RAIRO-Oper. Res. 56 (2022) 2475–2493
https://doi.org/10.1051/ro/2022103

RAIRO Operations Research
www.rairo-ro.org

DYNAMIC AIR TICKET PRICING USING REINFORCEMENT LEARNING METHOD

Jinmin Gao[1,*], Meilong Le[2] and Yuan Fang[2]

Abstract. This paper studies a dynamic air ticket pricing problem in a strategic and myopic passengers co-existence market. The strategic or myopic passengers can be further divided into high-valuation and low-valuation groups according to how they evaluate their purchases. The strategic passengers have different strategic levels. When the airline sets a ticket price, every passenger makes his or her purchase decision according to his or her type and the strategic level, or might select"wait" or "leave (the market)". The paper firstly proposes a dynamic pricing algorithm in which the utilities of both the airline and passengers are considered. The reinforcement learning (RL) is employed to deal with the progressive or dynamic decision-making framework, in which the dynamic pricing problem is formulated as a discrete finite Markov decision process (MDP) and the Q-learning is adopted to solve the problem. By using this method, the airline can adaptively decide the ticket price based on passengers strategic behaviors and the time-varying demand. The effects of the passenger type proportion and strategic level are analyzed. The computational results show the higher proportion of strategic passengers is, the smaller price increase the airline can adopt, and the higher proportion of high-valuation strategic passengers is, the larger price increase the airline can put to use under the same strategic level. If the proportion of low-valuation strategic passengers is higher, the price increase should be gentle and step by step when the price increase strategy is adopted. If the airline uses price-cut policy, the adjustment should be small. In addition, the high-valuation passenger mainly affects high-price periods and the low-valuation passenger mainly affects low-price periods. When the proportion of strategic passengers is fixed, the lower the passenger strategic level is, the larger the price slope is. These findings can provide some references for the airline to make more precise and flexible pricing decisions.

Mathematics Subject Classification. 90B22, 90C40.

Received June 26, 2020. Accepted June 14, 2022.

1. Introduction

In recent years, as the growth rate of GDP has declined, the growth rate of air transport demand has been weakened. However, the growth rate of the whole air transport capacity has not decreased, resulting in fierce competition between airlines. In such a situation, airlines are eager to seek new approaches to identify passenger demands more accurately and respond to them in a timelier manner.

Keywords. Dynamic pricing, strategic behavior, reinforcement learning, Q-learning, Markov decision process (MDP).

[1] School of Management, Shanghai University of Engineering Science, No. 333 Longteng Road, Shanghai 201620, P.R. China
[2] College of Civil Aviation, Nanjing University of Aeronautics & Astronautics, No. 29 Jiangjun Avenue, Nanjing 211106, P.R. China.
*Corresponding author: Jinmingao@foxmail.com

Revenue management (RM), a method to improve revenue, was firstly used and then introduced by American Airlines (AA) in the US in 1980s due to airlines' serious competition. The common approach for the airline to conduct RM is the seat inventory control. The airline sets several prices or classes for all seats on the flight, then controls the sale number of each class according to the market demand. Obviously, the more classes they set, the more revenue they can get. This leads to the ticket dynamic pricing in 2000s. The dynamic pricing is a method that implies that the airline dynamically changes the ticket price according to the demand. There are also two approaches for the airline to conduct the ticket dynamic pricing. One doesn't consider game between the airline and the passenger, and the other considers the game. Game theory is mainly used to study the rational decision-making behaviors of interdependent and mutually influencing participants in the game and the equilibrium results of these decisions. The game model can be expressed by players, actions, strategies, the information and the utility. Under the game scenario, the airline needs to consider not only their own inventory levels, costs and the impact of competitors, but also the possible choices made by passengers. The passenger will compare different products (tickets) vertically and horizontally, that is, flights at different departure time within one airline and flights in different airlines. Levin *et al.* [21] described the equilibrium at each moment as the Stackelberg equilibrium between the airline and the passenger. The passenger's goal is to maximize his or her utility while the airline wants to capture the passenger's surplus as much as possible to improve its total revenue.

Current studies of dynamic pricing mainly consider two or several time periods in the whole selling period. Such kind of methods cannot respond to demands in a timely manner. Meanwhile, most studies do not well consider passenger behaviors, which is more important to the ticket pricing in today's market. In this paper, we differentiate passengers into several categories, such as the strategic passenger or the myopic passenger. The strategic passenger, that is the passenger who evaluates the present and future purchase utility in order to maximize his or her benefit, may not choose immediate buy and wait for a lower price in the future, and the myopic passenger is the passenger who purchases immediately when the price is less than or equals to his or her valuation (willingness to pay). There could be no doubt that the existence of strategic passengers in the market is a difficult aspect for airlines to make pricing decisions. Shen and Su [32] pointed out the passenger choice behavior is mainly reflected in two aspects, when to buy and which to buy. Gonsch *et al.* [14] reviewed the related studies on dynamic pricing of strategic passengers. They found strategic passengers could continuously predict the airline's expected future price under e-commerce environment, and the airline should have according pricing policy based on the passenger categories and their consuming behaviors in the market. Thus, passengers' classification or categorization and their purchase behaviors become this paper's main focuses.

The rest of this paper is organized as follows. In Section 2, the literature review is presented. Section 3 is the problem description and assumptions. In Section 4, mathematical formulations for passengers and the airline are presented. In Section 5, the reinforcement learning (RL) algorithm is presented, which includes simulation of passenger behaviors and adoption of Q-learning to solve the decision-making problem. Section 6 is the computational experiments and analysis. Finally, the conclusion part is presented.

2. Literature review

In studies of dynamic pricing considering strategic customers, Mersereau and Zhang [28] assumed that the airline knows the total demand curve, but the proportion of strategic customers is unknown. They established a robust pricing model which is independent on the true proportion value. Su [33], Kremer *et al.* [18] divided the sale period into two periods and studied a pricing strategy to analyze the impact of different types of customers on their pricing strategies. Su [33] pointed out that customers are heterogeneous along two dimensions: their valuations for the product and the degree of patience. According to these two dimensions, customers were categorized into four different types, and the different influences on pricing policy were analyzed. Kremer *et al.* [18] suggested that the proportion of strategic customers influences the optimal pricing policy. Zhang and Zhang [37] considered a perishable product in a two-period model in which a retailer decides the order quantity and

price at the beginning of the first period. Customers pay full price in the first period and a marked down price in the second period. Dong and Wu [3] examined the impact of strategic and heterogeneous consumers on pricing and inventory decisions in a two-period model. They found that strategic consumers may yield more revenue in specific scenarios. Li *et al.* [24] established the two-period models to study a platform's discount pricing strategies with strategic consumers. The results show that the large discount will reduce the total demand under the instant strategy, the fraction of strategic consumers affects the platform's strategy choices, and the existence of strategic consumers will increase the product prices for two periods. Guan and Ren [15] divided the entire sale period into the normal sale period and the clearance period. In the paper, it is assumed that the price strategy relates to the proportion of strategic customers and the dependence degree of the reference price. Correa *et al.* [8] proposed a class of preannounced pricing policies in which the price path corresponds to a price menu contingent on the available inventory. Some other studies established two-period dynamic pricing models considering both myopic and strategic customers, and discussed the impact of different strategic customer ratios on pricing strategies [10,16,34,36,38]. All above studies only consider the dynamic pricing in two periods.

Most studies about multi-period dynamic pricing only consider strategic customers. Levin *et al.* [20, 21] established a pricing model for oligopolistic companies and monopoly companies respectively, and proved that the monotony of pricing strategies relates to the degree of the customer rationality. Through learning from the customer arrival rate and reservation price, Levina *et al.* [22] proposed a strategic waiting factor which was used in the customer choice model. Liu and Zhang [25] considered two companies which provide two vertical heterogeneous products. Customers choose a product according to its quality. When customers become more strategic, the company's revenue will reduce, and the company selling low-quality products suffers more loss than the company selling high-quality products. Chen and Farias [4] acknowledged that the customer is forward looking and his or her valuation of the product decreases with time. The robust pricing strategy can be obtained when the discount factor and the cost distribution are unknown to the customer. Li *et al.* [23] used the Bayesian posterior probability to update the arrival rate based on the past sale experience and the number of passengers' arrival. From their studies, we can find most studies assumed there are only strategic passengers in the market, and all passengers are homogeneous with the same valuation or the same value distribution to the same product. There are few studies considering strategic and myopic customers in multiple periods [17].

With the continuous development of artificial intelligence technology, more and more scholars have tried to use intelligent methods to solve the problem of dynamic pricing. The reinforcement learning (RL) is one of the most widely used methods [1, 5–7, 9, 11, 12, 29–31, 35]. Among them, Collins and Thomas [6] incorporated a variety of customer demand models within a simple airline pricing game to gauge the usefulness of three different RL approaches as a game theoretic solving mechanism. The results prove that the application of RL to the game is beneficial, and the benefit is both from solving games that are unsolvable using traditional methods and by giving an extra-dimension of insight into the game. In addition, Collins and Thomas [7], Dogan and Gner [11] used a decision framework of the Markov decision process (MDP) and Q-learning algorithm to study the dynamic pricing problem, but the customer behaviors were not considered in the learning process. In smart grids, RL was also applied in the dynamic pricing, where pricing strategies are learned in a customer simulation environment [26, 27].

Being different from above-mentioned studies, this paper analyzes passenger purchase behaviors more accurately, considers the variation of passenger arrival rates and reservation prices, and conducts dynamic pricing with myopic and strategic passengers' co-existence in multiple time periods. The heterogeneity of passengers is considered along two dimensions: the valuation and the strategic level, which jointly classify passengers into four categories or types. The pricing model and the RL algorithm are established to obtain the optimal dynamic price strategies. In the RL algorithm, passengers behaviors are simulated as a learning environment and the airline is set as the agent. Further, for different passenger types, the impacts of their strategic levels and their proportions on optimal price policies are analyzed.

3. Problem description and assumptions

3.1. Problem description

The airline sells a certain number of air tickets within a finite time period. This certain number Y is the total number of seats of a flight which can be sold. The objective is to maximize the total revenue of the flight. Owing to the special property of air tickets, the residual value of unsold tickets is zero after the sale time ends. The entire sale time is divided into T time periods, $t = \{1, 2, \ldots, T\}$. The time period should be small enough to guarantee at most one passenger arrives in each time period. p_t denotes the ticket price in time period t. Assume there are N passengers in the market, and n denotes the number of passengers who have purchased tickets. The airline and passengers are rational and they try to maximize their own utility. Let $U(t, n)$ and $V(t, n)$ denote the total utility of the passenger and the airline respectively from time period t to the end of the sale time with n passengers have purchased tickets.

There are two basic types of passengers called strategic and myopic passengers in the market. Each type has its own arrival rate. After a passenger arrives, he or she should determine whether to buy the priced ticket given by the airline. For myopic passengers, if the price is less than or equal to their valuation, they choose to buy. Otherwise, they choose to leave the market. For strategic passengers, they compare the current utility with the future utility, and then decide whether to buy immediately or wait for the lower price in the future or leave the market. If the current utility is more than the future utility, they choose to buy, otherwise, they keep in a "wait-and-see" state. The price, the left tickets and the waiting time are all under consideration of strategic passengers for their decisions.

(1) The impact of the left tickets on strategic passengers

Strategic passengers usually have strong willingness to buy tickets, which is represented by the purchase probability, but they are unwilling to buy if the price is higher than their valuation. In addition to the price, the seat inventory also plays a role in their decision-making. When the inventory is sufficient, they have no risk of future purchasing and will choose to wait until the price is equal or lower than their valuation. When the inventory becomes scarce, their purchase behaviors closely relate to their purchase willingness. If the willingness is strong, they choose to buy immediately. Otherwise, they still wait for the possible price cut.

(2) The impact of the waiting time on strategic passengers

The waiting time also affects the purchase behavior of strategic passengers. If the strategic passenger has a weak willingness to buy and the price does not fall below his or her valuation after a long-time waiting, he or she chooses to leave the market. How long the strategic passenger can wait relates to his or her willingness. If the willingness is strong, the waiting time is relatively long. In addition, some passengers may worry about the price increase after a long-time waiting, so they choose to buy as soon as possible.

The strategic passenger's behavior can be described by the purchase probability density function. Figure 1 is used to illustrate the relationship between price p, purchase probability density $f(p)$ and purchase probability $\bar{F}(p)$. The reservation price p reflects different passengers have different willingness to pay. To a particular setting price p, airlines can only capture the passengers whose reservation prices(willingness to pay) are higher than and equal to p. The area under the curve from p to the infinite is the purchase probability $\bar{F}(p)$ in the whole market, $F(p)$is the integral of $f(p)$ from zero to price p. Obviously, $\bar{F}(p) = 1 - F(p)$.

Curve 2 is gotten by moving curve 1 to the right to match the changed situation. It can be used to illustrate the effect of the remaining seats or the remaining selling periods. The decrease of remaining seats or selling periods leads to the increase of the willingness to pay (the passenger's reservation price). To a setting price p, the whole purchase probability increases (the area under the curve 2 is larger than the area under the curve 1) as the remaining seats or the selling periods decrease.

3.2. The assumptions

Assumption 3.1. *Passengers are heterogeneous along two dimensions, the ticket valuation and the strategic level. In the valuation dimension, we consider two types, which is the high valuation and the low valuation.*

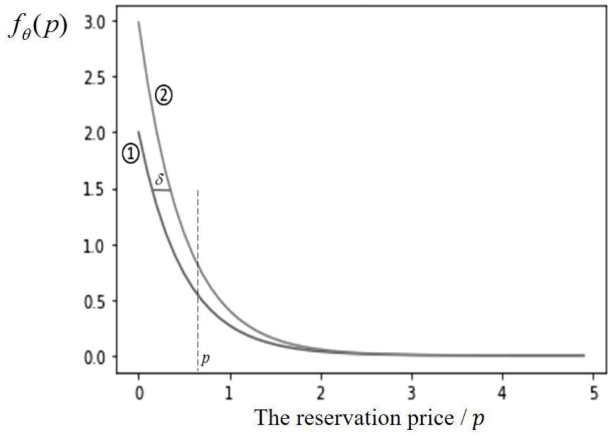

FIGURE 1. The reservation price distribution.

Let v_H and v_L denote the mean value of the high-valuation passengers and the low-valuation passengers, respectively. In the strategic level dimension, we divide passengers into "strategic"and "myopic". Combined the valuation and the strategic dimension together, we have four types of passengers who are the high-valuation strategic passengers (HS), the high-valuation myopic passengers (HM), the low-valuation strategic passengers (LS) and the low-valuation myopic passengers (LM). Let θ denote the passenger type, so $\theta = \{HS, HM, LS, LM\}$.

Assumption 3.2. *The total number of passengers, the proportion of each type and the strategic level can be deduced from the historical data.*

Assumption 3.3. *Passengers' arrival follows a non-stationary Poisson process. $\lambda_\theta(t)$ denotes passengers' arrival intensity of type θ, and it is a time-varying intensity. The probability of distribution is expressed as equation (3.1), which indicates the probability of k arrivals until time period t.*

$$P(N(t) = k) = \frac{e^{-m_\theta(t)} \left(m_\theta(t)\right)^k}{k!} \tag{3.1}$$

where $m_\theta(t) = \int_0^t \lambda_\theta(s)ds$, it is the mean function of the non-stationary Poisson process.

Assumption 3.4. *The airline and passengers have perfect knowledge of all market information, including the remaining capacities of the flight, the distributional and parametric characteristics of the passenger's reservation price. In addition, the reservation price distributions of all passenger types are independent.*

Assumption 3.5. *Team passengers (batch arrivals) are not considered.*

The above assumptions about the passenger type and the passenger arrival are common in this research field. The perfect information assumption is also reasonable. Nowadays, the airline has many ways to know the market information. First, through GDS such as TravelSky, Worldspan and Galileo, the airline can know the selling progresses of all airlines in the market. Second, through cooperated OTAs and its own EC website, the airline has ability to know the passenger's behavior (such as how many times he or she explored relative platforms, what is his or her interest, what is his or her focus, etc.), which contains many features that are very valuable for the airline's decision. The passenger also has some ways to sense the market through various platforms or personnel channels. In addition, the perfect information is an important limiting case with significant potential for management insight regarding the pricing and other policies.

4. Mathematical formulations

4.1. The passenger utility model

A fraction α of the total passengers are strategic passengers, and the remaining $\bar{\alpha} = 1 - \alpha$ are myopic passengers. Among strategic passengers, the proportion of high-valuation passengers is φ_s, and the proportion of low-valuation passengers is $\bar{\varphi}_s = 1 - \varphi_s$. Similarly, among myopic passengers, the proportion of high-valuation passengers is φ_m, and the proportion of low-valuation passengers is $\bar{\varphi}_m = 1 - \varphi_m$. w_θ denotes passengers' proportion of type θ in the total passengers. So, the proportions of the four types of passengers are $w_{\theta=HS,HM,LS,LM} = \{\alpha^*\varphi_s, \bar{\alpha} * \varphi_m, \alpha^*\bar{\varphi}_s, \bar{\alpha} * \bar{\varphi}_m\}$. The strategic level of the passengers of type θ is β_θ. It reflects how much the passenger values his/her future purchase. For a type θ passenger, the utility of buying a ticket in the future is discounted by $\beta_\theta \in [0, 1]$. $\beta_\theta = 0$ means that the passenger completely disregards the possibility of a future purchase. $\beta_\theta = 1$ means that the passenger values the current purchase the same as a purchase at any point in the future. Intermediate values of $\beta_\theta \in (0, 1)$ determine how long passengers can postpone their purchases without excessive loss of utility. Obviously, $0 < \beta_{LS}, \beta_{HS} \leq 1$, $\beta_{HM,LM} = 0$.

The passenger utility function is expressed as $u(.)$. For a particular ticket, the passenger utility is denoted as $u\left(p'_\theta - p_t\right)$. p'_θ represents the passenger valuation of type θ. p_t represents the ticket price in time period t as mentioned in Section 3.1. $u(.)$ is a strict increasing function of the price, and the inverse function is $u^{-1}(.)$. No matter whether the passenger is myopic or strategic, if he or she chooses to buy the ticket immediately when he or she arrives at the market in time period t, his or her expected utility is $E_{p'_\theta}\left[u\left(p'_\theta - p_t\right)\right]$. The future passenger utility is expressed as $U(t+1, n)$. For a strategic passenger, the goal of the model is to capture his or her intertemporal behavior. Then, the present value of the future passenger utility is represented as $\beta_\theta * U(t+1, n)$.

$\lambda_{t,\theta}$ denotes the arrival rate of passengers of type θ in time period t. The range of $\lambda_{t,\theta}$ is $[0, \bar{\lambda}]$, where $\bar{\lambda}$ is the maximum arrival intensity. According to Levina *et al.* [22], the purchase probability is related to the number of remaining seats (or the number of passengers who have purchased tickets), the current time period and the ticket price, which is expressed as $\Lambda^U_\theta\left(t, n, p_t\right)$. Then,$\Lambda^U\left(t, n, p_t\right) = E_\theta\left[\Lambda^U_\theta\left(t, n, p_t\right)\right]$, which means $\Lambda^U\left(t, n, p_t\right)$ is equal to the expected purchase probability of all passenger types. $U^p\left(t, n, p_t, p'_\theta\right)$ represents the present value of the passengers expected utility with price p_t, valuation p'_θ in time period t and n tickets purchased by passengers. The passenger utility is the function of $\lambda_{t,\theta}$.

$$\begin{aligned} U^p\left(t, n, p_t, p'_\theta\right) = \max_{0 \leq \lambda_{t,\theta} \leq \bar{\lambda}} \big\{ & \lambda_{t,\theta} u\left(p'_\theta - p_t\right) + \beta_\theta \Lambda^U\left(t, n, p_t\right) U(t+1, n+1) \\ & + \beta_\theta \left(1 - \lambda_{t,\theta} - \Lambda^U\left(t, n, p_t\right)\right) U(t+1, n) \big\}. \end{aligned} \tag{4.1}$$

The first item represents the utility if the passenger chooses to buy the ticket immediately. Both the second and the third item are the present value of the future utility. The difference between the second and the third item is the passengers choosing to wait or leave. By merging similar items, the expression can be rewritten as:

$$\begin{aligned} U^p\left(t, n, p_t, p'_\theta\right) = \max_{0 \leq \lambda_{t,\theta} \leq \bar{\lambda}} & \left\{\lambda_{t,\theta}\left(u\left(p'_\theta - p_t\right) - \beta_\theta U(t+1, n)\right)\right\} \\ & + \beta_\theta \Lambda^U\left(t, n, p_t\right)\left(U(t+1, n+1) - U(t+1, n)\right) + \beta_\theta U(t+1, n). \end{aligned} \tag{4.2}$$

Because $\lambda_{t,\theta} \geq 0$, $u\left(p'_\theta - p_t\right) \geq \beta_\theta U(t+1, n)$ is guaranteed to the passenger's instant purchase. It directly explains that if the immediate utility is greater than or equal to the present value of the expected future utility, it is ensured the passenger can maximize its own utility. Because the inverse function of the utility function exists, the condition can be transformed as $p'_\theta \geq p_t + u^{-1}\left(\beta_\theta U(t+1, n)\right)$.

Set $U^p\left(t, n, p_t\right) = E_{p'_\theta}\left[U^p\left(t, n, p_t, p'_\theta\right)\right]$, which represents the expected value of the passenger utility in all possible valuations. It is used as the passenger utility in the status (t, n).

$$U(t, n) = U^p\left(t, n, p_t\right), n = 0, \ldots, Y; t \in \{1, \ldots, T\}. \tag{4.3}$$

The termination condition is:

$$U(T,n)=0, n=0,\ldots,Y \tag{4.4}$$
$$U(t,Y)=0, t\in\{1,\ldots,T\}. \tag{4.5}$$

Equations (4.4) and (4.5) mean when the sale time ends or there is no seat left, the passenger utility is zero.

In reality, the following three purchase behaviors or cases will occur under condition $u\left(p'_\theta-p_t\right)\geq\beta_\theta U(t+1,n)$. Let $f^i_{\theta,t}$ denote the probability of purchase by the arrival passenger of type θ in case $i(i=1,2,3)$ in time period t.

Case 1. The myopic passengers will choose to buy tickets immediately when the current utility is greater than or equal to zero.

$$u\left(p'_\theta-p_t\right)\geq 0, \theta=HM,LM. \tag{4.6}$$

The purchase probability of the arrival myopic passenger is:

$$f^1_{\theta,t}=\bar{F}_{\theta,t}\left(p_t+u^{-1}(0)\right), \theta=HM,LM. \tag{4.7}$$

Case 2. The strategic passengers will choose to buy tickets only if the current utility is more than or equal to the present value of the utility that they may get in the future.

$$u\left(p'_\theta-p_t\right)\geq\beta_\theta U(t+1,n), \theta=HS,LS. \tag{4.8}$$

The purchase probability of the arrival strategic passenger is:

$$f^2_{\theta,t}=\bar{F}_{\theta,t}\left(p_t+u^{-1}\left(\beta_\theta U(t+1,n)\right)\right), \theta=HS,LS. \tag{4.9}$$

Case 3. When the utility is more than or equal to zero, but less than the present value of the future expected utility, the arrival strategic passenger will choose to wait.

$$0\leq u\left(p'_\theta-p_t\right)<\beta_\theta U(t+1,n), \theta=HS,LS. \tag{4.10}$$

The probability of the arrival strategic passenger who chooses to wait is:

$$f^3_{\theta,t}=F_{\theta,t}\left(p_t+u^{-1}\left(\beta_\theta U(t+1,n)\right)\right)-F_{\theta,t}\left(p_t+u^{-1}(0)\right), \theta=HS,LS. \tag{4.11}$$

Let $G_t=w_{HM}f^1_{HM,t}+w_{LM}f^1_{LM,t}+w_{HS}f^2_{HS,t}+w_{LS}f^2_{LS,t}$, the average arrival intensity in time period t is λ_t, so the probability that a myopic or strategic passenger arrives and chooses to buy the ticket in time period t is:

$$\lambda_t * G_t=\lambda_t *\left(\sum_{\theta=HM,LM} w_\theta\bar{F}_{\theta,t}\left(p_t+u^{-1}(0)\right)+\sum_{\theta=HS,LS} w_\theta\bar{F}_{\theta,t}\left(p_t+u^{-1}\left(\beta_\theta U(t+1,n)\right)\right)\right). \tag{4.12}$$

The probability that a strategic passenger arrives and chooses to wait in time period t is:

$$\lambda_t *\left(w_{HS}f^3_{HS,t}+w_{LS}f^3_{LS,t}\right)=\lambda_t *\left(\sum_{\theta=HS,LS} w_\theta F_{\theta,t}\left(p_t+u^{-1}\left(\beta_\theta U(t+1,n)\right)\right)-\sum_{\theta=HS,LS} w_\theta F_{\theta,t}\left(p_t+u^{-1}(0)\right)\right). \tag{4.13}$$

The probability that there is no passenger arrives or if he or she arrives, he or she chooses not to buy the ticket in time period t is:

$$1-\lambda_t *\left(G_t+w_{HS}f^3_{HS,t}+w_{LS}f^3_{LS,t}\right). \tag{4.14}$$

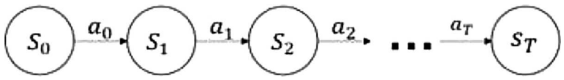

FIGURE 2. The state transition of MDP.

Obviously, the buying probability depends on the expected future utility $U(t+1,n)$, and the expected future utility depends on its rationality. According to Levin *et al.* [20], if a passenger is completely rational, he or she will adopt a balanced solution between the passenger utility and the airline utility. p_t in $U^p(t,n,p_t)$ should be a balanced solution $p^U(t,n)$ through gaming, and the passenger utility can be represented as $U^p\left(t,n,p^U(t,n)\right)$. However, passengers are not completely rational in reality and they are also impossible to know the complete pricing information. Hence, we suppose passengers are partially rational and they use their past purchase experience to estimate the expected value of the future utility.

4.2. The airline's pricing model

We use the Markov decision process (MDP) to construct a dynamic programming model for the airline's pricing behavior. MDP is a discrete time state transition process. It consists of five elements (S,A,P,R,γ). S denotes the set of states. A denotes the set of actions which the airline can select. R is the set of expected immediate rewards. P is the state transition probability given by $p(s',r \mid s,a)$, which means the probability that the state changes from s to s' under action a, $r \in R$. $\gamma \subset [0,1]$ is the discount factor, which represents the difference of importance between the future and present reward. The state transition of MDP must satisfy the Markov property, which is the next state s' only depends on the current state s and the decision-maker's action a. That is to say, given s and a, it is conditionally independent of all previous states and actions, for which we can use the following equation $p(s',r \mid s,a) = \Pr\{S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a\}$. It means the behavior in the time period t is only related to the state of time period $t-1$. For the dynamic pricing problem, the state is n, which denotes how many seats have been sold. It can be written as $S_t = n$, which means n seats sold in the current time period. The selectable action in this dynamic pricing problem is a set of prices A. The price in time period t is $p_t \in A$. $V(t,n)$ denotes the total revenue from the start state to the end of the sale time. Figure 2 is used to explain the state transition of MDP, it contains the state nodes s_t and the action nodes a_t, after implementing action a_t according to a certain strategy, the state will change from s_t to s_{t+1}. Actions here can be purchase, wait or leave. Figure 3 is used to explain the ticket sale process. Under each state-action pair, there are two possible results. If a ticket is sold, the reward is equal to the current price; if not, the reward is zero. When a ticket is purchased by a passenger, the number of passengers that purchased tickets is added by 1, otherwise, it keeps the same. In Figure 3, the right branch illustrates the action that the passenger bought a 30-dollar ticket set by the airline, the revenue increase is 30 dollars, the purchased number increases by 1, the time enters into the next period. It is the state changes from $S_{t=3}=2$ to $S_{t=4}=3$.

Considering the presence of all the four passenger types, if the current state is n at time period t, there are also three behaviors or cases.

Case 1. When a myopic or strategic passenger arrives with a certain probability and chooses to buy a ticket, the revenue state of the next time period is transferred to $V(t+1,n+1)$ and the gain obtained is $R_t = p_t$. Then, the revenue in this scenario is:

$$\lambda_t * G_t * [p_t + V(t+1,n+1)]. \tag{4.15}$$

Case 2. When a strategic passenger arrives and chooses to wait, the revenue state of the next time period is $V(t+1,n)$ and the return R_t is zero. Then, the revenue in this scenario is:

$$\lambda_t * \left(w_{HS} f^3_{HS,t} + w_{LS} f^3_{LS,t}\right) * V(t+1,n). \tag{4.16}$$

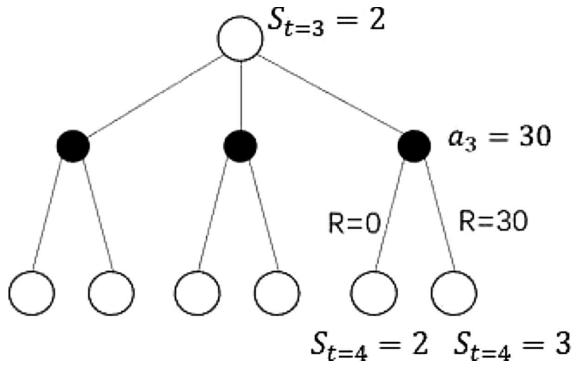

FIGURE 3. The state-action tree.

Case 3. When a passenger does not arrive or chooses to leave the market after his or her arrival, the revenue state of the next time period is also $V(t+1,n)$ and the return is zero. Then, the revenue in this scenario is:

$$\left[1-\lambda_t*\left(G_t+w_{HS}f^3_{HS,t}+w_{LS}f^3_{LS,t}\right)\right]*V(t+1,n). \tag{4.17}$$

On equations (4.16)–(4.17), $V(t+1,n)$ is numerically the same as $V(t,n)$ if we just consider the present revenue or income. No sale, no income. However, the purchase probability relates to the seats sold (or the remaining seats) and the current time period. Meanwhile, passengers' valuation of the ticket also relates to them. If we consider the future influence of passengers' purchase behaviors, the total revenues to the end of the sale time $V(t,n)$ and $V(t+1,n)$ are different.

So, the total revenue under all the three scenarios in state $S_t = n$ is:

$$\begin{aligned} V(t,n) &= \lambda_t*G_t*[p_t+V(t+1,n+1)]+\lambda_t*\left(w_{HS}f^3_{HS,t}+w_{LS}f^3_{LS,t}\right)*V(t+1,n)\\ &\quad+\left[1-\lambda_t*\left(G_t+w_{HS}f^3_{HS,t}+w_{LS}f^3_{LS,t}\right)\right]*V(t+1,n)\\ &= V(t+1,n)-\lambda_t*G_t*[V(t+1,n)-V(t+1,n+1)]+\lambda_t*G_t*p_t. \end{aligned} \tag{4.18}$$

Since the distribution $\bar{F}_{\theta,t}(p)$ is a monotonically decrease function, and G_t is also a monotonic decrease function, equation (4.18) can be expressed in the form of MDP.

$$V(t,n) = (1-\lambda_t G_t)\,V(t+1,n)+\lambda_t G_t\,[p_t+V(t+1,n+1)] \tag{4.19}$$

where,$n=0,\ldots,Y$; $t\in\{1,\ldots,T\}$.

The termination condition is:

$$V(T,n)=0, n=0,\ldots,Y \tag{4.20}$$

$$V(t,Y)=0, t\in\{1,\ldots,T\} \tag{4.21}$$

It means when the sale time ends or all the tickets are sold out, the sale process is over and the residual value is zero.

5. Algorithm design

Due to the large scale of possible states, it is difficult to solve the pricing model directly by an optimizer, so we designed a RL algorithm. RL is a sequence of decisions that maximize the total future rewards through a certain sequence of action choices. RL uses a trial-and-error approach. This algorithm contains two important

parts, namely the environment and the agent. The passenger purchase behaviors are used as the environment. The airline is the agent. The agent constantly interacts with the environment to get the best pricing policy. The environment is an external system, and the agent or airline makes decision action in the environment and gets a certain reward.

When the airline makes a decision, that is to set a price for the ticket in the current time period, the environment gives feedback, that is the passenger determines to buy or not. If the passenger buys the ticket, the immediate revenue of the airline is the current price p_t, and if not, it has no revenue. The airline updates the state based on the feedback from the environment. Through these interactions between the environment and the airline, the optimal pricing strategy can be reached. These kinds of interactions can be reflected by a RL process. Particularly, the $Q(\lambda)$ algorithm in RL is used in this study.

5.1. Simulation of passenger behaviors

Based on the method suggested by Campbell [2], we designed a simulation algorithm for passenger behaviors. The arrival time and the passenger type are determined according to the arrival intensity and the proportion, respectively. The entire sale time is divided into small time periods. The period should be small enough to guarantee there is at most one passenger arrival. The valuation distribution and the strategic behavior are used to judge whether the passenger will buy.

The detail simulation of the passenger purchase behavior is described as follows:

(1) Parameter setting

Set $\alpha, \varphi_s, \varphi_m, \lambda_{t,\theta}, \beta_\theta$, which relates the proportion of different passenger types, their arrival intensity and the strategic level. Different types have different purchase behaviors, which are reflected by their valuation distribution functions. As shown in Figure 1, the valuation distribution function is derivative of the probability, or the probability is the integral of valuation distribution function within certain range. The postpone purchase (wait) can be reflected by the right movement of the curve. For strategic passengers, the distance parameter $\delta(t, x)$ is set to reflect the effect of the waiting time t and the remaining seat inventory x.

(2) Generation of passengers of different types

According to the proportion of different passenger types, the interval $[0, 1]$ is divided into $[0, w_{HS})$, $[w_{HS}, w_{HS} + w_{HM})$, $[w_{HS} + w_{HM}, w_{HS} + w_{HM} + w_{LS})$ and $[w_{HS} + w_{HM} + w_{LS}, 1]$. A generated random decimal lies in $[0, 1]$ determines the passenger type according to which range the decimal belongs to.

(3) Simulation of passenger arrivals

This arrival process is simulated in time period sequence $0 = t_0 \leq t_1 \leq t_2 \leq \ldots$ by using a Poisson process. $\{N(t), t \geq 0\}$ represents a Poisson process. $\lambda = E[N(1)]$ represents the expected number of arrivals in one period. Because the length of one period is 1, this value represents the arrival intensity. Since the same type passengers do not arrive at the same intensity in different time periods, a non-stationary Poisson process is used. $\lambda(t)$ is used to represent the time-varying intensity. In this non-stationary Poisson process, it is still satisfied that at most one passenger arrives in one time period, and that a passengers arrival is an independent event.

The algorithm used is based on the sparse algorithm proposed by Gaver *et al.* [13]. Firstly, a stationary Poisson process with a constant rate λ^* and the arrival time $\{t_i^*\}$ is generated. Then, "sparse" is enacted to t_i^*. In order to get different arrival intensities, each t_i^* is discarded with a certain probability to avoid idle period as the following step (iv) indicates. λ^* is the maximum value in the arrival rate function $\lambda(t)$. If $\lambda(t_i^*)$ is large, it is more likely to generate a passenger arrival in period t_i^*, so we can achieve different arrival intensities in different time periods.

The algorithm can be described as the following steps.

(i) Set $t = t_{i-1}$;
(ii) Generate random decimals U_1 and U_2 respectively according to the uniform distribution $U(0, 1)$;
(iii) Replace t with $t - (1/\lambda^*) \ln U_1$;
(iv) If $U_2 \leq \lambda(t)/\lambda^*$, return $t_i = t$, otherwise, go to step (ii).

(4) Determination of passenger purchase behavior

A two dimensional array $W = [\delta(t_1, x_1), \delta(t_2, x_2) \ldots \delta(t_i, x_i) \ldots]$ is added for storing every strategic passenger in waiting. Each $\delta(t, x)$ contains the passenger waiting time and the number of remaining seats in the current time period. In each time period, when the airline interacts with the environment, the purchase behavior is determined by the current valuation distribution. If a passenger chooses to wait, he or she is added to W. Meanwhile, if a passenger in W chooses to buy the ticket or leave the market, the according $\delta(t_i, x_i)$ is deleted from W. W is dynamically updated as time elapses.

5.2. The $Q(\lambda)$ algorithm design

S represents a set of all possible states, $s = s_t, s' = s_{t+1}$. The policy or strategy of pricing is expressed as π, which can achieve the mapping from the state to the action, $\pi : S \to A$. Of course, every policy is under consideration. $\pi(a \mid s)$ is actually the probability of performing action a in the state s. $f(s' \mid s, p)$ is the probability that the state changes from s to s' when setting the ticket price to p. The probability can be gotten by $\lambda_t G_t$ and $1 - \lambda_t G_t$ which depends on passengers' choices. $\lambda_t G_t$ is the probability of one ticket sold, and $1 - \lambda_t G_t$ is the probability of no ticket sold. $V^\pi(s)$ is the total revenue that can be obtained from state s to the end of the sale time under policy π. Rewrite equation (4.18) as a state value function:

$$V^\pi(s) = \sum_p \pi(p \mid s) \sum_{s',r} f(s', r \mid s, p) [r + V^\pi(s')]. \tag{5.1}$$

Equation (5.1) shows that the expected revenue from state s is equal to the sum of the expected revenue from the next state s' and the revenue of this time, thus constructing a recurrence relation. Here, $\pi(p \mid s)$ is used to replace $\pi(a \mid s)$ since action here is pricing.

To learn the optimal pricing policy through time periods, the current state s and the price p should be given. The action-value function in case of the following strategy π is defined as $Q^\pi(s, p)$:

$$Q^\pi(s, p) = \sum_{s',r} f(s', r \mid s, p) \left[r + \sum_{p_t} \pi(p_t \mid s') Q^\pi(s', p_t) \right]. \tag{5.2}$$

Equation (5.2) indicates the expected revenue that can be obtained when the action p is selected by the strategy π in the current state s.

According to equations (5.1)–(5.2), we can get the relationship between the state value function and the action-value function.

$$V^\pi(s) = \sum_p \pi(p \mid s) Q^\pi(s, p) \tag{5.3}$$

$$Q^\pi(s, p) = \sum_{s',r} f(s', r \mid s, p) [r + V^\pi(s')]. \tag{5.4}$$

The Q-learning algorithm is an offline learning algorithm. Its basic idea is that the individual learns based on another policy although the individual has a strategy of its own. This policy can be a previous policy, or it can be some mature policies, such as human strategies. By observing behaviors based on such strategies, we can get some rewards. These rewards are used to update the action-value function. The strategy of updating the action-value function is different from the strategy of choosing the action. Q-table is used to calculate the final results. Q-table is a two-dimensional matrix. One dimension is the action and the other is the state. Q-value in Q-table is also called "quality". The greedy strategy is adopted in the action choice. That is, the update of the action-value function follows the principal that only maximum Q-value action in all next possible actions will be taken. The update formula for Q-learning is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta \left[r_t + \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]. \tag{5.5}$$

Here, η denotes the learning rate.

Particularly, the ε-greedy strategy is used to choose the action. The target of the ε-greedy strategy is to guarantee that each possible action will have a non-zero probability to be chosen in each time period. It is an exploitation of choosing the best action in the current state with the probability of $1-\varepsilon$, and an exploration of choosing other actions with the probability of ε.

$$\pi(a \mid s) = \begin{cases} 1-\varepsilon \text{ if } a = \arg\max_{a_t \in A} Q(s, a_t) \\ \frac{\varepsilon}{m} \qquad\qquad \text{else} \end{cases} \tag{5.6}$$

m denotes the action number. Here, it is the pricing or setting price times.

Proposition 5.1. *For any given strategy π, the strategy can be improved when using the ε-greedy strategy to evaluate it.*

Proof. Define $a^* = \arg\max_{a \in A} Q^{\pi}(s, a)$.

$$\begin{aligned} Q^{\pi}(s, a^*) &= \sum_{a \in A} \pi(a \mid s) Q^{\pi}(s, a) \\ &= \frac{\varepsilon}{m} * \sum_{a \in A} Q^{\pi}(s, a) + (1-\varepsilon) * \max_{a \in A} (Q^{\pi}(s, a)) \\ &\geq \frac{\varepsilon}{m} * \sum_{a \in A} Q^{\pi}(s, a) + (1-\varepsilon) * \sum_{a \in A} \frac{\pi(a \mid s) - \frac{\varepsilon}{m}}{1-\varepsilon} Q^{\pi}(s, a) \\ &= \sum_{a \in A} \pi(a \mid s) Q^{\pi}(s, a) = V^{\pi}(s). \end{aligned} \tag{5.7}$$

The $Q(\lambda)$ algorithm is an extension of the Q-learning algorithm. In this algorithm, the concept of the eligibility trace should be introduced. The eligibility trace is an additional attribute in each episode, which determines the relative degree between the current state and the update value of the current state.

The idea of the eligibility trace is very simple. When a state-action pair is selected in an episode, a short-term memory (called as a trace) is assigned and the trace does not decrease as time passes by. The meaning of the trace is to determine the size of each state-action pair's eligibility for learning. The eligibility trace can accelerate the learning process.

The update process of the eligibility trace is:

$$e_t(s, a) = I_{s=s_t} * I_{a=a_t} + \begin{cases} \gamma \lambda e_{t-1}(s, a) & \text{if } Q_{t-1}(s_t, a) = \max_{a_t \in A} Q_{t-1}(s_t, a_t) \\ 0 & \text{otherwise.} \end{cases} \tag{5.8}$$

$I_{s=s_t} * I_{a=a_t}$ means that the pair or element in the eligibility trace matrix increases 1 only when the state-action pair is the pair visited currently. We assume the future reward is as important as the present reward, then $\gamma = 1$. The eligibility traces are updated in two different ways. If a greedy-action is executed, that is, the corresponding action of maximum "quality" state is selected, all of the eligibility traces will decay at a parameterλ, and if an exploration action is taken, that is, the action is selected randomly, then the eligibility traces will be set to zero. The update process of quality depends on the value of the eligibility traces.

$$\delta_t = r_{t+1} + \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t), \forall s \in S, a \in A(s) \tag{5.9}$$

$$Q_{t+1}(s, a) \leftarrow Q_t(s, a) + \eta \delta_t e_t(s, a), \forall s \in S, a \in A(s). \tag{5.10}$$

The pseudo-code of the $Q(\lambda)$ algorithm is given in Figure 4. □

Initialize Q matrix, $Q(s,a)=0$;
for $k=1$: MaxEpisode:
  Initialize eligibility trace matrix, $e(s,a)=0$;
  Initialize state S;
  for step:
    choose an action a according to s from Q table with ε -greedy policy;
    take the action a;
    get reward r and the next state s' using Monte Carlo simulation method;
    choose an action a' corresponding to s' from Q table with ε -greedy policy;
    $a^* \leftarrow \arg\max_a Q(s',a)$;
    $\delta \leftarrow r+Q(s',a^*)-Q(s,a)$;
    for all $s_t \in S, a_t \in A(s_t)$;
      $Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\eta * \delta * e(s_t,a_t)$;
      if $a'=a^*$, then $e(s,a) \leftarrow \lambda e(s_t,a_t)$;
      else $e(s,a)=0$;
    $e(s,a) \leftarrow e(s,a)+1$;
    $s \leftarrow s'$;
    until the s is terminal state;
  until the Q matrix has convergence or up to the maximum iteration times.

FIGURE 4. $Q(\lambda)$*algorithm.*

6. Numerical test and analysis

In this section, we use numerical examples to verify the effectiveness of the algorithm proposed in this paper. The pricing problem is solved by python 3.6, and the compiler is Spyder.

Suppose that there are 18 seats that need to be sold in the remaining sale time. The set of prices that can be selected in each time period is $A=\{10,12,14,16,18,20,22,24,26,28,30\}$. Assume the number of passengers is 15. According to the simulation of passenger behaviors in Section 5.1, we generate the passenger arrival sequences as shown in Figure 5, which is obtained based on 100 runs. The sale time is divided into 20 periods according to the arrival sequences. From Figure 5, we can see the purchase behaviors of two different passenger types. Myopic passengers have some kinds of "rush in" behavior in the early stage of the sale period, and strategic passengers have more stable purchase behavior.

Considering pricing policies in various environments, we set α, φ_s and φ_m. The passenger utility functions and strategies are set by past studies [19,39]. The $Q(\lambda)$ algorithm is used to learn the dynamic pricing strategy in the passenger behavior simulation environment. In the simulation, we suppose the proportion is known, that is $\alpha=0.5,\varphi_s=0.5,\varphi_m=0.5$. The arrival intensities of different passenger types are different. The iterative convergence graph of this algorithm is shown in Figure 6. Obviously, the algorithm can converge well. After a certain number of iterations, an effective Q-value table can be obtained, which can be used to get the optimal pricing policy. Using this pricing policy for learning in the simulation environment, we can obtain the expected revenue. Figure 7 shows the average results of 100 runs. From Figure 7 we can see that, when the high-valuation passenger arrival intensity is lower than the low-valuation passenger's intensity in the previous time period but higher than it in the later stage, the price policy is an increasing policy. It can explain that why the high-

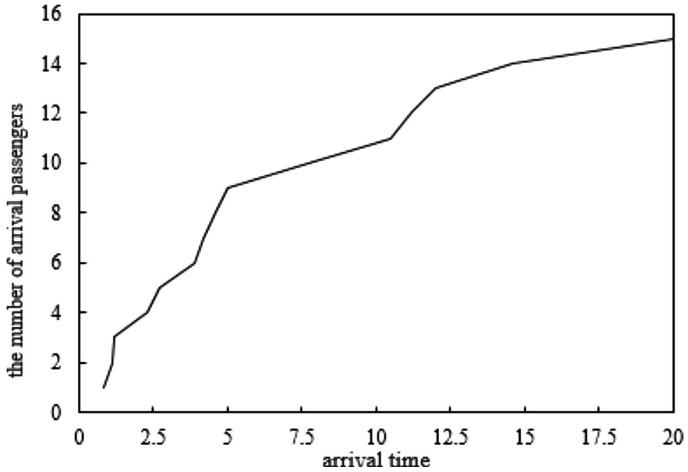

FIGURE 5. Simulated passenger arrival sequences.

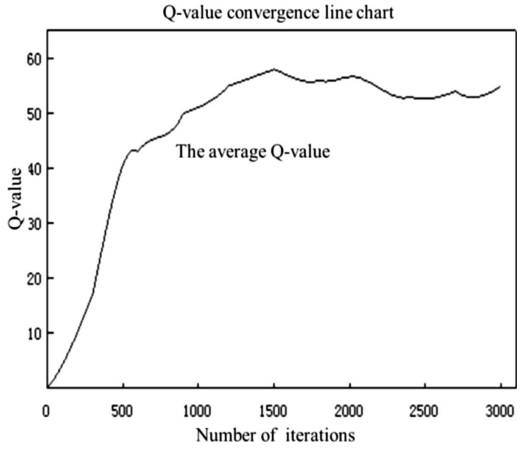

FIGURE 6. $Q(\lambda)$ iterative convergence graph.

valuation passengers have higher arrival intensities in the later periods, and hence the airline chooses to increase the price in the later periods to obtain the surplus value.

6.1. The proportion impact analysis

In this section, we test the effects of different passengers' proportion on the pricing policy.$\alpha = 0$ means all the passengers are myopic passengers, and $\alpha = 1$ means all the passengers are strategic passengers. Assume the strategic level of the same type is the same. $\beta_{HS} = 0.5$, $\beta_{LS} = 0.5$, $\beta_{LM} = \beta_{HM} = 0$, and parameters α, φ_s and φ_m are set differently. Based on 100 runs, we can get the results as Figures 8–10.

From Figures 8–10, we can find different proportions lead to different pricing policies if the strategic level is fixed. The higher proportion of the strategic passengers is, the smaller the price can be increased. Compared Figure 9 with Figure 8, we can find the increase is gentler and much stepwise. Compared Figure 10 with Figure 9, we can find the increase is relatively larger and faster. The higher proportion of the high valuation strategic passengers is, the larger the price can be increased. When the price-cut is adopted, the less continuous-decrease-

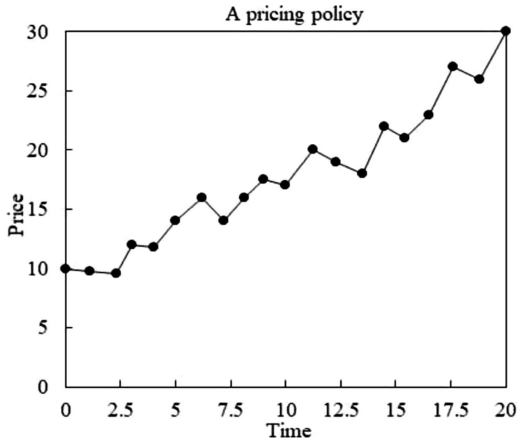

FIGURE 7. $\lambda_{HM,LM} < \lambda_{HS,LS}$.

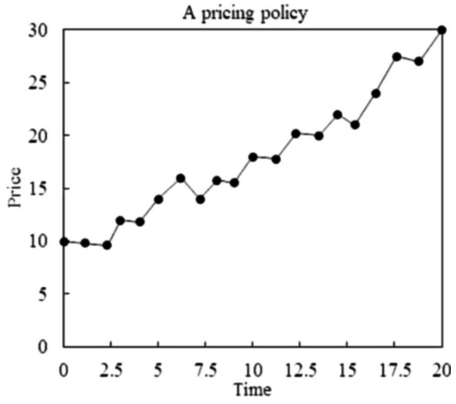

FIGURE 8. $\alpha = 0.4, \varphi_s = 0.4, \varphi_m = 0.5$.

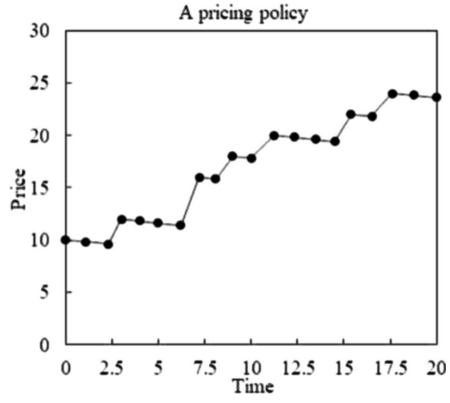

FIGURE 9. $\alpha = 0.6, \varphi_s = 0.4, \varphi_m = 0.5$.

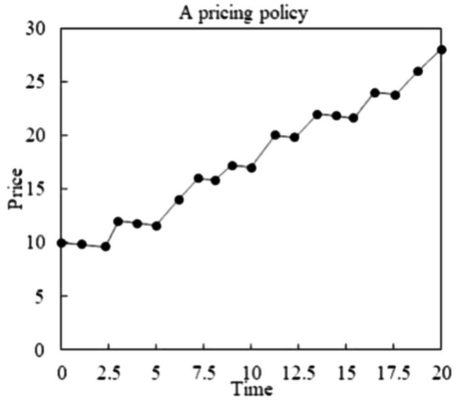

FIGURE 10. $\alpha = 0.6, \varphi_s = 0.6, \varphi_m = 0.5$.

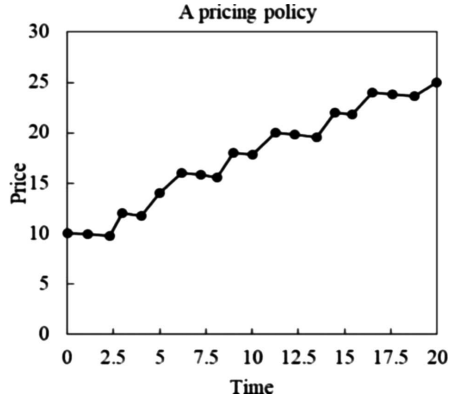

FIGURE 11. $\beta_{HS} = 0.6, \beta_{LS} = 0.4$.

policy is used. This can curb the strategic waiting since the strategic passenger always calculates the expected future utility before he or she makes a decision.

6.2. The strategic level impact analysis

Similarly, we change strategic level to analyze its effects on the pricing policy. Set $\alpha = 0.6, \varphi_s = 0.4, \varphi_m = 0.4$, we change the strategic level from 0.4 to 0.8 and get the average results of 100 runs as is shown from Figures 11–13. Of course, $\beta_{LM} = \beta_{HM} = 0$.

From Figures 11–13, we can find the strategic level also has its impact on the price policy. High-valuation passengers mainly affect high-price periods, and low-valuation passengers mainly affect low-price periods. In high-price periods, the lower the strategic level of the high-valuation passengers is, the larger the slope is in the price policy. Accordingly, in low-price periods, the lower the strategic level of low-valuation passengers is, the larger the slope is in the price policy. The degree of the strategic level affects the degree of the price increase. The higher the level is, the lower the price increases. This is because when the strategic level is higher, the significance of passengers' valuation to the future utility is greater. If the price increases too fast, these strategic passengers choose to leave the market. In addition, as high strategic level passengers dominate, the purchase behavior becomes more complex. The waiting or leaving behavior will generate additional plus or loss in demand, hence cause fluctuation in general price increase trend as is shown in Figure 13.

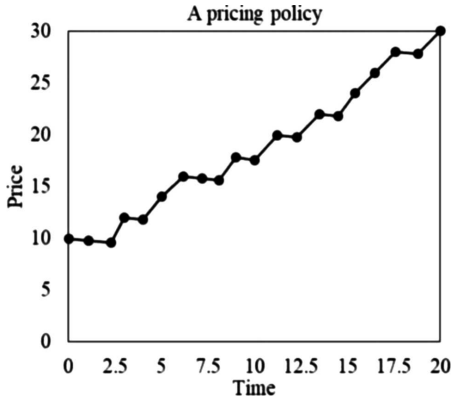

FIGURE 12. $\beta_{HS} = 0.4, \beta_{LS} = 0.4$.

FIGURE 13. $\beta_{HS} = 0.8, \beta_{LS} = 0.6$.

7. Conclusion

This paper studies a dynamic air ticket pricing algorithm for the airline RM in a strategic and myopic passengers' co-existence market, wherein the airline can adaptively decide the ticket price by using RL according to the passengers' arrival intensity, valuation distribution and strategic behaviors. We first formulate the dynamic pricing problem to a finite discrete MDP, and then employ Q-learning to solve the problem. By using RL, we separate the airline's behavior from passengers behaviors through the transformation of the airline's utility model. Passengers' behaviors are simulated as the learning environment. The airline is set as the agent. Through interactions between the environment and the agent by using the trial-and-error approach, the optimal pricing strategy of the airline can be reached. Using this approach, we avoid to directly solve a large-scale dynamic programming problem due to the large number of states and the existence of passengers strategic behaviors.

Based on the various percentage of strategic passengers, high-valuation passengers and different strategic level, a series of computations were conducted. The computational results show the pricing policy relates to the passenger composition. The airline should be more prudential because of the existence of strategic passengers. The higher proportion of the strategic passengers is, the smaller the price can be increased. The price increase should be adopted step by step, the increase quantity should be consistent to the percentage of high-valuation strategic passengers (or low-valuation strategic passengers). The higher percentage of high-valuation strategic

passengers is, the more increase of the price can be made. If the price-cut is taken, the change should be small. High-valuation passengers mainly affect high-price periods or later periods, and low-valuation passengers mainly affect low-price periods or early periods. The strategic level also plays an important role in how the airline can change the price. The lower the strategic level of high-valuation passengers is, the bigger the price increase can be made in the later periods, and the lower the strategic level of low-valuation passengers is, the bigger the price increase can be made in the early periods. These findings are helpful for airlines to set up pricing strategies in different scenarios.

Acknowledgements. This work was supported by Shanghai Planning Foundation of Philosophy and Social Science (No. 2020EGL014), Soft science project of Shanghai "Science and Technology Innovation Action Plan"(21692192600).

References

[1] Z. Bo and L. Meilong, The route efficiency evaluation method based on DEA. *Technol. Mark.* **26** (2019) 10–12.
[2] C. Campbell, MCMC simulation for modelling airline passenger choice behavior. Dublin City University (2001).
[3] J. Cao, Z. Liu and Y. Wu, Learning Dynamic Pricing Rules for Flight Tickets. In: Knowledge Science, Engineering and Management. Springer Publishing (2020).
[4] Y. Chen and V.F. Farias, Robust Dynamic Pricing with Strategic Customers. 16th ACM Conference on Economics & Computation, ACM Publishing (2015).
[5] B.D. Chung, J. Li and T. Yao, Demand learning and dynamic pricing under competition in a state-space framework. *IEEE Trans. Eng. Manag.* **59** (2012) 240–249.
[6] A. Collins and L. Thomas, Comparing reinforcement learning approaches for solving game theoretic models: a dynamic airline pricing game example. *J. Oper. Res. Soc.* **63** (2012) 1165–1173.
[7] A. Collins and L. Thomas, Learning competitive dynamic airline pricing under different customer models. *J. Revenue Pricing Manag.* **12** (2013) 416–430.
[8] J. Correa, R. Montoya, C. Thraves, Contingent Preannounced Pricing Policies with Strategic Consumers. *Oper. Res.* **64** (2016) 251–272.
[9] T.S. Cui, Y.Z. Wang and S. Nazarian, Profit maximization algorithms for utility companies in an oligopolistic energy market with dynamic prices and intelligent users. *AIMS Energy* **4** (2016) 119–135.
[10] P. Du and Q. Chen, Skimming or penetration: optimal pricing of new fashion products in the present of strategic consumers. *Ann. Oper. Res.* **257** (2017) 275–295.
[11] I. Dogan and A.R. Gner, A reinforcement learning approach to competitive ordering and pricing problem. *Expert Syst.* **32** (2013) 39–48.
[12] J. Feng, L. Liu and X. Liu, An optimal policy for joint dynamic price and lead-time quotation. *Oper. Res.* **59** (2011) 1523–1527.
[13] D.P. Gaver, P.A.W. Lewis and G.S. Shedler, Analysis of exception data in a staging hierarchy. *IBM J. Res. Dev.* **18** (1974) 423–435.
[14] J. Gonsch, R. Klein and M. Neugebauer, Dynamic pricing with strategic customers. *J. Bus. Econ.* **83** (2013) 505–549.
[15] Z. Guan and J. Ren, Optimal dynamic pricing policy in the presence of strategic consumers. *Syst. Eng. Theory Pract.* **34** (2014) 2018–2024.
[16] B. Josef and R. Paat, Dynamic pricing under a general parametric choice model. *Oper. Res.* **60** (2012) 965–980.
[17] M. Khouja and X. Liu, A price adjustment policy for maximizing revenue and countering strategic consumer behavior. *Int. J. Prod. Econ.* **236** (2021) 108–116.
[18] M. Kremer, B. Mantin and A. Ovchinnikov, Dynamic pricing in the presence of myopic and strategic consumers: theory and experiment. *Prod. Oper. Manag.* **26** (2017) 116–133.
[19] M. Le and M. Lu, Network revenue management with dependent demand under overlapping segments, In: ICNC-FSKD. IEEE Publishing (2018).
[20] Y. Levin, J. Mcgill and M. Nediak, Dynamic pricing in the presence of strategic consumers and oligopolistic competition. *Informs* **55** (2009) 32–46.
[21] Y. Levin, J. Mcgill and M. Nediak, Optimal dynamic pricing of perishable items by a monopolist facing strategic consumers. *Prod. Oper. Manag.* **19** (2010) 40–60.
[22] T. Levina, Y. Levin and J. Mcgill, Dynamic pricing with online learning and strategic consumers: an application of the aggregating algorithm. *Oper. Res.* **57** (2009) 327–341.
[23] H. Li, Q. Peng and M. Tan, Dynamic pricing strategy for airline tickets through demand learning with strategic passengers. *Oper. Res. Manag. Sci.* **27** (2018) 118–125.
[24] C. Li, M. Chu and C. Zhou, Two-period discount pricing strategies for an e-commerce platform with strategic consumers. *Comput. Ind. Eng.* **147** (2020) 1–13.
[25] Q. Liu and D. Zhang, Dynamic pricing competition with strategic customers under vertical product differentiation. *Manag. Sci.* **59** (2013) 84–101.

[26] R. Lu, S.H. Hong and X. Zhang, A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Appl. Energy* **220** (2018) 220–230.
[27] Q. Ma, F. Meng and X.J. Zeng, Optimal dynamic pricing for smart grid having mixed customers with and without smart meters. *J. Mod. Power Syst. Clean Energy* **6** (2018) 1244–1254.
[28] A.J. Mersereau and D. Zhang, Markdown pricing with unknown fraction of strategic customers. *Manuf. Serv. Oper. Manag.* **14** (2012) 355–370.
[29] C.V.L. Raju, Y. Narahari and K. Ravikumar, Learning dynamic prices in electronic retail markets with customer segmentation. *Ann. Oper. Res.* **143** (2006) 59–75.
[30] R. Rana and F.S. Oliveira, Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. *Omega* **47** (2014) 116–126.
[31] R. Rana and F.S. Oliveira, Dynamic pricing policies for interdependent perishable products or services using reinforcement learning. *Expert Syst. Appl.* **42** (2015) 426–436.
[32] Z.J.M. Shen and X. Su, Customer behavior modeling in revenue management and auctions: a review and new research opportunities. *Prod. Oper. Manag.* **16** (2010) 713–728.
[33] X. Su, Intertemporal Pricing with Strategic Customer Behavior. *Manag. Sci.* **53** (2007) 726–741.
[34] M. Tan, Optimal Pricing for Tickets with Myopic and Strategic Passengers. *Ind. Eng. Manag.* **23** (2018) 107–115.
[35] S. Yousefi, M.P. Moghaddam and V.J. Majd, Optimal real time pricing in an agent-based retail market using a comprehensive demand response model. *Energy* **36** (2011) 5716–5727.
[36] H. Zeng and Y. Zhang, Intertemporal Pricing of Substitutes under the Coexistence of Myopic and Strategic Consumers. *Syst. Eng.* **33** (2015) 33–39.
[37] Y. Zhang and J. Zhang, Strategic customer behavior with disappointment aversion customers and two alleviation policies. *Int. J. Prod. Econ.* **191** (2017) 170–177.
[38] J. Zhao, J. Qiu, and X. Hu, Vertically Differential Product Introduction and Pricing in the Presence of Strategic Consumers. *Syst. Eng. Theory Pract.* **37** (2017) 3098–3108.
[39] B. Zhu and M. Le, The route efficiency evaluation method based on DEA. *Technol. Mark.* **26** (2019) 10–12.