Drug Inventory Control: Human Decisions versus Deep Reinforcement Learning

Francesco Stranieri[1,3,*], Alberto Archetti[2,3], Enrico Robbiano[4], Chaaben Kouki[5] and Fabio Stella[1]

[1]University of Milano-Bicocca, Milan (20126), Italy
[3]Polytechnic of Turin, Turin (10129), Italy
[2]Polytechnic of Milan, Milan (20133), Italy
[4]Bristol Myers Squibb, Boudry (2017), Switzerland
[4]ESSCA School of Management, Angers (49000), France

Abstract

We investigate whether and how deep reinforcement learning (DRL) can be exploited for managing inventory systems with a specific reference to perishable pharmaceutical products. A real-world case study is formulated as a Markov decision process, where states, actions, and rewards are defined. We then developed a DRL agent based on the Proximal Policy Optimization algorithm and compared its performance with a human decision-maker with several years of experience. Our findings reveal that the DRL agent outperforms the human policy by 11%, optimizing storage space and leading to growing profitability. Such incremental improvements can translate into substantial value for pharmaceutical companies operating in complex scenarios, and patients also stand to benefit. Finally, the study highlights the strategic advantage of integrating DRL into inventory management business operations, particularly for its ability to estimate uncertainty and manage corresponding supply chain risks.

Keywords

inventory management, perishable products, deep reinforcement learning, business operations

1. Introduction

With the advent of artificial intelligence in transforming business operations, its potential in inventory management is worth exploring. In this study, we consider an inventory control system for perishable products. In detail, the product in question is a perishable pharmaceutical drug, and the system under consideration is derived from a *real-world case study* at Bristol Myers Squibb (BMS) [1], a pharmaceutical company offering several types of drugs.

The case we investigate involves a product manufactured by a third party, thereby resulting in longer replenishment lead times compared to direct production, and sold to one of BMS's primary markets through a central warehouse. Our study primarily focuses on various research

*3rd Italian Workshop on Artificial Intelligence and Applications for Business and Industries - AIABI | co-located with AI*IA 2023*

*Corresponding author.
francesco.stranieri@polito.it (F. Stranieri); alberto.archetti@polito.it (A. Archetti); enrico.robbiano@bms.com (E. Robbiano); chaaben.kouki@essca.fr (C. Kouki); fabio.stella@unimib.it (F. Stella)
0000-0002-5366-8499 (F. Stranieri); 0000-0003-3826-4645 (A. Archetti); 0000-0002-1394-0507 (F. Stella)

[1]https://www.bms.com/

trends. On one hand, we explore inventory management of *perishable products*. On the other hand, due to product disturbances, we are also in the context of inventory management with *random yields*.

The collaboration with BMS has two main objectives. The first is to evaluate the potential of deep reinforcement learning (DRL) in deriving competitive inventory policies. Should this approach prove cost-effective, BMS plans to exploit DRL to build simulations that mimic *human-like intelligence*. The second goal of the collaboration consists of objectively evaluating the probability and impact of *supply chain risk*. This assessment considers a range of uncertain factors, including forecast accuracy, supply reliability, quality issues, and external disruptions. The aim of these objectives is not only to minimize significant waste of time and human resources but also to overcome performance limitations, thus aligning with the principles of Industry 4.0.

2. Background and Related Work

DRL has been increasingly applied in inventory management [1, 2, 3], offering prominent solutions that dynamically adapt to varying supply chain settings and effectively address inherent limitations associated with traditional mathematical models [4].

However, since the 1970s, it has been recognized that determining an optimal inventory control policy for perishable products with a fixed lifetime is a significant challenge [5, 6]. While various studies have attempted to develop and characterize the optimal policy, the necessity of tracking the lifetime of items usually complicates the derivation of an effective solution [7]. As a result, numerous studies have focused on *approximate solutions* that perform closely to the optimal policy but often involve making several assumptions that may not correspond to real-world scenarios [8].

Considering the recent development of algorithms based on reinforcement learning (RL), it would be worthwhile to investigate their performance in comparison to traditional reordering policy and human decisions. RL algorithms are designed on the Markov decision process (MDP), which provides a mathematical framework for addressing sequential decision-making problems. Within this framework, an agent interacts with an environment at each time step t, observing a state s_t and responding with an action a_t guided by its policy $\pi(s_t)$. Following the action, the environment transitions to the next state s_{t+1}, rewarding the agent with a scalar value r_t. The ultimate goal of the agent is thus to maximize the long-term cumulative reward $\sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$, with $\gamma \in [0, 1)$, by learning and exploiting an optimal policy [9].

Specifically, RL constitutes a class of algorithms used for solving MDPs which has been significantly enhanced by the advent of *deep learning*. In DRL, neural networks are employed to handle high-dimensional state and action spaces through function approximation. In this study, we utilized a DRL algorithm known as Proximal Policy Optimization (PPO) due to its benefits such as training stability, high performance, and scalability [10].

In the field of periodic inventory control systems for perishable products with a fixed lifetime, there is limited research utilizing DRL. The study by [11] compares the performance of the Deep Q-Network (DQN) algorithm with a fixed reordering policy and other existing heuristics. They demonstrated that DQN outperforms all other methods in most of the considered experiments, underscoring the effectiveness of this class of algorithms. When considering pharmaceutical

perishable products in a healthcare supply chain, [12] found that DRL policies not only result in a reduced probability of product shortage and reduced risk of product expiration but also ensure a higher service level for patients.

3. Supply Chain Case Study

In our case study provided by BMS, orders for drugs produced by the third-party plant are dispatched to the central warehouse in batches of nQ, where n ranges from 0 to 6, while Q represents a fixed quantity of 20 items per batch. In this respect, the ordering costs are nonlinear. In particular, a single batch order incurs a cost of 5, a double batch order costs 8, a triple batch order is priced at 9, and orders ranging from four to six batches are fixed at 10. We assume the warehouse has infinite storage capacity and incurs a holding cost of 1 per item at each time step.

It is crucial to note that not every order received by the warehouse is flawless. In fact, items can be subject to product disturbances during production, rendering them unsaleable. Data provided by BMS suggest that up to 10% of each order can be deemed unsaleable. Items in stock are sold to customers at a price of 30, while any lost sale due to a shortage of items incurs a penalty cost of 10. The replenishment lead time L from the plant to the warehouse is 12 time steps, where each time step corresponds to one month. Items have a lifetime M of 12 time steps after which they expire, resulting in an expiration cost of 3 per item.

One of the main challenges of this study lies in *formulating the system as an MDP* and consequently identifying an inventory policy through DRL that maximizes cumulative reward. We define the reward r_t at time step t as the profit, calculated as sales revenue minus ordering, holding, penalties, and expiration costs. The purpose of the DRL agent is thus to determine, for each time step t, the optimal action a_t, which ranges from 0 to 6 (i.e., the domain of n). We designed the state of the MDP, s_t, as a vector composed of: (i) the current time step t; (ii) a vector of length M, where the m-th element indicates how many items have a lifetime of m time steps; (iii) a vector of length L, where the l-th element indicates how many items were shipped l time steps in the past; and (iv) a vector of length L which predicts the demand forecast between $t + L$ and $t + 2L$ (for simplicity, we assume that the forecast value amounts to the average demand value provided by the company). We made this choice for the vector of length L because the actions taken at time t only affect the system at time step $t + L$. Therefore, considering previous forecasts does not provide useful information to the DRL agent.

Note that all the information contained within the state is also accessible to the human decision-maker. Hence, our research aims to compare the performance of the PPO policy with the human one. It is worth noting that the latter mainly relies on the material requirement planning calculation under the supervision of a human decision-maker, i.e., a planning manager operating at BMS, who can override the suggested actions based on their *expertise*.

4. Experimental Setup and Results

To test the effectiveness of the DRL agent, we compare it against a human inventory control policy by simulating a demand function that mirrors a specific scenario, that is, a pharmaceutical drug *entering the market*, as illustrated in Figure 1. Each simulation covered a duration of 6

years, which is equivalent to 72 months (i.e., $T = 72$). The demand at time step t equals the company-provided average value, with a 15% standard deviation. To obtain this average value, the company supplied us with synthetic and anonymized data corresponding to a real-world case study.

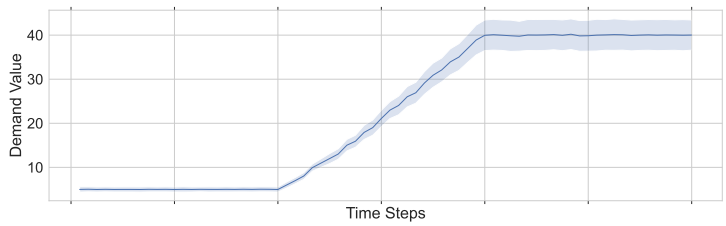


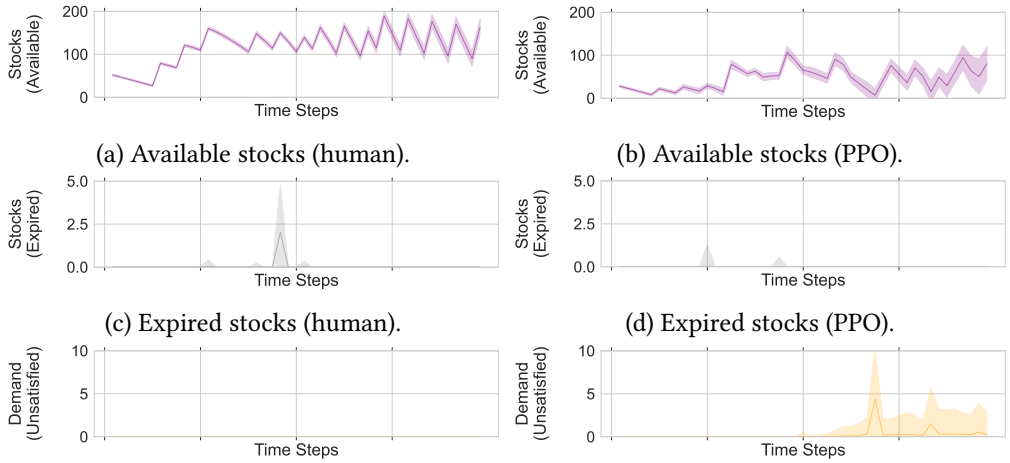
Figure 1: 95% confidence interval computed from 1000 demand realizations over a time horizon of 6 years (i.e., $T = 72$).

Regarding the results, we calculated the *average cumulative profit* simulating, for each of the two policies, the same 1000 independent episodes. However, during the evaluation phase, we excluded the first and last 12 months not to consider the effects of initial stocks and final actions. As reported in Table 1, the human policy yielded a profit of 26367 ± 688. In contrast, the DRL agent, optimized using the PPO algorithm, achieved a profit of 29249 ± 1044. Further analysis of the data indicates that the DRL agent outperforms the human policy by $11\% \pm 4$, as the average gap suggests.

Table 1
Average cumulative profit and standard deviation over 1000 episodes for the human and PPO policies, respectively. The higher the value, the better the algorithm.

Algorithm	Cumulative Profit
Human	26367 ± 688
PPO	29249 ± 1044

We then analyzed the *behavior* of the two policies to understand the reasons behind this difference in terms of profit. As Figure 2 illustrates, the improvement achieved by the DRL agent appears to be rooted in its ability to optimize the available storage space, leading to reduced storage costs. In fact, under the human policy, stocks consistently remain above 100, whereas they fall below this level when using the PPO policy (Figures 2a and 2b). This results the human policy experiencing more expired stocks than PPO, even though the values for both algorithms remain restricted to a few time steps (Figures 2c and 2d). Regarding unsatisfied demand, the human policy fully meets all customer demands for every single time step (Figure 2e). Conversely, during the final peak, PPO falls to entirely satisfying the demand, albeit by a margin that never exceeds 10 units (Figure 2f). From a practical perspective, this difference is both expected and intentional. Indeed, in the pharmaceutical market, it is generally more acceptable to incur additional storage costs than to risk failing to meet customer demand.



(a) Available stocks (human). (b) Available stocks (PPO).

(c) Expired stocks (human). (d) Expired stocks (PPO).

(e) Unsatisfied demand (human). (f) Unsatisfied demand (PPO).

Figure 2: Average value per time step of available stocks (Figures 2a and 2b), expired stocks (Figures 2c and 2d), and unsatisfied demand (Figures 2e and 2f) for the human and PPO policies, respectively.

5. Conclusions

In this study, we investigated the application of DRL to a real-world inventory control system. We formulated the problem as an MDP, defining states, actions, and rewards and implementing the PPO algorithm. Our goal was to benchmark DRL performance to the one achieved by an expert planning manager with several years of experience in the field.

The results indicate a significant improvement in *financial performance*–about an 11% increase–when employing a DRL agent due to its ability to optimize the available storage space, consequently reducing storage and expiration costs. In contexts like pharmaceuticals, where products have expiration dates and notable replenishment lead times, even minor performance improvements can translate into substantial value, yielding considerable benefits for both patients and companies. In fact, reduced obsolescence makes the product more financially sustainable and allows the company to invest more resources in research and development.

For future research, different *demand functions* could be tested, such as those pertaining to drugs already on the market or nearing the end of their exclusivity.

Lastly, to enhance the interplay between decision-making and artificial intelligence, integrating DRL policies into inventory management *business operations* would be a valuable next step; this represents a strategic advantage over traditional material requirement planning calculation, which often lacks accurate estimation of uncertainty and related risks associated with external disruptions, quality issues, and supply reliability.

References

[1] Z. Peng, Y. Zhang, Y. Feng, T. Zhang, Z. Wu, H. Su, Deep reinforcement learning approach for capacitated supply chain optimization under demand uncertainty, in: 2019 Chinese

Automation Congress (CAC), IEEE, 2019. URL: http://dx.doi.org/10.1109/CAC48633.2019.8997498. doi:10.1109/cac48633.2019.8997498.

[2] J. Gijsbrechts, R. N. Boute, J. A. Van Mieghem, D. J. Zhang, Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems, Manufacturing & Service Operations Management 24 (2022) 1349–1368. URL: http://dx.doi.org/10.1287/msom.2021.1064. doi:10.1287/msom.2021.1064.

[3] F. Stranieri, F. Stella, Comparing deep reinforcement learning algorithms in two-echelon supply chains, 2022. URL: https://arxiv.org/abs/2204.09603. doi:10.48550/ARXIV.2204.09603.

[4] F. Stranieri, E. Fadda, F. Stella, Combining deep reinforcement learning and multi-stage stochastic programming to address the supply chain inventory management problem, International Journal of Production Economics 268 (2024) 109099. URL: http://dx.doi.org/10.1016/j.ijpe.2023.109099. doi:10.1016/j.ijpe.2023.109099.

[5] S. Nahmias, Optimal ordering policies for perishable inventory—II, Operations Research 23 (1975) 735–749. URL: https://doi.org/10.1287/opre.23.4.735. doi:10.1287/opre.23.4.735.

[6] B. E. Fries, Optimal ordering policy for a perishable commodity with fixed lifetime, Operations Research 23 (1975) 46–61. URL: https://doi.org/10.1287/opre.23.1.46. doi:10.1287/opre.23.1.46.

[7] S. Nahmias, Perishable inventory theory: A review, Operations Research 30 (1982) 680–708. URL: https://doi.org/10.1287/opre.30.4.680. doi:10.1287/opre.30.4.680.

[8] V. Chaudhary, R. Kulshrestha, S. Routroy, State-of-the-art literature review on inventory models for perishable products, Journal of Advances in Management Research 15 (2018) 306–346. URL: https://doi.org/10.1108/jamr-09-2017-0091. doi:10.1108/jamr-09-2017-0091.

[9] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, MIT press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017. URL: https://arxiv.org/abs/1707.06347. doi:10.48550/ARXIV.1707.06347.

[11] B. J. D. Moor, J. Gijsbrechts, R. N. Boute, Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management, European Journal of Operational Research 301 (2022) 535–545. URL: https://doi.org/10.1016/j.ejor.2021.10.045. doi:10.1016/j.ejor.2021.10.045.

[12] E. Ahmadi, H. Mosadegh, R. Maihami, I. Ghalehkhondabi, M. Sun, G. A. Süer, Intelligent inventory management approaches for perishable pharmaceutical products in a healthcare supply chain, Computers & Operations Research 147 (2022) 105968. URL: https://doi.org/10.1016/j.cor.2022.105968. doi:10.1016/j.cor.2022.105968.