sustainability

MDPI

Article

Distributed Dynamic Pricing Strategy Based on Deep Reinforcement Learning Approach in a Presale Mechanism

Yilin Liang [1,2], Yuping Hu [1,2,*], Dongjun Luo [1,2], Qi Zhu [1,2], Qingxuan Chen [1,2] and Chunmei Wang [3]

[1] School of Informatics, Guangdong University of Finance & Economics, Guangzhou 510320, China
[2] Guangdong Intelligent Business Engineering Technology Research Center, Guangzhou 510330, China
[3] College of Internet Finance & Information Engineering, Guangdong University of Finance, Guangzhou 510521, China
* Correspondence: yphuyl@126.com

Abstract: Despite the emergence of a presale mechanism that reduces manufacturing and ordering risks for retailers, optimizing the real-time pricing strategy in this mechanism and unknown demand environment remains an unsolved issue. Consequently, we propose an automatic real-time pricing system for e-retailers under the inventory backlog impact in the presale mode, using deep reinforcement learning technology based on the Dueling DQN algorithm. This system models the multicycle pricing problem with a finite sales horizon as a Markov decision process (MDP) to cope with the uncertain environment. We train and evaluate the proposed environment and agent in a simulation environment and compare it with two tabular reinforcement learning algorithms (Q-learning and SARSA). The computational results demonstrate that our proposed real-time pricing learning framework for joint inventory impact can effectively maximize retailers' profits and has universal applicability to a wide range of presale models. Furthermore, according to a series of experiments, we find that retailers should not neglect the impact of the presale or previous prices on consumers' purchase behavior. If consumers pay more attention to past prices, the retailer must decrease the current price. When the cost of inventory backlog increases, they need to offer deeper discounts in the early selling period. Additionally, introducing blockchain technology can improve the transparency of commodity traceability information, thus increasing consumer demand for purchase.

Keywords: presale; dynamic pricing; deep reinforcement learning; revenue management; blockchain

Citation: Liang, Y.; Hu, Y.; Luo, D.; Zhu, Q.; Chen, Q.; Wang, C. Distributed Dynamic Pricing Strategy Based on Deep Reinforcement Learning Approach in a Presale Mechanism. *Sustainability* **2023**, *15*, 10480. https://doi.org/10.3390/su151310480

Academic Editor: Jeffrey Wurgler

Received: 4 May 2023
Revised: 12 June 2023
Accepted: 16 June 2023
Published: 3 July 2023

1. Introduction

With the development and progress of the Internet, e-commerce has changed people's traditional consumption patterns. A growing number of subscribers engage in online consumption, expanding the scale of online sales. To promote the purchase of more potential consumers, several firms have started to market their products using diversified methods. Among them, presale has emerged as the preferred sales strategy for the majority of retailers because it may assist both parties in exchanging information beforehand, reducing manufacturing risks and maximizing profits [1]. During the presale period, consumers can only decide whether to purchase based on their perception of the price if they are not informed of the formal price. The reference price and other factors influence customers' purchasing decisions once the formal sales season has begun [2]. Different pricing tactics impact the purchasing behavior of customers and the production, inventory, and profit of retailers. Therefore, optimizing the pricing strategy is a significant concern for merchants in the presale mode, which also plays a core role in efficient market operation [3]. While most presale mechanism research focuses on fixed-pricing mechanisms and two-stage decision making, the actual sales environment is complex and changeable, and this method is no longer applicable, especially when selling new products with no prior data to use as a reference. There is a certain amount of risk in price setting, which needs to be adjusted

in real time in accordance with market conditions. Therefore, it is necessary to research and provide a solution for multi-cycle dynamic pricing based on the impact of presale and inventory overstocking.

To obtain a competitive edge and maximize profits, retailers should gather consumer information and constantly readjust prices throughout the commodities' sales cycle at the lowest possible cost [4]. Instead of static pricing, dynamic pricing can utilize consumers' willingness to pay to readjust prices in response to various environmental effect factors [5]. Accordingly, dynamic pricing has become the preferred option for most retailers that can routinely alter prices according to inventory and demand information, which is also an essential component of price strategy [6,7]. Dynamic pricing strategy applies to various industries, such as airline and energy [8,9], which has extended to online retail [10]. E-commerce platforms have severe centralization, information islands, and trust issues as their scale grows, which has a negative impact on the supply chain's overall efficiency [11,12]. Furthermore, it is crucial to record the corresponding operation in each link for future investigation while storing many orders, logistics, inventory, and other information in a distributed environment during the transaction. Blockchain technology is a decentralized distributed database technology with decentralization and encryption security features, which can efficaciously solve these issues [13–15]. E-retailers utilize this technology to share and maintain data at a reduced cost to capture consumer behavior data and estimate demand. Furthermore, the data in the blockchain system are transparent, improving consumer willingness to pay by enabling them to track the origin of goods and acquire accurate information about them.

Considerable research efforts are devoted to establishing mathematical models with a known environment to seek the optimal price through dynamic programming (DP) and other methods. However, because of the dynamic nature of the actual environment, establishing known demand models will lead to unilateral results, which cannot accurately reflect the actual situation. Unstandardized models can lead to inconsistent estimations of price elasticity and poor pricing decisions [16]. Further, it is easy to fall into a dimensional disaster when using DP to solve the problem, leading to the calculation's complexity and consuming a significant expenditure of time. Consequently, the key issue is how to best optimize price strategy in a presale mode with the impact of uncertain demand and inventory backlog. As the technology of machine learning advances by leaps and bounds, the model-free reinforcement learning algorithm can be commendably applied to tackle the issue. We adopt the deep reinforcement learning method to optimize the strategy by interacting with the environment to cope with the indeterminacy of demand and insufficient computing power. Deep reinforcement learning (DRL) is an amalgamation of the perceptual ability of deep learning and the decision-making ability of reinforcement learning, which avoids dimension issues and result limitations [17,18]. It is predominantly used in the intelligent power grid, optimization, and scheduling [19,20]. Krichen et al. [21] proposed a state-of-the-art formal method for examining the verification and validation of machine-learning systems. Raman et al. [22] described a novel approach towards the application of machine-learning-based classifiers and formal methods for analyzing and evaluating emergent behavior of complex system of systems. However, DRL is rarely employed in revenue management and e-commerce presale environments. We aim to establish a discrete finite horizon MDP of the sales environment and propose an algorithm based on Dueling DQN (Dueling DQN-DP) to address the problem of maximizing retailers' accumulated revenue in different periods under the presale mode. The calculation of the Q-value is decentralized to improve the algorithm's stability. The main contributions of this paper are as follows:

- A general model for automatic real-time pricing in a dynamic presale environment is proposed to achieve optimal inventory and revenue.
- A Dueling DQN-DP algorithm is proposed to solve the dynamic pricing problem in a finite presale horizon, thus efficaciously improving the long-term profits of the

retailer. Experiments show that the algorithm used in this paper can learn better pricing strategies than the tabular Q-learning and SARSA algorithms.

- Currently, few works in the literature apply the theory of deep reinforcement learning to solve the dynamic pricing problem under presale mode. Therefore, this is an innovation to optimize the multi-period dynamic pricing strategy under presale mode in this article.
- A more realistic presale simulation environment is designed to train and evaluate the performance of the DRL dynamic pricing algorithm and prove that the model can be widely applied in a market environment with uncertain demand.
- The influence of different inventory cost and price deviation coefficients on decision making and profit is explored, and suggestions for retailers' pricing or service strategies are put forward.

The rest of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 introduces the relevant parameters and establishes a mathematical model of dynamic pricing under the presale mode. Section 4 proposes a dynamic pricing algorithm based on DRL to solve the model. In Section 5, we use several numerical experiments to evaluate the performance of the DRL algorithm and compare it with other tabular reinforcement learning methods. Finally, we summarize and prospect the research of this paper in Section 6.

2. Related Literature

We divide the previous literature into three categories: pricing strategy in presale mode; blockchain in supply chain management; and reinforcement learning technologies in revenue management.

2.1. Presale Strategies and Pricing Mechanism

The continuous development of e-commerce makes competition in the market particularly fierce, leading to the majority of enterprises beginning to adopt a presale strategy to increase sales. Numerous models for pricing strategy in presale mode have been developed [23–25].

Presale price includes discount, premium, and formal sales price [26]. Alexandrov et al. [27] demonstrated that after selecting appropriate costs and valuations, the presale model could be equivalent to bundling, which had a crucial impact on the profit of enterprises. In [28], the joint optimization problem of multiple dynamic marketing decisions under presale mode was studied and modeled as a deterministic Markov decision process to prove that increasing the sales cycle of commodities through presale strategy is conducive to improving enterprise profits. In recent years, many scholars have studied different presale mechanisms and pricing strategies, including static pricing, dynamic pricing, and whether to publish the price of the formal sales period in the presale period.

Dynamic pricing is the core technology in revenue management, so we mainly study the dynamic pricing strategy under the given inventory and undisclosed formal sales price. Chu et al. [29] found that several optimal pricing strategies depend on the information acquired at the time of purchase. Compared with releasing all price information in the presale period, merchants could obtain better profits when they publish part of the information. Ref. [30] introduced the price mechanism of not disclosing the official price during the presale period to study consumer behavior and the expected profits of enterprises. Shugan et al. [31] found that the model of limited sales would prompt consumers to purchase at a higher presale price to enable businesses to obtain greater profits. Because consumers have different criteria for judging the purchase price of commodities in different sales periods, their purchase behavior is not merely related to the price and quality of commodities, but also the dynamic reference price [32,33]. Rios et al. [34] provided a solution for retailers' dynamic pricing in the actual environment by establishing an optimization mechanism for seasonal commodity sales to develop the optimal dynamic pricing and inventory strategy. David et al. [35] studied the relationship between reference price, demand, and profit based

on a dynamic pricing mechanism. The results showed that dynamic pricing could help reduce the adverse effects of the reference price. Ref. [36] considered a dynamic pricing problem with an unknown potential demand model belonging to a finite set of demand functions.

2.2. Blockchain Technology in Supply Chain Management

Blockchain technology is applied in supply chain management to provide consumers with authentic and efficacious information and enhance the transparency and traceability of the supply chain [37]. Du et al. [38] utilized blockchain technology to manage the supply chain finance platform to improve the efficiency of capital flow and information flow at a lower cost and adopted homomorphic encryption to protect user privacy. Ref. [39] achieved full chain transparency and information sharing through the multi-signature method of blockchain, thereby improving the ability of commodity supply chain governance. Furthermore, regulatory issues cannot be ignored when blockchain is applied to supply chain management [40]. Li et al. [41] introduced an efficient consensus mechanism to improve the efficiency of the consensus process and leveraged the deployed blockchain network to store records securely. Han et al. [42] proposed a blockchain-based auditable access control system to ensure private data security in IoT environments and realize effective management of these data.

Considering that the characteristics of blockchain can increase the transparency and authenticity of commodity information, the application of blockchain in supply chain management can help participants trace the source effectively and provide convenience for consumers to obtain the information of commodities, thus increasing commodity sales.

2.3. Reinforcement Learning Techniques in Revenue Management

Numerous academics started researching the topic of intelligent revenue management as a result of the advancement of artificial intelligence technology. Revenue management research aims to maximize profits, which is consistent with the goal of reinforcement learning. Reinforcement learning (RL) is a branch of machine learning that interacts with the environment and maximizes reward through continuous exploration and trial and error. RL is a branch of machine learning that interacts with the environment continuously through trial and error to learn so that agents can obtain maximum benefits. A summary of previous studies on revenue management is shown in Table 1.

Table 1. Application of reinforcement learning (RL) in revenue management.

RL Techniques	Authors	Year	Environment
Q-learning	Rana et al. [43]	2014	Monopoly
	Chinthalapati et al. [44]	2006	Competition
	Kutschinski et al. [45]	2013	Competition
SARSA	Collins et al. [46]	2013	Competition
DQN	Bondoux et al. [47]	2020	Monopoly
	Zhou et al. [48]	2022	Monopoly
	Wang et al. [49]	2021	Monopoly

Q-learning and SARSA are classical algorithms in reinforcement learning, and many scholars have applied them to revenue management research. Rana et al. [43] modeled the pricing problem of maximizing revenue under limited inventory and non-stationary demand as a Markov decision process and used Q(λ) and Q-learning algorithms to solve it. In addition, Q-learning and SARSA are also used to solve the pricing problems in the electronic retail market and the aviation industry in the non-monopoly environment, taking into account price sensitivity and inventory, as well as other factors [44–46].

However, classical tabular algorithms are only suitable for small-scale state space, and they are liable to fall into the curse of dimensionality for large-scale systems. Some

studies have combined reinforcement learning with deep learning and used deep neural networks to approximate the value function to address this challenge. Bondoux et al. [47] proposed an RL-based airline revenue management system that did not require demand prediction and proved that this method could converge to the optimal solution. Moreover, some scholars have studied the pricing problem under joint inventory to solve the optimal ordering and pricing strategy [48,49].

Above all, current research on pricing under the presale model mainly focuses on two-stage decision making and the commitment price mechanism. However, merchants must adjust prices based on shifting environmental factors during the selling process. When consumers do not have sufficient information about the commodity, their demand is affected by the presale or upfront price. Consequently, we consider the dynamic pricing problem of limited sales in a multi-stage pricing presale model. To avoid the curse of dimensionality and the limitations of known demand models, we utilize a model-free deep reinforcement learning (DRL) method to solve the problem by fitting the value function with neural networks. Compared with traditional tabular reinforcement learning algorithms, the DRL algorithm used in this article has proven to be superior. This enriches the research on the dynamic pricing decision of merchants in the presale model combined with reinforcement learning.

3. Problem Description and Modeling

This section introduces the hypothesis and model of dynamic pricing under the presale model.

3.1. Problem Hypothesis and Parameter Description

We consider a pricing problem of a single new product sold to consumers by a retailer using the strategy of presale with limited quantities under the monopoly mode of flash sales. Customers' uncertain purchase decisions are a result of their unfamiliarity with the true quality and effect of the new products. Meanwhile, retailers' lack of access to historical data and the complexity of the environment create demand risks. Therefore, the retailer uses dynamic pricing strategies to maximize profits. We categorize the product sales cycle into presale, formal sales, and final sales phases. The entire time dimension is divided into T periods of price updating, as shown in Figure 1. The retailer observes and receives environmental information at the beginning of each period and then sets the current price according to consumer arrival rate and inventory. After the arrival of demand, the current revenue and inventory are updated, and the corresponding storage cost is generated until the end of the sales period or until the inventory is exhausted. We summarize the notations of the dynamic pricing model in Table 2. Specifically, time horizon $t \in T = \{0, 1, \ldots, T_1, T_1 + 1, \ldots, T_2\}$ is the set of finite discrete times at which pricing actions are executed, where T_2 is the last selling period.

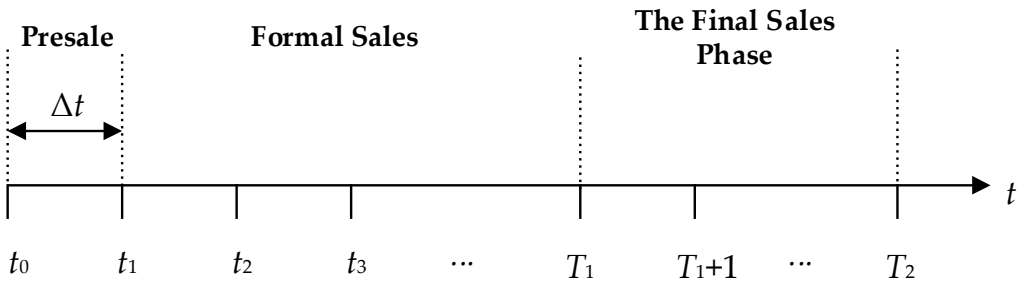

Figure 1. Sales period and pricing steps of commodities.

Table 2. Relevant notations and descriptions of dynamic pricing model.

Notation	Explanation
p_t	Sales price at period t
c	Order cost per unit
p_0	Initial price of commodities
c_i	Storage cost per unit
t	Time step for adjusting pricing
k_t	Initial inventory at period t
d_t	Sales volume at period t
δ	Parameters of reference price effects

3.2. Model Description

Because the retailer sells in limited quantities, there is a fixed subscription charge. In addition, consumers have different reference prices in each sales period. Specifically, the setting of the reference price in this article is as follows:

$$r_{t+1} = \begin{cases} p_0, if\ t \leq T_1 \\ p_t, if\ T_1 < t \leq T_2 \end{cases} \tag{1}$$

The reference price for consumers in the formal sales period and the final phase of the sales period are the presale price and the previous sale price, respectively. Consumers' reference prices are influenced by the presale period once the formal sales season begins, helping to direct their purchasing decisions. The impact of the presale period on consumers is reduced at the final sales phase, making them pay more attention to the price of the previous period. Assume that all consumers in the market are loss-neutral and the price deviation coefficient is δ. The price deviation coefficient represents the sensitivity of consumers to price deviation, which is the degree to which past prices have an impact on consumer purchasing behavior. Update the inventory at the end of each period, and the unsold commodities incur storage charges, as shown in Equation (2).

$$c_{t,inventory} = (k_t - d_t)c_i \tag{2}$$

Based on the above assumptions, the dynamic pricing model of this problem is as follows:

$$\begin{aligned} \max \quad & \sum_{t=0}^{T} (p_t d_t - c_{t,inventory}) - ck_0 \\ s.t. \quad & \sum_{t=0}^{T} d_t \leq k_0 \\ & c_{t,inventory} = (k_t - d_t)c_i \\ & p_t \in P \\ & d_t \geq 0, t = 0, 1, 2, \ldots, T \end{aligned} \tag{3}$$

3.3. Distributed Data Management

The essence of blockchain is a particular distributed database, where the data schema is globally unified and stored in each participating node as a copy. Data loss of a single node does not affect the whole system, as shown in Figure 2. The characteristics of blockchain technology also make it apply to supply chain management to improve traceability.

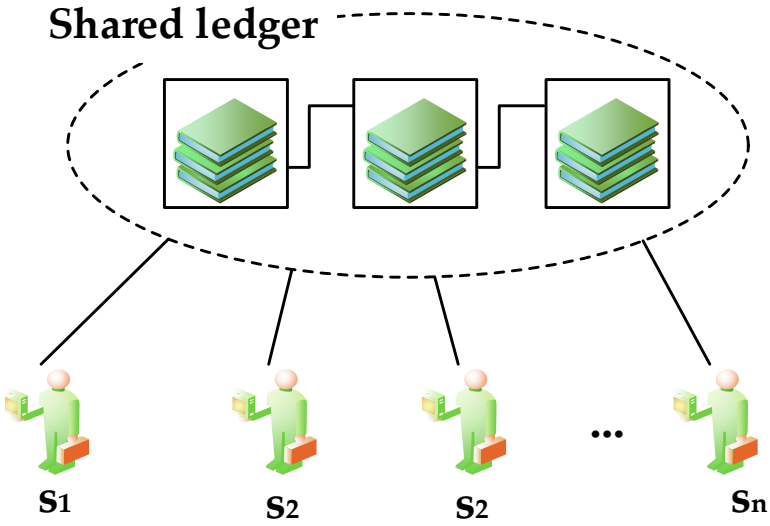

Figure 2. Distributed management of blockchain.

4. Dynamic Pricing Model Based on Deep Reinforcement Learning

4.1. Markov Decision Process

The Markov decision process (MDP) solves sequential reinforcement learning problems, providing a framework for agent and environment interaction [50]. In this study, we model the dynamic pricing problem as a discrete finite MDP and construct a quaternion $M = (S, A, P, R)$, where S represents the state space, A represents the action space, P represents the state transition function, and R represents the return function. The main components of dynamic pricing MDP are as follows:

- State space S: the collection of all states in the state space. $s_t = (\lambda_t, k_t)$, where λ_t represents the consumer's arrival state at the beginning of period t, which is the perception of the pricing environment, and the customer flow in different periods affects the customer arrival rate. k_t represents the inventory level, composed of the remaining inventory from the current time step to the end of the sales period.
- Action space A denotes the set of all executable actions in the action space. Assume the agent only takes discrete actions at fixed times to adjust the price. In this paper, we adopt the discount rate to adjust the price and define the action space as the discrete discount set that the retailer can choose, and a_t represents the discount selected at time t, $a_t \in A$. p_0 is the maximum reservation price acceptable to consumers, and $p_t = p_0 \cdot a_t$.
- State transition probability P: $p(s_t + 1|s_t, a_t)$ represents the probability of transferring from s_t to s_{t+1} after action a_t is executed in state s_t.
- Reward R: the immediate return obtained by the agent after executing action a_t in each decision step is as follows:

$$r_t = r_{income} - c_{t,inventory} = d_t p_t - (k_t - d_t)c_i \tag{4}$$

The retailer's ultimate goal is to find strategy $\pi : S \to A$, which represents the mapping from the state to the price of the selected commodity, allowing the retailer to maximize its cumulative profit over the entire sales period. The optimal strategy is as follows:

$$\pi^* = \text{argmax}_\pi E[\sum_{t=0}^{T} r_t|\pi] \tag{5}$$

Reinforcement learning means agents choose actions based on the rewards obtained by interacting with the environment to maximize the cumulative reward. The interaction process between the agent and the environment is shown in Figure 3.

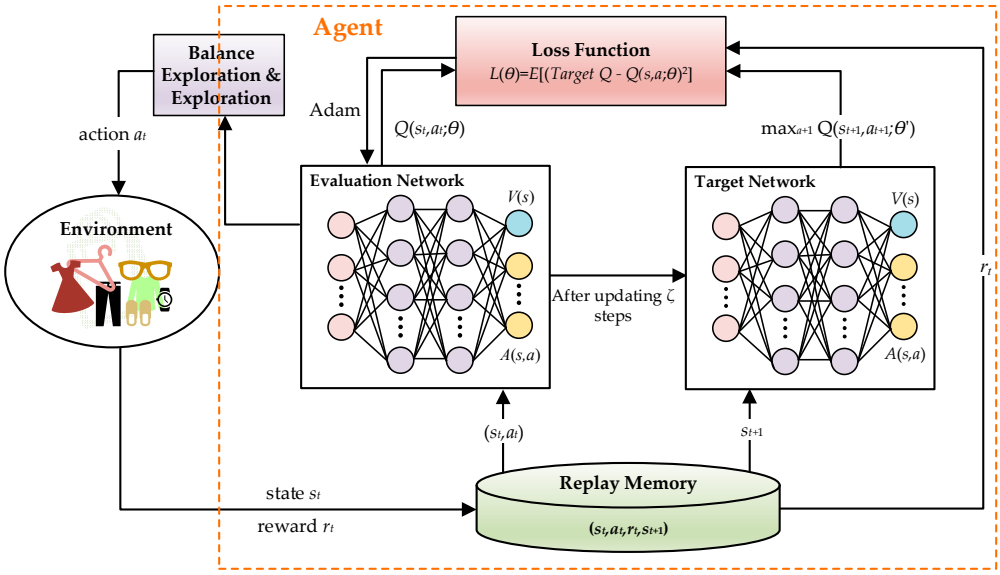

Figure 3. MDP framework for dynamic pricing.

4.2. Algorithm Description

Q-learning is a value-based reinforcement learning algorithm used to obtain the maximum reward by constructing a table to record the Q-value under different states and actions and selecting the optimal action [51]. The state action value function $Q_\pi(s,a)$ represents the expected reward of performing action a_t according to strategy π when the state is s_t.

$$Q_\pi(s,a) = E[\sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} | s_t = s, a_t = a, \pi] \quad (6)$$

Based on Equation (6), the state action value function follows the Bellman optimal equation under the optimal strategy. The optimal value function is as follows:

$$Q^*(s,a) = \sum_{s' \sim S} P(s'|s,a)[r + \gamma \max_{a'} Q^*(s',a')] \quad (7)$$

Q-learning algorithm converges $Q_\pi(s,a)$ through continuous iteration, and its state–action pair updating method is shown in Formula (8), where a represents the learning rate, γ represents the decay coefficient of the reward, and $a \in [0,1]$:

$$Q(s,a) \leftarrow Q(s,a) + a[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (8)$$

However, the large dimension of the state or action in the actual situation may fall into dimensional catastrophes that make computations difficult and time consuming. The Q-learning algorithm records the Q-value of all actions in each state. It takes a long time to estimate the Q-table and it is challenging to reach convergence when the space of states and actions is huge because this results in a very big Q-value table that must be built. In order to handle the issue of dimensional disaster, Mnih et al. [52] proposed the calculation of $Q(s,a)$, implemented by the deep neural network. The DQN algorithm is a combination of reinforcement learning and deep learning. Based on the Q-learning algorithm, the neural network is added to approximate the Q-value. Furthermore, the DQN algorithm adopts the experience playback method to store the data experienced by the agent in the replay memory and extracts a small batch of data from it during each update to break the correlation between the data. Wang et al. [53] proposed the Dueling DQN algorithm with a dual-network structure to improve the accuracy of value function estimation and divided Q-network into two parts: value function $V(s,a,m)$ and advantage function $A(s,a,\omega,l)$, as shown in Equation (9). We centralize the calculation of the Q-value, ensure that the relative

order of all dominant functions remains unchanged in each state, and remove redundant degrees of freedom to improve the algorithm's stability, as shown in Equation (10).

$$Q(s,a,\omega,m,l) = V(s,\omega,m) + A(s,a,\omega,l) \tag{9}$$

$$Q(s,a,\omega,m,l) = V(s,\omega,l) + [A(s,a,\omega,l) - \frac{1}{|A|}\sum_{a\in A} A(s,a,\omega,l)] \tag{10}$$

The Dueling DQN algorithm in this paper builds an evaluation and target network with the same structure. The calculation formula of the target network is shown in Equation (11). The target network is fixed in the updating process to make the algorithm training stable. After a certain number of times, the weight of the evaluation network is copied to the target network, thus reducing the correlation between the predicted Q-value and the target Q-value and the possibility of divergence of the loss value during training and improving the stability of learning. The loss function of DQN adopts the mean squared error (MSE) to minimize the loss between the target Q-value and the predicted Q-value, as shown in Equation (12). The MSE is a common regression estimation error-measurement method in machine learning, used to estimate the degree of inconsistency between the predicted value and the true value of the model. In the DQN algorithm, the loss function is to minimize the error between the predicted Q-value and the target Q-value, and then train the neural network based on the loss function through backpropagation.

$$TargetQ = r + \gamma max_{a'} Q(s', a'; \theta) \tag{11}$$

$$L(\theta) = E[(TargetQ - Q(s, a; \theta))^2] \tag{12}$$

In this algorithm, action is selected through strategy $\varepsilon - greedy$ to balance exploration and utilization so that it can generate higher returns. The value of exploration rate ε is attenuated exponentially, and ε_{decay} is used to control the attenuation rate, as shown in Equation (13) and Figure 4. A pseudo-code of the Dueling DQN algorithm is shown in Algorithm 1.

$$\varepsilon = \varepsilon_{end} + (\varepsilon_{init} - \varepsilon_{end}) exp(-\frac{step}{\varepsilon_{decay}}) \tag{13}$$

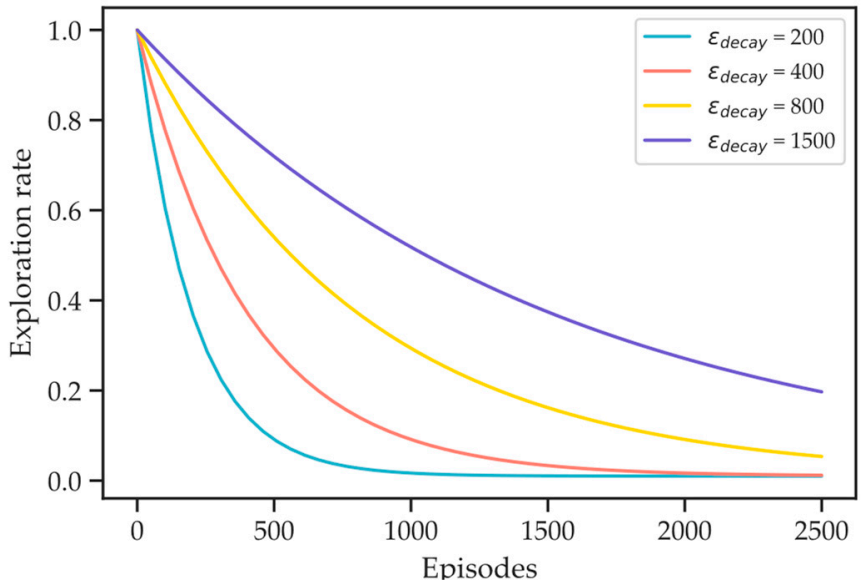

Figure 4. Exploration rate ε with different ε_{decay}.

Algorithm 1: Pseudo-code of the Dueling DQN-DP algorithm for dynamic pricing problem in presale mode

Input: Parameters of the env: c, c_i, p_0, k_0, δ, T;
parameters of the Dueling DQN-DP algorithm: α, B, γ, ζ, ε_{init}, ε_{end}, ε_{decay}
Output: Optimal pricing strategy π^*
Initialize experience replay memory D to capacity N
Initialize the Q-network weights θ
Initialize the target network weights θ^-
For *episode* = 1, M do
Reset the environment and initialize state $s_1 = (\lambda_1, k_1)$
For t = 1, T do
With probability ε select a random action a_t
otherwise select $a_t = \text{argmax}_a(s_t, a_t; \theta)$
Execute action a_t and observe reward r_t and s_{t+1}
Store transition (s_t, a_t, r_t, s_{t+1}) in D
Set $s_{t+1} = s_t$
Sample random minibatch B of transitions (s_t, a_t, r_t, s_{t+1}) from D
Set $y_i = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma max_{a'} \hat{Q}(s_{t+1}, a'; \theta^-) & \text{otherwise} \end{cases}$
Use gradient descent to train the loss function with respect to the network parameters θ
$L(\theta) = E\left[\left(TargetQ - Q\left(s, a; \theta\right)\right)^2\right]$
Every ζ steps update the target network parameters $\theta^- \leftarrow \theta$
End For
End For

5. Numerical Experiments

In this section, we evaluate the applicability of the Dueling DQN-DP algorithm in dynamic pricing under presale mode through numerical experiments and make a comparative analysis with other reinforcement learning algorithms.

5.1. Experimental Environment Settings

Based on the above-mentioned models and algorithms, our DNN is designed as a fully connected neural network with two hidden layers, each of which has 128 neurons and uses the ReLU activation function. The hyperparameter settings used in this experiment are shown in Table 3.

Table 3. The specific hyperparameter values for our numerical experiment.

Hyperparameter	Value	Explanation
N	10,000	Experience replay
B	32	Batch size
a	0.001	Learning rate
γ	0.95	Discount factor
ζ	100	Target network update frequency
ε_{init}	1.0	Initial exploration rate
ε_{end}	0.01	Final exploration rate
ε_{decay}	100	Decay factor

We adopt the Adam optimizer [54] to optimize the DNN, which can automatically adjust the learning rate so that the training can converge rapidly and learn accurately. Experiments were carried out in Python 3.9, and the results were obtained from a computer with an i5-7200U CPU 2.50 GHz and 8.00 GB of RAM. Experimental data were generated according to the following model:

Consumer arrival model. In this experiment, the arrival model of consumers refers to the literature [43]. $\lambda(t)$ refers to consumers' average arrival rate, a Poisson distribution

with discrete time, as shown in Equation (14). The initial arrival rate of consumers is randomly selected from the uniform distribution $[h_1, h_2]$.

$$\lambda(t) = \lambda(0) - \sigma t, t = 0, 1, \ldots, T \tag{14}$$

Demand function. Our depiction of a realistic sales environment takes into account the temporal fluctuations in demand, with consumers exhibiting varying purchasing needs across different time periods. Therefore, we divide the sales cycle into three parts: (1) Presale. The current price is inversely related to consumer demand; (2) Formal sales. The presale price has an impact on consumer demand, and the gap in price between the current and presale periods is inversely proportional; (3) The final sales phase. Due to the considerable duration between the final phase of sales and the presale period, the consumer's reference price is affected by the nearest price [32]. The description of the specific demand function is as follows:

$$D(t) = \begin{cases} \lambda(t) - b_1 p_t & t = 0 \\ \lambda(t) - b_2 p_t - \delta t(p_t - p_0) & t = 1, 2, \ldots, T_1 \\ \lambda(t) - b_2 p_t - \delta t(p_t - p_{t-1}) & t = T_1 + 1, T_1 + 2, \ldots, T_2 \end{cases} \tag{15}$$

The maximum number of pricing times in the sales cycle is T = 8, and the price discount set A = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. It is stipulated that the minimum price of goods should not be lower than its cost price. The primary environment parameter settings are shown in Table 4.

Table 4. The basic environmental parameter values setting.

Parameter	k_0	T_1	T_2	p_0	c	c_i	δ	$[h_1, h_2]$	σ	b_1	b_2
Value	2500	4	7	50	3	1	1.5	[400,420]	5	3	3

5.2. Algorithm Evaluation and Performance Comparison

In this subsection, we evaluate the performance of Dueling DQN, SARSA, Q-learning, and other algorithms by comparing the percentage of average revenue and revenue from the optimal strategy. The performance of these algorithms is measured by the percentage of deviations from the optimal policy. The optimal strategy of dynamic pricing is solved by dynamic programming. All rewards received by the agent are obtained by running 5000 episodes, and the average reward for each algorithm is summarized in Table 5. The main method for analyzing algorithm performance in reinforcement learning is through average return. The optimal revenue for this problem is 82,112, while the average revenue for Dueling DQN is 75,195.95. This indicates that the proposed method can achieve 91.58 percent of the optimal profit.

Table 5. Comparison of the average revenue from different algorithms.

Model	**Percentage of the Optimal Strategy**
Q-learning	77.84%
SARSA	78.41%
DQN	88.87%
Dueling DQN	91.58%

We adopt the same iteration steps and parameters to compare the performance of Dueling DQN, SARSA, Q-learning, and other algorithms. As concluded in Figure 5, compared with SARSA and Q-learning algorithms, Nature DQN and Dueling DQN algorithms improve continuously in continual environmental interaction, thus learning superior pricing strategies. However, the Dueling DQN algorithm converges faster and reaches stable

conditions with slight fluctuations after 500 training episodes, while Nature DQN needs 1200 steps. The Dueling DQN algorithm generally learns the state value function with advantage and affects the Q-value of all actions in each update. It is possible to complete more updates in fewer times, leading to a faster convergence rate and more accurate learning than the Nature DQN. Table 6 shows that DRL has better performance and stability and can improve the average rate of return by more than 10% compared with other tabular reinforcement learning algorithms under the same conditions.

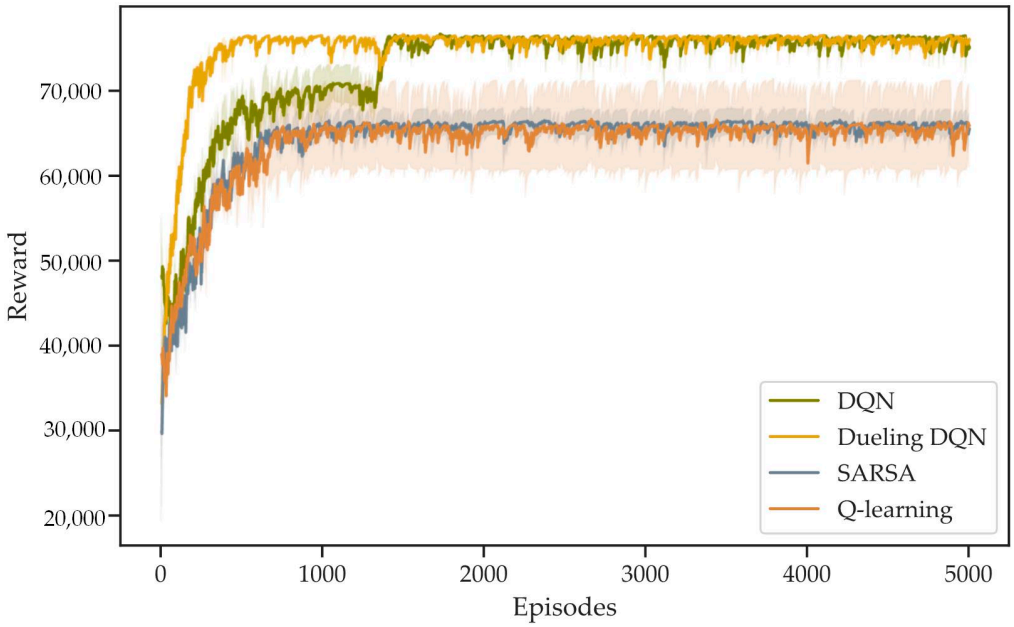

Figure 5. Convergence curve of the cumulative rewards.

Table 6. Percentage of optimal strategies compared between different algorithms.

Model	Average Revenue ± Standard Deviation	Max	Min	Ratio of Models to Dueling DQN
Q-learning	63,919.83 ± 5970.21	74,931	8115	85.01%
SARSA	64,374.69 ± 5895.57	74,153	11,284	85.61%
DQN	72,974.94 ± 7243.83	80,650	14,640	97.05%
Dueling DQN	75,195.95 ± 4936.19	81,177	15,285	100%

Figure 6 and Table 7 record the profit, sales volume, price, and remaining inventory of the four algorithms in each sales period under the optimal strategy. The two traditional RL algorithms often carry out promotions to attract consumers to reduce inventory when there is a cost of inventory backlog. Moreover, the profits of these two algorithms are lower than that of DRL because selling more commodities at a lower price reduces the current profit when backlog costs exist. Figure 7 shows the comparison between the optimal pricing, profit, and sales of dueling DQN-DP and dynamic programming. It can be further explained by the near-optimal characteristic of the method proposed in this paper. Although there are slight differences between the Dueling DQN-DP and the optimal strategy as the sales period ends, both achieve more sales volume and profits by reducing prices.

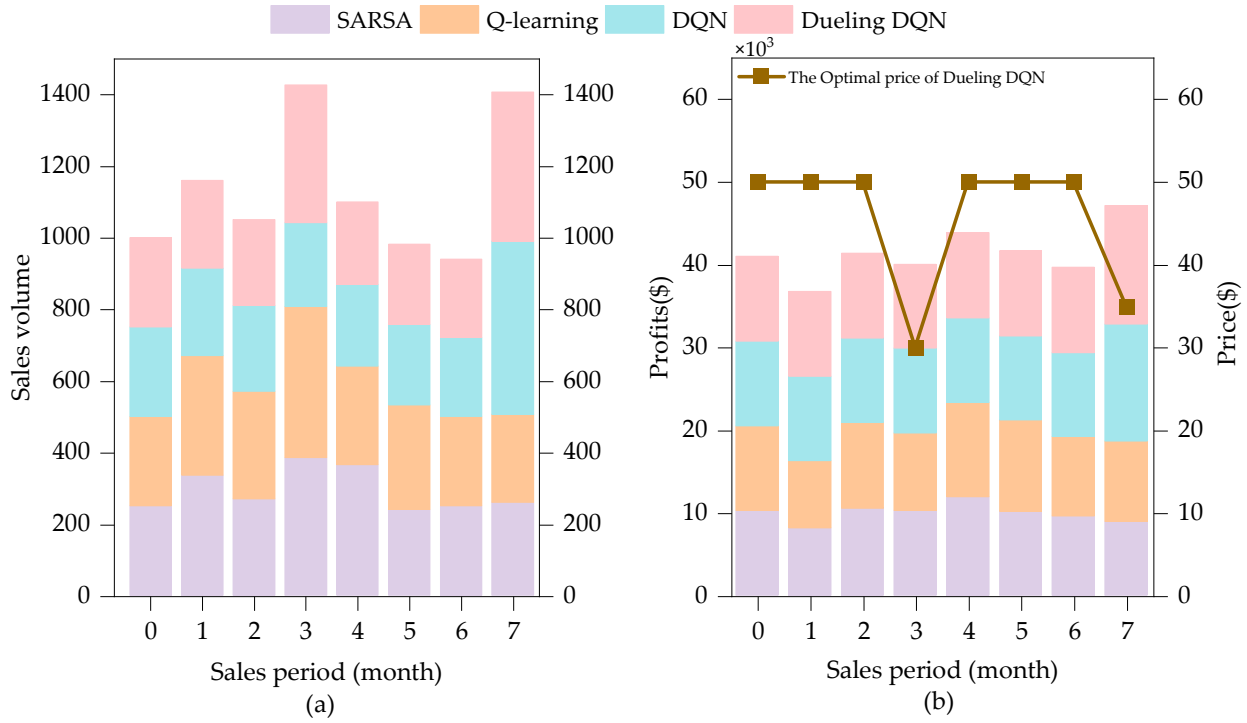

Figure 6. (**a**) Optimal sales volume in each period; (**b**) optimal price of Dueling DQN and the optimal profit of each algorithm in each period.

Table 7. Optimal price and remnant inventory of different algorithms in each period.

Period	Dueling DQN		DQN		Q-Learning		SARSA	
	Price (USD)	Inventory	Price (USD)	Inventory	Price (USD)	Inventory	Price (USD)	Inventory
0	50	2250	50	2250	50	2250	50	2250
1	50	2005	50	2005	30	1915	30	1915
2	50	1765	50	1765	40	1615	45	1645
3	30	1380	50	1530	35	1193	30	1260
4	50	1150	50	1300	45	918	35	895
5	50	925	50	1075	40	626	45	655
6	50	705	50	855	40	376	40	405
7	35	288	30	370	40	131	35	145

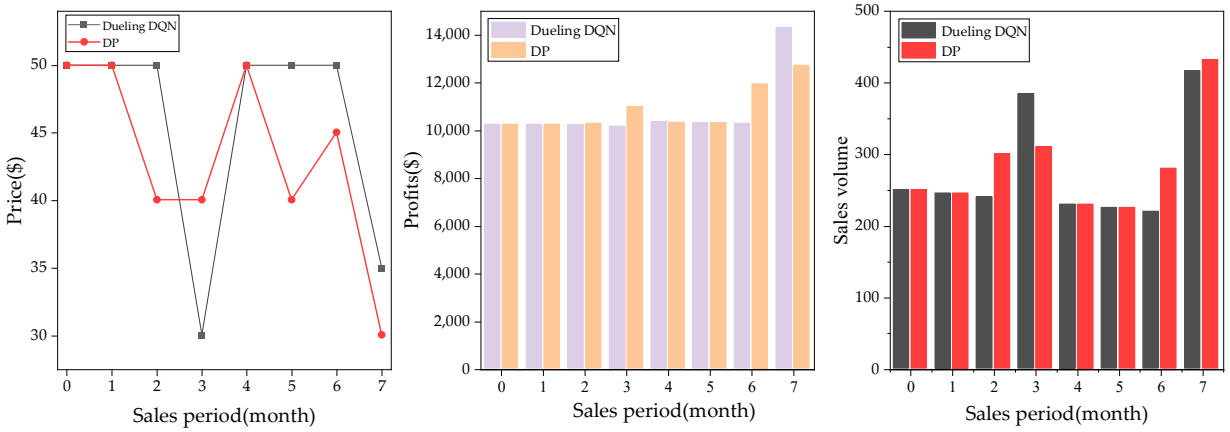

Figure 7. Optimal pricing, profit, and sales for dynamic programming and Dueling DQN.

5.3. Impact of Price Deviation Coefficient and Inventory Backlog

The inventory backlog cost affects retailers' profit in each period. Figure 7 displays the optimal cumulative profit and sales volume under the joint influence of the inventory backlog cost and the reference price coefficient. It is evident that retailers can sell more commodities and gain higher profits when the reference price has a greater influence on consumers. Therefore, retailers should not overlook the impact of reference prices when making pricing decisions. The findings demonstrate that retailers need to adopt a dynamic pricing strategy to maximize profit under the influence of reference price and reasonably reduce current pricing to increase sales when customers have a better memory of previous prices. Conversely, with the increase in inventory cost, the revenue gained by retailers decreases. The specific pricing and sales are shown in Figure 8. When the cost per unit of inventory rises, retailers sell at a discount in the lead-up to the sale, allowing them to sell more commodities to reduce the backlog of expenses.

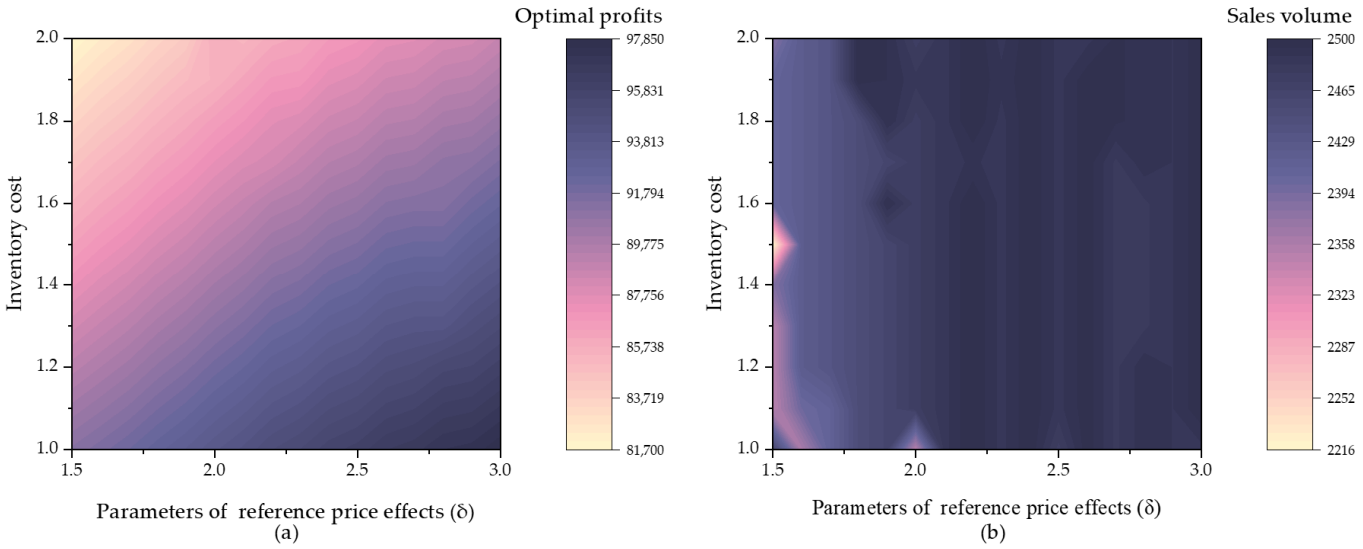

Figure 8. (**a**) Optimal profits under the influence of different inventory costs and reference prices; (**b**) sales volume of optimal strategy under the influence of different inventory costs and reference prices.

5.4. Applications in Complex Environments

In the practical sales environment, various interfering factors are constantly changing, so four different types of market environments are given in this subsection to test the application of DRL in different environments. The concrete settings are shown in Table 8, and the demand function at the final phase of sales in Market 4 is updated to the following:

$$D(t) = \lambda(t) - 15t \cdot exp(-\frac{p_t}{50_t}) \tag{16}$$

Table 8. Different market environment settings.

Market	b	f	δ
1	2	3	1.5
2	3	2	2
3	1.5	2	2.5
4	3	3	1.5

In this experiment, we carried out 5000 iterations for these market environments, and the results are shown in Figure 9. The Dueling DQN algorithm can converge in different environments and obtain better returns. It can be seen in Figure 10 that the average return can reach more than 90% of the optimal strategy, indicating that the method proposed in this paper can be applied to various environments and has universal applicability.

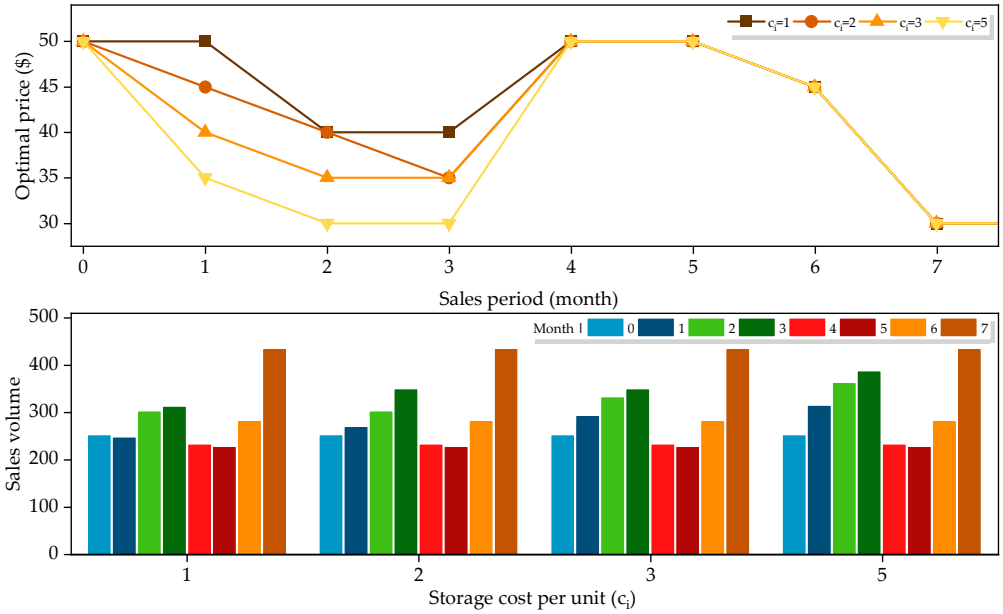

Figure 9. Optimal price and sales volume under different inventory costs.

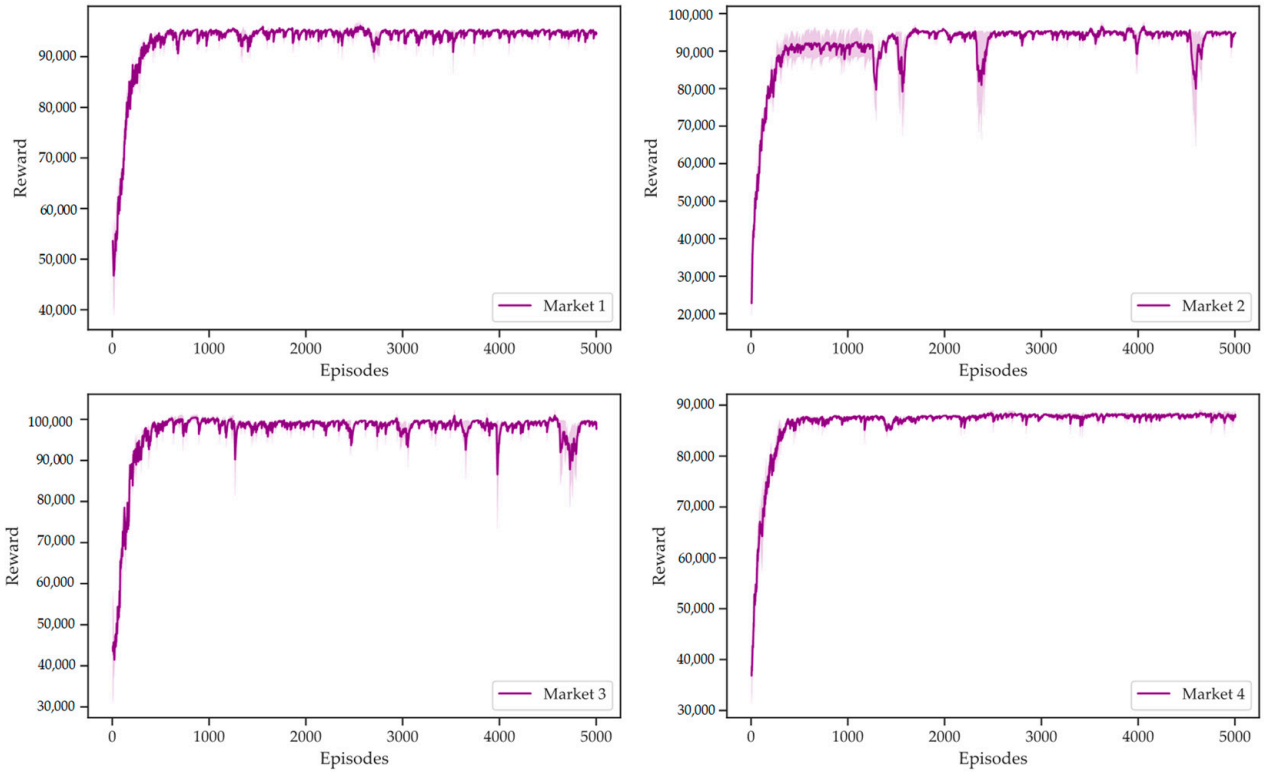

Figure 10. Convergence curve of the cumulative rewards in different markets.

5.5. Analysis of Distributed Data Management Effect Based on Blockchain

Retailers must manage large amounts of payment, inventory, and other information generated during the ordering and selling process. Therefore, we utilize blockchain technology's distributed storage, data sharing, and high-security features to manage data. Transactions, logistics, payment, inventory, and other data are stored in copies in participating nodes, which can accurately and quickly trace the source of commodities and obtain sales information. Data traceability is achieved through the design mechanism of timestamp and the chain connection between blocks, as shown in Figure 11.

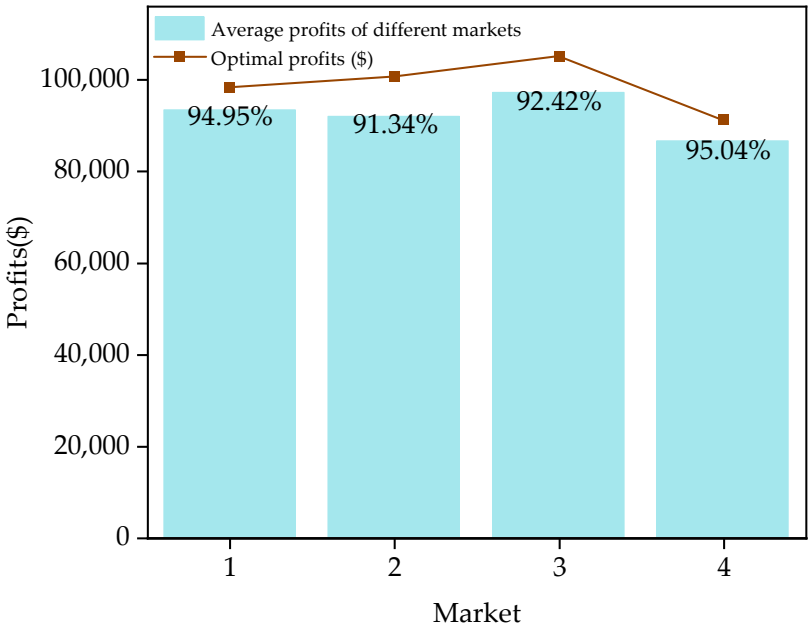

Figure 11. Optimal profits and average profits of DQN in different markets.

Blockchain is a decentralized distributed database that enables more extensive information sharing across regions and subjects. The technology allows e-retailers to share more private data, improving the efficiency of product traceability and inspection. Furthermore, it can enable consumers to access more transparent and realistic information about commodities, enhance corporate identity, and increase consumers' willingness to pay, as shown in Figure 12.

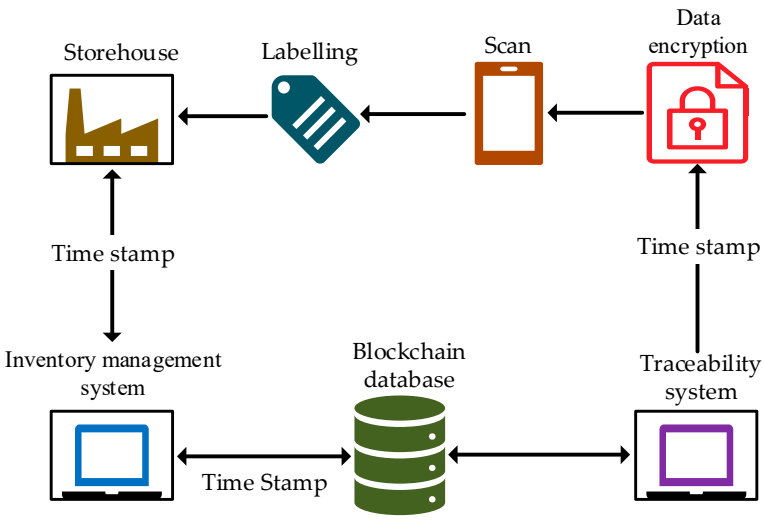

Figure 12. Block supply chain traceability.

6. Conclusions

In this paper, we investigate the dynamic pricing problem of limited sales in the presale environment with uncertain demand, establish the objective function of retailers, and transform the problem into discrete finite MDP. To optimize the strategy and maximize profit, we utilize deep reinforcement learning theory. We establish a learning framework based on the Dueling DQN-DP algorithm to solve dynamic pricing problems, reducing assumptions about the demand function and making the model more universal, which can address the uncertainty brought on by retailers adopting presale mode to sell new products. Utilizing RL to solve the optimal pricing problem can reduce the cost of dynamic pricing, avoid dimension disasters, and improve computing efficiency. Specifically, we evaluate the performance of our proposed algorithm by comparing it with the nature DQN algorithm and common tabular reinforcement learning algorithms (Q-learning and SARSA). Experimental results demonstrate that Dueling DQN can effectively figure out the dynamic pricing problem in the presale environment by learning preferred pricing strategies with a faster convergence rate. The Q-learning and SARSA algorithms can only achieve around 85.01% and 85.61% performance compared that of our Dueling DQN-DP algorithm, respectively. This paper enriches the research on dynamic pricing strategies considering the impact of inventory backlog and provides corresponding references for studying the impact of other marketing methods on dynamic pricing strategies. The method proposed in this paper does not require model configuration and can handle high-dimensional space problems. It has positive significance in exploring dynamic pricing strategies in uncertain environments. It contributes to the understanding of the impact of inventory overstocking on dynamic pricing strategy research and serves as a resource for the study of dynamic pricing strategies in other marketing technologies or fields. Furthermore, we consider the impact of reference price and inventory overstock costs on retailers' pricing. When consumers prioritize past prices, masterminds ought to formulate discount strategies to sell more commodities and thus acquire a higher profit. Additionally, it is shown that masterminds can use diverse discount strategies at the early stage of sales to stimulate consumers to buy and reduce inventory expenses when the inventory cost increases.

We improve the sophistication of the simulation environment model and introduce the time variance of consumer demand. According to the price reference effect and the amount of time left in the sales period, consumer demand fluctuates continually throughout time, which is more in line with reality. The experiment also includes a variety of market settings to corroborate the developed model's universal applicability. The results illustrate that the Dueling DQN algorithm remains valid and capable of learning excellent strategies. It has generalizability and can be applied to more realistic environments.

The following points are relevant for the expansion of this research direction in the future:

- In this paper, we focus on monopolistic pricing. However, there is more than one retailer in the physical sales environment. Businesses need to consider competitors' prices when setting their own prices because they influence customers' decisions to purchase the same goods. In future research, multiple agents can be considered to make pricing decisions in a competitive environment for the expansion of this research direction in the future.
- The blockchain system is transparent in data access, which leads to a disclosure risk of information. Consequently, the blockchain system needs to implement more efficient smart contracts.
- Assumptions about the strategic behavior of consumers should be made. Consumers make cross-period purchases based on price, quality, and other factors to maximize their expected utility.
- Risk-sensitive sellers and consumers could be considered.
- The return rate should be taken into account by merchants in the revenue management of e-commerce transactions.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L. and Y.H.; software, Y.L.; validation, Y.L. and D.L.; formal analysis, Y.L. and Y.H.; investigation, Y.L.; resources, Y.L. and Y.H.; data curation, Y.L. and Q.C.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and Y.H.; visualization, Y.L., C.W. and Q.Z.; supervision, Y.L., Y.H. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zer, O.; Boyaci, T. Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Soc. Sci. Electron. Publ.* **2009**, *58*, 1328–1349.
2. Li, J.; Guo, C.; Wang, P. Joint Pricing and Inventory Decision Considering the Reference Price Effect Based on Advance Selling. In Proceedings of the Twelfth International Conference on Management Science and Engineering Management, Melbourne, Australia, 1–4 August 2018; pp. 195–210. [CrossRef]
3. Nunan, D.; Domenico, M. Value creation in an algorithmic world: Towards an ethics of dynamic pricing. *J. Bus. Res.* **2022**, *150*, 451–460. [CrossRef]
4. Leloup, B.; Deveaux, L. Dynamic Pricing on the Internet: Theory and Simulations. *Electron. Commer. Res.* **2001**, *1*, 265–276. [CrossRef]
5. Keller, A.; Vogelsang, M.; Totzek, D. How displaying price discounts can mitigate negative customer reactions to dynamic pricing. *J. Bus. Res.* **2022**, *148*, 277–291. [CrossRef]
6. Lin, K.Y. Dynamic pricing with real-time demand learning. *Eur. J. Oper. Res.* **2006**, *174*, 522–538. [CrossRef]
7. Boer, A. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surv. Oper. Res. Manag. Sci.* **2015**, *20*, 1–18. [CrossRef]
8. Gao, J.; Le, M.; Fang, Y. Dynamic air ticket pricing using reinforcement learning method. *RAIRO—Oper. Res.* **2022**, *56*, 2475–2493. [CrossRef]
9. Zhong, S.; Wang, X.; Zhao, J.; Li, W.; Li, H.; Wang, Y.; Deng, S.; Zhu, J. Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating. *Appl. Energy* **2021**, *288*, 116623. [CrossRef]
10. Elmaghraby, W.; Keskinocak, P. Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices, and Future Directions. *Manag. Sci.* **2003**, *49*, 1287–1309. [CrossRef]
11. Gao, N.; Han, D.; Weng, T.-H.; Xia, B.; Li, D.; Castiglione, A.; Li, K.-C. Modeling and analysis of port supply chain system based on Fabric blockchain. *Comput. Ind. Eng.* **2022**, *172*, 108527. [CrossRef]
12. Li, D.; Han, D.; Zheng, Z.; Weng, T.-H.; Li, H.; Liu, H.; Castiglione, A.; Li, K.-C. MOOCsChain: A blockchain-based secure storage and sharing scheme for MOOCs learning. *Comput. Stand. Interfaces* **2021**, *81*, 103597. [CrossRef]
13. Li, J.; Han, D.; Wu, Z.; Wang, J.; Li, K.-C.; Castiglione, A. A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control. *Futur. Gener. Comput. Syst.* **2022**, *142*, 195–211. [CrossRef]
14. Li, D.; Han, D.; Weng, T.-H.; Zheng, Z.; Li, H.; Liu, H.; Castiglione, A.; Li, K.-C. Blockchain for federated learning toward secure distributed machine learning systems: A systemic survey. *Soft Comput.* **2022**, *26*, 4423–4440. [CrossRef]
15. Han, D.; Pan, N.; Li, K.-C. A Traceable and Revocable Ciphertext-Policy Attribute-based Encryption Scheme Based on Privacy Protection. *IEEE Trans. Dependable Secur. Comput.* **2020**, *19*, 316–327. [CrossRef]
16. Mila, N.; David, S.L.; He, W. Dynamic learning and pricing with model misspecification and endogeneity effect. *Manag. Sci. J.* **2019**, *65*, 4980–5000. [CrossRef]
17. Mousavi, S.S.; Schukat, M.; Howley, E. Deep Reinforcement Learning: An Overview. In *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*; Springer: Berlin, Germany, 2016; pp. 426–440. [CrossRef]
18. Singh, V.; Chen, S.; Singhania, M.; Nanavati, B.; Kar, A.; Gupta, A. How are reinforcement learning and deep learning algo-rithms used for big data based decision making in financial industries–A review and research agenda. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100094. [CrossRef]
19. Touzani, S.; Prakash, A.K.; Wang, Z.; Agarwal, S.; Pritoni, M.; Kiran, M.; Brown, R.; Granderson, J. Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency. *Appl. Energy* **2021**, *304*, 117733. [CrossRef]
20. Li, J.; Geng, J.; Yu, T. Multi-objective optimal control for proton exchange membrane fuel cell via large-scale deep rein-forcement learning. *Energy Rep.* **2021**, *7*, 6422–6437. [CrossRef]

21. Krichen, M.; Mihoub, A.; Alzahrani, M.Y.; Adoni, W.Y.H.; Nahhal, T. Are Formal Methods Applicable to Machine Learning and Artificial Intelligence? In Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 9–11 May 2022; pp. 48–53. [CrossRef]
22. Raman, R.; Gupta, N.; Jeppu, Y. Framework for formal verification of machine learning based complex system-of-system. *INCOSE Int. Symp.* **2021**, *26*, 91–102. [CrossRef]
23. He, B.; Pan, W.; Yang, Y. Joint pricing and overbooking policy in a full payment presale mechanism of new products. *Int. Trans. Oper. Res.* **2017**, *26*, 1810–1827. [CrossRef]
24. Wang, X.; Tian, J.; Fan, Z.-P. Optimal presale strategy considering consumers' preference reversal or inconsistency. *Comput. Ind. Eng.* **2020**, *146*, 106581. [CrossRef]
25. Gupta, V.; Chutani, A. Supply chain financing with advance selling under disruption. *Int. Trans. Oper. Res.* **2019**, *27*, 2449–2468. [CrossRef]
26. Zeng, C. Optimal Advance Selling Strategy under Price Commitment. *Pac. Econ. Rev.* **2013**, *18*, 233–258. [CrossRef]
27. Alexandrov, A.; Bedre-Defolie, Ö. The Equivalence of Bundling and Advance Sales. *Mark. Sci.* **2014**, *33*, 259–272. [CrossRef]
28. Cheng, Y.; Li, H.; Thorstenson, A. Advance selling with double marketing efforts in a newsvendor framework. *Comput. Ind. Eng.* **2018**, *118*, 352–365. [CrossRef]
29. Chu, L.Y.; Zhang, H. Optimal Preorder Strategy with Endogenous Information Control. *Manag. Sci.* **2021**, *57*, 1055–1077. [CrossRef]
30. Mei, W.; Du, L.; Niu, B.; Wang, J.; Feng, J. The effects of an undisclosed regular price and a positive leadtime in a presale mechanism. *Eur. J. Oper. Res.* **2016**, *250*, 1013–1025. [CrossRef]
31. Shugan, S.M.; Xie, J. Advance Selling for Services. *Calif. Manag. Rev.* **2004**, *46*, 37–54. [CrossRef]
32. Mazumdar, T.; Raj, S.; Sinha, I.; Arunraj, N.S.; Ahrens, D.; Koschate-Fischer, N.; Wüllner, K.; Chen, X.; Hu, P.; Hu, Z.; et al. Reference Price Research: Review and Propositions. *J. Mark.* **2005**, *69*, 84–102. [CrossRef]
33. Anton, R.; Régis, Y.; Chenavaz; Paraschiv, C. Dynamic pricing, reference price, and price-quality relationship. *J. Eco-Nomic Dyn. Control* **2022**, *146*, 104586. [CrossRef]
34. Rios, J.; Verab, J. Dynamic pricing and inventory control for multiple products in a retail chain. *Comput. Ind. En-Gineering* **2023**, *177*, 109065. [CrossRef]
35. David, P.; Martin, S. Dynamic pricing and reference price effects. *J. Bus. Res.* **2022**, *152*, 300–314. [CrossRef]
36. Cheung, W.C.; Simchi-Levi, D.; Wang, H. Technical Note—Dynamic Pricing and Demand Learning with Limited Price Experimentation. *Oper. Res.* **2017**, *65*, 1722–1731. [CrossRef]
37. Balzarova, M.A.; Cohen, D.A. The blockchain technology conundrum: Quis custodiet ipsos custodes? *Curr. Opin. Environ. Sustain.* **2020**, *45*, 42–48. [CrossRef]
38. Du, M.; Chen, Q.; Xiao, J.; Yang, H.; Ma, X. Supply Chain Finance Innovation Using Blockchain. *IEEE Trans. Eng. Manag.* **2020**, *67*, 1045–1058. [CrossRef]
39. Cao, S.; Foth, M.; Powell, W.; Miller, T.; Li, M. A blockchain-based multisignature approach for supply chain governance: A use case from the Australian beef industry. *Blockchain Res. Appl.* **2022**, *3*, 11. [CrossRef]
40. Bai, Y.; Liu, Y.; Yeo, W.M. Supply chain finance: What are the challenges in the adoption of blockchain technology? *J. Digit. Econ.* **2022**, *1*, 153–165. [CrossRef]
41. Li, H.; Han, D.; Tang, M. A Privacy-Preserving Storage Scheme for Logistics Data with Assistance of Blockchain. *IEEE Internet Things J.* **2021**, *9*, 4704–4720. [CrossRef]
42. Han, D.; Zhu, Y.; Li, D.; Liang, W.; Souri, A.; Li, K.-C. A Blockchain-Based Auditable Access Control System for Private Data in Service-Centric IoT Environments. *IEEE Trans. Ind. Inform.* **2022**, *18*, 3530–3540. [CrossRef]
43. Rana, R.; Oliveira, F.S. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. *Omega* **2014**, *47*, 116–126. [CrossRef]
44. Chinthalapati, V.; Yadati, N.; Karumanchi, R. Learning dynamic prices in MultiSeller electronic retail markets with price sensitive customers, stochastic demands, and inventory replenishments. *IEEE Trans. Syst. Man Cybern. Part C* **2006**, *36*, 92–106. [CrossRef]
45. Kutschinski, E.; Uthmann, T.; Polani, D. Learning competitive pricing strategies by multi-agent reinforcement learning. *J. Econ. Dyn. Control* **2013**, *27*, 2207–2218. [CrossRef]
46. Collins, A.; Thomas, L. Learning competitive dynamic airline pricing under different customer models. *J. Revenue Pricing Manag.* **2013**, *12*, 416–430. [CrossRef]
47. Bondoux, N.; Nguyen, A.Q.; Fiig, T.; Acuna-Agost, R. Reinforcement learning applied to airline revenue management. *J. Revenue Pricing Manag.* **2020**, *19*, 332–348. [CrossRef]
48. Zhou, Q.; Yang, Y.; Fu, S. Deep reinforcement learning approach for solving joint pricing and inventory problem with reference price effects. *Expert Syst. Appl.* **2022**, *195*, 116564. [CrossRef]
49. Wang, R.; Gan, X.; Li, Q.; Yan, X. Solving a Joint Pricing and Inventory Control Problem for Perishables via Deep Reinforcement Learning. *Complexity* **2021**, *2021*, 6643131. Available online: https://ideas.repec.org/a/hin/complx/6643131.html (accessed on 6 August 2022). [CrossRef]
50. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
51. Sutton, R.S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **1988**, *3*, 9–44. [CrossRef]

52. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. Available online: https://courses.cs.washington.edu/courses/cse571/16au/slides/dqn_nature.pdf (accessed on 10 October 2022). [CrossRef]
53. Wang, Z.; Freitas, N.D.; Lanctot, M. Dueling network architectures for deep reinforcement learning. In Proceedings of the 33rd International Conference on Machine Learning, New York City, NY, USA, 19–24 June 2016; pp. 1995–2003.
54. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. *arXiv* **2014**, arXiv:1412.6980.