Reinforcement Learning for Profit Optimization in Sustainable Inventory Management of Perishables Products

Limainla Kichu
Department of Science & Humanities(Mathematics)
National Institute of Technology Nagaland
Chumukedima, India
aienmkichu365@gmail.com

Prem Prakash Mishra
Department of Science & Humanities(Mathematics) National Institute of Technology Nagaland
Chumukedima, India
maths.prem79@gmail.com

***Abstract*— In response to growing environmental concerns, governments are increasingly imposing taxes on carbon emissions, prompting businesses to adopt green inventory management practices to mitigate the environmental impact. This study focuses on a sustainable inventory model for perishable goods, where demand is influenced by both the selling price and investments in environmentally friendly practices. To accurately represent the spoilage of perishable items over time, the deterioration rate is modeled using a Weibull distribution. The goal of this model is to maximize overall profit by determining the optimal cycle time and green investment level. This optimization is achieved through the application of Reinforcement Learning (RL), specifically utilizing the Q-learning algorithm to effectively adjust the decision variables for improved outcomes.**

***Keywords*— Inventory management, carbon-emissions, deterioration, reinforcement learning, Q-learning**

I. Introduction

Excessive carbon emissions are negatively impacting our environment. In response, the government has begun imposing fines and taxes on these emissions. Consequently, the need for sustainable management to reduce carbon dioxide and greenhouse gases has become a pressing issue [1]. For instance, the frequency of supply deliveries could be just as crucial in reducing carbon emissions as the energy efficiency of the vehicles used for those deliveries [2]. Adjusting operational factors such as optimal batch sizes or order quantities, have been shown to effectively reduce the impact of carbon emissions [3, 4]. Consequently, there has been extensive research in recent years on green inventory and sustainable supply chain management, as highlighted in [5-7]. In this study, we focus on sustainable inventory management where demand depends on the selling price and green investment cost. The spoilage rate of the perishable product is characterized using a Weibull distribution to accurately capture the time-dependent nature of deterioration [8]. The problem will be solved by with Reinforcement learning by utilizing the Q- learning algorithm [12]. In machine learning, reinforcement learning excels at tackling complex sequential decision-making tasks by leveraging prior knowledge. It adapts dynamically to changing environments without requiring a pre-defined environmental model [9]. [10] examines perishable goods, comparing stock and age-based ordering policies using Reinforcement Learning (RL) to minimize retailer costs, and demonstrates that age-based policies and RL algorithms outperform traditional approaches, especially with high demand variance and short product lifetimes. RL approach typically involves four core elements: the agent, the state, the action, and the reward [11].

The structure of the paper is outlined as follows. Section II provides the formulation of the sustainable inventory management. The components of reinforcement learning are explained in Section III. Details of the numerical experiment is given in Section IV and concludes the study in Section V.

II. FORMULATION OF THE SUSTAINABLE INVENTORY MANAGEMENT PROBLEM

The following assumptions are made in developing the model:

A. Notations

A: ordering cost per cycle (*\$/cycle*)
C_p: purchasing cost per unit item (*\$/unit item*)
C_h: holding cost per unit item (*\$/unit item)*
S_p: selling price per unit item (*\$/unit item*)
C_d: deterioration cost per unit item (*\$/unit item*)
G_c: green investment cost per unit time (*\$/unit time*)
C_v: variable amount of carbon emission cost for holding the inventory and for the deteriorated items
C_f: fixed amount of carbon emission for holding the inventory and deteriorated items
λ: extra tariff for carbon emission *(\$/unit carbon*)
$w(t)$: deterioration rate dependent on time
$D(S_p, G_c)$: demand dependent on unit selling price and green investment cost
T: total cycle time (in years)

B. Assumptions

- To minimize carbon emissions, investment in green technology is made. The fraction of the reduction of the average carbon emission is given by $V(G_c) = \rho(1 - e^{-\eta G_c})$, where $0 < \rho < 1$ denotes the fraction of the carbon emission after investment in green technology. $0 < \eta < 1$ is the sensitivity parameter reflecting the responsiveness of emissions reduction to the green technology investment cost [13].
- In our study our demand function is given as $D\left(S_p, G_c\right) = a - bS_p + \gamma\, V(G_c)$, where $a > 0$ is the constant demand of the market, $0 < b < 1$ is the selling price sensitivity coefficient, γ is the constant coefficient of $V(G_c)$ respectively.
- The items deteriorate at the rate of $w(t) = \alpha\beta t^{\beta-1}(\alpha > 0, \beta \geq 1, t > 0)$, which is a two parameter Weibull distribution [8]. The deterioration rate of the perishable products is an increasing function of time. The deteriorated items are not replaced or repaired during the period $[0, T]$.

- Replenishment time is instantaneous and shortage is not allowed.

C. Model Formulation

The inventory level with respect to time is governed by the differential equation

$$\frac{d\,I(t)}{dt} = -D(S_p, G_c) - w(t)I(t)$$

Solving the equation, and applying the initial condition at $t = 0, I = Q_0$ we get the solution of $I(t)$ as

$$I(t) = \left(-D(S_p, G_c)\left(t + \frac{\alpha t^{\beta+1}}{\beta+1}\right) + Q_0\right) e^{-\alpha t^{\beta}} \quad (1)$$

When the inventory reaches time total cycle time T, we apply the boundary condition for equation (1) at time $t = T, I = 0$.

$$\Rightarrow Q_0 = D(S_p, G_c)\left(T + \frac{\alpha T^{\beta+1}}{\beta+1}\right) \quad (2)$$

On which Q_0 gives the total order quantity. Substituting the value of Q_0 in equation (1) gives the maximum inventory level at time t.

$$\Rightarrow I(t) = D(S_p, G_c)\left((T - t) + \frac{\alpha}{\beta+1}\left(T^{\beta+1} - t^{\beta+1}\right)\right) e^{-\alpha t^{\beta}} \quad (3)$$

D. Cost Components

i. **Fixed ordering cost per order: A**

ii. **Holding cost of the inventory during one cycle time:**

$$C_h \int_0^T I(t)dt = C_h \int_0^T \left(-D(S_p, G_c)\left(t + \frac{\alpha t^{\beta+1}}{\beta + 1}\right) + Q_0\right) e^{-\alpha t^{\beta}}\, d$$

$$\Rightarrow C_h \int_0^T I(t)dt = C_h \left\{Q_0\left(T - \frac{\alpha T^{\beta+1}}{\beta + 1}\right) - D(S_p, G_c)\left(T^2 - \frac{\alpha\beta T^{\beta+2}}{(\beta + 1)(\beta + 2)} - \frac{\alpha^2 T^{2(\beta+1)}}{2(\beta + 1)}\right)\right\}$$

iii. **Deterioration cost**

$$= C_d w(t) \int_0^T I(t)dt$$

$$= C_d \alpha\beta \left\{\frac{Q_0 T^{\beta}}{\beta} - D(S_p, G_c)\left(\frac{T^{\beta+1}}{\beta + 1} + \frac{\alpha T^{2\beta+1}}{(\beta + 1)(2\beta + 1)}\right) - \alpha\frac{Q_0 T^{2\beta}}{2\beta} + \alpha D(S_p, G_c)\left(\frac{T^{2\beta+1}}{2\beta} + \frac{\alpha^2\, T^{3\beta+1}}{(\beta + 1)(3\beta + 1)}\right)\right\}$$

iv. **Green technology investment cost per cycle time: $\boldsymbol{G_c\, T}$**

v. **Total tax on carbon emission:** $\boldsymbol{\lambda\left(C_v \int_0^T I(t)dt\left(1+w(t)\right)+C_r\right)}$

The reduction of carbon emissions is facilitated by investments in green technology. The overall cost of carbon emissions resulting from these investments is as follows:

$$=\left(1-V(G_c)\right)\lambda\left(C_v \int_0^T I(t)dt\left(1+w(t)\right)+C_r\right)$$

vi. **Sales Revenue:** $\boldsymbol{S_p\, Q_0}$

vii. **Purchasing cost:** $\boldsymbol{C_p Q_0}$

Now, the objective of the proposed model is to maximize the total profit per cycle time with respect to the decision variables G_c and T through Reinforcement Learning.

The total profit per cycle time is represented by $TP(G_c,T) = Sales\ revenue - purchasing\ cost - fixed\ ordering\ cost - holding\ cost - deterioration\ cost - green\ investment\ cost - total\ carbon\ tax\ after\ green\ investment$

$$\begin{aligned}\Rightarrow TP(G_c,T) = {} & S_p\, Q_0 - C_p Q_0 - A \\ & - C_h\left\{Q_0\left(T-\frac{\alpha T^{\beta+1}}{\beta+1}\right)-D\left(S_p,G_c\right)\left(T^2-\frac{\alpha\beta T^{\beta+2}}{(\beta+1)(\beta+2)}-\frac{\alpha^2 T^{2(\beta+1)}}{2(\beta+1)}\right)\right\} \\ & - C_d\alpha\beta\left\{\frac{Q_0 T^{\beta}}{\beta}-D\left(S_p,G_c\right)\left(\frac{T^{\beta+1}}{\beta+1}+\frac{\alpha T^{2\beta+1}}{(\beta+1)(2\beta+1)}\right)-\alpha\frac{Q_0 T^{2\beta}}{2\beta}\right. \\ & \left. + \alpha D\left(S_p,G_c\right)+\left(\frac{T^{2\beta+1}}{2\beta}+\frac{\alpha^2\, T^{3\beta+1}}{(\beta+1)(3\beta+1)}\right)\right\} - G_c T \\ & - \left(1-V(G_c)\right)\lambda\left(C_v\int_0^T I(t)dt\left(1+w(t)\right)+C_r\right) \qquad (4)\end{aligned}$$

III. REINFORCEMENT LEARNING FOR SUSTAINABLE INVENTORY MANAGEMENT

The primary components of reinforcement learning include the state and action variables and the reward function aimed at maximizing average total profit. This study employs the Q-learning algorithm to address the proposed green inventory management problem.

State variable: The state represents a current situation or environment that the agent is currently in. In our study the two states variable defined are the green investment cost G_c and the total cycle time T. The state variable $S(t)$ at any time step t of G_{c_t} and T_t at any time step t respectively is defined by, $S(t) = \left[G_{c_t}, T_t\right]$ (5)

Equation (4) uniquely identifies the state in which the agent is in. In order to find the near optimal values which will give the maximum total profit $TP(G_c,T)$,the values of G_c and T are discretized and are given in the range $10 \le G_c \le 550$ and $0 \le T \le 0.5$.

Action variable: The action involves choosing a new pair of G_{c_t} and T_t. The agent either do exploration with a probability of ϵ and select a random action which is done by choosing random time step for G_{c_t} and T_t. Or the agent with a probability of 1- ϵ do exploitation by selecting the action that

ISBN : 978-81-978522-2-0

gives us the maximum current Q-value for the state it is in. After taking the action and transitioning to the new state, the Q-value for the previous state-action pair is updated using the Q-learning update rule:

$$Q(s,a) \leftarrow oldQ(s,a) + \alpha' \left(reward + \gamma' \max Q(s,a') - oldQ(s,a)\right) \qquad (6)$$

Here, α' is the learning rate, γ' is the discount factor and the term $\max Q(s,a')$ represents the highest Q-value in the new state by predicting the best possible future reward function.

Reward function: The reward function considered in this model is the average total profit $TP(G_c, T)$. To maximize the total profit, we must determine the near-optimal green cost and the total cycle time. Fig.1 describes the flowchart of the Q-learning.

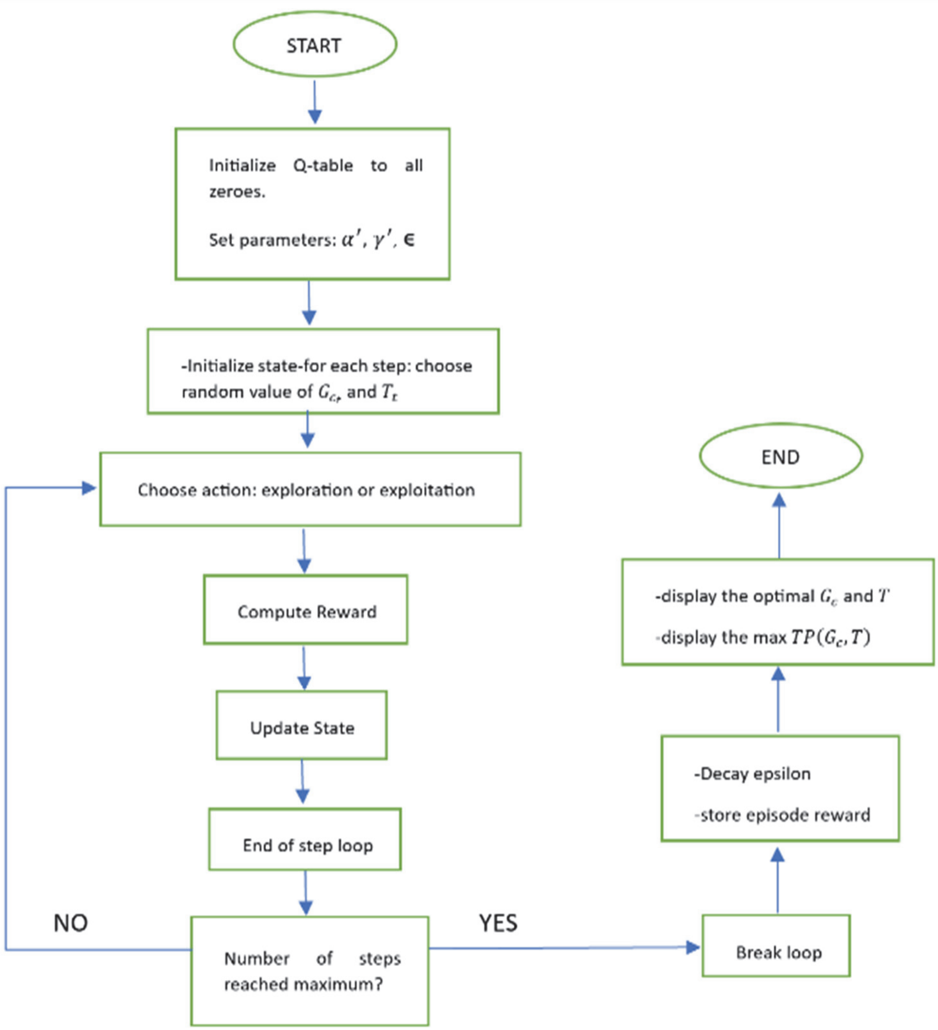

Fig.1. Flow Chart of the Q-learning algorithm to solve the sustainable inventory management

IV. NUMERICAL ILLUSTRATION

Following are the parameters used for experimenting our proposed model.

A=50\$, C_p=25\$/unit item, C_h =3(*\$/unit item), S_p*=100 (*\$/unit item*), C_d=2 (*\$/unit item*), $C_v = 0.5\$/item$, C_f=5\$, $\lambda = 10\$$, α=0.5, β=2, a=1000, b=0.2, η=0.01,ρ=0.2,γ=0.6. Using the Q-learning algorithm and setting the parameters $\alpha' = 0.4, \gamma' = 0.9, \epsilon = 0.995$ we get the optimal $G_c = 359.09\$$ and $T = 0.45(in\ years)$. Corresponding to the optimal values of G_c and T the average total profit $TP(G_c, T) = 10269.86\$$. Fig.2 shows the Q-learning trend of the total reward against the simulation time

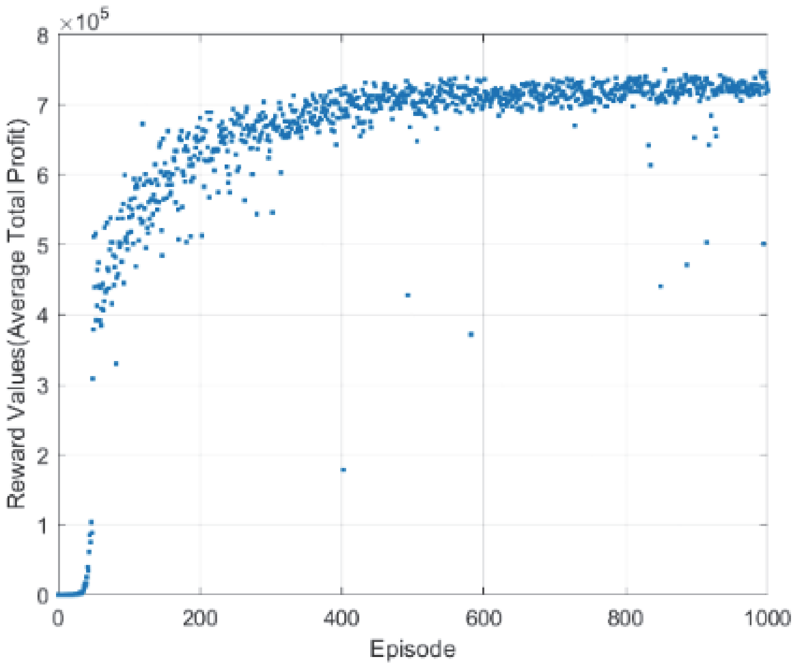

Fig.2 Training of Q-learning algorithm

V. CONCLUSION

We examined a sustainable inventory management problem where demand is dependent on the selling price and green-investment, considering a time-dependent deterioration rate for perishable products. Our proposed model focused on two decision variables such as green investment cost and total cycle time. By applying the Q-learning algorithm, a reinforcement learning technique, we successfully identified near-optimal solutions that maximize the average total profit. Through a numerical experiment, we demonstrated the effectiveness of our approach in determining the optimal green investment and cycle time, leading to sustainable and profitable inventory management.

For future scope, the model can be extended to consider multi-item inventory scenarios which are managed by vendors and additionally, exploring other reinforcement learning techniques like deepreinforcement learning and comparing their effectiveness with Q- learning, which may offer deeper insights into optimizing sustainable inventory systems.

REFERENCES

[1] Chen, Xi, Saif Benjaafar, and Adel Elomri. "The carbon-constrained EOQ." *Operations Research Letters* 41, no. 2 (2013): 172-179. https://doi.org/10.1016/j.orl.2012.12.003

[2] Benjaafar, Saif, Yanzhi Li, and Mark Daskin. "Carbon footprint and the management of supply chains: Insights from simple models." *IEEE transactions on automation science and engineering* 10, no. 1 (2012): 99-116. 10.1109/TASE.2012.2203304

[3] Bouchery, Yann, Asma Ghaffari, Zied Jemai, and Yves Dallery. "Including sustainability criteria into inventory models." *European Journal of Operational Research* 222, no. 2 (2012): 229-240. https://doi.org/10.1016/j.ejor.2012.05.004

[4] Cholette, Susan, and Kumar Venkat. "The energy and carbon intensity of wine distribution: A study of logistical options for delivering wine to consumers." *Journal of Cleaner Production* 17, no. 16 (2009): 1401-1413. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989. https://doi.org/10.1016/j.jclepro.2009.05.011

[5] Seuring, Stefan. "A review of modeling approaches for sustainable supply chain management." *Decision support systems* 54, no. 4 (2013): 1513-1520. https://doi.org/10.1016/j.dss.2012.05.053

[6] San-José, Luis A., Joaquín Sicilia, Leopoldo Eduardo Cárdenas-Barrón, and Manuel González-de-la-Rosa. "A sustainable inventory model for deteriorating items with power demand and full backlogging under a carbon emission tax." *International Journal of Production Economics* 268 (2024): 109098. https://doi.org/10.1016/j.ijpe.2023.109098

[7] Palanivel, Muthusamy, Murugesan Venkadesh, and Selvaraj Vetriselvi. "A Comprehensive Inventory Management Model with Weibull Distribution Deterioration, Ramp-Type Demand, Carbon Emission Reduction, and Shortages." *Supply Chain Analytics* (2024): 100069. https://doi.org/10.1016/j.sca.2024.100069

[8] Covert, Richard P., and George C. Philip. "An EOQ model for items with Weibull distribution deterioration." *AIIE transactions* 5, no. 4 (1973): 323-326. https://doi.org/10.1080/05695557308974918

[9] Rana, Rupal, and Fernando S. Oliveira. "Dynamic pricing policies for interdependent perishable products or services using reinforcement learning." *Expert systems with applications* 42, no. 1 (2015): 426-436. https://doi.org/10.1016/j.eswa.2014.07.007

[10] Kara, Ahmet, and Ibrahim Dogan. "Reinforcement learning approaches for specifying ordering policies of perishable inventory systems." *Expert Systems with Applications* 91 (2018): 150-158. https://doi.org/10.1016/j.eswa.2017.08.046

[11] Das, Tapas K., Abhijit Gosavi, Sridhar Mahadevan, and Nicholas Marchalleck. "Solving semi-Markov decision problems using average reward reinforcement learning." *Management Science* 45, no. 4 (1999): 560-574. https://doi.org/10.1287/mnsc.45.4.560

[12] KRÖSE, BJ A. "Learning from delayed rewards." *Robotics and autonomous systems* 15, no. 4 (1995): 233-235. https://doi.org/10.1016/0921-8890(95)00026-C . https://doi.org/10.1287/mnsc.45.4.560

[13] Mashud, A. H. M., Roy, D., Daryanto, Y., Chakrabortty, R. K., & Tseng, M. L. (2021). A sustainable inventory model with controllable carbon emissions, deterioration and advance payments. *Journal of Cleaner Production*, *296*, 126608. https://doi.org/10.1016/j.jclepro.2021.126608