Reinforcement Learning Provides a Flexible Approach for Realistic Supply Chain Safety Stock Optimisation

Edward Elson Kosasih * **Alexandra Brintrup** *

* *Institute for Manufacturing, University of Cambridge (e-mail: eek31@cam.ac.uk).*

Abstract: Although safety stock optimisation has been studied for more than 60 years, most companies still use simplistic means to calculate necessary safety stock levels, partly due to the mismatch between existing analytical methods' emphases on deriving provably optimal solutions and companies' preferences to sacrifice optimal results in favour of more realistic problem settings. A newly emerging method from the field of Artificial Intelligence (AI), namely Reinforcement Learning (RL), offers promise in finding optimal solutions while accommodating more realistic problem features. Unlike analytical-based models, RL treats the problem as a black-box simulation environment mitigating against the problem of oversimplifying reality. As such, assumptions on stock keeping policy can be relaxed and a higher number of problem variables can be accommodated. While RL has been popular in other domains, its applications in safety stock optimisation remain scarce. In this paper we investigate three RL methods, namely, Q-Learning, Advantage Actor-Critic and Multi-agent Advantage Actor-Critic for optimising safety stock in a linear chain of independent agents. We find that RL can simultaneously optimise both safety stock level and order quantity parameters of an inventory policy, unlike classical safety stock optimisation models where only safety stock level is optimised while order quantity is predetermined based on simple rules. This allows RL to model more complex supply chain procurement behaviour. However, RL takes longer time to arrive at solutions, necessitating future research on identifying and improving trade-offs between the use of AI and mathematical models are needed.

Keywords: Supply Chain, Safety Stock, Artificial Intelligence, Inventory Control, Simulation Optimisation, Reinforcement Learning

1. INTRODUCTION

Increasing supply chain complexity and economic uncertainty affects both supply and demand. Supply chain executives need to improve their inventory management strategy to handle emerging uncertainties (Humair et al. (2013)). A 2011 report from Chief Supply Chain Officer (CSCO) Insights found that 76% of surveyed executives listed inventory management excellence as either top priority or highly important (CSCO (2011)).

In this paper, we focus on the problem of end-to-end safety stock placement, which is a subset of the general inventory optimisation problem. The objective of end-to-end safety stock placement is to calculate how much and where to keep extra stock across the whole supply chain to mitigate stockout risks due to supply and demand uncertainty.

The safety stock placement problem, also known as safety stock optimisation, has been studied for more than 60 years, starting from the two seminal papers of Simpson Jr (1958) and Clark and Scarf (1960). Significant emphases have been put on deriving optimisation strategies with provable optimality, often at the expense of simplifying reality. Unfortunately, these limiting assumptions prevent industry adoption (Humair et al. (2013)). The CSCO 2011 report found that many companies, in practice, tend to still use rather simplistic means to calculate safety stock levels (CSCO (2011)). This is partly because companies are typically not as concerned about using the most optimally proven yet oversimplified policy but instead would like models that can reflect their reality (Humair et al. (2013)), pointing to a discrepancy between academic studies and practice.

We hypothesise that a recently emerging technique from the field of AI Reinforcement Learning (RL) could help address this requirement. In RL goal-directed agents learn policies that optimise the performance rewards they would receive from the black-box environment. The agents make limited assumptions about this black-box behaviour. The learning algorithm of RL is designed to be as generic as possible such that if the black-box environment is altered, no algorithmic design changes are required. The agents thus could utilise the same learning algorithm to adapt their behaviour to the new environment. In the case of inventory optimisation, the environment could be defined as the companies' supply chain while the agents represent their safety stock control software.

To test our hypothesis we examine the use of RL by contextualising the safety stock problem as Q-Learning, Ad-

vantage Actor-Critic and Multi-agent Advantage Actor-Critic RL approach. A baseline analytical model and a test problem are developed for cross-comparison of results. Our findings show that while RL is slightly suboptimal compared to the baseline results that are derived analytically, it is able to optimise both safety stock level and order quantity parameters of an inventory policy - requirement that is not possible to fulfill with classical safety stock optimisation modelling where only safety stock level is optimised whilst order quantity rule being a priori predetermined based on a simple rule (such as base stock, (s, S) or (s, nQ) inventory policy). On the other hand, we find that RL suffers from high computational complexity, thus the benefits obtained from RL must be balanced against the time required to obtain solutions.

This paper is organised as follows. First, a literature review is performed to identify existing safety stock optimisation methods and identify research gaps. RL is proposed as a potential method to address these problems. Second, we design a problem to test the RL methods and solve it analytically. Next, we elaborate on the RL algorithmic design used to tackle the problem. Experimental results with the RL algorithms are analysed and discussed, followed by suggestions for future avenues of research.

2. LITERATURE REVIEW

The safety stock placement problem is formulated as following: Minimise the total safety stock costs across all echelons in a supply chain, whilst ensuring that stockouts are prevented. Several comprehensive surveys on safety stock optimisation have been written by Diks et al. (1996), Simchi-Levi and Zhao (2011), Eruguz et al. (2016), and de Kok et al. (2018). There are two classes of methods in the literature, Guaranteed Service Model (GSM) and Stochastic Service Model (SSM), introduced by the two seminal papers Simpson Jr (1958) and Clark and Scarf (1960) respectively. GSM assumes that all demand will be fulfilled in fixed lead time, using extraordinary means like outsourcing or expediting if needed. Meanwhile, SSM allows some demand to be backordered, resulting in stochastic lead time. Klosterhalfen and Minner (2010) and Eruguz et al. (2016) have concluded that research on GSM has attracted more interests lately, and it performs slightly better than SSM given moderate cost for flexibility measures. Following this trend, we focus on GSM in this work.

GSM formulates agent's inventory as a sum of fixed mean demand (μ_j) and safety stock (SS_j). Since the former is fixed, only the latter part is optimised. Equation 1 shows the original GSM safety stock formula as proposed by Simpson Jr (1958). This equation is modified with various additional variables across the literature.

$$SS_j = z_j \times \sigma_j \times \sqrt{SI_j + T_j - S_j} \tag{1}$$

where z_j is service level, σ_j is standard deviation of demand seen at agent j, SI_j is service time of agent j's supplier, T_j is processing time and S_j is service time guaranteed to customer. In other words, safety stock is proportional to lead time ($SI_j + T_j - S_j$). Consequently, the total agent's inventory $I_j = \mu_j \times (SI_j + T_j - S_j) + z_j \times \sigma_j \times \sqrt{SI_j + T_j - S_j}$.

There are two main paradigms to solve this optimisation problem in the literature: analytical/mathematical programming and simulation-based models. Most literature used analytical/mathematical programming approaches. Simpson Jr (1958) and Inderfurth (1991) proved that in a serial and distribution supply chain, the optimal solution lies at the vertices of the solution set, hence one can enumerate all vertices and find the best value. Besides enumeration, various approaches have been proposed, such as dynamic programming (Graves and Willems (2000), Inderfurth (1991), and Minner (2012)), branch-and-bound (Lesnaia (2004)), piece-wise linear approximation (Magnanti et al. (2006), Shu and Karimi (2009)) and heuristics algorithms (Minner (2012), Li and Jiang (2012)). Meanwhile, alternative simulation-based optimisation methods have not been explored much in the literature (de Kok et al. (2018)). According to Glasserman and Tayur (1995) and Klemmt et al. (2009), simulation-based models work better with real system. In practice, this means that companies can use their ERP system as a black box plugged into the model, without the need to write down every subsystem interaction in a set of explicit equations. Some existing works on simulation-based optimisation in the literature included Monte Carlo (Sitompul et al. (2008)), hybrid (Glasserman and Tayur (1995)), upper limit simulation (Schoenmeyr (2008)), simulated annealing (Molinder (1997)) and Gaussian Process (Agarwal (2019)). The main limitation of simulation-based approach is, while they can find an optimal solution in realistic situation, they could not prove that this is the best possible value.

There is also an implicit modeling assumption that existing literature has made. In real-life, supply chain managers would not only decide the amount of safety stock (inventory policy), but also how and when should they order the stock (procurement policy). Existing safety stock optimisation models assumed that the procurement policy is fixed beforehand, usually with one of the most commonly defined policies in the literature (de Kok et al. (2018)), such as base stock (continuously place order to keep stock this base level); [s, S] (place order up to base level if inventory falls below reorder point s); or [s, nQ] (order with integer n multiplication of predetermined order quantity Q if inventory falls below reorder point s). While this implicit assumption can be proven to be optimal for simplified problems, in real complex system, such optimality might no longer be true.

We hypothesise that a family of machine learning algorithms called Reinforcement Learning (RL), could address the two aforementioned research gaps i.e. lack of simulation-based optimisation methods and predetermined inventory policy. RL trains agents to learn how to act in a given situation to optimise their rewards (Sutton and Barto (2018)). RL agents interact with simulation-based black box system to learn both the optimum state to be in, as well as the actions needed to achieve that state. This value-policy duality is analogous to our safety stock-procurement policy task; hence our proposal to investigate RL in this paper. Figure 1 shows the difference between

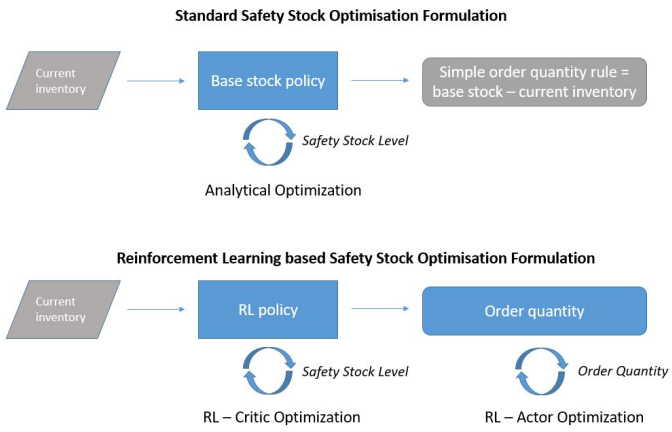

Fig. 1. In standard safety stock optimisation models, only the safety stock level is optimised, while the order quantity is based on a simple rule. Meanwhile, in RL-based models, both safety stock level and order quantity rule are optimised.

current safety stock optimisation model and our proposed RL approach.

While RL has been successfully in other difficult task, such as the game of Go (Silver et al. (2017)), its use in supply has been limited. Oroojlooyjadid et al. (2017) used Deep Q-Network to solve the beer distribution game. Kara and Dogan (2018) uses Q-Learning and SARSA to specify ordering policies of perishable inventory systems. Gijsbrechts et al. (2019) uses Asynchronous Advantage Actor-Critic (A3C) algorithm to solve multi-echelon inventory problem, and Jiang and Sheng (2009) uses case-based RL to learn procurement policies. To the best of our knowledge, none of the existing works explored safety stock optimisation.

We contribute to these research gaps by developing a simulation-based RL model that can optimise both safety stock level and procurement policy.

3. PROBLEM DESIGN

In order to test the efficacy of our proposed RL model, we select a standard serial supply chain problem whose analytical solution has been proven. This allows us to compare the gap in optimality of our solution. We benchmark the analytical solution with three commonly used RL algorithms: Q-Learning, Advantage Actor-Critic and Multi-agent Advantage Actor-Critic.

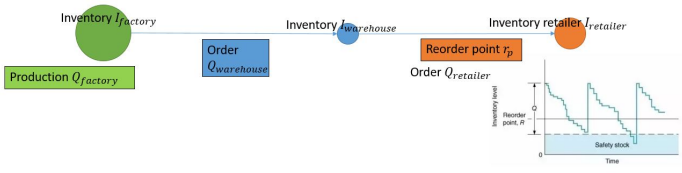

Fig. 2. Simulation Design

We develop a simulation of a pull-based (only place order upon receiving demand) serial supply chain, consisting of 3 agents: retailer, warehouse and factory. Retailer agent adopts a (Q, r_p) procurement policy with Q being order quantity and r_p reorder point. Both factory and warehouse agents learn the procurement policy with RL. Any extra items ordered will be kept as safety stock for the next cycle. Any orders unmet will incur stockouts.

The agents are trained using RL to minimise the total amount of safety stock in the system, while avoiding stockouts. We assume a cooperative game setup where agents work together. We perform two sets of experiments with different cost assignments. Using the inventory optimisation model typology developed by de Kok et al. (2018), our proposed method belongs to category 3, S, D, G | F, D | N, G, | O, F, U | M || S, O.

4. ANALYTICAL-BASED OPTIMISATION

We first derive the analytical solution to this problem. Based on Equation 1, we can formulate this as a GSM concave minimisation over bounded polyhedron, as following:

Minimise

$$h_f z_f \sigma \sqrt{0+1-S_f} + h_w z_w \sigma \sqrt{S_f+3-S_w} \quad (2)$$

subject to

$$S_f \leq 0+1 \quad (3)$$
$$S_w \leq S_f+3 \quad (4)$$
$$S_w \leq 3 \quad (5)$$

where h_f and h_w are inventory storage cost of factory f and warehouse w accordingly. We also assume σ is 1 and $z_w = z_h = 3$. The total inventory for factory and warehouse can be written as the following: $I_f = \mu \times (1 - S_f) + 3 \times 1 \times \sqrt{1-S_f}$ and $I_w = \mu \times (S_f + 3 - S_w) + 3 \times 1 \times \sqrt{S_f + 3 - S_w}$.

We consider two cases with different inventory storage cost assignments and solve the GSM problem using enumeration, as shown below.

Case 1: Let $h_f = 1000$ **and** $h_w = 5$. The optimal solution is $S_f = 1, S_w = 3$, or $I_f = 0, I_w = 13, r_p = 6$. In other words, factory keeps no stock.

Case 2: Let $h_f = 5$ **and** $h_w = 1000$. The optimal solution is $S_f = 0, S_w = 3$, or $I_f = 13, I_w = 0, r_p = 6$. Here, warehouse keeps no stock.

Observe that while this analytical-based approach provides the optimal safety stock allocation, it does not explicitly optimise for the procurement policy. This is because it implicitly assumes that procurement policy is predetermined as base stock (Eruguz et al. (2016)).

5. SIMULATION-BASED OPTIMISATION WITH RL

The serial supply chain simulation that we have built consists of cycles of agents procuring and fulfilling demands. At the end of every cycle, the following joint reward is provided to all agents. This reward is negatively proportional to inventory and stockouts (it costs η to deal with a stockout). Agents are trained to maximise rewards i.e. minimise inventory and stockouts.

$$\text{reward} = -\{h_f \times I_f + h_w \times I_w + \eta \times \text{stockouts}\} \quad (6)$$

5.1 Tabular Q-Learning

The agent sees its own inventory and demand from customer, then decides the amount to order from its supplier accordingly. Here, the agent is trained with Q-Learning, which is basically a lookup table that store the optimal order amount for each combination of inventory and demand.

5.2 Advantage Actor-Critic (A2C)

Q-Learning assumed discrete actions, hence it does not scale well with the problem size. We implement another policy gradient algorithm called A2C that works with continuous space, hence could scale better. Here, we represent all agents as a joint distribution using two neural networks, one called critic (to learn optimal safety stock level) and another called actor (to learn procurement policies for all agents). We chose default hyperparameters that are prescribed by the tensorflow-keras software package.

5.3 Multi-agent Advantage Actor-Critic

A2C modeled joint probability of all agents, hence the parameters scale exponentially with the number of agents. We investigated another model called multi-agent A2C where each agent is represented with individual actor neural networks, each seeing local information on inventory and demand. However, we still require a centralised critic that can see all agents' information. Developing decentralised critic is a potential future extension of this model.

6. EXPERIMENTAL RESULTS

Experiments were performed on a Dell laptop with Intel i9-9980HK CPU, 2.4 GHz processor and 16 GB RAM.

6.1 Tabular Q-Learning

Table 1 shows the resulting 95% CI average inventory level for both case 1 and case 2 when the agents are trained from scratch 10 times with different random initial conditions.

Table 1. Q-Learning Optimal Inventory Level

	Case 1 RL	Case 1 Analytical	Case 2 RL	Case 2 Analytical
r_p	[1.02, 1.12]	6	[3.58, 3.86]	6
I_w	[11.22, 13.34]	13	[1.29, 1.7]	0
I_f	[0.41, 0.98]	0	[11.91, 13.18]	13

As hypothesised, the result shows that the agent's behaviour is suboptimal but close to the analytically-derived solutions. However, the main caveat is the joint action space scales exponentially with the number of agents, and the model cannot generalise across states given the tabular entry modelling.

6.2 Advantage Actor-Critic

From Table 2 we can observe the resulting 95% CI average inventory level for both case 1 and case 2 when the agents are trained 10 times from different random initial conditions.

Table 2. A2C Optimal Inventory Level

	Case 1 RL	Case 1 Analytical	Case 2 RL	Case 2 Analytical
r_p	[2.11, 2.84]	6	[3.18, 3.77]	6
I_w	[12.58, 15.11]	13	[0.37, 0.85]	0
I_f	[1.66, 2.87]	0	[8.97, 11.24]	13

The results show that while the learn solutions are suboptimal, they are close to the analytical solutions. Besides inventory level, A2C also provides additional insights as we can visualise both the actor's and the critic's behavior as shown in Figure 3.

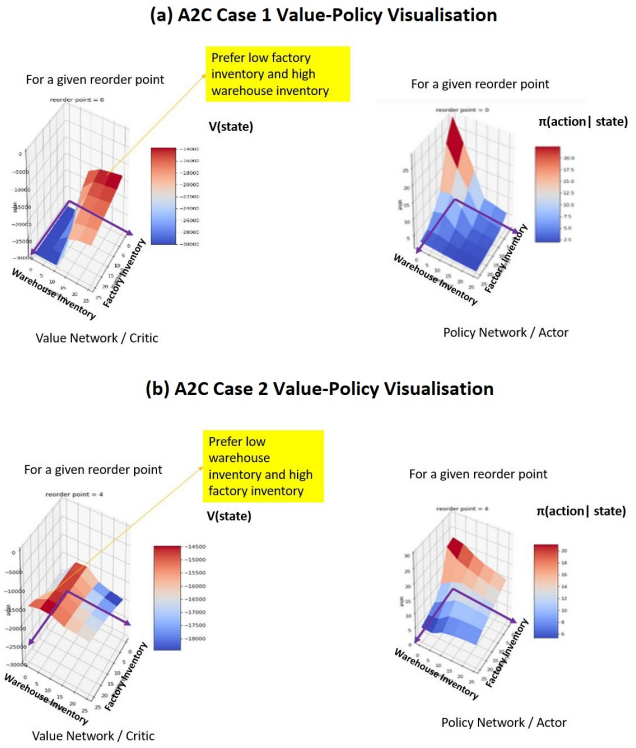

Fig. 3. A2C Value-Policy Visualisation

Figure 3(a) indicates that state $I_f = 0$ and $I_w > 13$ are preferable for case 1, similar to the analytical solution. Looking at the policy network, we see that Q_f is always kept to a minimum unless I_f and I_w approaches zero. This is similar to a lean procurement policy for the factory. Meanwhile, Figure 3(b) shows that for case 2 state $I_f > 13$ and $I_w = 0$ are preferable. The policy network shows that Q_f is always placed if I_f approaches zero, unless if I_f is too much; then Q_f will be cut down to zero, encouraging Factory to fulfill demand from existing stocks. This results in a consistent, innovative stock-keeping procurement policy.

6.3 Multi-agent Advantage Actor-Critic

Similar to Q-Learning and A2C, we train the agents 10 times from different random initial conditions for both cases 1 and 2, resulting in 95% CI average inventory level as seen in Table 3.

Table 3. Multi-agent A2C Optimal Inventory Level

	Case 1 RL	Case 1 Analytical	Case 2 RL	Case 2 Analytical
r_p	[2.73, 3.09]	6	[2.94, 3.46]	6
I_w	[13.98, 19.46]	13	[1.97, 2.96]	0
I_f	[0.58, 1.64]	0	[8.64, 10.98]	13

The optimisation results are close to the analytical solutions. Multi-agent A2C, akin to the previous A2C algorithm, also provides additional insights as we can visualise both actor's and critic's behavior as shown in Figure 4.

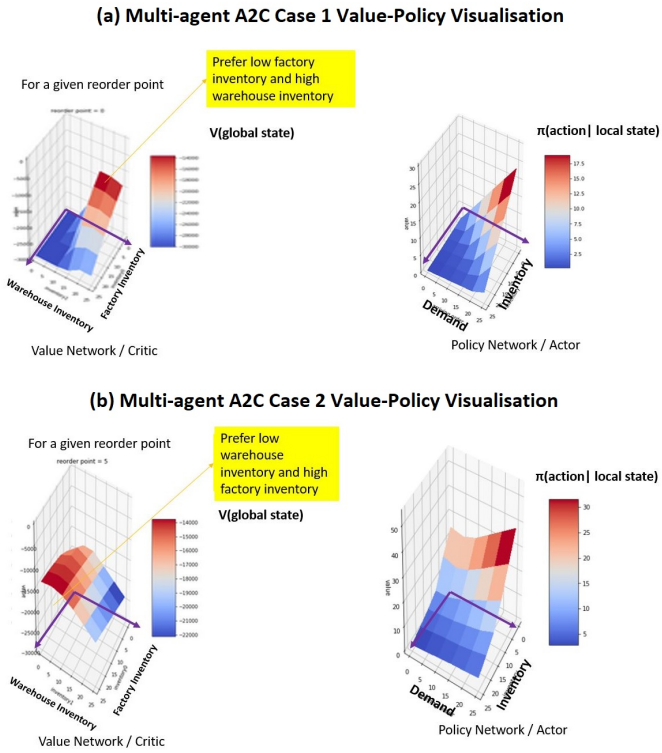

Fig. 4. Multi-agent A2C Value-Policy Visualisation

Figure 4(a) shows that agents prefer state $I_f = 0$ and $I_w > 13$, as predicted by the analytical solution. Similar to A2C, the policy network shows that Factory will only produce Q_f if $I_f = 0$ and it receives demand Q_w from Warehouse. This results in a lean procurement policy. Meanwhile, Figure 4(b) shows that state $I_f > 13$ and $I_w = 0$ are preferable. Here, Factory will always produce Q_f if I_f approaches zero regardless of demand, unless if I_f is too much then Q_f will be cut down to a minimum, encouraging Factory to use existing stocks to fulfill demand. Similar to A2C, this results in a consistent, innovative stock-keeping procurement policy.

6.4 Complexity Study

We compare the performance of the three algorithms: tabular Q-Learning (*single_Q*), A2C (*single_PG*) and multi-agent A2C (*multi_PG*) in terms of the time taken to do inference i.e. execution time. Figure 5 shows that Tabular Q-Learning is faster than both A2C and multi-agent A2C. This is expected since tabular Q-Learning performs a simple table lookup and update. Hence, more fundamental research is needed to speed up both A2C and multi-agent A2C.

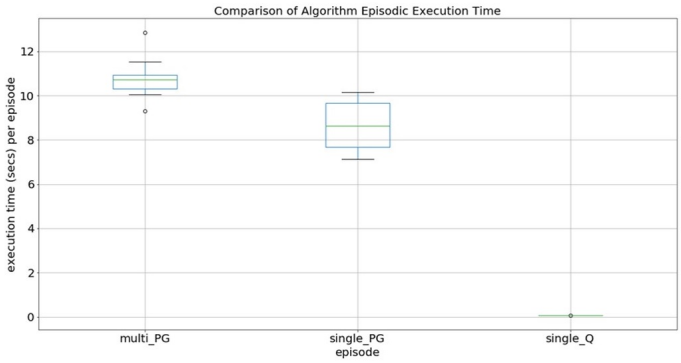

Fig. 5. Execution Time Comparison

7. DISCUSSION AND CONCLUSIONS

The safety stock optimisation problem has been studied for decades. Nevertheless, it is puzzling why many companies, in practice, tend to still use rather simplistic means to calculate safety stock level CSCO (2011). One reason is highlighted by Humair et al. (2013), where most emphases have been put on deriving algorithms with provable optimality, often at the expense of simplifying reality. However, companies are typically not as concerned about using the most optimally proven yet oversimplified policy. Instead, they prefer to utilise solutions that accommodates their actual pain points.

In this paper, we have shown that we could train RL algorithms to handle complex environments dynamically and as such, being able to optimise both safety stock level and order quantity rule simultaneously. With RL, we focus more on developing a practically adaptive solution, while not putting too much emphasis on proving optimality.

However, we have also encountered several challenges with RL-based optimisation algorithms that are worthy for future investigation. First, while neural network based policies could capture more complex behaviour, they are often difficult to interpret. In the wider literature of machine learning, this is often referred to as the explainability problem. Explainability is a concern for practitioners, as it reduces trust in the algorithm and its output, especially because optimality of an RL-based solution cannot be mathematically proven – it is typically assumed that convergence of reward is equivalent to reaching optimality. Second, the convergence rate and execution time of RL-based models are still slow relative to analytical-based mathematical programming approaches. It is currently still difficult to scale RL to a very large number of agents. However, researchers have actively been working on tackling this problem. Our experiment shows that certain classes of RL, like the Multi-Agent A2C, could scale better than traditional Q-Learning. Nevertheless, more investigation is needed to improve its' execution time. Therefore practically speaking, in the current state, there is a trade off between deploying an RL algorithm for a more realistic, out of the box solutions, and the time it takes to arrive at solutions.

REFERENCES

Agarwal, A. (2019). Multi-echelon Supply Chain Inventory Planning using Simulation-Optimization with Data Resampling. *arXiv preprint arXiv:1901.00090*.

Clark, A.J. and Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4), 475–490. ISBN: 0025-1909 Publisher: INFORMS.

CSCO (2011). Five strategies for improving inventory management across complex supply chain networks.

de Kok, T., Grob, C., Laumanns, M., Minner, S., Rambau, J., and Schade, K. (2018). A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research*, 269(3), 955–983. ISBN: 0377-2217 Publisher: Elsevier.

Diks, E.B., de Kok, A.G., and Lagodimos, A.G. (1996). Multi-echelon systems: A service measure perspective. *European Journal of Operational Research*, 95(2), 241–263. doi:10.1016/S0377-2217(96)00120-8.

Eruguz, A.S., Sahin, E., Jemai, Z., and Dallery, Y. (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics*, 172, 110–125. doi: 10.1016/j.ijpe.2015.11.017.

Gijsbrechts, J., Boute, R.N., Van Mieghem, J.A., and Zhang, D. (2019). Can Deep Reinforcement Learning Improve Inventory Management? Performance on Dual Sourcing, Lost Sales and Multi-Echelon Problems. *Performance on Dual Sourcing, Lost Sales and Multi-Echelon Problems (July 29, 2019)*.

Glasserman, P. and Tayur, S. (1995). Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Science*, 41(2), 263–281. ISBN: 0025-1909 Publisher: INFORMS.

Graves, S.C. and Willems, S.P. (2000). Optimizing strategic safety stock placement in supply chains. *Manufacturing & Service Operations Management*, 2(1), 68–83. ISBN: 1523-4614 Publisher: INFORMS.

Humair, S., Ruark, J.D., Tomlin, B., and Willems, S.P. (2013). Incorporating Stochastic Lead Times Into the Guaranteed Service Model of Safety Stock Optimization. *INFORMS Journal on Applied Analytics*, 43(5), 421–434. doi:10.1287/inte.2013.0699. Publisher: INFORMS.

Inderfurth, K. (1991). Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics*, 24(1-2), 103–113. ISBN: 0925-5273 Publisher: Elsevier.

Jiang, C. and Sheng, Z. (2009). Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Systems with Applications*, 36(3), 6520–6526. ISBN: 0957-4174 Publisher: Elsevier.

Kara, A. and Dogan, I. (2018). Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. *Expert Systems with Applications*, 91, 150–158. ISBN: 0957-4174 Publisher: Elsevier.

Klemmt, A., Horn, S., Weigert, G., and Wolter, K.J. (2009). Simulation-based optimization vs. mathematical programming: A hybrid approach for optimizing scheduling problems. *Robotics and Computer-Integrated Manufacturing*, 25(6), 917–925. ISBN: 0736-5845 Publisher: Elsevier.

Klosterhalfen, S. and Minner, S. (2010). Safety stock optimisation in distribution systems: a comparison of two competing approaches. *International Journal of Logistics: Research and Applications*, 13(2), 99–120. ISBN: 1367-5567 Publisher: Taylor & Francis.

Lesnaia, E. (2004). *Optimizing safety stock placement in general network supply chains*. Ph.D. thesis, Massachusetts Institute of Technology.

Li, H. and Jiang, D. (2012). New model and heuristics for safety stock placement in general acyclic supply chain networks. *Computers & Operations Research*, 39(7), 1333–1344. ISBN: 0305-0548 Publisher: Elsevier.

Magnanti, T.L., Max Shen, Z.J., Shu, J., Simchi-Levi, D., and Teo, C.P. (2006). Inventory placement in acyclic supply chain networks. *Operations Research Letters*, 34(2), 228–238. doi:10.1016/j.orl.2005.04.004.

Minner, S. (2012). *Strategic safety stocks in supply chains*, volume 490. Springer Science & Business Media.

Molinder, A. (1997). Joint optimization of lot-sizes, safety stocks and safety lead times in an MRP system. *International Journal of Production Research*, 35(4), 983–994. ISBN: 0020-7543 Publisher: Taylor & Francis.

Oroojlooyjadid, A., Nazari, M., Snyder, L., and Takáč, M. (2017). A Deep Q-Network for the Beer Game: A Deep Reinforcement Learning algorithm to Solve Inventory Optimization Problems. *arXiv preprint arXiv:1708.05924*.

Schoenmeyr, T.T.I. (2008). *Strategic inventory placement in multi-echelon supply chains: Three essays*. Ph.D. thesis, Massachusetts Institute of Technology.

Shu, J. and Karimi, I.A. (2009). Efficient heuristics for inventory placement in acyclic networks. *Computers & Operations Research*, 36(11), 2899–2904. doi: 10.1016/j.cor.2009.01.001.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359. doi: 10.1038/nature24270. Number: 7676 Publisher: Nature Publishing Group.

Simchi-Levi, D. and Zhao, Y. (2011). Performance Evaluation of Stochastic Multi-Echelon Inventory Systems: A Survey. doi:https://doi.org/10.1155/2012/126254. ISSN: 1687-9147 Library Catalog: www.hindawi.com Pages: e126254 Publisher: Hindawi Volume: 2012.

Simpson Jr, K.F. (1958). In-process inventories. *Operations Research*, 6(6), 863–873. ISBN: 0030-364X Publisher: INFORMS.

Sitompul, C., Aghezzaf, E.H., Dullaert, W., and Landeghem, H.V. (2008). Safety stock placement problem in capacitated supply chains. *International Journal of Production Research*, 46(17), 4709–4727. ISBN: 0020-7543 Publisher: Taylor & Francis.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement learning: An introduction*. MIT press.