Carbon Trading Supply Chain Management Based on Constrained Deep Reinforcement Learning

JAAMAS Track

Qinghao Wang
Institute for Artificial Intelligence, Peking University
Bejing, China
qinghw@pku.edu.cn

Yaodong Yang
Institute for Artificial Intelligence, Peking University
Bejing, China
yaodong.yang@pku.edu.cn

ABSTRACT

Reducing carbon emissions remains a formidable challenge in supply chain management. However, conventional numerical simulations based on heuristics often struggle to address the complex coupling of ordering decisions and carbon emissions—complicated further by high-dimensional observation, multiple constraints, and carbon quota requirements. Constrained DRL leverages the high-dimensional representation capabilities and robust decision-making under constraints, making it particularly suitable for supply chain management involving carbon trading. To address these challenges, this paper proposes a simulation framework grounded in Constrained Markov Decision Processes (CMDP), incorporating constrained deep reinforcement learning (DRL). Specifically, we develop a Double Order algorithm based on PPO-Lagrangian (DOPPOL) to simultaneously optimize business and carbon costs. Experimental results demonstrate that DOPPOL outperforms traditional (s, S) methods under fluctuating demand, effectively balancing cost optimization and emission reductions. Furthermore, integrating carbon trading into supply chain operations allows companies to adapt both ordering decisions and emissions, thereby enhancing operational efficiency. Then we highlight the pivotal role of carbon pricing in business contracts: rational pricing not only helps regulate carbon emissions but also reduces overall costs. Our findings contribute to the broader endeavor of mitigating climate change and promoting sustainable supply chain practices.

KEYWORDS

Carbon trading, Supply chain management, Deep reinforcement learning

ACM Reference Format:
Qinghao Wang and Yaodong Yang. 2025. Carbon Trading Supply Chain Management Based on Constrained Deep Reinforcement Learning: JAAMAS Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 4 pages.

1 INTRODUCTION

Climate change has received substantial global attention [12], prompting numerous strategies aimed at reducing carbon emissions [6, 15]. In this context, reconciling economic progress with environmental stewardship has become an increasingly prominent concern, elevating complex supply chain management to the forefront of sustainability challenges [7, 11]. Although considerable research has addressed multifaceted supply chain problems—ranging from joint replenishment [5] and transportation with shelf space management [3] to the bullwhip effect [16]—current work offers limited insight into how companies can jointly optimize ordering decisions and carbon trading. Such joint decision-making is essential for aligning corporate profitability with environmental objectives.

Deep reinforcement learning (DRL) has emerged as a powerful decision-making framework that can handle high-dimensional and dynamic environments [2, 4, 9]. Recent studies showcase the capability of DRL to solve a variety of supply chain issues involving lost sales [14], dual-sourcing [1], and multi-echelon configurations [8]. However, research applying DRL to supply chain operations constrained by carbon quotas and incorporating carbon trading remains scarce. Integrating carbon trading into supply chain decision-making holds significant potential for navigating trade-offs between economic outcomes and environmental requirements.

To bridge this gap, we formulate a comprehensive simulation framework considering both economic returns and sustainability. We develop a constrained DRL method called DOPPOL to address the combined ordering and carbon trading problem under carbon quota constraints, which demonstrates superior performance over conventional approaches. Moreover, we examine the effects of carbon trading policies and carbon price variations, thereby providing new insights for companies seeking to balance business profitability with ecological imperatives.

2 PROBLEM FORMULATION

We integrate carbon emissions and trading considerations into supply chain management, incorporating carbon trading along with carbon quota constraints. The problem is formulated as a Constrained Markov Decision Process (CMDP) over a period. A CMDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ represents actions, $\mathcal{P}$ denotes transition probabilities, $\mathcal{R}$ is the expected reward, and γ is the discount factor [13]. The objective is to find a policy $\pi(a|s)$ that maximizes the expected reward while adhering to the constraints[10]: $J^C(\pi) \doteq \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t c\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \leq d$, where $c\left(\mathbf{s}_t, \mathbf{a}_t\right)$: $\mathcal{S}\times\mathcal{A} \rightarrow \mathbb{R}$. Companies determine order quantities and engage in carbon trading based on demand, aiming to minimize total costs. The goal is to find a balance between maximizing the expected reward and satisfying the constraints and the training

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

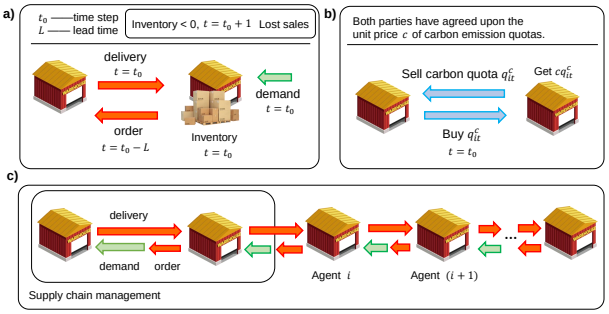

Figure 1: Supply chain management architecture.

objective is to maximize the return $J_r(\pi_\theta)$. $\max_{\pi_\theta \in \Pi_f} J_r(\pi_\theta)$ is calculated, where $\Pi_f \doteq \{\pi_\theta \in \Pi : J_c(\pi_\theta) \leq d\}$ is the feasible set.

We take into account the business costs (C_1) and environmental costs (C_2). At time step t, the order quantity for company i is denoted as q_{it}. d_{it} presents the demand. The unit order cost is denoted as p_i. The inventory is denoted as I_{it} and the holding cost is h_{it}. b_{it} is the stockout cost. Company i is allocated a certain amount of carbon quotas, denoted as I_{i0}, and it can buy extra allowances, denoted as q_{it}^c ($q_{it}^c > 0$). An indicator function is used to represent the purchase of carbon emission allowances $\mathbb{I}(q_{it}^c \geq 0)$ and the sale $\mathbb{I}(q_{it}^c < 0)$. A penalty of $B(B > 0)$ is imposed. The environmental profit is: $\text{Profit}_i^c = c \cdot \mathbb{I}(q_{it}^c < 0)|q_{it}^c| - C_{2t}$, where c represents the price of one unit of carbon quota. The constraint is $I_{i,t+1}^c = I_{i,t}^c + q_{it}^c - k \cdot C_{1t} \geq 0$, Let k ($k > 0$) represent the share of carbon emissions costs in business expenses. Firms enter contracts with a fixed unit price for all transactions. The overall profit (reward) is $Y_i = p_{i+1}d_{it} - p_i q_{it} - h_{it}I_{it} - b_i[d_{it} - q_{i,t-1} - I_{it}]^+ + c \cdot \mathbb{I}(q_{it}^c < 0)q_{it}^c - B \cdot \mathbb{I}(d_{it}^c - q_{i,t-1}^c - I_{it}^c > 0) - c \cdot \mathbb{I}(q_{it}^c \geq 0)|q_{it}^c|$. The interactive mode and profit calculation framework among supply chains considering carbon trading are illustrated in Fig. 1.

3 METHOD

In our decision-making framework, we formulate a CMDP wherein the state s encompasses a company's product inventory, historical order quantities, downstream demand, remaining carbon emission quotas, carbon emissions, and records of past carbon quota transactions. The action a is defined as (q, q^c).

To solve this CMDP, we propose the DOPPOL algorithm (Algorithm 1) based on PPO-Lagrangian. DOPPOL dynamically updates a Lagrange multiplier λ to flexibly address the evolving carbon emission constraints. Specifically, changes in a company's carbon emissions adjust its carbon quota, influencing subsequent trading decisions. Consequently, λ is recalculated at each step to reflect these variations, as shown in $\hat{\lambda}_{t+1} = \max(0, \hat{\lambda}_t + \alpha_\lambda(J^c(\pi) - I_{it}^c))$.

4 RESULTS

In our study, the results indicate that DOPPOL outperforms traditional method (s, S) (Fig. 2), and companies achieve optimized ordering with varying distributions by utilizing DOPPOL. As shown in Fig. 3, our method effectively controls carbon trading due to carbon emission quota limitations. Fig. 4 illustrates the impact of carbon pricing on a company's overall returns, showing that as carbon pricing increases in contracts, companies incur higher carbon costs.

Algorithm 1 DOPPOL: double order based on PPO-Lagrangian

```
Initialize environment, policy and value network
Set hyperparameters: clipping parameter ϵ, value coefficient β,
  entropy coefficient η, learning rate for Lagrange multiplier α_λ
for epochs do
    Estimate J^c(π_θ)
    Sample action [q, q_c]
    Calculate cumulative reward
    for each optimization step do
        Update Lagrange multiplier λ
        Update value function:
            Minimize L^V(θ_v) = 𝔼[(V_θv(s) − V_target)^2]
        Update policy:
            Compute J^r(π_θ) − λJ^c(π_θ)
            Minimize L(θ) = L^r(θ) + λL^c(θ)
            Update policy parameters
    end for
end for
```

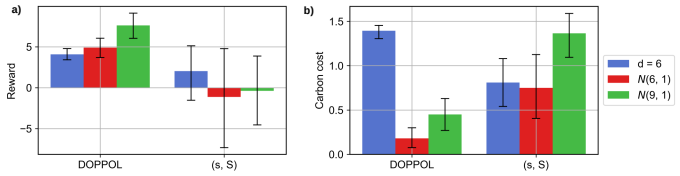

Figure 2: Comparison between (s, S) and DOPPOL.

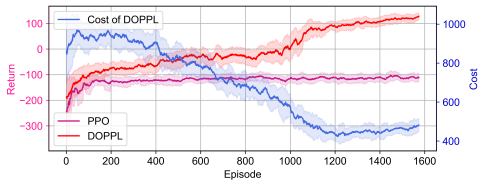

Figure 3: A comparison of the PPO and DOPPOL.

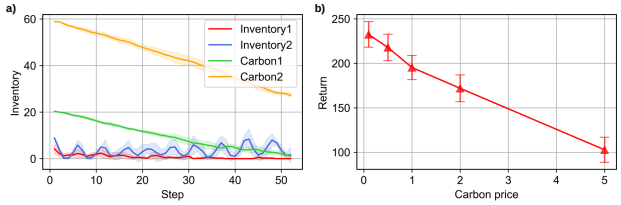

Figure 4: In accordance with the terms of the contract for carbon trading.

5 CONCLUSION

In conclusion, we propose DOPPL, a novel supply chain management approach grounded in constrained DRL that strategically integrates carbon trading and ordering decisions. By incorporating

carbon prices into operational considerations, our framework significantly enhances cost efficiency, maximizes profitability, and improves overall supply chain performance. We further examine how contract-based carbon pricing significantly influences profitability. This paper offers critical insights into advancing sustainable supply chain management and facilitating transformational change.

REFERENCES

[1] Lucas Böttcher, Thomas Asikis, and Ioannis Fragkos. 2023. Control of Dual-Sourcing Inventory Systems Using Recurrent Neural Networks. *INFORMS Journal on Computing* 35, 6 (2023), 1215–1532.

[2] Robert N Boute, Joren Gijsbrechts, Willem Van Jaarsveld, and Nathalie Vanvuchelen. 2022. Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research* 298, 2 (2022), 401–412.

[3] Gerard Cachon. 2001. Managing a Retailer's Shelf Space, Inventory, and Transportation. *Manufacturing & Service Operations Management* 3 (07 2001), 211–229.

[4] Joren Gijsbrechts, Robert N Boute, Jan A Van Mieghem, and Dennis J Zhang. 2022. Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management* 24, 3 (2022), 1349–1368.

[5] Suresh K. Goyal and Ahmet T. Satir. 1989. Joint replenishment inventory control: Deterministic and stochastic models. *European Journal of Operational Research* 38, 1 (1989), 2–13.

[6] Donald Huisingh, Zhihua Zhang, John C Moore, Qi Qiao, and Qi Li. 2015. Recent advances in carbon emissions reduction: policies, technologies, monitoring, assessment and modeling. *Journal of cleaner production* 103 (2015), 1–12.

[7] Richard Lamming and Jon Hampson. 1996. The environment as a supply chain management issue. *British journal of Management* 7, 1 (1996).

[8] Xiaotian Liu, Ming Hu, Yijie Peng, and Yaodong Yang. 2022. Multi-agent deep reinforcement learning for multi-echelon inventory management. *Production and Operations Management* (2022), 10591478241305863.

[9] Afshin Oroojlooyjadid, MohammadReza Nazari, Lawrence V Snyder, and Martin Takáč. 2022. A deep q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management* 24, 1 (2022), 285–304.

[10] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* 7, 1 (2019), 2.

[11] Hartmut Stadtler. 2015. *Supply Chain Management: An Overview.* Springer Berlin Heidelberg, Berlin, Heidelberg, 3–28. https://doi.org/10.1007/978-3-642-55309-7_1

[12] Wilfried Thuiller. 2007. Climate change and the ecologist. *Nature* 448, 7153 (2007), 550–552.

[13] Akifumi Wachi and Yanan Sui. 2020. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*. PMLR, 9797–9806.

[14] Qinghao Wang, Yijie Peng, and Yaodong Yang. 2022. Solving Inventory Management Problems through Deep Reinforcement Learning. *Journal of Systems Science and Systems Engineering* 31, 6 (2022), 677–689.

[15] John P Weyant. 1993. Costs of reducing global carbon emissions. *Journal of Economic Perspectives* 7, 4 (1993), 27–46.

[16] Y. Yang, J. Lin, G. Liu, and L. Zhou. 2021. The behavioural causes of bullwhip effect in supply chains: A systematic literature review. *International Journal of Production Economics* 236 (2021), 108120.