Structure-Informed Deep Reinforcement Learning for Inventory Management

Alvaro Maggiar [*]
Amazon.com
maggiar@amazon.com

Sohrab Andaz
Amazon.com
sandaz@amazon.com

Akhil Bagaria
Amazon.com
akhilbg@amazon.com

Carson Eisenach
Amazon.com
ceisen@amazon.com

Dean Foster
Amazon.com
foster@amazon.com

Omer Gottesman
Amazon.com
omergott@amazon.com

Dominique Perrault-Joncas
Amazon.com
joncas@amazon.com

Abstract

This paper explores the application of deep reinforcement learning (DRL) to classical inventory management problems while incorporating theoretical insights from traditional operations research. We demonstrate that a simple DRL implementation using DirectBackprop [12] can effectively handle diverse scenarios including multi-period systems with lost sales, lead times, perishability, dual sourcing, and joint procurement-removal decisions. Through extensive experiments, we show that our approach performs competitively against established benchmarks while naturally learning many structural properties of optimal policies that were previously derived analytically. We introduce a Structure-Informed Policy Network technique that explicitly incorporates these analytical insights into the learning process, enhancing generalization and robustness. Using realistic retail demand data, we demonstrate how this approach helps with extrapolation and provides robustness on out-of-sample data.

1 Introduction and Background

Inventory management optimization represents a cornerstone challenge in operations research and supply chain management, exemplifying the tension between model-based rigor and data-driven flexibility that characterizes modern ML-OR integration challenges. Traditional approaches have relied on analytical methods and heuristics [26, 20], providing valuable solutions for simplified settings such as the seminal newsvendor problem [17] or Economic Order Quantity [7]. These model-based approaches offer interpretability and theoretical guarantees but often struggle with the complexity and scale of real-world systems. However, real-world applications involve complex dynamics, stochastic demand patterns, and intricate constraints that often render these methods intractable. Even in relatively simple settings such as one with stationary demand, lost sales and lead times, the problem is notoriously difficult [27].

Deep Reinforcement Learning (DRL) offers a promising alternative, capable of learning effective policies directly from data without relying on explicit modeling of system dynamics. This model-free

[*] Corresponding Author

Preprint.

paradigm aligns with recent advances in AI/ML that have achieved success by eschewing traditional model-based assumptions. However, this creates the challenge of leveraging the rich theoretical insights from decades of OR research while harnessing the adaptive power of modern ML techniques.

Recent work has demonstrated DRL's potential in addressing practical inventory control problems, with notable successes in applications at major retailers like JD.com [18], Alibaba [11], and Amazon [12]. However, most approaches suffer from limitations when it comes to representing practical alternatives and fail to adequately address the uncertainty mitigation challenges inherent in converting data into reliable operational decisions. Common approaches apply RL algorithms at an instance level [21, 16, 6], learning policies for individual products by leveraging information across multiple realized scenarios. This contrasts with practical settings where one typically has access to single realized scenarios for individual products and must leverage information across products. Additionally, many studies allow policies to consume distribution parameters directly, while in practice we only have access to historical realized demand, creating additional layers of uncertainty propagation from data prediction errors into operational decisions.

2 Approach and Methodology

Our work addresses these limitations through several key innovations that exemplify effective ML-OR synergization:

- We apply DRL as it would be implemented in practice, mimicking state information available to practitioners. This involves learning policies across products using only historical information, without access to demand distribution parameters. This approach directly addresses the uncertainty mitigation challenge by learning robust policies that can handle distributional shifts and model uncertainty without requiring explicit uncertainty quantification.
- We leverage the DirectBackprop algorithm [12], which allows for the reduction of the problem to supervised learning with minimal hyperparameter tuning. This algorithm has shown success in practical applications and benefits from learnability results [25]. The differentiable formulation enables efficient gradient-based optimization while maintaining computational tractability for large-scale applications.
- We demonstrate strong performance across diverse scenarios: multi-period systems with lost sales (both with and without lead times), perishable product management, dual sourcing, and joint inventory procurement and removal.
- We propose a Structure-Informed Neural Network technique that incorporates analytically-derived characteristics of optimal policies into the learning process, inspired by Physics Informed Neural Networks [19]. This approach represents a novel framework for integrating operational domain knowledge into ML algorithms, balancing model-based insights with data-driven flexibility. The technique uses differential penalties to enforce structural properties derived from classical OR theory, ensuring that learned policies conform to known theoretical properties while maintaining the adaptability of neural networks.

3 Key Results

Our experimental results demonstrate several significant findings:

3.1 Performance on Classical Problems

We evaluate our approach on five classical inventory management problems:

- **Basic Lost Sales:** Our DRL approach achieves within 0.5% of the optimal (omniscient) policy despite having access to only historical information rather than full distributional knowledge. This is notable as the optimal policy is known to be of base-stock type [8, 22].
- **Lead Times with Lost Sales:** In this notoriously difficult setting [27], where optimal policies are generally unknown, our approach outperforms established heuristics including vector base-stock policies [14], with the performance gap increasing with lead time length.

- **Perishable Inventory:** For products with finite shelf life, where optimal policies depend on the full $m-1$ dimensional state space [15, 3], the DRL policy significantly outperforms naive base-stock policies for short shelf lives while matching the performance of optimized base-stock policies as shelf life increases.
- **Dual Sourcing:** Our approach learns near-optimal policies in settings with known solutions [4, 24] and outperforms heuristics in more complex scenarios with non-consecutive lead times.
- **Joint Procurement-Removal:** The DRL agent successfully learns interval-stock policies [13] for inventory management with returns, demonstrating its ability to discover complex policy structures.

3.2 Structure-Informed Learning

A key finding is that the DRL approach naturally learns many structural properties of optimal policies that were previously derived through operations research methods, including monotonicity properties [14] and sensitivity results derived through $L^\natural$-convexity [28] and multimodularity [10]. However, these properties may not hold uniformly across the state space, particularly in regions rarely visited during training, creating potential reliability issues when policies encounter out-of-distribution states.

Our Structure-Informed Policy Network technique addresses this by explicitly incorporating these properties through differential penalties during training, representing a principled approach to uncertainty mitigation that leverages the rich theoretical arsenal of OR. The method computes gradients of policy outputs with respect to state variables and penalizes violations of known structural properties such as monotonicity in inventory levels or convexity in cost functions. This regularization approach ensures that learned policies maintain theoretical consistency even in rarely-visited regions of the state space, providing robustness guarantees that are crucial for operational deployment.

Figure 1 illustrates the regularization effect for 4 sample products in the case of a multiperiod inventory problem with lost sales, and a lead time of 5 periods. The contour plots are slices of the learned policies in a given period, and show the order quantity as a function of on-hand inventory and inventory arriving 1 period hence. The top row corresponds to the unregularized policy, while the bottom row corresponds to the regularized one.

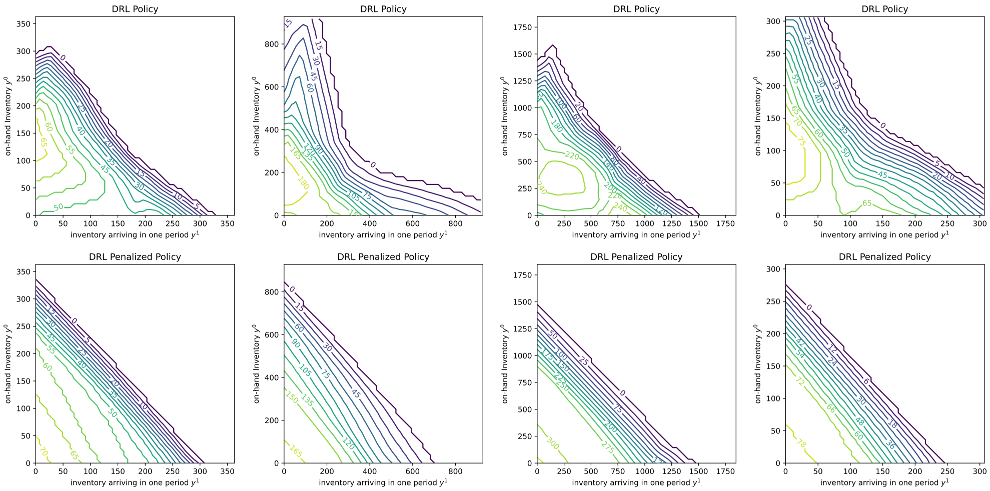

Figure 1: Example contour plots of the unpenalized (top), and penalized (bottom) policies for a few given product in the case of $L = 5$ as a function of the endogenous state $\mathbf{y} = (y^0, y^1, 0, 0, 0)$ at time $t = 0$.

We show empirically that this approach:

- Improves generalization to out-of-sample states by enforcing known structural properties throughout the state space

- Provides robustness against demand shocks and non-stationary patterns by maintaining policy coherence with established OR principles
- Maintains or improves performance while ensuring policy conformance with theoretical properties, demonstrating that OR insights enhance rather than constrain ML performance
- Results in more interpretable policies that are easier to validate and deploy in practice, bridging the gap between "black box" ML systems and the transparency requirements of operational systems
- Reduces the computational overhead of uncertainty quantification by embedding structural constraints directly into the policy architecture

3.3 Real-World Demand Data

Using retail data from Corporacion Favorita, we compare our end-to-end DRL approach against a traditional predict-then-optimize baseline using state-of-the-art demand forecasting [23] in a simplified zero-lead time setting. The DRL approach:

- Achieves higher average reward over the test period
- Operates with consistently lower inventory levels (reducing working capital requirements)
- Shows better adaptability to demand spikes and seasonal patterns, particularly around Christmas periods
- Maintains higher demand-weighted service levels while using less inventory
- Demonstrates superior recovery from major demand shocks like the 2016 Ecuador earthquake

This outperformance aligns with theoretical results suggesting benefits of end-to-end approaches over separate prediction and optimization [2] and matches industry experiences at major retailers [12, 18, 11].

4 Implications and Future Work

Our work demonstrates that combining modern deep reinforcement learning with classical operations research insights yields practical and robust solutions for real-world inventory management challenges. This bridges the gap between theoretical understanding and data-driven approaches, offering several promising directions for future research:

- Extending the Structure-Informed Policy Network approach to incorporate additional types of structural properties, particularly those derived from convex analysis tools like supermodularity and $L^\natural$-convexity [1].
- Investigating scalability to larger networks with multiple echelons or more complex constraints, building on recent work in network inventory management [5, 9].
- Developing theoretical guarantees for the convergence and optimality of structure-informed policies, extending existing results on learnability [12] and VC theory [25].
- Exploring applications to other classical operations problems where theoretical insights exist but practical implementation remains challenging.

The success of our approach in handling diverse scenarios while maintaining theoretical properties suggests broad applicability across the operations research domain. By combining the flexibility of deep learning with classical theoretical insights, we provide a framework for developing practical solutions that benefit from decades of analytical research while adapting to real-world complexities. This work contributes to the broader goal of ML-OR synergization by demonstrating how principled integration of domain knowledge can enhance both the performance and reliability of data-driven operational systems.

References

[1] Xin Chen. $L^\natural$-convexity and its applications in operations. *Frontiers of Engineering Management*, 4(3):283–294, 2017.

[2] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

[3] Bruce E Fries. Optimal ordering policy for a perishable commodity with fixed lifetime. *Operations research*, 23(1):46–61, 1975.

[4] Yoichiro Fukuda. Optimal policies for the inventory problem with negotiable leadtime. *Management Science*, 10(4):690–708, 1964.

[5] Kevin Geevers, Lotte van Hezewijk, and Martijn RK Mes. Multi-echelon inventory optimization using deep reinforcement learning. *Central European Journal of Operations Research*, 32(3):653–683, 2024.

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

[7] Ford W Harris. How many parts to make at once. *Operations research*, 38(6):947–950, 1990.

[8] Samuel Karlin. Dynamic inventory policy with varying stochastic demands. *Management Science*, 6(3):231–258, 1960.

[9] Illya Kaynov, Marijn van Knippenberg, Vlado Menkovski, Albert van Breemen, and Willem van Jaarsveld. Deep reinforcement learning for one-warehouse multi-retailer inventory management. *International Journal of Production Economics*, 267:109088, 2024.

[10] Qing Li and Peiwen Yu. Multimodularity and its applications in three stochastic dynamic inventory problems. *Manufacturing & Service Operations Management*, 16(3):455–463, 2014.

[11] Jiaxi Liu, Shuyi Lin, Linwei Xin, and Yidong Zhang. Ai vs. human buyers: A study of alibaba's inventory replenishment system. *INFORMS Journal on Applied Analytics*, 53(5):372–387, 2023.

[12] Dhruv Madeka, Kari Torkkola, Carson Eisenach, Anna Luo, Dean P. Foster, and Sham M. Kakade. Deep inventory management. *arXiv preprint arXiv:2210.03137*, 2022.

[13] Alvaro Maggiar and Ali Sadighian. Joint inventory and revenue management with removal decisions. *Available at SSRN 3018984*, 2017.

[14] Thomas E Morton. Bounds on the solution of the lagged optimal inventory equation with no demand backlogging and proportional costs. *SIAM review*, 11(4):572–596, 1969.

[15] Steven Nahmias and William P Pierskalla. Optimal ordering policies for a product that perishes in two periods subject to stochastic demand. *Naval Research Logistics Quarterly*, 20(2):207–229, 1973.

[16] Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *IISE Transactions*, 52(4):444–463, 2020.

[17] Nicholas C Petruzzi and Maqbool Dada. Newsvendor models. In James J. Cochran, Louis A. Cox, Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Colen Smith, editors, *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.

[18] Meng Qi, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, and Zuo-Jun Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, 69(2):759–773, 2023.

[19] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[20] Lawrence V Snyder and Zuo-Jun Max Shen. *Fundamentals of supply chain theory*. John Wiley & Sons, 2019.

[21] Nathalie Vanvuchelen, Joren Gijsbrechts, and Robert Boute. Use of proximal policy optimization for the joint replenishment problem. *Computers in Industry*, 119:103239, 2020.

[22] Arthur F Veinott Jr. Optimal policy for a multi-product, dynamic, nonstationary inventory problem. *Management science*, 12(3):206–222, 1965.

[23] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A Multi-Horizon Quantile Recurrent Forecaster, 2017.

[24] Alice S Whittemore and SC Saunders. Optimal inventory under stochastic demand with two supply options. *SIAM Journal on Applied Mathematics*, 32(2):293–305, 1977.

[25] Yaqi Xie, Will Ma, and Linwei Xin. VC theory for inventory policies. *arXiv preprint arXiv:2404.11509*, 2024.

[26] Paul Zipkin. *Foundations of Inventory Management*. McGraw-Hill, 2000.

[27] Paul Zipkin. Old and new methods for lost-sales inventory systems. *Operations research*, 56(5):1256–1263, 2008.

[28] Paul Zipkin. On the structure of lost-sales inventory models. *Operations research*, 56(4):937–944, 2008.