Iranian Electric Industry Journal
Quality & Productivity
P.P 73-87

Vol. 15/ No. 1/ Spring 2026
ISSN 2322-2344
EISSN 2717-1639

Dynamic Order Allocation Optimization in Multi-Supplier Supply Chains under Cost Uncertainty: A Comparative Study of Dynamic Programming and Reinforcement Learning Abstract

Leila Hosseini, Mohammad Saber Fallah Nezhad, Vali Derhami, Mohammad Saleh Owlia
Leila Hosseini, PhD Student, Department of Industrial Engineering, Faculty of Engineering, Yazd University, Yazd, Iran, Email: lysahoseini@stud.yazd.ac.ir
Mohammad Saber Fallah Nezhad, Professor, Department of Industrial Engineering, Faculty of Engineering, Yazd University, Yazd, Iran,Email: Fallahnezhad@yazd.ac.ir
Vali Derhami,Professor, Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran, Email: vderhami@yazd.ac.ir
Mohammad Saleh Owlia,Professor, Department of Industrial Engineering, Faculty of Engineering, Yazd University, Yazd, Iran, Email: owliams@yazd.ac.ir

Abstract:

Dynamic Order Allocation Optimization in Multi-Supplier Supply Chains under Cost Uncertainty: A Comparative Study of Dynamic Programming and Reinforcement Learning

1. Introduction and Research Objective

Order management in multi-source supply chains under stochastic cost fluctuations poses significant structural challenges. In recent years, the complexity of supply chains has increased considerably, and uncertainty has become a structural feature of operational environments. Fluctuations in ordering costs, raw material prices, and supply conditions create serious challenges for decision-making. This issue is particularly critical in industries with high dependence on raw materials, such as polymers. In such conditions, determining the order quantity and its allocation among multiple suppliers becomes a dynamic, stochastic, and multi-level problem.

The key innovation of this research lies in the different definition of the problem. Unlike most existing models that assume the decision-maker has access to the probability distribution of future costs or can accurately estimate it from historical data, this study defines the problem with a different understanding. Although historical cost data are available, due to the inherent volatility of fluctuating markets (such as polymers), the distribution extracted from past data cannot reliably represent the future. Sudden changes in exchange rates, tariff policies, and supply shocks make the predictability of cost distribution in future periods highly uncertain. Therefore, the decision-maker does not have access to a stable and deterministic distribution of future costs at the time of choice.

The main objective of this research is to provide an advanced decision-making framework for multi-period ordering problems in multi-supplier supply chains under cost uncertainty. By focusing on a real and complex multi-supplier problem, this research aims to bridge the gap between model-based methods and data-driven approaches.

2. Research Methodology

The problem studied in this research is defined as a multi-period optimization model with a finite horizon within the framework of a Markov Decision Process. In this problem, the decision-maker, aiming to minimize the total ordering and inventory holding costs, determines the total order quantity and its allocation among suppliers in each period by observing the system state (including inventory level and supplier cost levels).

Data: The research data includes 188 weekly observations from March 2022 to October 2025, with main data including demand, ordering costs from two suppliers, and inventory holding costs. These data were extracted from a real industrial unit in the polymer industry and formed the basis of the empirical analysis.

Data Preprocessing: Due to significant inflationary fluctuations, all costs were converted to constant prices using the Consumer Price Index. To enable Markovian modeling, continuous costs were mapped to three levels: low, medium, and high. The median absolute deviation from the median was used to determine the thresholds for cost level differentiation.

Markov Decision Process Framework: The state space includes the combination of cost levels of two suppliers, initial inventory, and period number. The action space includes the total order volume and the share allocated to each supplier. The reward function is defined as the negative of the total cost.Solution Methods: Two complementary approaches were used to solve the problem:- Dynamic Programming: As a model-based method that assumes access to the system's probabilistic structure (transition matrices extracted from historical data), it derives the optimal policy by solving Bellman equations.- Q-learning Reinforcement Learning: As a model-free method that learns the decision policy incrementally

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

through interaction with the environment without explicit modeling of cost distribution. Learning parameters include a learning rate of 0.1, a discount factor of 1, and 20,000 episodes.

3. Research Findings

To evaluate the performance of the proposed framework, the model was validated using real data from an industrial unit active in the production of polymer disposable containers. Three policies were compared: the company's current policy (reference policy), the dynamic programming-based policy, and the reinforcement learning-based policy.

Total Cost Comparison: The results show that in all 47 planning horizons, the costs from dynamic programming and reinforcement learning were lower than the reference policy. The average savings for dynamic programming was 0.23% and for reinforcement learning was 0.21%. Considering the enormous scale of supply chain costs, this improvement leads to very significant absolute savings (average of approximately 73,000 monetary units per horizon). In some horizons, absolute savings exceeded 124 million Rials. In scenarios with significant cost differences between suppliers, these savings increased to over 40%.Notably, the performance of reinforcement learning was very close to that of dynamic programming. In many horizons, the total costs of these two methods were exactly equal (difference less than 0.01%). This finding confirms that the proposed data-driven approach, despite not requiring modeling of future cost distribution, performs very close to the optimal model-based method.Order Allocation Analysis: The analysis of order allocation to suppliers shows that the reference policy follows a relatively stable and predetermined approach. In contrast, dynamic programming performs allocation in a state-dependent manner, proportional to the cost conditions of each period. The reinforcement learning policy also reproduced patterns similar to dynamic programming in many horizons. In conditions where one supplier has a clear cost advantage, both data-driven policies focus orders toward that supplier, while the current policy shows less tendency toward such concentration.Convergence and Stability of Reinforcement Learning: Examination of the learning process confirms numerical stability and convergence of the extracted policy. The average Q-values decreased sharply in the early stages of training and converged to a stable level after approximately 10,000 episodes. The management of the exploration-exploitation balance through exponential decay of the ε parameter from 1 to 0 indicates the algorithm's successful transition from extensive exploration to exploitation. The simultaneous convergence of average cost, reduction in standard deviation, and decrease in variance indicate that the final policy is not only optimized in terms of efficiency but has also reached an acceptable and stable level in terms of risk and dispersion.

4. Conclusion and Recommendations

This research was conducted to develop an advanced decision-making framework for multi-period ordering problems in multi-supplier supply chains under cost uncertainty. Unlike previous studies that often assume the probability distribution of costs is known or estimable from historical data, this research, by accepting the fact that in volatile markets the cost distribution is inherently unstable, presented a novel framework based on Markov Decision Processes.

Empirical findings showed that both dynamic programming and reinforcement learning approaches outperform the reference policy. The main strength of the framework was revealed in sensitivity analyses, where the savings potential increased significantly up to 40% as the cost difference between suppliers widened. The Q-learning algorithm, despite not requiring modeling of future cost distribution, performed very close to dynamic programming and demonstrated high flexibility in facing cost fluctuations.

The proposed framework can serve as the core of decision support systems in volatile industries such as petrochemicals, steel, and polymers. It is recommended that organizations strengthen their data infrastructure for accurate and real-time recording of supply costs. For future studies, extending the model to dependent products, combining with deep learning, simultaneous consideration of demand uncertainty, and adding operational constraints such as supplier capacity and lead time are suggested.

Keywords: Supply Chain, Dynamic Programming, Reinforcement Learning, Q-learning, Markov Decision Process, Order Management

Keywords: Supply Chain, Dynamic Programming, Reinforcement Learning, Q-learning, Markov Decision Process, Order Management

Date of sending the article: 2026/04/30
Acceptance date of the article: 2026/07/22

Name of the Corresponding Author: Mohammad Saber Fallah Nezhad
Corresponding Author's Address: Department of Industrial Engineering, Faculty of Engineering, Yazd University, Yazd,.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

Iranian Electric Industry Journal
Quality & Productivity
ص.ص ۸۷-۷۳

Vol. 15 / No. 1/ Spring 2026
ISSN 2322-2344
EISSN 2717-1639

بهینه‌سازی پویای تخصیص سفارش در زنجیره تأمین چندتأمین‌کننده تحت عدم‌قطعیت هزینه: مقایسه برنامه‌ریزی پویا و یادگیری تقویتی

نوع مطالعه:پژوهشی

لیلا حسینی[۱]، محمد صابر فلاح نژاد[۲] ، ولی درهمی[۳] ، محمد صالح اولیا [۴]

۱- دانشجوی دکتری، گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه یزد، یزد، ایران lysahoseini@stud.yazd.ac.ir

۲- استاد، گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه یزد، یزد، ایران Fallahnezhad@yazd.ac.ir

۳- استاد، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه یزد، یزد، ایران vderhami@yazd.ac.ir

۴- استاد، گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه یزد، یزد، ایران owliams@yazd.ac.ir

چکیده:

مدیریت سفارش‌دهی در زنجیره‌های تأمین چندمنبعی تحت تأثیر نوسانات تصادفی هزینه‌ها، چالش‌های ساختاری جدی را ایجاد می‌کند؛ در این پژوهش، مسئله به‌عنوان یک تصمیم‌گیری راهبردی در شرایط عدم‌قطعیت تعریف شده است؛ به‌گونه‌ای که اگرچه داده‌های تاریخی هزینه‌ها در دسترس است، اما به‌دلیل نوسانات ساختاری و تغییرات پیوسته‌ی بازار، توزیع حاصل از داده‌های گذشته نمی‌تواند نماینده‌ی دقیقی برای آینده باشد و تصمیم‌گیرنده در لحظه‌ی انتخاب، به توزیع پایدار و قطعی هزینه‌های آتی دسترسی ندارد. در رویکرد برنامه‌ریزی پویا، ماتریس‌های احتمال انتقال از داده‌های تاریخی استخراج شده و به‌عنوان معیار مرجع بهینه (مدل‌محور) عمل می‌کنند. در مقابل، رویکرد یادگیری تقویتی مبتنی بر Q بدون نیاز به برآورد صریح توزیع احتمالی و صرفاً از طریق تعامل با محیط شبیه‌سازی‌شده (بر پایه‌ی داده‌های واقعی)، سیاست تصمیم‌گیری را می‌آموزد. اعتبارسنجی مدل در صنعت پلیمر و با استفاده از داده‌های واقعی در ۴۷ افق زمانی نشان می‌دهد که چارچوب پیشنهادی، حتی در شرایط پایه با هزینه‌های نزدیک تأمین‌کنندگان، به‌طور میانگین ۰٫۰۲۳٪ صرفه‌جویی معنادار (p-value<۰٫۰۰۰۱) ایجاد می‌کند که معادل کاهش بیش از ۷۳ هزار واحد پولی در هر افق است. افزون بر این، تحلیل‌های حساسیت نشان می‌دهد که با افزایش اختلاف هزینه‌ای میان تأمین‌کنندگان، پتانسیل صرفه‌جویی به‌طور چشمگیری تا ۴۰٪ افزایش می‌یابد. اگرچه برنامه‌ریزی پویا بهینه‌ترین عملکرد نظری را ارائه می‌دهد، اما روش یادگیری Q باوجود عدم نیاز به مدل‌سازی توزیع آینده، تطبیق‌پذیری سریع و پایداری بالایی در برابر شوک‌های بازار از خود نشان می‌دهد. این پژوهش با پر کردن شکاف میان بهینه‌سازی تئوریک و روش‌های داده‌محور، مبنایی عملی و مقاوم برای پیاده‌سازی سیستم‌های پشتیبانی تصمیم در زنجیره‌های تأمین با اطلاعات ناقص و هزینه‌های تصادفی ناپایدار فراهم می‌آورد.

واژه‌های کلیدی: زنجیره تأمین، برنامه‌ریزی پویا، یادگیری تقویتی، Q-learning، تصمیم‌گیری مارکوف، مدیریت سفارش‌دهی

تاریخ ارسال مقاله:۱۴۰۵/۰۲/۱۰

تاریخ پذیرش مقاله:۱۴۰۵/۰۴/۳۱

نام نویسنده مسئول: محمد صابر فلاح نژاد

نشانی نویسنده مسئول: گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه یزد، یزد، ایران

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

۱- مقدمه

در سال‌های اخیر، پیچیدگی زنجیره‌های تأمین به طور قابل توجهی افزایش یافته و عدم قطعیت به یک ویژگی ساختاری محیط‌های عملیاتی تبدیل شده است. نوسانات در هزینه‌های سفارش، قیمت مواد اولیه و شرایط تأمین، چالش‌های جدی برای تصمیم‌گیری ایجاد می‌کند (Chen and Rossi, 2021, Kawakami et al., 2021). این مسئله در صنایع با وابستگی بالا به مواد اولیه، مانند فولاد و پلیمرها، حیاتی‌تر است. در چنین شرایطی، تعیین مقدار سفارش و تخصیص آن بین چندین تأمین‌کننده به یک مسئله پویا، تصادفی و چندسطحی تبدیل می‌شود (Li et al., 2009).

نکتهٔ کلیدی و نوآوری این پژوهش در تعریف متفاوت مسئله نهفته است. برخلاف اغلب مدل‌های موجود که فرض می‌کنند تصمیم‌گیرنده به توزیع احتمالی هزینه‌های آینده دسترسی دارد یا می‌تواند آن را با دقت از داده‌های تاریخی برآورد کند، در پژوهش حاضر مسئله با درکی متفاوت تعریف می‌شود. اگرچه داده‌های تاریخی هزینه‌ها در دسترس است، اما به‌دلیل نوسان‌پذیری ذاتی بازارهای پرنوسان (نظیر پلیمر و فولاد)، توزیع استخراج‌شده از داده‌های گذشته نمی‌تواند نمایندهٔ معتبری برای آینده باشد. تغییرات ناگهانی نرخ ارز، سیاست‌های تعرفه‌ای و شوک‌های عرضه، پیش‌بینی‌پذیری توزیع هزینه‌ها را در دوره‌های آتی با عدم‌قطعیت جدی مواجه می‌سازد. ازاین‌رو، تصمیم‌گیرنده در لحظهٔ انتخاب، به توزیع پایدار و قطعی هزینه‌های آتی دسترسی ندارد. این وضعیت که در بسیاری از محیط‌های صنعتی واقعی رخ می‌دهد، توسط مدل‌های کلاسیک برنامه‌ریزی پویا یا فرآیندهای تصمیم مارکوف که بر ایستایی توزیع هزینه‌ها استوارند، به‌خوبی قابل مدیریت نیست. بنابراین، ضرورت این پژوهش ایجاد چارچوبی است که بدون اتکای صرف به توزیع برآوردشده از گذشته، بتواند سیاست‌های سفارش‌دهی نزدیک به بهینه را در مواجهه با ناایستایی بازار تولید کند.

مدل‌های کلاسیک مدیریت موجودی، مانند مدل مقدار سفارش اقتصادی، بر فرض ثبات پارامترها استوار هستند و در محیط‌های پایدار عملکرد مناسبی دارند (Scarf, 1960).

اگرچه پژوهش‌های گسترده‌ای مدل مدل مقدار سفارش اقتصادی را تحت عدم‌قطعیت پارامترهایی مانند تقاضای تصادفی، هزینه‌های متغیر و نرخ‌های بهره نامعین توسعه داده‌اند (Scarf, 1960, Thompson et al., 2009)، اما این مدل‌ها عمدتاً بر ساختار تک‌تأمین‌کننده و تصمیم‌گیری تک‌دوره‌ای یا با افق نامحدود متمرکز هستند و توانایی مدل‌سازی هم‌زمان چندتأمین‌کنندگی، تخصیص پویای سفارش و وابستگی‌های بین‌دوره‌ای را در افق محدود ندارند (Thompson et al., 2009). در ادامه، روش‌هایی مانند برنامه‌ریزی پویا و الگوریتم واگنر–ویتین برای بهبود دقت تصمیم‌گیری توسعه یافتند (Wagner and Whitin, 2004, Wagner and Whitin, 1958, Wagelmans et al., 1992). این روش‌ها اگرچه بهینه هستند، اما به دانش کامل از محیط نیاز دارند و در مسائل با ابعاد بزرگ با محدودیت محاسباتی مواجه می‌شوند (Van Roy et al., 1997).

چارچوب فرآیند تصمیم‌گیری مارکوف ابزار مناسبی برای مدل‌سازی مسائل تصمیم‌گیری چندمرحله‌ای تحت عدم‌قطعیت فراهم می‌کند (Azaron et al., 2009). در ادامه، پژوهش‌ها به سمت روش‌های مقیاس‌پذیرتر حرکت کردند. به‌طور خاص، برنامه‌ریزی پویای تقریبی و روش‌های تقریب تابع ارزش امکان حل مسائل با ابعاد بزرگ را فراهم کردند (Pontrandolfo et al., 2002, De Farias and Van Roy, 2003). همچنین، کاربرد برنامه‌ریزی پویا تطبیقی در مسائل تخصیص منابع در مقیاس بزرگ نشان داد که می‌توان در حضور محدودیت‌های محاسباتی، به راه‌حل‌های باکیفیت دست یافت (Powell et al., 2002). در همین راستا، روش‌های مبتنی بر سیاست‌های نگاه به آینده، مانند الگوریتم‌های گرادیان دانش، عملکرد رقابتی در محیط‌های تصمیم‌گیری آنلاین ارائه دادند (Ryzhov et al., 2008). با این حال، بخش عمده این مطالعات به ساختارهای ساده یا تک‌تأمین‌کننده محدود بوده و فرض ثبات یا سادگی در ساختار هزینه‌ها را حفظ کرده‌اند (Barnes-Schuster et al., 2002). در مقابل، یادگیری تقویتی به‌عنوان یک رویکرد مدل‌آزاد، امکان یادگیری سیاست‌های تصمیم‌گیری از طریق تعامل با محیط را فراهم کرده است (Sutton and Barto, 1998). این ویژگی، آن را به گزینه‌ای مناسب برای محیط‌های پیچیده و با اطلاعات ناقص تبدیل کرده است. مطالعات اخیر نشان داده‌اند که یادگیری تقویتی می‌تواند در مدیریت موجودی چندسطحی، انتخاب تأمین‌کننده و کاهش هزینه‌های عملیاتی مؤثر باشد (Geevers et al., 2024, Zhou et al., 2024, Piao et al., 2023). همچنین، ترکیب یادگیری تقویتی با بهینه‌سازی مقاوم به کاهش ریسک ناشی از نوسانات بازار کمک کرده است (Kharvi and Pakkala, 2022, Luo et al., 2021) و می‌توان میان اهداف اقتصادی و زیست‌محیطی توازن ایجاد کرد (Hosseini et al., 2022, Nayeri et al., 2023). با وجود این پیشرفت‌ها، بسیاری از مدل‌ها همچنان از در نظر گرفتن هم‌زمان پویایی هزینه‌های سفارش‌دهی، چندتأمین‌کنندگی و وابستگی‌های بین‌دوره‌ای ناتوان هستند. از سوی دیگر، برخی مطالعات به بررسی انتخاب پویای تأمین‌کنندگان با استفاده از یادگیری تقویتی پرداخته‌اند، اما عدم‌قطعیت هزینه‌ها و ساختار چندمرحله‌ای تصمیمات را به‌طور کامل لحاظ نکرده‌اند (Bilsel and Kumara, 2007). همچنین، توسعه رویکردهای چندعاملی و مدل‌های مقاوم نشان داده است که می‌توان اثر شلاقی را کاهش داد و تاب‌آوری زنجیره تأمین را بهبود بخشید (Zhao and Sun, 2010). در سال‌های اخیر، استفاده از چارچوب‌های مارکوفی برای مسائل چندتأمین‌کننده نیز گسترش یافته و نشان داده است که این رویکردها در محیط‌های پیچیده کارایی مناسبی دارند (Maulidi et al., 2022). با این حال، هماهنگی میان تأمین‌کنندگان و مدل‌سازی

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

دقیق نوسانات تصادفی هزینه‌ها همچنان یک چالش باز باقی مانده است.

پژوهش‌های جدید در صنایع مختلف نیز بر اهمیت توسعه مدل‌های دقیق‌تر تأکید دارند. برای مثال، استفاده از برنامه‌ریزی پویای تقریبی در صنعت فولاد منجر به کاهش هزینه‌ها و افزایش پاسخ‌پذیری شده است(Wu et al., 2024). افزون بر این، کاربرد فرآیندهای تصمیم‌گیری مارکوف و یادگیری تقویتی در زنجیره‌های تأمین حلقه‌بسته، به بهبود هماهنگی و کارایی تصمیم‌گیری کمک کرده است (Zhang et al., 2024). در حوزه پلیمر، ترکیب روش‌های تصمیم‌گیری چندمعیاره با مدل‌های بهینه‌سازی چندهدفه توانسته است توازن مناسبی میان هزینه، ریسک و پایداری ایجاد کند (Cheraghalipour and Farsad, 2018, Sulistyoningarum et al., 2020). با این وجود، این رویکردها اغلب فاقد یک چارچوب یکپارچه برای تصمیم‌گیری پویا در محیط‌های چندمرحله‌ای و تصادفی هستند. در همین راستا، مطالعاتی با استفاده از ترکیب روش جمع‌سپاری و تکنیک تصمیم‌گیری چندمعیاره آراس به شناسایی و رتبه‌بندی ریسک‌های زنجیره تأمین برق پرداختند و نشان دادند که این رویکرد ترکیبی می‌تواند به افزایش دقت در اولویت‌بندی ریسک‌ها کمک کند(dabbagh et al., 2025). علاوه بر این، روندهای اخیر نشان می‌دهند که استفاده از داده‌های واقعی، یادگیری ماشین و روش‌های شبیه‌سازی در حال افزایش است و می‌تواند دقت تصمیم‌گیری را بهبود دهد (Savvopoulou, 2024). همچنین، روش‌های بهینه‌سازی پیشرفته مانند برنامه‌ریزی عدد صحیح مختلط، الگوریتم‌های فراابتکاری و تجزیه بندرز در مسائل تخصیص منابع کاربرد یافته‌اند (Chauhan et al., 2023, Mohamed et al., 2023). با این حال، این روش‌ها معمولاً بر ساختارهای ایستا تکیه دارند و انعطاف‌پذیری لازم برای مواجهه با محیط‌های کاملاً پویا و نامطمئن را ندارند (Milani, 2024, Wulan, 2023).

در حوزه‌های مجاور نیز پیشرفت‌های قابل‌توجهی در توسعه الگوریتم‌های یادگیری تقویتی حاصل شده است. برای مثال، کاربرد یادگیری تقویتی مبتنی بر Q در مسائل مالی و معاملاتی نشان داده است که این الگوریتم می‌تواند در محیط‌های پرنوسان عملکرد قابل قبولی داشته باشد (Darezereshki and Derhami, 2024). همچنین، توسعه روش‌های ترکیبی مانند یادگیری تقویتی فازی و الگوریتم‌های مبتنی بر نظریه آشوب، به بهبود تعادل میان اکتشاف و بهره‌برداری کمک کرده‌اند (Nadi et al., 2024, Khodadadi and Derhami, 2025). با این حال، انتقال این دستاوردها به مسائل مدیریت سفارش‌دهی صنعتی همچنان نیازمند چارچوب‌سازی و انطباق با ویژگی‌های خاص این حوزه است.

علیرغم پیشرفت‌های قابل توجه در ادبیات موضوع، یک شکاف اساسی همچنان پابرجاست. مقایسه نظام‌مند بین رویکردهای مبتنی بر مدل (مانند برنامه ریزی پویا) و رویکردهای بدون مدل (مانند یادگیری تقویتی) در یک چارچوب یکسان و تحت عدم قطعیت چندتأمین‌کننده و چندسطحی به ندرت انجام شده است. به عبارت دیگر، مشخص نیست که آیا روش‌های بدون مدل (یادگیری تقویتی) می‌توانند در شرایطی که توزیع پایدار هزینه‌ها در دسترس نیست، به عملکردی نزدیک به روش‌های بهینهٔ مبتنی بر مدل دست یابند یا خیر. علاوه بر این، بسیاری از مدل‌های موجود به طور همزمان تأثیر متقابل چندین تأمین‌کننده، پویایی هزینه‌ها و وابستگی‌های بین دوره‌ای را به طور جامع مدل نمی‌کنند (Zhang et al., 2024). این محدودیت‌ها مانع از ارائه راهکارهای عملی و قابل اعتماد برای محیط‌های پیچیده صنعتی می‌شود و درک روشنی از مزایا و معایب نسبی هر رویکرد در شرایط واقعی فراهم نمی‌کند.

این پژوهش با تمرکز بر یک مسئله واقعی و پیچیده در زمینه چندتأمین‌کننده، سعی در پر کردن شکاف فوق دارد و یک چارچوب تحلیلی تعمیم‌پذیر ارائه می‌دهد. مشارکت اصلی در سه سطح خلاصه می‌شود. نخست طراحی فضای حالت هوشمند در چارچوب فرایند تصمیم‌گیری مارکوف که بدون نیاز به دانستن توزیع احتمالی پایدار هزینه‌های آینده، پویایی تغییرات تصادفی را مدل می‌کند، دوم مقایسه نظام‌مند و تحلیلی دو رویکرد مکمل (برنامه‌ریزی پویا به عنوان معیار بهینه با فرض دانستن کامل پویایی سیستم، و یادگیری Q به عنوان روشی بدون مدل که فقط از طریق تعامل با محیط یاد می‌گیرد) و سوم اعتبارسنجی مدل در دو صنعت نوسانی (پلیمر و فولاد) که تعمیم‌پذیری چارچوب را فراتر از یک مورد خاص نشان می‌دهد. یافته‌ها نه تنها بینش روشنی درباره عملکرد نسبی روش‌ها ارائه می‌دهد، بلکه راهنمایی عملی برای تصمیم‌گیرندگان در انتخاب سیاست‌های سفارش‌دهی در محیط‌های پویا و نوسانی فراهم می‌آورد. به این ترتیب، این مقاله به طور معناداری به توسعه روش‌شناختی و کاربرد عملی در حوزه مدیریت زنجیره تأمین و تصمیم‌گیری تحت عدم قطعیت کمک می‌کند.

۲- روش تحقیق

۲-۱- چارچوب کلی مسئله و منطق تصمیم گیری

مسئله مورد مطالعه در این پژوهش به‌صورت یک مدل بهینه‌سازی چندمرحله‌ای با افق محدود و در چارچوب فرآیند تصمیم‌گیری مارکوف تعریف می‌شود. در این مسئله، تصمیم‌گیرنده با هدف کمینه‌سازی مجموع هزینه‌های سفارش‌دهی و نگهداری موجودی، در هر دوره با مشاهده وضعیت سیستم (شامل میزان موجودی و سطوح هزینه‌ی تأمین‌کنندگان)، مقدار کل سفارش و نحوه‌ی تخصیص آن میان تأمین‌کنندگان را تعیین می‌کند.

نکته‌ی کلیدی در تعریف مسئله، ماهیت عدم‌قطعیت هزینه‌هاست. اگرچه داده‌های تاریخی هزینه‌ها در دسترس است، اما به‌دلیل تغییرات پیوسته و پیش‌بینی‌ناپذیر بازار بازارهای پرنوسان (نظیر پلیمر و فولاد)، توزیع استخراج‌شده از داده‌های گذشته نمی‌تواند نماینده‌ی معتبری برای آینده باشد. تغییرات ناگهانی نرخ ارز، سیاست‌های تعرفه‌ای و شوک‌های عرضه، پیش‌بینی‌پذیری توزیع هزینه‌ها را در دوره‌های آتی با عدم‌قطعیت جدی مواجه می‌سازد. ازاین‌رو، تصمیم‌گیرنده در لحظه‌ی انتخاب، به

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

توزیع پایدار و قطعی هزینه‌های آتی دسترسی ندارد. این رویکرد، تفاوت بنیادین این پژوهش با مدل‌های کلاسیک است که معمولاً بر ایستایی توزیع هزینه‌ها استوارند.

در رویه‌های سنتی مدیریت موجودی، مانند مدل مقدار سفارش اقتصادی یا سیاست‌های دوره‌ای ثابت، تصمیم‌گیری در هر دوره اغلب مستقل از سایر دوره‌ها انجام می‌شود. در این رویه‌ها، تصمیم‌گیرنده بر اساس میزان موجودی فعلی، نقطه‌ی سفارش ثابت و مقدار سفارش اقتصادی ازپیش‌محاسبه‌شده عمل می‌کند و تأثیر سفارش امروز بر هزینه‌های دوره‌های آینده را به‌صورت پویا در نظر نمی‌گیرد.

در مقابل، رویکرد پویای این پژوهش بر دو مؤلفه‌ی اساسی استوار است. نخست، در نظر گرفتن پیامدهای آتی تصمیمات فعلی؛ به‌این معنا که تصمیم‌گیرنده آگاه است سفارش امروز چه تأثیری بر میزان موجودی و هزینه‌های دوره‌های آینده خواهد داشت و این وابستگی بین‌دوره‌ای را در تصمیم خود لحاظ می‌کند. به‌عبارتی، انتخاب مقدار سفارش در هر دوره، با آگاهی از پیامدهای آن بر موجودی و هزینه‌های بعدی صورت می‌پذیرد. دوم، واکنش‌پذیری به تغییرات لحظه‌ای هزینه‌ها؛ بدین معنا که برخلاف رویه‌های سنتی که در آنها تخصیص سفارش میان تأمین‌کنندگان به‌صورت ثابت و ازپیش‌تعیین‌شده انجام می‌شود، در رویکرد پویا، تخصیص سفارش بر اساس آخرین اطلاعات هزینه‌ای و شرایط لحظه‌ای بازار، به‌صورت تطبیقی صورت می‌گیرد. این ویژگی به مدل اجازه می‌دهد در مواجهه با شوک‌های قیمتی، به‌سرعت واکنش نشان داده و سفارش را به‌سمت تأمین‌کننده‌ی با هزینه‌ی کمتر سوق دهد.

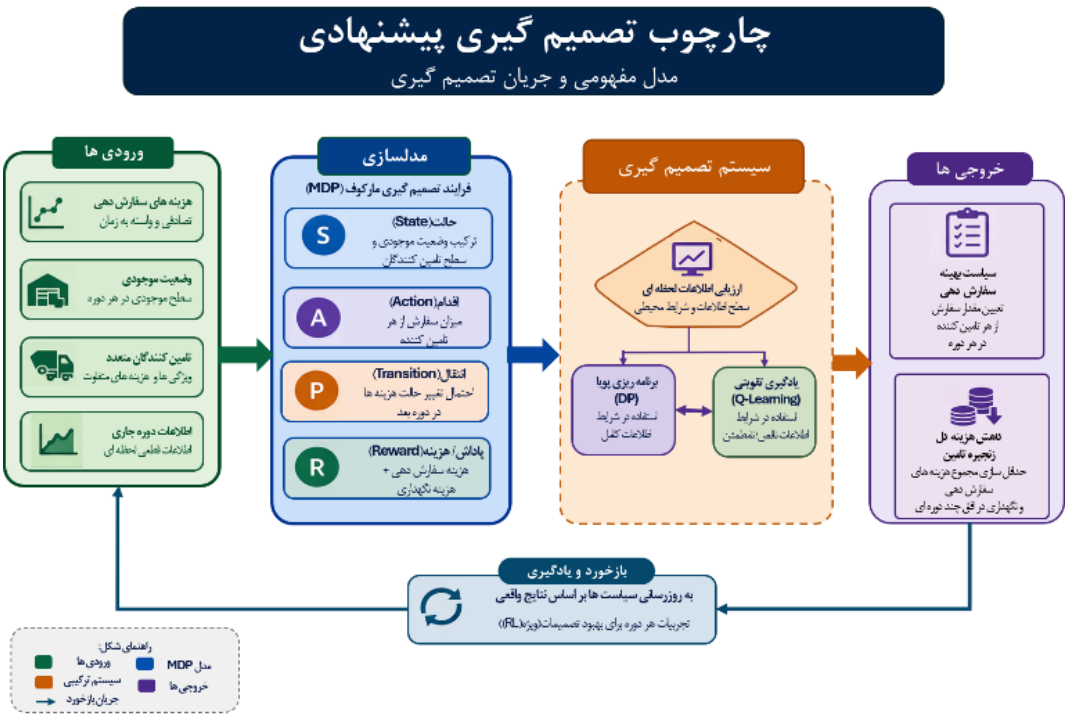
این دو مؤلفه، تمایز اساسی رویکرد پویای پژوهش حاضر را با رویه‌های ایستا و ازپیش‌تعیین‌شده شکل می‌دهند. در واقع، آنچه یک مدل را «پویا» می‌سازد، نه صرفاً تغییر پارامترها در طول زمان، بلکه نگاه به جلو و تطبیق‌پذیری در واکنش به اطلاعات جدید است. مدل‌های سنتی معمولاً فاقد این دو ویژگی هستند؛ زیرا یا تصمیمات را مستقل از آینده اتخاذ می‌کنند و یا قابلیت تنظیم مجدد بر اساس تغییرات لحظه‌ای را ندارند.

سیاست فعلی شرکت (سیاست مرجع) مبتنی بر رویه‌های تجربی و قضاوت مدیریتی است و به‌طور صریح در چارچوب یک مدل بهینه‌سازی هزینه طراحی نشده است. ازاین‌رو، ارزیابی عملکرد آن در مقایسه با رویکردهای تحلیلی، امکان سنجش میزان بهبود بالقوه ناشی از به‌کارگیری روش‌های نظام‌مند را فراهم می‌سازد.

در این راستا، دو رویکرد جایگزین توسعه داده شده است؛ برنامه‌ریزی پویا به‌عنوان روش مدل‌محور که با فرض دسترسی به ساختار احتمالاتی سیستم (ماتریس‌های انتقال استخراج‌شده از داده‌های تاریخی)، سیاست بهینه را از طریق حل معادلات بلمن استخراج می‌کند و یادگیری تقویتی مبتنی بر Q به‌عنوان روش بدون‌مدل که از طریق تعامل با محیط و بدون نیاز به مدل‌سازی صریح توزیع هزینه‌ها، سیاست تصمیم‌گیری را به‌صورت تدریجی می‌آموزد.

تمرکز اصلی بر مقایسه عملکرد هر یک از این رویکردها نسبت به سیاست مرجع و همچنین مقایسه دوجانبه آن‌ها است.

شکل (۱) نمای شماتیک از ساختار مسئله و جریان اطلاعات را نشان می دهد.



شکل ۱: نمای شماتیک از ساختار مسئله و جریان اطلاعات

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

ارزیابی تجربی در یک محیط عملیاتی واقعی با ویژگی‌هایی نظیر روندهای تورمی، نوسانات ساختاری و تغییر رفتار تأمین‌کنندگان انجام می‌شود. عملکرد سه سیاست از منظر هزینه کل، پویایی بین‌دوره‌ای سفارش‌دهی و الگوی تخصیص میان تأمین‌کنندگان در افق چهاردوره‌ای مقایسه می‌شود. این طراحی، امکان ارزیابی اعتبار بیرونی مدل و کارایی عملی رویکردهای پیشنهادی را فراهم می‌سازد.

اجزای مسئله سفارش‌دهی مطابق با مدل فرایند تصمیم گیری مارکوف در ادامه تشریح می‌گردند.

۲-۲- داده ها و پیش پردازش

داده‌های پژوهش شامل ۱۸۸ مشاهده هفتگی در بازه فروردین ۱۴۰۱ تا مهر ۱۴۰۴ بوده و داده های اصلی آن شامل تقاضا، هزینه‌های سفارش‌دهی از دو تأمین‌کننده و هزینه نگهداری موجودی است. این داده‌ها از یک واحد صنعتی واقعی استخراج شده و مبنای تحلیل تجربی قرار گرفته‌اند.

با توجه به نوسانات تورمی قابل‌توجه، تمامی هزینه‌ها با استفاده از شاخص قیمت مصرف‌کننده و با در نظر گرفتن فروردین ۱۴۰۱ به‌عنوان دوره پایه، به قیمت‌های ثابت تبدیل شدند تا اثرات پولی تورم حذف شود. رابطه تعدیل به‌صورت رابطه (۱) اعمال شده است:

$$\text{قیمت تعدیل شده} = \text{واقعی قیمت} \times \frac{CPI \text{ مبنا}}{CPI \text{ فعلی}} \qquad (۱)$$

که در آن قیمت تعدیل‌شده هزینه پس از حذف اثر تورم (به قیمت ثابت دوره‌ی پایه)، قیمت واقعی هزینه ثبت‌شده در دوره‌ی مورد نظر، CPI مبنا شاخص قیمت مصرف‌کننده در دوره‌ی پایه (فروردین ۱۴۰۱) و CPI فعلی شاخص قیمت مصرف‌کننده در دوره‌ی جاری است.

برای امکان‌پذیرسازی مدل‌سازی مارکوفی، هزینه‌های پیوسته به سه سطح پایین(L) ، متوسط (M) و بالا (H) نگاشت شدند. این گسسته‌سازی، پیش‌نیاز تعریف فضای حالت محدود و محاسبه‌ی ماتریس‌های احتمال انتقال در چارچوب فرآیند تصمیم‌گیری مارکوف است.

برای تعیین آستانه‌های تفکیک سطوح هزینه، از معیار آماری میانه‌ی قدرمطلق انحرافات از میانه استفاده شده است. این معیار در مقایسه با انحراف معیار، به داده‌های پرت و مقادیر دورازاندازه حساسیت کمتری دارد و بنابراین برای داده‌های اقتصادی با نوسانات شدید، انتخاب مناسبی محسوب می‌شود (Leys et al., 2013). مقدار میانه‌ی قدرمطلق انحرافات از میانه به‌صورت رابطه (۲) محاسبه می‌شود:

$$MAD = Median(|x_i - Median(x)| \qquad (2)$$

سپس با استفاده از ضریب تبدیل ۱.۴۸۲۶ (معادل انحراف معیار برای توزیع نرمال)؛ که در رابطه (۳) نشان داده شده ا است، آستانه نوسان قابل‌قبول ε به‌صورت رابطه (۴) به‌دست می‌آید (Leys et al., 2013):

$$\frac{1}{\Phi^{-1}(0.75)} = 1.4826 \qquad (3)$$

و

$$\varepsilon = 1.4826 \times MAD \qquad (4)$$

برای هر تأمین‌کننده s در دوره t، تغییرات درصدی هزینه نسبت به دوره قبل (t-۱) مطابق رابطه (۵) محاسبه می‌گردد:

$$\Delta_t = \frac{C_t - C_{t-1}}{C_{t-1}} \qquad (5)$$

که در آن C_t هزینه سفارش‌دهی در دوره t و C_{t-1} هزینه دوره قبلی است. بر اساس تغییرات نسبی هزینه‌ها و آستانه نوسان ε، سطح هزینه‌ی هر دوره به‌صورت رابطه (۶) تعیین می‌گردد:

$$Level_t = \begin{cases} L & \Delta_t < -\varepsilon \\ M & -\varepsilon \le \Delta_t \le \varepsilon \\ H & \Delta_t > \varepsilon \end{cases} \qquad (6)$$

پس از تعیین سطوح هزینه برای تمام دوره‌ها، به هر سطح گسسته باید یک مقدار عددی نماینده نسبت داده شود تا در محاسبات برنامه‌ریزی پویا به‌کار رود. برای این منظور، از یک پنجره‌ی زمانی لغزان چهار دوره‌ای استفاده می‌شود که شامل دوره‌ی جاری و سه دوره‌ی قبل از آن است. بر روی داده‌های هزینه‌ی واقعی در این پنجره، سه معیار آماری محاسبه می‌گردد: صدک ۰.۳۳ برای سطح پایین، میانگین برای سطح متوسط، و صدک ۰.۶۶ برای سطح بالا. این مقادیر، موقعیت نسبی هزینه‌ها را در بازه‌ی زمانی جاری منعکس می‌کنند و به‌دلیل لغزنده بودن پنجره، اثر روندهای بلندمدت صعودی یا نزولی از فرایند گسسته‌سازی حذف می‌شود. به‌این ترتیب، همواره رابطه‌ی (L <M < H) برقرار بوده و ساختار فضای حالت در طول زمان پایدار می‌ماند.

پس از تعیین سطوح هزینه و مقادیر نماینده، ماتریس‌های احتمال انتقال برای هر تأمین‌کننده به‌صورت جداگانه محاسبه می‌شوند. این ماتریس‌ها، احتمال جابه‌جایی هزینه‌ی سفارش‌دهی از یک سطح به سطح دیگر را در یک دوره نشان می‌دهند. برای محاسبه‌ی ماتریس انتقال تأمین‌کننده s ابتدا فراوانی انتقال‌های مشاهده‌شده بین سطوح مختلف در داده‌های تاریخی شمارش شده و سپس با نرمال‌سازی سطری، احتمالات مطابق با رابطه (۷) به‌دست می‌آیند:

$$P_s(i,j) = \frac{N_s(i,j)}{\sum_j N_s(i,j)} \quad \forall i,j \epsilon \{L,M,H\} \qquad (7)$$

که در آن $N_s(i,j)$ تعداد دفعاتی است که هزینه‌ی تأمین‌کننده s از سطح i به سطح j منتقل شده است.برای نمونه ماتریس احتمال انتقال برای تأمین‌کننده اول در رابطه (۸) نشان داده شده است.

$$P_1(i,j) = \begin{bmatrix} ۰.۰۲۷۷۸ & ۰.۸۰۵۵۵ & ۰.۱۶۶۶۷ \\ ۰.۲۳۷۲۹ & ۰.۵۵۹۳۲ & ۰.۲۰۳۳۹ \\ ۰.۲۱۲۱۲ & ۰.۶۹۶۹۷ & ۰.۰۹۰۹۱ \end{bmatrix} \qquad (8)$$

برای تأمین‌کننده‌ی دوم نیز ماتریس مشابهی به‌همین روش

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

محاسبه می‌شود. این ماتریس‌ها در روش برنامه‌ریزی پویا به‌عنوان ورودی شناخته‌شده و در روش یادگیری تقویتی مبتنی بر Q به‌صورت ضمنی و از طریق تعامل با محیط یاد گرفته می‌شوند.

۲-۳- تعریف فضای حالت

در چارچوب فرآیند تصمیم‌گیری مارکوف، وضعیت سیستم در هر دوره، نمایشی از تمام اطلاعات موردنیاز برای اتخاذ تصمیم بهینه است. در مسئله‌ی پژوهش حاضر، فضای حالت در دوره‌ی t به‌صورت ترکیب سطوح هزینه‌ی دو تأمین‌کننده، موجودی ابتدای دوره و شماره‌ی دوره، مطابق با رابطه (۹) تعریف می‌شود:

$$S_t = \left\{c_i^{(w_i)}, I_t, t\right\} \quad t = ۱, ۲, ۳, ۴ \qquad (9)$$

که در آن $c_i^{(w_i)}$ بیانگر هزینه‌ی نماینده‌ی تأمین‌کننده‌ی i در سطح هزینه‌ی w_i ، I_t میزان موجودی در ابتدای دوره t و t دوره تصمیم‌گیری را نشان می‌دهد.

۲-۴- تعریف فضای اقدام

در هر دوره، تصمیم‌گیرنده حجم کل سفارش و نسبت تخصیص سفارش میان دو تأمین‌کننده را تعیین می‌کند. بر این اساس، فضای اقدام در هر دوره به‌صورت یک زوج مرتب مطابق با رابطه (10) تعریف می‌شود:

$$A_t = (\lambda, X_t) \qquad (10)$$

که در آن X_t حجم کل سفارش در دوره‌ی t، λ سهم تأمین‌کننده اول از سفارش است و سهم تأمین‌کننده دوم برابر با $(\lambda - ۱)$ می‌باشد. λ به صورت گسسته در بازه $[۰,۱]$ با گام ۰.۱ مطابق با (۱۱) تعریف شده است:

$$\lambda \in \{0, 0.1, \ldots, 1\} \qquad (11)$$

براین اساس میزان سفارش از هر تأمین‌کننده به‌صورت روابط (۱۲) و(۱۳) محاسبه می‌شود:

$$x_{۱t} = \lambda \times X_t \qquad (12)$$
$$x_{۲t} = (۱ - \lambda) \times X_t \qquad (13)$$

به‌منظور تضمین معناداری اقتصادی تصمیم‌ها و کنترل پیچیدگی محاسباتی، هر دو متغیر تصمیم X_t و λ به‌صورت گسسته تعریف شده‌اند. ازآنجاکه در سیاست بهینه، مقدار سفارش تنها می‌تواند مقادیر گسسته‌ای از جنس مجموع تقاضای دوره‌های متوالی را اختیار کند، گسسته‌سازی فضای اقدام بر همین مبنا انجام شده و از انتخاب مقادیر پیوسته و فاقد تفسیر اقتصادی اجتناب شده است.

مجموعه‌ی مقادیر مجاز برای حجم سفارش به‌صورت رابطه (۱۴) تعریف می‌شود:

$$\{d_1, d_1 + d_2, d_1 + d_2 + d_3, d_1 + d_2 + d_3 + d_4, d_2 + d_2 + d_3, d_2 + d_3 + d_4, d_3, d_3 + d_4, d_4, 0\} \qquad (14)$$

این گسسته‌سازی امکان مدل‌سازی طیفی از سیاست‌های تخصیص، از تخصیص کامل به یک تأمین‌کننده تا تقسیم متعادل سفارش را فراهم می‌سازد.

۲-۵- تابع پاداش

تابع پاداش، معیار عملکرد سیستم در هر گام تصمیم‌گیری است. ازآنجاکه هدف مدل کمینه‌سازی هزینه است تابع پاداش به‌صورت منفی هزینه کل تعریف‌شده است تا بیشینه‌سازی پاداش معادل با کمینه‌سازی هزینه باشد.. تابع پاداش در دوره t مطابق با معادله (۱۵) است:

$$R(s_t, a_t) = -\left(\sum_{s=۱}^{S} c_s^{(w_s)} x_{st} + h_t I_t\right) \qquad (15)$$

که در آن $c_s^{(w_s)}$ هزینه واحد سفارش‌دهی از تأمین‌کننده‌ی s در سطح هزینه‌ی w_s ، x_{st} میزان سفارش از تأمین‌کننده‌ی s در دوره‌ی t و h_t هزینه‌ی واحد نگهداری است.

۲-۶- تابع انتقال

تابع انتقال، احتمال جابه‌جایی سیستم از یک حالت به حالت دیگر را در اثر اقدام انجام‌شده بیان می‌کند. با توجه به خاصیت مارکوفی، وضعیت آینده S_{t+1} تنها به وضعیت فعلی S_t و اقدام A_t وابسته است و مستقل از تاریخچه‌ی قبل از t می‌باشد. در این پژوهش، این خاصیت بر دو اساس برقرار است. نخست تغییرات هزینه‌های سفارش‌دهی بر اساس فرضیه‌ی گام تصادفی، تنها به سطح هزینه‌ی دوره‌ی جاری وابسته است و ماتریس انتقال مارکوفی، این وابستگی را مدل‌سازی می‌کند؛ دوم تراز موجودی نیز به‌صورت قطعی محاسبه می‌شود که تنها به موجودی و سفارش دوره‌ی جاری وابسته است. اما ازآنجاکه مسئله در یک افق چنددوره‌ای تعریف شده است، برای محاسبه‌ی ارزش مورد انتظار آینده در معادله‌ی بلمن، به احتمال انتقال در چند دوره نیاز داریم. به‌همین دلیل، از توان ماتریس انتقال یا همان انتقال چندمرحله‌ای مارکوف استفاده می‌شود که احتمال رفتن از سطح i در دوره‌ی j به سطح n در دوره‌ی k (با فاصله‌ی k-j دوره) را محاسبه می‌کند. احتمال اینکه هزینه‌ی تأمین‌کننده s در u دوره‌ی آینده از سطح i به سطح j منتقل شود (ماتریس انتقال مرتبه u)، از رابطه‌ی (۱۶) به‌دست می‌آید:

$$(p_s)_{ij}^{u} = \left(C_{s(t+u)} = C_s^j \middle| C_{st} = C_s^i\right) \qquad (16)$$

که در آن، C_{st} هزینه‌ی سفارش‌دهی تأمین‌کننده‌ی s در دوره‌ی t (که می‌تواند یکی از سطوح L , M یا H باشد)، $C_{s(t+u)}$ هزینه‌ی سفارش‌دهی تأمین‌کننده‌ی s در دوره‌ی u، $(p_s)_{ij}^{u}$ توان u_ام ماتریس انتقال (ماتریس انتقال مرتبه u)، که در واقع احتمال گذار از حالت i به حالت j در طی u گام (دوره) را نشان می‌دهد. است. این توان ماتریس، امکان محاسبه‌ی ارزش مورد انتظار آینده را در معادله‌ی بلمن را فراهم می‌سازد، زیرا جمله‌ی سوم رابطه‌ی (۱۹)، دقیقاً از همین توان ماتریس‌های انتقال برای محاسبه‌ی احتمال انتقال در(k-j) دوره استفاده می‌کند. بنابراین، با دانستن وضعیت جاری S_t و با استفاده از توان ماتریس انتقال برای پیش‌بینی چنددوره‌ای، وضعیت آینده S_{t+u} به‌طور کامل مشخص می‌شود و خاصیت مارکوفی در مدل برقرار است.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

۲-۷- روش های حل

به‌منظور استخراج سیاست بهینه سفارش‌دهی و تخصیص سفارش بین تأمین‌کنندگان، دو رویکرد مکمل الگوریتمی مورد استفاده قرار گرفتند.

برنامه‌ریزی: برای حالت‌هایی که مدل (ماتریس‌های انتقال) به‌صورت کامل از داده‌های تاریخی قابل استخراج است و به‌عنوان معیار مرجع بهینه عمل می‌کند.

یادگیری تقویتی: برای شرایط واقعی و مدل‌آزاد که در آن توابع انتقال یا توزیع هزینه‌ها به‌طور دقیق شناخته‌شده نیستند و عامل باید از طریق تعامل با محیط یاد بگیرد.

۲-۷-۱- حل مدل با برنامه ریزی پویا

برنامه‌ریزی پویا یکی از روش‌های کلاسیک برای حل مسائل تصمیم‌گیری چندمرحله‌ای است که در آن وضعیت سیستم از خاصیت مارکوف برخوردار است (Bellman, 1957). در این پژوهش، با فرض شناخت کامل مدل (دسترسی به ماتریس‌های انتقال استخراج‌شده از داده‌های تاریخی)، از معادله‌ی بلمن برای محاسبه‌ی تابع ارزش و استخراج سیاست بهینه استفاده شده است.

نسخه خاص‌شده تابع ارزش نهایی و رابطه بازگشتی بلمن برای مدل دوتأمین‌کننده (S=۲)با هزینه‌های تصادفی مارکوفی مطابق معادلات (۱۷) و (۱۸) تعریف شده است.

$$V_T\left(\mathrm{c}_1^{(\mathrm{w}_1)}, \mathrm{c}_2^{(\mathrm{w}_2)}, \mathrm{I_j}, j\right) = 0 \tag{17}$$

$$V_j\left(\mathrm{c}_1^{(\mathrm{w}_1)}, \mathrm{c}_2^{(\mathrm{w}_2)}, \mathrm{I_j}, \mathrm{j}\right) = \min_{j<k\leq T}\left\{\begin{array}{c}\left[\lambda \mathrm{c}_1^{(\mathrm{w}_1)} + \left((1-\lambda)\mathrm{c}_2^{(\mathrm{w}_2)}\right)\right] \times \sum_{r=j+1}^{k} d_r \\ + \sum_{t=j+1}^{k-1} h_t\left(\sum_{r=t+1}^{k} d_r\right) \\ + \sum_{\mathrm{m}=1}^{\mathrm{w}_1}\sum_{\mathrm{n}=1}^{\mathrm{w}_2} (P_1)_{\mathrm{w}_1 m}^{(k-j)} \times (P_2)_{\mathrm{w}_2 \mathrm{n}}^{(k-j)} \times V_k(\mathrm{c}_1^m, \mathrm{c}_2^{\mathrm{n}}, \mathrm{I_k}, \mathrm{k})\end{array}\right\} \tag{18}$$

$$i = 1,2,\dots,\mathrm{w}_1\ , l = 1,2,\dots,\mathrm{w}_2$$

,

$$\mathrm{I}_{T-1} + \sum_{s=1}^{S}\lambda_s \mathrm{x}_{sT} - d_t = 0 \quad (t = 2,3,\dots,T-1) \tag{21}$$

$$\mathrm{I}_t \geq 0\ ,\ \mathrm{x}_{st} \geq 0 \quad (s = 1,2), (t = 1,2,\dots,T-1) \tag{22}$$

در این رابطه λ_s سهم تخصیص یافته به تأمین‌کنندهs ام، $\mathrm{c}_s^{(\mathrm{w}_s)}$ هزینه سفارش‌دهی برای تأمین‌کنندهs ام در سطح هزینه w_s، $(P_s)_{\mathrm{w}_s \mathrm{n}_s}^{(k-j)}$ احتمال انتقال از سطح هزینه w_s به سطح n_s برای تأمین‌کننده s ام در فاصله زمانی k-j ، d_r تقاضا در دورهr و h_t هزینه نگهداری در دورهt است.

رابطه‌ی (۱۸) به‌صورت پویا از دوره‌ی آخر T به دوره‌ی اول حل می‌شود. در هر دوره‌ی j، برای تمام ترکیبات ممکن سطوح هزینه‌ی دو تأمین‌کننده و سطوح موجودی، مقدار بهینه‌ی تابع ارزش محاسبه می‌شود. جمله‌ی اول نشان‌دهنده‌ی هزینه‌ی سفارش‌دهی، جمله‌ی دوم هزینه‌ی نگهداری موجودی، و جمله‌ی سوم ارزش مورد انتظار آینده است که با استفاده از توان ماتریس‌های احتمال انتقال P_1 و P_2 محاسبه می‌شود. خروجی این فرایند، تابع ارزش V_j برای هر حالت، که حداقل هزینه‌ی قابل‌دستیابی را نشان می‌دهد و سیاست بهینه که برای هر حالت، تعیین بهترین اقدام است، می باشد.

برای تضمین عملیاتی بودن مدل، محدودیت‌های تراز موجودی (معادلات ۱۹ تا ۲۲) اعمال می‌شوند(Azaron et al., 2009):

$$\sum_{s=1}^{S}\lambda_s \mathrm{x}_{s1} - \mathrm{d}_1 = \mathrm{I}_1 \tag{19}$$

$$\mathrm{I}_{t-1} + \sum_{s=1}^{S}\mathrm{x}_{st}\ \mathrm{d}_t = \mathrm{I}_t\ \ (s = 1,2), (t = 1,2,\dots,T-1) \tag{20}$$

۲-۷-۱- حل مدل با یادگیری تقویتی مبتنی بر Q

الگوریتم یادگیری تقویتی مبتنی بر Q یک روش یادگیری تقویتی بدون مدل است که در آن عامل (تصمیم‌گیرنده) بدون داشتن هیچ اطلاعاتی از ماتریس‌های احتمال انتقال یا تابع پاداش، از طریق تعامل مکرر با محیط، مقادیر Q را به‌روزرسانی می‌کند (Sutton and Barto, 2018, Watkins and Dayan, 1992).

تابع Q(s, a) ارزش مورد انتظار انجام اقدام a در حالت s و سپس پیروی از سیاست بهینه را نشان می‌دهد و به‌صورت رابطه (۲۳) تعریف می‌شود:

$$Q\left(\mathrm{s}_t, \mathrm{a}_t\right) = \mathbb{E}\left[\mathrm{R}_t + \gamma \max_a Q(\mathrm{s}_{t+1}, \dot{a})\right] \tag{23}$$

که در آن γ ضریب تنزیل برای لحاظ پاداش‌های آینده، $\max_a Q(\mathrm{s}_{t+1}, \dot{a})$ تخمین بهترین پاداش آینده از حالت بعدی و R_t پاداش دریافتی در گام t است.

تابع Q از طریق رابطه‌ی تکراری (۲۴) به‌روزرسانی می‌شود:

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[R_t + \gamma . \max_{á}(s_{t+1}, á) - Q(s_t, a_t)\right] \quad (24)$$

که در آن α نرخ یادگیری عددی بین ۰ و ۱(در این پژوهش ۰.۱= α) و عبارت داخل کروشه، خطای یادگیری نامیده می‌شود و نشان‌دهنده اختلاف بین تخمین فعلی و پاداش واقعی مشاهده‌شده است.

در فرایند یادگیری، در هر اپیزود (یک اجرای کامل از ابتدا تا انتهای افق ۴ دوره‌ای)، عامل از وضعیت اولیه شروع کرده و در هر دوره:

- با استفاده از سیاستε - greedy اقدام a_t را انتخاب می‌کند (با احتمال ε اقدام تصادفی برای اکتشاف و با احتمال $1-\varepsilon$) بهترین اقدام بر اساس Q فعلی)

- اقدام را اعمال کرده، پاداش R_t را از محیط دریافت می‌کند و به وضعیت جدید s_{t+1} منتقل می‌شود

- مقدار Q را مطابق رابطه (۲۴) به‌روز می‌کند

مقدار ε را به‌تدریج کاهش می‌دهد تا تعادل میان اکتشاف و بهره‌برداری برقرار شود.

پس از طی ۲۰،۰۰۰ اپیزود، تابع $Q(s, a)$ به‌تدریج به مقدار بهینه همگرا می‌شود. سیاست بهینه بر اساس این تابع طبق رابطه (۲۵) به‌دست می‌آید:

$$\pi^*(s_t) = \arg\min_{a \in A(s)} Q(s_t, a_t) \quad (25)$$

تابع پاداش در دوره t مطابق با معادله‌های(۲۶) است:

$$R(s_t, a_t) = -\left[\left[\lambda c_{1t}^{(w_1)} + (1-\lambda) c_{2t}^{(w_2)}\right] \times \sum_{r=j+1}^{k} d_r + + \sum_{t=j+1}^{k-1} h_t\left(\sum_{r=t+1}^{k} d_r\right)\right] \quad (26)$$

برخلاف برنامه‌ریزی پویا که نیاز به محاسبه‌ی مستقیم ماتریس‌های انتقال و حل معادلات بلمن دارد، یادگیری تقویتی مبتنی بر Q این توزیع‌ها را به‌صورت ضمنی و از طریق نمونه‌برداری از محیط می‌آموزد. این ویژگی، آن را برای شرایطی که توزیع هزینه‌های آینده به‌طور کامل شناخته‌شده نیست یا دستخوش تغییرات می‌شود، مناسب‌تر می‌سازد. در این پژوهش، یادگیری تقویتی مبتنی بر Q به‌عنوان رویکردی مکمل برای برنامه‌ریزی پویا به‌کار رفته است تا نشان داده شود که در شرایط واقعی و با اطلاعات ناقص، این روش می‌تواند به سیاستی نزدیک به بهینه همگرا شود. این مقایسه، هسته‌ی اصلی نوآوری پژوهش را تشکیل می‌دهد.

پارامترهای یادگیری شامل نرخ یادگیری $\alpha = ۰.۱$، ضریب تنزیل $\gamma = ۱$ و تعداد اپیزودها برابر با۲۰۰۰۰ به صورت تجربی تنظیم شدند. با اجرای سیاست ε- greedy ، مقدار ε از ۰.۹۹۵شروع شد و به‌تدریج کاهش یافت تا در پایان فرایند، الگوریتم عمدتاً از سیاست بهره‌برداری پیروی کند.

برای ارزیابی و مقایسه‌ی سیاست‌های حاصل از روش‌های مختلف، معیارهای زیر مورد استفاده قرار گرفته‌اند:

- هزینه‌ی کل مورد انتظار در پایان افق برنامه‌ریزی (شامل هزینه‌های سفارش‌دهی و نگهداری موجودی)

- میزان تخصیص سفارش بین تأمین‌کنندگان و الگوی تخصیص در طول افق

- انحراف معیار هزینه‌ها به‌عنوان معیار پایداری و ریسک‌پذیری سیاست

- درصد بهبود نسبت به سیاست مرجع (سیاست فعلی شرکت)

این معیارها امکان ارزیابی هم‌زمان کارایی اقتصادی، پایداری و قابلیت اطمینان تصمیمات را فراهم می‌سازند.

۳- نتایج و بحث

به‌منظور ارزیابی عملکرد چارچوب پیشنهادی، مدل با استفاده از داده‌های واقعی یک واحد صنعتی فعال در حوزه تولید ظروف یکبار مصرف پلیمری اعتبارسنجی گردید. سه سیاست مورد مقایسه قرار گرفتند. این سیاست‌ها شامل سیاست فعلی شرکت به‌عنوان سیاست مرجع، سیاست مبتنی بر برنامه‌ریزی پویا ، و سیاست مبتنی بر یادگیری تقویتی مبتنی بر Q بودند. افق برنامه‌ریزی چهار دوره و تعداد تأمین‌کنندگان دو عدد در نظر گرفته شد. معیارهای ارزیابی شامل هزینه کل شامل هزینه سفارش‌دهی و نگهداری موجودی، درصد بهبود نسبت به سیاست مرجع، و پایداری عملکرد در ۴۷ افق تصمیم‌گیری بود.

۳-۱- مقایسه هزینه کل سیاست ها

نتایج مقایسه هزینه کل سه سیاست نشان می‌دهد که در تمامی ۴۷ افق برنامه‌ریزی، هزینه حاصل از برنامه‌ریزی پویا و یادگیری تقویتی مبتنی بر Q کمتر یا مساوی سیاست مرجع (موجود) بوده است. میانگین صرفه‌جویی برای برنامه‌ریزی پویا برابر ۰.۲۳ درصد و برای یادگیری تقویتی مبتنی بر Q برابر ۰.۲۱ درصد است. اگرچه این ارقام ممکن است از نظر درصدی اندک به نظر برسند، اما با توجه به حجم عظیم هزینه‌های زنجیره تأمین در عمل (ارقام میلیاردی)، حتی این بهبود ظاهراً کوچک به صرفه‌جویی مطلق بسیار بزرگی منجر می‌شود میانگین حدود ۳۱.۸ میلیون واحد پولی در هر افق). برای نمونه، در برخی افق‌ها صرفه‌جویی مطلق بیش از ۱۲۴ میلیون ریال بوده است که برای هر بنگاه اقتصادی قابل توجه است. حداکثر درصد بهبود در برخی افق‌ها به ۰.۸۳ درصد رسیده که معادل صدها میلیون ریال صرفه‌جویی است. همچنین، در سناریوهای با اختلاف هزینه‌ی شدید بین تأمین‌کنندگان (سناریوی۶ در بخش تحلیل حساسیت)، این صرفه‌جویی به بیش از ۴۰٪ افزایش می‌یابد

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

که بیانگر پتانسیل بالای چارچوب در صنایع با نوسانات شدید قیمتی است.

نکته قابل‌توجه، نزدیکی بسیار زیاد عملکرد یادگیری تقویتی مبتنی بر Qبه برنامه‌ریزی پویا است. در بسیاری از افق‌ها، هزینه‌ی کل این دو روش دقیقاً برابر بوده است (تفاوت کمتر از ۰.۰۱٪ در میانگین هزینه). این یافته تأیید می‌کند که رویکرد داده‌محور پیشنهادی با وجود عدم نیاز به مدل‌سازی توزیع هزینه‌های آینده، عملکردی بسیار نزدیک به روش بهینه‌ی مبتنی بر مدل برنامه ریزی پویا ارائه می‌دهد.

جدول (۱): نمونه ای از مقایسه هزینه کل سه سیاست در طول ۴۷ افق تصمیم‌گیری

افق	هزینه کل (سیاست فعلی)	هزینه کل (برنامه‌ریزی پویا)	هزینه کل (یادگیری تقویتی مبتنی برQ)	صرفه‌جویی (فعلی - DP)	صرفه‌جویی (فعلی - QL)	تفاوت (QL - DP)
۱	۷۲۳۵۰۰۰	۷۱۹۲۱۵۰	۷۲۲۱۲۰۰	۴۲۸۵۰	۱۳۸۰۰	۲۹۰۵۰
۲	۱۱۴۱۲۳۰۰	۱۱۲۹۴۵۰۰	۱۱۲۸۸۲۰۰	۱۱۷۸۰۰	۱۲۴۱۰۰	۶۳۰۰-
۳	۱۷۶۵۶۷۰۰	۱۷۶۰۴۶۰۰	۱۷۶۱۴۰۰۰	۵۲۱۰۰	۴۲۷۰۰	۹۴۰۰
۴	۷۲۳۵۰۰۰	۷۱۹۲۱۵۰	۷۲۲۱۲۰۰	۴۲۸۵۰	۱۳۸۰۰	۲۹۰۵۰
.	…	…	…	…	…	…
۴۶	۵۰۱۶۲۱۰۰	۵۰۱۳۸۱۰۰	۵۰۱۳۸۱۰۰	۲۴۰۰۰	۲۴۰۰۰	۰
۴۷	۴۹۲۸۲۴۰۰	۴۹۲۴۷۰۰۰	۴۹۲۴۷۰۰۰	۳۵۴۰۰	۳۵۴۰۰	۰

جدول (۲) آمار توصیفی هزینه‌های سه سیاست را نشان می‌دهد. طبق جدول (۲)، میانگین هزینه سیاست برنامه‌ریزی پویا حدود ۷۳,۱۵۰ واحد و سیاست یادگیری تقویتی مبتنی بر Q حدود ۶۶,۵۹۱ واحد کمتر از سیاست مرجع است. انحراف معیار هزینه‌ها در سه سیاست تقریباً یکسان است. این موضوع نشان‌دهنده پایداری مشابه آن‌ها از نظر پراکندگی هزینه‌ها است. لذا هر دو روش بهینه‌سازی عملکرد بهتری از سیاست فعلی دارند.

جدول (۲): آمار توصیفی هزینه سیاست ها

سیاست ها	میانگین هزینه ها	حداکثر هزینه ها	حداقل هزینه ها	انحراف معیار هزینه ها
هزینه کل (سیاست فعلی)	۳۱,۸۰۲,۰۷۰	۵۰,۱۶۲,۱۰۰	۷,۲۳۵,۰۰۰	۹,۸۲۴,۳۶۵
هزینه کل (برنامه‌ریزی پویا)	۳۱,۷۲۸,۹۲۰	۵۰,۱۳۸,۱۰۰	۷,۱۹۲,۱۵۰	۹,۸۲۳,۷۱۸
هزینه کل (یادگیری تقویتی مبتنی بر Q)	۳۱,۷۳۵,۴۷۹	۵۰,۱۳۸,۱۰۰	۷,۲۲۱,۲۰۰	۹,۸۲۴,۳۰۴

۳-۲- تحلیل تخصیص سفارش به تامین کنندگان

تحلیل تخصیص سفارش به تأمین‌کنندگان نشان می‌دهد که سه سیاست مورد مقایسه (سیاست مرجع، برنامه‌ریزی پویا و یادگیری تقویتی مبتنی بر Q) الگوهای رفتاری متفاوتی دارند. همان‌طور که در جدول (۳) مشاهده می‌شود، سیاست مرجع رویکردی نسبتاً پایدار و از پیش‌تعیین‌شده را دنبال می‌کند. در مقابل، برنامه‌ریزی پویا ، تخصیص را به‌صورت حالت‌وابسته و متناسب با شرایط هزینه‌ای هر دوره انجام می‌دهد. سیاست یادگیری تقویتی مبتنی بر Q نیز، در بسیاری از افق‌ها الگوی مشابه برنامه‌ریزی پویا را بازتولید کرده و گاهی تخصیص ترکیبی میان دو تأمین‌کننده داشته است. به‌طور کلی، رفتار دو سیاست داده‌محور (برنامه‌ریزی پویا و یادگیری تقویتی مبتنی بر Q) نشان‌دهنده وابستگی تصمیم تخصیص به شرایط هزینه‌ای هر دوره است، در حالی که سیاست مرجع انعطاف‌پذیری کمتری در واکنش به تغییر سطوح هزینه دارد. می توان گفت درزمینهٔ مدیریت ریسک هزینه‌ای، دو سیاست داده‌محور در شرایطی که یک تأمین‌کننده دارای مزیت هزینه‌ای مشخص است، تمرکز سفارش را به سمت همان تأمین‌کننده سوق می‌دهند و از مزیت نسبی بهره‌برداری می‌کنند، درحالی‌که سیاست فعلی گرایش کمتری به چنین تمرکزی دارد. درمجموع، استفاده از چارچوب‌های تصمیم‌گیری پویای مبتنی بر داده می‌تواند فرآیند تخصیص را از یک رویه تجربی و شهودی به یک قانون تصمیم نظام‌مند و بهینه تبدیل کند.

[Downloaded from ieijqp.ir on 2026-09-12]

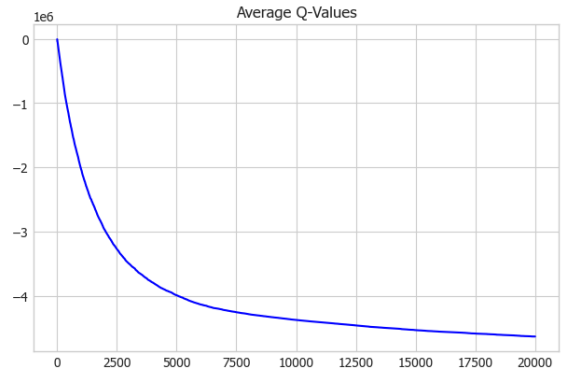
[DOR: 20.1001.1.23222344.1405.15.1.5.5]

جدول (۳): میزان و درصد تخصیص سفارش به تأمین‌کنندگان در هر دوره

افق	دوره	تقاضا	R_x۱	R_x۲	DP_x۱	DP_x۲	QL_x۱	QL_x۲	R_x۱	R_x۲	DP_x۱	DP_x۲	QL_x۱	QL_x۲
۱	۱	۷۱	۳۶	۳۵	۰	۱۵۴	۰	۷۱	%۵۱	%۴۹	%۰	%۱۰۰	%۰	%۱۰۰
۱	۲	۸۳	۰	۸۳	۰	۰	۵۸	۲۵	%۰	%۱۰۰	%۰	%۰	%۷۰	%۳۰
۱	۳	۳۳	۲۰	۱۳	۰	۳۳	۰	۳۳	%۶۱	%۳۹	%۰	%۱۰۰	%۰	%۱۰۰
۱	۴	۴۶	۴۶	۰	۰	۴۶	۰	۴۶	%۱۰۰	%۰	%۰	%۱۰۰	%۰	%۱۰۰
…	…	…	…	…	…	…	…	…						
۴۷	۱	۱۳۲	۸۰	۵۲	۱۳۲	۰	۱۳۲	۰	%۶۱	%۳۹	%۱۰۰	%۰	%۱۰۰	%۰
۴۷	۲	۱۴۲	۸۰	۶۲	۰	۱۴۲	۰	۱۴۲	%۵۶	%۴۴	%۰	%۱۰۰	%۰	%۱۰۰
۴۷	۳	۱۴۲	۸۰	۶۲	۰	۱۴۲	۰	۱۴۲	%۵۶	%۴۴	%۰	%۱۰۰	%۰	%۱۰۰
۴۷	۴	۱۴۲	۸۰	۶۲	۱۴۲	۰	۱۴۲	۰	%۵۶	%۴۴	%۱۰۰	%۰	%۱۰۰	%۰

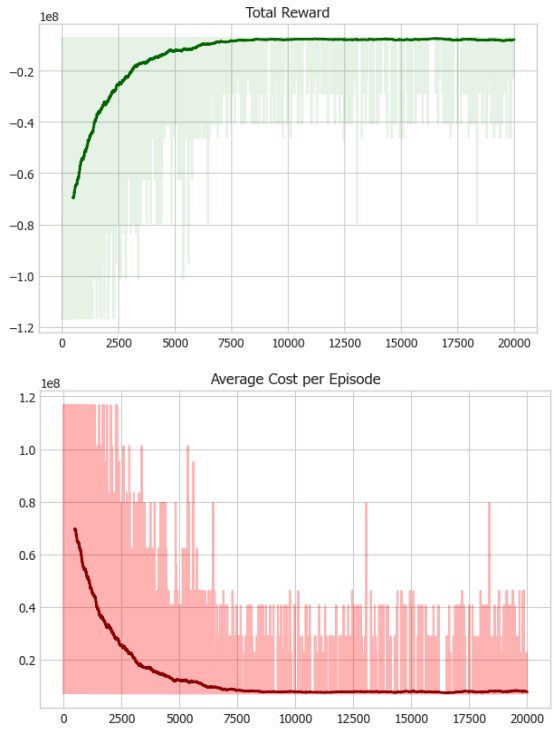
۳-۳- تحلیل همگرایی و پایداری یادگیری تقویتی

بررسی روند یادگیری الگوریتم یادگیری تقویتی مبتنی بر Q، پایداری عددی و همگرایی سیاست استخراج‌شده را تأیید می‌کند. همان‌طور که در شکل (۲) مشاهده می‌شود، میانگین مقادیر Q در مراحل ابتدایی آموزش با شیب تندی کاهش یافته و پس از حدود ۱۰،۰۰۰ اپیزود به سطحی پایدار همگرا شده است. عدم وجود نوسانات شدید در مراحل پایانی، نشان‌دهنده ورود الگوریتم به ناحیه تثبیت و سازگاری با نرخ یادگیری تنظیم‌شده است.



شکل ۲ نمودار روند تغییرات میانگین مقادیر Q در طول اپیزودها

همچنین، شکل (۳) روند هزینه تجمعی (پاداش منفی) را نشان می‌دهد. اگرچه در اپیزودهای اولیه به دلیل اکتشاف گسترده فضای عمل، نوسانات زیاد بود، اما با پیشرفت آموزش، مسیر تغییرات هموارتر شده و هزینه متوسط به‌صورت نزولی همگرا گردید. این رفتار بیانگر کاهش تصمیم‌های پرهزینه و تثبیت سیاست بهینه است.

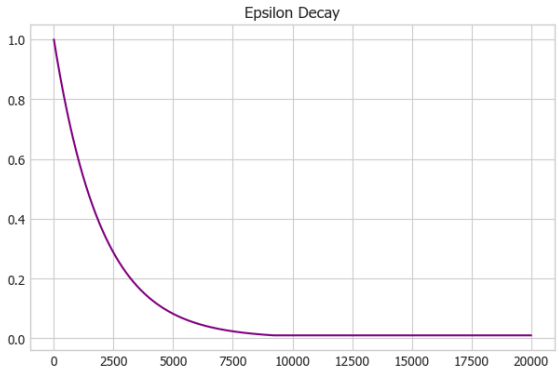


شکل ۳: روند تغییرات هزینه تجمعی (پاداش) در طول فرآیند آموزش

مدیریت تعادل اکتشاف-بهره‌برداری نیز در شکل (۴) تحلیل شده است. پارامتر ε از مقدار ۱ آغاز شده و به‌صورت نمایی به سمت صفر میل کرده است. این کاهش هم‌زمان با تثبیت مقادیر Q (شکل ۱) و کاهش واریانس، نشان‌دهنده گذر موفقیت‌آمیز الگوریتم از فاز اکتشاف گسترده به فاز بهره‌برداری از دانش کسب‌شده است.

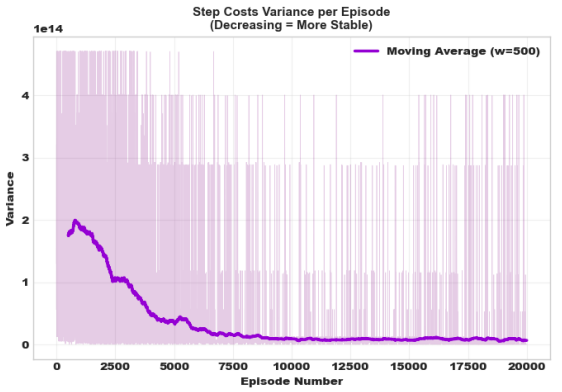
[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]



شکل ۴: تحلیل اکتشاف–بهره‌برداری

در نهایت، شکل (۵) روند تغییرات واریانس و انحراف معیار هزینه را نمایش می‌دهد. واریانس در مراحل اولیه به دلیل تنوع در انتخاب اعمال در سطوح بالا قرار داشت، اما با پیشرفت آموزش و خروج از فاز اکتشاف، به‌طور معناداری کاهش یافت. همگرایی هم‌زمان میانگین هزینه، کاهش انحراف معیار و افت واریانس، حاکی از آن است که سیاست نهایی نه‌تنها از نظر کارایی بهینه شده، بلکه از نظر ریسک و پراکندگی نیز به سطحی قابل قبول و پایدار رسیده است که این ویژگی برای کاربردهای عملی در زنجیره تأمین حیاتی است.



شکل ۵: تحلیل واریانس و انحراف معیار هزینه در طول فرآیند یادگیری

به طور کلی ارزیابی مدل در صنایع پلیمر و فولاد نشان می‌دهد که سیاست‌های مبتنی بر برنامه‌ریزی پویا و یادگیری Q، عملکردی برتر از سیاست مرجع دارند. میانگین صرفه‌جویی هزینه به‌ترتیب ۰.۲۳٪ برای برنامه‌ریزی پویا و ۰.۲۱٪ برای یادگیری Q برآورد شد، هرچند در برخی افق‌ها این میزان به ۰.۸۳٪ نیز رسیده است (جدول ۳). در برخی افق‌های برنامه‌ریزی، این میزان بهبود به ۰.۸۳٪ رسید. با توجه به مقیاس عملیاتی، این درصدها منجر به صرفه‌جویی میلیاردی ریالی (بیش از ۱۲۴ میلیارد ریال در برخی افق‌ها) می‌شوند که از نظر اقتصادی کاملاً توجیه‌پذیر است.

نکته کلیدی این پژوهش آن است که الگوریتم یادگیری Q، با وجود عدم نیاز به مدل‌سازی توزیع هزینه‌های آینده، عملکردی بسیار نزدیک به برنامه‌ریزی پویا (با اختلاف کمتر از ۰.۰۱٪ در میانگین هزینه) ارائه داده است. این یافته تأیید می‌کند که رویکرد داده‌محور پیشنهادی، انعطاف‌پذیری بالایی در مواجهه با نوسانات هزینه دارد و می‌تواند به‌عنوان یک سیستم پشتیبان تصمیم‌گیری مؤثر در محیط‌های با اطلاعات ناقص مورد استفاده قرار گیرد.

۳-۴- تحلیل حساسیت ساختاری

تحلیل حساسیت به‌منظور ارزیابی پایداری مدل پیشنهادی و بررسی انطباق نتایج با منطق اقتصادی، بر روی داده‌های واقعی و تحت شش سناریوی متفاوت هزینه‌ای انجام شده است. این تحلیل برای هر دو رویکرد برنامه‌ریزی پویا و یادگیری تقویتی مبتنی بر Qصورت گرفته است.

سناریوی ۱ (شرایط پایه): در حالت عادی، سیاست بهینه هر دو روش عمدتاً به سمت تأمین‌کننده دوم متمایل است (تخصیص ۶۱ درصد به تأمین‌کننده دوم در برنامه‌ریزی پویا و ۶۷ درصد در یادگیری تقویتی). برنامه‌ریزی پویا رفتار قطعی‌تری نشان می‌دهد، در حالی که یادگیری تقویتی مبتنی بر Q دارای نوسانات بیشتری است. هزینه‌های نهایی دو روش بسیار نزدیک به یکدیگر است.

سناریوی ۲ (دو برابر شدن هزینه تأمین‌کننده اول): با افزایش اختلاف هزینه‌ای، هر دو روش به‌طور کامل به سمت تأمین‌کننده ارزان‌تر (دوم) تغییر جهت می‌دهند (تخصیص ۱۰۰ درصد). برنامه‌ریزی پویا به‌صورت قطعی (λ=۰) عمل می‌کند، در حالی که یادگیری تقویتی با وجود نوسانات اولیه، جهت‌گیری صحیح اقتصادی را تشخیص می‌دهد.

سناریوی ۳ (حذف هزینه نگهداری): در غیاب هزینه نگهداری، برنامه‌ریزی پویا اقدام به پیش‌سفارش و انباشت موجودی می‌کند (تخصیص ۶۱ درصد به تأمین‌کننده دوم). یادگیری تقویتی اگرچه رفتار مشابهی نشان می‌دهد، اما شدت انباشت کمتر است (تخصیص ۵۷ درصد به تأمین‌کننده دوم). این تفاوت ناشی از ماهیت تقریبی و نیاز به اپیزودهای بیشتر برای همگرایی کامل است.

سناریوی ۴ (برابری هزینه تأمین‌کنندگان): در شرایط بی‌تفاوتی هزینه‌ای، برنامه‌ریزی پویا به تخصیص کاملاً متقارن (λ=۰.۵) همگرا می‌شود (۵۰-۵۰). یادگیری تقویتی نوسانات بیشتری نشان می‌دهد (۵۳.۵ درصد به تأمین‌کننده اول)، اما هزینه نهایی نزدیک به برنامه‌ریزی پویا است.

سناریوی ۵ (افزایش شدید هزینه نگهداری): با افزایش چشمگیر هزینه نگهداری، هر دو روش به سیاست سفارش برابر با تقاضای همان دوره روی می‌آورند. موجودی پایان دوره تقریباً صفر می‌شود و هزینه‌های نهایی یکسانی ثبت می‌گردد.

سناریوی ۶ (اختلاف هزینه‌ی بیشتر از قبل بین تأمین‌کنندگان): در این سناریو، اختلاف هزینه‌ی تأمین‌کنندگان به‌گونه‌ای افزایش می‌یابد که یک تأمین‌کننده به‌طور قابل‌توجهی ارزان‌تر از دیگری می‌شود.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

به‌طور کلی، نتایج نشان می‌دهد که برنامه‌ریزی پویا از پایداری ساختاری بالاتر و رفتار قطعی‌تری برخوردار است. یادگیری تقویتی مبتنی بر Q با وجود ماهیت اکتشافی و تصادفی، در تمامی سناریوها به جواب‌های بسیار نزدیک یا مشابه با جواب بهینه دست یافته است. میزان تخصیص به تأمین‌کنندگان در دو روش در اکثر موارد هم‌راستا است، هرچند شدت تمایل در یادگیری تقویتی بیشتر متغیر است. به‌طور کلی، سیاست یادگیری تقویتی مبتنی بر Q توانسته است کارایی بالایی از خود نشان دهد و با وجود پراکندگی بیشتر در تخصیص‌ها، به همگرایی رفتاری قابل‌قبولی با روش بهینه برسد.

۳-۵- تحلیل حساسیت پارامترهای یادگیری تقویتی

در این بخش، تأثیر پارامترهای کلیدی یادگیری شامل نرخ یادگیری (α)، ضریب تنزیل (γ)، نرخ کاهش اکتشاف (ε-decay) و تعداد اپیزودها بر همگرایی و پایداری سیاست بررسی شده است. سناریوی پایه($\alpha = ۰.۱,\ \gamma = ۱, \varepsilon - decay =$ ۰.۹۹۹۵ و ۲۰۰۰اپیزود): با ضریب تنزیل برابر با یک، عامل کل پاداش‌های آتی را بدون تنزیل لحاظ می‌کند. روند کاهش هزینه به‌صورت یکنواخت و پایدار آغاز شده و همگرایی در بازه ۳۰۰۰ تا ۵۰۰۰ اپیزود حاصل می‌شود. پس از این مرحله، نوسانات محدود و کنترل‌شده حول مقدار تعادلی مشاهده می‌شود که بیانگر پایداری سیاست است.

یادگیری سریع ($\alpha = ۰.۵$):افزایش نرخ یادگیری موجب کاهش سریع‌تر هزینه در مراحل ابتدایی می‌شود. با این حال، نوسانات قابل‌توجه‌تری در هزینه و پاداش مشاهده می‌شود که بیانگر افزایش واریانس و کاهش پایداری همگرایی است. این رفتار نشان می‌دهد که همگرایی سریع‌تر حاصل می‌شود، اما سیاست نهایی از نظر پایداری در سطح پایین‌تری قرار می‌گیرد.

یادگیری آهسته ($\alpha = ۰.۰۱$): کاهش نرخ یادگیری منجر به فرآیند به‌روزرسانی تدریجی‌تر مقادیر Q شده است. روند کاهش هزینه با شیب ملایم‌تری انجام می‌شود و همگرایی در بازه زمانی طولانی‌تری حاصل می‌گردد. نوسانات به‌طور محسوسی کاهش یافته و مسیر یادگیری از یکنواختی بیشتری برخوردار است. این الگو نشان‌دهنده افزایش پایداری آماری و کاهش حساسیت به نوسانات تصادفی است.

دیدگاه کوتاه‌مدت ($\gamma = ۰.۵$): با کاهش ضریب تنزیل، عامل تمرکز بیشتری بر پاداش‌های فوری پیدا می‌کند. همگرایی اولیه با سرعت بیشتری رخ می‌دهد، اما مقدار نهایی هزینه در سطح بالاتری تثبیت می‌شود. این رفتار نشان‌دهنده آن است که بی‌توجهی به پیامدهای بلندمدت منجر به اتخاذ تصمیم‌های کوتاه‌بینانه و استخراج سیاستی زیر بهینه می‌شود.

دیدگاه میانمدت ($\gamma = ۰.۹$): با ضریب تنزیل ۰.۹۰، عامل به پاداش‌های آتی و آنی توجه دارد. فرآیند همگرایی نسبت به حالت کوتاه‌مدت کندتر اما پایدارتر است و در نهایت هزینه به سطح پایین‌تری همگرا می‌شود. این نتیجه نشان می‌دهد که کاهش جزئی در توجه به آینده می‌تواند تعادلی میان سرعت و کیفیت همگرایی ایجاد کند.

کاهش سریع اکتشاف ($\varepsilon - decay = ۰.۹۹$): افزایش نرخ کاهش ε موجب می‌شود که عامل سریع‌تر از فاز اکتشاف به بهره‌برداری منتقل شود. کاهش هزینه با سرعت بیشتری انجام شده و همگرایی زودتر حاصل می‌شود. با این حال، مقدار نهایی هزینه در سطح بالاتری تثبیت می‌شود. این رفتار نشان‌دهنده احتمال گیر افتادن در بهینه‌های محلی به دلیل عدم کاوش کافی فضای حالت-عمل است.

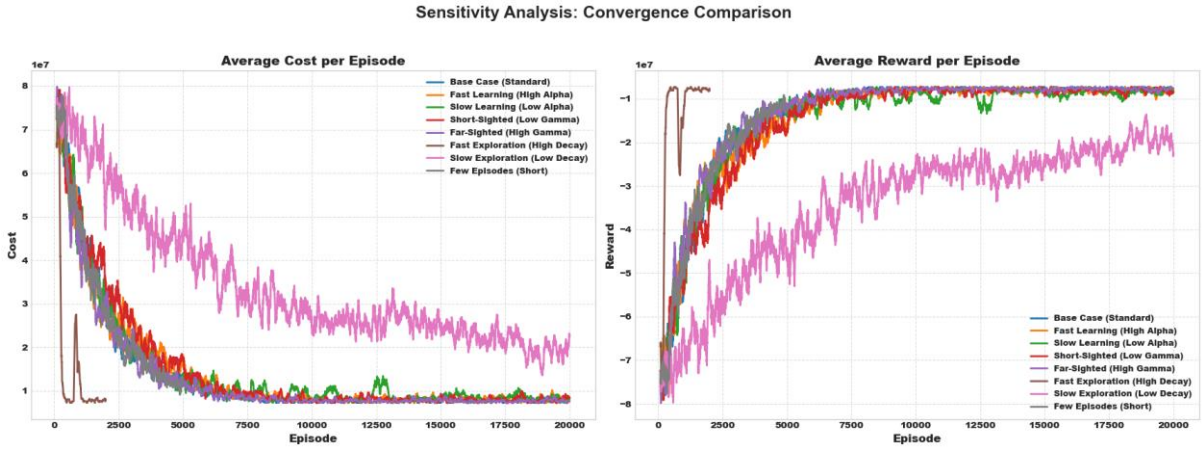
اکتشاف آهسته ($\varepsilon - decay = ۰.۹۹۹۹$): کاهش آهسته ε باعث تداوم فاز اکتشاف در طول بخش عمده‌ای از فرآیند یادگیری می‌شود. نوسانات در مراحل ابتدایی بیشتر بوده و همگرایی با تأخیر همراه است. با این حال، عامل قادر به کشف مسیرهای تصمیم‌گیری بهینه‌تر شده و به هزینه‌های پایین‌تری دست می‌یابد. این الگو نشان‌دهنده اهمیت اکتشاف کافی در جلوگیری از همگرایی زودرس و بهبود کیفیت سیاست نهایی است.

تعداد اپیزود کم (۵۰۰ اپیزود): با محدود شدن تعداد اپیزودها، فرآیند یادگیری پیش از رسیدن به ناحیه تثبیت متوقف می‌شود. روند کاهش هزینه همچنان ادامه‌دار بوده و به مقدار پایدار نرسیده است. سیاست حاصل از نظر عملکرد و پایداری در سطح پایین‌تری قرار دارد. این یافته نشان می‌دهد که کفایت افق آموزش شرط ضروری برای دستیابی به همگرایی است.

به طور کلی می توان گفت عملکرد الگوریتم به‌شدت تحت تأثیر تعامل میان نرخ یادگیری، ضریب تنزیل و سیاست اکتشاف قرار دارد. در نظر گرفتن $\gamma = ۱$ در سناریوی پایه منجر به بیشترین توجه به پاداش‌های بلندمدت و دستیابی به سیاستی پایدارتر شده است. افزایش نرخ یادگیری و کاهش سریع اکتشاف، اگرچه همگرایی را تسریع می‌کند، اما با افزایش نوسانات و کاهش کیفیت سیاست نهایی همراه است. در مقابل، کاهش نرخ یادگیری و حفظ اکتشاف در بلندمدت، پایداری بیشتری ایجاد کرده و امکان دستیابی به راه‌حل‌های باکیفیت‌تر را فراهم می‌سازد. تنظیم بهینه پارامترها مستلزم ایجاد توازن میان سرعت یادگیری، میزان اکتشاف و افق تصمیم‌گیری است که در شکل(۶) قابل مشاهده است.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]



شکل ۶: نتایج حاصل از تغییر پارامترهای یادگیری

۴- نتیجه‌گیری

این پژوهش با هدف توسعه‌ی یک چارچوب تصمیم‌گیری پیشرفته برای مسئله‌ی سفارش‌دهی چندمرحله‌ای در زنجیره‌های تأمین چندتأمین‌کننده تحت عدم‌قطعیت هزینه‌ها انجام شد. برخلاف مطالعات پیشین که اغلب فرض می‌کنند توزیع احتمال هزینه‌ها شناخته‌شده یا قابل‌برآورد از داده‌های تاریخی است، این پژوهش با پذیرش این واقعیت که در بازارهای پرنوسان (نظیر پلیمر و فولاد)، توزیع هزینه‌ها به‌دلیل تغییرات ناگهانی نرخ ارز، سیاست‌های تعرفه‌ای و شوک‌های عرضه، ماهیتی ناپایدار دارد، چارچوبی نوین مبتنی بر فرآیند تصمیم‌گیری مارکوف ارائه داد. این چارچوب با تکیه بر دو رویکرد مکمل، یعنی برنامه‌ریزی پویا به‌عنوان معیار بهینه‌ی مدل‌محور و یادگیری تقویتی مبتنی بر Q به‌عنوان روشی مدل‌آزاد، امکان استخراج سیاست‌های سفارش‌دهی را در شرایطی فراهم می‌سازد که تصمیم‌گیرنده به توزیع پایدار و قطعی هزینه‌های آتی دسترسی ندارد.

یافته‌های تجربی در صنعت پلیمر بر اساس داده‌های واقعی یک واحد صنعتی در ۴۷ افق برنامه‌ریزی نشان داد که هر دو رویکرد برنامه‌ریزی پویا و یادگیری تقویتی مبتنی بر Q، عملکردی برتر از سیاست مرجع (رویه‌های تجربی و قضاوت مدیریتی) ارائه می‌دهند. در شرایط پایه با هزینه‌های نزدیک تأمین‌کنندگان، میانگین صرفه‌جویی هزینه‌ای معادل ۰.۲۳٪ برای برنامه‌ریزی پویا و ۰.۲۱٪ برای یادگیری Q حاصل شد (با کاهش میانگین هزینه‌ای حدود ۷۳ هزار واحد پولی در هر افق) حاصل شد که از نظر آماری معنادار است (p-value<۰،۰۰۱). بااین‌حال، در برخی افق‌ها، حداکثر صرفه‌جویی تا ۰.۸۳٪ نیز مشاهده شده است .نقطه‌ی قوت اصلی چارچوب در تحلیل‌های حساسیت آشکار شد. به‌گونه‌ای که با افزایش اختلاف هزینه‌ای میان تأمین‌کنندگان، پتانسیل صرفه‌جویی به‌طور چشمگیری تا مرز ۴۰٪ افزایش یافت (سناریوی اختلاف هزینه‌ی شدید). این یافته بیانگر آن است که در صنایع با نوسانات شدید قیمتی (نظیر فولاد، پتروشیمی و پلیمر)، ارزش عملی چارچوب به‌طور قابل‌توجهی افزایش می‌یابد و می‌تواند به‌عنوان یک ابزار مؤثر مدیریت ریسک هزینه مورد استفاده قرار گیرد.

تحلیل‌های حساسیت ساختاری نشان داد که مدل پیشنهادی نسبت به تغییرات پارامترهای هزینه‌ای و نگهداری، رفتار منطقی و اقتصادی دارد؛ به‌گونه‌ای که در شرایط برابری هزینه‌ی تأمین‌کنندگان، به تخصیص متقارن (۵۰-۵۰) و در شرایط افزایش شدید هزینه‌ی نگهداری، به سیاست سفارش برابر با تقاضای همان دوره روی می‌آورد. همچنین، تحلیل حساسیت پارامترهای یادگیری تقویتی نشان داد که عملکرد الگوریتم به‌شدت تحت تأثیر تعامل میان نرخ یادگیری، ضریب تنزیل و سیاست اکتشاف قرار دارد. تنظیم بهینه‌ی این پارامترها (نرخ یادگیری ۰.۱، ضریب تنزیل ۱، و کاهش نمایی اکتشاف) منجر به همگرایی پایدار و سیاستی با کیفیت بالا گردید؛ به‌گونه‌ای که افزایش نرخ یادگیری یا کاهش سریع اکتشاف، اگرچه همگرایی را تسریع می‌کند، اما با افزایش نوسانات و کاهش کیفیت سیاست نهایی همراه است.

از منظر پایداری تصمیمات، چارچوب پیشنهادی با کاهش انحراف معیار هزینه‌ها از ۹,۸۲۴,۳۶۵ در سیاست فعلی به ۹,۸۲۳,۷۱۸ در برنامه‌ریزی پویا و ۹,۸۲۴,۳۰۴ در یادگیری تقویتی، نشان داد که تأثیر نامطلوبی بر پراکندگی هزینه‌ها ندارد و در بدترین حالت، عملکردی هم‌تراز با سیاست فعلی دارد. همچنین، شناسایی ۱۰۰٪ سیاست‌های بهینه در مواجهه با شوک‌های قیمتی (با استفاده از مکانیزم تطبیقی پیشنهادی)، قابلیت اطمینان سیستم را در شرایط بحران به‌طور چشمگیری افزایش می‌دهد.

چارچوب ارائه‌شده در این پژوهش می‌تواند به‌عنوان هسته‌ی مرکزی سیستم‌های پشتیبان تصمیم‌گیری در صنایع پرنوسان مانند پتروشیمی، فولاد و داروسازی مورد استفاده قرار گیرد. مدیران زنجیره‌ی تأمین می‌توانند با جایگزینی روش‌های سنتی

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

پیش‌بینی‌محور با سیستم‌های یادگیرنده و تطبیقی، ریسک تصمیم‌گیری در شرایط بحران و نوسانات قیمتی را به‌طور قابل‌توجهی کاهش دهند. پیشنهاد می‌شود سازمان‌ها زیرساخت‌های داده‌ای خود را برای ثبت دقیق و لحظه‌ای هزینه‌های تأمین تقویت کنند تا الگوریتم‌های یادگیری بتوانند از داده‌های تاریخی برای استخراج الگوهای پنهان و بهبود تدریجی سیاست‌های خود بهره ببرند.

۴-۱- پیشنهادها برای مطالعات آینده

اگرچه این پژوهش با محدودیت‌هایی نظیر مدل‌سازی تک‌محصولی و فرض استقلال هزینه از حجم سفارش مواجه بود، اما با توجه به کارایی مدل برای محصولات مستقل، توسعه آن به سمت محصولات وابسته (استراتژیک و مکمل)، ترکیب با یادگیری عمیق ، در نظرگیری هم‌زمان عدم‌قطعیت تقاضا و افزودن قیود عملیاتی نظیر ظرفیت تأمین‌کنندگان و زمان تحویل، به‌عنوان مسیرهای اصلی برای پژوهش‌های آتی پیشنهاد می‌شود تا این رویکرد مقاوم و مقرون‌به‌صرفه به یک سیستم جامع‌تر برای مدیریت زنجیره تأمین در شرایط عدم‌قطعیت شدید ارتقا یابد.

۴-۲- جمع بندی

این پژوهش با هدف توسعه‌ی یک چارچوب تصمیم‌گیری پیشرفته برای مسئله‌ی سفارش‌دهی چندمرحله‌ای در محیط‌های چندتأمین‌کننده و نامطمئن انجام شد. نتایج نشان داد که ترکیب فرآیند تصمیم‌گیری مارکوف، برنامه‌ریزی پویا و یادگیری تقویتی، یک رویکرد مؤثر برای استخراج سیاست‌های بهینه و پایدار فراهم می‌آورد. مدل پیشنهادی توانست با لحاظ نمودن پویایی سیستم، وابستگی بین‌دوره‌ای تصمیمات و عدم‌قطعیت هزینه‌ها، عملکردی برتر نسبت به سیاست‌های سنتی ارائه دهد. یافته‌های این پژوهش نشان داد که الگوریتم‌های داده‌محور مانند یادگیری Q، با وجود عدم نیاز به مدل‌سازی دقیق توزیع هزینه‌های آینده، می‌توانند به عملکردی نزدیک به روش‌های بهینه‌ی مبتنی بر مدل دست یابند و در عین حال، انعطاف‌پذیری و تطبیق‌پذیری بالاتری در مواجهه با نوسانات بازار داشته باشند. این پژوهش با ارائه‌ی یک رویکرد یکپارچه و داده‌محور، گامی مؤثر در جهت بهبود تصمیم‌گیری در زنجیره‌های تأمین با اطلاعات ناقص و هزینه‌های تصادفی ناپایدار برداشته است و چارچوبی مقاوم برای پیاده‌سازی سیستم‌های پشتیبان تصمیم در محیط‌های عملیاتی پرنوسان فراهم می‌آورد.

مراجع-۵

Azaron, A., Tang, O., & Tavakkoli-Moghaddam, R. (2009). Dynamic lot sizing problem with continuous-time Markovian production cost. International Journal of Production Economics, 120(2), 607–612.

Barnes-Schuster, D., Bassok, Y., & Anupindi, R. (2002). Coordination and flexibility in supply contracts with options. Manufacturing & Service Operations Management, 4(3), 171–207.

Bellman, R. (1957). A Markovian decision process. Journal of Mathematics and Mechanics, 679–684.

Bilsel, R. U., & Kumara, S. R. (2007). A reinforcement learning approach for dynamic supplier selection. In 2007 IEEE International Conference on Service Operations and Logistics, and Informatics (pp. 1–6). IEEE.

Chauhan, V. K., Mak, S., Parlikad, A. K., Alomari, M., Casassa, L., & Brintrup, A. (2023). Real-time large-scale supplier order assignments across two-tiers of a supply chain with penalty and dual-sourcing. Computers & Industrial Engineering, 176, 108928.

Chen, Z., & Rossi, R. (2021). A dynamic ordering policy for a stochastic inventory problem with cash constraints. Omega, 102, 102378.

Cheraghalipour, A., & Farsad, S. (2018). A bi-objective sustainable supplier selection and order allocation considering quantity discounts under disruption risks: A case study in plastic industry. Computers & Industrial Engineering, 118, 237–250.

Darezereshki, F., & Derhami, V. (2024). A reinforcement learning approach to determine when and how many stocks to buy in stock trading. [Unpublished manuscript].

De Farias, D. P., & Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. Operations Research, 51(6), 850–865.

Geevers, K., Van Hezewijk, L., & Mes, M. R. (2024). Multi-echelon inventory optimization using deep reinforcement learning. Central European Journal of Operations Research, 32(3), 653–683.

Hosseini, Z. S., Flapper, S. D., & Pirayesh, M. (2022). Sustainable supplier selection and order allocation under demand, supplier availability and supplier grading uncertainties. Computers & Industrial Engineering, 165, 107811.

Kawakami, K., Kobayashi, H., & Nakata, K. (2021). Seasonal inventory management model for raw materials in steel industry. INFORMS Journal on Applied Analytics, 51(4), 312–324.

Kharvi, S., & Pakkala, T. (2022). An optimal ordering inventory policy when the purchase price follows a discrete-time Markov chain. Communications in Statistics-Simulation and Computation, 1–18.

Khodadadi, H., & Derhami, V. (2025). Employing chaos theory for exploration–exploitation balance in deep reinforcement learning. Tabriz Journal of Electrical Engineering, 55(1), 113–121.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764–766.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]

Li, Q., Wu, X., & Cheung, K. L. (2009). Optimal policies for inventory systems with separate delivery-request and order-quantity decisions. Operations Research, 57(3), 626–636.

Luo, S., Ahiska, S. S., Fang, S.-C., King, R. E., Warsing Jr., D. P., & Wu, S. (2021). An analysis of optimal ordering policies for a two-supplier system with disruption risk. Omega, 105, 102517.

Maulidi, I., Radhiah, R., Hayati, C., & Apriliani, V. (2022). Optimal raw material inventory analysis using Markov decision process with policy iteration method. JTAM (Jurnal Teori dan Aplikasi Matematika), 6(3), 638–650.

Milani, H. (2024). A Markov decision processes model for stochastic inventory problem under uncertainty in supply. Available at SSRN 4833769.

Mohamed, I. B., Klibi, W., Sadykov, R., Şen, H., & Vanderbeck, F. (2023). The two-echelon stochastic multi-period capacitated location-routing problem. European Journal of Operational Research, 306(2), 645–667.

Nayeri, S., Khoei, M. A., Rouhani-Tazangi, M. R., GhanavatiNejad, M., Rahmani, M., & Tirkolaee, E. B. (2023). A data-driven model for sustainable and resilient supplier selection and order allocation problem in a responsive supply chain: A case study of healthcare system. Engineering Applications of Artificial Intelligence, 124, 106511.

Piao, M., Zhang, D., Lu, H., & Li, R. (2023). A supply chain inventory management method for civil aircraft manufacturing based on multi-agent reinforcement learning. Applied Sciences, 13(13), 7510.

Pontrandolfo, P., Gosavi, A., Okogbaa, O. G., & Das, T. K. (2002). Global supply chain management: A reinforcement learning approach. International Journal of Production Research, 40, 1299–1317.

Powell, W. B., Shapiro, J. A., & Simão, H. P. (2002). An adaptive dynamic programming algorithm for the heterogeneous resource allocation problem. Transportation Science, 36(2), 231–249.

Ryzhov, I., Powell, W., & Frazier, P. (2008). The knowledge-gradient algorithm for online learning. [Submitted for publication].

Savvopoulou, A. (2024). Inventory management: Application at a packaging materials manufacturer. [Master's thesis].

Scarf, H. (1960). The optimality of (S, s) policies in the dynamic inventory problem. In Mathematical Methods in the Social Sciences.

Sulistyoningarum, R., Rosyidi, C. N., & Rochman, T. (2020). Supplier selection and order allocation of recycled plastic materials: A case study in a plastic manufacturing company. International Journal of Information and Management Sciences, 31(4), 315–330.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

Thompson, S., Nunez, M., Garfinkel, R., & Dean, M. D. (2009). OR practice—efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. Operations Research, 57(2), 261–273.

Van Roy, B., Bertsekas, D. P., Lee, Y., & Tsitsiklis, J. N. (1997). A neuro-dynamic programming approach to retailer inventory management. In Proceedings of the 36th IEEE Conference on Decision and Control (Vol. 4, pp. 4052–4057). IEEE.

Wagelmans, A., Van Hoesel, S., & Kolen, A. (1992). Economic lot sizing: An O(n log n) algorithm that runs in linear time in the Wagner-Whitin case. Operations Research, 40(1), S145–S156.

Wagner, H. M., & Whitin, T. M. (1958). Dynamic version of the economic lot size model. Management Science, 5(1), 89–96.

Wagner, H. M., & Whitin, T. M. (2004). Dynamic version of the economic lot size model. Management Science, 50(12), 1770–1774.

Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3), 279–292.

Wu, J., Su, L., Wang, G., & Yang, Y. (2024). Approximated dynamic programming for production and inventory planning problem in cold rolling process of steel production. Mathematics, 12(24), 3922.

Wulan, Q. (2023). Order scheduling optimization in manufacturing enterprises based on MDP and dynamic programming. Scientific Reports, 13(1), 9783.

Zhang, H., Li, N., & Lin, J. (2024). Modeling the decision and coordination mechanism of power battery closed-loop supply chain using Markov decision processes. Sustainability, 16(11), 4329.

Zhao, G., & Sun, R. (2010). Application of multi-agent reinforcement learning to supply chain ordering management. In 2010 Sixth International Conference on Natural Computation (Vol. 7, pp. 3830–3834). IEEE.

Zhou, Y., Guo, K., Yu, C., & Zhang, Z. (2024). Optimization of multi-echelon spare parts inventory systems using multi-agent deep reinforcement learning. Applied Mathematical Modelling, 125, 827–844.

دباغ، ر.، حق‌شناس، ا.، و جودت‌نیا، س. (۱۴۰۴). شناسایی و اولویت‌بندی ریسک‌های زنجیره تأمین برق با رویکرد تحلیلی FMEA و ARAS (مطالعه موردی: شرکت توزیع برق استان آذربایجان غربی). کیفیت و بهره‌وری صنعت برق، ۱۴(۱)، ۲۹-۱۹.

دزفولیان، ح.، شیخی، م.، و سموئی، پ. (۱۴۰۴). شناسایی و رتبه‌بندی ریسک‌های زنجیره تأمین برق با استفاده از جمع‌سپاری و روش تصمیم‌گیری چندمعیاره آراس. کیفیت و بهره‌وری صنعت برق، ۱۴(۱)، ۳۸.

نادی، ف.، درهمی، و.، و عالمیان هرندی، ف. (۱۴۰۳). یادگیری تقویتی فازی مبتنی بر ارزش‌تکرار در ربات دنبال‌کننده هدف. مجله کنترل، ۱۸(۲)، ۱۲-۱.

[Downloaded from ieijqp.ir on 2026-09-12]

[DOR: 20.1001.1.23222344.1405.15.1.5.5]