University of Wollongong

Research Online

SMART Infrastructure Facility - Papers

Faculty of Engineering and Information Sciences

2010

Intelligent negotiation behaviour model for an open railway access market

S K. Wong
Hong Kong Polytechnic University

T K. Ho
Hong Kong Polytechnic University, markho@uow.edu.au

Follow this and additional works at: https://ro.uow.edu.au/smartpapers

Part of the Engineering Commons, and the Physical Sciences and Mathematics Commons

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Intelligent negotiation behaviour model for an open railway access market

Abstract

In an open railway access market, the provisions of railway infrastructures and train services are separated and independent. Negotiations between the track owner and train service providers are thus required for the allocation of the track capacity and the formulation of the services timetables, in which each party, i.e. a stakeholder, exhibits intelligence from the previous negotiation experience to obtain the favourable terms and conditions for the track access. In order to analyse the realistic interacting behaviour among the stakeholders in the open railway access market schedule negotiations, intelligent learning capability should be included in the behaviour modelling. This paper presents a reinforcement learning approach on modelling the intelligent negotiation behaviour. The effectiveness of incorporating learning capability in the stakeholder negotiation behaviour is then demonstrated through simulation.

Keywords

model, behaviour, negotiation, open, intelligent, market, railway, access

Disciplines

Engineering | Physical Sciences and Mathematics

Publication Details

Wong, S. K. & Ho, T. K. (2010). Intelligent negotiation behaviour model for an open railway access market. Expert Systems with Applications, 37 (12), 8109-8118.

This journal article is available at Research Online: https://ro.uow.edu.au/smartpapers/62

Intelligent Negotiation Behaviour Model for An Open Railway Access Market

S.K. Wong and T.K. Ho*
Department of Electrical Engineering
Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
* Corresponding author, email: eetkho@polyu.edu.hk; tel: +852 2766 6146; fax: +852 2330 1544

Abstract

In an open railway access market, the provisions of railway infrastructures and train services are separated and independent. Negotiations between the track owner and train service providers are thus required for the allocation of the track capacity and the formulation of the services timetables, in which each party, i.e. a stakeholder, exhibits intelligence from the previous negotiation experience to obtain the favourable terms and conditions for the track access. In order to analyse the realistic interacting behaviour among the stakeholders in the open railway access market schedule negotiations, intelligent learning capability should be included in the behaviour modelling. This paper presents a reinforcement learning approach on modelling the intelligent negotiation behaviour. The effectiveness of incorporating learning capability in the stakeholder negotiation behaviour is then demonstrated through simulation.

Keywords

Reinforcement learning; Open railway access markets; Negotiation behaviour; Intelligent transportation system

1. Introduction

For two centuries, railways have been the backbones for public transportation and pivotal to urban development. Because of the substantial capital costs on the infrastructure, the railway service provision may easily become natural monopoly. As a result, the train services are generally non-competitive and they require heavy regulations (ECMT, 2001; Cantos & Campos, 2005). In order to introduce competition and improve efficiency, the railway industries in many countries have undergone substantial reform in the last two decades. The market structure has been revamped with the introduction of open railway access market (ABTRE, 2003). Independent service providers are allowed to enter the railway market to create competition in an attempt to improve service quality. An open railway access market usually consists of an Infrastructure Provider (IP) and a group of Train Service Providers (TSPs). The IP is responsible for providing infrastructure resources, negotiating prices and selling track capacity to the TSPs. The TSPs operate train services and hence they are required to negotiate with the IP to obtain the rights of track access. As the track recourses are limited, TSPs may compete among themselves for the track access.

This new reform poses a new challenge to the railway management. In the traditional railway services market, the infrastructure and train service provisions are managed by a single company which has the overall and integrated responsibility for devising train service timetables while satisfying various operational objectives. The trade-off between objectives is relatively easy to achieve. For example, the railway operator is able to adjust with flexibility the number of train services to balance between cost recovery (i.e. revenue intake) and traffic demand (i.e. maintaining the services of low or even no profit). However, in the new market environment, the stakeholders are distributed and self-interested individuals and they have different business goals (Pietrantonio & Pelkmans, 2004). Their objectives are likely in conflict. A TSP may pull out from train service provisions which are low in profit despite public demand while the IP is keen on maximising the infrastructure utilisation for the interest of the public. Therefore, repetitive negotiations and concessions among the stakeholders to ensure reasonable service provision are required. Allocating the resource efficiently becomes a demanding job for the IP as the behaviour of the stakeholders during negotiation is difficult to predict.

In the analysis of the quality of service provision in an open railway market, the traditional post-operation evaluation techniques, such as statistical analysis and case studies, are not particularly useful as the stakeholder behaviour model, which is unique in a given market configuration, is not included. Moreover, the traditional techniques do not allow the study on the performance of different possible market configurations since the cost or risk of experimenting with those parameters in a real market is too high. The above issues are indeed the limitations in the traditional methods for the purpose of analysing stakeholder behaviour.

With the new open market and complicated negotiation processes among the stakeholders, it is necessary to evaluate the resulting resource allocation and timetables, and the quality of services and infrastructure capacity utilisation. Simulation is regarded as the most conceivable approach to address the above problems because it provides the systematic ways to evaluate and analyse the behaviour of different stakeholders. The change of the stakeholder behaviour and under different market configurations can be studied by adjusted the behaviour models flexibly. However, the traditional simulation modelling techniques, such as rule-based approaches, recursive and iterative algorithms, are not adequate to model the interactive market environment and self-interested stakeholder behaviour as they are too complicated to be pre-defined and approximated. In a previous study, a multi-agent system is employed for the open railway access market (Tsang, 2006; Tsang & Ho, 2007; Tsang & Ho, 2008). Multi-agent system is a powerful tool to represent the behaviour and interactions of the stakeholders which are distributed in nature (Shen et al., 2006). The complex and interactive behaviour among stakeholders has been demonstrated even when only a simple model of the individual stakeholder behaviour is adopted. However, the stakeholder behaviour modelling in this study is not yet adequate as the agents lack the learning capability that their human counterparts possess. Introducing learning capability allows agents to demonstrate goal-oriented behaviour and enhance the problem-solving ability. In this study, the learning capability is incorporated in the stakeholder model through reinforcement learning.

Reinforcement learning techniques focus on discovering the optimal actions in a specific environment so that the long-term reward is maximised. They enable an agent (or a learner) to learn the optimal policy that maps directly from environment states to actions and to achieve the goal of the agent (Kaelbling et al., 1996). Reinforcement learning allows more realistic simulation studies on the stakeholder interactions and their negotiations in an open railway access market.

This paper demonstrates the application of reinforcement learning techniques to enhance stakeholder behaviour modelling for negotiation simulation in open railway access market. Section 2 describes the reinforcement learning framework in an open market. Section 3 gives the negotiation models for IP and TSP. Section 4 introduces the reinforcement learning algorithm adopted and Section 5 presents the simulation results of the case studies. Conclusions are then made in Section 6.

2. Reinforcement learning framework

It has been shown that the Multi-Agent System (MAS) approach is suitable for modelling the independent, self-interested and autonomous features of the stakeholders in an open railway access market (Tsang, 2007). MAS manifests the complex interactions among independent entities and their behaviour by modelling the individuals as agents. However, this simple agent model adopted is very much on a rule-based approach and hence not sufficient to demonstrate the more complex negotiation behaviour that occurs in real-life situations as it lacks the intelligence for learning and bargaining. For example, the strategies for the IP to determine the charge for track access; and for the TSP to set service proposal, are pre-defined prior to the negotiation, but they should have been revised intelligently according to the experiences from the previous rounds of negotiations. In order to facilitate a more realistic behaviour model, intelligent learning should be incorporated in the agent modelling so that the stakeholder agents in an open railway access market are able to select the most suitable strategies among the available options from their own experiences.

Although game theory is one of the possible approaches to model and govern the agent's negotiation behaviour (Binmore & Vulkan, 1997), it is not applicable in this study because the complete game model (i.e. payoff and penalty of each strategy) is not available. The agent should learn the reward or penalty of each strategy before solving the game.

Machine learning is one of the topical areas of artificial intelligence, which allows computers to learn. Supervised learning, unsupervised learning and reinforcement learning are the three common types of machine learning. With supervised learning, a function is generated to map input objects to desired outputs from a set of training data (Zhao & Liu, 2007). A typical application of supervised learning is classification, in which the input objects are classified into different categories according to specific criteria. Unsupervised learning focuses on capturing the inherent data organisation structures (Zhao & Liu, 2007). Clustering is an example of unsupervised learning problems, in which a hierarchical structure is sorted and established with a given set of unlabelled data. Reinforcement learning enables the process of learning a policy on taking an action in an environment in order to maximise the long-term reward (Sutton & Barto, 1998) and it has been applied in multi-agent systems as a learning algorithm (Crites & Barto, 1998; Suematsu & Hayashi, 2002; Oo et al., 2002). Reinforcement learning algorithms have been successfully deployed in various decision-making processes, such as setting prices in a competitive marketplace (Tesauro & Kephart, 2002).

The basic reinforcement learning consists of three main components. They are environment states, action sets and reward (or value) functions. An advantage of reinforcement learning is that pre-defined training sets are not required (Ribeiro, 1999). The learning is achieved through real-time interactions with the environment. At each interaction with the environment, a learner first perceives the environment states and the available actions and then selects an action which is based on the current action-taking policy, the environment states and the value functions.

Subsequently, the learner revises the action-taking policy and the value functions by the observed new environment states and the immediate reward. Having experienced these interactions, the learner is able to learn the action-taking policy that leads to higher long-term reward according to the pre-defined value function.

During stakeholder negotiation in an open railway market, pre-defined actions and reward data sets for training are usually unavailable. The stakeholder can only gain knowledge and experience of achieving goals through the interactions with other stakeholders in the environment of open railway access market. The learning objective focuses on selecting suitable action rather than clustering different actions to different objectives. Therefore, reinforcement learning is the most promising learning algorithm for stakeholder negotiation modelling in an open railway access market. In this study, reinforcement learning is introduced as the learning algorithm to the stakeholder negotiation behaviour model.

The reinforcement learning approach is integrated into the negotiation behaviour model by mapping stakeholders and their behaviour to the reinforcement learning framework, in which the individual stakeholders are the learners. The stakeholders should be able to learn and choose the suitable action or strategy which satisfies their own objectives. The environment states consist of the parameters that are to be changed by the decisions the stakeholders make. The action sets describe the negotiation strategies available, such as price adjustment. Reward function indicates how well the stakeholder objectives are achieved.

For example, the IP determines which pricing strategies or actions (increasing, decreasing or keeping the prices) should be applied during the negotiation with different TSPs such that the overall revenue intake can be further improved. The decision-making variables include the degree of pricing level that a train service is currently charged, and the features of the train services that are favourable to the TSP. The reward function records the revenue improvement resulting from applying pricing strategies under certain conditions. Upon completing the learning, the IP is equipped with the best pricing strategy with respect to the reward function.

Fig. 1 shows the proposed intelligent stakeholder negotiation behaviour models.

Maximising profit is likely to be the common objective of IP and TSP. This paper will present a negotiation problem that is intended solely on maximising the profit for both IP and TSP. In practice, IP and TSP may carry multiple and different objectives.

3. Open market negotiation models

During each negotiation between IP and TSPs to allocate the track capacity and generating train services timetables in an open railway markets, the IP and one of the TSPs attempt to reach a deal on the track access right which specifies the conditions and costs for track usage. The negotiation is an iterative process in which the two stakeholders take turns to express their requirements or restrictions on the track usage. Since certain terms or conditions of the timetable requested by a stakeholder are not likely to be favourable to another, the negotiation inevitably contains a sequence of bargaining and/or concession made by both parties. The negotiation is completed until a mutually acceptable agreement is reached, or one of parties withdraws from the process. This section discusses the negotiation models, procedures and criteria of reaching a train service timetable.

3.1 Track access rights

The track access rights consists of four components, a track access charge (TAC) that TSP has agreed to pay; the type of rolling stock to be operated on tracks; a parameter called flex level to denote the right for IP to revise the train schedule when track and station capacity become scarce (Gibson et al., 2002); and a time schedule specifying the details of the train service. A track access rights R is defined below, where c is the track access charge, ω is the type of rolling stock, Φ is the flex level and Ψ is the train schedule.

$$R = < c, \omega, \Phi, \Psi > \qquad (1)$$

The train schedule Ψ is a set of stations, S, describing the sequence of the train stations the train service goes through. The service commencement time, dwell time at each station and inter-station runtimes between stations are given by ζ, T_D and T_R respectively.

$$\Psi = < S, \zeta, T_D, T_R > \qquad (2)$$

3.2 Negotiation procedures

One approach to classify negotiations is by the number of parties involved (Luo, 2003). The negotiation is regarded as bilateral when there are only two participating parties. A simple case is that a single IP negotiates with a TSP (i.e. IP-TSP negotiation). When more than one TSPs are involved, the negotiation becomes multilateral (i.e. IP-nTSP negotiation).

The negotiation is initialised by the IP who requests all TSPs to submit services bids of their preferred train schedules. After collecting all the bids, the IP derives a feasible service timetable by taking into consideration of all constraints imposed by the TSPs (i.e. combinatorial timetable generation) or negotiating with the TSPs one by one according to a specific sequence and then deriving the train schedules (i.e. sequential timetable generation).

Sequential negotiation is further classified into transaction-based and round-based. In transaction-based negotiation, the IP proposes offers to other TSP's only after completing the negotiation and conflict-resolving process with the current TSP (i.e. upon completion of a transaction). On the other hand, with round-based negotiation, the IP temporarily adjourns the negotiation with the current TSP and proposes an offer to another TSP immediately if there is a conflict between the current TSP's proposed schedule and those committed by the IP so far. The IP then determines another negotiation sequence and starts negotiations with the other TSPs again. The optimisation problem in combinatorial timetable generation clearly requires higher computation demand since the problem scale is substantially larger. Sequential timetable generation is however a more practical approach since it only considers one IP-TSP transaction (i.e. a bilateral negotiation) at a time. Furthermore, the transaction-based negotiation only requires one negotiation-sequence and it is thus adopted in this study. The negotiation sequence can be derived from simple rules, such as first-come, first-served (Tsang, 2007), or intelligent ranking of the service bids from the TSPs (Ho, et al., 2009).

With the negotiation sequence, the IP evaluates the bid content and proposes a train service schedule to each TSP; or requests the TSP to revise the bid specification if necessary. In bid revision, the TSP first determines the possibility of revising the bid and then re-submits a bid to IP if a revision is possible. Otherwise, the TSP withdraws from negotiation. An IP-TSP transaction is only completed if TSP satisfies with the train service schedule or withdraws from negotiation. Upon the completion of the IP-TSP transaction, IP starts a new negotiation process and negotiates with the next TSP in the sequence. This IP-nTSP negotiation continues until all the IP-TSP transactions are completed. The procedures of an IP-nTSP transaction are illustrated in Fig. 2.

3.3 Acceptance criteria

In this study, the objective of the IP and TSPs is to maximise profit. The following timetable acceptance criteria are adopted to satisfy this objective.

3.3.1 IP

IP collects revenue by selling the track capacity to the TSPs. The revenue intake is determined by the track access charge (TAC). The profit is thus the difference between the TAC and the expenses on setting up the train services. In general, the cost for the IP to run an additional service (i.e. marginal cost) is complex and it depends on a number of factors, such as energy consumption, maintenance and traffic demands. The marginal cost is derived from the sub-charges on track usage, traction energy, peak power demand and traffic congestion (Tsang, 2007). IP accepts a train service timetable only if the TAC is higher than the marginal cost and the timetable is free from conflict.

3.3.2 TSP

The revenue intake of TSP is determined by the service charges and service usage. The demand of the TSP train services depends on several factors, such as service quality, price elasticity, seasons and market share. A simple model has been established to calculate the revenue intake of the TSP (Tsang, 2007), in which the TSP is assumed to operate the passenger services only and the revenue intake is determined directly by the passenger traffic demand.

The daily passenger traffic demand is required to determine the usage of train services. Several factors, such as population, employment, trip rates, ticket costs and transportation time, are considered. A simple daily passenger traffic demand model *D* is adopted here.

$$D(t) = C_d f(t) \tag{3}$$

f(t) describes the variation of traffic demand in one day. C_d is set to 1 if the TSP operates a train service which satisfies the requirements of the passengers, such as reasonable ticket prices and expected commuting times, or C_d is set to 0. C_d thus imposes an additional constraint on the TSP's train service provision. In real-life operations, there are several more accurate but complex models to determine the passenger demand (Chen, 2007; Li et al., 2002; Thamizh et al., 1997). This simple model is adequate to allow demonstration of the proposed negotiation behaviour model. The revenue intake $R(t)$ of the TSP is then obtained by multiplying the ticket cost C_T to the passenger traffic demand.

$$R(t) = C_T D(t) \tag{4}$$

A TSP is willing to operate train services only if the services are profitable, which implies that the TSP is satisfied with the train service timetable. In other words, the degree of satisfaction is measured by the deviation of the original TSP's proposed train schedule from IP's proposed schedule. The TSP is willing to pay higher TAC if the IP can provide the TSP's preferred service timetable. Two decision thresholds, TSP_AC (HS) and TSP_AC (LS), are introduced to indicate the TSP's maximum acceptable TAC, corresponding to the highest and lowest satisfaction levels of the schedule. The actual value of TSP acceptance cost (TSP_AC) is determined from the timetable satisfaction level of the train schedule which has been proposed by the IP. The TSP accepts the track access right only if the proposed TAC is lower than the TSP acceptance cost. Fig. 3 shows a typical example of revenue intake and TSP acceptance cost over a period of starting time of the train service.

4. Reinforcement learning framework

Reinforcement learning in a multi-agent system provides agents with the intelligence capabilities to transform the experience of interacting with other agents to the knowledge of achieving their goals. The reinforcement learning mechanism can be regarded as a generalised policy iteration (GPI) (Sutton & Barto, 1998), which includes two processes, policy evaluation and policy improvement; and they are interleaved. In the policy evaluation process, the value function is reviewed and updated according to the environment states, the selected action and the observed reward. This process enables learning the reward of selecting an action in a given environment state. The policy improvement process is to summarise the knowledge of the value function and to learn the greedy policy (i.e. the policy which always takes the action with the highest short-term reward). Reinforcement learning has found numerous successful applications over various areas in recent years (Ernst, 2004; Wang & Usher, 2005; Jiang & Sheng, 2009; Gomez-Perez et al., 2009)

4.1 Computational algorithm

4.1.1 Q-Learning

Q-Learning is a computational algorithm to solve reinforcement learning problem (Watkins, 1989). One of the advantages of Q-Learning is that no actual environment models are required. An agent learns the optimal action by determining the optimal action-value function (Q-function).

The Q-function is the expected return of a given state and action under certain policy π:

$$Q^{\pi}(s,a) = E\{\sum_{j=0}^{\infty} \gamma^{j} r_{i+j+1} \mid s_i = s, a_i = a, \pi\} \qquad (5)$$

where $s \in S$, and s is environment state of the finite environment set S; $a \in A$ and A is the action set, r is the reward and γ is the discount factor of future rewards.

The simplest form of one-step Q-Learning is defined by (6). The Q-function is updated by the immediate reward r, the selected action a, the greedy Q-value in the next state $Q^{\pi}(s',a')$, the discount factor γ and the learning rate α.

$$Q^{\pi}(s,a) \leftarrow Q^{\pi}(s,a) + \alpha[r + \gamma \max_{a'} Q^{\pi}(s',a') - Q^{\pi}(s,a)] \qquad (6)$$

4.1.2 Exploration vs exploitation

One important issue in reinforcement learning is the balance between exploration and exploitation. Adequate exploration action helps the agent to discover the alternative non-greedy actions which yields better result over time. In this study, the soft-max selection function (Buşoniu et al., 2008) is applied to ensure sufficient exploring actions.

$$P(s,a) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a_i} \exp(Q(s,a_i)/\tau)} \qquad (7)$$

The soft-max selection provides a simple method to balance the exploration and exploitation actions by allowing the highest action selection probability for the highest Q-value action while keeping lower Q-value actions with lower probability. $P(s,a)$ denotes the probability of selecting action a in state s. The positive parameter τ is the temperature which controls the randomness of exploration. High temperature leads to more arbitrary action selections and hence higher exploration rate. Low temperature results in a wider variation in action selection probability so that the higher-valued actions are more likely to be selected. When the temperature is zero, the function output is equivalent to the greedy action selection.

4.1.3 Q-Planning

Q-Planning is one of the methodologies for an agent to predict how the environment responds to its actions by modelling the environment (Sutton & Barto, 1998). Given a certain state and policy, the model generates a simulated experience which contains the information of the observed state; and the reward which is used to update the value function and policy. The learning can be hastened by providing more such simulated experience.

4.2 Reinforcement learning model for open railway market

4.2.1 IP model

Environment Set: The strategy for the IP to maximise profit is that the IP's offered TAC should be the same as the TSP's maximum acceptable TAC which is highly dependent on the business goal of individual TSP and its proposed train service timetable. Therefore, the environment set should include the factors to evaluate the TSP's maximum acceptable TAC. The proposed train service start-time is a suitable candidate for determining the maximum TAC as it is based on the proposed TSP traffic demand and revenue. The environment set should also include the price charging levels (i.e. the offered TAC) of different train services. A simple definition of the charging level is based on the IP's marginal cost (MC).

$$Charging\ Level\ = \frac{TAC - MC}{TAC} * 100\% \qquad (8)$$

In this study, the environment set of the IP is determined by the combination of individual charging levels and defined as $S_{IP} : (c_1, \ldots, c_i, \ldots, c_n)$, where c_i is the charging level for the i^{th} TSP and n is the total number of TSPs. The value of c_i is given by $f_i(t_{start-time}) \in \{0,$ $1,$ $2, \ldots, c_{\max}\}$, where $c_{\max}$ is the maximum charging level which is set as 2*MC. $f_i(t_{start-time})$ is a function of the i^{th} TSP's proposed train service time and $t_{start-time} \in \{0, 1, 2, \ldots, t_{\max}\}$. $t_{\max}$ denotes the last possible service start-time permissible in the timetable. For example, $S_{IP} : (f_1(0) = 0,\ f_2(3) = 2\ ,\ f_3(5) = 1)$ means the IP is charging the three TSPs 1, 2 and 3 at levels "0", "2" and "1" respectively. The proposed train service start-times of the TSPs 1, 2 and 3 are “0”, “3” and “5”. The charging levels and train service start-times are set as follows.

Charging level “0” = the price is set at marginal cost.

Charging level “1” = the price is set at marginal cost +5%.

Charging level “2” = the price is set at marginal cost +10%; and so on.

Train service start-time “0” = service start-time set at the earliest start-time.

Train service start-time “1” = service start-time set at the earliest start-time +5 mins.

Train service start-time “2” = service start-time set at the earliest start-time +10 mins; and so on.

Action Set: The action set *A* of the IP to be taken on the *i*th TSP is defined by (9), where A_{in} is increasing the charging level by one step, A_{un} is keeping the charging level unchanged and A_{de} is decreasing the charging level by one step.

$$
\begin{aligned}
&A{:}(A_{in}, A_{un}) && \text{for } c_i = 0 && (9)\\
&A{:}(A_{un}, A_{de}) && \text{for } c_i = c_{\max} &&\\
&A{:}(A_{in}, A_{un}, A_{de}) && \text{otherwise} &&
\end{aligned}
$$

For example, the initial charging levels on the services proposed by TSPs 1, 2 and 3 are “0”, “2” and “1” respectively, i.e. S_{IP}:(0, 2, 1). If the IP is taking action A_{in} to the TSP-1, A_{un} to the TSP-2 and A_{de} to the TSP-3, the charging levels of TSPs 1, 2 and 3 become 1, 2 and 0 and the state of IP is S_{IP}:(1, 2, 0).

Reward Function: The objective of this learning process is to set suitable TAC in order to maximise the profit in the negotiation with the TSP. IP’s profit is the difference between the TAC and the marginal cost required to provide the track access of the agreed service. The reward function indicates the possible improvement in profit when the learning process goes through the environment states of the IP. The IP should propose a reasonable TAC, or it loses the goodwill in further negotiation with the TSP. A penalty factor *p* = max(0, TAC–TSP_AC) is introduced in the reward function to ensure proper negotiation behaviour.

$$r = \text{Profit (observed state) - Profit (previous state) - } p \qquad (10)$$

4.2.2 TSP model

Environment Set: In order to maximise profit, TSP should actively respond to traffic demand. From the traffic demand model derived in the previous section, a TSP should set up a train service at the peak hours to maximise the revenue intake. Therefore, the train service start-time should be included in the environment set. The environment set of the i^{th} TSP is determined by the service start-time and defined as $S_{TSPi} \in \{0, 1 \ldots, t_{\max}\}$, where $t_{\max}$ is last possible service start-time permissible in the timetable. When S_{TSP1}=3, TSP-1 is proposing a train service schedule to IP, in which the service start-time is "3". The train service start-times are interpreted as follows:

Train service start-time "0" = service start-time set at the earliest time.

Train service start-time "1" = service start-time set at the earliest time +5 mins.

Train service start-time "2" = service start-time set at the earliest time +10 min, and so on.

It should be noted that the notations of train service start-times in the environment sets of IP and TSP can be different. Individual agents in a MAS are allowed to have their own models.

Action Set: The action set A of the i^{th} TSP is defined below, where T_{de} is delaying the service start-time by one step, T_{un} is keeping the service start-time unchanged and T_{ad} is advancing the service start-time by one step.

$$A{:}(T_{de}, T_{un}) \quad \text{for } S_i = 0 \qquad (11)$$

$$A{:}(T_{un}, T_{ad}) \quad \text{for } S_i = t_{\max}$$

$$A{:}(T_{de}, T_{un}, T_{ad}) \quad \text{otherwise}$$

For example, if TSP-1 initially sets up a train service with start-time "3", i.e. S_{TSP1}=3, and subsequently it adjusts the service start-time by taking action T_{de}, the service start-time of TSP-1 becomes "4" and hence it is in the state S_{TSP1}=4.

Reward Function: The objective of this learning process is to set suitable service start-time to maximise the profit in the negotiation with the IP. TSP's profit the difference between its income, as derived by (4), and the TAC paid for the agreed service. The reward function r

depicts the possible improvement on profit when the learning process is taken through the environment states of the TSP.

$$r = \text{Profit (observed state) - Profit (previous state)} \qquad (12)$$

5. Simulation results and discussions

A Multi-Agent System for Open Railway Access Markets (MAS-ORAM) developed in a previous study is adopted here to facilitate the stakeholder negotiations and to evaluate the agent behaviour (Tsang, 2007). MAS-ORAM is implemented through the platform of JADE (Java Agent Development Framework) which allows easy realisation of fully FIPA-compliant multi-agent systems (Bellifemine et al., 2003). MAS-ORAM acts as a platform for the agents to interact and negotiate with others for resource allocation and price setting; and also to demonstrate the intellectual capabilities of the agent induced by reinforcement learning.

A hypothetical open railway access market has been set up to demonstrate the effectiveness of reinforcement learning. This market consists of one IP and five TSPs while the railway services cover five stations (A, B, C, D and E). The details of the TSPs and the inter-station distances are given in Tables 1 and 2. The earliest start-time (i.e. train service start-time interval ID="0") is 06:00 and each time-step is a 5-min interval. Hence, the initial environment states of the TSPs are S_{TSP1}=1 (6:05), S_{TSP2}=3 (6:15), S_{TSP3}=7 (6:35), S_{TSP4}=9 (6:45) and S_{TSP5}=5 (6:25).

In this study, three learning scenarios have been implemented to demonstrate how reinforcement learning enhances the stakeholders' intellectual capabilities. They are individual IP learning only; TSP learning only; and both IP and TSP learning. In each scenario, there are two phases, training and testing phases. In the training phase, the stakeholder negotiates and interacts with others to acquire the knowledge for achieving their objectives. A substantial number of training rounds are given before performance evaluation is conducted (i.e. testing phase). The actual number of training rounds required is determined by observing the agent behaviour or the greedy action policy (i.e. the policy which always takes the action with the highest short-term reward). The training phase is completed if the greedy policy is unchanged after 15 consecutive rounds and 90% of the environment states have been visited. In each training round, the IP agent negotiates with 5 TSP agents and generates a complete train service schedule (completing one round of IP-nTSP negotiation). Upon the commencement of each new training round, the IP agent negotiates with 5 TSP agents again with the preserved knowledge and state values (i.e. the value functions, state values and action-taking policy are identical to those at the end of the previous round). However, the state values are arbitrarily reset after completing a fixed number of training rounds (say, 40 training rounds). It is to allow the agents to explore more environment states and actions.

In the testing phase, there are no arbitrary state values resetting and the agents always take the greedy action. Before conducting the testing phase, the environment states of the agents are reset to the initial values (i.e. S_{IP} =(0, 0, 0, 0, 0); and S_{TSP1}=1, S_{TSP2}=3, S_{TSP3}=7, S_{TSP4}=9 and S_{TSP5}=5) while the value functions and action-taking policies remain. At the end of the testing phase, the learning is deemed effective if the agent is able to take the appropriate actions and achieve its objective of maximising profit.

5.1 IP learning

The reinforcement learning is applied to IP only. 280 training rounds are completed in the training phase. Table 4 summaries the IP's revenue intake upon the agreed services with the TSPs at the end of the testing phase, the calculated marginal cost of the services and hence the profit. Having possessed no knowledge on TSPs initially, the IP is able to set TACs which are close to the TSPs maximum acceptance cost so that its own profit is maximised in the agreed services. As further illustrated in Fig. 4, the revenue intake and the profit attained by the IP are improved through the testing rounds as it gradually learns to take the suitable actions to maximise its profit.

5.2 Individual TSP learning

240 training rounds are taken in the training phase for the individual TSP learning, in which reinforcement learning is given to TSP-1 only. To highlight the effect of learning, the service retained by TSP-1 and hence the corresponding revenue intake with and without learning are compared in Table 4. TSP-1, equipped with knowledge from learning, improves its revenue from $4,800 to $5,400 by changing the service start-time at the end of the testing phase. Fig. 6 illustrates the TSP-1's revenue and profit through the testing rounds and it is evident that TSP-1 takes appropriate actions swiftly and settles on a better profit in the testing phase.

5.3 IP and TSP learning

Learning is adopted by both IP and TSPs here and 680 training rounds are taken. Table 5 summarises the IP's estimated TAC limits of the TSPs at the end of the testing phase. Figs. 6 - 10 show the IP's performance on the negotiation with each of the 5 TSPs, including its marginal cost (MC), profit and estimated acceptance cost of the TSP over the possible the range of service start-times. The highest TSP acceptance cost limits, i.e. TSP_AC(HS) are also given for comparison. In general, the IP is able to set TAC close to the TSP_AC limit. The cases where the IP's estimated TSP acceptance cost is not close to the TSP limit are mostly on service start-times with lower revenue intake. The corresponding states are thus less attractive with lower reward values and hence they are visited less frequently during the learning process. Therefore, the IP does not have the sufficient experience to determine the maximum schedule cost at these service start-times. On the other hand, when these service start-times do not give high revenue intake, they usually favour neither IP nor TSP in their pursuit of maximum profit. As a result, it is not necessary for the IP to give good estimate of the TSP's acceptance cost over such service start-times.

For the performance of the TSPs upon learning, the service start-times proposed by the TSPs in the testing phase (represented by Time Interval ID) are given in Fig. 11 while Figs. 12 and 13 show the revenue intakes and profits of TSPs in the testing phase. Table 6 summarises revenue intakes of the TSPs on the agreed services at the end of the testing phase.

The TSPs are able to identify the states leading to the maximum revenue and to revise the service start-times accordingly, which is reflected by their increasing profit and changes of service-start-times during the testing phase. From Table 6, it is interesting to note that TSP-3 settles on the service start-time interval ID “9” rather than the time interval ID “7” which gives a higher revenue intake. It is due to the fact that the IP TAC setting at the time interval ID “9” is substantially lower than that at “7”. As a result, TSP-3 can make more profit by operating the service at time interval ID “9” than that at “7”. This example depicts the interactive behaviour of the IP and TSPs during negotiation as they look for the possible improvement of the reward through the experience of interaction. As a whole, the simulation results suggest that the reinforcement learning significantly helps both IP and TSPs to maximise profit.

6. Conclusions

The application of reinforcement learning in intelligent behaviour modelling on the stakeholders for the open access market train-service schedule negotiation is presented. The reinforcement learning models for the stakeholders as a learning mechanism in the realistic negotiation behaviour have been discussed. A simple railway open market with stakeholders equipped with reinforcement learning capability is set up and the simulation results suggest that the learning algorithm enhances the problem-solving abilities of the stakeholders with respect to achieving their goals. They have gained the experience and learnt how to take the most suitable actions to achieve the goals through interactions in the environment. The IP is able to determine the suitable track access charge and hence maximise the profit while the TSPs improve the profit by adjusting the train service start-time.

This study provides the platform for further development on stakeholder behaviour modelling to facilitate more realistic simulation of the railway open markets. The TSPs compete with others through negotiations with the IP in order to optimise their own objectives. However, in a railway market, cooperation among TSPs to attain a mutually favourable train service is possible. Through cooperation, a TSP is able to eliminate or reduce the conflicts with other TSP's proposed train schedules by negotiating with them directly. Therefore, those proposed train schedules are more likely to be accepted by the IP because of better track utilisation (from the IP perspective). As a result, the TSPs benefit from the cooperation as their objectives are fulfilled. This cooperative negotiation behaviour can also be achieved by the introduction of reinforcement learning. For example, the environment states of a TSP include the identity of the collaborators and their credit records. The TSP then accepts or rejects the schedule adjustment request according to the credit record and the margin of schedule adjustment. A positive reward is obtained if other TSPs also demonstrate the cooperative behaviour or vice versa. Through the interactions with others, the TSP learns to determine the most suitable strategy in the cooperative negotiations (i.e. accepting or rejecting the schedule adjustment request) and the corresponding concessions.

While this study does not apply the game theory techniques to the schedule negotiation, game theory can be integrated in the reinforcement learning framework. It is possible to employ reinforcement learning techniques to explore actions and obtain the action rewards in an interactive environment, and to adopt game theory techniques to solve the negotiation (game). In other words, reinforcement learning techniques extend the applicability of game theory in the negotiation behaviour modelling.

In addition, the reinforcement learning knowledge has been stored and represented in tabular format in this study. This representation is not the most efficient as the reuse of the knowledge is not particularly flexible. Further studies on the improvement of knowledge management are necessary for more realistic behaviour models.

Acknowledgments

The authors would like to express their sincere gratitude to the Hong Kong Polytechnic University for its financial support on this work.

References

Australian Bureau of Transport and Regional Economics [of Australia], (2003). *Rail infrastructure pricing: principles and practice.* Report 109.

Bellifemine, F., Caire, G., Poggi, A., and Rrimassa, G.. (2003). JADE - a white paper', *EXP in Search of Innovation*, 3, (3), pp. 6-19.

Binmore, K. & Vulkan, N. (1998). Applying game theory to automated negotiation. *Netnomics*, 1, 1-9.

Buşoniu, L., Babuška, R. & Schutter, B. D. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2), 156-172.

Cantos, P. & Campos, J. (2005). Recent changes in the global rail industry: facing the challenge of increased flexibility. *European Transport*, 29, 1-21.

Chen, N. (2007). Modelling demand for rail transport with dynamic econometric approaches. *International Review of Business Research Papers*, 3(2), 85-96.

Crites, R.H. & Barto, A.G. (1998) Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33(2-3), 235-262.

ECMT [European Conference of Ministers of Transport], (2001). Railway reform: regulation of freight transport markets. *OECD Publication Service*, Paris.

Ernst, D., Glavic, M., & Wehenkel, L. (2004). Power systems stability control: reinforcement learning framework. *IEEE Trans. on Power Systems*, 19(1), 427-425.

Gibson, S., Cooper, G. & Ball, B. (2002). Developments in transport policy: the evolution of capacity charges on the UK rail network. *Journal of Transport Economics and Policy*, 36(2), 341-354.

Gomez-Perez, G., Martin-Guerrero, J.D., Soria-Olivas, E., Balaguer-Ballester, E., Palomares, A. and Casariego, N. (2009). Assigning discounts in a marketing campaign by reinforcement learning, *Expert Systems with Applications*, 36, 8022-8031.

Ho, T.K., Ip, K.H. & Tsang, C.W. (2009). Service bid comparisons by fuzzy ranking in open railway market timetabling. *Expert System with Applications*, 36, 10334-10343.

Jiang, C. & Sheng, Z. (2009). Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Systems with Applications*, 36, 6520-6526.

Kaelbling, L.P., Littman, M.L. & Moore, A.W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237-285.

Li, T., van Heck, E., Vervest, P., Voskuilen, J., Hofker, F. & Jansma, F. (2006). A passenger travel behavior model in railway network simulation. *Proc. of Winter Simulation Conference*, USA.

Luo, X., Jennings, N.R., Shadbolt, N., Leung, H. & Lee, J.H. (2003). A fuzzy constraint based model for bilateral multi-issue negotiations in semi-competitive environments. *Artificial Intelligence*, 148(1/2), 53-102.

Oo, J., Lee, J.W. & Zhang, B.T. (2002). Stock trading system using reinforcement learning with cooperative agents. *19th International Conference on Machine Learning*, 451-458.

Pietrantonio, L.D. & Pelkmans, J. (2004). The economics of EU railway reform. *Bruges European Economic Policy (BEEP) Briefing*.

Ribeiro, C.H.C. (1999). A tutorial on reinforcement learning techniques. *International Conference on Neural Network*, INNS Press, Washington, DC, USA.

Shen, W., Ghenniwa, H. & Li, Y.S. (2006). Agent-based service-oriented computing and applications. *1st International Symposium on Pervasive Computing and Applications*.

Suematsu, N. & Hayashi, A. (2002). A multiagent reinforcement learning using extended optimal response. *1st International Joint Conference on Autonomous Agents and Multiagent Systems,* 370-377.

Sutton, R.S. & Barto, A.G.. (1998). *Reinforcement learning: an introduction*. MIT Press, Cambridge.

Tesauro, G. & Kephart, J.O. (2002). Pricing in agent economies using multi-agent Q-learning.

Autonomous Agents and Multiagent Systems, 5(3), 289-304.

Thamizh Arasan, V. & Selvaraj, R. (1997). Development of demand model for long-haul intercity rail travel. *Journal of the Institution of Engineers, India Civil Engineering Division*, 78, 53-57.

Tsang, C.W. & Ho, T.K. (2006). Conflict resolution through negotiation in a railway open access market: a multi-agent system approach", *Transportation Planning & Technology*, 29(3), 157-182.

Tsang, C.W., (2007). *Modelling negotiations in open railway access market for resource allocation*. PhD Thesis, The Hong Kong Polytechnic University.

Tsang, C.W. & Ho, T.K. (2008). Optimal track access rights allocation for agent negotiation in an open railway market, *IEEE Transactions on Intelligent Transportation Systems*, 9(1), 68-82.

Wang, Y.C. & Usher, J.M. (2005). Application of reinforcement learning for agent-based production scheduling. *Engineering Applications of Artificial Intelligence*, 18, 73-82.

Watkins, C.J.C.H. (1989). *Learning from delayed rewards*. PhD Thesis, Cambridge University.

Zhao. Z. & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. *24th International Conference on Machine Learning*.

Figures

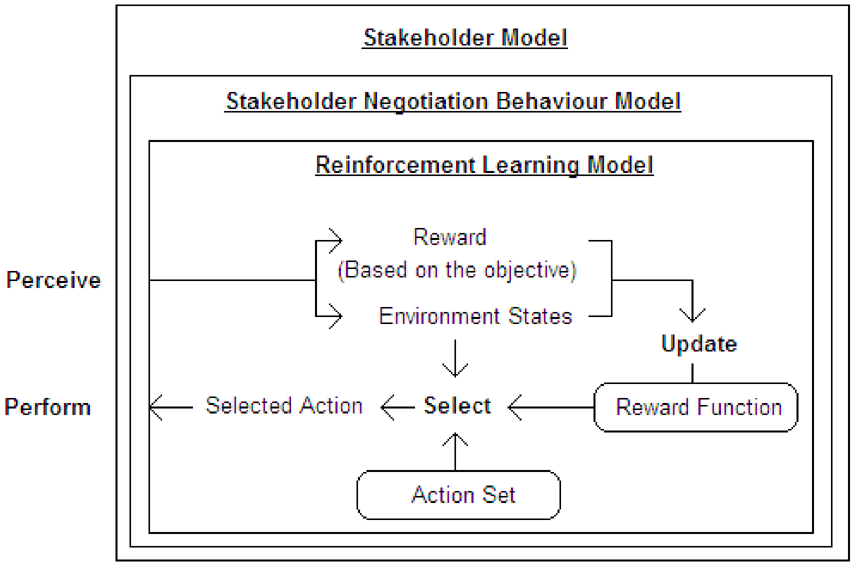

Fig. 1. Intelligent stakeholder negotiation behaviour model for an open railway access market

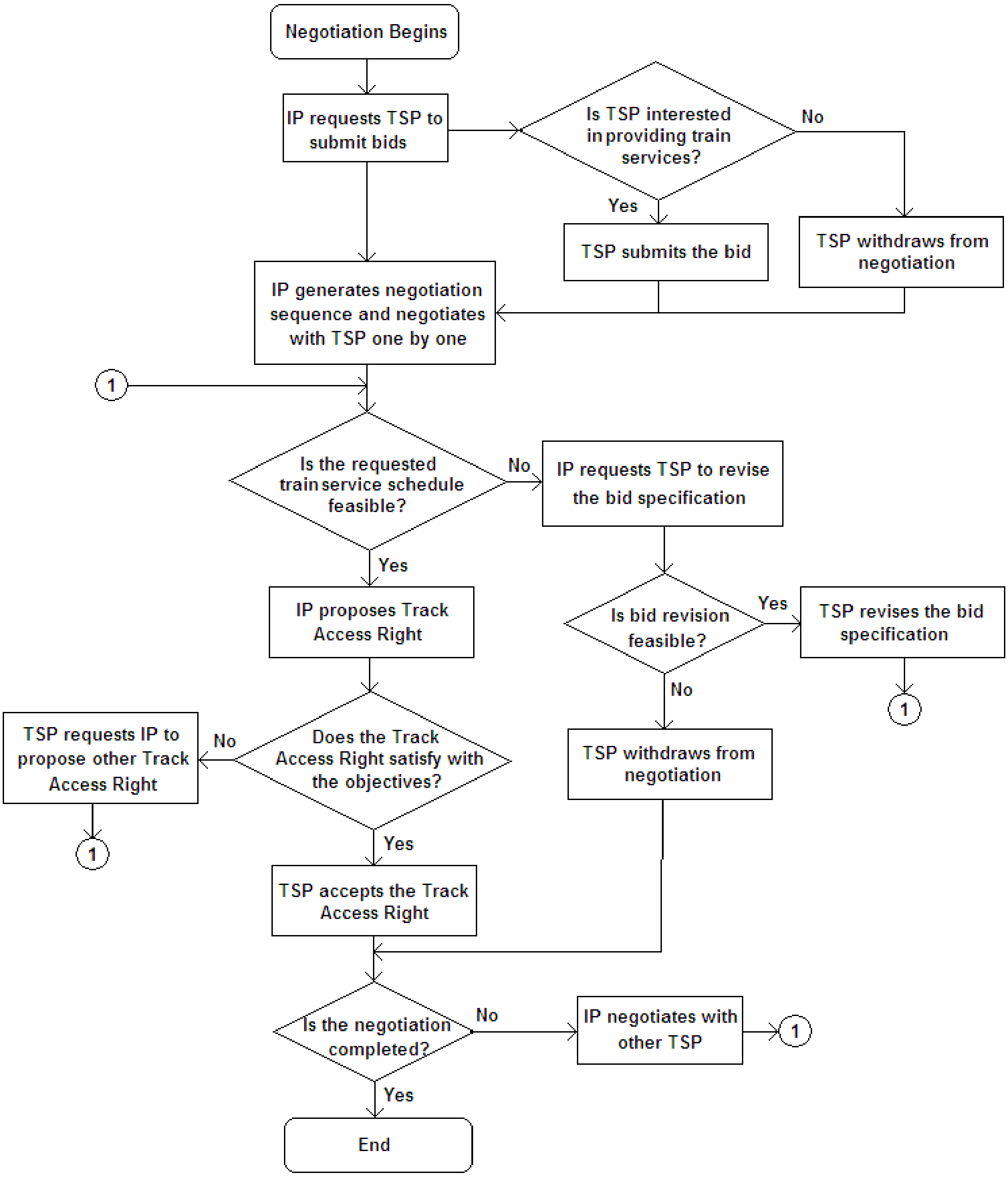
Negotiation Begins
IP requests TSP to submit bids
Is TSP interested in providing train services?
No
Yes
TSP submits the bid
TSP withdraws from negotiation
IP generates negotiation sequence and negotiates with TSP one by one
1
Is the requested train service schedule feasible?
No
IP requests TSP to revise the bid specification
Yes
IP proposes Track Access Right
Is bid revision feasible?
Yes
TSP revises the bid specification
1
No
TSP requests IP to propose other Track Access Right
No
Does the Track Access Right satisfy with the objectives?
TSP withdraws from negotiation
1
Yes
TSP accepts the Track Access Right
Is the negotiation completed?
No
IP negotiates with other TSP
1
Yes
End

Fig. 2. Negotiation procedures between IP and TSPs

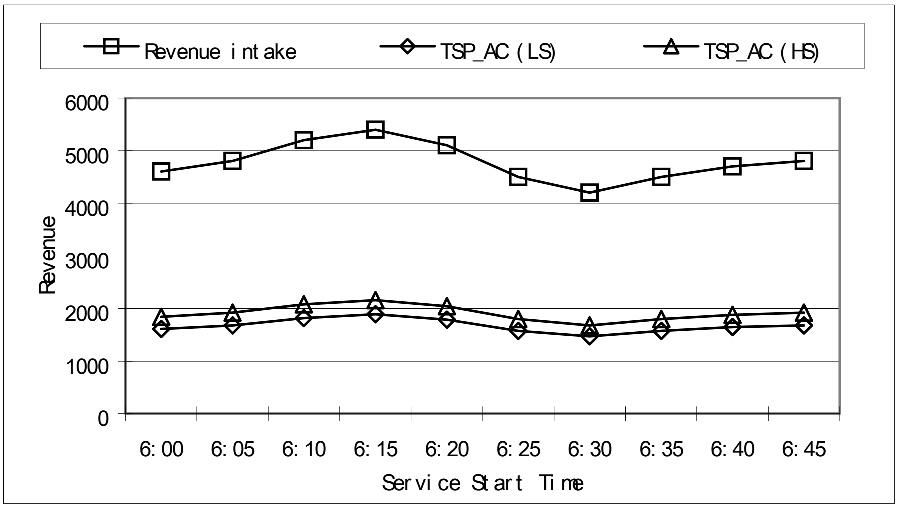

Fig. 3. An example of revenue intake and the highest and lowest acceptance cost of a TSP

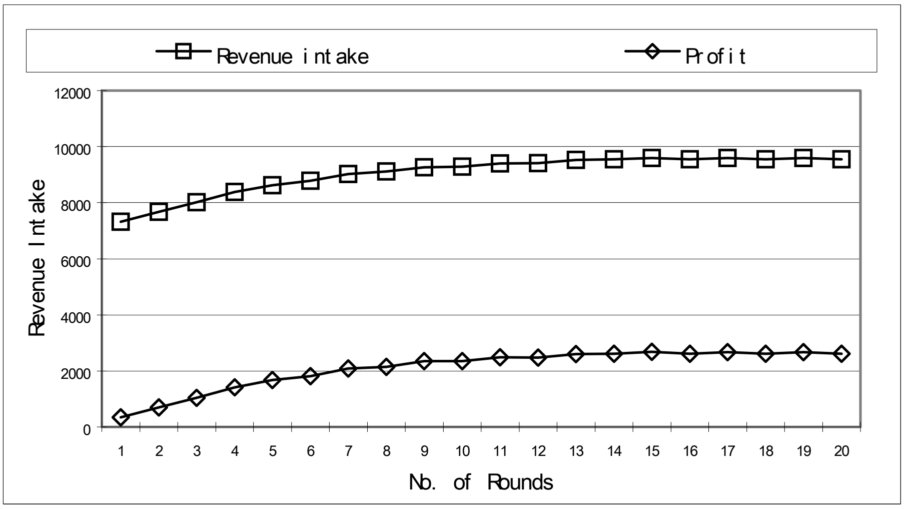

Fig. 4. Revenue intake and profit attained by the IP in the testing phase

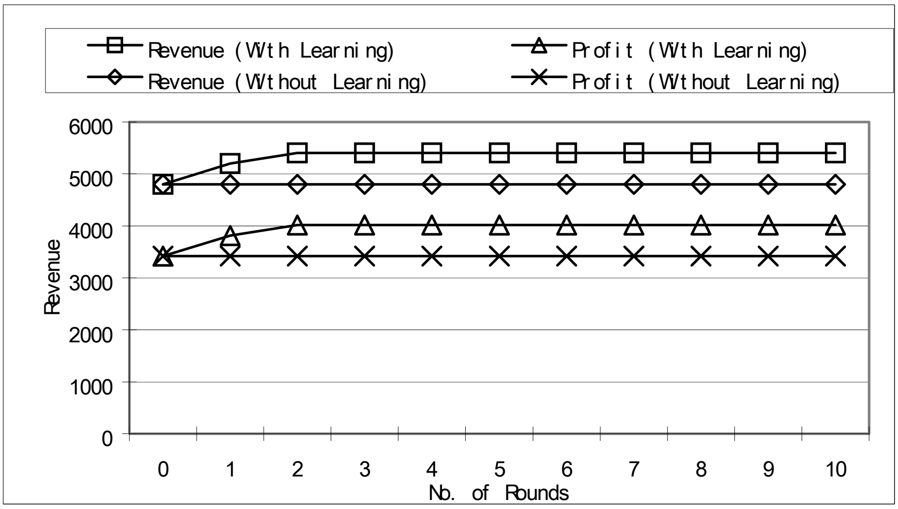

Fig. 5. Revenue intake and profit attained by TSP-1 in the testing phase

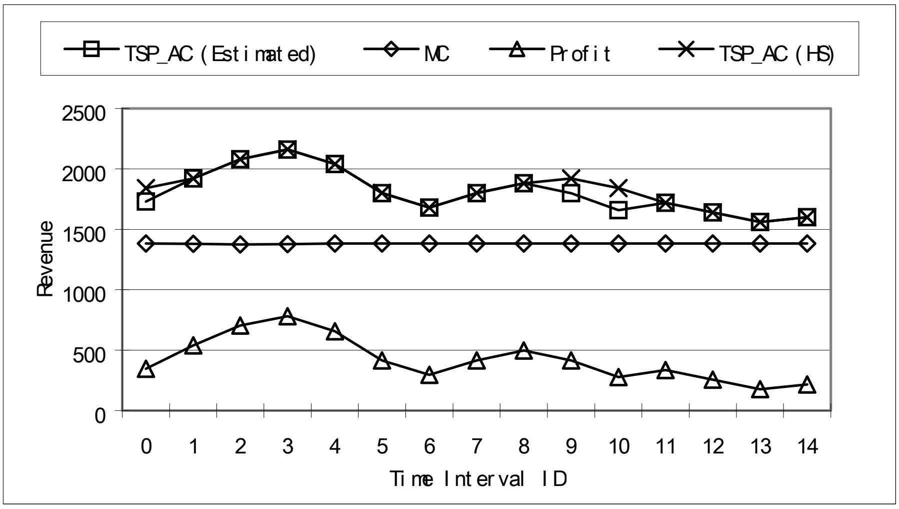

Fig. 6. Estimated TSP-1 acceptance cost, marginal cost, profit and TSP-1_AC (HS)

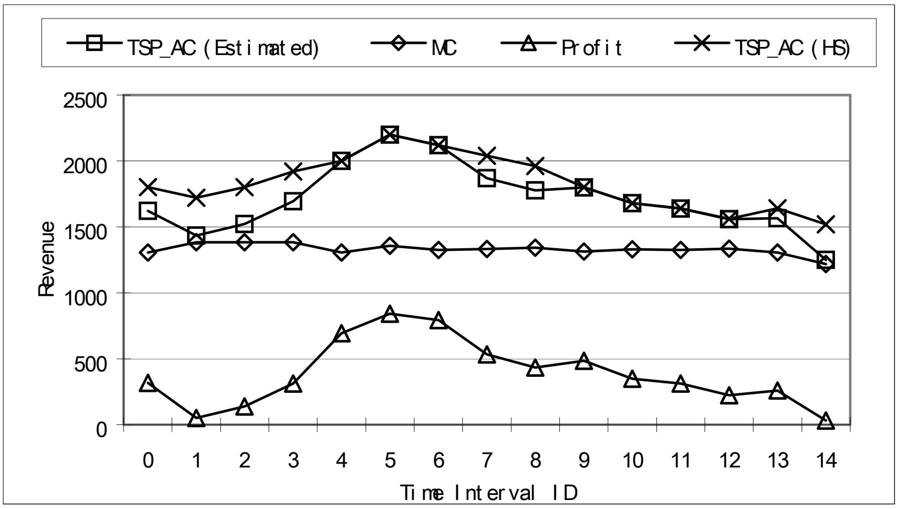

Fig. 7. Estimated TSP-2 acceptance cost, marginal cost, profit and TSP-2_AC (HS)

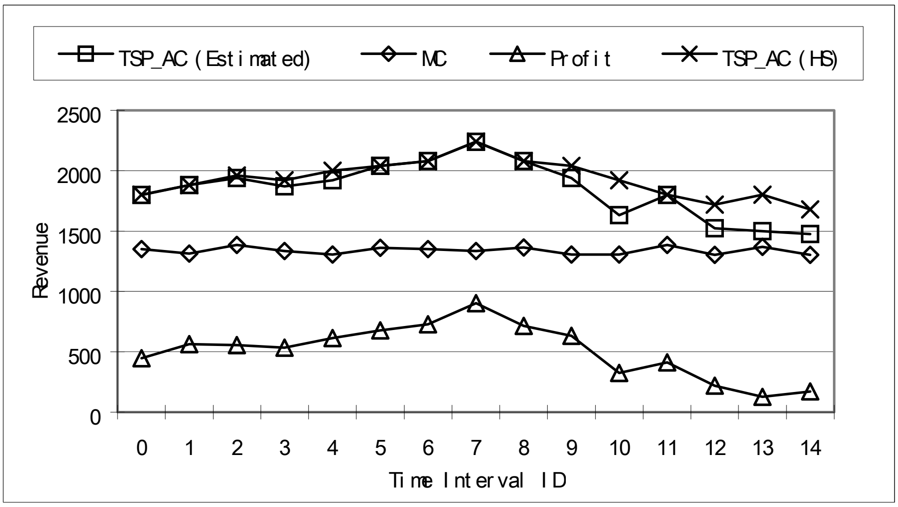

Fig. 8 Estimated TSP-3 acceptance cost, marginal cost, profit and TSP-3_AC (HS)

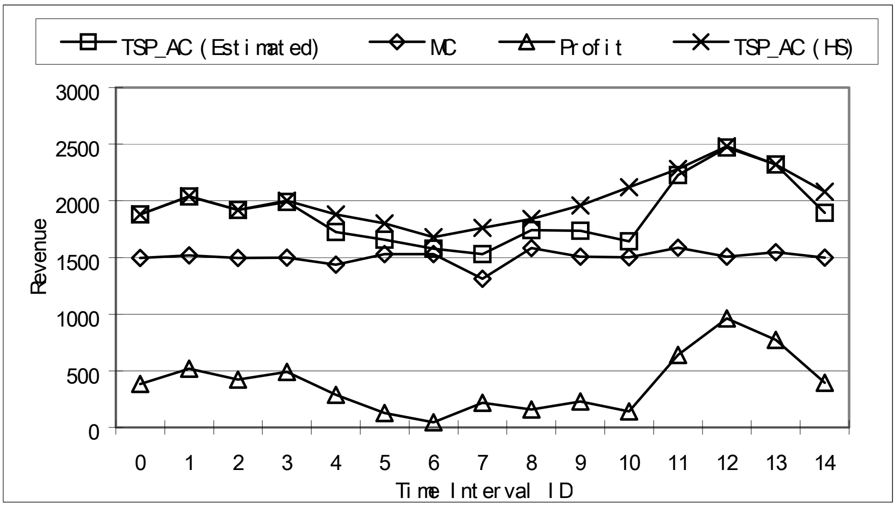

Fig. 9. Estimated TSP-4 acceptance cost, marginal cost, profit and TSP-4_AC (HS)

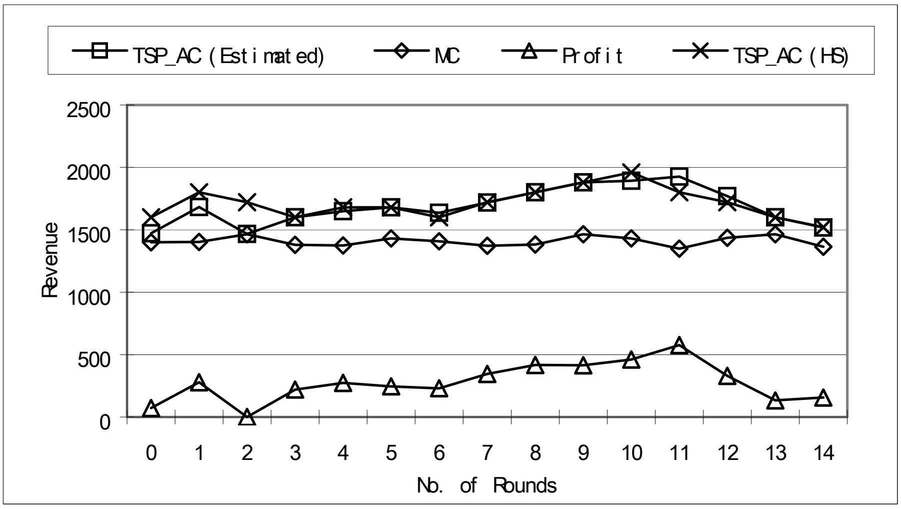

Fig. 10. Estimated TSP-5 acceptance cost, marginal cost, profit and TSP-5_AC (HS)

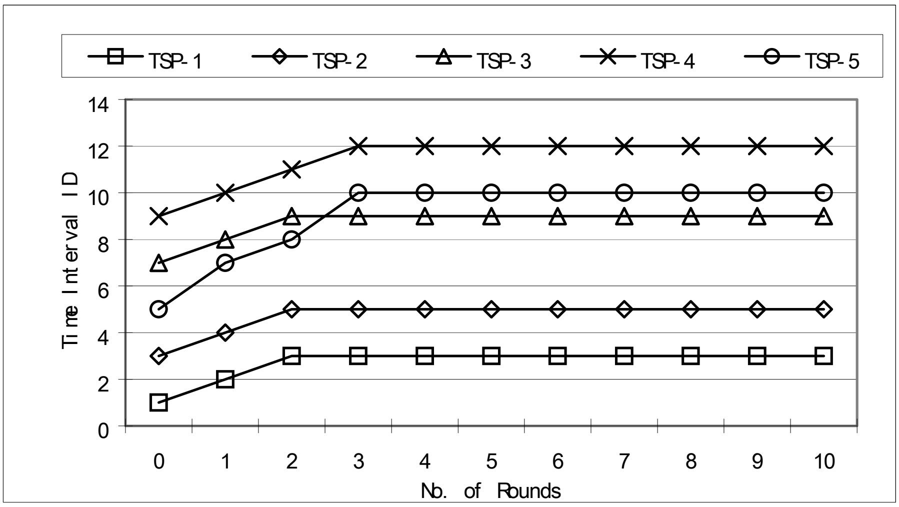

Fig. 11. TSP proposed service start-time in the testing phase

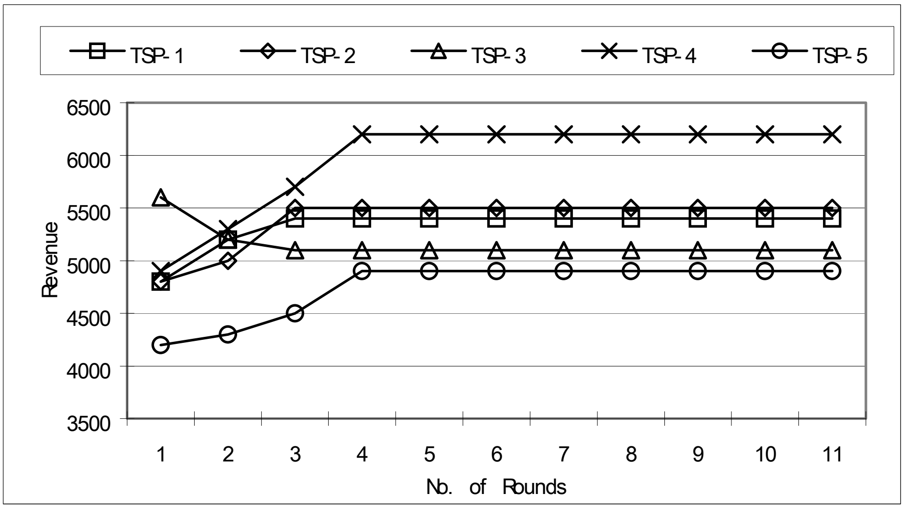

Fig. 12. Revenue intake of TSPs in the testing phase

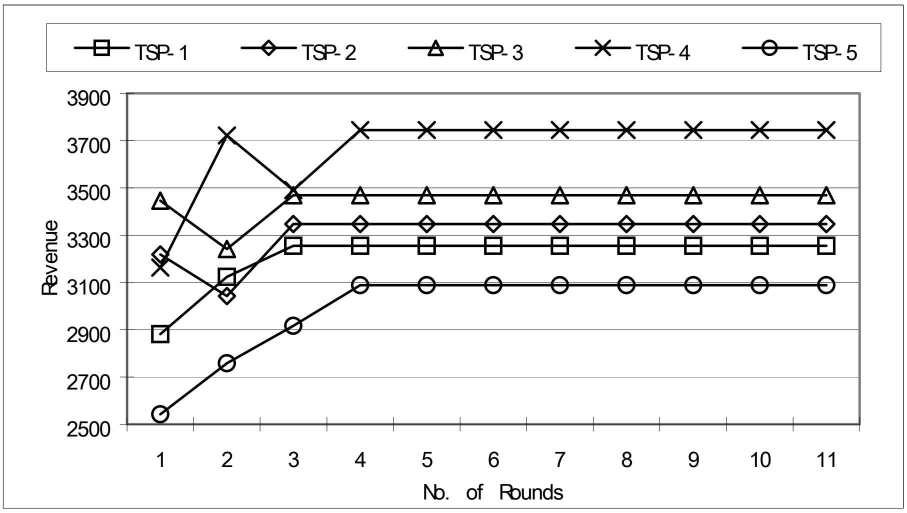

Fig. 13. Profit of TSPs in the testing phase

Tables

Table 1 Initial TSP proposed train schedules and their maximum acceptance cost

ID	TSP-1	TSP-2	TSP-3	TSP-4	TSP-5
ζ (hh:mm)	06:05	06:15	06:35	06:45	06:25
T_D (min)	{1,1,1,1,1}	{1,1,1,1,1}	{1,1,1,1,1}	{5,0,0,0,3}	{13,12,0,13,16}
T_R (min)	{16,24,14,15}	{16,24,14,15}	{16,24,14,15}	{11,16,9,11}	{24,34,23,23}
TSP_AC($)	1920	1920	2240	1960	1680

Table 2 Distances between the stations

Origin station	Destination station	Inter-station station (km)
A	B	20
B	C	30
C	D	15
D	E	20

Table 3 Collected revenue intake and the final cost upon TSP acceptance

	Testing phase results				
	TSP-1	TSP-2	TSP-3	TSP-4	TSP-5
Collected Revenue (TAC)	1920	1826	2220	1924	1680
Marginal Cost (IP Expense)	1362	1305	1306	1568	1379
Profit	558	521	914	356	301

Table 4 Agreed service and revenue intake attained by TSP-1

Initial service start-time	6:05	
Initial time interval ID	1	
Initial revenue intake	4800	
	With Learning	Without Learning
Agreed service start-time	6:15	6:05
Agreed time interval ID	3	1
Revenue intake	5400	4800
Agreed TAC	1384	1384
Profit	4016	3416

Table 5 IP performance during IP-TSP learning

	TSP Proposed Service Start-Time ID	TSP_AC (HS)	IP Estimated Max. TSP_AC
TSP-1	3	2160	2159
TSP-2	5	2200	2199
TSP-3	9	2040	1939
TSP-4	12	2480	2469
TSP-5	10	1960	1892

Table 6

	Time Interval ID with Max. Revenue Intake	Revenue Intake	Time Interval ID of Agreed Services	Revenue Intake	Profit
TSP-1	3	5400	3	5400	3255
TSP-2	5	5500	5	5500	3347
TSP-3	7	5600	9	5100	3468
TSP-4	12	6200	12	6200	3744
TSP-5	10	4900	10	4900	3088