (b) yields $UB_{i,2} = 5$ and $LB_{i,2} = 3$.

More formally, the lead time shift method consists of the following steps.

We first divide the historical timeline into pre-specified time periods of duration P . At ASML, this duration is determined by the mean lead time across all SKUs and the frequency of updates to the installed base information. Let us denote the set of these pre-specified time periods by $Q = \{1, \dots, |Q|\}$.

For each SKU $i \in \mathcal{I}$ and period $q \in Q$, we create multiple counts of demand occurrences by shifting the lead time window t_i across period q , where t_i is the lead time of SKU i . We define a vector $\vec{e}_{i,q} = [(e_{i,q})_1, (e_{i,q})_2, \dots, (e_{i,q})_{|t_i-P|}]$ that contains the $|t_i - P|$ number of elements, where each element $(e_{i,q})_l$ represents the number of demand occurrences counted when the lead time window for SKU i is shifted by l time units relative to the start of the period q . This shifting process ensures that we capture all possible ways in which a lead time period could overlap with our pre-specified time periods.

From these vectors, we compute the demand bounds for each SKU and period. The upper ($UB_{i,q}$) and lower ($LB_{i,q}$) bounds for each SKU $i \in \mathcal{I}$ are computed as:

$$UB_{i,q} = \max_{1 \leq l \leq |t_i-P|} (e_{i,q})_l, \quad \forall i \in \mathcal{I}, q \in Q,$$

$$LB_{i,q} = \min_{1 \leq l \leq |t_i-P|} (e_{i,q})_l, \quad \forall i \in \mathcal{I}, q \in Q.$$

Since these bounds are now calculated over identical time periods q , we can calculate the aggregate lower and upper bounds, Γ_α and $\bar{\Gamma}_\alpha$, respectively, across all SKUs and periods as follows:

$$\Gamma_\alpha = \min_{q \in Q} \sum_{i \in \alpha} \frac{LB_{i,q}}{t_i}, \quad \bar{\Gamma}_\alpha = \max_{q \in Q} \sum_{i \in \alpha} \frac{UB_{i,q}}{t_i}.$$

An alternative to constructing the bounds of the uncertainty set is the *unit shift method*. This method divides the historical timeline into fixed time intervals (e.g., days or weeks) and counts demand occurrences within these fixed time intervals. This approach offers computational simplicity. However, for SKUs with long lead times, the unit shift method can underestimate demand volatility because the analysis window is too short to capture the full pattern of variability throughout their entire replenishment cycle. In addition, for SKUs with short lead times, the method can overestimate risks by including demand fluctuations outside of their actual replenishment cycle, potentially leading to excessive inventory allocations. The lead time shift method avoids the pitfalls of the unit shift method by tailoring the analysis windows to each SKU's replenishment cycle.

The following example illustrates these methodological differences.

Example 4.3 (Comparison of Methods for One Period) Consider a 10-day period with two SKUs having different lead times: SKU 1 ($t_1 = 2$ days) and SKU 2 ($t_2 = 5$ days). The demand pattern is:

Day	1	2	3	4	5	6	7	8	9	10
SKU 1	1	0	0	1	0	0	0	1	0	0
SKU 2	0	1	0	0	1	0	1	0	0	1

Lead Time Shift Method: For SKU 1, we shift a 2-day window through the period:

4

Window Position	Demand Count
Days 1-2	1
Days 2-3	0
Days 3-4	1
Days 4-5	1
Days 5-6	0
Days 6-7	0
Days 7-8	1
Days 8-9	1
Days 9-10	0

This gives $LB_1 = 0$ and $UB_1 = 1$. Thus, we have $\frac{LB_1}{t_1} = 0$ and $\frac{UB_1}{t_1} = 0.5$ demands per day.

For SKU 2, we shift a 5-day window:

Window Position	Demand Count
Days 1-5	2
Days 2-6	2
Days 3-7	2
Days 4-8	2
Days 5-9	2
Days 6-10	2

This gives $LB_2 = 2$ and $UB_2 = 2$. Thus, we have $\frac{LB_2}{t_2} = 0.4$ and $\frac{UB_2}{t_2} = 0.4$ demands per day. The aggregate bounds are: $\underline{\Gamma}_{\{1,2\}} = 0 + 0.4 = 0.4$ and $\bar{\Gamma}_{\{1,2\}} = 0.5 + 0.4 = 0.9$ demands per day.

Unit Shift Method: Using 1-day units, we have $\underline{\Gamma}_{\{1,2\}} = 0$ and $\bar{\Gamma}_{\{1,2\}} = 2$ demands per day.

The lead time shift method accounts for replenishment cycles, yielding tighter bounds (0.4 to 0.9 demands per day). The unit shift method overestimates volatility, producing a conservative upper bound ($\bar{\Gamma}_{\{1,2\}} = 2$ demands per day). This mismatch occurs because fixed units fail to capture demand patterns over actual lead times, especially when SKUs have lead times with no shared time intervals (e.g., 2 and 5 days).

4.5. ASML Case study

This case study evaluates the practical implementation of the ARO Problem for optimizing spare parts inventory control at ASML. While the case studies in Chapters 2 and 3 assume identical lead times across SKUs for computational simplicity, real-world spare parts operations exhibit substantial variations in lead times. Figure 4.3 presents the distribution of the average lead times of different SKUs for a specific machine type at ASML, demonstrating a variability ranging from 14 to 793 days. To better reflect this operational reality, our case study incorporates these lead times in our analysis.

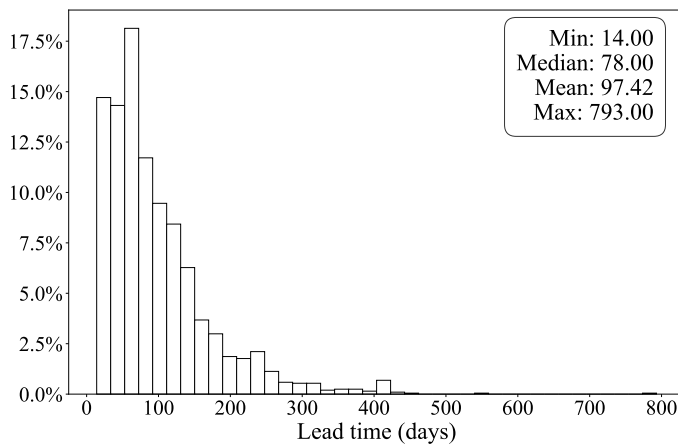


Figure 4.3: Distribution of average lead time for a specific machine type at ASML.

ASML currently employs an SO approach for its spare parts inventory management system, which first assumes Poisson demand, then estimates demand rates, and finally applies a greedy algorithm to solve Problem (4.1). Our analysis compares the performance of the proposed RO model against the SO approach used at ASML.

Our dataset comprises 1,597 SKUs for a specific machine type. To construct meaningful demand uncertainty sets, we include only SKUs with either historical demand data from the past 3 years or IFR estimates. Given ASML's commitment to maintaining high service levels, we examine multiple fill rate targets: $\beta^{\text{obj}} \in \{0.85, 0.90, 0.95, 0.99\}$. We set c_i^{bo} at 1,000 Euro per occurrence of a backorder, reflecting a simplified but reasonable estimate based on practical scenarios. In practice, c_i^{bo} varies across supply chain scenarios. When both central and local warehouses do not have stock, sourcing components directly from the factory significantly increases costs. Our 1,000 Euro estimate appropriately exceeds the 750 Euro emergency shipment cost discussed in Chapter 3, reflecting the higher expenses typically associated with backorder situations. For a detailed analysis of how c_i^{bo} impacts ARO model outputs, we refer the reader to Section 7.2.2 of Pessers (2024).

For the SO problem, we implement a greedy algorithm. For the RO model, we use our three-step approach from Section 4.3. First, we find an estimated lower bound using the hybrid approach applied to the lost sales problem (2.2). Second, we determine a tighter upper bound using the ConGA algorithm. These preprocessing steps determine near-optimal stock levels for approximately 90% of SKUs. For the remaining SKUs, we approximate stock levels using the affine decision rule approach.

For the lead time shift method, we set the time period P to 6 months based on practical considerations. Similarly to the case study in Chapter 2, we use a training-testing approach to evaluate performance. Specifically, we use the first two years of historical demand data as our training set to generate solutions and then use the third year of data as our test set to evaluate performance.

The presented results in this section are normalized for confidentiality reasons. Table 4.1 compares the performance of the RO solution with the SO solution (SO-Greedy). The RO solution exhibits more stable performance, maintaining simulated fill rates between 0.895 and 0.912 in different β^{obj} . In contrast, the SO-Greedy approach shows greater variability, with fill rates ranging from 0.646 to 0.879. It should be noted that the simulated fill rate of the RO solution plateaus at 0.912 even when

Table 4.1: Simulated fill rate and normalized annual cost for RO and SO solutions

β^{obj}	Method	Aggregate Fill Rate	Normalized annual cost
0.85	Robust	0.895	0.41
	SO-Greedy	0.646	0.14
0.90	Robust	0.903	0.44
	SO-Greedy	0.790	0.22
0.95	Robust	0.909	0.46
	SO-Greedy	0.877	0.99
0.99	Robust	0.912	0.57
	SO-Greedy	0.879	1

Note: Costs are normalized to the maximum observed value ($\beta^{\text{obj}} = 0.99$ for SO-Greedy).

$\beta^{\text{obj}} = 0.99$. This limitation is attributed to a change in demand patterns for some SKUs in the test period (third year) compared to the training data (first two years). Our analysis reveals that more than 27% of the SKUs exhibited a faster increase in demand than the trend of increase predicted on the basis of historical data and sales. This unexpected acceleration in demand growth exceeds the conservative bounds established by our worst-case scenario modeling, highlighting the challenges of prediction in highly volatile spare parts environments.

Figure 4.4 further illuminates the performance dynamics of solutions through a visualization of the trade-off between the cost and service level. The RO solution exhibits better stability, clustering at higher fill rates while maintaining relatively controlled increases in simulated total costs. Most notably, the SO-Greedy solution shows a dramatic cost escalation as the simulated fill rates approach 0.88.

In general, while the SO approach shows competitive performance at lower service levels, the RO model proves substantially more cost-efficient at higher service levels, which are typically required in critical spare parts management. Most importantly, the stability of the RO solution in simulated fill rates and controlled cost scaling makes it especially valuable for service providers requiring stringent service levels, offering a more predictable and economical approach to spare parts inventory management.

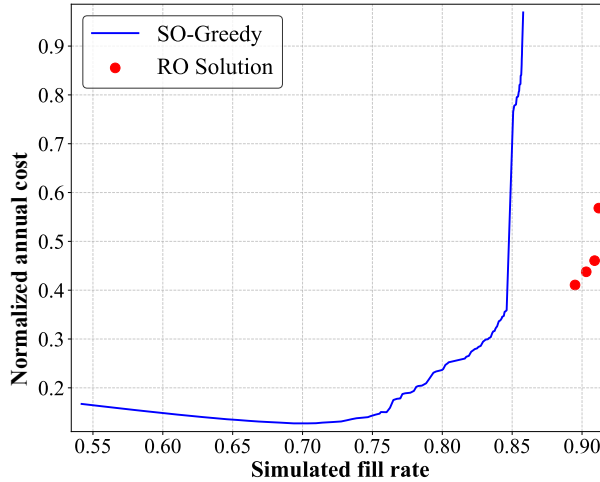


Figure 4.4: The trade-off between the simulated fill rate and the normalized annual cost of solutions using SO-Greedy and RO. Costs are normalized to the observed value ($\beta^{\text{obj}} = 0.99$ for SO-Greedy).

4

4.6. Conclusion

In this chapter, we develop an ARO model for managing spare parts inventory at the central warehouse. Unlike existing literature, which only focuses on periodic review inventory policy, we develop the first continuous review inventory model that incorporates backorders through robust optimization.

To solve the ARO problem, we first reformulate it as its deterministic counterpart and then develop a three-step solution approach. The first step establishes solution bounds by deriving an approximate lower bound through a lost sales problem (Chapter 2) and an upper bound through conservative estimation based on worst-case demand scenarios. For components where the bounds differ, our second step introduces a tighter upper bound through a relaxation of the original ARO problem. In the third step, we employ additional approximation methods to determine near-optimal solutions for the remaining components.

We introduce a lead time shift method that handles different lead times across different components when constructing uncertainty sets, addressing a practical complexity that previous chapters do not consider.

We validate our approach through a comprehensive case study at ASML with 1,597

SKUs. Our results demonstrate that the robust model consistently outperforms the SO model, particularly in practically important scenarios with stringent service requirements.

4.A. Appendix

This chapter includes four appendices. In Appendix 4.A.1, we present the proof of Theorem 4.1. Appendix 4.A.2 explores alternative robust optimization models with backorders. Appendix 4.A.3 offers a numerical comparison between Problems (4.2) and (4.14). Finally, Appendix 4.A.4 introduces an efficient algorithm for computing bounds in the budget uncertainty set.

4.A.1 Proof of Theorem 4.1

Proof. We first reformulate Problem (4.2) into a fixed-recourse ARO problem. For a given $i \in \mathcal{I}$ and $(\zeta^1, \zeta^2) \in \mathcal{D}$, we define $\epsilon_i(\zeta^2) := \beta_i(\zeta^2)\zeta_i^2$. Now, Problem (4.2) can be reformulated to:

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}, \eta_i: \mathbb{R}^n \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i + \eta \quad (4.9a)$$

$$\text{s.t. } \frac{c_i^{\text{bo}}}{t_i} (\zeta_i^2 - \epsilon_i(\zeta^2)) \leq \eta_i(\zeta^2), \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.9b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) \leq \eta, \quad \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.9c)$$

$$\epsilon_i(\zeta^2) \leq S_i - (\zeta_i^1 - S_i)^+, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.9d)$$

$$\sum_{i \in \mathcal{I}} \frac{\epsilon_i(\zeta^2)}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i}, \quad \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.9e)$$

$$0 \leq \epsilon_i(\zeta^2) \leq \zeta_i^2, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \quad (4.9f)$$

$$\eta_i(\zeta^2) \geq 0, \quad \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.9g)$$

We then show how we can reduce the number of $\epsilon_i(\zeta^2)$ of Problem (4.9) in the following lemma.

Lemma 4.1 *Given $k \in \mathcal{I}$, let $\mathcal{I}^{n-k+1} = \{n-k+1, \dots, n\}$. Problem (4.9) is equivalent to Problem (4.10):*

$$\min_{\substack{S \in \mathbb{N}_0^n \\ \epsilon_i: \mathbb{R}^{n-k} \rightarrow \mathbb{R} \\ \eta \in \mathbb{R} \\ \eta_i: \mathbb{R}^n \rightarrow \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h S_i + \eta$$

$$\begin{aligned}
s.t. \quad S_i - (\zeta_i^1 - S_i)^+ &\geq \zeta_i^2 - \frac{t_i \eta_i(\zeta^2)}{c_i^{bo}}, & \forall i \in \mathcal{I}^{n-k+1}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\zeta_i^2 - \epsilon_i(\zeta^2) &\leq \frac{t_i \eta_i(\zeta^2)}{c_i^{bo}}, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) &\leq \eta, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\epsilon_i(\zeta^2) &\leq S_i - (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
S_i &\geq (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I}^{n-k+1}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i=1}^{n-k} \frac{\epsilon_i(\zeta^2)}{t_i} + \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} &+ \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} \frac{(S_i - (\zeta_i^1 - S_i)^+)}{t_i} \geq \beta^{obj} \sum_{i=1}^n \frac{\zeta_i^2}{t_i}, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
\zeta_i^2 &\geq \epsilon_i(\zeta^2) \geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k+1}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\eta_i(\zeta^2) &\geq 0, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}.
\end{aligned} \tag{4.10}$$

Let $k = 1$, then eliminating $\epsilon_n(\zeta_n^2)$ using Fourier–Motzkin elimination (FME) results in

$$\begin{aligned}
\epsilon_i(\zeta^2) &\geq \zeta_i^2 - \frac{t_i \eta_i(\zeta^2)}{c_i^{bo}}, & \forall i \in \mathcal{I} \setminus \{n\}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
S_n - (\zeta_n^1 - S_n)^+ &\geq \zeta_n^2 - \frac{t_n \eta_n(\zeta^2)}{c_n^{bo}}, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\epsilon_i(\zeta^2) &\leq S_i - (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I} \setminus \{n\}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
S_n &\geq (\zeta_n^1 - S_n)^+, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) &\leq \eta, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i=1}^{n-1} \frac{\epsilon_i(\zeta^2)}{t_i} + \frac{S_n}{t_n} - \frac{(\zeta_n^1 - S_n)^+}{t_n} &\geq \beta^{obj} \sum_{i=1}^n \frac{\zeta_i^2}{t_i}, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i=1}^{n-1} \frac{\epsilon_i(\zeta^2)}{t_i} + \frac{\zeta_n^2}{t_n} &\geq \beta^{obj} \sum_{i=1}^n \frac{\zeta_i^2}{t_i} & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
0 \leq \frac{\epsilon_i(\zeta^2)}{t_i} &\leq \zeta_i^2, & \forall i \in \mathcal{I} \setminus \{n\}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\eta_i(\zeta^2) &\geq 0, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}.
\end{aligned}$$

This proves that Lemma 4.1 holds if $k = 1$. Let us assume that Lemma 4.1 holds for a given k . We show that it holds for $k + 1$, too. Eliminating $\epsilon_{n-k}(\zeta_{n-k}^2)$ from Problem (4.9) using Fourier-Motzkin elimination results in

$$\begin{aligned}
S_i - (\zeta_i^1 - S_i)^+ &\geq \zeta_i^2 - \frac{t_i \eta_i(\zeta^2)}{c_i^{\text{bo}}}, & \forall i \in \mathcal{I}^{n-k}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\epsilon_i(\zeta^2) &\geq \zeta_i^2 - \frac{t_i \eta_i(\zeta^2)}{c_i^{\text{bo}}}, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\epsilon_i(\zeta^2) &\leq S_i - (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
S_i &\geq (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I}^{n-k}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) &\leq \eta, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\sum_{i=1}^{n-k-1} \frac{\epsilon_i(\zeta^2)}{t_i} + \frac{S_{n-k}}{t_{n-k}} - \frac{(\zeta_i^1 - S_i)^+}{t_i} \\
&+ \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} \frac{(S_i - (\zeta_i^1 - S_i)^+)}{t_i} \geq \beta^{\text{obj}} \sum_{i=1}^n \frac{\zeta_i^2}{t_i}, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
\sum_{i=1}^{n-k-1} \frac{\epsilon_i(\zeta^2)}{t_i} + \frac{\zeta_{n-k}}{t_{n-k}} \\
&+ \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \mathcal{I}^{n-k+1} \setminus \alpha} \frac{(S_i - (\zeta_i^1 - S_i)^+)}{t_i} \geq \beta^{\text{obj}} \sum_{i=1}^n \frac{\zeta_i^2}{t_i}, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \alpha \subseteq \mathcal{I}^{n-k+1}, \\
\zeta_i &\geq \epsilon_i(\zeta^2) \geq 0, & \forall i \in \mathcal{I} \setminus \mathcal{I}^{n-k}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\eta_i(\zeta^2) &\geq 0, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}.
\end{aligned} \tag{4.11}$$

By the principle of mathematical induction, Lemma 4.1 holds for all $k \in \mathcal{I}$. \square

We apply Lemma 4.1 with $k = n$. This eliminates all $\epsilon_i(\zeta^2)$, i in \mathcal{I} , from Problem (4.9), resulting in Problem (4.12):

$$\begin{aligned}
\min_{\substack{S \in \mathbb{N}_0^n \\ \eta_i \in \mathbb{R}^n}} & \sum_{i \in \mathcal{I}} c_i^h S_i + \eta \\
\text{s.t.} & S_i - (\zeta_i^1 - S_i)^+ \geq \zeta_i^2 - \frac{t_i \eta_i(\zeta^2)}{c_i^{\text{bo}}}, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
& \sum_{i \in \mathcal{I}} \eta_i(\zeta^2) \leq \eta, & \forall (\zeta^1, \zeta^2) \in \mathcal{D}, \\
& \sum_{i \in \mathcal{I} \setminus \alpha} \frac{(S_i - (\zeta_i^1 - S_i)^+)}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \alpha} \frac{\zeta_i^2}{t_i}, & \forall \alpha \subseteq \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D},
\end{aligned}$$

$$\begin{aligned}
S_i &\geq (\zeta_i^1 - S_i)^+, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}, \\
\eta_i(\zeta^2) &\geq 0, & \forall i \in \mathcal{I}, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.12)
\end{aligned}$$

One can easily eliminate all $\eta_i(\zeta_n^2)$, for any i in \mathcal{I} of Problem (4.12) to prove Theorem 4.1.

4.A.2 Alternative Robust Optimization Models with Backorders

In this section, we first present an adaptation of the emergency shipment model of Chapter 3 to the backorder case and discuss its limitations. We then improve it with a safety stock formulation to handle historical backorders and analyze why this approach remains conservative.

Although Chapter 3 assumes a simplified setting with identical repair lead times across all SKUs, this chapter advances our model to reflect real-world complexity by incorporating distinct lead times for each SKU. This extension requires a modification of the aggregate fill rate constraint, where each term is normalized by dividing by the corresponding lead time t_i .

The first robust optimization model, discussed in the Master thesis of Pessers (2024), adapts Problem (3.5) in Chapter 3 to the backorder case by replacing c_i^{em} by c_i^{bo} . For each SKU i , the demand uncertainty ζ_i lies within an uncertainty set \mathcal{D} .

$$\begin{aligned}
&\min_{\substack{S \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}^n}} \sum_{i \in \mathcal{I}} c_i^{\text{h}} S_i + \eta_i \\
&\text{s.t.} \quad \frac{1}{t_i} \zeta_i (1 - \beta_i(\zeta)) c_i^{\text{bo}} \leq \eta_i, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
&\quad \beta_i(\zeta) \zeta_i \leq S_i, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
&\quad \sum_{i \in \mathcal{I}} \frac{\beta_i(\zeta) \zeta_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i}{t_i}, & \forall \zeta \in \mathcal{D}, \\
&\quad 0 \leq \beta_i(\zeta) \leq 1, & \forall i \in \mathcal{I}, \zeta \in \mathcal{D}, \\
&\quad \eta_i \geq 0, & \forall i \in \mathcal{I}.
\end{aligned} \quad (4.13)$$

Problem (4.13) can serve as an approximation for backorder situations and be solved using our algorithm proposed for Problem (3.5). However, in a backorder situation, unmet demand carries over through time, creating temporal dependencies that Problem (4.13) cannot capture. This makes the model overly conservative as it essentially requires independent full fulfillment of worst-case demand in each lead time. In addition, the model fails to accurately represent the backorder dynamics

since it does not account for how backordered demand affects available stock and subsequent fill rates, leading to incorrect calculations of both on-hand inventory and fill rates.

To address this limitation, we propose an alternative formulation that introduces an additional safety stock variable $s_i (\geq 0)$ for each SKU $i \in \mathcal{I}$ to cover backorders from the previous period. The process unfolds as follows. During the lead time t_i , we observe the demand $\zeta_i^1 (\geq 0)$, which is first partially satisfied from the base stock level S_i , and any backordered portion is then satisfied using the safety stock s_i . At the end of the lead time, we encounter demand $\zeta_i^2 (\geq 0)$ and determine β_i , the fraction of ζ_i^2 that can be satisfied from the base stock level S_i . The unfilled demand $(1 - \beta_i)\zeta_i^2$ becomes backorder and will again be satisfied by s_i . If s_i is insufficient, we calculate the new backorder and update s_i to ensure that it covers any remaining backorders. This process repeats cyclically, adapting to uncertain demand scenarios within the set \mathcal{D} . So, the ARO Problem (4.14) is:

$$\min_{\substack{S, s \in \mathbb{N}_0^n \\ \beta_i: \mathbb{R}^n \rightarrow \mathbb{R} \\ \eta \in \mathbb{R}}} \sum_{i \in \mathcal{I}} c_i^h (S_i + s_i) + \eta \quad (4.14a)$$

$$\text{s.t. } \zeta_i^1 \leq S_i + s_i, \quad \forall \zeta^1 \in \mathcal{D}, \quad (4.14b)$$

$$\sum_{i \in \mathcal{I}} \eta_i(\zeta^2) \leq \eta, \quad \forall \zeta^2 \in \mathcal{D}, \quad (4.14c)$$

$$\frac{1}{t_i} (\zeta_i^2 - S_i) c_i^{\text{bo}} \leq \eta_i(\zeta^2), \quad \forall i \in \mathcal{I}, \zeta^2 \in \mathcal{D}, \quad (4.14d)$$

$$\beta_i(\zeta^2) \zeta_i^2 \leq S_i, \quad \forall i \in \mathcal{I}, \zeta^2 \in \mathcal{D}, \quad (4.14e)$$

$$\sum_{i \in \mathcal{I}} \frac{\beta_i(\zeta^2) \zeta_i^2}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i}, \quad \forall \zeta^2 \in \mathcal{D}, \quad (4.14f)$$

$$0 \leq \beta_i(\zeta^2) \leq 1, \quad \forall i \in \mathcal{I}, \zeta^2 \in \mathcal{D}, \quad (4.14g)$$

$$\eta_i(\zeta^2) \geq 0, \quad \forall i \in \mathcal{I}, \zeta^2 \in \mathcal{D}. \quad (4.14h)$$

Here, η is the total amount of backordered cost per time unit, which is introduced to move the uncertainty from the objective function to Constraint (4.14c), η_i is the amount of backordered cost per time unit per SKU, which depends on the realization of the actual demand ζ^2 . Constraint (4.14b) ensures sufficient stock to cover historical demand for each SKU independently. Constraint (4.14e) ensures that the stock level is sufficient to meet part of the demand ζ^2 . Constraint (4.14g) ensures that β_i lies between 0 and 1, and Constraint (4.14h) guarantees that the backorder amount is non-negative.

Although Problem (4.14) provides better control over backorders compared to Problem (4.13), it remains conservative as it requires full coverage of worst-case historical demand for each SKU individually. In addition, Problem (4.14) ignores correlations between ζ^1 and ζ^2 , which may lead to unnecessary stock. To achieve a less conservative approach that better reflects the dynamic nature of continuous review systems, we propose Problem (4.2).

As demonstrated by our numerical comparison in Appendix 4.A.3, Problem (4.2) allows backorders to persist over time, while Problem (4.14) requires full fulfillment of historical demand through additional safety stock s_i .

4.A.3 Comparing Problems (4.2) and (4.14): A Numerical Example

To compare Problems (4.2) and (4.14), we analyze an illustrative example of nine SKUs using the same input parameters as shown in Table 2.1, with c_i^{BO} of 750 Euros per SKU. We generate solutions for both problems using the affine decision approach.

Figure 4.5 presents the resulting stock levels, demonstrating that Problem (4.14) consistently prescribes higher stock levels than Problem (4.2), particularly for SKUs with medium and high prices. In particular, the stock levels of Problem (4.14) remain constant in different β^{obj} , reflecting its conservativeness to maintain additional safety stock to cover historical demand. In contrast, Problem (4.2) exhibits more flexible stock level adjustments as β^{obj} varies, especially for SKUs with medium to high demand rates. Although problem (4.14) provides stronger protection against stockouts through higher stock levels, this conservative approach may result in unnecessary holding costs.

4.A.4 An Algorithm for the Budget Uncertainty Set

In this section, we present an efficient approximation algorithm for computing the right-hand side values of Constraint (4.8b), motivated by the computational complexity considerations discussed in Appendix 2.A.6. We begin by reformulating Constraint (4.8b) in an equivalent form

$$\sum_{i \in \alpha} \frac{S_i}{t_i} \geq \beta^{\text{obj}} \sum_{i \in \mathcal{I}} \frac{\zeta_i^2}{t_i} - \sum_{i \in \mathcal{I} \setminus \alpha} \frac{\zeta_i^2}{t_i} + \sum_{i \in \alpha} \frac{\zeta_i^1}{t_i}, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset, (\zeta^1, \zeta^2) \in \mathcal{D}. \quad (4.15)$$

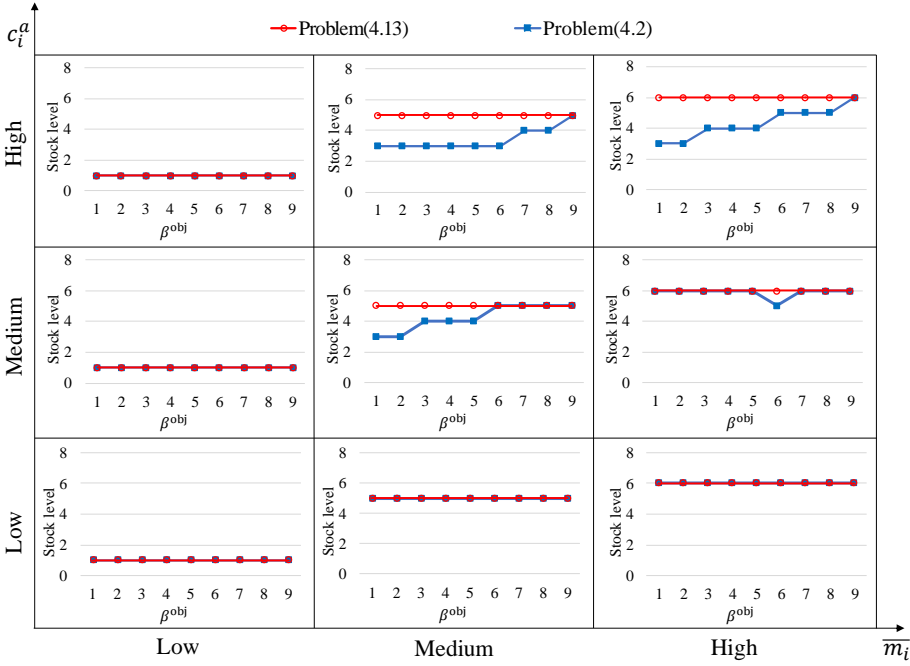


Figure 4.5: Solutions to Problems (4.2) and (4.14) for the illustrative example.

Let \tilde{b}_α denote the values on the right-hand side of Constraint (4.15). We propose the following approximation:

$$\tilde{b}_\alpha := \left\lceil \sum_{i \in \alpha} \frac{\beta^{\text{obj}} \overline{d}_i^{12} + (1 - \beta^{\text{obj}}) \overline{d}_i^1}{t_i} + (\beta^{\text{obj}} - 1) \underline{\Gamma}_{\mathcal{I} \setminus \alpha} \right\rceil, \quad \forall \alpha \subseteq \mathcal{I}, \alpha \neq \emptyset. \quad (4.16)$$

Our approximation works by decomposing the right-hand side into three parts. For any SKU in set α , we use \overline{d}_i^{12} to account for the maximum aggregate demand over two consecutive lead times, weighted by β^{obj} . For these same SKUs, we add $(1 - \beta^{\text{obj}}) \overline{d}_i^1$ to account for the maximum demand in the first period. For SKUs not in set α (i.e., $\mathcal{I} \setminus \alpha$), we include $\underline{\Gamma}_{\mathcal{I} \setminus \alpha}$ to account for the minimum aggregate demand, weighted by $(\beta^{\text{obj}} - 1)$. We then take the ceiling of this sum to ensure integer values, as fractional stock levels are not allowed.

Chapter 5

Conclusions

In this thesis, we develop robust optimization approaches for spare parts inventory control across different warehouse settings. Section 5.1 revisits our main findings and contributions, demonstrating four key advances in spare parts inventory control. We then discuss promising directions for future research in Section 5.2, focusing on both modeling and algorithmic improvements.

5.1. Research Topics Revisited

We study spare parts inventory control under high demand uncertainty. In Chapters 2 and 3, we propose a robust optimization approach for spare parts inventory control, developing algorithms for a local warehouse with lost sales and one with emergency shipments, respectively. In Chapter 4, we consider another setting and develop a robust optimization model for a central warehouse with backorders. These three chapters demonstrate that properly addressing demand uncertainty through robust optimization can lead to substantial cost savings and service level improvements for service providers. The methodology in each chapter is illustrated in Figure 5.1, which shows the key components of each chapter: input, problem formulation, solution methods, and performance evaluation. Our research demonstrates four key advances in spare parts inventory management.

First, we show that robust optimization has great potential to handle demand uncertainty in different warehouse settings. Starting with a lost sales model at local warehouses (Chapter 2), we extend our approach to incorporate emergency shipments (Chapter 3) and finally address the complexities of backorders at central

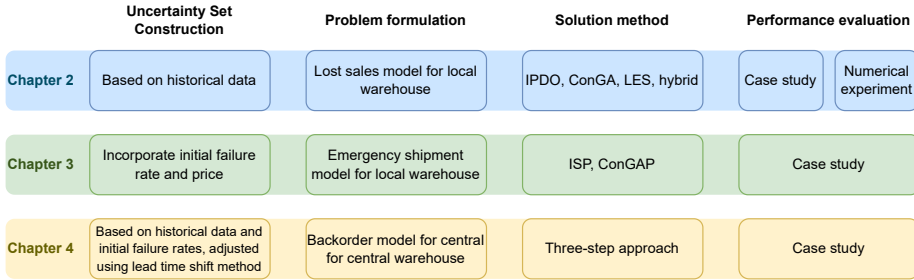


Figure 5.1: Methodology development and implementation in thesis chapters.

warehouses (Chapter 4). Each model demonstrates superior performance compared to the stochastic optimization model under the assumption of Poisson failures, particularly when dealing with non-Poissonian demand or limited historical data.

Second, we address the computational challenges of implementing robust optimization in spare parts inventory control. Our solution methods evolve from the basic Iterative Projection in Descending Order (IPDO) algorithm to specialized approaches for emergency shipments (Iterative Stocking including Preprocessing (ISP) algorithm, ConGA including Preprocessing (ConGAP) algorithm), and finally to a three-step approach for the central warehouse problem. These methods make our models computationally tractable for large-scale industrial applications with thousands of components.

Third, we develop increasingly sophisticated methods for constructing uncertainty sets. We progress from basic historical data-based sets to incorporating initial failure rates for the new product introduction phase, and finally to handling varying lead times through our lead time shift method. This progression reflects the increasing complexity of the demand interaction between components in real-world spare parts inventory systems and shows how robust optimization can adapt to different operational contexts.

Fourth, the case studies at ASML demonstrate the practical value of these methodological advances. Our robust optimization models achieve higher service levels than conventional approaches in all three warehouse settings while maintaining cost efficiency. Our research contributes to both the theoretical advancement of robust optimization techniques and their practical application in spare parts inventory management. The models and algorithms developed in this thesis can be applied

to various capital goods industries where service providers face similar challenges of high demand uncertainty and stringent service level requirements.

5.2. Future Research Directions

While we validate our models through case studies in each chapter, a comprehensive validation framework is needed to measure real-world performance. This includes designing pilot studies, establishing appropriate performance metrics, and creating monitoring systems to track long-term benefits. The framework should account for both quantitative measures, such as cost savings, and qualitative aspects, such as ease of implementation. Our models have potential applications across various industries beyond semiconductor equipment. In sectors such as aerospace, healthcare, and oil and gas, where equipment downtime can lead to significant consequences, our robust optimization approach could be adapted to meet industry-specific service requirements.

We now provide an outlook for further research, structured into two main categories: Model Extensions and Algorithmic Improvements.

Model Extensions: Our research opens several opportunities to improve inventory models for real-world applications.

First, our current models assume linear cost structures in the objective function, where costs increase proportionally with quantities. For example, we assume that the holding costs are linear in the base stock levels. While this simplification enables tractable analysis, it fails to capture the nonlinear cost behaviors pervasive in practice. In practice, many cost components exhibit nonlinear behavior. As shown in Johansson and Olsson (2017) and Lamghari-Idrissi et al. (2021), in many practical scenarios, service providers face fixed penalty fees when delivery times exceed certain thresholds, creating stepwise cost functions. Ignoring these nonlinearities risks flawed decisions, such as underestimating penalties when delays exceed contractual thresholds (e.g., 15% delay triggers a fee, while 10% does not). Despite their practical relevance, extending such nonlinear cost structures to robust optimization remains an open challenge despite their practical relevance. Incorporating these nonlinear cost structures would require new theoretical developments in nonlinear robust optimization. These developments would need to focus on creating tractable reformulations when both the constraints and objective function contain uncertain-

ties, designing efficient solution methods that can handle nonlinear objectives while maintaining computational feasibility for large-scale problems.

Second, while our current framework constructs uncertainty sets based on historical data and initial failure rates, it could be extended to incorporate real-time sensor data and installed base information. Elwany and Gebraeel (2008) pioneer sensor-driven prognostic models for component replacement and spare parts inventory, showing how condition-based sensor data can enable dynamic updating of decisions based on the physical condition of the equipment. Van der Auweraer et al. (2019) provide a comprehensive review of how installed base information can improve spare parts demand forecasting. Building on these works, future research could develop methods that dynamically adjust uncertainty sets based on both sensor data and installed base information. This would require new theoretical developments in robust optimization to efficiently update models as new data arrive and to incorporate equipment-specific degradation characteristics into the uncertainty sets.

5

Third, while we have developed separate models for local and central warehouses, future research could extend to multi-location systems with lateral transshipments under uncertainty. This would involve developing robust optimization models that simultaneously optimize stock levels throughout the entire network of central and local warehouses. The challenge lies in capturing the complex interactions between different echelons, including how demand uncertainty propagates through the network, how lateral transshipments between local warehouses affect system performance, and how to balance the trade-off between central and local stock levels. To address this challenge, a promising direction would be to explore decoupling approximations inspired by Section 5.4 of Van Houtum and Kranenburg (2015). By decomposing the network into subsystems (e.g., separating local warehouses from the central warehouse), we can model emergency shipments at the local level independently while capturing their aggregate impact on central warehouse replenishment. This approach avoids the need for closed-form solutions for the entire network. Other new theoretical developments are also worth exploring to maintain computational tractability while handling the complexity of network-wide optimization under uncertainty.

Fourth, our models could be enhanced to support customized service levels for different customer segments, allowing service providers to differentiate their offerings

based on customer priorities and willingness to pay. This differentiation could enable more efficient resource allocation while maintaining appropriate service levels for each customer category.

Fifth, while our research compares the robust optimization approach with the commonly used stochastic optimization model that assumes a Poisson demand process, future work should conduct comprehensive comparisons with other methods that effectively handle demand uncertainty. The commonly used model assumes a known demand rate, but demand rates are often uncertain in practice. Van Wingerden (2019, Chapter 2) show that when demand follows a Poisson process with an uncertain rate, this uncertainty has an equivalent impact on lead time demand as lead time variability. Moreover, they demonstrate that higher demand rates amplify the effect of additional demand uncertainty, whereas longer lead times reduce their relative impact. Some studies extend the Poisson framework using compound distributions, which allow for modeling demand rates and inter-demand intervals as separate variables. Feeney and Sherbrooke (1966) generalize Palm's theorem of Poisson demand (Palm, 1938) to compound Poisson demand, proving that steady-state probabilities maintain the same compound Poisson form while still depending only on the mean resupply lead time. Kiesmüller et al. (2004) then develop analytical approximations for divergent N-echelon networks under compound renewal demand, allowing for arbitrary distributions of inter-arrival times and demand rates. Grob and Bley (2018) further evaluate different wait time approximations in distribution networks with compound renewal customer demand processes. Future research could systematically compare these approaches by analyzing their relative strengths in handling demand uncertainty, computational efficiency, and practical implementation challenges.

Finally, our robust optimization framework could be extended to incorporate human decision-making behavior. De Kok (2018) emphasizes the importance of distinguishing between intervention-independent performance measures and human-driven adjustments, highlighting how empirical validation of inventory policies can align with planners' tacit knowledge while maintaining mathematical rigor. Käki et al. (2019) show that analyzing deviations between model recommendations and actual decisions can provide valuable insight into improving decision support systems. In the context of spare parts inventory, future research could investigate how inventory managers deviate from robust optimization recommendations, particularly under high uncertainty during the new product introduction phase.

Understanding these deviations could help improve both model design and implementation. Our observations at ASML suggest that RO solutions better align with inventory planners' decision-making patterns. The SO solution may recommend high stock levels for inexpensive parts and zero stock levels for expensive ones under Poisson demand assumptions, while the RO solution typically suggests more balanced inventory levels. Although planners' intuitions about inventory decisions may not always be correct, analyzing these decision patterns and the reasoning behind them can provide valuable insights for developing more practical inventory models.

Algorithmic Improvements: Our solution methods progress from basic algorithms (IPDO in Chapter 2) to more sophisticated approaches (ISP and ConGAP in Chapter 3) and finally to the three-step solution approach in Chapter 4. Building on this evolution, we identify several directions for algorithmic improvements.

First, modern machine learning techniques could enhance our solution algorithms in multiple ways. For example, deep learning models could be trained to predict good initial solutions for our iterative algorithms, potentially reducing computation time. Recent research demonstrates how end-to-end deep learning models can directly generate inventory decisions (Qi et al., 2023) and how recurrent neural networks can efficiently handle large-scale production networks (Wang and Hong, 2023). Using these techniques to understand the relationship between the problem parameters and optimal solutions, we could develop fast approximation schemes that enable real-time decision support.

Second, our hybrid approach in Chapter 2 demonstrates the benefits of combining different solution methods. This concept could be extended through metaheuristic algorithms, particularly useful for the complex constraints in our emergency shipment model in Chapter 3 and the model with backorders in Chapter 4. For example, genetic algorithms could be developed to efficiently search the solution space for large-scale problems, using crossover and mutation operators specifically designed for inventory problems.

Third, the dynamic nature of our uncertainty set construction, especially evident in Chapter 3's phased approach, suggests the need for online algorithms that efficiently update solutions as new data arrive. This would involve creating incremental update methods that avoid solving the full problem from scratch, which is particularly important for dynamic uncertainty set adjustments. The challenge lies

in maintaining solution quality while meeting strict computational time requirements for real-time decision-making.

Conclusion: We have provided directions for future research in robust optimization for spare parts inventory control. Our work has opened up several promising theoretical extensions, from incorporating lead time uncertainty to developing multi-echelon models. We have also identified opportunities for algorithmic improvements through machine learning and metaheuristics, along with practical applications across different industries and customer segments. Through these research directions, we hope to inspire both practitioners and researchers. For practitioners, our work provides a foundation for implementing robust optimization in real-world inventory management, particularly in environments with high demand uncertainty. For researchers, we hope that this thesis encourages further exploration of robust optimization approaches in inventory control, especially in addressing the theoretical and computational challenges we have identified.

Bibliography

- Air Canada. Annual report 2023. <https://aircanada.investorroom.com/annual-reports>, 2024. Accessed: 2024-05-28.
- A. Ardestani-Jaafari and E. Delage. Robust optimization of sums of piecewise linear functions with application to inventory problems. *Operations Research*, 64(2):474–494, 2016.
- K.-P. Aronis, I. Magou, R. Dekker, and G. Tagaras. Inventory control of spare parts using a bayesian approach: A case study. *European Journal of Operational Research*, 154(3):730–739, 2004.
- Asia Financial. TSMC predicts \$60m hit from taiwan’s biggest quake in 25 years, 4 2024. URL <https://www.asiafinancial.com/tsmc-predicts-60m-hit-from-taiwans-biggest-quake-in-25-years>.
- ASML. ASML - supply chain management, Apr 2014. URL https://www.youtube.com/watch?v=7JwSF_6sdyo.
- ASML. ASML 2022 annual report, Feb 2023. URL <https://www.asml.com/en/investors/annual-report/2022/>.
- ASML. ASML customer support, Apr 2025. URL <https://www.asml.com/en/products/customer-support>.
- S. Axsäter. Modelling emergency lateral transshipments in inventory systems. *Management Science*, 36(11):1329–1338, 1990.
- M. Z. Babai, H. Chen, A. A. Syntetos, and D. Lengu. A compound-poisson bayesian approach for spare parts inventory forecasting. *International Journal of Production Economics*, 232:107954, 2021.
- R. Basten and G.-J. van Houtum. Spare parts inventory planning. In *Research*

- Handbook on Inventory Management*, pages 455–475. Edward Elgar Publishing, 2023.
- R. J. I. Basten and G.-J. Van Houtum. System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science*, 19(1):34–55, 2014.
- M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.
- A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.
- D. S. Bernstein. Matrix mathematics. In *Matrix Mathematics*. Princeton University Press, 2009.
- D. Bertsimas and C. Caramanis. Finite adaptability in multistage linear optimization. *IEEE Transactions on Automatic Control*, 55(12):2751–2766, 2010.
- D. Bertsimas and F. J. De Ruiter. Duality in two-stage adaptive linear optimization: Faster computation and stronger bounds. *INFORMS Journal on Computing*, 28(3): 500–511, 2016.
- D. Bertsimas and D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas LLC, 2022. ISBN 9781733788526. URL https://books.google.nl/books?id=V_RPzweECAAJ.
- D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.
- D. Bertsimas and A. Thiele. A robust optimization approach to inventory theory. *Operations Research*, 54(1):150–168, 2006.
- D. Bertsimas, V. Goyal, and X. A. Sun. A geometric characterization of the power of finite adaptability in multistage stochastic and adaptive optimization. *Mathematics of Operations Research*, 36(1):24–54, 2011.
- D. Bertsimas, I. Dunning, and M. Lubin. Reformulation versus cutting-planes for robust optimization: A computational study. *Computational Management Science*, 13:195–217, 2016.
- D. Bienstock and N. Özbay. Computing robust basestock levels. *Discrete Optimiza-*

- tion, 5(2):389–414, 2008.
- T. Burgin. The gamma distribution and inventory control. *Journal of the Operational Research Society*, 26(3):507–525, 1975.
- Business Insider. Spare parts shortages are forcing airlines to ground planes, 7 2022. URL <https://www.businessinsider.com/airlines-ground-planes-due-to-shortage-of-plane-parts-report-2022-7?international=true&r=US&IR=T>.
- D. Caglar, C.-L. Li, and D. Simchi-Levi. Two-echelon spare parts inventory system subject to a service constraint. *IIE Transactions*, 36(7):655–666, 2004.
- C. Chaves and A. Gosavi. On general multi-server queues with non-poisson arrivals and medium traffic: a new approximation and a COVID-19 ventilator case study. *Operational Research*, 22(5):5205–5229, 2022.
- L. Chen, M. Gendreau, M. H. Hà, and A. Langevin. A robust optimization approach for the road network daily maintenance routing problem with uncertain service time. *Transportation Research Part E: Logistics and Transportation Review*, 85:40–51, 2016.
- Y. Chen, G. Iyengar, and C. Wang. Robust inventory management: A cycle-based approach. *Manufacturing & Service Operations Management*, 25(2):581–594, 2023.
- Z. Chen, M. Sim, and P. Xiong. Robust stochastic optimization made easy with R SOME. *Management Science*, 66(8):3329–3339, 2020.
- CNBC. Inside ASML, the company advanced chipmakers use for EUV lithography. <https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html>, 2022. Accessed: 2024-05-12.
- Companies Market Cap. Largest tech companies by market cap. <https://companiesmarketcap.com/tech/largest-tech-companies-by-market-cap/>, 2024. Accessed: 2024-05-12.
- F. Costantino, G. Di Gravio, R. Patriarca, and L. Petrella. Spare parts management for irregular demand items. *Omega*, 81:57–66, 2018.
- J.-F. Côté, M. Gendreau, and J.-Y. Potvin. An exact algorithm for the two-dimensional orthogonal packing problem with unloading constraints. *Operations*

- Research*, 62(5):1126–1141, 2014.
- J.-F. Côté, M. Haouari, and M. Iori. Combinatorial benders decomposition for the two-dimensional bin packing problem. *INFORMS Journal on Computing*, 33(3): 963–978, 2021.
- T. De Kok. Inventory management: Modeling real-life supply chains and empirical validity. *Foundations and Trends® in Technology, Information and Operations Management*, 11(4):343–437, 2018. ISSN 1571-9545. doi: 10.1561/0200000057. URL <http://dx.doi.org/10.1561/0200000057>.
- J. R. do Rego and M. A. De Mesquita. Demand forecasting and inventory control: A simulation study on automotive spare parts. *International Journal of Production Economics*, 161:1–16, 2015.
- M. Drent and J. Arts. Expediting in two-echelon spare parts inventory systems. *Manufacturing & Service Operations Management*, 23(6):1431–1448, 2021.
- O. El Housni and V. Goyal. On the optimality of affine policies for budgeted uncertainty sets. *Mathematics of Operations Research*, 46(2):674–711, 2021.
- A. H. Elwany and N. Z. Gebraeel. Sensor-driven prognostic models for equipment replacement and spare parts inventory. *IIE Transactions*, 40(7):629–639, 2008.
- G. J. Feeney and C. C. Sherbrooke. The $(s-1, s)$ inventory policy under compound poisson demand. *Management Science*, 12(5):391–411, 1966.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian data analysis third edition. *Chapman and Hall/CRC*, 2013.
- B. Ghodrati. Weibull and exponential renewal models in spare parts estimation: a comparison. *International Journal of Performability Engineering*, 2(2):135, 2006.
- C. Grob and A. Bley. Comparison of wait time approximations in distribution networks using (r, q) -order policies. *arXiv preprint arXiv:1801.09617*, 2018.
- L. Gurobi Optimization. Gurobi optimizer reference manual, 2018.
- M. S. Hamada, H. F. Martz, C. S. Reese, and A. G. Wilson. *Bayesian Reliability*, volume 15. Springer, 2008.
- Q. Hu, J. E. Boylan, H. Chen, and A. Labib. OR in spare parts management: A review. *European Journal of Operational Research*, 266(2):395–414, 2018.

- L. Johansson and F. Olsson. Quantifying sustainable control of inventory systems with non-linear backorder costs. *Annals of Operations Research*, 259:217–239, 2017.
- A. Käki, K. Kempainen, and J. Liesiö. What to do when decision-makers deviate from model recommendations? empirical evidence from hydropower industry. *European Journal of Operational Research*, 278(3):869–882, 2019.
- G. P. Kiesmüller, T. G. de Kok, S. R. Smits, and P. J. van Laarhoven. Evaluation of divergent n-echelon (s, nq)-policies under compound renewal demand. *Or Spectrum*, 26:547–577, 2004.
- KLM. Financial statement 2023. <https://www.klmanualreport.com>, 2024. Accessed: 2024-05-28.
- D. Lamghari-Idrissi, R. Basten, and G.-J. van Houtum. Reducing risks in spare parts service contracts with a long downtime constraint. *IIEE Transactions*, 53(10):1067–1080, 2021.
- D. Lamghari-Idrissi, R. van Hugten, G.-J. van Houtum, and R. Basten. Increasing chip availability through a new after-sales service supply concept at asml. *INFORMS Journal on Applied Analytics*, 52(5):460–470, 2022.
- Y. F. Lim and C. Wang. Inventory management based on target-oriented robust optimization. *Management Science*, 63(12):4409–4427, 2017.
- Y. Liu and A. I. Abeyratne. *Practical applications of Bayesian reliability*. John Wiley & Sons, 2019.
- A. Marandi and D. Den Hertog. When are static and adjustable robust optimization problems with constraint-wise uncertainty equivalent? *Mathematical Programming*, 170:555–568, 2018.
- D. Murthy and D. Nguyen. Study of a multi-component system with failure interaction. *European Journal of Operational Research*, 21(3):330–338, 1985.
- E. Özkan and G.-J. van Houtum. Joint inventory and scheduling control in a repair facility. *Operations Research*, 2023.
- C. Palm. Analysis of the erlang traffic formula for busy-signal arrangements. *Ericsson Technics*, 5(9):39–58, 1938.
- H. Pessers. Implementation of adaptive robust optimization in the spare parts

- service network of asml. Master's thesis, Eindhoven University of Technology, 2024.
- A. A. Pessoa, M. Poss, R. Sadykov, and F. Vanderbeck. Branch-cut-and-price for the robust capacitated vehicle routing problem with knapsack uncertainty. *Operations Research*, 69(3):739–754, 2021.
- M. Qi, Y. Shi, Y. Qi, C. Ma, R. Yuan, D. Wu, and Z.-J. Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, 69(2):759–773, 2023.
- C. C. Sherbrooke. Metric: A multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968.
- C. C. Sherbrooke. *Optimal inventory modeling of systems: multi-echelon techniques*, volume 72. Springer Science & Business Media, 2006.
- SIEMENS. The true cost of downtime 2022, April 2023. URL <https://assets.new.siemens.com/siemens/assets/api/uuid:3d606495-dbe0-43e4-80b1-d04e27ada920/dics-b10153-00-7600truecostofdowntime2022-144.pdf>.
- Y. Tan, A. A. Paul, Q. Deng, and L. Wei. Mitigating inventory overstocking: Optimal order-up-to level to achieve a target fill rate over a finite horizon. *Production and Operations Management*, 26(11):1971–1988, 2017.
- K. Tarasov. ASML is the only company making the \$200 million machines needed to print every advanced microchip., March 2022. URL <https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html>.
- R. H. Teunter, A. A. Syntetos, and M. Z. Babai. Stock keeping unit fill rate specification. *European Journal of Operational Research*, 259(3):917–925, 2017.
- U. W. Thonemann, A. O. Brown, and W. H. Hausman. Easy quantification of improved spare parts inventory policies. *Management Science*, 48(9):1213–1225, 2002.
- A. Thorsen and T. Yao. Robust inventory control under demand and lead time uncertainty. *Annals of Operations Research*, 257:207–236, 2017.

- L. Turrini and J. Meissner. Spare parts inventory management: New evidence from distribution fitting. *European Journal of Operational Research*, 273(1):118–130, 2019.
- S. Van der Auweraer, R. N. Boute, and A. A. Syntetos. Forecasting spare part demand with installed base information: A review. *International Journal of Forecasting*, 35(1):181–196, 2019.
- G.-J. Van Houtum and B. Kranenburg. *Spare parts inventory control under system availability constraints*, volume 227. Springer, 2015.
- E. Van Wingerden. *System-focused spare parts management for capital goods*. PhD thesis, Technische Universiteit Eindhoven, 2019.
- E. Van Wingerden, T. Tan, and G.-J. Van Houtum. The impact of an emergency warehouse in a two-echelon spare parts network. *European Journal of Operational Research*, 276(3):983–997, 2019.
- T. Wang and L. J. Hong. Large-scale inventory optimization: A recurrent neural networks–inspired simulation approach. *INFORMS Journal on Computing*, 35(1):196–215, 2023.
- H. Wong, G.-J. van Houtum, D. Cattrysse, and D. Van Oudheusden. Multi-item spare parts systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research*, 171(3):1071–1093, 2006.
- H. Wong, B. Kranenburg, G.-J. Van Houtum, and D. Cattrysse. Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR Spectrum*, 29:699–722, 2007.
- M. Wood. Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods*, 8(4):454–470, 2005.
- J. Zhen, D. Den Hertog, and M. Sim. Adjustable robust optimization via fourier–motzkin elimination. *Operations Research*, 66(4):1086–1100, 2018.
- J. Zhen, A. Marandi, D. de Moor, D. den Hertog, and L. Vandenbergh. Disjoint bilinear optimization: A two-stage robust optimization perspective. *INFORMS Journal on Computing*, 34(5):2410–2427, 2022.

Summary

Robust Spare Parts Inventory Management

This thesis advances the field of spare parts inventory management by developing robust optimization approaches to handle high demand uncertainty. We address three key settings in the spare parts supply chain: local warehouses with lost sales, local warehouses with emergency shipments, and central warehouses with backorders.

Local Warehouse with Lost Sales: We address this research topic in Chapter 2. We consider the base case that when the local warehouse cannot fulfill demand immediately, the demand is considered lost as service providers seek alternative solutions. We develop an adjustable robust optimization (ARO) model for this multi-component, single-location inventory control problem. The key challenge lies in efficiently solving the computationally intensive ARO model. We first prove that the ARO model can be reformulated into a deterministic counterpart, though this reformulation initially yields an exponential number of constraints. By analyzing the structure of the optimal solution, we develop an algorithm called Iterative Projection in Descending Order (IPDO), which achieves optimal solutions under some conditions.

Recognizing that large-scale inventory problems require more efficient computational methods, we develop two heuristic algorithms based on IPDO's foundation. The Constraint Generation (ConGA) algorithm provides near-optimal solutions efficiently, while the Linear Equation System (LES) algorithm offers exceptional computational speed, capable of processing hundreds of components in seconds. We then develop a hybrid approach that dynamically combines these methods, offering a flexible framework that balances solution quality with computational efficiency.

Our extensive computational studies, including both simulation experiments with various demand patterns and a real-world case study at ASML involving 710 components, demonstrate the superiority of our approach: The results show that our

ARO model consistently outperforms the conventional method in achieving target fill rates, especially when dealing with non-Poissonian demand patterns. This research topic establishes the methodological foundation for the settings addressed in subsequent chapters.

Local Warehouse with Emergency Shipments: We address this research topic in Chapter 3. We consider emergency shipments as an alternative fulfillment option at local warehouses. We develop an ARO model for spare parts inventory control with emergency shipments. To make the model computationally tractable, we reformulate the ARO model into a deterministic counterpart and then decompose it into two mixed-integer optimization problems. Building on this reformulation, we propose the Iterative Stocking including Preprocessing (ISP) and ConGA including Preprocessing (ConGAP) algorithms that can efficiently solve problems involving thousands of components.

We propose a phased approach to construct uncertainty sets when historical demand data are limited. This approach incorporates the initial failure rate (IFR) provided by reliability engineers. When there is very little demand data, the uncertainty set is built using only the IFR. As more demand data becomes available over time, we gradually decrease the weight given to the IFR in the uncertainty set construction.

The case study at ASML shows the advantages of the ARO model over the stochastic optimization model. With the same total cost, the ARO model reduces the mean waiting time by up to 3.5 hours. This reduction in waiting time could save more than €250,000 in lost production costs per lithography system breakdown. The results demonstrate that the ARO model offers robust and cost-effective solutions for the control of spare parts inventory.

Central Warehouse with Backorders: We address this research topic in Chapter 4. We study spare parts inventory control at the central warehouse. The central warehouse plays a crucial role, as it receives parts directly from suppliers and serves as the emergency shipment source for local warehouses. At the central warehouse, when stockouts occur, orders will be backordered because emergency shipments are not possible.

To the best of our knowledge, we are the first to formulate a continuous review inventory model with backorders using robust optimization. We develop an ARO model for the central warehouse and solve it using a three-step approach. In the first

step, we find an approximate lower bound of stock levels from the lost sales problem in Chapter 2 and an upper bound from considering worst-case demand. This step allows us to quickly find near-optimal stock levels for about 90% of the spare parts. For components where bounds differ, we introduce a tighter upper bound in the second step through a relaxed version of the ARO problem, which shares structural similarities with the lost sales problem and can be solved using established methods like IPDO and ConGA. Finally, we employ additional approximation methods in the third step for any remaining components. Unlike previous chapters, where we assume identical repair lead times for all SKUs, in Chapter 4, we consider different lead times for each SKU and introduce a lead time shift method to address this reality when constructing uncertainty sets.

The ASML case study shows that our robust optimization model is more cost-efficient than the conventional method as service level requirements increase. This advantage is particularly evident when stringent service levels are needed for critical spare parts management.

Key Advances: Overall, our research demonstrates four fundamental advances in spare parts inventory control. First, we show the potential of the robust optimization approach in different warehouse settings. Second, our progressive development of solution methods makes these ARO models computationally tractable for large-scale industrial applications. Third, we advance increasingly sophisticated methods for uncertainty set construction, from considering only historical data to incorporating initial failure rates and different lead times. Fourth, comprehensive case studies at ASML demonstrate that our approaches achieve higher service performance and maintain cost efficiency in all warehouse settings.

Future Research Directions: This research provides both theoretical advances and practical solutions for spare parts inventory control under uncertainty, creating a strong foundation for future developments in the field. Our models could be extended to include lead-time uncertainty and nonlinear cost structures. For multi-location networks, developing integrated models for central and local warehouses together would be valuable. On the algorithmic front, machine learning techniques could help predict good initial solutions and create adaptive policies. The development of faster solution methods would be particularly valuable for large-scale applications. These advances would further strengthen the practical application of robust optimization in spare parts control.

Acknowledgments

This thesis would not have been possible without the support of many wonderful people. First and foremost, I extend my deepest gratitude to Ahmadreza Marandi, my daily supervisor. As his first Ph.D. student, our collaborative journey through challenges and victories has made this research experience both enriching and memorable.

I am profoundly thankful to Rob Basten for his unwavering support and encouragement. His ability to celebrate every milestone, from manuscript submissions to small breakthroughs, coupled with his invaluable guidance during challenging times, has been instrumental in my growth.

My sincere appreciation goes to Ton de Kok for his exceptional mentorship. I vividly remember his passionate explanation of spare parts inventory concepts during my job interview, which sparked my interest in this field. His wisdom and guidance have been invaluable throughout these years.

I am grateful to Jiawei Zhang for hosting me at NYU, fulfilling my aspiration to experience research in the United States. Special thanks to Joan Stip, Hugo Pessers, and my colleagues at ASML, whose industry insights and collaboration have enriched my research tremendously.

To my fellow OPAC group members, I feel incredibly fortunate to have worked alongside such a diverse and talented group of people from all corners of the world. Our shared experiences and camaraderie have made this journey special.

My heartfelt thanks go to my parents and friends, whose unwavering faith and support have been my cornerstone throughout this journey. Your belief in me has been my greatest motivation in reaching this milestone.

Finally, I acknowledge my own determination in choosing and persevering through this Ph.D. journey. As I look toward new horizons, I will carry forward this

achievement with pride, knowing that my journey, like the ocean ahead, holds boundless possibilities.

About the author

Zhao Kang was born in Shangluo, People's Republic of China, on July 6, 1995. In 2017, he obtained his Bachelor's degree in Transportation Engineering from Chang'an University in China. He then pursued further studies abroad and completed his master's degree in Transport and Planning at Delft University of Technology in 2019.

From 2020 to 2025, he pursued his Ph.D. at the Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, under the supervision of dr. Ahmadreza Marandi, dr. Rob Basten, and prof.dr. Ton de Kok. His research focused on developing robust optimization approaches for spare parts inventory management, with particular emphasis on addressing demand uncertainty in complex supply chain networks. During his Ph.D., he had the opportunity to conduct research visits at New York University Stern School of Business.

In collaboration with ASML and under the supervision of Joan Stip, Zhao successfully translated his theoretical developments into practical applications for managing spare parts inventory in advanced semiconductor manufacturing equipment. His research contributions have been recognized internationally. For example, he received the 2024 Best Student Paper Award from the POMS Supply Chain Management College.

This thesis presents the culmination of his research, which bridges the gap between theoretical advances in robust optimization and their practical implementation in industry settings. Upon completion of his PhD project, Zhao is working as a senior research scientist at Huawei.

