Research on transfer discount for public transportation based on deep reinforcement learning and game theory

Qinan Wang*[a], Suyu Zhu[b]
[a]College of Transportation Engineering, Chang'an University, Xi'an 710064, China
[b]College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China
*Corresponding author: wangqinan@chd.edu.cn

ABSTRACT

Reasonable transit transfer discount policies incentivize people to take public transportation and alleviate the increase in vehicle usage. This research establishes and analyzes a dynamic game model with perfect information for public transportation operator and travelers as players. By treating the optimization process of discount as a learning process for operator to continuously explore and obtain the optimal price in a complex traveler environment, an adaptive discount adjustment method based on deep reinforcement learning is proposed. The result shows that transfer discount can increase the travel share of public transportation; Operator can use the first mover advantage to and achieve the most favorable Nash equilibrium for himself; The applicability of the proposed method in transfer discount scenarios is tested via simulation; According to the reinforcement learning DDPG algorithm, the Nash equilibrium is reached when the fare change interval is between (-0.68, -0.52) yuan, resulting in an increase in the usage of public transportation.

Keywords-Public transportation; Deep reinforcement learning; Game theory; Transfer discount

1. Introduction

China's urbanization rate has experienced rapid development. The huge demand for vehicle travel, while bringing mobility and convenience, has led to increasing traffic congestion, air pollution, road accidents and disconnection between residents and communities. Attracting residents who choose vehicle travel to take more high-capacity public transport is becoming increasingly crucial. Urban rail transit has developed rapidly in China due to its large volume, long distance and fast speed. By 2022, 9584 kilometers of urban rail transit have been put into operation. However, due to its lack of reachability, it is necessary to properly use the bus routes with high reachability and develop reasonable transfer discount policies while improving the connectivity of facilities, to encourage public transport travel.

Liu et al.[1] optimized fares and transfer discount with the aim of maximizing social welfare, and concluded that travel time uncertainty and station spacing differences play an important role. From the perspective of operator and travelers, Ding[2] simulated the implementation of transfer discount in Dalian, and demonstrated discount has a promoting effect on guiding transfer travel. Lee et al.[3] used a metamodel-based approach to optimize discounts, taking into account actual travel times, congestion and fare, resulting in a 12% increase in operator's profit. Liao[4] established a multi-objective model for discount policy, and concluded the impedance of the road network decreases most significantly when the discount range is 15%-35%. Liu[5] jointly optimized the fare system of rail transit and bus, obtained the best combination fare system, and predicted the distribution of travelers in different public transport lines. Jin et al.[6] analyzed the influence of different discount policies on the choice of travel modes of different groups in Chengdu. The deep reinforcement learning algorithm can provide reasonable and stable public transport price policies for complex travel demands, significantly increasing the overall revenue of public transport under different demand models[7]. Xia et al.[8] constructed a TSCA interaction mathematical model using game theory and used distributed reinforcement Q-learning to build payoff values, causing interaction between action selections between intersection Agent and management Agent. Bouton et. al.[9] proposed a combination of reinforcement learning and game theory to learn the best vehicle merging behaviors in dense traffic. Lopez et. al.[10] studied lane-changing decision-making and payoff learning for autonomous vehicles based on Nash Q-learning by formulating multiple games for pairs of agents, to improve the traffic efficiency and safety of vehicles in complex road environments.

International Conference on Smart Transportation and City Engineering (STCE 2023),
edited by Miroslava Mikusova, Proc. of SPIE Vol. 13018, 1301806
 · 0277-786X · doi: 10.1117/12.3024156

Although many scholars have conducted in-depth research on the transfer discount, there has been not much game analysis between discount and travel mode conversion. This paper takes public transport operator and vehicle travelers as game players, and uses the dynamic game with perfect information to analyze the implementation of discount. The discount formulation is regarded as a learning process to achieve price optimization through constant exploration and adjustment in the complex travelers environment, and an adaptive discount adjustment method based on deep reinforcement learning with game theory is proposed. The effectiveness and feasibility of the algorithm are verified by theoretical derivation and numerical simulation, and the Nash equilibrium strategy is obtained.

2. Construction of dynamic game model with perfect information

2.1. Construction of travel network structure

As multimodal subway-bus networks typically involve departure, destination and transfer points, a transport network consisting of three stations is constructed. Station A represents a point within the subway's radiation, station N represents a transfer point, and station B represents a point outside the subway's radiation. The distance between station A and station N includes both subway and bus, while only bus is available between station N and station B. Passengers traveling between A and B do not have direct access to public transport (PT) and must transfer at station N. In addition, travelers can opt for direct travel by driving their own vehicle. The travel network's structure is illustrated in Figure 1.

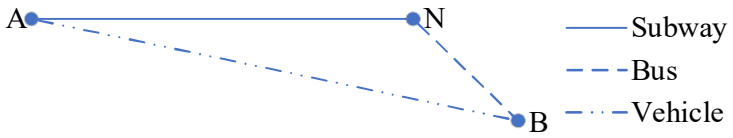

Figure 1. Travel network's structure.

2.2. Model assumptions for transfer discount dynamic game

In a dynamic game with perfect information, players make decisions sequentially rather than simultaneously. The first mover has an advantage in predicting the rational responses of the later actor and make decisions based on the responses to achieve the most favorable Nash equilibrium. During the process of implementing the transit transfer discount policy, in the player set O = {PT operator, travelers who originally chose vehicle travel}, PT operator makes the decision first, and travelers then choose their own strategy based on the operator's decision. The strategy set for the PT operator is G = {large discount, small discount}, while the strategy set for travelers is H = {most travelers switch to public transport, a few travelers switch to public transport}. PT operator provides transport services and charges to generate revenue, while travelers' choices are heterogeneous, depending on their sociodemographic characteristics, trip routes and preferences for fares and quality of service. In the game, it is assumed that both players are rational economic beings who pursue utility maximization.

Assuming that PT operator provides a large discount, and considering the discount also applies to travelers who originally took public transport, the cost is assumed to be 2w, and if a large/small number of vehicle travelers switch to take public transport due to the transfer discount, operator will receive a financial benefit of 2b/b. Assuming the discount is small, the cost is w, and if a large/small number of travelers switch to PT, the operator receives a financial benefit 3b/1.5b. The transfer discount aiming at reducing environmental pollution and alleviating road congestion, which is a matter of public welfare, so it is necessary to consider the environmental protection benefits. It is assumed that if a large/small number of travelers switch to PT, operator will receive environmental protection benefits of 2e/e.

Travel service can be viewed as a special commodity, and travel as a consumption. Therefore, the travelers' travel utility is a measure of their satisfaction with this consumption. As rational consumers, travelers always tend to choose the travel mode with the greatest utility. They decide whether to switch from vehicle to PT transfer based on their own utility function, which takes into account the service quality and fare of the travel mode. The travel utility function is jointly determined by the transport service characteristics of subway-bus public transport and passengers' own preferences. An increase in the quality of public transport services or favorable fares for PT results in an increase in travel utility. Suppose that the payoff utility U of passengers taking public transport in the game is a combination of the fare transfer discount ΔP and the change of PT service quality S, as shown in equation:

$$U = \Delta P + S \tag{1}$$

Suppose when the discount is large, most vehicle travelers switch to PT transfer, ΔP=2d; When the discount is small, most travelers switch to PT transfer, ΔP=d; If most travelers still choose to travel by vehicle, they will not enjoy the discount, ΔP=0. During the discount is implemented, if most travelers switch to PT transfer, there may be a decline in PT service quality, S= -s; If most travelers still choose vehicle travel, it will not have a great impact on the quality of PT service, S=0. There is a discount dynamic game tree model with perfect information, as shown in Figure 2.

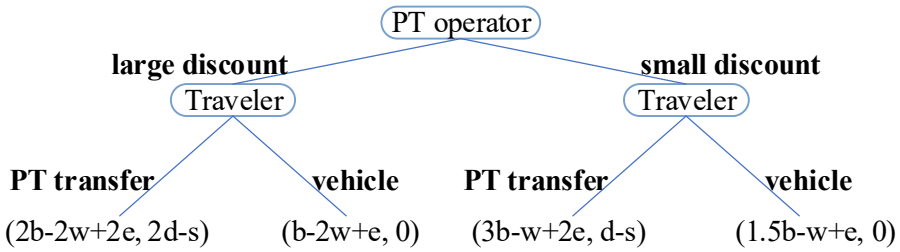

Figure 2. Transfer discount dynamic game model with perfect information.

In this game, PT operator acts first and has only one information set with two optional actions, denoted as SA = {large discount, small discount}. The travelers act after observing the decision of PT operator and have two information sets: PT operator's transfer discount is large and discount is small. There are four feasible pure strategies for them, which are (transfer | large discount, transfer | small discount), (transfer | large discount, vehicle | small discount), (vehicle | large discount, transfer | small discount), (vehicle | large discount, vehicle | small discount), where "|" indicates the condition in.

2.3. Game model analysis and nash equilibrium solution

Calculating game players' payoffs in real-world scenarios can often be a complex process, Therefore, this section solely demonstrates the methodology of utilizing inverse (backward) induction to determine the Nash equilibrium solution of the proposed game model under hypothetical circumstances.

Assuming d<s<2d, the analysis shows that, for travelers, if operator executes a large discount, most travelers will choose PT transfer to obtain a higher payoff of 2d-s>0, compared to vehicle. Conversely, if operator executes a small discount, most travelers will opt for the vehicle as the payoff of choosing PT transfer is lower at d-s. This leaves two rational strategies for travelers: (large discount | transfer) and (small discount | private car), which are rational for travelers. Then, for PT operator, larger discount results in a payoff of 2b-2w+2e compared to 1.5b-w+e for smaller discount. Therefore, assuming 2b-2w+2e > 1.5b-w+e, PT operator will choose to implement larger discount for greater payoff.

Thus, according to the inverse (backward) induction method, the Nash equilibrium of this model is (large discount, (transfer, vehicle)), and the equilibrium result is: PT operator implements large discount, and most vehicle travelers choose PT transfer instead. The derivation process is shown in Figure 3.

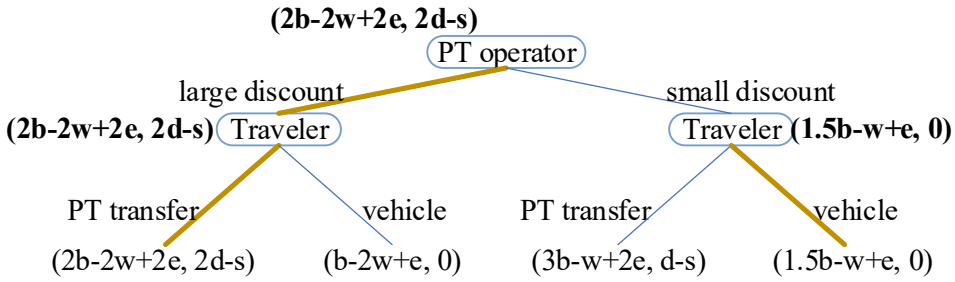

Figure 3. Transfer discount game flowchart.

3. Construction of transfer discount game model based on deep RL

3.1. Construction of transfer discount game model based on the DDPG algorithm

The principle behind the deep reinforcement learning algorithm is that agents interact with the complex environment, receive the environment states, execute actions, and then update the policy network based on the reward received from the environment. This process gradually trains agent to improve the quality of decision-making and achieve the goal, which is

similar to the idea of dynamic game with perfect information. This paper draws on the adaptation process of agent under a complex environment in the DDPG algorithm to design the response function between fare adjustment policy and travelers' travel demand. The DDPG algorithm is a deep deterministic policy gradient algorithm, which is a reinforcement learning algorithm proposed to solve the continuous action control problem.

During each training time t, the transfer discount adjustment policy of public transport trip between OD pair i is set as the PT operator's action variable Δp_t, where the action of Δp_t is the discount. After adjusting the fare discount, the proportion of travelers who switch to the public transfer travel mode in the vehicle travel group q_t (obtained by the travel mode selection model), and the PT fare p_t constitute state variables $\vec{s}_t$ in the algorithm. The payoff of PT operator is taken as the reward value $r(\vec{s}_t, \Delta p_t)$.

In reinforcement learning, the objective of agent is to maximize the cumulative reward value, which includes the current reward value and the future cumulative reward value. With the discount factor γ, the cumulative reward value is expressed as:

$$R_{\mathrm{t}} = \sum_{k=0}^{\infty} \gamma^k r_{t+k} = r_t + \gamma r_{t+1} + \cdots \tag{2}$$

Policy-based reinforcement learning is realized by seeking the optimal policy to maximize the cumulative reward value. The policy of reinforcement learning can be defined as the $\pi(\Delta p \mid \vec{s})$ function relationship between the state variable s and the action variable Δp, and the $\pi(\Delta p \mid \vec{s})$ is expressed as:

$$\pi(\Delta p \mid \vec{s}) = P(\Delta p_t = \Delta p \mid \vec{s}_t = \vec{s}) \tag{3}$$

This policy function calculates the probability value of the action. As the DDPG algorithm used in this paper is a deterministic algorithm, the policy function directly outputs a specific value representing the specific action to control agent to execute the action. To avoid agent from falling into a local optimal solution when selecting the action, the exploration noise ε_t is added in, which can be expressed as:

$$\Delta p_t = \pi(\Delta p \mid \vec{s}_t) + \varepsilon_t \tag{4}$$

As the policy function π is unknown, following the method widely used in deep reinforcement learning, a neural network $\hat{\pi}(\Delta p \mid \vec{s}; \theta)$ is introduced to approximate the policy function, and θ is updated with the policy gradient algorithm to train $\hat{\pi}(\Delta p \mid \vec{s}; \theta)$.

The state will be updated after agent executes an action. To get the long-term impact of the current state on the policy π, the state-value function V_π is introduced, which is the expected value of the reward obtained by agent's choice according to the policy π. The state-value function V_π is expressed as follows:

$$V_\pi(\vec{s}) = E_\pi\left[R_t \mid \vec{s}_t = \vec{s}\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \vec{s}_t = \vec{s}\right] \tag{5}$$

Under state $\vec{s}_t$, by adding the action Δp_t, the value function $Q(\vec{s}_t, \Delta p_t)$ is expressed as:

$$Q_\pi(\vec{s}_t, \Delta p_t) = E_\pi\left[R_t \mid \vec{s}_t = \vec{s}, \Delta p_t = \Delta p\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \vec{s}_t = \vec{s}, \Delta p_t = \Delta p\right] \tag{6}$$

To accurately represent the value function of different fare adjustment policy under complex travel demand evolution conditions, a neural network $\hat{Q}(\vec{s}_t, \Delta p_t; w)$ is introduced to approximate the value function, and trains it with the gradient descent algorithm.

The reward $r(\vec{s}_t, \Delta p_t)$ is composed of the payoff obtained by PT operator with the implementation of the discount adjustment policy Δp_t. To avoid the algebraic loop, agent's action should be delayed by one time unit before being input into the reward function, expressed as:

$$r(\vec{s}_t, \Delta p_t) = p_t \cdot q_{t-1} \cdot g_2 + \Delta p_{t-1} \cdot (g_1 + g_3) + \zeta \cdot q_{t-1} \cdot g_2 \tag{7}$$

where g_1 and g_2 respectively represent the number of people whose initial travel mode in the system is PT transfer and vehicle, g_3 represents the number of people whose travel mode is changed from vehicle to PT under the effect of fare discount, ζ represents the environmental protection factor, q_t is obtained from the evolution of the travel demand of the vehicle traveler group:

$$q_t = \frac{1}{\exp\left[-\left(-\Delta p_t + \mathrm{c}\right)\right] + 1} \tag{8}$$

where c represents the change of PT service quality under the discount system, affected by q_t. Also, a delay module is added to eliminate the algebraic loop: $c = c_0 - \xi \cdot q_{t-1}$, where ξ is the service quality penalty factor. The structure and implementation process of the fare optimization model proposed in this paper are shown in Figure 4, illustrating PT operator adjust the discount margin by predicting the reaction of travelers to the transfer discount, to maximize his own payoff ultimately.

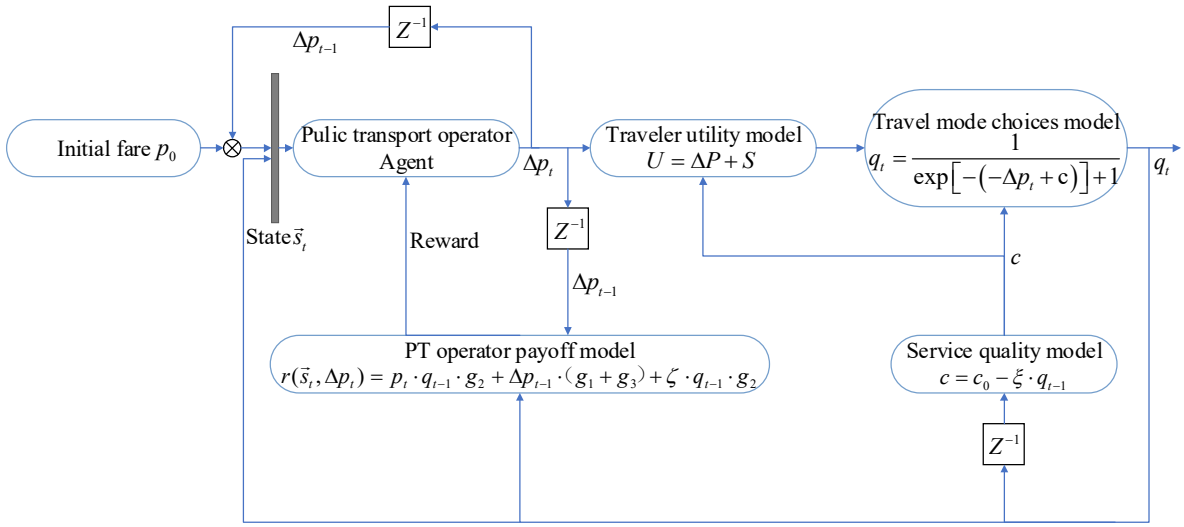

Figure 4. Structure and implementation process of fare optimization model based on DDPG.

3.2. Transfer discount policy training method based on DDPG

This paper utilizes DDPG to train the action variables for fare discount adjustment. Through empirical replay, each state variable and corresponding reward are saved to update the policy network and value network, ultimately resulting in the optimal fare through continuous iteration of the fare discount adjustment policy. In order to enhance training stability, the DDPG algorithm establishes new target policy network $\hat{\pi}'(\vec{s};\theta)$ and target value network $\hat{Q}'(\vec{s}_t, \Delta p_t; w')$ on the basis of current policy network $\hat{\pi}(\vec{s};\theta)$ and current value network $\hat{Q}(\vec{s}_t, \Delta p_t; w)$, and updates the weights of the target networks by slowly tracking the learning networks, to reduce fluctuations and overfitting during the update process. The specific training steps are as follows:

S1. Randomly initialize the policy network $\hat{\pi}(\vec{s};\theta)$ and value network $\hat{Q}(\vec{s}_t, \Delta p_t; w)$, initialize the target networks $\hat{\pi}'(\vec{s};\theta)$ and $\hat{Q}'(\vec{s}_t, \Delta p_t; w')$, initialize the noise ε, and initialize the fare discount adjustment replay buffer D for storing experience samples.

S2. Reset initial state $\vec{s}_t = [p_t, q_t]$. Taking s_t as input, randomly select action $\Delta p_t = \hat{\pi}(\vec{s}_t;\theta) + \varepsilon_t$ with policy network and noise.

S3. Execute the action Δp_t, $p_{t+1} = p_t + \Delta p_t$, then new fare impact on vehicle travelers group decision, environment feedback new state $\vec{s}_{t+1} = \left[p_{t+1}, q_{t+1} \right]$ and reward $r(\vec{s}_t, \Delta p_t)$.

S4. Store transition $\left[\vec{s}_t, \Delta p_t, r(\vec{s}_t, \Delta p_t), \vec{s}_{t+1} \right]$ in D.

S5. In order to make the training data relatively independent and improve stability in the network update process, sample a random minibatch of N transitions $\left[\vec{s}_t, \Delta p_t, r(\vec{s}_t, \Delta p_t), \vec{s}_{t+1} \right]$, to compute:

$$Q_{target} = r(\vec{s}_t, \Delta p_t) + \gamma \cdot \hat{Q}'(\vec{s}_{t+1}, \hat{\pi}'(\vec{s}_{t+1}; \theta'); w') \tag{9}$$

S6. Update value network parameter w by minimizing the Loss function:

$$L(w) = \frac{1}{N} \sum_{t=1}^{N} \left[\hat{Q}(\vec{s}_t, \Delta p_t; w) - Q_{target} \right]^2 \tag{10}$$

that is, update by gradient descent $w \leftarrow w - \alpha \cdot \frac{\partial L(w)}{\partial w}$.

S7. Update policy network parameter θ by maximizing $\hat{Q}(\vec{s}_t, \Delta p_t; w)$, that is, update by the deterministic policy gradient algorithm $\theta \leftarrow \theta + \beta \cdot \frac{1}{N} \cdot \sum_{t=1}^{N} \left[\frac{\partial \Delta p_t}{\partial \theta} \cdot \frac{\partial \hat{Q}(\vec{s}_t, \Delta p_t; w)}{\partial \Delta p_t} \right]$.

S8. Using the soft update method with an update rate ρ, update the target networks $w' \leftarrow \rho w + (1-\rho) w'$; $\theta' \leftarrow \rho \theta + (1-\rho) \theta'$.

S9. Complete one iteration and proceed to Step 2 until training requirements are met.

4. Design and verification of simulation experiment

4.1. Setting of basic parameters of simulation

In the proposed game model of PT operator and vehicle owner based on reinforcement learning, the traveler utility model, travel mode choices model, service quality model and operator utility model usually are typically obtained through data fitting. As this paper is only an introduction to this methodology, the relevant model functions are just assumed based on the models' input-output physical meanings. Table 1 outlines the basic parameters of the model.

Table 1. Basic parameter setting.

Parameter	Symbol	Value	Parameter	Symbol	Value
Episode	ep	200	Sample time	T_s	0.2
Fare discount range	[Δp_{min}, Δp_{max}]	[-1,1]	Experience buffer size	-	10^6
			Experience mini-batch size	N	128
Initial value of the PT fare	p_0	7	Reward discount factor	γ	0.99
The number of neurons in the hidden layer (value network)	-	128,200 (state path)	Target network smoothing factor	ρ	10^{-3}
			Initial noise ε_0 standard deviation	σ	0.3
	-	200 (action path)	Decay rate of standard deviation	-	10^{-5}
			Initial number of people by PT	g_1	100
Value network learning rate	α	10^{-3}	Initial number of people by vehicle	g_2	80
The number of neurons in the hidden layer (policy network)	-	128, 200	Environmental protection factor	ζ	0.5
			Initial quality of service	c_0	0.2
Policy network learning rate	β	5×10^{-4}	Service quality penalty factor	ξ	0.5

4.2. Transfer discount simulation optimization results

Based on the custom models developed in Chapter 3, the simulation experiments of the transfer discount optimization are conducted. The fare variation is set as a continuous variable, and changes within the discount range. Figure 5 shows the evolution of the possibility of vehicle travelers changing to public transport. When the fare is reduced by 1 yuan, the discount is the largest, and the probability of vehicle owner changing to public transportation is the highest, at about 0.93. With the reduction of fare discount, the probability decreases. Thus, the game environment model, which consists of the traveler utility model, travel mode choices model, service quality model and operator utility model in Chapter 3, is reasonable.

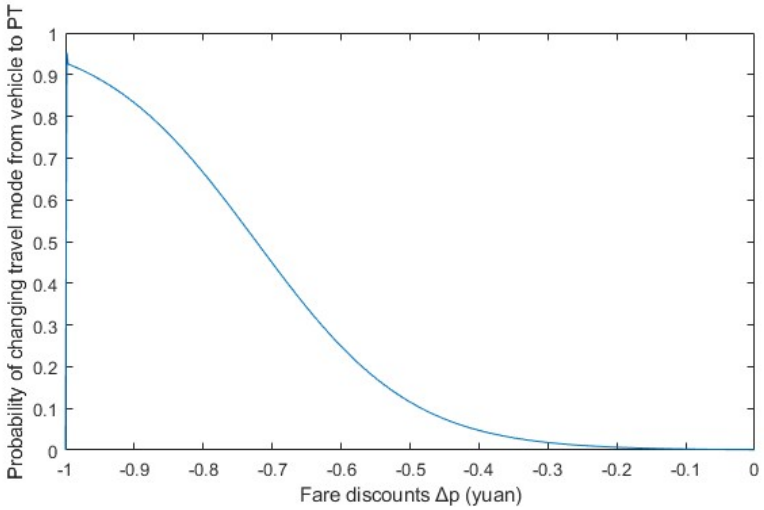

Figure 5. The relationship between discount and the probability of changing the travel mode from vehicle to PT.

As the PT operator agent model is connected to the system, the iterative process of optimization simulation is shown in Table 2, Figure 6 and 7. After approximately 110 iterations, it becomes clear that fare stabilizes at (6.36, 6.48), while the PT operator reward remains roughly stable at (370, 390).

Table 2. Record of training process.

Iterations	Fares	Reward value	Iterations	Fares	Reward value
0-25	[9.70, 9.91]	[-2202, -2037]	67-95	[6.40, 6.56]	[311.8, 385.4]
26-33	[4.54, 8.58]	[-1103, 1851]	96-106	[4]	[-1653]
34-66	[6.46, 7.21]	[-138.5, 341.9]	107-200	[6.13, 6.54]	[103.1, 414.8]

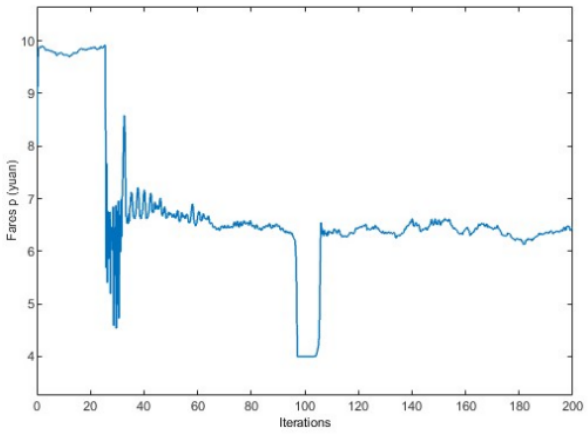

Figure 6. PT fare in optimization process

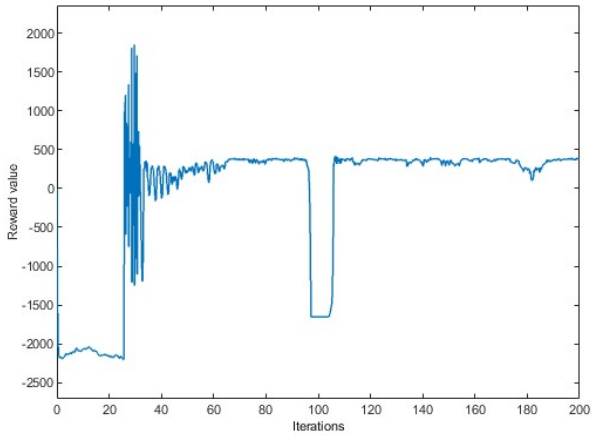

Figure 7. Reward value in training process

The reliability of the effect can be verified as follows. Within the fare discount range, the payoff of PT operator is shown in Figure 6, indicating that there is a maximum value of PT operator's payoff. In Figure 7, the fare obtained by DDPG algorithm is stable in the range (6.36, 6.48), which is converted into the fare change range of (-0.64, -0.52). By marking it in Figure 8, it can be proven that the maximum payoff of the PT operator can indeed be obtained within this range, and the maximum payoff is within the reward range (370, 390) obtained by reinforcement learning. Therefore, the optimal fare obtained based on the reinforcement learning DDPG algorithm is reliable. Moreover, if the fare discount increases or decreases on the basis of the optimal fare obtained by DDPG algorithm, the PT operator's reward will not increase, which proves that the Nash equilibrium has been reached.

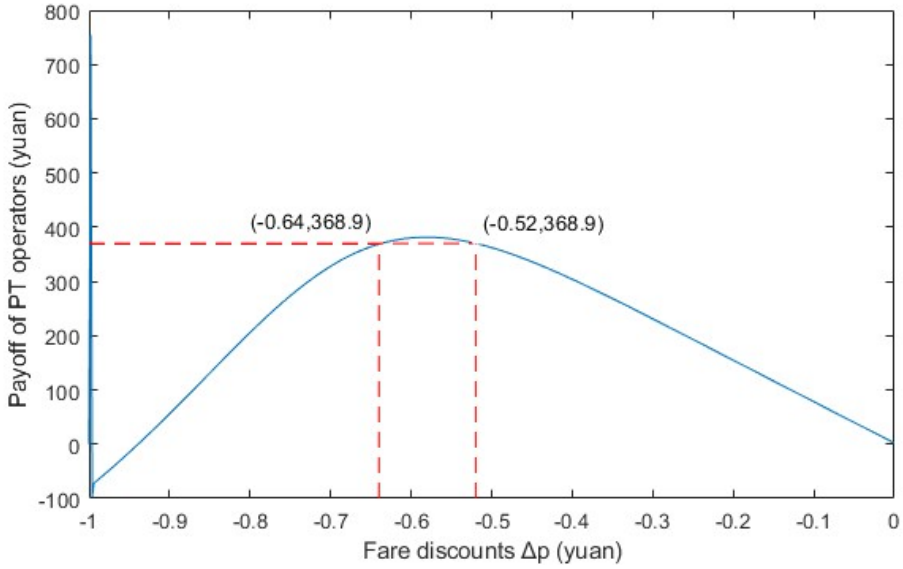

Figure 8. Payoff of PT operator.

5. Conclusions

Urban rail transit and conventional buses complement each other in terms of capacity, accessibility distance and speed, so it is essential to develop reasonable transfer discount policies to encourage public transportation.

(1) Travelers' choices are heterogeneous, if they believe that the benefits of switching to PT outweigh than the inconvenience caused by public transfer, they will opt for public transportation, otherwise, they will still drive. When calculating the payoff of PT operator in the game, the benefits of environmental protection should be taken into account. In the dynamic game with perfect information used in this paper, PT operator, as the first decision-maker, have the strategic advantage. Operator can predict the rational response of later-acting travelers based on their prior information, so as to achieve the most favorable Nash equilibrium for themselves.

(2) The reinforcement learning DDPG algorithm offers the advantages of model-free learning and continuous action spaces handling, making it a suitable approach for optimizing decision-making and performance in public transport systems with varying sizes and complexities, so this research uses it to simulate the game process of operator adjusting fare policy in complex traveler environment. Simulation experiments have verified that the transfer discount game model proposed in this research, is reasonable and reliable. The results show that under ideal circumstance, transfer discount can influence travelers' mode choices and eventually reach an equilibrium state. Assuming an initial fare of 7 yuan, the model reaches a sub-game perfect Nash equilibrium when the discount interval is between -0.68 and -0.52 yuan. As depicted in Figure 5, approximately 14%-30% of vehicle travelers will switch to public transport at this equilibrium. This demonstrates that the transfer discount policy can effectively encourge residents to travel by public transportation and promote the sustainable development of urban transportation.

To a certain extent, the research results provide a theoretical reference for optimizing the transfer discount policy of public transport using reinforcement learning algorithm and game theory. However, in modeling and simulation here, the main purpose is only to apply the algorithm and game theory. Due to the lack of actual data, the assignment of utility function

for each decision-making player is based on subjective inference, while the actual situation is more complex, thus the transfer discount game in the actual traffic environment needs to be further studied. In practical applications, the model parameters in this methodology can be fitted using actual travelers' travel data and PT operation data, and this methodology can serve as a solver for Nash equilibrium to provide a reference for the formulation of PT transfer discount.

REFERENCES

[1] Liu B, Ge Y, Cao K, Jiang X and Meng L. (2022) Optimizing fares and transfer discounts for a bus-subway corridor. *Transport*, 34(6): 672–683.

[2] Ding Y, 2020. Research on Residents' Travel Behavior Choice and Bus Preferential Treatment Based on the Prospect Theory. Jilin: Jilin University.

[3] Lee E, Patwary A, Huang W and Lo H. (2020) Transit interchange discount optimization using an agent-based simulation model. *Procedia Computer Science*, 170: 702-07.

[4] Liao X, 2018. A study of public transportation transfer fare considering travel mode metastasis. Xi'an: Chang'an University.

[5] Liu B, 2017. Integrated Optimization of Fare Strategy and Fare Levels for Subway and Bus in a Transit Corridor. Dalian: Dalian University of Technology.

[6] Jin J, Zhang D, Yang D and Li Y, 2009. Psychological Behavior Analysis of Bus Fare System. *Int. Conf. on Transportation Engineering (ICTE)*. 3142-47.

[7] Li X, Zhang H, Li J and Qiu H. (2022) A Model for Optimizing Urban Public Transport Ticket Prices Based on Deep Reinforcement Learning. *Journal of Industrial Engineering Management*, 36(06): 144-55.

[8] Xia X and Xu L, 2009. Traffic Signal Control Agent Interaction Model Based on Game Theory and Reinforcement Learning. *Int. Forum on Computer Science-Technology and Applications*, 164-8.

[9] Bouton M, Nakhaei A, Isele D, Fujimura K and Kochenderfer M, 2020. Reinforcement Learning with Iterative Reasoning for Merging in Dense Traffic. *IEEE 23rd Int. Conf. on Intelligent Transportation Systems (ITSC)*, 1-6.

[10] Lopez V, Lewis F, Liu M, Wan Y, Nageshrao S and Filev D. (2022) Game-Theoretic Lane-Changing Decision Making and Payoff Learning for Autonomous Vehicles. *IEEE Transactions on Vehicular Technology*, 71(4): 3609-20.