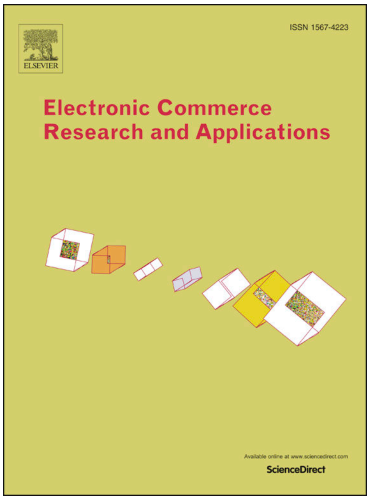
Journal Pre-proofs


ISSN 1567-4223
ELSEVIER
Electronic Commerce Research and Applications
Available online at www.sciencedirect.com
ScienceDirect

Learning Pareto Optimal Solution of a Multi-Attribute Bilateral Negotiation Using Deep Reinforcement

Mina Montazeri, Hamed Kebriaei, Babak N. Araabi

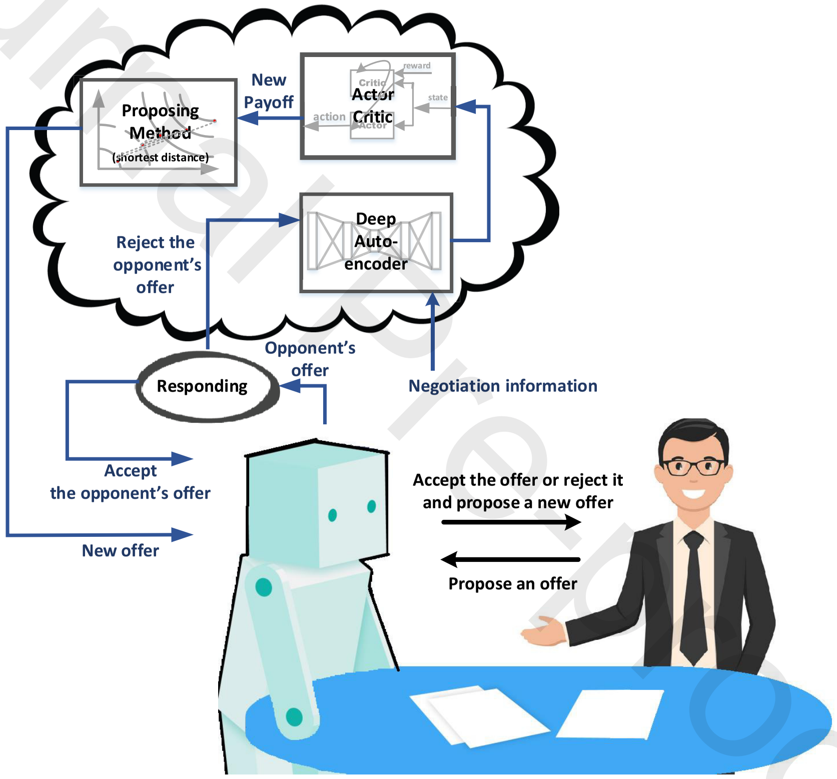
PII: S1567-4223(20)30064-8
DOI: https://doi.org/10.1016/j.elerap.2020.100987
Reference: ELERAP 100987

To appear in: *Electronic Commerce Research and Applications*

Received Date: 25 November 2019
Accepted Date: 12 June 2020

Please cite this article as: M. Montazeri, H. Kebriaei, B.N. Araabi, Learning Pareto Optimal Solution of a Multi-Attribute Bilateral Negotiation Using Deep Reinforcement, *Electronic Commerce Research and Applications* (2020), doi: https://doi.org/10.1016/j.elerap.2020.100987

Graphical Abstract (for review)



Highlights

- The agent learns to decide in a multi-attribute negotiation via Deep Reinforcement.
- The buyer (learner) recognizes the sellers type online using K-means clustering.
- A mediator excludes the unreasonable offers from the feasible set of the offers.
- The existence and the uniqueness of the Nash Bargaining Solution are proven.
- The Bargaining Power of agents is examined.

Learning Pareto Optimal Solution of a Multi-Attribute Bilateral Negotiation Using Deep Reinforcement

Mina Montazeri[a], Hamed Kebriaei[a,*], Babak N. Araabi[a]

[a] *School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.*

Abstract

This paper aims to design an intelligent buyer to learn how to decide in an incomplete information multi-attribute bilateral simultaneous negotiation. The buyer does not know the negotiation strategy of the seller and only have access to the historical data of the previous negotiations. Using the historical data and clustering method, the type of seller is identified online during the negotiation. Then, the deep reinforcement learning method is utilized to support the buyer to learn its optimal decision. In the complete information case, we prove that the negotiation admits a unique Nash bargaining solution with possibly asymmetric negotiation powers. In comprehensive simulation studies, the efficiency of the proposed learning agent is evaluated in different scenarios and we show that the learning negotiation with incomplete information is converged to a Pareto optimal solution. Then, using the concept of the Nash bargaining solution, the negotiation power of the buyer is assessed in negotiation.

Keywords: Multi-attribute negotiation, Deep auto encoder, Actor-Critic, Nash bargaining solution, Bargaining power.

1. Introduction

In bilateral contracts, multi-attribute negotiation is a common practice to reach an agreement (Hao and Leung, 2012), not only due to the fact that two parties usually have to negotiate on multiple issues of a trading good (Yu and Wong, 2015), but also to take the advantage of reaching possibly to a "win-win" solution (Lai et al., 2008b). For example, in the navy problem (Lai et al., 2004), commands and sailors usually have to negotiate multi issues, like payment rate, projection rotation date, length of service training. In bilateral contracts of energy in the electricity market (Kebriaei and Majd, 2009), loads and generators usually have to negotiate the quantity and price of energy. Also in the company's

[*] Corresponding author, email address: Kebriaei@ut.ac.ir.

This work is supported in part by the Institute for Research in Fundamental Sciences under Grant CS 1398-4-256.

contract (Lai et al., 2008a), an employer and a union usually need to negotiate wage level, health care and vacations.

The essence of bilateral negotiation is the exchange of proposals (Hao and Leung, 2012). Both the buyer and seller submit their offers and counteroffers in order to converge to a mutually acceptable agreement. Research studies on bilateral negotiation have mainly employ tools from Game Theory (Lai et al., 2006; Chen and Weiss, 2015; Lin and cho T. Chou, 2004), and Artificial Intelligence (AI) (Camerer et al., 2017; Zhan et al., 2017; Chen et al., 2017; Francisco et al., 2019; Eshragh et al., 2019; Buffett and Spencer, 2007). These approaches analyze the possible optimal solutions of the negotiation (Game Theory) or design an intelligent agent that learns to maximize its profit during the negotiation process (AI).

In many researches in the AI field like (Hajimiri et al., 2014; Baek and Kim, 2007; Jamali and Faez, 2012; li Huang and ren Lin, 2008), only single attribute bilateral negotiation is studied, while in many applications, agents need to go through multi-attribute negotiation. Furthermore, the single attribute negotiation naturally results in a ‘win-lose’ situation. As a result, in recent years, studies have focused more and more on multi-attribute negotiation (Zhan et al., 2018; Kersten et al., 2013; Kolomvatsos et al., 2016). Moreover, many researchers assumed that the opponent follows a predefined strategy (e.g. time-dependent) during the negotiation while depending on the negotiation’s condition, the opponent’s strategy may change during the negotiation. This research aims to design a buyer’s representative learning agent to deal with the bilateral negotiation over the price and the quantity. An agent can observe the price and the quantity offered by its opponent but it does not know the opponent’s payoff function and conceding strategy. In this work, the buyer learns how to negotiate with the seller to maximize its long-term expected payoff using Deep Reinforcement Learning (DRL) approach and then, the results are assessed through Game Theory and the concept of Nash bargaining solution with asymmetric negotiation power. We propose a state-action-reward framework that is learned by Deep Auto-encoder via the Actor-Critic algorithm (Lange and Riedmiller, 2010; de Bruin et al., 2018; Maldonado-Ramirez et al., 2018) so that the buyer agent learns how to negotiate with the seller, without having information about the model of seller. The deep auto-encoder used in this approach helps the agent to learn a suitable low dimensional state space which is reduced from a high dimensional input space (Hinton and Salakhutdinov, 2006). Moreover, Actor-Critic as the RL method helps the agent to learn in a continuous environment and also avoids the curse of dimensionality (Van Hasselt and Wiering, 2007; Lillicrap et al., 2015).

Since the buyer has no information about the seller’s behavior, in this paper; a K-means clustering is designed to help the buyer to recognize the type of opponent from its historical offers. To speed up the learning process, we have also used a mediator to exclude the unreasonable offers from the feasible set of the negotiation offers by the proposed algorithm. Unreasonable offers are those less beneficial for both of the agents. In this paper, the mediator does not interfere with the agents’ offers and it does not require any information about the agents' strategies. To analyze the results of the learning process, we obtain the Pareto optimal solutions of the negotiation with complete information using the Nash bargaining solution with asymmetric bargaining power. The existence and the uniqueness of the solution to our problem are also proven. Then, by comparing the agreement resulted from the DRL method and the Nash bargaining solution with different bargaining powers, we can verify that the DRL method maintains the Pareto optimality while the bargaining power of the agents can be assessed by this comparison.

The main contributions of this paper are as follows:

- The opponent's type is identified online adaptively by using K-means clustering according to the seller's offers during the negotiation process.
- The buyer agent is designed to learn how to concede using a DRL method in a continuous environment, by extracting effective elements as a state from high-dimensional input via deep auto-encoder, without having information about the model of the seller agent.
- In order to increase the learning speed, a mediator-based algorithm is proposed to exclude the unreasonable offers (less beneficial offers for both of the agents) from the feasible set of the negotiation offers. The mediator does not influence the agents' offering strategies nor enquire any information about the agents' payoff functions.
- The existence and the uniqueness of the Nash bargaining solutions for the proposed bilateral negotiation have been proven.
- The Pareto efficiency of the learning mechanism and the bargaining power of the negotiating agents are examined by comparing the agreement resulted from the DRL method and that of the Nash bargaining solution with asymmetric bargaining powers.

The rest of the paper is organized as follows: In Section 2, we review the literature about negotiation. Preliminaries needed in this paper are given in Section 3. Section 4 provides the problem statement and proposes the assumed model. In Section 5 Nash bargaining solution with complete information is described. Automated negotiation with incomplete information is described in Section 6. Learning in negotiation is studied in Section 7. Section 8 includes a case study, the simulation study of negotiation with the proposed deep actor-critic agent and their analysis. The conclusion is presented in the final section.

2. Literature review

There are three types of multi-attribute negotiation; separate, simultaneous and sequential/hierarchical negotiation (Lai et al., 2004). Separate negotiation means that agents negotiate each attribute, independently. While in simultaneous negotiation, the agents negotiate a complete package on all the

attributes, concurrently. In the sequential/hierarchical negotiation, agents negotiate "attribute-by-attribute".
Separate negotiation is like multiple single attribute negotiation. So, it is not possible to achieve a win-win solution in this type of negotiation. On the other hand, in sequential negotiation, an agent needs to determined its desired order of attributes. Therefore, if the agents have different order of attributes, we cannot use this type of negotiation (Gerding et al., 2000b; Fershtman, 2000). In addition, if the attributes are interdependent, it would be probable that we cannot either prioritize the attributes or prioritization of attributes leads to sub-optimal solution. Nevertheless, in such cases, using simultaneous multi attribute negotiation can lead to a Pareto-optimal win-win solution (Lai et al., 2004). The problem under study in this paper is the simultaneous multi-attribute negotiation.
The study on simultaneous multi-attribute negotiation is mainly conducted by cooperative multi-attribute negotiation in Game Theory. Cooperative Game Theory aims to specify the number of fair axioms in which agreement on those axioms naturally leads to a fair optimal solution. Among different axioms, Pareto optimality is the main condition that guarantees the bargaining solution lay on the Pareto frontier. The Nash bargaining solution is the most popular one (Nash, 1950) which is generalized also to the asymmetric Nash bargaining solution depending on the bargaining powers of the agents (Gerding et al., 2000a). Kalai and Smorodinsky also proposed some modifications on the Nash bargaining solution (Kalai and Smorodinsky, 1975).

In practice, a negotiator can observe the offers proposed by the opponent, but it doesn't have access to the information of payoff function and other parameters of the opponent, like reservation payoff or negotiation deadline. Negotiations with incomplete information in the Game Theory field are divided into two broad categories: mediated negotiation (Chen et al., 2014; Lin and cho T. Chou, 2004; Ehtamo et al., 1999; Klein and Faratin, 2003) and non-mediated negotiation (Lai et al., 2008b). Sycara in (Lai et al., 2006) and Ehtamo in (Ehtamo et al., 1999) have considered a mediator into the negotiation procedure to help the agents to reach the Pareto optimal solution with incomplete information. In this method, a mediator generates constraints in each step of negotiation and asks agents to propose an offer under those constraints. Although the Pareto optimality is achieved, however, due to the influence of the mediator on the agent's offers, the agents cannot make a decision independently based on their own negotiation strategy. Klein and Faratin (Klein and Faratin, 2003) propose a more tractable and decentralized mediating approach for binary valued but complex negotiations where the mediator generates an offer in each step and proposes it to both of the agents. The agents then decide whether to accept the offer based on their own strategies. Two types of negotiation strategies are proposed and the equilibrium outcomes are examined as well as the system efficiency. In the non-mediated negotiation, Sycara in (Lai et al., 2008b) suggests that agents propose an offer on their indifferent curve, which has the shortest distance to the offer made by the opponent in the previous step. Sycara verifies that the result obtained by this method is close to Pareto frontier but not exactly on the Pareto frontier. In this work, the conceding behavior of both agents is modeled by the fixed time-dependent strategy.

Research on Artificial Intelligence has mainly focused on two areas: learning an opponent model and learning the negotiation method. In the literature, Bayesian-learning (Eshragh et al., 2019; Buffett and Spencer, 2007), non-linear regression (Yu et al., 2013),and neural network (Lee and Ou-Yang, 2009) have been used to learn different opponent attribute like acceptance strategy, deadline, preference profile and bidding strategy (Baarslag et al., 2016). In learning the negotiation method, Sycara (Sycara, 1991) (Sycara, 1990) has presented a plausible approach in the negotiation where the agents make their offer based on the similarity of the current negotiation to previous negotiations. In (Faratin et al., 2002) Faratin has proposed a trade-off strategy to increase the chance of reaching an agreement without decreasing the payoff. In this method, agents have fuzzy criteria to approximate the preference structure of the other agents. In (Carbonneau et al., 2008), an Artificial Neural Network (ANN) is constructed with three layers for predicting the opponent's offers during the negotiation process. The network exploits information from offers made by agents in prior negotiations to predict future offers. In 2017, Zhan (Zhan et al., 2017) designed a fuzzy algorithm to update the agents' payoff in each negotiation round. The paper shows how different features like regret, risk, and patience influence on agents' payoff. Further to the above AI studies, Reinforcement learning (RL) has become an important part of AI that has been applied to the negotiation problem in recent years (Hajimiri et al., 2014; Chen et al., 2014; Takadama, 2008). Takadama (Takadama, 2008) studied the agent behavior in a single state negotiation model, where only the opponent's previous proposal was considered as the agent's state in the Q-learning method. In (Lao and Zhong, 2010) Lao and Zhong proposed another version of the Q-learning method for the negotiation. The main promotion of Lao's paper is considering two parameters to define the agent's state vector at each step of the negotiation. The first parameter was a distance between the last offers of the agents and the second parameter was the current time in the negotiation. Since the environment is continuous, using discrete methods would increase the number of states and make convergence of the algorithm difficult. Huang and Lin (li Huang and ren Lin, 2008) proposed a Temporal Different learning (TD) method in the negotiation. They used Neural Network (N.N) as a function to approximate Q-values. Hajimiri (Hajimiri et al., 2014) proposed a Fuzzy-Sarsa Learning (FSL) method for a single attribute negotiation. In this method, different parameters can be used to define the agent's state vector without the threat of the curse of dimensionality. A distance between the last offer of the agents and the reservation price and time are three parameters those define the state vector.

3. Preliminaries

3.1. Deep Reinforcement Learning

RL is a semi-supervised learning method to enable the agent to learn its optimal decision by interacting with the environment. The RL makes realistic assumptions about the information available to the agent where an agent does not have any information about the payoff function or the parameters of the opponent (R. S. Sutton and Barto, 1998). The agent may only have some sensory information from the previous rounds of the negotiations. The RL agent learns its optimal action in each state, by moving from random decision making to greedy one, while receiving reward and punishment from the environment in different states, and updating the value of each state-action pairs after each decision making (R. S. Sutton and Barto, 1998).

The advent of deep learning has had a significant impact on RL due to the property of finding compact low-dimensional representations of high-dimensional data. This property enables DRL to face with decision-making problems that were previously intractable, i.e., the curse of dimensionality in state and action spaces (Bengio et al., 2013). Solving DRL tasks is usually divided into two steps. The first is, mapping high-dimensional input data into a low-dimensional representation (which here, our focus is using the unsupervised learning methods of deep architectures). The second is, designing an agent to learn how to propose an offer to get the most payoff. These two steps are depicted in Fig. 1 and explained in more detail below.

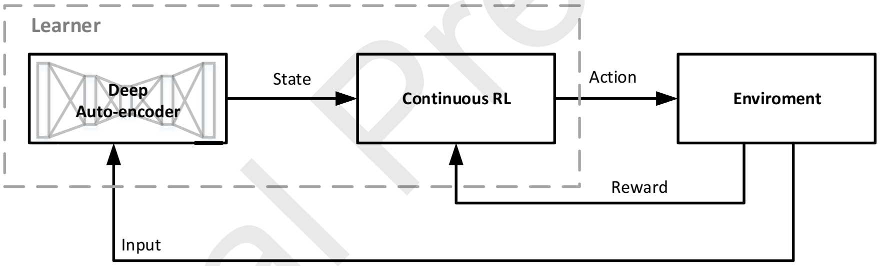


Figure 1: Overview of Proposed DRL method.

3.1.1. Deep Auto-Encoder

To be able to learn properly, the RL agent needs a state vector as input that sufficiently includes information about the output, and yet, it should be minimal in size. A possible solution is to find a transformation that turns the high dimensional input data to a low dimensional representation that encapsulates all the necessary details and feeds it as the RL input. In this paper, a "deep auto-encoder" was proposed as a tool to find the transformation between input and a state of a reasonable size, so that the RL system would be able to learn the system in a reasonable time (Hinton and Salakhutdinov, 2006). A deep auto-encoder can be defined as a deep neural network capable of learning efficient data coding in an unsupervised way (Bengio et al., 2007). A deep auto-encoder consists of two main parts, a deep encoder and a deep decoder, which can e

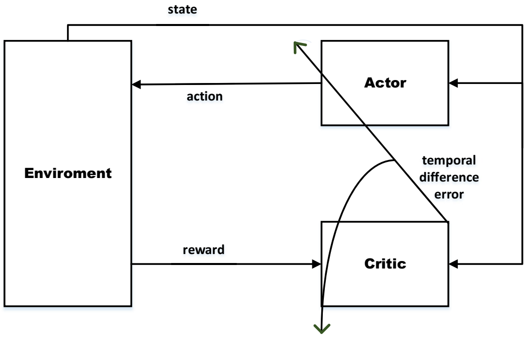


Figure 2: Schematic diagram of actor-critic method.

defined as the following mapping function:

$$f, g = \arg\min_{f,g} \|x - fog(x)\|^2 \tag{1}$$

where x is the input feature vector, $g : R^n \to R^m$ is the encoding function that maps the input vector to its encoded representation and $f : R^m \to R^n$ is the decoding function that maps the encoded representation to a reconstruction of the input vector. n is the dimension of the input vector and m is the dimension of encoded representation of the input.

3.1.2. Actor-Critic

The actor-critic method is one of the continuous RL methods that is base on a temporal difference learning. This method consists of two neural networks, actor and critic (Barto et al., 1983). The actor network is responsible for selecting the action according to the current policy, which is the decision mechanism that chooses an action from the state. The critic network receives the reinforcement feedback and comparing it with its estimation of the feedback, and then it updates the parameters of itself and actor network to minimize the error. The overview of the actor-critic method is depicted in Fig.2.

Critic network assigns a value to each state by a neural network, which is formulation in Equation 2:

$$V(s) = \sum_{i=1}^{h} w_i \phi_i(s) \tag{2}$$

where w_i is the weight of the network and $\phi_i(s)$ is the Radial basis function (RBF). Equation 3 shows the RBF .

$$\phi_j(s) = \exp(-\frac{\|s - c_j\|^2}{2{\sigma_j}^2}) \tag{3}$$

where c_j and σ_j are the mean and the standard deviation of RBF.

The actor network employs a neural network as shown in Equation 4 to evaluate the probability of selection of action in a particular state.

$$p(s) = \sum_{i=1}^{h} v_i \phi_i(s) \tag{4}$$

where v_i is the weight of the network. In each state, the critic network calculates the error according to Equation 5:

$$\varepsilon = r_k + \gamma V(s_{k+1}) - V(s_k) \tag{5}$$

and then updates the parameters in both actor and critic networks to minimize the error.

3.2. Nash Bargaining Solution

In cooperative bargaining, the agents are supposed to have access to all joint feasible payoffs and the solution of bargaining is obtained by some rational axioms which are called axiomatic solution. The main feature of the cooperative bargaining solution which is common in different axiomatic methods is Pareto optimality. The Pareto optimal solution is the point that none of the agents can be made better off without making at least one agent worse off. The all of these Pareto optimal points create a boundary which is called Pareto frontier. The Nash bargaining solution is a well-known axiomatic solution to the bargaining that satisfies the following axioms (Nash, 1950).

1. Pareto efficiency: the agreement will represent a situation that none of the agents can be made better off without decreasing the payoff of at least one agent.
2. Summitry: the payoff should not discriminate between the indistinguishable agents.
3. Invariant to affine transformation: An affine transformation of the payoff and disagreement point does not change the shape of the solution although it changes the numerical value of the solution
4. Independent of irrelevant alternative: if s is Nash bargaining solution for bargaining set x then for any subset of x like y $(y \subseteq x)$ that contains s, s is still a Nash bargaining solution.

Based on (Nash, 1950), when the payoff set of the agents is convex and closed, the Nash bargaining solution (v_1^*, v_2^*) can be obtained by solving the following optimization problem.

$$(v_1^*, v_2^*) = \arg\max_{v_1,v_2} (v_1 - d_1).(v_2 - d_2) \ \textit{subject to} (v_1, v_2) \in U \tag{6}$$

where d_1 and d_2 are the disagreement payoff pair and U is the payoff set of the agents. In (Binmore et al., 1986), Binmore demonstrates that different factors like the asymmetry in the agent's payoffs, different treat points, different beliefs about the environment and different bargaining procedures, cause an asymmetry in the bargaining procedure. Therefore, the asymmetric version of the Nash bargaining solution can be characterized by the pair of payoff (v_1^*, v_2^*) that solve the following problem.

$$(v_1^*, v_2^*) = \arg\max_{v_1, v_2} (v_1 - d_1)^{\alpha}.(v_2 - d_2)^{1-\alpha} \ \textit{subject to}(v_1, v_2) \in U \tag{7}$$

where α represents the bargaining power of agents.

4. Problem statement

We consider a two attribute bilateral negotiation with a seller and a buyer as the negotiator and the price and the quantity as the attributes of trading goods, respectively. The negotiation procedure is according to the general alternating-offer negotiation protocol (Rubinstein, 1982).

In the alternating-offer protocol, the buyer (proposer) suggests an offer to the seller, by specifying its desired values for each attribute. Then, the seller (responder) reacts to that offer by either accepting or rejecting the buyer's proposal. If the responder accepts the offer, the negotiation comes to an end; otherwise, the agents exchange their roles and the seller proposes a counter-offer in the next step of negotiation. This process continues until an agreement occurs or the negotiation deadline is met for an agent.

Accepting or rejecting a proposal, and also making an offer/counter-offer is decided based on the payoff functions of the agents. For a buyer and a seller, we can generically model the payoff functions, respectively as follows:

$$payoff_b(p, q) = u(q) - p.q \tag{8}$$

$$payoff_s(p, q) = p.q - c(q) \tag{9}$$

where p is the price and q is the quantity of the trading good that $q \in \Delta = [q_{\min}, q_{\max}]$ and $p \in \Gamma = [p_{\min}, p_{\max}]$. $u(q)$ is the utility function of the buyer and $c(q)$ is the cost function of the seller which $u(q)$ is concave, $c(q)$ is convex and both of them are positive.

In our methodology, first, we perform a thorough theoretical analysis of the negotiation with complete information, which gives the ideal results for the more complex incomplete information case. To this aim, we derive the Nash bargaining solutions of the proposed bargaining model with asymmetric negotiation powers. In addition, the existence and uniqueness of the solution have been proven. In the next step, assuming incomplete information about the type and payoff function of the opponent, we employ tools from clustering and deep reinforcement learning to design an intelligent agent (buyer) to be able to negotiate with the opponent (seller) over two attributes (price and quantity). In the final step, we study the results of both approaches by comparing the agreement achieved by designing an intelligent agents with incomplete information and the Nash bargaining solution with asymmetric negotiation powers with complete information.

5. Nash bargaining solution with complete information

In this paper, we begin by analyzing bargaining with complete information. We show that the proposed negotiation has a unique Nash bargaining solution. The payoff values for the buyer and the seller can be calculated from equations (8) and (9). Since the agents are not willing to reach an agreement with a negative payoff, the following constraint is imposed on the agreement price :

$$p \in \Gamma(q) = \{p | \frac{c(q)}{q} \leq p \leq \frac{u(q)}{q}\} \tag{10}$$

The payoff set of agents is defined as follows:

$$U = \{(u_1, u_2) | u_1 = payoff_b(p, q),\ \ u_2 = payoff_s(p, q)\ \forall p \in \Gamma, \forall q \in \Delta\} \tag{11}$$

In the following theorem, we prove that the set of possible joint payoffs of the negotiation defined above i.e. U is closed and convex, and therefore the Nash bargaining solution exists and is unique for this bargaining problem (Nash, 1950).

Theorem 1. *The set U is convex and closed.*

Proof. The total payoff of a seller and a buyer which only depends on the quantity of the trading good (q) is as follows:

$$payoff_b + payoff_s = u(q) - c(q) \tag{12}$$

First, we show that U is equal to a set S which is defined as follows:

$$S = \{(s_1, s_2) | s_1 > 0, s_2 > 0, s_1 + s_2 = u(q) - c(q)\ \ \forall q \in \Delta\} \tag{13}$$

To show $S = U$ one can verify $S \subseteq U$ and $U \subseteq S$.
If $(u_1, u_2) \in U$, $(s_1, s_2) \in S$ then there is $q_1 \in \Delta, p_1 \in \Gamma$ such that $u_1 = payoff_b(p_1, q_1), u_2 = payoff_s(p_1, q_1)$. As a result

$$u_1 + u_2 = u(q_1) - c(q_1) \tag{14}$$

Therefore, based on 13 and 14 , $(u_1, u_2) \in S$.
On the other hand, if $(s_1, s_2) \in S$ then there is $q \in \Delta$ such that:

$$s_1 = u(q) - c(q) - s_2 \tag{15}$$

By considering s_1 as a feasible payoff for the buyer, we have:

$$s_1 = u(q) - c(q) - s_2 = u(q) - p.q \tag{16}$$

By solving equation 16, we obtain:

$$p = \frac{c(q) + s_2}{q} \leq \frac{c(q) + s_1 + s_2}{q} = \frac{c(q) + u(q) - c(q)}{q} = \frac{u(q)}{q} \in \Gamma \tag{17}$$

Which is a feasible value for p. Similar to previous part, by considering s_2 as a feasible payoff of the seller, we have:

$$s_2 = u(q) - c(q) - s_1 = p.q - c(q) \tag{18}$$

By solving equation 18, we obtain:

$$p = \frac{u(q) - s_1}{q} \geq \frac{u(q) - s_1 - s_2}{q} = \frac{u(q) + c(q) - u(q)}{q} = \frac{c(q)}{q} \in \Gamma \tag{19}$$

According to equation 17 and 19 we find $(s_1, s_2) \in U$. As a result $S = U$ and the proof is completed . Now by considering:

$$q_m = \arg\max_{q \in \Delta} (u(q) - c(q)) \tag{20}$$

$$q_n = \arg\min_{q \in \Delta} (u(q) - c(q)) \tag{21}$$

The upper and lower bound of the total payoff functions of a buyer and the seller are obtained as follow:

$$J_{\max} = \max(s_1 + s_2) = \max(payoff_b + payoff_s) = u(q_m) - c(q_m) \tag{22}$$

$$J_{\min} = \min(s_1 + s_2) = \min(payoff_b + payoff_s) = u(q_n) - c(q_n) \tag{23}$$

The set of S is shown in Fig. 3. The upper and lower bounds of $s_1 + s_2$ is determined according to equations 22 and 23. As a result, S is a trapezoid. Since trapezoid is a polygon with all interior angles less than or equal to $\angle 180$, it is a convex quadrilateral. So S and therefore U are convex and closed. □

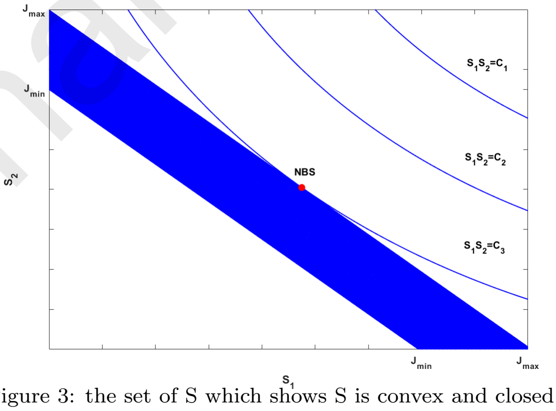


Figure 3: the set of S which shows S is convex and closed.

6. Automated negotiation with incomplete information

Here we consider the bargaining with incomplete information. In this case, the agreement may not occur in one-step and therefore, the agents try to reach an agreement through a negotiation process which can be looked as a repeated game with incomplete information. An agent can observe only the price and the quantity offered by its opponent but it doesn't know the opponent's payoff function.

In bilateral negotiation with incomplete information, each agent has three decision-making tasks including conceding, proposing and responding. Conceding means how much an agent needs to concede from its current payoff toward reaching an agreement, considering its reservation utility. The proposing task determines the suggesting offer (the value of attributes) to the opponent associated with the value of the current (conceded) payoff. It should be noted that in a multi-attribute negotiation it is likely to have many feasible offers with the same payoff for an agent. These offers constitute an indifference curve in the attributes' space. Therefore, it is important for an agent to develop an effective strategy to propose the best point on its indifferent payoff curve to the opponent. The responding strategy as the third task determines whether the agent accepts or rejects the opponent's proposal. Naturally, an agent accepts the offer that gives a payoff greater or equal to its last offer.

The general procedure of the negotiation is presented in Fig.4. In what follows we explain the different components of this scheme. At the end of this section, we give an overview of our approach shown in Algorithm 1.

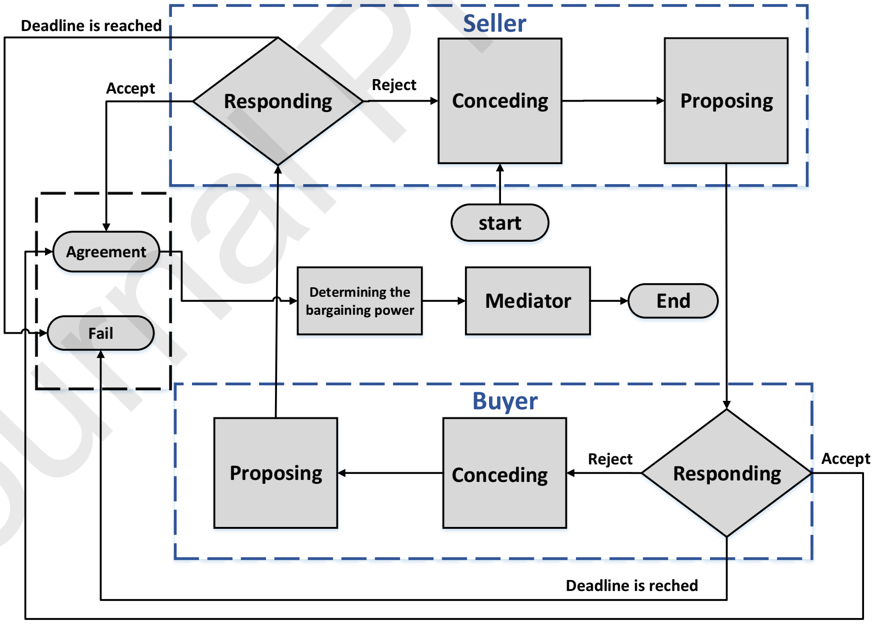


Figure 4: The general flowchart of the proposed method in the paper.

6.1. Conceding

In this paper, the seller concedes its payoff by its own strategy which is not known to the buyer, while the buyer is designed using a DRL method to learn how to concede, considering the current state of negotiation, without having information about the model of the seller. The details are discussed in what follows.

6.1.1. The buyer agent

In this paper, the buyer is designed to learn the policy of how to concede in each step of negotiation to maximize its payoff gained from the agreement. But as the buyer may face different types of opponents in different negotiations, employing a static policy is not helpful (Monteserin and Amandi, 2013). Therefore, the buyer needs to adapt its conceding policy to the different types of sellers. DRL is a powerful approach that can accomplish this goal. In this paper, the buyer is designed using a DRL method to learn how to negotiate with different types of sellers, without having information about the payoff function of the opponent. The deep actor-critic method as a deep and continuous RL method helps the buyer to maximize its payoff via the appropriate definition of the state vector, action, and rewards. Actor-critic, as one of the famous RL methods, helps the buyer to learn the continuous state-action space without facing the curse of dimensionality. Moreover, the deep auto-encoder which is added to actor-critic helps the buyer to consider a variety of inputs that sufficiently includes information about the negotiation to be able to learn the property of negotiation and make the right decision. The number of steps passed from starting the negotiation and left until its deadline, the distance between the buyer's payoff of its offer and its reservation payoff, the distance between the buyer's payoff of its offer and the buyer's payoff of the seller's offer, the distance between the buyer's payoff from the two consecutive seller's offer, the distance between the two successive seller's offers and the class of the opponent are considered as an input of deep auto-encoder in this paper which are explained in detail in Section 7. The performance of the deep actor-critic learner as the buyer (compared to other well-known traditional negotiation techniques) is studied in the case study.

6.1.2. The seller agent

In this paper, the seller concedes its payoff by its own strategy. The buyer does not have the information about the function of the negotiation strategy of the seller and tries to recognize the type of the seller only based on the available historical data of the previous rounds of negotiation. To this end, a K-means clustering is designed to help the buyer to cluster the seller's conceding behavior into three classes (types): boulware, normal, and conceder behavior. The boulware agents look towards to the maximum profits. Therefore, they maintain the offered value until the deadline is near up. The conceder agents normally like to make deal with opponents as soon as possible. So the conceder agents go to their reservation value very quickly (Faratin et al., 1998). If the

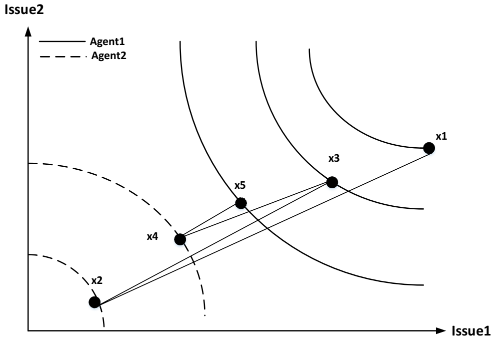


Figure 5: The shortest distance algorithm that Sycara introduces in (Lai et al., 2008b) for proposing the payoff in each step of negotiation.

concession rate of an agent is rather smooth during the negotiation process, we call this type of agent the normal one. The normal agents have an intermediate behavior between boulware and conceder agents (Ho and Zhou, 2008). Contrary to the previous articles, those have considered a fixed behavior for the seller during negotiation, in this paper; the seller can also change its negotiation strategy, depending on the state of the negotiation.

6.2. Proposing

In the proposing part, the agent determines the offer to be proposed to the opponent in each step of negotiation. In a multi-attribute negotiation, usually there exist many proposals with the same payoff for the agent (the indifferent curve). Therefore, the agent needs to follow an effective strategy to select a point from the indifference curve and propose it to the opponent. In this paper we have used the method by Sycara introduces in (Lai et al., 2008b). The heuristic in this approach is that the agent chooses a point from its current indifference curve which has the shortest distance to the previous offer made by the opponent. The intuition behind this heuristic is that such an offer has a higher probability to be accepted by the opponent in comparison with the other offers on the indifference curve.

Fig. 5 presents an example of a two-attribute (Issue1 and Issue2) negotiation where the agents follow the shortest distance strategy to the opponent. The dashed curves are the indifferent curves of agent 2 and the solid curves are the indifferent curves of agent 1. In the first step, agent 1 makes the offer x_1, but agent 2 rejects it. Then in the second step, agent 2 finds an offer x_2 which has the shortest distance to the point x_1 on its indifferent curve. Agent 1 rejects this offer and it concedes to the second left solid curve in the third step. It finds offer x_3 which is the closest one to x_2 on this curve. This procedure continues until an agreement or the deadline is met. The proposed strategy is shown in Fig. 5 and formulized as follows:

$$x_b^t = \arg\min_{x \in c} \left\| x - x_s^{t-1} \right\| \tag{24}$$

where c is the set of offers that have the special payoff (indifferent curve), x_s^{t-1} is the sellers offer that proposed at $t-1$, x_b^t is the buyers offer that proposes at t and $\|x-y\|$ represents the distance between x and y.

6.3. Responding

The agent decides to accept or reject the opponent's offer by comparing the opponent's offer with its own offer in the next step. In this kind of responding strategy, the agent accepts the opponent's offer, if the payoff gained from the opponent's offer is greater than or equal to the payoff which is expected from the acceptance of its offer in the next step. Otherwise, the agent rejects the opponent's offer and proposes a new offer to the opponent. Moreover, the agent leaves the negotiation, if the deadline is met. This responding strategy is formulized as follows:

$$a_b^t = \begin{cases} \text{accept} & \text{if } payoff_b(x_b^{t+1}) < payoff_b(x_s^t) \\ \text{reject} & \text{if } payoff_b(x_b^{t+1}) > payoff_b(x_s^t) \\ \text{leave the negotiation} & \text{if } payoff_b(x_b^{t+1}) > payoff_b(x_s^t) \;\; and \;\; t > deadline_b \end{cases} \tag{25}$$

where t is the time that the seller proposes an offer to the buyer, a_b^t is the buyer's decision about the seller's offer, $payoff_b(x_s^t)$ is the payoff of the buyer from the seller's offer at time step t and $payoff_b(x_b^{t+1})$ is the payoff of the buyer from its offer at step $t+1$ and $deadline_b$ is the deadline of the buyer.

6.4. The mediator

To speed up the learning process, a mediator is implemented to exclude the offers that are less beneficial for both of the agents. Most of the previous studies about a mediator in the negotiation can limit the authority of the agents in decision-making by manipulating the agents' offers or needing private information of the agents (Chen et al., 2014; Lai et al., 2006; Ehtamo et al., 1999). But in this paper, the mediator does not manipulate the agents' offers. Moreover, it does not enquire any information about the agents' strategies and payoff functions.

At each round of negotiation which is equivalent to a learning episode for the buyer agent, the seller and buyer inform the mediator about all of the offers/counteroffers they have proposed/received to/from the opponent and their corresponding payoffs. Among those, for any offer/counteroffer if there exist at least another offer/counteroffer with larger payoff for both of the agents, then the mediator removes that dominated offer/counteroffer from feasible strategy set of the agents. In this way, in each round, the strategy space of the agents are reduced to more effective one, which leads to faster learning process of the Pareto optimal solution.

6.5. Pareto optimality and Negotiation Power Assessment

At the end of the learning process, by comparing the agreement offer with Pareto frontier we expect that, if the learning process is successful, the agreement occurs close to Pareto frontier. After that, the bargaining power of the agents can be determined using the concept of the Nash bargaining solution with the asymmetric negotiation power which was explained in Section 3.2. In other words, we can arrange an equivalency between the result of the negotiation process with incomplete information and the corresponding complete information bargaining. Through this, the relationship between the bargaining power and the agents' behavior during negotiation can be assessed. This argument is studied in the case study in Section 8.

The proposed methodology for bilateral negotiation is summarized in Algorithm 1.

7. Learning Negotiation through Reinforcement

As mentioned in Section 3 states, action, and rewards play an important role in the DRL method. This section provides meaningful definitions for states, action, and rewards to utilize the DRL method for the buyer in the negotiation process.
In the previous works, when RL in the negotiation is employed (Takadama, 2008; Lao and Zhong, 2010) the last offer of the opponent is used as the state of the agent. However, it's not a comprehensive state definition and moreover, adding more factors into the state definition results in the curse of dimensionality. To solve this problem, Hajimiri (Hajimiri et al., 2014) used fuzzy RL. The fuzzy feature of the proposed method helps the agent to select three different factors as the elements of the state vector. Nonetheless, it is possible to further improve by adding more factors than three elements to the state vector. With this goal, the paper uses deep auto-encoder. Deep auto-encoder facilitates the inclusion of more factors into the state definition, without emerging the curse of dimensionality (Maldonado-Ramirez et al., 2018). We include four categories of elements as the deep auto-encoder input which are explained as follows:

- Time: The agreement should occur before the deadline. Moreover, the payoff is commonly discounted over time, which makes the time an important factor that should be taken into account in constructing the state of the agent. Hence, both the number of steps passed from the starting of the negotiation and the number of steps left until the deadline is considered as the auto-encoder input.

- Reservation payoff: The distance between the buyer's payoff out of its current offer and its reservation payoff (RP) is considered as another input of the deep auto-encoder. This value determines the flexibility of the buyer agent for the next concessions/offers.

- opponent's offer: The history of the opponent offers is one of the useful elements in the decision making of the agent. There are different representations of the opponent's offer that the agent can use as a state vector. In this paper, some of these representations that could do more to help the agent decision making has been considered as the input of deep auto-encoder. These representations are as follows:

Algorithm 1 The proposed methodology for bilateral negotiation

```
Input: feasible set of offers
Output: payoff$_{buyer}$,payoff$_{seller}$,agreement offer, bargaining power,feasible set of offers
  $t \leftarrow 1$
  $c_s \leftarrow \{\varnothing\}$
  $k \leftarrow a_b^1(x_s^1)$
  if $k == accept$ then
      $payoff_{buyer} \leftarrow payoff_b(x_s^1)$
      $payoff_{seller} \leftarrow payoff_s(x_s^1)$
      $agreement$ offer $\leftarrow x_s^1$
      $Go$ to the Game Analysis
  else if $k == leave$ the negotiation then
      $payoff_{buyer} \leftarrow$ punishment of not reaching the reaching the agreement
      $payoff_{seller} \leftarrow$ punishment of not reaching the reaching the agreement
      $agreement$ offer $\leftarrow \varnothing$
      $Go$ to the Game Analysis
  end if
  while $k == reject$ do
      $t \leftarrow t + 1$
      $sellers$ behavior $\leftarrow k - means$ clustering($x_s^t$)
      $u_b^t \leftarrow$ deep RL function($current$ state)
      $x_b^t \leftarrow \arg\min_{x \in c} \left\| x - x_s^{t-1} \right\|$ c is the set of offers that have the utility equal to u$_b^t$
      $G \leftarrow a_s^t(x_b^t)$
      if $G == accept$ then If $G == accept$
          $payoff_{buyer} \leftarrow payoff_b(x_b^t)$
          $payoff_{seller} \leftarrow payoff_s(x_b^t)$
          $agreement$ offer $\leftarrow x_b^t$
      else if $G == leave$ the negotiation then
          $payoff_{buyer} \leftarrow$ punishment of not reaching the reaching the agreement
          $payoff_{seller} \leftarrow$ punishment of not reaching the reaching the agreement
          $agreement$ offer $\leftarrow \varnothing$
      else if $G == reject$ then
          $t \leftarrow t + 1$
          $u_s^t \leftarrow seller's$ conceding function($t$)
          $x_s^t \leftarrow \arg\min_{x \in c} \left\| x - x_b^{t-1} \right\|$ c is the set of offers that have the utility equal to u$_s^t$
          $k \leftarrow a_b^t(x_s^t)$
          if $k == accept$ then
              $payoff_{buyer} \leftarrow payoff_b(x_s^t)$
              $payoff_{seller} \leftarrow payoff_s(x_s^t)$
              $agreement$ offer $\leftarrow x_s^t$
          else if $k == leave$ the negotiation then
              $payoff_{buyer} \leftarrow$ punishment of not reaching the reaching the agreement
              $payoff_{seller} \leftarrow$ punishment of not reaching the reaching the agreement
              $agreement$ offer $\leftarrow \varnothing$
          end if
      end if
      feasible set of offers $\leftarrow$ function for excluding undesired offer from feasible set (Algorithm ??)
  end while
  $Game$ Analysis
  Bargaining power $\leftarrow$ arg $\underset{Bargaining \text{ power}}{equal}$ ($Nash$ Bargaining Solution(Bargaining power)=agreement offer)
```

  - The distance between the buyer's payoff from its offer and the buyer's payoff from the seller's offer is an effective element in the decision making of the buyer. Based on the value of this distance and the time of the negotiation, the buyer may change its behavior to concede more or less in the current step.
  - The distance between the buyer's payoff from the two consecutive seller's offers is considered as another effective element in the auto-encoder input. Based on this distance in different steps of negotiation, the buyer can learn the behavior of the seller and predict the seller's next offer.
  - The distance between the two successive seller's offers of each attribute is intended as another input of auto-encoder. As the above element, this distance in different steps of negotiation helps the buyer to learn the behavior of the seller.

- Seller's behavior: The type of the seller which is determined by K-means is considered as another input of auto-encoder. This element represents the behavior of the seller

Therefore, the input vector of deep auto-encoder is determined as follows.

$$\begin{aligned} s_t = [&t, deadline_b - t, \|payoff_b(x_b^{t-2}) - RP\|, \|payoff_b(x_b^{t-2}) - payoff_b(x_s^{t-1})\|, \\ &\|payoff_b(x_s^3) - payoff_b(x_s^1)\|, ..., \|payoff_b(x_s^{t-1}) - payoff_b(x_s^{t-3})\|, \\ &\|x_s^3 - x_s^1\|, ..., \|x_s^{t-1} - x_s^{t-3}\|, class(seller)] \end{aligned} \tag{26}$$

where t and $deadline_b$ is the current time-step of the negotiation and the deadline of the buyer, Respectively. x_s^{t-1} is the sellers offer that proposed at $t-1$, $payoff_b(x_s^{t-1})$ is the payoff the buyer from the sellers offer at time step $t-1$ and $payoff_b(x_b^{t-2})$ is the payoff the buyer from its offer at time step $t-2$, $\|x-y\|$ represents the distance between x and y. $class(seller)$ refers to the class of the opponent which is determined by K-means clustering. The agent uses deep auto-encoder output (low dimensional data) as its state and decides on its action according to its state. The action of the buyer in the actor-critic is defined as the amount of concession payoff. When the buyer proposes an offer to the seller, the seller can respond in three different ways: it can accept the offer, reject the offer and respond to it with a counter-offer, or reject the offer and leaves the negotiation table because the deadline is passed. The learning algorithm should assign appropriate rewards/punishments with respect to these three incidents. In this paper, three different scenarios are provided for the definition of rewards/punishments as follows.

- In the first scenario, the algorithm assigns a reward proportional to the agreement payoff, just at the end of the negotiation. Moreover, if the deadline is passed and the buyer loses the negotiation, the agent gets a big punishment. This scenario is formulized as follows:

$$\text{Reward} = \begin{cases} payoff_b(x_{agreement}) & \text{if agents reach an agreement} \\ -\alpha.payoff_b(x_b^{t=deadline}) & \text{if t=deadline and agents can not reach an agreement} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

where $payoff_b(x_{agreement})$ is the payoff of the buyer from the agreement offer. $payoff_b(x_b^{t=deadline})$ is the payoff of the buyer from its offer at the deadline. $\alpha > 10$ is a parameter that determines the power of punishment when the buyer loses the negotiation.

- In the second scenario, besides the reward that the buyer receives at the end of the negotiation, it gets a minor reward based on the suitability proposed offer in each step of negotiation. The same as the first scenario, if the deadline is passed and the agent cannot reach an agreement, it gets a big punishment.

$$\text{Reward} = \begin{cases} payoff_b(x_{agreement}) & \text{if agents reach an agreement} \\ -\alpha.payoff_b(x_b^{t=deadline}) & \text{if t=deadline and agents can not reach an agreement} \\ \gamma.payoff_b(x_b^t) & \text{otherwise} \end{cases} \tag{28}$$

where $\gamma < 1$ is a discount factor that specifies the amount of the reward in each step of negotiation.

- In the last scenario, in addition to the two mentioned kinds of rewards and punishments, the buyer gets a minor punishment as a result of postponing the negotiation to the next step. Because in this situation, the agent still has a chance to reach an agreement, the assigned punishment should not be as high as the punishment of situations in which the deadline is passed and the agent cannot reach an agreement.

$$\text{Reward} = \begin{cases} payoff_b(x_{agreement}) & \text{if agents reach an agreement} \\ -\alpha.payoff_b(x_b^{t=deadline}) & \text{if t=deadline and agents can not reach an agreement} \\ \gamma.payoff_b(x_b^t) \ - \mu.\text{t} & \text{otherwise} \end{cases} \tag{29}$$

where $\mu > 1$ is a parameter that determines the amount of the punishment the agent receives as a result of postponing the negotiation to the next step.

These three different scenarios are compared to each other in the case study.

8. Case study

This section presents a case study for bilateral contracts of energy to show the applicability of the proposed negotiation method. Assume that we have a pair of generator and load as the seller and buyer, respectively. In this situation, there are two methods to purchase/sell electrical energy in the wholesale market. The first one is through the electricity pool, on which the generators and loads submit their bids to the market and then, the independent system operator (ISO) clears the price and power to be exchanged in the grid. In the second approach, each pair of generator and load can sign a bilateral contract of energy (Kirschen and Strbac, 2004a). That is an agreement between two parties to exchange electric power under a set of specified attributes like price, quantity, penalty to be paid by the generator to the load if the generator declines to supply the agreed power and penalty to be paid by the load if it withdraws the contract (Chung et al., 2003). In this kind of contract, the parties have more control over the cleared attributes which is a hedge against the risks in the spot electricity market. In many power grids the major share of power exchange is covered by bilateral contracts; e.g., in PJM bilateral contracts account for about twice the volume traded in its spot market, and in Britain's NETA system they are nearly 100% (Chao et al., 2008). Therefore, the buyer and seller need to learn their optimal negotiation strategies in bilateral contracts to maximize their profit. Further, since such negotiation involves multiple attributes and also incomplete information of the seller and buyer from each other, the proposed negotiation method in this paper can be used as an effective tools to deal with this problem.

8.1. Two attribute negotiation

To implement this negotiation, first, we assume the parties negotiate on two attributes including power and price. The cost function of the generator as the seller and the utility function of the load as the buyer in equation 8 and 9 are defined as follows (Samadi et al., 2010; Huang et al., 2014):

$$c(q) = a_1.q^2 + b_1.q + c_1 \tag{30}$$

$$u(q) = -a_2.q^2 + b_2.q + c_2 \tag{31}$$

where $a_i, b_i, c_i > 0$ are the coefficients of the cost function and the utility function of the seller and the buyer.

In the simulations, the generator's offer is generated using the time-dependent strategy at each step of negotiation. Nevertheless, the load is not aware of the strategy of the generator. In the time-dependent strategy, the acceptable value of payoff for an agent to reach an agreement changes from its maximum value to the reservation value according to the following time-dependent function:

$$payoff(t) = \alpha(t).ru + [1 - \alpha(t)].mu \tag{32}$$

$$\alpha(t) = (\frac{t}{T})^{\frac{1}{\beta}} \tag{33}$$

where $payoff(t)$ is the target payoff of the generator at time t, ru is the reservation utility obtained using the marginal price of the generator, mu is the maximum utility, t is the current time and T is the negotiation deadline. β represents the behavior of the agent from boulware to conceder. In time-dependent strategy, constant and the time-varying strategies are called fixed time-dependent strategy and variable time-dependent strategy, respectively.

8.1.1. Learning analysis

To show the learning process of the DRL, the negotiations are simulated in two different situations. In the first situation, the load negotiates with the generators that concede using the fixed time-dependent strategy in each round. In this situation, when the load is faced with a new generator, based on the initial clustering, it determines the class of the generator and then negotiates with the generator using the deep actor-critic algorithm. In the second situation, the load negotiates with the generators concede by the variable time-dependent strategy in each round. In this strategy, the generator's behavior changes during the negotiation. In this situation when the load faces a new generator it determines the class of the generator and negotiates until the special step of negotiation. The load will then revise the class of the generator by considering the new generator's behavior and continue the negotiation. This process will be performed once again until the end of the negotiation.

To compare the learning process of the buyer in the three scenarios described in Section 7, three different markets are simulated. The load is rewarded in each market according to one of the scenarios. If the generator concedes using a fixed time-dependent strategy, each market will be 1000 rounds of negotiations. Otherwise, when the generator concedes by a variable time-dependent strategy, each market will be 4000 rounds of negotiations. For a more accurate comparison, each market is run 10 times with the same stored data, the average payoff of these 10 times is considered as the mean payoff. The load has a different behavior with the different classes of fixed time-dependent generators. Therefore, to see the progress of learning it's better to analyze the learning progress of the load when dealing with each class of the generator separately. In 1000 rounds of negotiations, the load determines 320, 300 and 380 generators as the conceder, normal and the boulware generators, respectively. The average of the load's payoffs in the negotiation with the 1000 fixed time-dependent generators in each market is shown in Fig. 6.

To analysis the learning progress when the load negotiates with the variable time-dependent generators, we divide generators into 24 categories and observe the learning progress of the load in dealing with each category separately. The number of each generator in each category in the 4000 rounds that determined by the load and the load's payoff in the last round when it is rewarded by the various scenarios outlined in Table 1.

The simulation results show that the load performs better when it is rewarded by the third scenario (the load receives the reward and the punishment in each step of negotiation). This is because of the minor reward and the punishment in each step of the negotiation encourage the load to reach an agreement as quickly as possible. Moreover, due to the low complexity of the fixed time-dependent strategy in comparison with the variable time-dependent strategies, the load can learn better and faster when faced with a generator that has a fixed time-dependent strategy. Therefore, as illustrated in Fig. 6 and Table 1, the average payoff of the load when the generators concede by the fixed time-dependent strategy is more than when the generators concede by variable

Table 1: The average of the achieved payoff when the load negotiates with the different class of the variable time-dependent generator and rewarded by various scenarios.

	First behavior	Second behavior	Third behavior	Number of each category in 4000 rounds	Loads's payoff in the last round when it rewarded by the first scenario	Load's payoff in the last round when it rewarded by the second scenario	Load's payoff in the last round when it rewarded by the third scenario
1	boulware	boulware	conceder	152	830	831	843
2	boulware	boulware	Normal	203	816	825	830
3	normal	normal	boulware	148	841	848	850
4	normal	normal	conceder	184	861	869	881
5	conceder	conceder	Normal	153	876	880	889
6	conceder	conceder	boulware	180	862	876	880
7	conceder	boulware	boulware	152	832	834	861
8	normal	boulware	boulware	173	852	854	838
9	boulware	normal	Normal	200	852	861	869
10	conceder	normal	Normal	137	868	871	882
11	normal	conceder	conceder	164	881	883	891
12	boulware	conceder	conceder	180	862	876	880
13	boulware	conceder	boulware	144	831	836	841
14	boulware	normal	boulware	162	811	820	832
15	normal	boulware	Normal	143	832	828	841
16	normal	conceder	Normal	167	858	861	872
17	conceder	normal	conceder	170	863	869	880
18	conceder	boulware	conceder	178	862	865	869
19	boulware	normal	conceder	195	860	866	873
20	boulware	conceder	Normal	150	835	843	850
21	normal	conceder	boulware	175	852	857	861
22	normal	boulware	conceder	162	830	841	849
23	conceder	boulware	Normal	158	846	849	861
24	conceder	normal	boulware	170	833	848	852

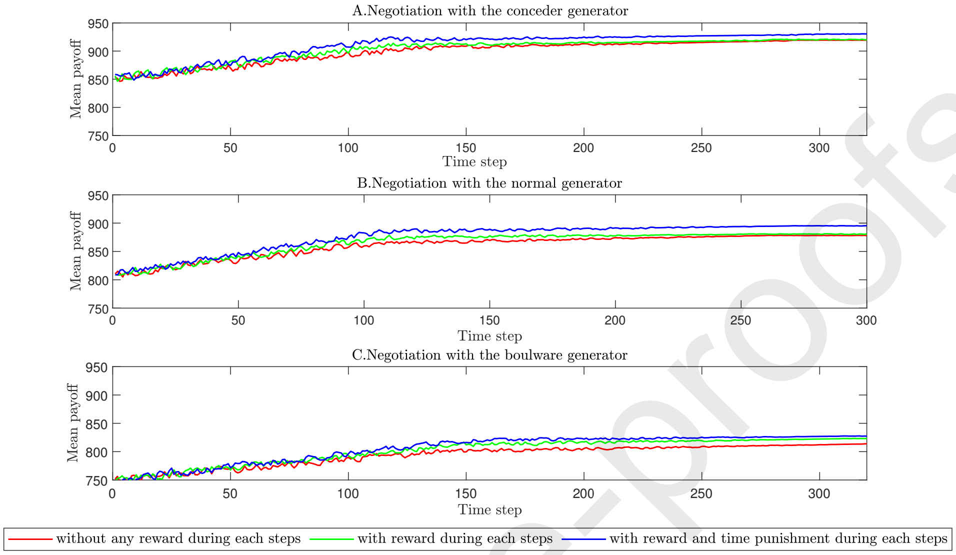


Figure 6: The average of the achieved payoff when the load negotiates with the different class of fixed time-dependent generators and rewarded by various scenarios

time-dependent strategies.

The variance of payoffs in 1000 rounds for the fixed time-dependent generators and 4000 rounds for variable time-dependent generators are shown in Fig. 7 and Fig. 8. According to these figures, as the learning process continues, the variance of payoffs gradually reduces.

The negotiation procedures for both the fixed time-dependent and the variable time-dependent generators are presented in Fig. 9 and Fig. 10, respectively. The curves show the indifference curves of the agents. The stars marked on the indifference curve are the points that the agents propose as the offer in each step of negotiation. The green circles are the points that the mediator excludes them from the feasible set of offers during all rounds with the algorithm **??**. Finally, the red bold point is the agreement offer. As shown in Fig. 10, the generator changes its behavior after step 3 from the boulware to the conceder because the generator thinks that it cannot reach an agreement by continuing this process.

8.1.2. Game-Theory analysis

In this section, we will compare the results of the negotiation process with incomplete information and complete information. First, it is shown that the agreement of the negotiation with incomplete information occurs close to the Pareto frontier. Then, the relationship between the bargaining power of the agents and the behavior of the agents is studied. To this aim, the agreements resulted from the DRL method and the Nash bargaining solution are compared

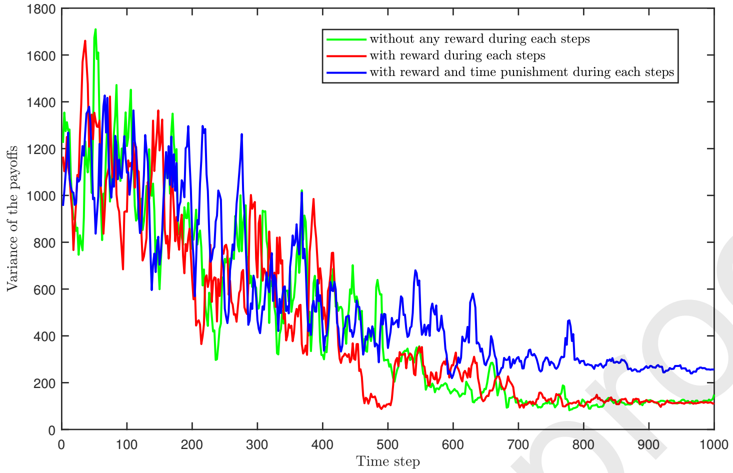


Figure 7: The variance of the achieved payoff for the fixed time-dependent generators when the generator is rewarded by the various scenarios

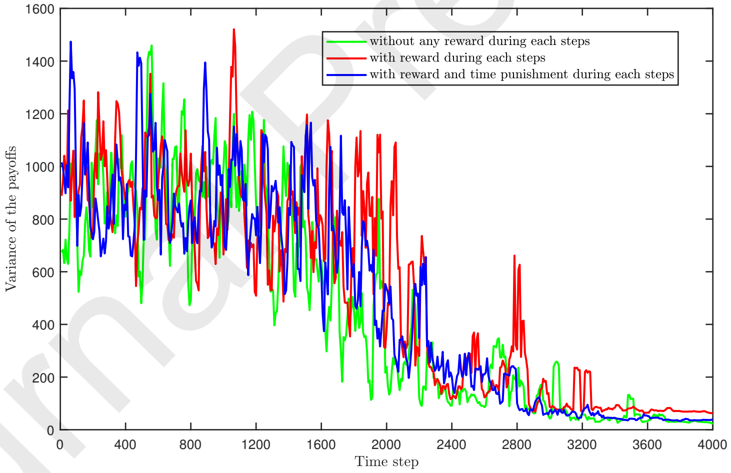


Figure 8: The variance of the achieved payoff for variable time-dependent generators when the load is rewarded by the various scenarios

and the bargaining power of the generators is assessed.

After 1000 rounds of learning negotiations, the results of the negotiation with the normal, conceder and the boulware generators are studied, separately. Blue stars in Fig.11 represent the joint load's and the generator's feasible payoff set. The outer boundary of blue stars represents the Pareto frontier, which means that no agent can make its payoff better without making the opponent's payoff worse off. As shown in Fig.11, the payoffs of all three negotiations are close

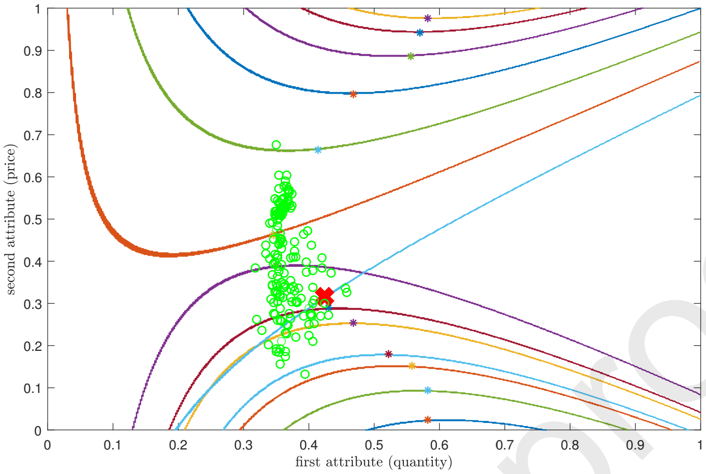


Figure 9: The negotiation procedure for fixed time-dependent generator and Learner load

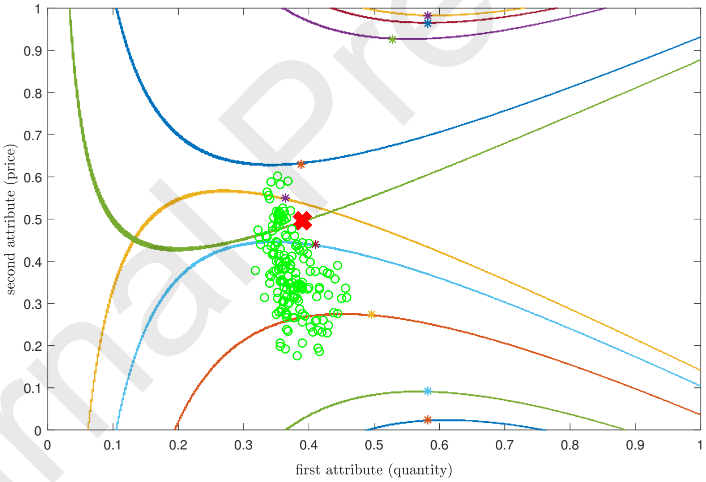


Figure 10: The negotiation procedure for variable time-dependent generator and Learner load

to the Pareto frontier curve. Therefore, we can verify that the DRL method maintains the Pareto optimality.

For more accurate studies, different generators negotiate with the learner (the load that uses the deep actor-critic algorithm) 400 rounds of negotiations. By comparing the generator's and the load's payoffs (achieved by the DRL method) with that from the equation 7 for complete information bargaining, the equivalent bargaining power of the agents for the Nash bargaining solution for each round of negotiation is determined. If the agents cannot reach an agreement before the deadline, we determine the bargaining power of the agents by

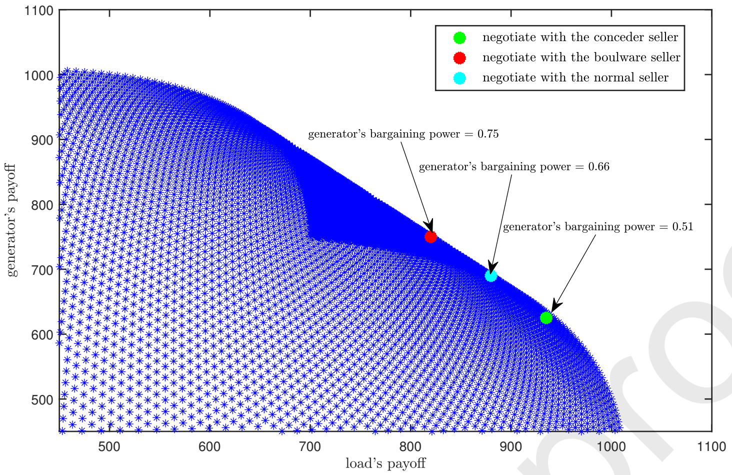


Figure 11: The load's and the generator's payoff and the Pareto frontier of negotiation

Table 2: The payoff and the bargaining power that fits the different groups of fixed time-dependent generators

Type of generator	Generator's accumulated payoff	Load's accumulated payoff	Bargaining power of generator	Bargaining power of load (learner)
boulware	217573	261049	0.75	0.25
normal	263829	297753	0.62	0.38
conceder	236885	318398	0.51	0.49

taking the time to reach an agreement. The average of the bargaining power in 400 rounds of negotiations is considered as the bargaining power that fits the agent's behavior. The results are shown in Table2.

As shown in Table 2, the different behavior of the generator causes an asymmetry in the bargaining procedure and therefore it causes a change in the bargaining power of the agents. The analysis indicates that although the boulware generator (whose concessions are only made when the deadline is almost reached, otherwise the proposals are only slightly changed) has more bargaining power than the conceder generator, it does not gain the highest accumulated payoff in comparison with the other generators, as it cannot reach an agreement most of the time and therefore, faced with a big punishment. In contrast, the conceder generator reaches an agreement in most of the negotiations, but due to conceding during the primary steps of the negotiation, the payoff of the agreement offer and its bargaining power is not noticeable. The normal generator has a tradeoff between gaining a higher utility and reaching an agreement before the deadline. Therefore, although it does not have the most bargaining power, it gains the highest accumulated payoff.

8.1.3. Comparison

To show the superiority of the deep actor-critic in comparison with the other conceding algorithms, that are commonly used in literature (Lai et al., 2006, 2008b; Hajimiri et al., 2014; Takadama, 2008), seven different markets (each market consisting of 1000 rounds of negotiations) are simulated. In the first, second, and third markets, the load uses the fixed time-dependent algorithm respectively as the conceder, the boulware and the normal (Lai et al., 2008b). In the fourth to eighth markets, the load learns how to concede its payoff by different kinds of Rl methods. In the fourth market, the load learns its strategy by Q-learning method (Takadama, 2008). In this market, the last offer of the opponent is considered as the only element of the state vector in the Q-learning method. In the fifth and the sixth markets, the load learns how to concede its payoff by the proposed FSLA method introduces in (Hajimiri et al., 2014). A distance between the last offer of the agents, the reservation price and time are three elements that define the state vector. The main difference between these two markets is in the presence of the mediator. There is a mediator(as introduced in (Lai et al., 2006)) in the sixth market that changes the users' offers in such a way that the agreement offer converges to the Pareto frontier (Lai et al., 2006), but there is no mediator in the fifth market. Finally, in the last market, the load learns how to concede its payoff in each step of negotiation using the deep actor-critic algorithm proposed in this paper. In this case, all effective elements are defined as the deep auto-encoder inputs, and the deep auto-encoder makes the state vector of the actor-critic by decreasing these inputs' dimensions. In all of these markets, the load proposes the offer from the indifference curve by the shortest distance algorithm (Lai et al., 2008b).

For a more accurate comparison, the load negotiates with the same group of generators concede by the fixed time-dependent strategy in all seven markets. The accumulative payoff due to each of the markets in the 1000 rounds is calculated. Each of these markets is run 5 times with the same stored data. The average of the accumulative payoffs of the seventh different markets in the 1000 rounds is shown in Fig.12.

As shown in Fig.12, the accumulative payoff of the learner who uses the deep actor-critic algorithm is more than the other buyers. The conceder agent concedes a lot during the primary steps, so although it reaches an agreement in most of the negotiations, due to conceding during the primary steps of the negotiation the payoff of the agreement offer is not noticeable. The boulware agent has the weakest performance in comparison with the other agents that use the time-dependent algorithm; as a result of not conceding during the primary steps of negotiation. Although there is a higher payoff when agents reach an agreement, they cannot always reach an agreement most of the time and therefore faced the big punishment. The learner (the load in the forth to the seventh markets) keeps the balance between gaining higher utility and reaching an agreement before the deadline. After almost 200 rounds, it learns how to negotiate such that

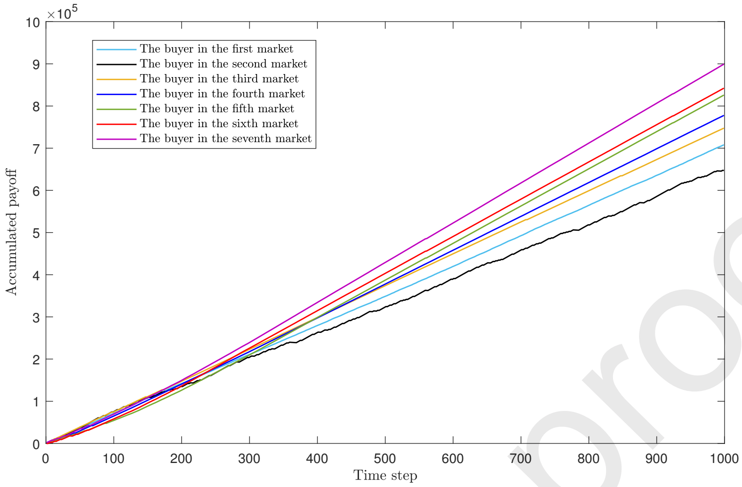


Figure 12: The accumulative payoff of seven different markets

it can receive a noticeable payoff, and is able to reach an agreement in most negotiations. Therefore, the accumulative payoff curves of the learner in the forth to the seventh markets are above the other agents' curves.

Comparing the fourth to the seventh markets (the markets where the load proposes an offer by the RL methods), in the markets 5 and 6, since the agent uses FSLA method, it can increase the number of elements in its state vector to more than one, therefore, other effective elements would be placed in the state vector. This leads to better cumulative payoff for the learner in markets 5 and 6, comparing to market 4, which has only one element in its state vector. However, using the FSLA method cannot increase the state vector dimensions to any optional value; thus, in the market seven auto-encoder is used to place all the effective elements in the state vector. That's why the cumulative payoff of the learner at the market 7 is higher than the rest of the markets. In this market, all effective elements are defined as the auto-encoder inputs, and the auto-encoder decreases these inputs' dimensions and fed them to the system as the state vector.

A statistical T-test was used to ensure the competency of the deep actor-critic learner against the other buyers in all six markets. The results show that there is a relatively significant statistical difference between the performance of the deep actor-critic learner and the other loads. The accumulative payoff of the deep actor-critic learner compared to other loads in six markets studies has a confidence percentage exceeding 99 % .

In Table 3 The average of accumulative payoff and the bargaining power of these seven different markets are shown. In each market, the load negotiates with the same group of generators (concede by the fixed time-dependent

Table 3: The payoff and the bargaining power that fits the different type of loads in different markets

Number of market	Type of load	Bargaining power that fits the load's behavior	Load's accumulated payoff
1	Conceder	0.34	251135
2	Boulware	0.43	225940
3	Normal	0.39	269889
4	Learner proposed in (Takadama, 2008)	0.43	278022
5	Learner proposed in (Lai et al., 2008b; Hajimiri et al., 2014)	0.45	283710
6	Learner proposed in (Lai et al., 2006; Hajimiri et al., 2014)	0.45	297211
7	Learner (deep actor-critic)	0.47	326032

strategy) 400 rounds of negotiations.

As shown in Table 3, when the buyer uses the fixed time-dependent algorithm, it earns the most accumulative payoff if it has a normal behavior. As explained above, due to the trade off between gaining higher utility and reaching an agreement before the deadline, the normal load earns the most accumulative payoff, although it does not have the highest bargaining power. When the load uses the RL algorithm for conceding (market 4 to 7), it learns how to concede the payoff and can keep a balance between gaining higher utility and reaching an agreement before the deadline. In this regard, the proper definition for the state vector has a significant impact on the learning process and as mentioned in the 8.1.1, the load in Market 7 has the most comprehensive state vector. Therefore, it gains a higher accumulative payoff and bargaining power than when it uses the other conceding algorithms.

8.2. Multi attribute negotiation

To generalize the negotiation to more than two attributes, three different scenarios are provided as follows.

- In the first scenario, the load and generator negotiate two attributes including the power quantity and the price.

- In the second scenario, the load and generator negotiate three attributes including the power quantity, the price and the penalty to be paid by the generator to the load if the generator declines to supply the agreed power in the contract. In this case, the payoff functions of the load and generator can modeled , respectively as follows (Chung et al., 2003; Kirschen and Strbac, 2004b):

$$payoff_l(p,q) = (1-\alpha).(u(q) - p.q) + \alpha.k_g.q \tag{34}$$

$$payoff_g(p,q) = (1-\beta).(p.q - c(q)) + \beta.(p_g^m.q - k_g.q - c(q)) \tag{35}$$

where p and q are the price and the quantity of energy, respectively. k_g is the monetary penalty per unit of energy to be paid by the generation to the load if the generator decline to supply the contract. α and β are the load and generator's forecast about the probability of the generator to contract cancellation. p_g^m is the generator's forecast about the price of energy in the market when it decides to sell its energy to the market instead of sells to the load.

- In the third scenario, we add the penalty to be paid by the load if it withdraws the contract as the fourth attribute. In this case, the payoff function of the load and generator are modeled as follows (Chung et al., 2003; Kirschen and Strbac, 2004b):

$$payoff_l(p,q) = (1-\alpha_1-\alpha_2).(u(q)-p.q)+\alpha_1.k_g.q+\alpha_2.(u(q)-p_l^m.q-k_l.q) \quad (36)$$

$$payoff_g(p,q) = (1-\beta_1-\beta_2).(p.q-c(q))+\beta_1.(p_g^m.q-k_g.q-c(q))+\beta_2.k_l.q \quad (37)$$

  where k_l is the monetary penalty per unit of energy to be paid by the load to the generator if the load decline to buy the energy from generator. α_1 and α_2 are the load forecast about the probability of the generator and the load to contract cancellation, respectively. β_1 and β_2 are the generator forecast about the probability of the generator and the load to terminate the contract, respectively. p_l^m are the load's forecast about the price of energy in the market when it decides to buy its energy from the market.

To show the generalization of negotiation to more than two attributes, the aforementioned three scenarios are compared to each other. 1000 generators with different time-dependent behavior are considered and the average load's payoff in each negotiation scenario is shown in Fig.13. In 1000 round of negotiation, the load determines 320,300 and 380 generators as the concider, normal and boulware generator, respectively. To better compare the learning speed of the three different scenarios in Fig. 13, each market's reward is normalized. The simulation results show that as the number of attributes increases the buyer learns more slowly. This is because by adding attribute to the negotiation, the complexity of the problem increases. However, as shown in Fig. 13, the load learns how to bid to maximize its payoff in all of the three scenarios.

9. Conclusion

This paper firstly proposes an intelligent buyer (using the DRL approach) to negotiate on the multi-attribute (the price and the quantity) simultaneously under incomplete information. In the negotiation protocol, the proposer suggests an offer to the responder. The responder then reacts to that offer by either accepting or rejecting the buyer's proposal. If the responder accepts the offer, the negotiation comes to an end; otherwise, the agents will exchange their roles and the negotiation proceeds to the next step. In this paper, the seller uses a conceding strategy in which depending on the negotiation conditions, its

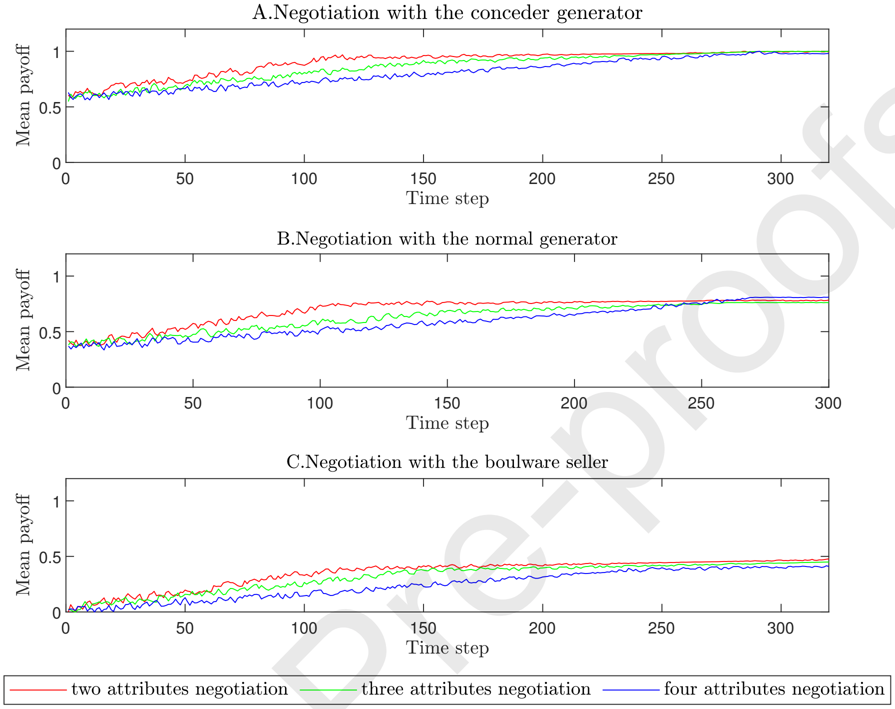


Figure 13: The average of the achieved payoff when the load negotiates with the different classes of fixed time-dependent generators on various attributes.

behavior can be changed during the negotiation. While the buyer learns how to concede each step to maximize its long-term payoff using the deep actor-critic algorithm. The existence of deep auto-encoder in this algorithm and consequently the ability to consider the comprehensive and appropriate definition of the state vector, rewards and action will result in the buyer (learner) to achieve higher payoff during negotiations. Considering the buyer has no information about the seller's behavior, a K-means clustering is designed to help the buyer to determine the type of the seller. To speed up the learning process, we have also used a mediator to exclude the unreasonable offers from the feasible set of negotiation offers. The superior performance of the deep actor-critic learner is illustrated through different market simulation studies and statistical T-test (the buyer uses different conceding algorithms in each market). We have also discussed the effects of the seller's conceding strategy and assigning the appropriate reward functions in the deep actor-critic algorithm for the learning process. Moreover, we generalize the negotiation to more than two attributes in the case study.

Finally, our research presents a useful Game Theory analysis based on the empirical results to compare the result of the negotiation process with incomplete information and complete information. First, the paper proves the existence and the uniqueness of the Nash bargaining solution for the proposed negotiation with complete information. The experimental results show that this model can help agents reach a Pareto-efficient agreement very closely. Moreover, the relationship between the bargaining power of the agents and the behavior of the agents is studied. By reviewing the behavior of different sellers, it became clear that although the boulware seller has more bargaining power than the conceder seller, it does not gain the highest accumulated payoff in comparison with other sellers. This is because the boulware seller cannot reach an agreement in most of the negotiations. The highest accumulated payoff can be received by the normal seller that has a tradeoff between gaining higher payoff and reaching an agreement before the deadline. To show the ability of the deep actor-critic in comparison with the other conceding algorithm in literature, buyers with different behaviors negotiate with the same group of sellers. Simulation results show the superiority of the deep actor-critic learner performance in terms of accumulated payoff and power factor compared to other buyers.

Acknowledgments

This research in part supported by a grant from the Institute for Research in Fundamental Sciences (IPM) under grant number: CS 1397-4-56.

References

References

T. Baarslag, M. J. Hendrikx, K. V. Hindriks, and C. M. Jonker. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. 30(5):849–898, 2016. ISSN 15737454. doi: 10.1007/s10458-015-9309-1.

J.-G. Baek and C. O. Kim. Learning single-issue negotiation strategies using hierarchical clustering method. *Expert Systems with Applications*, 32(2):606 – 615, 2007. ISSN 0957-4174.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, Sep. 1983. doi: 10.1109/TSMC.1983.6313077.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, 2007.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013. doi: 10.1109/TPAMI.2013.50.

K. Binmore, A. Rubinstein, and A. Wolinsky. The Nash Bargaining Solution in Economic Modelling. *The RAND Journal of Economics*, 17(2):176, 1986.

S. Buffett and B. Spencer. A bayesian classifier for learning opponents' preferences in multi-object automated negotiation. *Electronic Commerce Research and Applications - ECRA*, 6, 09 2007. doi: 10.1016/j.elerap.2006.06.008.

C. F. Camerer, G. Nave, and A. Smith. Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. *Management Science*, 2017. doi: 10.1287/mnsc.2017.2965.

R. Carbonneau, G. E. Kersten, and R. Vahidov. Predicting opponent's moves in electronic negotiations using neural networks. *Expert Systems with Applications*, 34(2):1266–1273, 2008. ISSN 09574174. doi: 10.1016/j.eswa.2006.12.027.

H. p. Chao, S. Oren, and R. Wilson. Reevaluation of vertical integration and unbundling in restructured electricity markets. In *Competitive electricity markets*, pages 27–64. Elsevier, 2008.

L. Chen, H. Dong, and Y. Zhou. A reinforcement learning optimized negotiation method based on mediator agent. *Expert Systems with Applications*, 41(16): 7630–7640, 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.06.003.

S. Chen and G. Weiss. An approach to complex agent-based negotiations via effectively modeling unknown opponents. *Expert Systems with Applications*, 42(5):2287–2304, 2015. ISSN 09574174. doi: 10.1016/j.eswa.2014.10.048.

S. Chen, J. Hao, S. Zhou, and G. Weiss. Negotiating eith unknown opponents toward multi lateral agreement in real time domains. *Computational Intelligence*, 674:219–229, 2017. doi: 10.1007/978-3-319-51563-2.

T. Chung, S. Zhang, C. Yu, and K. Wong. Electricity market risk management using forward contracts with bilateral options. *IEE Proceedings-Generation, Transmission and Distribution*, 150(5):588–594, 2003.

T. de Bruin, J. Kober, K. Tuyls, and R. Babuška. Integrating state representation learning into deep reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1394–1401, 2018.

H. Ehtamo, R. Hamalainen, and P. Heiskanen. Generating Pareto solution in two party setting: constraint proposal methods. *Managment science*, 45: 1697–1709, 1999.

F. Eshragh, M. Shahbazi, and B. Far. Real-time opponent learning in automated negotiation using recursive bayesian filtering. *Expert Systems with Applications*, 128:28 – 53, 2019.

P. Faratin, C. Sierra, and N. Jennings. Negotiation decision functions for autonomous agents. 1998.

P. Faratin, C. Sierra, and N. R. Jennings. Using similarity criteria to make issue trade-offs in automated negotiations. *Artificial Intelligence*, 142(2):205–237, 2002. ISSN 00043702. doi: 10.1016/S0004-3702(02)00290-4.

C. Fershtman. A note on multi-issue two-sided bargaining: bilateral procedures. *Games and Economic Behavior*, 30(2):216–227, 2000.

M. Francisco, Y. Mezquita, S. Revollar, P. Vega, and J. F. D. Paz. Multiagent distributed model predictive control with fuzzy negotiation. *Expert Systems with Applications*, 129:68 – 83, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.03.056.

E. H. Gerding, D. D. B. van Bragt, and J. A. La Poutré. *Scientific approaches and techniques for negotiation: a game theoretic and artificial intelligence perspective*. Centrum voor Wiskunde en Informatica, 2000a.

E. H. Gerding, D. D. B. van Bragt, and J. A. La Poutré. *Scientific approaches and techniques for negotiation: a game theoretic and artificial intelligence perspective*. Centrum voor Wiskunde en Informatica, 2000b.

M. H. Hajimiri, M. Nili Ahmadabadi, and A. Rahimi-Kian. An intelligent negotiator agent design for bilateral contracts of electrical energy. *Expert Systems with Applications*, 41(9):4073–4082, 2014. ISSN 09574174. doi: 10.1016/j.eswa.2013.12.034.

J. Hao and H. F. Leung. ABiNeS: An adaptive bilateral negotiating strategy over multiple items. *Proceedings - 2012 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2012*, 2:95–102, 2012. doi: 10.1109/WI-IAT.2012.72.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

T.-B. Ho and Z.-H. Zhou. *PRICAI 2008: Trends in Artificial Intelligence: 10th Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, December 15-19, 2008, Proceedings*, volume 5351. Springer, 2008.

W.-T. Huang, K.-C. Yao, and C.-C. Wu. Using the direct search method for optimal dispatch of distributed generation in a medium-voltage microgrid. *Energies*, 7(12):8355–8373, 2014.

S. Jamali and K. Faez. Applying saq-learning algorithm for trading agents in bilateral bargaining. pages 216–222, 2012. doi: 10.1109/UKSim.2012.39.

E. Kalai and M. Smorodinsky. Other Solutions to Nash ' s Bargaining Problem. *Econometrica*, 43(3):513–518, 1975.

H. Kebriaei and V. J. Majd. A simultaneous multi-attribute soft-bargaining design for bilateral contracts. *Expert Systems with Applications*, 36:4417–4422, 2009. ISSN 09574174. doi: 10.1016/j.eswa.2008.05.003.

G. E. Kersten, R. Vahidov, and D. Gimon. Concession-making in multi-attribute auctions and multi-bilateral negotiations: Theory and experiments. *Electronic Commerce Research and Applications*, 12(3):166 – 180, 2013. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2013.02.002. Negotiation and E-Commerce.

D. S. Kirschen and G. Strbac. Fundamentals of power system economics. pages 52–58, 2004a.

D. S. Kirschen and G. Strbac. Fundamentals of power system economics. pages 33–40, 2004b.

M. Klein and P. Faratin. Protocols for Negotiating Complex Contracts. *IEEE Intelligent Systems*, 2003.

K. Kolomvatsos, K. Panagidi, I. Neokosmidis, D. Varoutas, and S. Hadjiefthymiades. Automated concurrent negotiations: An artificial bee colony approach. *Electronic Commerce Research and Applications*, 19:56 – 69, 2016. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2016.09.002.

G. Lai, C. Li, K. Sycara, and J. Giampapa. Literature review on multi-attribute negotiations. *Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-04-66*, 2004.

G. Lai, C. Li, and K. Sycara. Efficient multi-attribute negotiation with incomplete information. *Group Decision and Negotiation*, 15(5):511–528, 2006. ISSN 09262644. doi: 10.1007/s10726-006-9041-y.

G. Lai, C. Li, and K. Sycara. A general model for pareto optimal multi-attribute negotiations. In *Rational, Robust, and Secure Negotiations in Multi-Agent Systems*, pages 59–80. Springer, 2008a.

G. Lai, K. Sycara, and C. Li. A decentralized model for multi-attribute negotiations with incomplete information and general utility functions. *Rational, Robust, and Secure Negotiations in Multi-Agent Systems*, 89:39–57, 2008b. ISSN 1860949X. doi: 10.1007/978-3-540-76282-9_3.

S. Lange and M. Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

G. Lao and C. Zhong. A Negotiation Model and Its Simulation Based on Q-Learning Algorithm. *2010 International Conference on Computational Intelligence and Software Engineering*, pages 1–4, 2010. doi: 10.1109/CISE.2010.5676783.

C. C. Lee and C. Ou-Yang. A neural networks approach for forecasting the supplier's bid prices in supplier selection negotiation process. *Expert Systems with Applications*, 36(2, Part 2):2961 – 2970, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.01.063.

S. li Huang and F. ren Lin. Using temporal-difference learning for multi-agent bargaining. *Electronic Commerce Research and Applications*, 7(4):432–442, 2008. ISSN 15674223. doi: 10.1016/j.elerap.2007.04.001.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

R. J. Lin and S. cho T. Chou. Mediating a bilateral multi-issue negotiation. *Electronic Commerce Research and Applications*, 3(2):126 – 138, 2004. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2004.02.002.

A. Maldonado-Ramirez, I. Lopez-Juarez, and R. Rios-Cabrera. Deep convolutional autoencoders as a minimal state representation for reinforcement learning in industrial robot manipulators. In *2018 XX Congreso Mexicano de Robótica (COMRob)*, pages 1–6. IEEE, 2018.

A. Monteserin and A. Amandi. A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. *Expert Systems with Applications*, 40(6):2182–2188, 2013. ISSN 09574174. doi: 10.1016/j.eswa.2012.10.045.

J. Nash. The Bargaining Problem. *The Econometric Society*, 18(2):155–162, 1950. ISSN 00129682. doi: 10.3982/ECTA10813.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998. ISBN 6123456718139.

B. Y. A. Rubinstein. Perfect Equilibrium in a Bargaining Mode. *Econometrica*, 50(1):97–109, 1982. ISSN 00129682. doi: 10.2307/1912531.

P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. Wong, and J. Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 415–420. IEEE, 2010.

K. P. Sycara. Persuasive argumentation in negotiation. *Theory and Decision*, 28(3):203–242, 1990. ISSN 00405833. doi: 10.1007/BF00162699.

K. P. Sycara. Problem Restructuring in Negotiation. *Management Science*, 37 (10):1248–1268, 1991. ISSN 0025-1909. doi: 10.1287/mnsc.37.10.1248.

K. Takadama. in Agent-based simulation : Reproduction of Human-like Behaviors and Thinking in a Sequential Bargaining Game. *Artificial socieities and social simulation*, 2008.

H. Van Hasselt and M. A. Wiering. Reinforcement learning in continuous action spaces. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279. IEEE, 2007.

C. Yu and T. N. Wong. An agent-based negotiation model for supplier selection of multiple products with synergy effect. *Expert Systems with Applications*, 42(1):223–237, 2015. ISSN 09574174. doi: 10.1016/j.eswa.2014.07.057.

C. Yu, F. Ren, and M. Zhang. *An Adaptive Bilateral Negotiation Model Based on Bayesian Learning*, volume 435, pages 75–93. 01 2013. doi: 10.1007/978-3-642-30737-9_5.

J. Zhan, X. Luo, C. Feng, and M. He. A multi-demand negotiation model based on fuzzy rules elicited via psychological experiments. *Applied Soft Computing Journal*, 2017. ISSN 15684946. doi: 10.1016/j.asoc.2017.07.013.

J. Zhan, X. Luo, and Y. Jiang. An Atanassov intuitionistic fuzzy constraint based method for offer evaluation and trade-off making in automated negotiation. *Knowledge-Based Systems*, 139:170–188, 2018. ISSN 09507051. doi: 10.1016/j.knosys.2017.10.020.

Credit author statement

All authors (Mina Montazeri, Hamed Kebriaei and Nanak Araabi) have contributed to Conceptualization (Ideas; formulation or evolution of overarching research goals and aims), and Writing - Original Draft (Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision).

Mina Montazeri and Dr, Kebriaei have contributed to Development or design of methodology. Mina Montazeri also contributed to programming and simulating the results.

Dr. Kebriaei and Dr. Araabi also contributed to Supervision (Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team).

Sincerely,

Dr. Hamed Kebriaei

Corresponding Author

Conflicts of Interest Statement

Manuscript title: Learning Pareto Optimal Solution of a Multi-Attribute Bilateral Negotiation Using Deep Reinforcement

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Mina Montazeri, Hamed Kebriaei and Babak Nadjar Araabi