Cost Optimization Method for Procurement and Inventory of NEV Manufacturers Based on Deep Reinforcement Learning

Qun Lu[1], Weichen Zhou[2*]

[1]College of Business Administration, Xuzhou College of Industrial Technology, Xuzhou 221000, China,ywslqyfn_03@163.com

[2]College of Business Administration, Xuzhou College of Industrial Technology, Xuzhou 221000, China,19905175037@163.com

[*]Corresponding author, E-mail:19905175037@163.com

Abstract

The new energy vehicle (NEV) supply chain faces significant challenges stemming from highly uncertain end-user demand and sharp fluctuations in key raw material prices. These factors make procurement costs and inventory levels difficult to control, directly impacting supply chain stability and profitability. Traditional methods, such as stochastic dynamic programming (DP) and standard reinforcement learning (RL) models, which primarily respond only to historical and current state information, often prove insufficient for effectively addressing these complexities. To address these limitations, this paper proposes a Proactive Reinforcement Learning (Pro-RL) framework for joint procurement and inventory decision-making. By integrating a predictive information module into the sequential decision-making process of a Soft Actor-Critic (SAC) agent, the framework constructs an enhanced state space that incorporates predicted future information. This allows the agent to move beyond traditional passive response patterns, enabling proactive utilization of market information to achieve a better balance between immediate costs and long-term risks through iterative learning. To validate its effectiveness, this study develops a supply chain simulation platform aligned with NEV industry characteristics, and comparisons with multiple benchmarks were conducted. Experimental results demonstrate that this end-to-end decision-making policy, which integrates predictive information with deep reinforcement learning, offers advantages in responding to market volatility and achieving coordinated optimization of cost and service levels. This provides NEV enterprises with a theoretical model and practical approach for building flexible and efficient smart supply chains.

Keywords: Deep Reinforcement Learning; Proactive Decision-Making; Supply Chain Cost Optimization; New Energy Vehicle (NEV); Soft Actor-Critic (SAC)

1 Introduction

Amid the global energy structure transition, the new energy vehicle (NEV) industry is experiencing explosive growth. However, the complexity and vulnerability of its supply chain have become increasingly pronounced. On the demand side, the impact of technological iteration, changes in consumer preferences and competitive environment on the market leads to strong nonlinear and high-frequency fluctuation characteristics of market demand (Christensen, 1997; Kotler & Keller, 2016; Porter, 1990). On the supply end, critical raw materials like lithium carbonate face price fluctuations and supply disruptions influenced by multiple factors including geopolitics, speculative behavior, and supply-demand mismatches (Shi et al., 2023). This high dual uncertainty from both ends of the supply chain amplifies stage by stage through the "bullwhip effect," causing NEV manufacturers a dilemma where inventory overstock coexists with stockout risks, while procurement costs spiral out of control (Simchi-Levi et al., 2003), which urgently demands integrated cost optimization strategies.

To address uncertainties, the field of Operations Research (OR) has developed mature optimization paradigms. Among them, Stochastic Programming models the probability distribution of uncertainties through scenario trees. However, when the sources of uncertainties are diverse and the process is non-stationary (as in the NEV market), the number of scenarios grows exponentially, making it difficult to solve the model (Shapiro, 2009). On the one hand, though classical Dynamic Programming (DP) serves as the theoretical foundation for solving sequential decision-making problems, it encounters the "curse of dimensionality" when applied to high-dimensional continuous state-space problems like NEV supply chain, rendering direct computational implementation impractical (Powell, 2011). Robust Optimization, on the other hand, focuses on guaranteeing performance under worst-case scenarios, but its decisions often prove overly conservative, potentially sacrificing significant average performance (Ben-Tal et al., 2009). More importantly, when dealing with the high-dimensional, dynamic, and non-stationary characteristics of NEV market, the reliance on precise probability distribution and high computational costs of these model-based OR methods have become major obstacles to their application.

As a model-free sequential decision-making paradigm, Reinforcement learning (RL) offers a promising approach to solving such problems. It eliminates the need for prior probability assumptions about environmental randomness (e.g., price fluctuations or demand variations), instead learning optimal strategies through direct interaction and trial-and-error with the environment. This inherently aligns with the dynamic uncertainty inherent in NEV supply chains (Sutton & Barto, 2018). Recent advancements in deep reinforcement learning (DRL) have demonstrated significant potential in single inventory control problems (Oroojlooy & Nazari, 2021). However, most existing research focuses on passively responding to historical and current states, with few frameworks actively integrating future prediction information into decision-making processes. This results in strategies lacking foresight, leading to suboptimal decisions when addressing anticipated seasonal promotions or known supply risks.

This has created a critical gap in research: the lack of an integrated optimization method that avoids the "dimensional disaster" and model dependency of traditional OR approaches while transcending the passive response pattern of conventional RL, and actively incorporates future prediction information into decision-making processes. To address this gap, this paper proposes a Proactive Reinforcement Learning (Pro-RL) framework. Its core contributions and innovations include: (1) Integrated decision-making framework: Achieving end-to-end joint optimization of procurement and inventory within a unified DRL framework while addressing dual uncertainties in demand and pricing. (2) Proactive learning mechanism: Innovatively embedding probabilistic prediction information directly into the agent's state space, transforming decision-making from "passive response to historical data" to "active adaptation to future scenarios"–an extension of classical MDP representations. (3)Efficient Solution Paradigm: The Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018) was employed. It effectively handles high-dimensional, continuous decision spaces, and benchmark experiments have verified its superiority over mainstream Deep Reinforcement

Learning (DRL) algorithms such as DDPG (Lillicrap et al., 2015) and PPO (Schulman et al., 2017) in terms of sample efficiency and policy stability (Henderson et al., 2018). This study provides a novel theoretical model and a practical pathway for the frontier scientific question of "how to efficiently integrate predictive information into sequential decision-making processes to navigate deep uncertainty."

The paper is structured as follows: Part II provides formal modeling of the problem and details the Pro-RL framework; Part III outlines the computational experiment design, including newly added benchmarks; Part IV presents and thoroughly discusses the comparative experimental results; Finally, Part V summarizes the paper and outlines future research directions.

2 Problem Formulation and Methodology

2.1 Formalization of the Joint Optimization Problem

This study focuses on a three-echelon supply chain system comprising raw material suppliers, an NEV manufacturer, and the end-market. The manufacturer, as the central decision-making entity, faces the joint optimization problem of procurement and inventory management in a non-stationary market environment. This problem can be modeled as a stochastic dynamic programming problem. In each decision period t (e.g., weekly), the manufacturer's objective is to minimize the expected total discounted cost over a finite horizon T. The problem is formulated as a stochastic dynamic program:

$$\min E\left[\sum_{t=0}^{T-1} \gamma^t C_t(I_t, q_t, m_t | d_t, p_t)\right]$$

where:

E [.] is the expectation operator. $\gamma \in (0,1]$ is the discount factor. C_t(.)denotes the total cost incurred in period t. I_t represents the inventory level of raw materials at the beginning of period t. q_t denotes the procurement quantity decided and ordered in period t. m_t is a target inventory or service level parameter decided in period t, which influences the shortage cost. d_t and p_t represent the observed demand information and raw material price in period t, respectively.

The periodic total cost C_t is composed of the procurement cost C_{pur}、the inventory holding cost C_{hold} and the shortage cost C_{short}. defined as:

$$C_t(I_t, q_t, m_t | d_t, p_t) = C_{pur}(q_t, p_t) + C_{hold}(I_{t+1}) + C_{short}(d_t, m_t)$$

The model is subject to standard inventory balance constraints, physical capacity constraints, and service level requirements.

The core challenge lies in the fact that demand and prices in the NEV market are not simple stationary stochastic processes; rather, they are non-stationary time series driven by complex macroeconomic factors, exhibiting time-varying drift and heteroscedasticity（Co kun &Tümer，2023）. This makes the problem difficult to solve using either methods that require precise stochastic models (like DP) or reactive reinforcement learning models that only respond to historical and current states（Wang et al., 2025）. This fundamental challenge is the primary motivation for our pursuit of a proactive reinforcement learning approach as a breakthrough solution.

2.2 Proactive Reinforcement Learning (Pro-RL)

To address the challenges posed by the aforementioned non-stationary environment, this study proposes a Proactive Reinforcement Learning (Pro-RL) framework. The core concept involves integrating uncertainty-aware prediction with reinforcement learning decision-making, enabling agents to optimize current strategies not only by leveraging historical data but also by considering anticipated future system state changes.

2.2.1 Framework Overview

As shown in Figure 1, the Pro-RL framework consists of three core modules: the supply chain environment simulator, the prediction module, and the RL decision module. At each decision step, the prediction module first generates forecasts for exogenous market factors, such as demand and prices, for the upcoming H time steps, where H denotes the prediction horizon, using prediction techniques informed by the latest market data. The RL decision module then integrates this forward-looking information with real-time supply chain status data from the simulator to form an augmented state vector. This enables the module to output strategy-driven procurement and production decisions, generating an immediate reward. These decisions and rewards induce state transitions in the simulator, allowing the agent to continuously refine its decision-making strategy through these (state, action, reward, next state) transition samples.

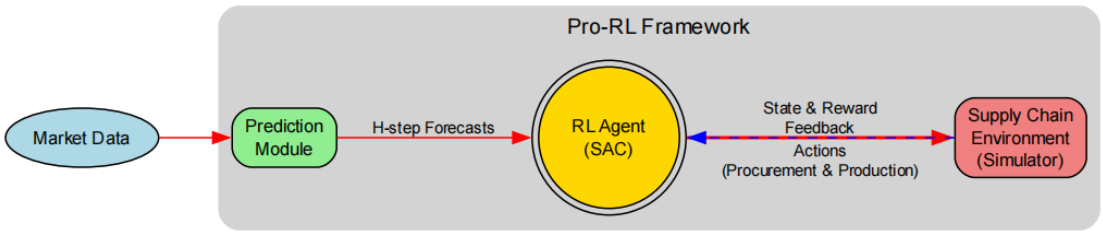

Figure1 Pro-RL Framework

2.2.2 Forecasting Module: Decision-Oriented Uncertainty Quantification

Traditional forecasting models, which provide only point estimates, neglect the inherent uncertainty of the predictions—a critical oversight for risk-sensitive supply chain decisions. In this paper, we design probabilistic forecasting models capable of quantifying this uncertainty:

Demand Forecasting: To account for the trend and seasonality in demand, we employ a hybrid approach combining deep learning and statistical methods. Specifically, a Long Short-Term Memory (LSTM) network is used to capture the nonlinear dynamics of demand, and it is integrated with quantile regression. This allows the model to directly output demand forecast quantiles (e.g., 10%, 50%, 90%) for the next H periods, thereby providing a comprehensive characterization of future demand uncertainty.

Price Forecasting: The prices of NEV raw materials often exhibit volatility clustering. To address this, construct a GARCH-LSTM hybrid model. A GARCH(1,1) component is employed to capture and predict price volatility, while the LSTM component learns the conditional mean dynamics of the prices. This model is particularly well-suited for characterizing the dynamic behavior of raw material prices, such as lithium carbonate, and can output a conditional probability distribution of future prices.

2.2.3 Supply Chain Environment Simulator: a repeatable and controllable training environment

The supply chain environment simulator serves as a benchmark platform for agents to learn and interact, encapsulating the core logic and rules of supply chain operations. This simulator receives decision actions from agents (such as purchase quantities and production volumes), then calculates the system's next- state and real-time costs and rewards based on internal business rules (like inventory dynamics and constraints) and external random factors (such as demand fluctuations). Through this process, it simulates the dynamic evolution of real-world supply chains.

Its core role is to provide a repeatable and controllable training environment for model-free reinforcement learning, enabling agents to learn optimal decision strategies under uncertainty through extensive trial-and-error interactions, without incurring the high costs of real-world experimentation.

2.2.4 Reinforcement Learning (RL) Decision Module: Forward-looking Markov Decision Process (MDP) Reconstruction and Solution

The Reinforcement Learning (RL) decision module, serving as the intelligent core of this framework, consists of two key steps: first, formalizing the supply chain decision problem as a Markov Decision Process (MDP) reconstruction; second, employing an efficient algorithm, Soft Actor-Critic (SAC), to solve this task. The "forward-looking" aspect of the MDP reconstruction leverages the probabilistic forecasts from the preceding module, enabling the design of states and rewards that inherently account for future uncertainty.

2.2.4.1. Proactive MDP Reengineering

To achieve proactive decision-making, this study reengineers the traditional Markov decision process into a five-element tuple (S, A, R, P, γ):

State space (S): The state vector not only contains conventional information reflecting the "past and present", but also incorporates future information from the prediction module in Section 2.2.2. The specific construction is as follows:

$$s_t = [I_t, p_{t-1}, d_{t-1}, ..., \mathbf{F}^d_{t\to t+H}, \mathbf{F}^p_{t\to t+H}]$$

Here, I_t represents the current inventory level; p_{t-1} and d_{t-1} are historical data; The parameters $\mathbf{F}^d_{t\to t+Hd}$ and $\mathbf{F}^p_{t\to t+Hp}$ represent the key parameters of future demand and price forecast distributions over distinct forecast horizons,Hd and Hp respectively. This enhanced state representation transforms agents into decision-makers capable of 'anticipating' future scenarios, providing a foundational information base for developing proactive strategies.

Action Space (A): The agent's actions are represented as a two-dimensional continuous vector $a_t = (a_t^{procure}, a_t^{produce})$, where the two components respectively denote the current raw material purchase quantity and product production quantity. To address practical constraints (such as minimum order quantity (MOQ) and maximum production capacity (MPC)), this paper employs a structured action-space reparameterization method. It decouples the policy network's learning space (a standardized unconstrained space) from the physical execution space of the environment. Through a transformation function f(·), the network's outputs are mapped to a valid decision interval, ensuring the feasibility of decisions in real-world implementation.Specifically, for each action dimension, the function f(·) maps an unconstrained network output to its corresponding feasible interval. For instance, procurement quantity is mapped to $[a_t^{\{procure,min\}}, a_t^{\{procure,max\}}]$ and production quantity to $[a_t^{\{produce,min\}}, a_t^{\{produce,max\}}]$. These per-period operational constraints, such as minimum order quantities (MOQ) and maximum production capacities (MPC), define these feasible intervals.

Reward function (R): The reward is designed as the negative of the periodic total cost, i.e., $R_t = -C_t$, to drive the agent to learn cost-minimizing behaviors.

Transition probability (P) and discount factor (γ): The state transition probability P is implicitly defined by the supply chain environment simulator in Section 2.2.3, without requiring an exact model. The discount factor γ balances short-term and long-term benefits.

2.2.4.2. SAC Algorithm-Based Policy Solving

To address the continuous action and high-dimensional state spaces within the reconstructed MDP, the Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018) is employed as the core policy optimization method. SAC, an advanced maximum entropy off-policy reinforcement learning algorithm, offers theoretical and practical advantages well-suited to the demands of this study.

Excellent continuous control performance: It is specifically designed for continuous action spaces, exhibiting superior stability and sample efficiency.

Intelligent and efficient exploration mechanism: Its maximum entropy objective function

$$J(\pi) = E_{\{\tau\sim\pi\}}[\Sigma\, \gamma^{yt}\,(R(s_t, a_t) + \alpha\, H(\pi(\cdot|s_t)))]$$

encourages agents to maintain strategic randomness while pursuing high returns. This feature is crucial for discovering robust strategies and avoiding local optima in an uncertain supply chain environment.

Efficient off-policy learning: It can utilize a replay buffer to repeatedly learn from historical data, significantly improving data utilization and accelerating convergence.

The SAC algorithm implementation in this study relies on the collaboration of three core components:

Policy network: This network takes the enhanced state (derived from the MDP reconstruction, which integrates predictive information) as input and outputs a stochastic policy for purchasing and production quantities.

Value network: Using dual/twin Q networks to accurately evaluate the value of state-action pairs.

Automatic entropy coefficient α: This coefficient dynamically balances exploration and exploitation through an automatic adjustment mechanism.

Through MDP reconstruction, which effectively integrates predictive information into the state representation, the RL decision module profoundly embeds forecasting into the decision-making process. By employing SAC, the module efficiently learns optimal strategies within this complex, enhanced state space. The agent then undergoes continuous training within the supply chain simulator, thereby developing the capacity to make forward-looking decisions that adeptly balance immediate benefits with long-term strategic objectives.

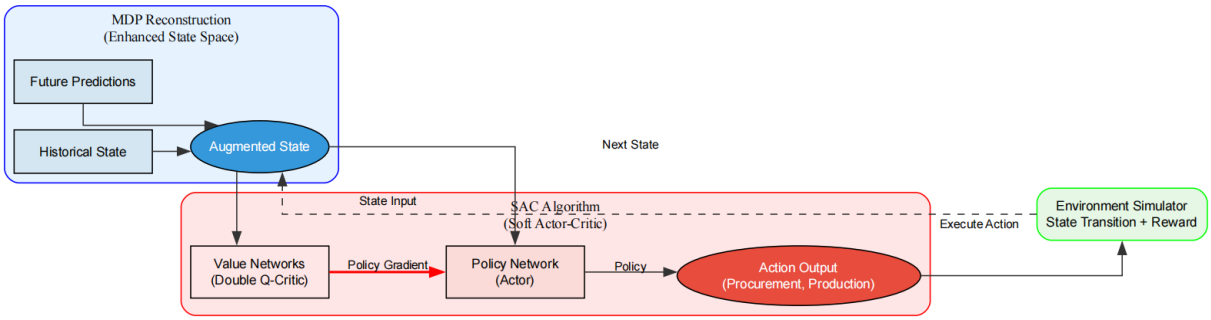

Figure2 MDP AND SAC Algorithm

3 Computational Experiments and Design

To quantitatively evaluate the effectiveness and performance of the proposed Pro-RL framework, this section details a series of computational experiments. The experimental design follows a logical hierarchy, encompassing environment construction, benchmark selection, and a metric system. Through systematic comparison with relevant baseline methods, the experiments aim to validate the proposed method's performance, robustness, and stability.

3.1 Experimental Environment and Parameter Settings

This subsection focuses on the construction of a high-fidelity supply chain simulation environment, whose key parameters are set to reproduce the typical market uncertainty of the new energy vehicle (NEV) industry. The simulation spans 200 weeks, with the agent trained over the first 150 weeks and evaluated on the subsequent 50 weeks. A fixed random seed (e.g., 42) was used to ensure reproducibility.

Market dynamics parameters: To ensure the practical relevance of the experiment, this paper bases the simulation on weekly NEV sales data and spot price data of key raw materials (such as lithium carbonate) publicly available in the Chinese market from 2020 to 2023. Through time series analysis, the mean, variance, autocorrelation, and other characteristics are extracted, forming the basis for generating a simulated data stream conforming to NEV industry dynamics. The demand process is characterized by a baseline cyclical component (e.g., seasonal trend) overlaid with stochastic innovations. These innovations are modeled

as following a normal distribution with a mean of 10 units and a standard deviation of 3 units. Raw material prices, after accounting for any mean process, are modeled such that their innovations follow a normal distribution with a mean of 5 units (e.g., $/kg) and a standard deviation of 1.5 units, with their conditional variance modeled by a GARCH(1,1) process to simulate realistic volatility.

Supply chain environmental parameters: The cost coefficients are based on industry reports and relevant literature (e.g., Syntetos et al., 2016). The unit holding cost is $Ch = 0.2$ (e.g., $/unit/week), the unit stock-out cost is $Cs = 5.0$ (e.g., $/unit/week), and the unit procurement cost Cp is influenced by market prices. Other physical parameters are set as: maximum inventory capacity $ Imax= 100$ units, initial inventory $ I0 = 20$ units.

SAC Algorithm Parameters: To ensure reproducibility and fair comparison, the algorithm hyperparameters adopted a widely recognized robust configuration within the field (Haarnoja et al., 2018). Specific settings include: discount factor $γ= 0.99$, learning rates for both the policy network (Actor) and value network (Critic) set to $3×10 $, experience replay buffer capacity of $10 $ transitions, and a batch size of 256.

Network Architecture: Both Actor and Critic networks employ fully connected networks (FCNs) with two hidden layers of 256 neurons each, utilizing ReLU activation functions. This configuration demonstrates a favorable trade-off between complex nonlinear function approximation and computational efficiency (Golabi et al., 2024).

3.2 Benchmark Strategies

To comprehensively evaluate the performance of the Pro-RL framework, this paper designs three representative benchmark categories for comparison: classical inventory models, advanced Deep Reinforcement Learning (DRL) algorithms, and traditional Operations Research (OR) optimization methods. To ensure fair comparisons, all key parameters within these categories are systematically optimized using an independent validation dataset (e.g., through grid search on a validation set) to achieve optimal performance within their respective frameworks.

3.2.1 Classical Inventory Models

This section details the classical inventory models used as benchmarks.

(s, S) inventory strategy: A time-honored approach involving periodic stocktaking and replenishment. When the end-of-period inventory level It falls below the reorder point s, an order of S-It is placed to restore inventory to the target level S. Through grid search on historical data, this study optimizes the (s, S) parameters, ultimately determining (s, S) = (20,60) to ensure robust baseline performance.

Reorder Point (ROP) Strategy: A continuous monitoring and fixed replenishment approach. When inventory reaches the ROP, an order is placed with a fixed batch size (typically the Economic Order Quantity, EOQ). The parameters (ROP, EOQ) are optimized to 30 and 40 respectively.

3.2.2 Deep Reinforcement Learning Benchmarks

This section introduces Deep Reinforcement Learning (DRL) algorithms as benchmarks to assess the relative performance of the proposed SAC algorithm.

To assess the relative performance of the SAC algorithm in this problem, this study introduces two additional widely used deep reinforcement learning algorithms for comparison:

Deep Deterministic Policy Gradient (DDPG) (Lilicrap et al., 2015). As an early successful DRL algorithm for continuous control, DDPG serves as a crucial baseline for evaluating subsequent improvements like SAC. However, it is susceptible to issues such as overestimation of values and training instability (Fujimoto et al., 2018).

Policy-Optimized Proximal Policy Optimization (PPO) (Schulman et al., 2017): A mainstream on-poli-

cy algorithm that stabilizes training by clipping the objective function to constrain policy updates. While renowned for its robustness and simplicity, PPO typically requires more samples than offline -policy algorithms.

By comparing SAC with DDPG and PPO, this study can systematically evaluate the superiority of SAC's algperforithm in ce across multiple dimensions, including convergence speed, sample efficiency, final policy performance, and training stability.

3.2.3 Operations Research Benchmark (OR Benchmark)

This section describes the Operations Research (OR) benchmark, employing a two-stage stochastic programming model to assess Pro-RL's advantages over traditional optimization methods.

Two-Stage Stochastic Programming (SP): To assess Pro-RL's advantages over traditional OR methods, this study constructed a two-stage stochastic programming model (Campbell, 2011) as a benchmark. The first stage (current procurement and production decisions) is made before uncertainties (future demand and prices) are revealed. The second stage calculates expected holding and shortage costs based on a set of discrete stochastic scenarios (e.g., five discrete scenarios covering high, medium, and low levels, generated by K-means clustering). This model aims to identify a robust decision with the lowest expected total cost across all scenarios, representing a strong representative of traditional model-based optimization.

3.3 Performance Evaluation Framework

In this paper, the following evaluation framework is designed, considering both economic and resilience aspects. To mitigate the impact of randomness, all learning-based algorithms (SAC, DDPG, PPO) are subjected to 10 independent training runs, each initialized with a distinct random seed. The reported performance metrics are presented as mean values from these 10 runs, accompanied by their 95% confidence intervals (CI), to assess stability and result reliability.

Economic Indicators:

Average Total Cost: This metric represents the average total cost per test cycle across all test scenarios, serving as the primary measure of the strategy's economic benefits.

Cost Composition Analysis:The total cost is decomposed into three components: procurement, holding, and shortage costs, to facilitate the analysis of cost drivers for different strategies.

Resilience Indicators:

Extreme Shock Response: This indicator is measured by the inventory-level trajectories of each strategy over time under simulated extreme demand shocks (e.g., bullwhip effect triggering events) or supply disruptions. It visually assesses their ability to resist disturbances and restore balance.

Inventory Level Distribution: The statistical distribution of inventory levels for each strategy across all test cycles is visualized using violin charts to comprehensively measure the stability and risk exposure of inventory management.

4 Results and Discussion

Figure 3 shows the dynamic learning trajectories of various deep reinforcement learning algorithms, alongside their average cost against a non-learning benchmark strategy.

The Pro-RL framework (blue solid line) demonstrates exceptional learning efficiency, achieving the fastest convergence speed. A noticeable flattening of its average cost curve after approximately 400 iterations indicates that the strategy has essentially converged, approaching a performance plateau. The shaded area around the average cost curve represents the standard deviation (±1 standard deviation) of 10 independent experiments. Pro-RL exhibits the narrowest shaded area, indicating high training stability and reproducibility of results. In contrast, the wider shaded area of DDPG reflects its unstable training process.

After convergence, the Pro-RL framework achieves a stable average cost of 140.5, substantially outperforming all comparison methods. Compared to the suboptimal stochastic programming (SP) model (cost 175.2) and the traditional (s,S) strategy (cost 210.5), Pro-RL achieves 19.8% and 33.3% cost reductions, respectively. These quantitative results indicate a clear advantage in cost-effectiveness for the framework. This combination of rapid convergence, superior stability, and significant cost reduction underscores Pro-RL's robustness and efficiency, making it highly promising for practical applications where consistent performance and cost optimization are critical.

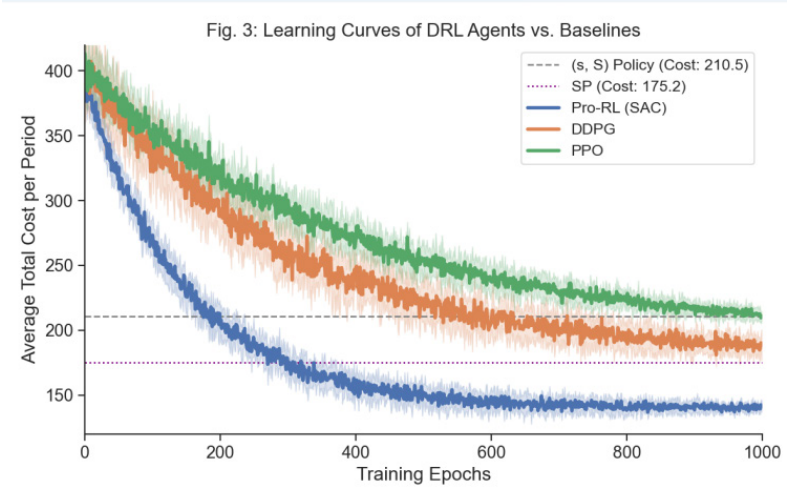

Figure 3: Dynamic learning process of different deep reinforcement learning algorithms

To elucidate the source of Pro-RL's cost advantage, this paper thoroughly analyzes its cost structure and evolutionary process, as presented in Figure 4. Specifically, Figure 4a details the cost structure for each strategy at steady state, while Figure 4b illustrates the dynamic learning trajectory of Pro-RL's optimization strategy. The final cost structure (Figure 4a) confirms the findings from Figure 3: Pro-RL consistently achieves the lowest total cost. Its primary strength lies in virtually eliminating high inventory shortage costs (red), which represents a major challenge for all other strategies. To accomplish this, Pro-RL incurs slightly higher holding costs (green) compared to the (s, S) strategy, thereby demonstrating an ability to intelligently balance different cost components.

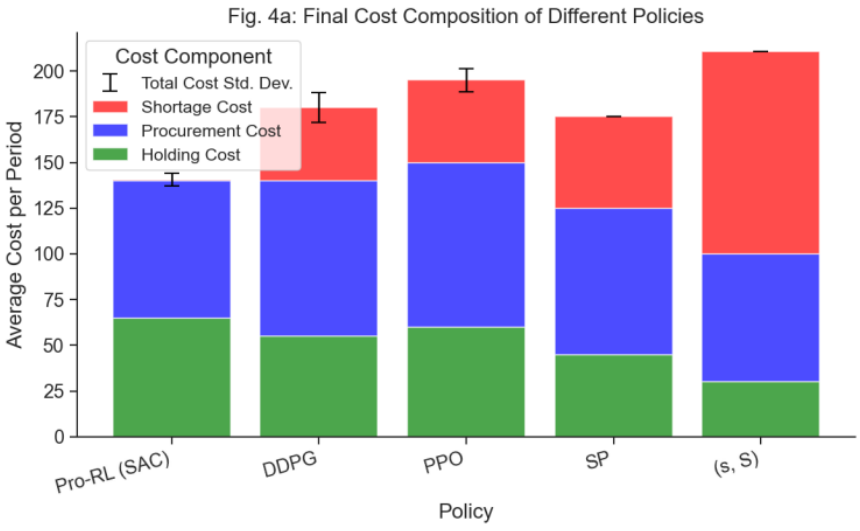

Figure 4a: Cost composition after stabilization of each strategy

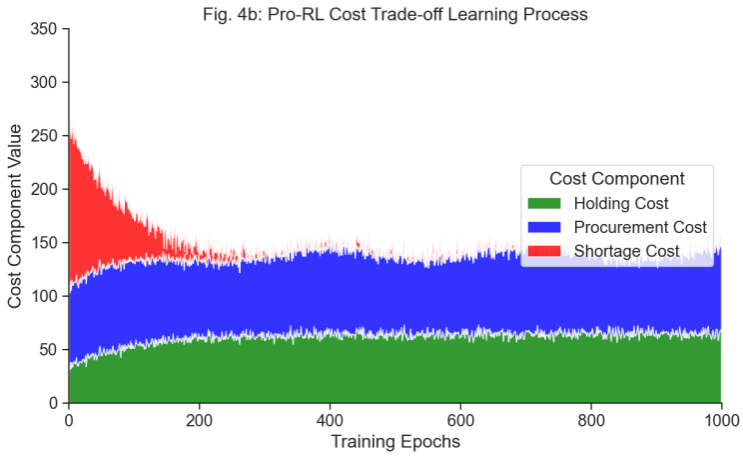

Figure 4b: Dynamic learning process

Figure 5 illustrates the Pro-RL framework's response to an extreme demand shock, highlighting its resilience and foresight. In this visualization, the red area indicates periods of proactive inventory adjustment, while the black dashed line represents the critical shortage threshold.

Proactive adaptation: Unlike conventional reactive strategies, the Pro-RL framework's inventory levels begin to rise modestly multiple cycles before the shock. This anticipatory behavior highlights Pro-RL's predictive module's ability to detect future risks and proactively increase inventory holdings.

Shock resilience and recovery: During the shock period, the Pro-RL framework maintained safe positive inventory levels without ever experiencing shortages, consistently remaining above the black dashed line. In contrast, the inventories of both Proximal Policy Optimization (PPO) and the (s,S) inventory policy plummeted, leading to severe shortages due to their delayed responses. Post-shock, the Pro-RL framework quickly rebounded to normal levels, demonstrating significant resilience.

Evidence of a paradigm shift: This diagram suggests how the Pro-RL framework may fundamentally transform supply chain decision-making paradigms, shifting from 'reactive post-event responses' to 'proactive pre-event adaptation'.

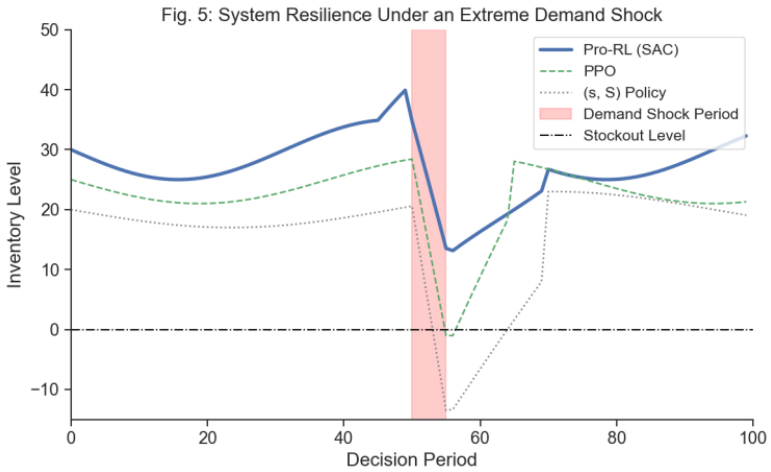

Figure 5: Conclusions on system resilience

Figure 6, a violin plot, visually presents a statistical analysis of the internal inventory control logic and stability for each strategy, outlining a three-tier comparison: Pro-RL's inherent characteristics, its performance against benchmark strategies, and an internal comparison among DRL methods. Pro-RL (SAC) exhibits a sharply peaked density with a narrow interquartile range (IQR), suggesting a precise and stable

control policy that effectively maintains the system within an optimal operational range. In contrast, inventory distributions of benchmark strategies such as (s,S) and the Service Level Policy (SP) show broader, fat-tailed patterns with wide spread, indicating a coarse-grained control that is associated with pronounced fluctuations in system states. The (s,S) strategy most prominently illustrates this, with inventory levels fluctuating wildly between its predefined thresholds s and S, highlighting potential limitations in this setting. Within DRL methods, Pro-RL demonstrates a more concentrated distribution than both DDPG and PPO, underscoring its superior control precision. Collectively, these results demonstrate Pro-RL's enhanced stability and refined inventory management capabilities compared to both traditional and other DRL approaches.

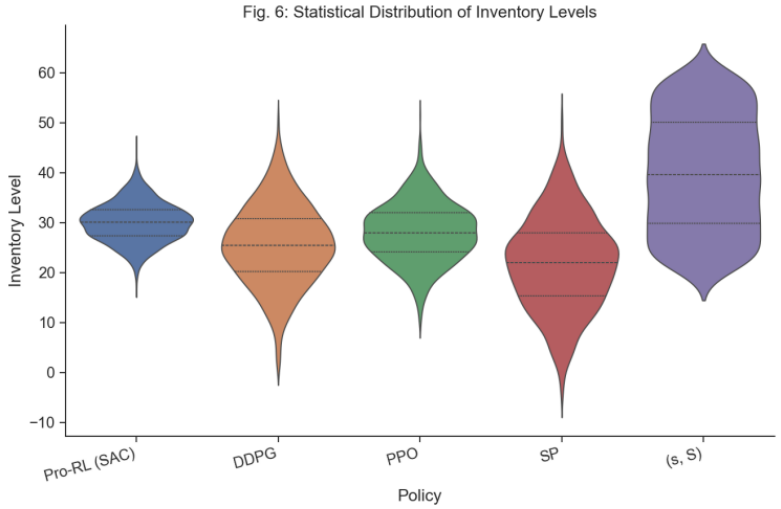

Figure 6: Inventory inherent in each strategy

In conclusion, computational experiments demonstrate the Pro-RL framework's comprehensive superiority across multiple dimensions including economic efficiency, cost structure, impact response, and statistical stability. Its core advantage lies in empowering agents with "future prediction" capabilities, enabling a paradigm shift in supply chain decision-making from "passive response" to "active adaptation." This transformation significantly enhances the system's resilience and robustness in uncertain environments.

5 Conclusion

This study develops a Pro-RL framework for joint procurement and inventory decision-making in new energy vehicle (NEV) supply chains under dynamic uncertainty. By integrating a predictive module into the sequential decision-making process of deep reinforcement learning, the framework constructs an enhanced state space capable of anticipating future dynamics. The joint decision-making problem is then solved using the Soft Actor-Critic (SAC).

To validate the efficacy of the proposed framework, extensive experimental evaluations were conducted. The results consistently demonstrate the framework's superiority across several key dimensions:

Economic Performance and Stability:** The framework not only outperforms all benchmarks in overall economic performance but also exhibits high training stability. Compared with optimized traditional (s,S) strategies and two-stage stochastic programming (SP) models, Pro-RL achieves significant cost reductions of 33.3% and 19.8%, respectively. Additionally, it quantitatively demonstrates superior performance over other mainstream deep reinforcement learning algorithms such as Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO).

Intelligent Trade-Offs and System Resilience: In-depth analysis reveals that Pro-RL's success stems from its adaptive trade-off mechanism—by proactively accepting moderate inventory holding costs to nearly

eliminate high out-of-stock costs. This strategy of stockpiling goods using predictive information demonstrates exceptional system resilience during extreme demand shocks, effectively avoiding the severe stockouts faced by other approaches.

Control Accuracy and Precision: By comparing the statistical distribution of inventory levels under different strategies, Pro-RL demonstrates a highly accurate and stable control pattern, continuously maintaining the system within the optimal operating range, thereby enabling precise inventory management.

In conclusion, this study not only validates the effectiveness of integrating predictive information with sequential decision-making processes but also provides a new theoretical perspective and robust technical framework for addressing other operational management challenges involving "prior knowledge." Furthermore, it offers actionable solutions for enterprises to develop more resilient and cost-effective supply chain strategies in complex and dynamic market environments.

However, this study also has certain limitations. Future research could focus on the following aspects:

Integrating more sophisticated prediction techniques (such as Transformer or LSTM time series models) into the framework to explore the profound impact of prediction accuracy on decision quality.

Extending the framework to multi-level, multi-agent supply chain networks involving battery manufacturers, material suppliers, and vehicle manufacturers to investigate collaborative and competitive decision-making issues in multi-agent settings.

Incorporating more realistic constraints such as procurement lead times, supplier capacity limitations, and multi-product portfolios into the model to further validate the framework's applicability and robustness.

Funding

[1]University-Level "Integrated Curriculum Development" Special Fund (SG-JW-2024)

References

[1] Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). Robust optimization. Princeton University Press.

[2] Campbell, G. M. (2011). A two-stage stochastic program for scheduling and allocating cross-trained workers. Journal of the Operational Research Society, 62(6), 1038–1047. https://doi.org/10.1057/jors.2010.16

[3] Christensen, C. M. (1997). The innovator's dilemma: When new technologies cause great firms to fail. Harvard Business School Press.

[4] Co kun, K., & Tümer, B. (2023). Learning under concept drift and non-stationary noise: Introduction of the concept of persistence. Engineering Applications of Artificial Intelligence, 123, Article 106363.

[5] Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. https://doi.org/10.48550/ARXIV.1802.09477

[6] Golabi, A., Erradi, A., Qiblawey, H., Tantawy, A., Bensaid, A., & Shaban, K. (2024). Optimal operation of reverse osmosis desalination process with deep reinforcement learning methods. Applied Intelligence, 54(8), 6333–6353. https://doi.org/10.1007/s10489-024-05452-8

[7] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. https://doi.org/10.48550/ARXIV.1801.01290

[8] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Article AAAI-18-0245.

[9] Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson Education.

[10] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv:1509.02971.

[11] Oroojlooy, A., & Nazari, M. (2021). A review of deep reinforcement learning in operations research and management science. European Journal of Operational Research, 293(2), 401–418.

[12] Porter, M. E. (1990). The competitive advantage of nations. Free Press.

[13] Powell, W. B. (2011). Approximate dynamic programming: Solving the curses of dimensionality (2nd ed.). John Wiley & Sons.

[14] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. https://doi.org/10.48550/ARXIV.1707.06347

[15] Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). Lectures on stochastic programming: Modeling and theory (3rd ed.). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611976595

[16] Simchi-Levi, D., & Zhao, Y. (2003). The value of information sharing in a two-stage supply chain with production capacity constraints. Naval Research Logistics, 50(8), 888–916. https://doi.org/10.1002/nav.10094

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

[18] Ta, T. A., Mai, T., Bastin, F., & L'Ecuyer, P. (2020). On a multistage discrete stochastic optimization problem with stochastic constraints and nested sampling. Mathematical Programming, 190(1–2), 1–37. https://doi.org/10.1007/s10107-020-01518-w

[19] Wang, B., Wang, L., Zhong, S., Xiang, N., & Qu, Q. (2022). Assessing the supply risk of geopolitics on critical minerals for energy storage technology in China. Frontiers in Energy Research, 10, Article 1032000. https://doi.org/10.3389/fenrg.2022.1032000

[20] Wang, M., Li, X., He, Y., Li, Y., Bennis, M., Islam, R., & Wang, H. (2024). Wavelet predictive representations for non-stationary reinforcement learning. arXiv. https://doi.org/10.48550/ARXIV.2510.04507

[21] Wang, Y., Geng, S., & Gao, H. (2018). A proactive decision support method based on deep reinforcement learning and state partition. Knowledge-Based Systems, 143, 248–258. https://doi.org/10.1016/j.knosys.2017.11.005