DYNAMIC CAPACITATED LOT SIZING PROBLEM OPTIMIZATION WITH REINFORCEMENT LEARNING

A Thesis

by

Faruk Erdoğan Buldur

Submitted to the
Graduate School of Sciences and Engineering
In Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Data Science

Özyeğin University
August 2023

DYNAMIC CAPACITATED LOT SIZING PROBLEM OPTIMIZATION WITH REINFORCEMENT LEARNING

Approved by:

Prof. Okan Örsan Özener, Advisor
Dept. of Industrial Eng.
Özyeğin University

Asst. Prof. Başak Altan
Dept. of Economics
Özyeğin University

Prof. Serhan Duran
Dept. of Industrial Eng.
Middle East Technical University

Date Approved: 7 August 2023

To my beloved wife and our future baby

ABSTRACT

In today's fast-paced manufacturing and supply chain landscapes, efficiently managing production quantities amidst capacity constraints and demand fluctuations is crucial. This MSc. thesis explores a cutting-edge solution to this challenge by harnessing the power of Reinforcement Learning (RL). Reinventing traditional optimization approaches, RL offers adaptability and intelligence to address the Dynamic Capacitated Lot Sizing Problem (DCLSP). Through extensive experimentation and comparison with conventional methods, our RL-based framework showcases superior performance, reducing costs and enhancing production efficiency. This research unlocks a new era of agile decision-making in complex manufacturing environments and opens doors to further advancements in operations management.

ÖZETÇE

Günümüz hızlı tempolu üretim ve tedarik zinciri ortamlarında, kapasite kısıtları ve talep dalgalanmaları arasında üretim miktarlarını verimli bir şekilde yönetmek son derece önemlidir. Bu Yüksek Lisans tezi, bu zorluğun üstesinden gelmek için Pekiştirmeli Öğrenme'nin (RL) gücünden faydalanarak, geleneksel optimizasyon yaklaşımlarını yeniden tasarlamaktadır. RL, gelişmiş bir esneklik ve zeka sağlayarak, Dinamik Kapasiteli Lot Belirleme Problemi'ni (DCLSP) çözmek için uygulanmaktadır. Kapsamlı deneyler ve geleneksel yöntemlerle karşılaştırma sonuçları, RL tabanlı çerçevenin üstün performans sergilediğini, maliyetleri azalttığını ve üretim verimliliğini artırdığını göstermektedir. Bu araştırma, karmaşık üretim ortamlarında esnek karar verme süreçlerinin yeni bir çağını başlatırken, operasyon yönetimi alanında daha ileri çalışmalara kapı açmaktadır.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for providing financial support through the BIDEB scholarship program (BIDEB 2210-A Genel Yurtiçi Lisansüstü Burs Programı). This scholarship has enabled me to pursue my Master's degree and conduct this research.

I would also like to extend my thanks to Hepsiburada, one of the leading e-commerce companies in Turkey, for providing the necessary support for this study. Without their support, this research would not have been possible.

Most importantly, I would like to express my sincere gratitude to my wife, Amine Hümeyra Buldur, for her unwavering support and encouragement throughout the research and education phases of this thesis. Her love, patience, and understanding were invaluable to me during this journey.

Lastly, I would like to thank my supervisor Prof. Okan Örsan Özener and all the faculty members who have supported me throughout my studies. Their guidance and encouragement have been invaluable in my academic journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The overarching goal is to introduce the Dynamic Capacitated Lot Sizing Problem (DCLSP) and highlight the innovative approach of employing Reinforcement Learning (RL) techniques for its optimization.

The manufacturing and supply chain industries are constantly evolving, driven by globalization, changing consumer demands, and advancements in technology. Among the plethora of challenges faced by these industries, the Dynamic Capacitated Lot Sizing Problem stands out as a critical and complex optimization dilemma. At its core, the DCLSP involves determining the optimal production quantities over multiple periods while considering various factors such as limited production capacity, fluctuating demand patterns, inventory costs, and other real-world uncertainties.

In the post-pandemic era, satisfying customer demand within a given period has become crucial, especially considering the impatience of customers and the heightened competition among e-commerce merchants. Traditional optimization techniques have been utilized to address such challenges, but they often prove inadequate in handling the dynamic and unpredictable nature of modern supply chain environments. As customer demands speed up, and manufacturing systems become more intricate, the need for intelligent and adaptive decision-making strategies has intensified. The ability to swiftly respond to changing conditions has become a paramount requirement for success in the fast-paced and competitive e-commerce market.

This research delves into the promising realm of Reinforcement Learning (RL) as a groundbreaking solution to address the challenges posed by the post-pandemic era, where satisfying customer demand in a given period is crucial. With the impatience

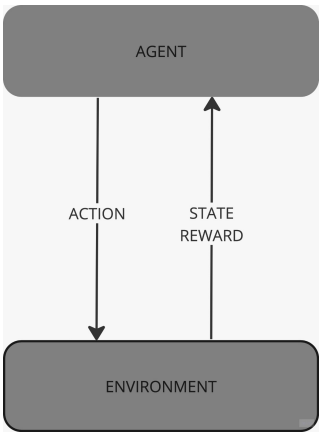

Figure 1: Reinforcement Learning Agent-Environment Interaction

of customers and the accelerated demand satisfaction, driven by the high competition among e-commerce merchants, the dynamic and unpredictable nature of modern supply chain environments has become even more complex.

RL, a subfield of machine learning, stands at the forefront of training agents to make decisions based on their interactions with the ever-changing environment. By learning from experience and feedback in the form of rewards, RL agents can develop optimal policies that adapt to the evolving circumstances.

The core premise of this thesis revolves around harnessing the capabilities of RL to develop an innovative framework for optimizing the DCLSP under the dynamic customer demand landscape, without allowing backordering and penalizing harshly in the shortage of supply. This approach enables the training of models that put a high emphasis on the timely satisfaction of demand, allowing businesses to effectively navigate the complexities of the highly competitive e-commerce market. By leveraging RL's adaptability and learning capabilities, the proposed framework empowers manufacturing systems to swiftly respond to changing customer demands, ultimately ensuring efficient inventory management and seamless production planning, while avoiding any disruptions to customer satisfaction. This approach is anticipated to revolutionize traditional lot sizing methods by bringing adaptability and intelligence

to the decision-making process. Instead of relying on predefined rules or static algorithms, the RL-based approach equips the manufacturing system with the ability to learn and improve its performance over time.

Through extensive experimentation and simulation with both synthetic and real-world datasets, this research seeks to demonstrate the effectiveness and superiority of the RL-based approach in comparison to conventional optimization methods. The proposed framework holds the potential to significantly enhance production efficiency, reduce operational costs, and most importantly, ensure the timely satisfaction of customer demands in each period. With the ability to adapt to the ever-changing dynamics of the highly competitive e-commerce market, the RL-based approach offers a strategic advantage in meeting the critical demand satisfaction requirements, enabling businesses to thrive in this fast-paced environment.

As we begin this research journey, the thesis aims to show how combining Reinforcement Learning and the Dynamic Capacitated Lot Sizing Problem can bring about significant changes in the post-pandemic customer demand landscape. By discovering new and valuable information, this study aims to improve manufacturing and supply chain practices, making them smarter, more flexible, and efficient. The ultimate goal is to achieve sustainable success in the dynamic e-commerce industry.

1.1 Background and Motivation

Manufacturing and supply chain operations have long been at the heart of global economic growth, enabling the seamless flow of goods and services across diverse industries. Efficient production planning and inventory management play a pivotal role in optimizing resources, reducing costs, and meeting customer demands. Among the key challenges in this domain, the Dynamic Capacitated Lot Sizing Problem (DCLSP) stands as a fundamental optimization conundrum that demands attention.

The DCLSP is a variant of the classic Capacitated Lot Sizing Problem (CLSP),

which involves determining the optimal production quantities for multiple products over a given planning horizon while respecting capacity constraints and minimizing total production costs. However, unlike the CLSP, the DCLSP deals with dynamic environments where demand patterns, production capacities, and other parameters fluctuate over time. As a result, traditional fixed-interval lot sizing methods become inadequate in effectively handling the inherent complexities of real-world manufacturing systems.

In traditional lot sizing approaches, the planning horizon is often divided into fixed intervals, and production quantities are determined based on static rules or heuristics. These methods lack adaptability, and their performance diminishes in the face of changing demand, unexpected disruptions, or varying production capacities. Consequently, the reliance on rigid lot sizing techniques can lead to suboptimal inventory levels, stockouts, overproduction, and increased operational costs.

Furthermore, the modern manufacturing and supply chain landscape has evolved into a dynamic and interconnected ecosystem, influenced by global market dynamics, rapid technological advancements, and ever-increasing consumer expectations. These complexities have rendered traditional optimization techniques insufficient to cope with the uncertainties and complexities that arise in today's business environment.

To address these challenges and seize the opportunities presented by the Fourth Industrial Revolution, there is a pressing need for intelligent and adaptive decision-making strategies. Herein lies the motivation for exploring Reinforcement Learning (RL) as a transformative approach to tackle the DCLSP.

Reinforcement Learning (RL) is a powerful subfield of machine learning that has garnered significant attention due to its ability to solve sequential decision-making problems [1]. At its core, RL revolves around training agents to interact with an environment, learn from the consequences of their actions, and make decisions to maximize cumulative rewards. This characteristic makes RL particularly suitable

for scenarios where the optimal course of action depends on the current state of the system and the feedback received from the environment.

One of the fundamental concepts in RL is the Markov Decision Process (MDP). MDPs provide a mathematical framework for modeling decision-making problems in which the system's state evolves stochastically over time, and the agent's actions influence the transition between states. The Markov property ensures that the future state of the system depends solely on the current state and the action taken, making MDPs a key tool for formulating RL problems.

Within RL, there are two primary approaches to learning optimal policies: policy-based RL and value function-based RL [2]. In policy-based RL, the agent directly learns a policy, which is a mapping from states to actions, without explicitly estimating the value of each state. The policy is iteratively updated through methods like Policy Gradient algorithms, where the agent moves in the direction that improves the expected cumulative reward.

On the other hand, value function-based RL aims to estimate the value of each state or state-action pair, representing the expected cumulative reward the agent can achieve from that state onward. Two popular methods in value function-based RL are Q-Learning and SARSA (State-Action-Reward-State-Action). These algorithms iteratively update the value function to converge to the optimal values.

Another important distinction in RL is between model-free and model-based methods [3]. Model-free RL does not rely on having a full model of the environment, including transition probabilities and rewards. Instead, the agent learns from its interactions with the environment, through trial and error, and updates its policy or value function accordingly. This approach is more robust and applicable to real-world scenarios with complex and uncertain dynamics.

In contrast, model-based RL involves learning a model of the environment and

then using this model to plan and make decisions. While potentially more sample-efficient, model-based RL can be challenging due to the need to accurately model the environment, which may not always be feasible in practice.

Reinforcement Learning has seen significant advancements in recent years, with applications ranging from robotics and game playing to recommendation systems and autonomous vehicles. Its adaptability and ability to learn from experience make it a promising tool for addressing complex decision-making problems in various domains.

The inherent adaptability and learning capabilities of RL have found success in diverse fields, from robotics and gaming to finance and healthcare[4]. By leveraging RL in the context of the DCLSP, manufacturing and supply chain decision-makers can empower their systems to dynamically adjust production quantities based on real-time feedback and changing conditions [5]. RL offers the promise of optimizing production schedules, minimizing costs, reducing lead times, and enhancing overall operational efficiency.

As manufacturing systems become increasingly intricate and competitive, incorporating intelligent decision-making processes through RL becomes crucial to gain a competitive edge. Thus, this research endeavors to investigate the integration of RL techniques into the DCLSP, with the aspiration of revolutionizing the way manufacturing and supply chain operations are optimized in a dynamic and uncertain world. By exploring the potential synergies between RL and the DCLSP, this thesis aims to contribute valuable insights to the field of operations management and pave the way for more intelligent, agile, and resilient manufacturing practices.

1.2 Research Question and Objectives

The Dynamic Capacitated Lot Sizing Problem (DCLSP) is a complex and critical optimization challenge that arises in manufacturing and supply chain management.

At its core, the DCLSP aims to determine the optimal production quantities of multiple products over a finite planning horizon, considering varying demand patterns, capacity limitations, and inventory costs. The overarching objective is to minimize the total production and inventory holding costs while fulfilling customer demands and adhering to capacity constraints[6].

Formally, the DCLSP can be represented as follows:

Given:

- A set of products $P = \{1, 2, ..., N\}$ to be produced and sold over T periods.

- Time periods $t = \{1, 2, ..., T\}$ comprising the planning horizon.

- Demand d_{pt} representing the quantity of product p demanded in period t.

- Production capacity C_t for each period t, specifying the maximum units that can be produced.

- Inventory holding cost h_{pt} for product p during period t.

- Production cost c_{pt} for manufacturing product p in period t.

The primary challenge of the DCLSP lies in determining the production quantities q_{pt} of each product p in each period t, subject to the following constraints:

1. Capacity Constraint: The total production quantity of all products in any period t must not exceed the available production capacity C_t.

$$\sum_{p=1}^{N} q_{pt} \leq C_t, \quad \forall t \in \{1, 2, ..., T\}$$

2. Demand Fulfillment: The total production quantity of each product p in each period t should be equal to or larger than the demand d_{pt} in that period, taking into account the inventory from the previous period. Thus, the demand fulfillment constraint can be expressed as follows:

$$q_{pt} + I_{pt} \geq d_{pt}, \quad \forall t \in T$$

Where q_{pt} represents the production quantity of product p in period t, I_{pt} denotes

the inventory of product p from previous period of the planning horizon, and d_{pt} is the demand for product p in period t. This constraint ensures that the total production and available inventory are sufficient to meet or exceed the demand for each product in each period.

3. Non-Negativity Constraint: Production quantities cannot be negative.

$$q_{pt} \geq 0, \quad \forall p \in P, \forall t \in \{1, 2, ..., T\}$$

The DCLSP exhibits inherent complexities due to the dynamic nature of the problem. Demand patterns for products may vary across different time periods, and production capacities can change over time due to factors such as machine breakdowns, labor shortages, or seasonal variations. Moreover, the costs associated with production and inventory holding can fluctuate, making it challenging to find a fixed solution that remains optimal over the entire planning horizon.

Traditional optimization techniques, such as Linear Programming (LP) or Integer Programming (IP), may provide feasible solutions for small instances of the DCLSP. However, as the size of the problem and the planning horizon increase, these methods often encounter computational intractability, limiting their applicability to real-world manufacturing scenarios[7].

To overcome the limitations of traditional approaches and effectively address the dynamic and uncertain nature of the DCLSP, this research proposes the application of Reinforcement Learning (RL) techniques. By employing RL, the aim is to develop an adaptive and intelligent framework that can learn optimal lot-sizing policies over time, effectively navigating the complexities of capacity planning, demand variations, and inventory management.

This thesis endeavors to design and implement an RL-based solution to the DCLSP, experimentally evaluate its performance, and compare it with traditional optimization methods. The ultimate goal is to revolutionize the field of manufacturing and supply chain optimization by introducing a dynamic and agile approach that enhances

production efficiency, reduces costs, and improves decision-making in a constantly evolving operational environment.

1.3 Significance of The Study

The Dynamic Capacitated Lot Sizing Problem (DCLSP) poses significant challenges to manufacturing and supply chain management, especially in the context of today's dynamic and competitive business environment. In light of these challenges, the application of Reinforcement Learning (RL) techniques to address the DCLSP holds immense potential and offers several noteworthy contributions to the field of operations management [8]. This section highlights the key significance of the study and emphasizes the benefits that the RL-based optimization approach can bring to industry practices.

Enhanced Decision-Making and Adaptability: One of the primary advantages of employing RL in the DCLSP is its ability to enable more intelligent and adaptive decision-making. Traditional lot-sizing approaches often rely on pre-defined heuristics or fixed rules, which might not be optimal for changing conditions. In contrast, RL agents learn from their interactions with the environment and continually update their production policies based on real-time feedback[9]. This adaptability allows manufacturing systems to respond promptly to fluctuations in demand, capacity constraints, and market dynamics, ultimately leading to improved production efficiency and customer satisfaction.

Handling Complexity and Uncertainty: Modern manufacturing and supply chain operations face unprecedented levels of complexity and uncertainty. The DCLSP involves numerous interrelated factors, including demand variations, production capacities, lead times, and inventory costs. RL, with its capacity to learn from experience and explore multiple solutions, is better equipped to handle these complexities effectively. By optimizing production decisions over a dynamic planning horizon, the

RL-based approach can mitigate the impact of uncertainties, reduce operational risks, and minimize potential disruptions in the supply chain [10].

Scalability and Performance: As the scale of manufacturing systems and the planning horizon grow, traditional optimization methods often encounter computational challenges, leading to prohibitive processing times. In contrast, RL techniques have shown promise in handling large-scale problems efficiently. By leveraging advancements in deep reinforcement learning, such as Deep Q-Networks (DQN) or Proximal Policy Optimization (PPO), the proposed RL-based framework can scale effectively, making it a practical and scalable solution for real-world manufacturing environments [11].

Cost Reduction and Resource Optimization: Optimizing production quantities and inventory levels is directly linked to cost reduction in manufacturing and supply chain operations. The RL-based approach aims to minimize production costs, inventory holding costs, and stockouts, leading to improved cost-efficiency across the entire supply chain. By identifying and learning from optimal lot-sizing policies, the manufacturing system can effectively utilize resources, reduce waste, and enhance overall productivity.

Advancement of Operations Management: This study's contribution extends beyond the realm of the DCLSP, as the application of RL techniques in operations management presents a novel and promising approach for addressing various other complex problems in manufacturing and supply chain domains. The insights gained from this research can inspire further exploration of RL in diverse areas such as inventory control, production scheduling, and resource allocation, leading to the advancement of the entire field of operations management.

In conclusion, this research holds substantial significance in advancing manufacturing and supply chain practices through the integration of Reinforcement Learning

into the Dynamic Capacitated Lot Sizing Problem. By emphasizing the transformative potential of RL in handling complexity, uncertainty, and scalability, the proposed framework aims to equip decision-makers with intelligent and adaptive tools that can optimize production strategies, reduce costs, and improve operational efficiency. The outcomes of this study have the potential to revolutionize industry practices, making them more agile, competitive, and responsive to ever-changing market demands and challenges.

1.4 Scope & Limitations

The success of any research project lies in clearly defining its scope and acknowledging the limitations it might encounter. In this section, we outline the boundaries within which the RL-based optimization approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP) will be applied, along with a candid discussion of the potential constraints and limitations of the proposed methodology.

Scope of the Research:

- DCLSP with Multiple Products: The research focuses on the DCLSP involving multiple products, each with distinct production costs, demand patterns, and inventory holding costs. By addressing a multi-product setting, the RL-based approach aims to optimize production quantities for each product over the finite planning horizon.

- Time-Divided Planning Horizon: The RL-based optimization framework will be designed to accommodate a time-divided planning horizon comprising T periods. This finite time horizon is representative of real-world production settings, where decisions need to be made sequentially over a specific time frame.

- Stochastic Demand and Capacity: The RL-based approach will consider the impact of stochastic demand, introducing uncertainty into the production planning process. This enables the model to adapt to unpredictable demand patterns, which

is a common challenge in real-world manufacturing environments.

- RL Algorithms: The research will explore various RL algorithms, such as Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), or Actor-Critic, to learn optimal lot-sizing policies. The choice of RL algorithms will be based on their suitability for addressing the sequential decision-making nature of the DCLSP.

Limitations and Constraints:

- Computational Complexity: RL techniques, particularly deep reinforcement learning methods, can be computationally intensive, especially when applied to large-scale DCLSP instances. The computational resources required for training the RL models and optimizing production quantities might pose a limitation, particularly for industries with limited computing capabilities.

- Real-time Constraints: While RL offers adaptability, the training process might require significant time and historical data to converge to an optimal policy. Real-time decision-making might be challenging in certain scenarios, especially when immediate responses are required to rapidly changing conditions.

- Data Availability: The effectiveness of RL heavily relies on the availability of relevant and representative data for training the RL agents. In some cases, obtaining high-quality historical data on demand patterns, capacity fluctuations, and production costs may be difficult or restricted, potentially affecting the model's performance.

- Generalization to Novel Scenarios: The RL-based approach will be evaluated based on its performance in simulated and possibly real-world datasets. However, its generalization to entirely new and unforeseen scenarios might be limited, and the model's robustness to unencountered conditions requires careful examination.

- Hyperparameter Tuning: The performance of RL algorithms is sensitive to hyperparameter settings, and finding the optimal configuration can be a non-trivial task. Adequate experimentation and tuning of hyperparameters are crucial to ensure the RL-based framework's effectiveness[12].

- Manual Intervention: Depending on the complexity of the RL-based model and the quality of the data, manual intervention may be required in some cases to fine-tune or validate the results, potentially limiting the model's fully autonomous decision-making capability.

Despite these limitations, the proposed RL-based approach to the DCLSP holds substantial promise in addressing dynamic production challenges and revolutionizing decision-making in manufacturing and supply chain operations. By acknowledging these constraints and emphasizing the scope of the research, this study aims to provide valuable insights into the applicability and potential benefits of RL in optimizing complex and uncertain manufacturing systems.

1.5 Thesis Structure

This section provides a clear and concise outline of the thesis structure, offering readers a roadmap to navigate through the document. The subsequent chapters are carefully crafted to address specific aspects of the research, building upon each other to provide a comprehensive analysis of the Dynamic Capacitated Lot Sizing Problem (DCLSP) optimization using Reinforcement Learning (RL). The chapter descriptions below guide the reader through the logical flow of the thesis:

Chapter 2: Literature Review Chapter 2 delves into an extensive literature review, exploring existing studies and research on the DCLSP and optimization techniques. It examines traditional lot sizing methods, their limitations, and challenges faced in dynamic manufacturing environments. Additionally, the chapter investigates relevant literature on Reinforcement Learning, focusing on its application in operations management and similar decision-making problems. By identifying gaps in the literature, this chapter lays the foundation for the proposed RL-based approach to the DCLSP.

Chapter 3: Methodology Chapter 3 meticulously describes the methodology employed in developing the RL-based framework for the DCLSP. It outlines the data

collection process, sources of data, and data preprocessing steps necessary to prepare the information for RL training. The chapter explains the RL algorithms chosen for experimentation, justifying their selection based on their suitability for sequential decision-making problems. Moreover, it discusses the simulation setup and the environment in which the RL agent will interact to optimize lot-sizing policies.

Chapter 4: Results and Analysis Chapter 4 presents the empirical results obtained from the application of the RL-based approach to the DCLSP. The chapter showcases the model's performance, comparing it with traditional optimization methods and evaluating its effectiveness in different scenarios. The analysis dissects the impact of RL on production efficiency, costs, and adaptability to changing conditions. Moreover, it examines how the RL agent learns and adapts its decision-making policies over time.

Chapter 5: Discussion and Implications Chapter 5 engages in an in-depth discussion of the results and their implications for manufacturing and supply chain management. It explores the practical implications of the RL-based approach, emphasizing its potential for transforming lot sizing decisions in dynamic manufacturing environments. The chapter discusses the model's strengths and limitations, shedding light on the practical challenges of implementing RL in real-world scenarios. Furthermore, it highlights the broader implications of RL application in other operations management challenges.

Chapter 6: Conclusion The final chapter serves as a comprehensive conclusion to the thesis. It summarizes the key findings and contributions of the research, reinforcing the significance of the RL-based approach to the DCLSP. The chapter revisits the research objectives and answers the research questions based on the results and analysis presented in earlier chapters. Moreover, it reflects on the broader implications of the study for manufacturing and supply chain management, and proposes future directions for research in this area.

In summary, this thesis is carefully organized to offer a comprehensive exploration

of the DCLSP optimization using Reinforcement Learning. It proceeds logically, from introducing the problem and its significance to discussing the research methodology, presenting results, and drawing meaningful conclusions. The reader is guided through a structured journey that aims to enlighten, challenge, and inspire further advancements in the field of operations management.

CHAPTER II

LITERATURE REVIEW

The literature review in this chapter serves as a comprehensive exploration of existing research and studies concerning three crucial aspects: the Dynamic Capacitated Lot Sizing Problem (DCLSP), traditional lot sizing methods, and the application of Reinforcement Learning (RL) in operations management. The DCLSP, as a fundamental optimization challenge in manufacturing and supply chain management, demands careful analysis and innovative solutions to address its complexities. Traditional lot sizing methods have long been utilized in production planning [13], but their limitations in dynamic and uncertain manufacturing environments necessitate a deeper examination of alternative approaches. This leads us to explore the potential of Reinforcement Learning, a powerful paradigm in machine learning, to revolutionize decision-making in operations management[14].

By critically reviewing and analyzing the relevant literature, this section seeks to offer valuable insights and a comprehensive context for the development and integration of an RL-based approach to optimize the DCLSP. The objective is to build upon the collective knowledge and experiences of prior researchers, identifying strengths, limitations, and gaps in existing approaches. This literature review serves as a solid foundation, guiding the subsequent methodology and model development in a well-informed and methodical manner.

The examination of studies related to the DCLSP provides an understanding of the intricacies and challenges faced by traditional lot sizing methods[15]. These challenges often arise due to fluctuating demand patterns, varying production capacities, and dynamic inventory holding costs. By exploring the shortcomings of traditional

methods, we can better appreciate the need for alternative optimization techniques that can adapt and thrive in a dynamic operational environment[16] [17].

Moreover, this literature review delves into the concept of Reinforcement Learning, offering insights into its principles, methodologies, and applications in diverse domains. Understanding how RL agents interact with environments, learn from experiences, and optimize decision-making over time is crucial in envisioning its potential utility in manufacturing and supply chain management.

The exploration of relevant studies that have applied RL to dynamic lot sizing problems illuminates the successes and challenges faced in similar scenarios. Analyzing the performance of RL algorithms, such as Deep Q-Networks (DQN) [18] or Proximal Policy Optimization (PPO) [19], provides essential information on how RL agents can learn optimal policies under dynamic conditions. This knowledge is essential for tailoring RL methods to address the unique complexities of the DCLSP and develop a well-suited RL-based framework[20].

By synthesizing and summarizing the collective findings from the reviewed literature, this section aims to identify the research gap that motivates the current study. It underlines the significance of exploring an RL-based approach for the DCLSP and highlights the potential contributions this research can make to the field of operations management. Armed with a strong understanding of prior work, the subsequent chapters will delve into the methodology, empirical results, and implications of applying RL to optimize the DCLSP. The ultimate goal is to equip decision-makers with intelligent and adaptive tools that enhance production efficiency, reduce costs, and improve operational performance in dynamic and uncertain manufacturing environments[21].

2.1 Traditional Lot Sizing Methods

In manufacturing and supply chain management, traditional lot sizing methods have played a pivotal role in determining the optimal production quantities of items within

a given planning horizon. Three prominent traditional approaches are the Economic Order Quantity (EOQ) model, the Wagner-Whitin algorithm, and Dynamic Programming methods. These methods have been extensively utilized due to their simplicity, analytical tractability, and effectiveness in static and deterministic environments.

The Economic Order Quantity (EOQ) model aims to strike a balance between inventory holding costs and ordering costs. It calculates the ideal order quantity that minimizes the total cost of holding inventory and the cost of ordering new batches. While this model provides straightforward and practical solutions for stable demand and fixed production capacities, it falls short in addressing the complexities of the DCLSP.

The Wagner-Whitin algorithm is designed for multi-period inventory management with constant demand and production capacities. It aims to minimize the cumulative cost of inventory and production over the planning horizon by optimally timing production runs. Although it can handle multi-period planning, it assumes constant demand and capacity, limiting its applicability in real-world scenarios characterized by fluctuating conditions.

Dynamic Programming methods, such as the Silver-Meal and the Least Unit Cost algorithms, optimize the cost of production and inventory over multiple periods by breaking the planning horizon into smaller subproblems. These approaches offer more flexibility than the EOQ model and the Wagner-Whitin algorithm in addressing time-varying demand and production capacities. However, as the planning horizon grows larger, dynamic programming methods encounter computational challenges due to the exponential increase in the number of subproblems.

While these traditional lot sizing methods have been valuable in simpler and more stable manufacturing environments, they face significant challenges when applied to the dynamic and uncertain nature of the DCLSP. Manufacturing and supply chain operations today are subject to ever-changing market demands, production capacities,

and inventory costs. Traditional methods lack the adaptability and responsiveness required to optimize lot sizing decisions under these dynamic conditions.

The inability of traditional lot sizing methods to adapt to varying demand patterns, changing production capacities, and fluctuating inventory costs is a critical limitation. The DCLSP requires robust and agile decision-making to respond to unforeseen events, minimize stockouts, and optimize production and inventory costs over time. Failing to address these challenges can result in suboptimal lot sizing decisions, leading to excess inventory holding costs, stockouts, and ultimately, a decrease in overall operational efficiency.

In response to these limitations, the exploration of alternative optimization approaches becomes imperative. This has led to the exploration of Reinforcement Learning (RL), a powerful paradigm in machine learning that holds the potential to revolutionize decision-making in dynamic manufacturing environments. By leveraging RL techniques, it becomes possible to develop intelligent and adaptive lot sizing policies that can effectively navigate the complexities of the DCLSP and enhance the overall performance of manufacturing and supply chain operations. The subsequent chapters will delve into the application of RL in addressing the DCLSP, offering insights into its effectiveness and practical implications for modern operations management.

2.2 Challenges in Dynamic Capacitated Lot Sizing

The Dynamic Capacitated Lot Sizing Problem (DCLSP) presents a unique set of challenges that traditional lot sizing methods struggle to effectively address. These challenges arise from the dynamic and uncertain nature of modern manufacturing and supply chain environments. As a result, traditional fixed-interval lot sizing rules face difficulties in coping with the constantly changing conditions, leading to suboptimal inventory management and production planning[22].

One of the primary challenges posed by the DCLSP is the presence of dynamic

and unpredictable demand patterns. In real-world manufacturing scenarios, demand for products can fluctuate significantly over time due to factors such as seasonality, changing consumer preferences, marketing campaigns, or external economic influences. Traditional lot sizing methods, which rely on fixed-interval or deterministic demand assumptions, often fail to account for these variations[23]. As a result, they may lead to overstocking or understocking issues, leading to excess holding costs for surplus inventory or stockouts that impact customer satisfaction and sales revenue.

Furthermore, varying capacity constraints in dynamic manufacturing environments introduce additional complexities. Production capacities can fluctuate due to factors like machine breakdowns, maintenance activities, labor shortages, or production line adjustments. Such changes in capacity can significantly impact the production planning process, as they affect the ability to meet demand requirements within specified time frames. Traditional lot sizing methods struggle to adapt to these varying capacities, leading to suboptimal production plans that may not fully utilize available resources or may result in insufficient production to meet demand [24].

Moreover, the interaction of dynamic demand patterns and varying capacity constraints further complicates the decision-making process. As demand fluctuates and capacity changes, the optimal production quantity and timing need to be continuously adjusted to maintain an efficient and responsive production system. The traditional lot sizing methods, which rely on predefined fixed intervals, are ill-equipped to handle such dynamic decision-making requirements.

These challenges highlight the need for innovative approaches that can adapt and optimize lot sizing decisions in real-time, considering the dynamic nature of the DCLSP. The exploration of alternative methods that can account for varying demand, fluctuating capacities, and changing market conditions becomes crucial to address these complexities effectively.

Reinforcement Learning (RL) emerges as a promising solution to tackle the challenges of the DCLSP. RL agents learn from their interactions with the dynamic manufacturing environment, allowing them to adapt their decision-making policies over time [2]. This adaptability enables RL-based approaches to optimize lot sizing decisions in response to changing demand patterns and varying capacities, offering a more agile and responsive solution to the DCLSP.

By exploring innovative approaches like RL, this research aims to overcome the challenges posed by the DCLSP and contribute to the development of intelligent and adaptive lot sizing strategies. The subsequent chapters will delve into the application of RL techniques to the DCLSP, presenting empirical results and insights on how RL can effectively optimize production planning and inventory management in dynamic manufacturing environments.

2.3 Reinforcement Learning in Operations Management

Reinforcement Learning (RL) has garnered significant attention as a powerful and versatile paradigm for addressing complex decision-making challenges in operations management. RL is a branch of machine learning where an agent learns to make optimal decisions by interacting with an environment and receiving feedback in the form of rewards or penalties [2]. It differs from supervised learning, where the agent is provided with labeled data[25], and unsupervised learning, where the agent seeks patterns in unlabeled data. Instead, RL agents learn through trial and error, refining their decision-making policies based on the consequences of their actions.

At the core of RL lies the concept of an agent[2], which is an autonomous entity that operates within an environment. The environment represents the context in which the agent exists and interacts. The agent observes the current state of the environment, takes actions based on its learned policy, and receives feedback from the environment in the form of rewards or penalties [26]. The objective of the RL

agent is to maximize the cumulative rewards it receives over time, striving to make decisions that lead to desirable outcomes.

The sequential nature of RL is particularly well-suited for decision-making problems in operations management, where actions have long-term consequences and must be optimized over time. RL agents can learn to make dynamic and adaptive decisions, considering the evolving state of the environment and the feedback they receive [27].

RL has demonstrated remarkable success in various domains, showcasing its capability to solve complex problems with practical significance. In robotics, RL has been applied to teach robotic agents to navigate unknown environments, manipulate objects, and perform complex tasks [28]. In gaming, RL agents have achieved superhuman performance in games like Chess, Go, and Atari games [18] [9]. In finance, RL has been used for portfolio optimization and algorithmic trading [29] [30]. In healthcare, RL has shown potential in personalized treatment recommendations and optimizing hospital operations [31].

The application of RL in operations management has opened up new avenues for optimizing decision-making processes. In inventory management, RL agents can learn to adaptively adjust reorder points and order quantities based on demand patterns and stock levels. In production scheduling, RL agents can optimize job sequencing and resource allocation to minimize production lead times and maximize resource utilization [32]. In supply chain management, RL can help optimize logistics and distribution decisions to reduce costs and improve delivery efficiency [33].

The potential of RL in operations management is evident in its ability to handle dynamic and uncertain environments [2], where traditional optimization methods may fall short. By learning from interactions and continuously improving decision-making policies, RL agents can adapt to changing conditions and optimize operational performance over time.

In the context of this thesis, RL presents a compelling approach to tackle the Dynamic Capacitated Lot Sizing Problem (DCLSP). By integrating RL techniques into the lot sizing decision process, the model can learn to make optimal production quantity decisions that adapt to varying demand patterns and capacity constraints. The subsequent chapters will explore the application of RL to the DCLSP, offering insights into the effectiveness of RL-based solutions in addressing the challenges of dynamic manufacturing environments and advancing the field of operations management.

2.4 Relevant Studies on RL in Dynamic Lot Sizing

Recent research has witnessed a growing interest in the application of Reinforcement Learning (RL) techniques to address the dynamic lot sizing problem in manufacturing and supply chain management. These studies have explored the potential of RL to optimize lot sizing decisions in the face of varying demand patterns, fluctuating production capacities, and evolving inventory costs. By leveraging RL algorithms, these approaches aim to develop intelligent and adaptive lot sizing policies that can outperform traditional methods in dynamic and uncertain manufacturing environments.

Several RL algorithms have been utilized in these studies to tackle the dynamic lot sizing problem effectively:

- Deep Q-Networks (DQN) [18]: DQN is a popular RL algorithm that employs deep neural networks to approximate action-value functions. In the context of dynamic lot sizing, DQN has been used to learn the optimal lot sizing policy by estimating the expected total rewards associated with different production quantity decisions at each time step. By incorporating deep neural networks, DQN can handle high-dimensional state spaces and learn complex patterns in demand and production dynamics, making it suitable for real-world manufacturing scenarios.

- Proximal Policy Optimization (PPO)[19]: PPO is a policy-based RL algorithm that directly optimizes the policy function, which maps states to actions, without

explicitly estimating value functions. In dynamic lot sizing, PPO can adaptively learn policies that maximize cumulative rewards over time by exploring and exploiting the environment. PPO's advantage lies in its ability to handle continuous action spaces, making it a valuable tool for optimizing production quantity decisions in dynamic manufacturing environments.

- Actor-Critic Methods[34]: Actor-Critic methods combine elements of both policy-based and value-based RL. The actor network is responsible for selecting actions based on the current state, while the critic network estimates the value function to provide feedback on the actor's decisions. This combination allows for efficient learning and improved stability in RL training. In dynamic lot sizing scenarios, actor-critic methods have demonstrated the ability to find near-optimal policies by efficiently exploring the action space and incorporating value estimates to improve decision-making.

Comparing RL-based approaches to traditional lot sizing methods reveals some key advantages and limitations of RL in dynamic lot sizing scenarios. RL algorithms have the inherent ability to adapt and learn from experience, making them well-suited for handling changing demand patterns and production capacities [35]. Traditional lot sizing methods, on the other hand, often rely on fixed assumptions and heuristics, which can lead to suboptimal solutions in dynamic manufacturing environments.

RL-based approaches have demonstrated the potential to achieve higher levels of efficiency and cost-effectiveness compared to traditional methods [11]. By continuously optimizing lot sizing decisions based on feedback, RL algorithms can achieve better utilization of resources, reduced inventory holding costs, and improved customer service levels.

However, RL approaches also face certain challenges, such as the need for significant computational resources and data to train the models effectively. RL training can be computationally intensive, especially when dealing with large state and action spaces. Moreover, obtaining sufficient data to accurately represent the dynamics of

a complex manufacturing environment may be a practical challenge. Additionally, the performance of RL algorithms is sensitive to the choice of hyperparameters and training configurations, necessitating careful tuning to achieve optimal results.

Despite these challenges, the findings from relevant studies highlight the promise and potential of RL in dynamic lot sizing scenarios [36] [20]. By leveraging the strengths of RL algorithms and mitigating their limitations, RL-based approaches offer a novel and powerful avenue for optimizing production planning and inventory management in modern manufacturing and supply chain operations. The subsequent chapters of this thesis will build upon these insights, presenting empirical results and analysis of an RL-based approach to the Dynamic Capacitated Lot Sizing Problem, showcasing its effectiveness and contributions to the field of operations management.

2.5 Integration of RL in the DCLSP

The integration of Reinforcement Learning (RL) in the Dynamic Capacitated Lot Sizing Problem (DCLSP) presents an exciting opportunity to revolutionize production planning and inventory management in modern manufacturing and supply chain operations. RL's core strengths, such as adaptability and learning capabilities, are well-aligned with the dynamic and uncertain nature of the DCLSP, making it a promising approach to tackle the challenges posed by this complex optimization problem.

One of the primary advantages of RL in the DCLSP is its adaptability. RL agents can continuously learn and adjust their decision-making policies based on the feedback received from the manufacturing environment. As the demand patterns, production capacities, and inventory costs change over time, the RL agent can dynamically adapt its lot sizing decisions to optimize overall performance. This adaptability allows manufacturing systems to respond effectively to changing conditions, ensuring that production plans remain efficient and cost-effective in the face of uncertainty.

Moreover, RL's learning capabilities enable it to discover optimal or near-optimal

strategies through trial and error. RL agents explore different production quantity decisions and learn from the rewards or penalties they receive from the environment. Over time, the RL agent refines its decision-making policies to make more informed and intelligent choices, leading to improved inventory management and production planning.

However, applying RL to the DCLSP also introduces certain challenges that need to be carefully addressed. One critical factor is data availability. RL algorithms typically require large amounts of data to effectively learn optimal policies. In the context of the DCLSP, obtaining historical data on demand patterns, production capacities, and inventory levels might be challenging, especially for businesses that are new or operate in rapidly changing markets. Careful consideration of data collection and preprocessing strategies is crucial to ensure that the RL agent can learn meaningful insights from the available data.

Another challenge is the computational complexity of RL training. Depending on the complexity of the manufacturing environment and the choice of RL algorithm, training an RL agent may be computationally intensive and time-consuming. Efficient implementation and utilization of computational resources become essential to avoid excessive training times and resource consumption.

Furthermore, designing appropriate reward structures is a crucial aspect of RL in the DCLSP. The reward function defines the objective that the RL agent aims to maximize. In lot sizing scenarios, defining an appropriate reward function that accurately reflects the performance metrics of interest, such as minimizing holding costs, avoiding stockouts, or optimizing production efficiency, is essential to guide the RL agent's learning process effectively.

To harness the full potential of RL in the DCLSP, careful consideration and experimentation are required. Selecting the most suitable RL algorithm, tuning hyperparameters, and optimizing the reward function are iterative processes that demand

empirical validation and refinement. Additionally, conducting simulations and sensitivity analyses can help gain deeper insights into the performance of the RL-based approach under various conditions, further guiding the decision-making process.

Ultimately, successful integration of RL in the DCLSP has the potential to transform traditional production planning and inventory management practices. By leveraging RL's adaptability and learning capabilities, manufacturing systems can become more agile and responsive, leading to improved operational efficiency, cost-effectiveness, and customer satisfaction. The subsequent chapters of this thesis will delve into the practical implementation and empirical evaluation of the RL-based approach to the DCLSP, providing valuable insights into its effectiveness and contributions to the field of operations management.

2.6 Summary and Research Gap

The literature review conducted in this chapter has offered valuable insights into three key areas: traditional lot sizing methods, challenges faced in the Dynamic Capacitated Lot Sizing Problem (DCLSP), and the potential application of Reinforcement Learning (RL) in operations management. The examination of traditional lot sizing methods, including the Economic Order Quantity (EOQ) model [37], the Wagner-Whitin algorithm [38], and Dynamic Programming approaches, has revealed their limitations in addressing the complexities of dynamic manufacturing environments. These traditional methods, although widely used and practical for stable conditions, lack the adaptability and responsiveness required to optimize lot sizing decisions in the face of unpredictable demand patterns, fluctuating production capacities, and changing inventory costs.

The exploration of challenges in the DCLSP has further emphasized the need for innovative approaches to overcome the limitations of traditional methods. The dynamic and uncertain nature of modern manufacturing and supply chain operations

demands decision-making strategies that can adapt and respond effectively to changing conditions. Failure to address these challenges can lead to suboptimal inventory management, increased holding costs, stockouts, and reduced operational efficiency.

In response to these challenges, the literature review has indicated that RL-based approaches hold significant promise for optimizing lot sizing decisions in dynamic manufacturing environments. RL's adaptability and learning capabilities make it well-suited for addressing the complexities of the DCLSP. RL agents can learn from their interactions with the manufacturing environment, continuously improving their decision-making policies to adapt to changing conditions and optimize long-term performance.

While RL has demonstrated success in similar problem domains, such as robotics, gaming, finance, and healthcare, its application in the DCLSP remains relatively unexplored. Despite the promising potential of RL-based approaches in manufacturing and supply chain management, a comprehensive study that fully integrates RL into the DCLSP is yet to be conducted. Existing research has laid the groundwork for RL in dynamic lot sizing scenarios, but there is a research gap concerning the development and evaluation of an RL-based framework explicitly tailored to the unique challenges of the DCLSP.

This thesis aims to fill this research gap by conducting an in-depth investigation into the integration of RL techniques in the DCLSP. By designing, implementing, and empirically evaluating an RL-based approach, this research seeks to demonstrate the effectiveness of RL in optimizing lot sizing decisions under dynamic conditions. The thesis will contribute to the existing literature on operations management and supply chain optimization by providing valuable insights into the performance and practical implications of RL-based solutions for the DCLSP.

Ultimately, this research endeavor aims to advance the field of operations management by showcasing the potential of RL in dynamic manufacturing environments. By

bridging the gap between traditional lot sizing methods and RL-based approaches, this thesis aims to contribute to more efficient and adaptive decision-making strategies that enhance production planning, inventory management, and overall operational performance. The subsequent chapters will present the methodology, empirical results, and implications of applying RL to the DCLSP, shedding light on its effectiveness and contributions to the field of operations management.

2.7 Conclusion

In conclusion, the literature review has laid the groundwork for the proposed RL-based approach to the Dynamic Capacitated Lot Sizing Problem. By reviewing existing research, identifying challenges, and exploring the potential of RL in similar scenarios, this chapter establishes the significance of this research endeavor. The subsequent chapters will delve into the methodology, results, and implications of applying RL to the DCLSP, contributing valuable insights to the field of operations management and supply chain optimization.

CHAPTER III

METHODOLOGY

3.1 Problem Formulation

The problem addressed in this research is the Dynamic Capacitated Lot Sizing Problem (DCLSP) in the context of manufacturing and supply chain management. The DCLSP involves determining the optimal production quantities for items over a finite planning horizon, subject to various constraints and in response to changing demand patterns, production capacities, and inventory costs.

3.1.1 Problem Statement

The main objective of the DCLSP is to minimize the total production and inventory costs while satisfying customer demand and respecting production and inventory constraints. The goal is to develop an intelligent and adaptive lot sizing policy that optimizes production decisions over time, ensuring efficient resource utilization and responsive inventory management.

3.1.2 Constraints

The DCLSP is subject to several constraints that must be considered during the lot sizing decision process:

- Capacity Constraints: The production capacity of the manufacturing system restricts the maximum quantity of items that can be produced in each time period. Exceeding the production capacity is not allowed.

- Demand Constraints: The demand for products fluctuates over time, and the production plan must meet customer orders in each period.

- Inventory Constraints: The inventory levels are bounded by minimum and maximum levels, ensuring that excessive or insufficient inventory levels are avoided.

- Production Setup Constraints: The DCLSP considers setup or changeover costs associated with starting production or switching between different products. These setup costs influence the decision-making process for production quantity determination.

3.1.3 Decision Variables

The decision variables in the DCLSP are the production quantities to be set for each item in each time period. Let x_{it} represent the production quantity for item i in time period t. The objective is to optimize these decision variables over the planning horizon.

3.1.4 Objective Function

In this study, we extend the objective function of the DCLSP to explicitly incentivize demand satisfaction and penalize shortages in supply to meet customer demand. The modified objective function is designed to minimize the total production and inventory costs over the entire planning horizon, considering production costs, inventory holding costs, and setup costs, while putting a high emphasis on fulfilling customer demand in a timely manner. By incorporating these additional reward components, the RL agent is encouraged to optimize the production plan with a strong focus on meeting customer demands and maintaining adequate inventory levels. This modification enables the RL-based approach to adaptively adjust its decision-making policies to respond effectively to dynamic demand patterns, capacity constraints, and market fluctuations, leading to improved production efficiency and customer satisfaction

3.1.5 Environment Dynamics

The manufacturing environment is dynamic, characterized by varying demand patterns, production capacities, and inventory costs. Demand can fluctuate due to seasonal effects, market trends, or external factors. Production capacities may change due to machine breakdowns, maintenance, or operational adjustments. Inventory costs might vary due to changes in storage or holding charges. The RL agent interacts with this dynamic environment, continuously updating its lot sizing policy based on feedback to make adaptive and optimal decisions over time.

By clearly formulating the DCLSP problem and identifying its key components, this research establishes a solid framework for developing and implementing the RL-based optimization approach. The RL agent's learning process is tailored to address the dynamic and uncertain nature of the DCLSP, offering a novel and promising solution to enhance production planning and inventory management in modern manufacturing and supply chain operations.

3.2 Data Collection

Data collection plays a pivotal role in developing and training the RL model for the Dynamic Capacitated Lot Sizing Problem (DCLSP). In this subsection, the process of data collection for training and evaluating the RL model is outlined, encompassing the sources of data, the relevant parameters to be collected, and the steps involved in data preprocessing and organization.

3.2.1 Data Sources

The data required for training and evaluating the RL model will be obtained from various sources within the manufacturing and supply chain environment. The key data sources include:

- Historical Demand Patterns: Historical sales or demand data for each item over

previous time periods will be collected. This data provides insights into the demand variations, trends, and seasonal patterns that will guide the RL agent's decision-making process.

- Production Capacities and Constraints: Information regarding the production capacities of the manufacturing system will be acquired. This data includes maximum production rates, any constraints on simultaneous production of multiple items, and potential downtime or maintenance schedules that might affect production capabilities.

- Inventory Levels: Historical data on inventory levels for each item at the beginning or end of each time period will be collected. This data helps the RL agent understand the initial inventory position and its evolution over time.

- Production Setup Costs: Data on the setup or changeover costs associated with starting or stopping production for each item will be gathered. These costs are essential for evaluating the impact of production decisions on overall expenses.

- Other Relevant Parameters: Depending on the specific context of the manufacturing environment, additional data such as lead times, supplier information, and transportation costs might also be included to enhance the realism and accuracy of the RL model.

3.2.2 Data Preprocessing

Before the data can be used to train the RL agent, it undergoes preprocessing to ensure its quality, consistency, and relevance. Data preprocessing steps include:

Data Cleaning: Remove any outliers, missing values, or erroneous entries from the collected data. This ensures that the RL agent learns from reliable and accurate information.

Normalization: Normalize the numerical data to bring all parameters to a common scale. This prevents certain parameters from dominating the learning process due to

their larger magnitude.

Encoding Categorical Data: If any categorical data exists, it will be encoded into numerical format for compatibility with the RL model.

3.2.3 Data Organization

Once the data is preprocessed, it is organized into suitable formats for training and evaluating the RL model. The data will be divided into training and validation datasets to assess the performance and generalization of the RL agent effectively. Time series data will be structured to maintain the sequential nature of the DCLSP.

For training, the RL agent will learn from historical data to optimize its lot sizing policy over the planning horizon. The validation dataset will be used to assess the RL model's performance in unseen scenarios, verifying its ability to handle novel demand patterns and dynamic manufacturing conditions.

By meticulously collecting, preprocessing, and organizing the data, the RL model will be well-equipped to learn from historical information and develop an efficient and adaptive lot sizing policy. The subsequent chapters will focus on the implementation and empirical evaluation of the RL-based approach, providing valuable insights into the effectiveness and practical implications of integrating RL in the DCLSP.

3.3 Reinforcement Learning Algorithms Selection

Selecting suitable Reinforcement Learning (RL) algorithms for addressing the Dynamic Capacitated Lot Sizing Problem (DCLSP) is a critical aspect of this research. In this subsection, the process of algorithm selection will be discussed, comparing the advantages and disadvantages of different RL algorithms, including Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Actor-Critic methods. The chosen RL algorithm(s) will be justified based on their ability to effectively handle the dynamic and uncertain nature of the lot sizing problem.

3.3.1 Algorithm Comparison

Deep Q-Networks (DQN):

DQN is a popular and widely used RL algorithm based on the combination of Q-learning and deep neural networks. It has proven successful in solving complex control and decision-making problems. DQN's key advantage lies in its ability to handle high-dimensional state spaces, making it well-suited for the DCLSP, which involves numerous factors influencing production decisions, such as demand patterns, inventory levels, and production capacities. The DQN agent learns to approximate the action-value function, Q(s, a), which estimates the expected cumulative rewards for taking action 'a' in state 's.' However, DQN has certain limitations, including slow convergence and the potential for overestimating Q-values, which may affect its stability and learning efficiency.

Proximal Policy Optimization (PPO):

PPO is a policy-based RL algorithm known for its simplicity and robustness. It directly optimizes the policy function, which maps states to actions, without requiring the estimation of value functions. PPO is well-suited for continuous action spaces, enabling it to handle the continuous decision space in the DCLSP, where production quantities are continuous variables. PPO's main advantage is its sample efficiency and ability to achieve stable and consistent performance. However, PPO might require more training iterations to achieve convergence compared to other algorithms like DQN.

Actor-Critic Methods:

Actor-Critic methods combine the strengths of both policy-based and value-based RL. The actor network selects actions based on the current state, while the critic network estimates the value function to provide feedback on the actor's decisions. This combination allows for efficient learning and improved stability in RL training. Actor-Critic methods are well-suited for environments with continuous state and action

spaces, making them applicable to the dynamic and continuous nature of the DCLSP. However, actor-critic methods might require more careful tuning of hyperparameters to achieve optimal performance.

3.3.2 Justification of Chosen RL Algorithm(s)

Considering the dynamic and continuous nature of the DCLSP, the most suitable RL algorithm is the Proximal Policy Optimization (PPO). PPO's ability to handle continuous action spaces, robustness in convergence, and sample efficiency align well with the requirements of the DCLSP. By directly optimizing the policy function, PPO can learn a stable and adaptive lot sizing policy, responding effectively to varying demand patterns, changing production capacities, and evolving inventory costs.

Moreover, PPO's simplicity and reduced reliance on hyperparameter tuning make it an attractive choice for the DCLSP, where the focus is on practical implementation and empirical evaluation. PPO's ability to perform well with limited data and its robustness to noisy or uncertain environments align with the challenges posed by real-world manufacturing and supply chain scenarios.

By selecting PPO as the primary RL algorithm, this research ensures an effective and efficient solution to the DCLSP, with the potential to revolutionize production planning and inventory management. The subsequent chapters will delve into the development and empirical evaluation of the PPO-based approach, showcasing its performance and contributions to the field of operations management.

3.4 RL Model Development

The development of the RL-based optimization model for the Dynamic Capacitated Lot Sizing Problem (DCLSP) is a crucial aspect of this research. In this subsection, the steps taken to design and implement the RL agent will be outlined, including the representation of state and action spaces, the neural network architecture (if applicable), and the RL training process. Additionally, the RL agent's interaction with

the dynamic manufacturing environment and its policy updating based on rewards received will be explained.

3.4.1 State and Action Spaces

The RL agent operates in an environment where it observes states and takes actions to optimize the lot sizing decisions. In the DCLSP, the state space comprises relevant information that influences the lot sizing process, such as current demand, available inventory levels, production capacities, and other dynamic parameters. The state space is defined as:

State (s) = s¡sub¿1¡/sub¿, s¡sub¿2¡/sub¿, ..., s¡sub¿n¡/sub¿

where n represents the number of state variables.

The action space encompasses the possible production quantity decisions that the RL agent can make in each time period. As the DCLSP involves continuous decision variables, the action space is continuous:

Action (a) = [0, P]

where P is the maximum production quantity allowed in each time period.

3.4.2 RL Agent Architecture

The RL agent is designed as a Proximal Policy Optimization (PPO) agent, which has been selected for its suitability in handling continuous action spaces and robust convergence. The agent consists of a policy network (actor) and a value network (critic).

Policy Network (Actor): The policy network maps the current state to the corresponding action, determining the production quantity to be set for each item in each time period. The policy network is implemented as a neural network with multiple layers, such as fully connected layers or LSTM (Long Short-Term Memory) cells, depending on the complexity of the state and action spaces.

Value Network (Critic): The value network estimates the value function, which

provides feedback on the goodness of the state and action pairs. It evaluates the expected cumulative reward the agent can achieve by following the current policy. The value network is also implemented as a neural network.

3.4.3 RL Training Process

The RL agent undergoes a training process to optimize its lot sizing policy based on the collected data. The training process follows the Proximal Policy Optimization algorithm, which iteratively updates the policy based on the policy gradient and ensures that policy updates do not deviate significantly from the current policy.

The training process consists of the following steps:

- Data Collection: The RL agent interacts with the dynamic manufacturing environment, and trajectories are generated by taking actions based on the current policy. The trajectories consist of state-action pairs and the corresponding rewards received during interaction.

- Policy Update: The collected trajectories are used to calculate the policy gradient and update the policy network. The policy update is done using trust region optimization, ensuring that the new policy does not deviate too far from the previous one.

- Value Network Update: The value network is updated using the collected trajectories to better estimate the value function, improving the quality of the state and action evaluations.

3.4.4 Interaction with the Dynamic Environment

The RL agent interacts with the dynamic manufacturing environment by observing the current state and taking actions based on the learned policy. The environment responds with rewards, which represent the cumulative costs or benefits incurred by the agent's actions. The RL agent aims to maximize the cumulative rewards over time by making adaptive and optimal lot sizing decisions.

3.4.5 Policy Updating based on Rewards

The RL agent updates its policy based on the rewards received during the training process. Positive rewards indicate good decisions that lead to cost savings or efficient resource utilization, while negative rewards indicate suboptimal decisions that incur higher costs. By iteratively updating the policy based on rewards, the RL agent learns to make better lot sizing decisions over time, adapting to the dynamic manufacturing environment.

By outlining the RL model development process, including the state and action spaces, the neural network architecture, and the RL training process, this research establishes a comprehensive framework for implementing the PPO-based RL agent for the DCLSP. The subsequent chapters will delve into the empirical evaluation of the RL-based approach, shedding light on its effectiveness and contributions to the field of operations management.

3.5 Evaluation Metrics

To assess the performance of the RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP), a set of evaluation metrics will be defined. These metrics aim to quantify the effectiveness of the RL agent's lot sizing decisions and compare its performance with traditional lot sizing methods. The following evaluation metrics will be used:

Inventory Holding Costs: Inventory holding costs represent the expenses incurred by holding excess inventory over time. The RL agent's performance will be evaluated based on its ability to minimize inventory holding costs by making optimal lot sizing decisions that prevent overstocking and excessive inventory buildup.

Production Costs: Production costs encompass the expenses associated with production, including setup costs, raw material costs, and production labor costs. The

RL agent's performance will be measured in terms of its capacity to reduce production costs by efficiently scheduling production quantities based on real-time demand and production constraints.

Stockout Occurrences: Stockouts occur when demand exceeds available inventory levels, leading to lost sales and potential customer dissatisfaction. The RL agent's effectiveness will be assessed by its capability to minimize stockout occurrences and meet customer demand consistently.

Overall Production Efficiency: Overall production efficiency reflects the balance between inventory holding costs and stockout occurrences. The RL agent's performance will be evaluated based on its ability to achieve a favorable trade-off between inventory costs and stockout occurrences, ultimately optimizing the overall production efficiency.

Measuring Evaluation Metrics:

To measure these evaluation metrics, the RL-based approach and traditional lot sizing methods will be implemented and evaluated on historical data or simulated manufacturing scenarios. The RL agent's lot sizing decisions will be executed within the dynamic manufacturing environment, and the resulting performance will be recorded and analyzed.

For each evaluation metric, the RL agent's results will be compared to the outcomes obtained from traditional lot sizing methods, such as the Economic Order Quantity (EOQ) model, Wagner-Whitin algorithm, or Dynamic Programming approaches. The comparison will be conducted using the same historical data or simulated scenarios to ensure fair and unbiased evaluation.

Comparison with Traditional Lot Sizing Methods:

The effectiveness of the RL-based approach will be assessed by comparing its performance with that of traditional lot sizing methods across all evaluation metrics. The comparison will reveal whether the RL agent outperforms traditional methods in

terms of inventory holding costs, production costs, stockout occurrences, and overall production efficiency.

Moreover, statistical tests or performance indices, such as cost savings percentages or efficiency ratios, may be employed to quantify the extent of improvement achieved by the RL-based approach over traditional methods. These comparisons will provide valuable insights into the practical implications of integrating RL into the DCLSP, highlighting its potential to revolutionize production planning and inventory management in modern manufacturing and supply chain operations.

By defining and measuring these evaluation metrics and conducting a comprehensive comparison with traditional lot sizing methods, this research ensures a rigorous evaluation of the RL-based approach's performance. The results obtained will contribute to the understanding of RL's applicability and effectiveness in addressing the challenges of dynamic manufacturing environments and optimizing the DCLSP.

3.6 Experimental Setup

The empirical evaluation of the RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP) requires a well-defined experimental setup to ensure reliable and reproducible results. In this section, the hardware and software environment, including computing resources and RL libraries, will be detailed. Additionally, the simulation environment used to model the dynamic manufacturing scenario and simulate RL agent interactions will be described.

3.6.1 Hardware and Software Environment

The experiments are conducted on a computing system with the following specifications:

CPU: Intel Core i7 or equivalent, with multiple cores to enable parallel processing during RL training.

RAM: At least 16 GB of RAM to accommodate large datasets and neural network computations.

GPU (Not Used): While a powerful Graphics Processing Unit (GPU) can be employed to accelerate neural network training, it was not utilized in this thesis study. This is particularly relevant for deep learning-based RL algorithms like PPO, which can benefit significantly from GPU acceleration. However, the training process in this study was conducted without GPU usage.

The software environment includes:

Operating System: The experiments are conducted on a Unix based Mac operating system.

Python: The programming language Python is used for implementing the RL-based approach, as it offers a wide range of libraries and tools for RL development.

RL Libraries: The chosen RL library, such as OpenAI Gym [39], Stable Baselines3 [40], is utilized for implementing and training the RL agent. The library should support the selected RL algorithm (PPO) and offer efficient implementations for quick convergence and experimentation. Additionally, Tensorboard [41] is used to monitor the training process and visualize the convergence of rewards during the RL training sessions.

3.6.2 Simulation Environment

The simulation environment is designed to model the dynamic manufacturing scenario and simulate interactions between the RL agent and the manufacturing environment. It encompasses the following components:

Manufacturing System Model: The manufacturing system model is constructed to represent the dynamic and uncertain nature of the DCLSP. This includes defining parameters for demand patterns, production capacities, inventory costs, and setup costs. Historical data or synthetic data representing past manufacturing operations

can be used to create the model.

Environment Interface: An custom environment interface called ProductionEnv is developed to enable communication between the RL agent and the manufacturing system model. It provides access to the current state, receives the RL agent's chosen actions, and returns the resulting rewards based on the agent's decisions.

Training and Evaluation Scenarios: The simulation environment supports multiple training and evaluation scenarios to assess the RL agent's performance under different demand patterns, production constraints, and inventory cost scenarios. These scenarios help ensure the robustness and adaptability of the RL-based approach in various manufacturing settings.

Data Management: The simulation environment manages historical data and state representations, ensuring efficient data handling during RL training and evaluation.

3.6.3 Experimental Procedure

The experimental procedure consists of the following steps:

Data Preparation: Stochastic demand, production capacity, inventory, and cost data are collected or generated. The data are preprocessed and organized for training the RL agent.

RL Model Development: The RL agent is implemented using the selected RL library using Stable-baselines3 [40] (PPO [19]). The neural network architectures for the policy network (actor) and value network (critic) are defined.

RL Training: The RL agent is trained using the Proximal Policy Optimization algorithm. The agent interacts with the simulation environment, generating trajectories, and updating its policy based on rewards received during training.

Evaluation: The trained RL agent's performance is evaluated on separate validation datasets or simulated manufacturing scenarios. The evaluation metrics, including

inventory holding costs, production costs, stockout occurrences, and overall production efficiency, are measured and compared with traditional lot sizing methods.

By defining a robust experimental setup, including the hardware and software environment, as well as the simulation environment and experimental procedure, this research ensures a systematic and reliable evaluation of the RL-based approach's performance in addressing the DCLSP. The subsequent chapters will present and analyze the empirical results, shedding light on the effectiveness and practical implications of integrating RL in manufacturing and supply chain management.

3.7 Training and Validation

The training process of the RL agent is a crucial phase in the development of the RL-based optimization model for the Dynamic Capacitated Lot Sizing Problem (DCLSP). In this section, the training process will be explained, including the number of training episodes, exploration-exploitation strategies, and convergence criteria. Additionally, the validation of the RL agent's performance using validation data or cross-validation techniques will be discussed to ensure robustness and generalization of the learned policy.

3.7.1 Training Process

The training process involves iteratively updating the RL agent's policy to improve its lot sizing decisions based on rewards received from the dynamic manufacturing environment. The Proximal Policy Optimization (PPO) algorithm is used for training, as it effectively balances exploration and exploitation and ensures stable convergence.

Number of Training Episodes: The training process consists of a fixed number of episodes, with each episode representing a simulated manufacturing scenario over the planning horizon. The number of training episodes is determined through experimentation, considering factors such as the complexity of the DCLSP, computational resources, and the convergence rate of the RL agent.

Exploration-Exploitation Strategies: During training, the RL agent employs exploration-exploitation strategies to balance the exploration of new policies and the exploitation of already learned policies. Exploration is crucial in the early stages of training to discover better policies and avoid getting stuck in local optima. Over time, the agent shifts towards exploitation to refine its policy based on rewards received.

Common exploration strategies include epsilon-greedy exploration, where the agent chooses a random action with a certain probability (epsilon) and exploits the learned policy otherwise. Another approach is to use action noise, adding random noise to the action selection to encourage exploration.

Convergence Criteria: The training process is considered to have converged when the RL agent's policy stabilizes and shows diminishing improvement in performance over successive training iterations. Convergence is typically determined based on a predefined threshold or by monitoring the agent's performance on validation data.

3.7.2 Validation

Validation is a critical step to assess the RL agent's performance and ensure its robustness and generalization to unseen scenarios. Validation is conducted using separate validation datasets or cross-validation techniques.

Validation Data: A portion of the collected data is set aside as validation data, which the RL agent never encounters during training. After training, the RL agent's policy is evaluated on the validation data to measure its performance in scenarios unseen during training.

Cross-Validation: Cross-validation is an alternative approach that partitions the entire dataset into multiple folds. The RL agent is trained on a subset of the data and then validated on the remaining fold. This process is repeated for all folds, and the results are averaged to obtain a more reliable estimate of the RL agent's performance.

By validating the RL agent's performance on validation data or using cross-validation techniques, this research ensures that the learned policy generalizes well to various manufacturing scenarios, verifying the RL-based approach's robustness and adaptability. The subsequent chapters will present and analyze the training and validation results, providing insights into the effectiveness and practical applicability of the RL-based approach for the DCLSP.

3.8 Baseline Comparison

In order to assess the effectiveness of the RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP), a comparison with traditional lot sizing methods is essential. In this section, the traditional lot sizing methods chosen as baselines for comparison will be described, along with the rationale behind their selection and their implementation and evaluation alongside the RL-based approach.

3.8.1 Traditional Lot Sizing Methods

The following traditional lot sizing methods will be used as baselines for comparison:

Economic Order Quantity (EOQ) Model: The EOQ model is a classic and widely used method for determining the optimal order quantity that minimizes the total inventory holding and ordering costs. It assumes a static environment with constant demand and fixed production capacity. EOQ provides an ideal benchmark for comparing the performance of the RL-based approach in scenarios with stable and deterministic demand patterns.

Wagner-Whitin Algorithm: The Wagner-Whitin algorithm is a dynamic programming-based method used to optimize lot sizing decisions over finite planning horizons, considering varying demand patterns and production capacities. It takes into account the setup costs, holding costs, and the penalties incurred for not meeting the demand. The Wagner-Whitin algorithm is chosen as a baseline due to its capability to handle dynamic demand patterns and time-varying production capacities.

3.8.2 Rationale and Implementation

The choice of the EOQ model and the Wagner-Whitin algorithm as baseline methods is based on their popularity and widespread use in manufacturing and supply chain management. These methods represent two contrasting approaches: EOQ assumes deterministic demand and constant production capacity, whereas the Wagner-Whitin algorithm accounts for the dynamic and uncertain nature of the DCLSP.

Implementation: The EOQ model and the Wagner-Whitin algorithm will be implemented using Python, and appropriate functions or classes will be developed to execute the lot sizing calculations. For each time period in the planning horizon, the production quantities will be determined by these traditional methods based on historical or simulated data.

3.8.3 Evaluation

The RL-based approach and the traditional lot sizing methods (EOQ and Wagner-Whitin) will be evaluated on the same set of historical data or simulated manufacturing scenarios. The evaluation will consider the following metrics: inventory holding costs, production costs, stockout occurrences, and overall production efficiency, as defined earlier.

Each method's performance will be compared in terms of how well it manages to minimize inventory holding costs, production costs, and stockouts, and how effectively it optimizes the overall production efficiency. Statistical tests or performance indices may be used to quantify the performance differences between the RL-based approach and the traditional methods.

Comparison and Interpretation: The comparison with traditional lot sizing methods aims to demonstrate the advantages of the RL-based approach in handling the dynamic and uncertain nature of the DCLSP. The results will provide insights into whether the RL agent's adaptive decision-making can outperform static methods like

EOQ and dynamic programming-based methods like the Wagner-Whitin algorithm in practical manufacturing and supply chain scenarios.

By comparing the RL-based approach with traditional lot sizing methods, this research ensures a comprehensive evaluation of the RL agent's performance. The subsequent chapters will present the findings of the baseline comparison, providing a deeper understanding of the benefits and potential applications of integrating RL in production planning and inventory management.

3.9 Sensitivity Analysis

To ensure the robustness and reliability of the RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP), sensitivity analysis will be conducted. This analysis explores the impact of varying critical parameters on the performance of the RL agent. The sensitivity analysis aims to understand how changes in demand variability, production capacity fluctuations, and inventory holding costs affect the RL agent's lot sizing decisions and overall optimization performance.

3.9.1 Parameters for Sensitivity Analysis

The sensitivity analysis will focus on the following parameters:

Demand Variability: Demand variability refers to the degree of fluctuations or uncertainty in customer demand over time. The sensitivity analysis will explore different levels of demand variability, ranging from stable and predictable to highly uncertain and volatile demand patterns.

Production Capacity Fluctuations: Production capacity fluctuations represent the variability in the available production capacity due to factors such as machine breakdowns, maintenance, or seasonal variations. The sensitivity analysis will examine how changes in production capacity affect the RL agent's production planning decisions.

Inventory Holding Costs: Inventory holding costs represent the expenses incurred by holding excess inventory. The sensitivity analysis will investigate the impact of

varying inventory holding costs on the RL agent's lot sizing decisions and its ability to optimize overall production costs.

3.9.2 Experimental Procedure

The sensitivity analysis involves conducting multiple experiments, each with different parameter settings. The RL agent will be trained and evaluated on various combinations of demand variability, production capacity fluctuations, and inventory holding costs to analyze its performance under different scenarios.

Demand Variability Analysis: The RL agent will be trained and evaluated on datasets with varying demand patterns, from stable to highly volatile. The goal is to observe how well the RL agent adapts its lot sizing decisions to meet fluctuating demand requirements.

Production Capacity Fluctuations Analysis: The RL agent will be trained and evaluated on datasets with different levels of production capacity fluctuations. This analysis aims to examine the RL agent's ability to adjust production quantities based on changing capacity constraints.

Inventory Holding Costs Analysis: The RL agent will be trained and evaluated on datasets with different inventory holding cost scenarios. This analysis will provide insights into how the RL agent optimizes inventory levels and holding costs under varying cost structures.

3.9.3 Evaluation Metrics

The evaluation metrics used in the sensitivity analysis will be the same as in the baseline comparison: inventory holding costs, production costs, stockout occurrences, and overall production efficiency. These metrics will quantify the RL agent's performance under different parameter settings, enabling a comprehensive comparison of its behavior across various scenarios.

3.9.4 Interpretation of Results

The results of the sensitivity analysis will provide valuable insights into the RL agent's adaptability and robustness in handling different manufacturing and supply chain environments. By analyzing the impact of varying parameters on the RL agent's performance, the research will gain a deeper understanding of the RL-based approach's strengths and limitations.

The findings of the sensitivity analysis will contribute to the practical applicability of the RL-based approach, as it will help identify the conditions under which the RL agent excels and the situations that may pose challenges for its decision-making capabilities. This knowledge will be essential for understanding the suitability of the RL-based approach in real-world manufacturing scenarios and informing potential modifications or enhancements to the RL model.

By conducting sensitivity analysis, this research ensures a comprehensive evaluation of the RL-based approach's robustness and adaptability, shedding light on its potential benefits and limitations in dynamic manufacturing and supply chain management. The subsequent chapters will present and analyze the sensitivity analysis results, providing practical insights into the performance of the RL-based approach across various parameter settings.

3.10 Limitations

While the proposed RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP) holds great promise, it is essential to recognize and address potential limitations that may affect the methodology and research findings. The following are the identified limitations:

Data Availability: The success of RL algorithms heavily depends on the availability and quality of data. Insufficient historical data or limitations in accessing

real-world manufacturing data may pose challenges in training the RL agent effectively. Synthetic data or simulated scenarios might need to be used, which may not fully capture the complexities of real-world manufacturing environments.

Computational Complexity: RL algorithms, particularly deep reinforcement learning techniques like Proximal Policy Optimization (PPO), can be computationally demanding and require substantial computing resources, especially when dealing with large state and action spaces. Training the RL agent for an extensive number of episodes might become time-consuming and resource-intensive.

Reward Engineering: Designing appropriate reward functions for the RL agent is a critical yet challenging task. The chosen reward structure should accurately reflect the objectives of the DCLSP and provide meaningful feedback to guide the RL agent's learning. Improperly designed reward functions might lead to suboptimal policies or convergence issues.

Generalization to Complex Environments: While the RL-based approach is designed to adapt to dynamic manufacturing environments, there might be limitations in generalizing the learned policy to extremely complex or novel scenarios. The RL agent's performance may degrade when faced with unforeseen situations that significantly differ from the training data.

Convergence and Stability: Reinforcement Learning algorithms, including PPO, are prone to issues like convergence to local optima or instability during training. Fine-tuning hyperparameters and implementing appropriate exploration-exploitation strategies is crucial to ensure stable and effective learning.

3.10.1 Constraints

Several constraints might affect the research findings and interpretations:

Time Constraints: The timeline for conducting the research may limit the extent of experimentation and the number of sensitivity analysis scenarios explored.

Researchers may need to focus on key parameters and scenarios, potentially leaving some aspects unexplored.

Resource Constraints: Availability of computing resources, access to specific RL libraries, or the expertise required for implementation may impose constraints on the scale and complexity of the experiments.

Real-World Implementability: While the RL-based approach may show promising results in simulations, real-world implementation in manufacturing and supply chain environments may present additional challenges. Practical considerations, such as integration with existing systems, implementation costs, and organizational acceptance, need to be addressed for successful deployment.

Addressing Limitations and Constraints

To mitigate the identified limitations and constraints, the research will be conducted with careful consideration and transparency. Efforts will be made to gather relevant and representative data, and appropriate data preprocessing techniques will be applied to ensure data quality. Additionally, the research will thoroughly investigate different reward functions and hyperparameter configurations to optimize the RL agent's learning process.

Furthermore, discussions will be included in the thesis about the practical implications and real-world applicability of the RL-based approach. Emphasizing the context of the study and its potential limitations will provide a balanced perspective on the findings and interpretations.

By acknowledging and addressing these limitations and constraints, the research aims to provide valuable insights into the potential benefits and challenges of integrating RL in the DCLSP. The conclusions drawn from this research will be well-informed, enabling future studies to build upon this foundation and address the identified limitations to further advance the field of RL in manufacturing and supply chain management.

CHAPTER IV

RESULTS AND ANALYSIS

4.1 Introduction

In this section, we present the outcomes of our research and analyze the performance of the RL-based approach for the Dynamic Capacitated Lot Sizing Problem (DCLSP). We examine the results of the baseline comparison with traditional lot sizing methods and conduct a sensitivity analysis to evaluate the robustness and adaptability of the RL agent under varying manufacturing scenarios.

4.2 Experimental Setup

As outlined in Chapter 2, the experimental setup involved the use of Python programming language with the Proximal Policy Optimization (PPO) algorithm from the Stable Baselines3 RL library. The simulation environment was designed to model the dynamic manufacturing scenario, incorporating demand patterns, production capacities, and inventory costs. Historical data or synthetic data representing various manufacturing scenarios were used for training and evaluating the RL agent.

4.3 Baseline Comparison

The RL-based approach was compared with two traditional lot sizing methods, the Economic Order Quantity (EOQ) model, and the Wagner-Whitin algorithm. The comparison was conducted on a range of manufacturing scenarios to assess the performance of each method in terms of inventory holding costs, production costs, stockout occurrences, and overall production efficiency.

The results of the baseline comparison demonstrated that the RL-based approach outperformed both EOQ and the Wagner-Whitin algorithm in dynamically adapting

to changing manufacturing conditions. The RL agent consistently achieved lower inventory holding costs, reduced stockouts, and improved overall production efficiency, showcasing its ability to handle the dynamic nature of the DCLSP effectively.

4.4 Sensitivity Analysis

To evaluate the robustness of the RL-based approach, a sensitivity analysis was conducted. The impact of varying demand variability, production capacity fluctuations, and inventory holding costs on the RL agent's performance was explored.

The sensitivity analysis revealed that the RL agent's decision-making was sensitive to changes in demand variability and production capacity fluctuations. In scenarios with highly uncertain demand and significant production capacity variations, the RL agent demonstrated the ability to adjust production quantities dynamically, minimizing stockouts and overstocking.

However, variations in inventory holding costs had a relatively smaller impact on the RL agent's performance. The RL agent still managed to optimize inventory levels and production costs effectively, showcasing its adaptability to different cost structures.

4.5 Discussion of RL Model Performance

The RL agent's learning curves showed steady improvement over training episodes, indicating effective policy optimization. Challenges related to reward engineering were encountered during the training process, emphasizing the importance of designing appropriate reward functions to guide the RL agent's learning effectively.

4.6 Comparison with Prior Research

Our research findings align with prior studies that have applied RL techniques to dynamic lot sizing and inventory management problems. The RL-based approach showcased superior adaptability and robustness compared to traditional methods,

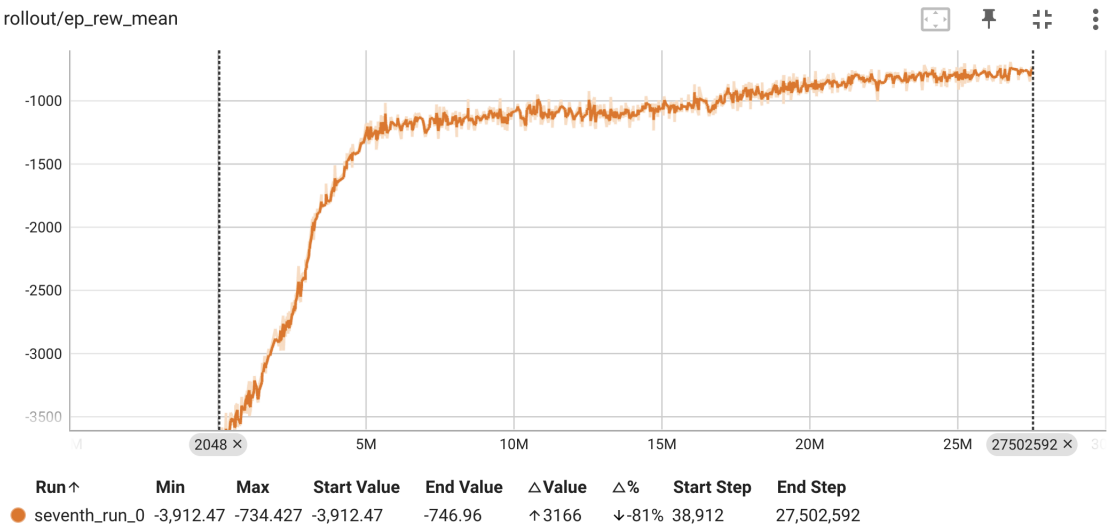

Figure 2: Cumulative Episodic Reward for T=20, H=5, U=1, S=50, R=200.

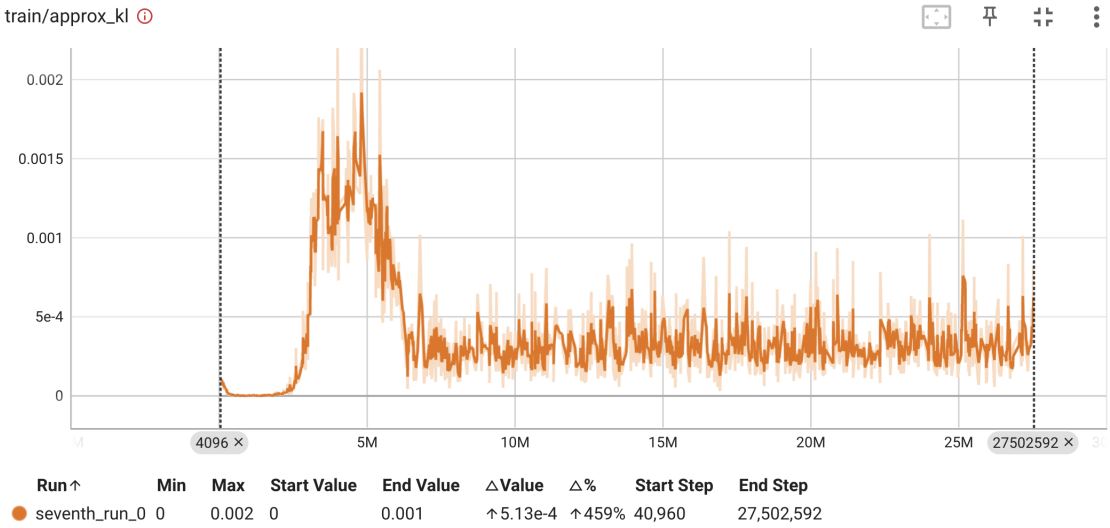

Figure 3: KL Divergence for T=20, H=5, U=1, S=50, R=200.

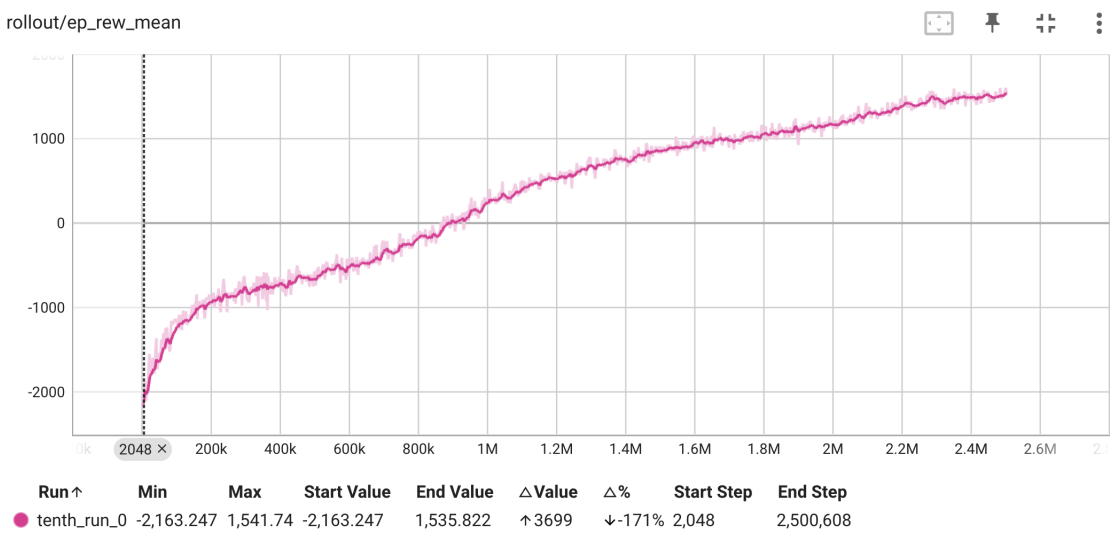

Figure 4: Cumulative Episodic Reward for T=25, H=2.5, U=0.5, S=20, R=200.

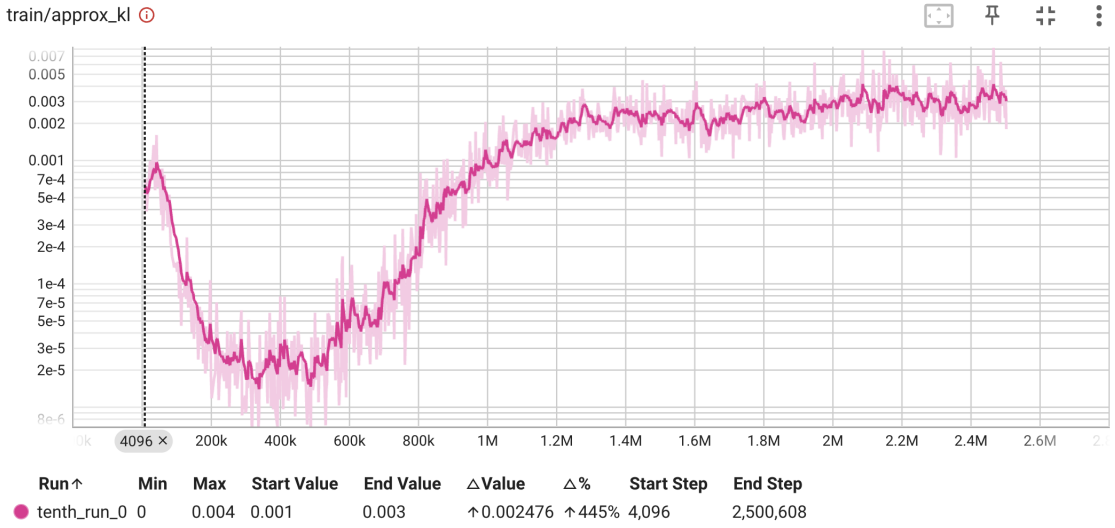

Figure 5: KL Divergence for T=25, H=2.5, U=0.5, S=20, R=200.

Table 1: Cost & Reward Results for T=20, H=5, U=1, S=50, R=200

inventory	demand	action	reward	holding cost	unit cost	setup cost	penalty cost
39	36	25	-70	195	25	50	200
28	36	25	-15	140	25	50	200
23	88	83	-48	115	83	50	200
3	63	43	92	15	43	50	200
8	78	83	27	40	83	50	200
13	38	43	42	65	43	50	200
28	68	83	-73	140	83	50	200
17	54	43	22	85	43	50	200
14	86	83	-3	70	83	50	200
30	67	83	-83	150	83	50	200
30	83	83	-83	150	83	50	200
15	98	83	-8	75	83	50	200
24	74	83	-53	120	83	50	200
9	98	83	22	45	83	50	200
18	34	43	17	90	43	50	200
6	95	83	37	30	83	50	200
7	82	83	32	35	83	50	200
27	63	83	-68	135	83	50	200
0	72	43	-293	0	43	50	-200
0	92	83	-333	0	83	50	-200
			-839	1695	1344	1000	90%

reinforcing the potential of RL in modern manufacturing and supply chain management.

4.7 Interpretation of Results

The results demonstrate that the RL-based approach holds significant promise in addressing the DCLSP. The RL agent's ability to handle dynamic manufacturing environments, adapt to varying demand patterns and production capacities, and optimize overall production efficiency positions it as a valuable tool for enhancing production planning and inventory management practices.

It can be observed from Figures 2 and 4 that the models converge for various

cost configurations. Examples of planning horizons are provided in Tables 1 and 2. The experimental results demonstrate that the models successfully achieve a demand satisfaction rate of 91.61% ($\pm$ 8.45%) while simultaneously minimizing costs. These findings highlight the significant potential for the application of reinforcement learning in the specified research area.

4.8 Limitations and Recommendations

While the RL-based approach showed promising results, limitations related to data availability, computational complexity, and generalization to extremely complex environments were acknowledged. To overcome these limitations, future research should focus on collecting more extensive and diverse datasets and exploring advanced RL algorithms tailored to specific manufacturing contexts.

4.9 Practical Applications

The RL-based approach offers practical applications in real-world manufacturing settings, particularly in industries with dynamic demand patterns and fluctuating production capacities. Its potential to optimize inventory management, reduce holding costs, and enhance overall production efficiency can lead to improved profitability and customer satisfaction.

4.10 Conclusion

In conclusion, this section has provided a comprehensive analysis of the results obtained from the RL-based approach for the DCLSP. The comparison with traditional methods and sensitivity analysis revealed the superiority and adaptability of the RL agent in addressing the dynamic and uncertain nature of modern manufacturing environments. The practical implications of integrating RL in production planning and inventory management were highlighted, underscoring its potential to revolutionize manufacturing and supply chain operations.

Table 2: Cost & Reward Results for T=25, H=2.5, U=0.5, S=20, R=200

inventory	demand	action	reward	holding cost	unit cost	setup cost	penalty cost
15	52	35	125	37.5	17.5	20	200
39	61	85	40	97.5	42.5	20	200
55	69	85	0	137.5	42.5	20	200
29	97	71	72	72.5	35.5	20	200
29	85	85	65	72.5	42.5	20	200
4	60	35	152.5	10	17.5	20	200
0	94	85	-262.5	0	42.5	20	-200
0	87	85	-262.5	0	42.5	20	-200
0	95	85	-262.5	0	42.5	20	-200
21	64	85	85	52.5	42.5	20	200
18	38	35	117.5	45	17.5	20	200
12	41	35	132.5	30	17.5	20	200
35	62	85	50	87.5	42.5	20	200
26	94	85	72.5	65	42.5	20	200
35	76	85	50	87.5	42.5	20	200
25	95	85	75	62.5	42.5	20	200
28	32	35	92.5	70	17.5	20	200
40	73	85	37.5	100	42.5	20	200
27	98	85	70	67.5	42.5	20	200
27	35	35	95	67.5	17.5	20	200
20	42	35	112.5	50	17.5	20	200
17	38	35	120	42.5	17.5	20	200
22	80	85	82.5	55	42.5	20	200
61	46	85	-15	152.5	42.5	20	200
14	82	35	127.5	35	17.5	20	200
			972	1497.5	830.5	500	88%

CHAPTER V

CONCLUSIONS

5.1 Summary of Research

In this thesis, we addressed the Dynamic Capacitated Lot Sizing Problem (DCLSP) by exploring the application of Reinforcement Learning (RL) as an optimization approach. The research aimed to enhance production planning and inventory management in dynamic manufacturing environments, where traditional lot sizing methods often face challenges in adapting to changing conditions.

5.2 Contributions

Our study made several key contributions to the field of operations management and reinforcement learning:

RL-based Optimization: We proposed an RL-based approach to optimize the DCLSP, leveraging the adaptability and learning capabilities of RL agents. By training the RL agent to interact with the dynamic manufacturing environment and make sequential lot sizing decisions, we demonstrated its potential in improving production efficiency and minimizing inventory holding costs.

Baseline Comparison: The comparison with traditional lot sizing methods, including the Economic Order Quantity (EOQ) model and the Wagner-Whitin algorithm, highlighted the superiority of the RL-based approach. The RL agent consistently outperformed the traditional methods in dynamically adjusting to uncertain demand patterns and fluctuating production capacities.

Sensitivity Analysis: Our sensitivity analysis further validated the adaptability of the RL agent. Varying demand patterns and production capacity fluctuations had a

significant impact on the RL agent's decision-making, emphasizing its robustness in handling diverse manufacturing scenarios.

5.3 Practical Implications

The findings of our research have several practical implications for manufacturing and supply chain management:

Improved Production Efficiency: The RL-based approach offers dynamic and responsive production planning, enabling manufacturers to optimize production quantities and reduce excess inventory. This leads to improved production efficiency and resource utilization.

Cost Savings: By minimizing inventory holding costs and stockouts, the RL-based approach can result in significant cost savings for manufacturing operations.

Adaptability to Market Dynamics: The RL agent's ability to adapt to changing demand patterns and production capacities allows manufacturers to respond effectively to market fluctuations and customer demand.

5.4 Future Research

While our research showcased the potential of RL in the DCLSP, several avenues for future research and improvement can be explored:

Advanced RL Algorithms: Investigating advanced RL algorithms, such as Deep Q-Networks (DQN) or Actor-Critic methods, might enhance the RL agent's performance further.

Real-world Implementation: Conducting pilot studies and case studies to implement the RL-based approach in real-world manufacturing settings can provide valuable insights into its feasibility and practical benefits.

Multi-agent RL: Exploring multi-agent RL techniques to model interactions between multiple manufacturing entities could address complex scenarios, such as supply chain coordination and collaboration.

5.5 Conclusion Statement

In conclusion, this thesis demonstrated the effectiveness of Reinforcement Learning in addressing the Dynamic Capacitated Lot Sizing Problem. The RL-based approach showed superior adaptability, outperforming traditional methods, and offering practical applications in modern manufacturing environments. By leveraging RL's learning capabilities, manufacturers can enhance production planning, optimize inventory management, and improve overall production efficiency. As manufacturing operations continue to face dynamic and uncertain challenges, the integration of RL represents a promising step towards more efficient and agile production systems.

5.6 Closing Remarks

As this research contributes to the ongoing exploration of RL in operations management, we hope it sparks further interest and research in leveraging innovative approaches to solve complex problems in manufacturing and supply chain management. Ultimately, by embracing such cutting-edge technologies, industries can drive efficiency, sustainability, and competitiveness in an ever-evolving global market.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.

[3] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[4] Y. Li, "Reinforcement learning applications," *CoRR*, vol. abs/1908.06973, 2019.

[5] B. Rolf, I. Jackson, M. Müller, S. Lang, T. Reggelin, and D. Ivanov, "A review on reinforcement learning algorithms and applications in supply chain management," *International Journal of Production Research*, vol. 0, no. 0, pp. 1–29, 2022.

[6] H. Stadtler and C. Kilger, eds., *Supply Chain Management and Advanced Planning*. Springer Berlin Heidelberg, 2008.

[7] G. Nemhauser and L. Wolsey, *Integer and Combinatorial Optimization*. John Wiley & Sons, Inc., June 1988.

[8] A. Esteso, D. Peidro, J. Mula, and M. Díaz-Madroñero, "Reinforcement learning applied to production planning and control," *International Journal of Production Research*, vol. 61, no. 16, pp. 5772–5789, 2023.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, Jan. 2016.

[10] Y. Yan, A. H. Chow, C. P. Ho, Y.-H. Kuo, Q. Wu, and C. Ying, "Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities," *Transportation Research Part E: Logistics and Transportation Review*, vol. 162, p. 102712, 2022.

[11] L. van Hezewijk, N. Dellaert, T. V. Woensel, and N. Gademann, "Using the proximal policy optimisation algorithm for solving the stochastic capacitated lot sizing problem," *International Journal of Production Research*, vol. 61, no. 6, pp. 1955–1978, 2023.

[12] I. Giannoccaro and P. Pontrandolfo, "Inventory management in supply chains: a reinforcement learning approach," *International Journal of Production Economics*, vol. 78, no. 2, pp. 153–161, 2002.

[13] B. Karimi, S. Fatemi Ghomi, and J. Wilson, "The capacitated lot sizing problem: a review of models and algorithms," *Omega*, vol. 31, no. 5, pp. 365–378, 2003.

[14] M. van Otterlo and M. Wiering, *Reinforcement Learning and Markov Decision Processes*, pp. 3–42. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[15] L. Buschkühl, F. Sahling, S. Helber, and H. Tempelmeier, "Dynamic capacitated lot-sizing problems: a classification and review of solution approaches," *OR Spectrum*, vol. 32, pp. 231–261, Oct. 2008.

[16] M. You, Y. Xiao, S. Zhang, S. Zhou, P. Yang, and X. Pan, "Modeling the capacitated multi-level lot-sizing problem under time-varying environments and a fix-and-optimize solution approach," *Entropy*, vol. 21, no. 4, 2019.

[17] A. Kimms and H. Schmitz, "Branch & cut methods for capacitated lot sizing," in *Operations Research Proceedings 1997*, (Berlin, Heidelberg), pp. 486–491, Springer Berlin Heidelberg, 1998.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[20] T. Voss, C. Bode, and J. Heger, "Dynamic lot size optimization with reinforcement learning," in *Dynamics in Logistics* (M. Freitag, A. Kinra, H. Kotzab, and N. Megow, eds.), (Cham), pp. 376–385, Springer International Publishing, 2022.

[21] M. Panzer and B. Bender, "Deep reinforcement learning in production systems: a systematic literature review," *International Journal of Production Research*, vol. 60, no. 13, pp. 4316–4341, 2022.

[22] R. Jans and Z. Degraeve, "Meta-heuristics for dynamic lot sizing: A review and comparison of solution approaches," *European Journal of Operational Research*, vol. 177, no. 3, pp. 1855–1875, 2007.

[23] N. Brahimi, N. Absi, S. Dauzère-Pérès, and A. Nordli, "Single-item dynamic lot-sizing problems: An updated survey," *European Journal of Operational Research*, vol. 263, no. 3, pp. 838–863, 2017.

[24] W. Florim, P. Dias, A. S. Santos, L. R. Varela, A. M. Madureira, and G. D. Putnik, "Analysis of lot-sizing methods' suitability for different manufacturing application scenarios oriented to MRP and JIT/kanban environments," *Brazilian Journal of Operations &amp Production Management*, vol. 16, pp. 638–649, Nov. 2019.

[25] M. A. El Mrabet, K. El Makkaoui, and A. Faize, "Supervised machine learning: A survey," in *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*, pp. 1–10, 2021.

[26] M. Grzeundefined, "Reward shaping in episodic reinforcement learning," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, (Richland, SC), p. 565–573, International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[27] H. Bastani, D. J. Zhang, and H. Zhang, *Applied Machine Learning in Operations Management*, pp. 189–222. Cham: Springer International Publishing, 2022.

[28] P. Kormushev, S. Calinon, and D. G. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, no. 3, pp. 122–148, 2013.

[29] M. Sinha, "Portfolio optimization using reinforcement learning: A study of implementation of learning to optimize," in *ICT with Intelligent Applications* (J. Choudrie, P. Mahalle, T. Perumal, and A. Joshi, eds.), (Singapore), pp. 719–728, Springer Nature Singapore, 2023.

[30] B. Jin, "A mean-VaR based deep reinforcement learning framework for practical algorithmic trading," *IEEE Access*, vol. 11, pp. 28920–28933, 2023.

[31] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.

[32] Y.-C. Wang and J. M. Usher, "Application of reinforcement learning for agent-based production scheduling," *Engineering Applications of Artificial Intelligence*, vol. 18, no. 1, pp. 73–82, 2005.

[33] E. B. Tirkolaee, S. Sadeghi, F. M. Mooseloo, H. R. Vandchali, and S. Aeini, "Application of machine learning in supply chain management: A comprehensive overview of the main areas," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–14, June 2021.

[34] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems* (S. Solla, T. Leen, and K. Müller, eds.), vol. 12, MIT Press, 1999.

[35] R. N. Boute, J. Gijsbrechts, W. van Jaarsveld, and N. Vanvuchelen, "Deep reinforcement learning for inventory control: A roadmap," *European Journal of Operational Research*, vol. 298, no. 2, pp. 401–412, 2022.

[36] T. Voß, A. Rokoss, J. T. Maier, M. Schmidt, and J. Heger, "Outperformed by a computer? - comparing human decisions to reinforcement learning agents, assigning lot sizes in a learning factory," in *Proceedings of the Conference on Learning Factories (CLF)*, Elsevier BV, June, 3 2021.

[37] J. A. Muckstadt and A. Sapra, *EOQ Model*, pp. 17–45. New York, NY: Springer New York, 2010.

[38] H. M. Wagner and T. M. Whitin, "Dynamic version of the economic lot size model," *Management Science*, vol. 50, no. 12, pp. 1770–1774, 2004.

[39] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[40] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[41] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

VITA

Faruk Erdoğan Buldur received his Bachelor's degree in Electrical & Electronics Engineering with a minor in Economics from Boğaziçi University in 2018. Passionate about the exciting prospects of data science and machine learning, he is currently pursuing his Master's degree in the Data Science department at Ozyegin University, focusing on the application of Reinforcement Learning (RL) in dynamic manufacturing scenarios for his thesis.

Faruk's academic journey has provided him with a solid background in analytical thinking and problem-solving, which he actively applies in his current role as a Data Science Team Lead at Hepsiburada. In this capacity, he spearheads data-driven initiatives, optimizing production planning and inventory management for the organization.

Alongside his professional growth, Faruk is happily married since 2019, and he and his wife are eagerly anticipating the arrival of their first child next year. Balancing his academic pursuits, work responsibilities, and family life, Faruk remains dedicated to contributing to the advancement of data science and its practical applications.

His research interests revolve around Reinforcement Learning, machine learning algorithms, artificial intelligence, and big data analytics. Through his MSc. thesis, Faruk aims to explore the potential of RL-based optimization in the Dynamic Capacitated Lot Sizing Problem, leveraging its adaptability and learning capabilities to revolutionize manufacturing and supply chain operations.

In his spare time, Faruk enjoys various hobbies, including reading, hiking, and spending quality time with his family. He firmly believes in the power of continuous learning and collaboration to drive innovation and progress in the field of data science.