Reinforcement Learning Based Optimal Bidding Strategy Learning Framework With Risk Preference in Spot Electricity Market

First A. YUANYU, Fellow, IEEE, Second B. Author, and Third C. Author, Jr., Member, IEEE

***Abstract*—Great effort has been made to restructure the traditional monopoly power industry, introducing fair competition. The deregulation of market allows the electricity price to form based on power plants offers indicating generation willing at corresponding bidding price. This paper proposes reinforcement learning(RL) methods to devise optimal bidding strategy maximizing the profit with consideration of risk preference in spot electricity market.The problem is formulated in the framework of Markov decision process (MDP), a discrete stochastic optimization method. The cumulative profit over the span is the objective function to be optimized. The temporal difference technique and actor-critic learning algorithm are employed. The Smart-Market market-clearing system and Gaussian distribution is included in the formulation. Two different environment conditions of the spot electricity market, static and dynamic, are applied in simulation for analysis completeness. Only the target plant can adjust bidding strategy in the static environment while all plants can adjust bidding strategy in the dynamic environment. Simulation cases of nine participants are considered and the obtained results are analyzed.**

***Index Terms*—Bidding strategy, risk preference analysis, Spot electricity market, inverse reinforcement learning**

I. Introduction

DEREGULATION of the electricity industry has become an established practice in many parts of the world. In the day-ahead market, the power-exchange bidding mechanism requires each participant to submit bids for all 24 h of a day as a block. However, in the Spot electricity market, the supply of electricity is matched from power stations with real time consumption by households and businesses. All electricity in the spot market is bought and sold at the spot price.

The bid is in the form of points of piecewise linear curve on the axes, with energy level in megawatt hours and price in dollars per megawatt hour. Thus, it is necessary for participants to propose advantageous and rational bidding strategies, the hourly generation schedule coupled with the competitive bid price, in a deregulated electricity market. Advantageous strategy means maximizing ones own profits which denoting success transaction with great difference between revenue and cost considering self-generation cost, clear-price prediction, competitors behavior forecasting. Rational strategy means satisfying security boundary conditions of generation and minimizing the risk involved.

A strategy is defined to be optimal for a plant at the moment, if it can maximize profit with safety guarantee. Study on finding the optimal bidding strategy is significant in three aspects. Firstly, it guides participants to bid sensibly, notably improving transaction efficiency. Secondly, it contributes to market investigation, identifying the potential for abuse of market power through loopholes that can be exploited in market structure and management rules since these results have important policy implications.Thirdly, ideally the electricity market structure and management mechanisms will grows better designed directing the operation of the market towards maximizing social welfare.

Several optimal bidding strategy models focusing on the market clearing-price forecast have been proposed, while others concentrate on the bidding behaviors prediction of competitors. Meanwhile, some actual market bidding strategies are well modeled to evaluate the accuracy of market simulation and load forecasting. A basic price-based auction mechanism is proposed in [1]. An auctioneer matches the buyers and sellers bids to find the market clearing price (MCP) [2]. Game theory is used for optimal bidding [3] and [4] for hourly auction. Dynamic programming is used in [2] for revenue adequate bidding. A genetic-algorithm based method is described in [6]. This approach is effective only if the market is not volatile. Optimization-based bidding strategies are proposed in [8]. The optimal bidding is divided into two optimization problemsone each for a participant and the independent system operator (ISO). The ISO sub problem is deterministic and the participants sub problem is stochastic. In [9], the problem of optimal bidding is modeled as a Markov decision process (MDP) where load on a weekly basis with peak and off peak loads is considered.

On conclusion, up to now, study on finding optimal bidding strategy are limited in conventional simulation methods, mainly focusing on the day-ahead market. Reinforcement Learning, one of the most important machine learning technology has not been applied as model. Other problems include market environment analysis deficiency since simulation under static and dynamic conditions belongs to different agent models. Meanwhile, sampling process of load is not convincing with distinction only in weekly peak and off-peak. Whats more, risk preference analysis is mostly absent from market simulation, which is a key component in studying the behavior of power generation company. Risk preference analysis always indicates the acceptance degree to risks of power plants during bidding. Not only the market participants and policy-makers can benefit from risk preference analysis of power plants, but also the electricity market can become mature since the prediction framework grows complete considering risk preference. Even though, some models take risk into consideration, the data collection approach relies on traditional investigation or data-driven method, which is sometimes inaccurate.

In this paper, we propose reinforcement learning (RL) methods to devise optimal bidding strategy, maximizing the profit with consideration of risk preference in spot electricity

market. The problem is formulated in the framework of Markov decision process (MDP) under two different environment conditions of the spot electricity market, static and dynamic. The temporal difference technique and actor-critic learning algorithm are employed. The Smart-Market market-clearing system and Gaussian distribution sampling in loads forecasting is included in the formulation. Major contributions of this paper are summarized as follows.

1) Based on RL, optimal strategy considering loads, competitors offers, historical bidding will be learned and evaluated from market simulation under static and dynamic environment respectively. The static environment, where only target power plant can decide offers, simplifies the complex transaction process while the dynamic environment where all power plants can decide their bidding, being closer to the reality. In the static environment, single agent Deep Deterministic Policy Gradient (DDPG) algorithm is used to produce the optimal bidding strategy. In the dynamic environment multiple agents DDPG algorithm is applied and corresponding risk for each competitor will be calculated. DDPG is a deep reinforcement learning algorithm based on policy selection action, which is an improvement of Deep Policy Gradient (DPG) and the Actor-Critic algorithm.The Actor-Critic algorithm is a model-free, off-policy method where the critic acts as a value-function approximator to estimate the action value, while the actor updates the policy distribution in the direction suggested by the critic. In our framework, the actor is our target plants whose action is to decide its next bidding price with others strategy fixed. Whenever the actor decides its bidding price in current state, the clear result will be calculated by the Smart-Market and the net profit will be regarded as rewards. The critic is executed to judge the action based on the rewards correspondingly while updating the overall evaluation mechanism.
2) Data set of risk preference analysis based on Markov Decision Processes and Gaussian distribution will be generated from the market simulation framework to well improve the original methods. Frequently, great gap exists between the expected outcomes of these optimal strategies predicted from conventional model and the actual transaction results. The deviation indicates that the deficiency of precise risk preference analysis in the model will cause prediction inaccuracy. Generally, there are three aspects in the motivation of risk preference analysis. Firstly, market participants can check the consistence between their actual bidding behavior and their predetermined strategy. Secondly, market participants can acquire risk preferences information of their competitors contributing to better bidding strategy design. Thirdly, the policy-makers can conduct better policy reforms to improve price stability in the electricity market utilizing the risk preference information.

There are mainly two approaches in risk preference information gathering, traditional investigation and data-driven method. In traditional questionnaires and experiments investigation, participants are required to make choices in different well-designed dilemma and the risk preference index will be evaluated based on the results. However, there are some inevitable disadvantages of questionnaires and experiments. Firstly, it is intricate and time-consuming to design and conduct investigation. Secondly, individuals risk preferences is influenced by many factors thus, will change time to time. Even though respondents can be investigated under different scenarios, it still can not guaranteed the comprehensiveness of market situation resulting in the inaccuracy. Thirdly, since risk preference is relatively personalized information, competitors may not well cooperate with the experiment under such competitive market environment.

Thus, data-driven method is applied later in information collecting whose accuracy and reliability have been recognized generally. However, ground truth information is greatly needed to evaluate the precision of risk preference analysis in data-driven approach, which is actually insufficient in reality. The deficient of benchmark data set limits the utility of the data-driven method. Thus, the spot electricity market simulation framework generating data set of risk preference analysis based on Markov Decision Processes and Gaussian distribution can well improve it.

The proposed method not only solves the problem of insufficient ground truth information in the data-driven method by producing risks preference data under well design market simulation environment. But also the data set generated can be utilized to train the RL market simulation model contributing to learn optimal bidding strategy that is more accurate with risk preference analysis.

The remaining sections of this paper are organized as follows In Section II, we briefly present the background and notation of the established Markov Decision Processes (MDPs), RL, and OPF algorithm. In Section III, the RL methods of learning optimal bidding policy under different environment conditions considering risk preference for the spot electricity market. In Section IV, we verify the method in a simulation environment and present the results of bidding. Conclusion and further discussion are provided in the last section.

II. BACKGROUND

This section introduce the related work of Markov Decision Process, reinforcement learning and optimal power flow. .

A. *Markov decision process*

As a popular and attractive way of modelling the decision processes with uncertainties, Markov decision process (MDP) works as a foundation framework in reinforcement learning. The MDP, also referred to as controlled Markov

chain, describes a multi-state problem in which decision-making agent, either single or multiple, must choose action at every node of the chain in order to maximize some reward-based optimization criterion. In the decision process, the state transition is stochastic, with the probabilities called transition probabilities. Every state transition is associated with some reward.

A standard MDPs model is composed of five factors in a tuple $(\mathbf{S}, \mathbf{A}, \{\mathbf{P_{sa}}\}, \gamma, \mathbf{R})$,where

- S is a finite set of N states.
- $\mathbf{A} = \{\mathbf{a_1}...\mathbf{a_k}\}$ is a set of k actions.
- p_{sa} are the state transition probabilities upon taking action *a in state s.*
- $\gamma \in [0, 1)$ *is the discount factor.*
- $\mathbf{R} : \mathbf{S} \rightarrow \Re$ *is the reinforcement function, bounded in absolute value by* R_{max}

A policy is defined as any map $\pi : \mathbf{S} \rightarrow \mathbf{A}$.The aim is to find an optimal policy maximizing the accumulated reward in MDPs, with specific transition function and reward function. The value function $V^{\pi}(s)$ for a policy π, evaluated at any state is given by

$$V^{\pi}(s_1) = E[R(s_1) + \gamma R(S_2)] + \gamma^2 R(s_3) + ...|\pi] \tag{1}$$

Where the expectation is over the distribution of the state sequence passed through during the execution of the policy π. We also define action-value function, Q-function according to

$$Q^{\pi}(s, a) = E^{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | S_t = s, A_t = a\} \tag{2}$$

Where the notation s' $P_{sa}(.)$ refers to the expectation is with respect to s' distributed according to $P_{sa}(.)$.

According to the Bellman equation for Q-function,$Q^{\pi}(s, a)$ satisfies Equation(3) for all $s \in \mathbf{S}$ and $a \in \mathbf{A}$.

$$Q^{\pi}(s, a) = R(s) + \gamma E_{s' \sim P_{sa}(.)}[V^{\pi}(s')] \tag{3}$$

Combine (3) with Bellman optimal, proves that (4) must be satisfied in a finite state space $\mathbf{S} = \{\mathbf{s_1}, ..., \mathbf{s_n}\}$ and a set of actions $\mathbf{A} = \{\mathbf{a_1}, ..., \mathbf{a_n}\}$ when the i^{th} element is the reward at i^{th} state in a N-dimensional vector.

$$(P_{a_1} - P_a)(I - \gamma P_{a_1})^{-1} R \geq 0 \tag{4}$$

Where a_1 is the optimal action in actions set $\mathbf{A}$ and a refers to other actions in $\mathbf{A}$. P_a is a N by N transition probability matrix where the element at the (i,j) position gives the probability of transitioning to state j upon taking action a in state i. The generalization of (4) with the value function is :

$$E_{s' \sim P_{sa_1}}[V^{\pi}(s')] \geq E_{s' \sim P_{sa}}[V^{\pi}(s')] \tag{5}$$

Using a linear function to approximate the reward function, the optimization problem can be formulated as :

$$\begin{aligned} &\max_{\beta} \sum_{s \in \mathbf{S_0}} \min_{a \in \mathbf{A}/\mathbf{a_1}} \{p(\Delta E)\} \\ &\Delta E = E_{s' \sim P_{sa_1}}[V^{\pi}(s')] - E_{s' \sim P_{sa}}[V^{\pi}(s')] \end{aligned} \tag{6}$$

Where $\mathbf{S_0}$ is the initial state set, β is the coefficient of the reward function and p is the penalty weight given by $p(X) = x$ if $x > 0$; $p(x) = 2x$ otherwise.

The Bellman optimality equation represents a finite set of equations for a finite MDP. If there are N states, then there are N equations and N unknowns. If there are dynamics of the environment, transition probabilities and rewards that are known, then one can solve this system of equations for $V^{\pi}(s)$ using methods such as dynamic programming. But in circumstances when complete dynamics of the system are not known, simulation methods like Monte Carlo estimation or temporal difference learning are used [10]. The following sections present the reinforcement learning method in general and the actor-critic learning algorithm in particular.

B. Reinforcement learning and DDPG Learning Algorithm

Reinforcement learning simulates an intelligent agent that can learn how to make good decisions by observing its own behavior. It uses built-in mechanisms for improving the actions through a reinforcement mechanism. It essentially maps situations to actions in order to maximize a numerical reward [10].

To realize learning strategy through reinforcement learning, several approaches such as Temporal Difference Algorithm [3], Q-Learning [4] and Classifier Systems [5], are commonly proposed. In this paper, we will mainly solve the problem according to DDPG algorithm.

Inspired by deterministic policy gradient [8,9], DDPG is the algorithm that combines DPG and actor-critic framework, which utilizes the success of Deep Q-learning to the continuous action spaces, DDPG uses experience replay and has four neural networks: critic network $Q(s, a|\theta^Q)$, actor network $\mu(s, a|\theta^{\mu})$, target critic network $Q'(s, a|\theta^{Q'})$, target actor network $\mu'(s, a|\theta^{\mu'})$.

DQN shows a great learning effect in many practical problems, but it is a value-based algorithm, which can only select actions in discrete space and cannot output continuous actions. When the dimension of environment state and action is low, a Q-table is enough to store the value of action under a certain state. However, when environment states and actions are high-dimensional and continuous, it is not reality to store Q value in Q tables. So, a function is used to fit Q value. When similar states are input into the function, the similar actions should be output, the function is as follow:

$$Q^{(}\mathbf{S}, \mathbf{A}, \theta) \approx Q'(\mathbf{S}, \mathbf{A}) \tag{7}$$

In the (7), Q represents the value of action under the state S with the parameter θ. Q function is approximated to the optimal Q value by updating parameters. Deep neural network can automatically extract complex features and is suitable as a fitting function. It combines the perceptual ability of neural networks with the decision-making ability of reinforcement learning. In order to ensure the convergence and stability of DQN algorithm, the following two mechanisms are added into DQN:

- Experience Replay: the experience acquired in the learning process is saved as experience samples, and then a fixed number of experience samples are randomly selected for training, which destroys the correlation between training samples.
- Target Network: there is a network whose architecture is corresponding to the main network structure, but update frequency of this network is different with the main network.

The framework of DQN algorithm is shown in Figure 1.

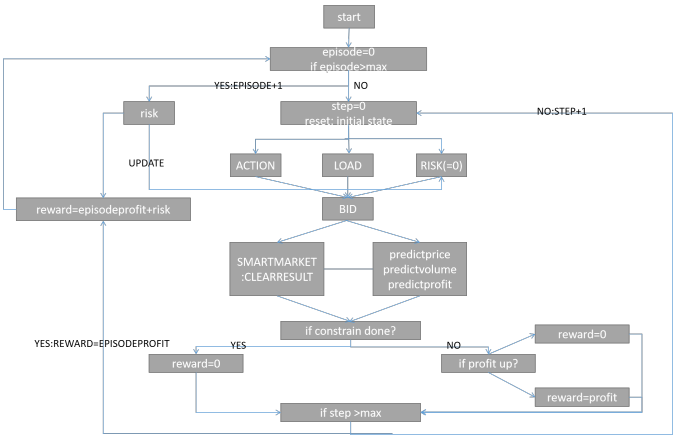

Fig. 1. Framework of DQN algorithm

Actor-critic method is TD method that has a separate memory structure to explicitly represent the policy independent of the value function. The temporal difference (TD) learning is a novel method in a group of reinforcement learning methods of solving large-stage MDP [10]. It learns from experience to solve the prediction problem. The general schematic diagram of an agent employing reinforcement learning is presented in Fig. 1. The agent selects some action based on the current state and obtains a reward from the environment. The environment makes a transition into a new state due to this action. The agent updates the state values depending upon the immediate reward and the next state which results. The actor-critic method proposes two learning agents that work in two loops. The outer loop consists of a reinforcement learning agent, which selects the action in accordance with the current policy and receives the reinforcement feedback. The policy structure is called the actor. The inner loop constructs a more informative evaluation function. The state value function of each state is estimated from the reinforcement feedback. The estimated value function of each state is called the critic because it criticizes the action made by the actor. The actorcritic method can be represented schematically as shown in Fig. 1.

The actor-critic method works by selecting an action from the existing policy. The reward obtained from the transition is used to update estimates of the state value of the current state and the preference of selection of the action next time. After selecting each action, the TD error δ_t is calculated as:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{8}$$

Where V(s) is the current value function implemented by the critic. It is called TD error because it estimates the difference between the current estimated state value and the actual state value for the present policy. The TD error is used to evaluate the action just selected (i.e., the action a_t taken in state s_t). If TD error is positive, then the tendency to select action a_t in the future should be encouraged. If it is negative, then the tendency to select action a_t in the future should be discouraged. This is implemented by updating the preference p as follows:

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t \tag{9}$$

Where $\beta, (0 < \beta \leq 1)$ is a step-size parameter. The policy is derived from the preferences using the softmax method

$$\pi(s, a) = \frac{exp(\frac{p(s,a)}{\gamma})}{\sum_{b=1}^{B} exp(\frac{p(s,b)}{\gamma})} \tag{10}$$

Where B is the total number of actions available in state. The parameter γ is called the temperature. It is used to control the relative probabilities of selection of the states. Selection of γ should be made judiciously. A large value of γ makes all actions equally probable and a small value of γ increases the probability of marginally better actions disproportionately, leading the agent in the wrong direction. A simple but effective choice of γ is made by making it the function of the mean of the preferences.

$$\gamma = F(\frac{\sum_{b=1}^{B} p(s, b)}{B}) \tag{11}$$

The state values are updated according to:

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \tag{12}$$

Where $\alpha, (0 < \alpha \leq 1)$, is a step-size parameter.

On conclusion, the framework of DDPG algorithm is shown in Figure 2.

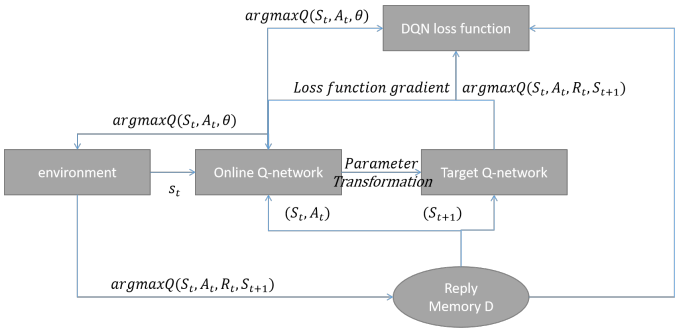

Fig. 2. Framework of DDPG algorithm

The policy thus obtained after sufficient iterations is a sub optimal but vastly improved policy. In the following section, we present the formulation of the bidding problem as an MDP and its solution based upon the actor-critic learning method.

C. The Electricity Market Bidding Problem

Since the 1980s much effort has been made to restructure the traditional monopoly power industry with the objectives of introducing fair competition and improving economic efficiency. The creation of well designed mechanisms for power suppliers, and sometimes for large consumers, to openly trade electricity is at the core of this change. There are mainly two kinds of transaction mode in the electricity market, the day-ahead market and the spot market. We focus on the spot market in this paper.

In the spot electricity market nowadays, the bidding problem in electricity markets is related to pool trading in which the sealed auction is widely employed and power suppliers, and sometimes large consumers also, are required to offer price and quantity bids to a market operator (ISO) who is responsible for clearing the whole bidding process.

This approach leads to a centralization of the unit commitment decisions at the market operators level: ISO is required to send all the relevant information, including market historical records and power grid information, such as the load forecasts for the next hour, system operational states, and other net safety constraints. After receiving the information, market participants need to build their bidding strategies according to the published information and predictions of competitors. All bids must be submitted before the deadline, otherwise, the ISO will adopt default values. Then, ISO proposes the spot clearing-price and operation plan through solving the security-constrained unit commitment (SCUC) and security-constrained economic dispatch (SCED) problem. The spot clearing-price and operation plan are finally published to all participants. The timing diagram of the spot market bidding process is shown in Fig. 1.

The bidding file for the power plant is pairs of price-volume value, and is comprehensively determined by the market rules. The bidding volume represents the additional power output that the power plant is willing to generate at the corresponding bidding price. Table I gives a bidding example with five price-volume pairs. The bidding strategy of each power plant can only be supported by publicly available market information and its private information. The bidding price offered by plants must increase monotonically and be limited by the upper and lower bounds, called price cap and price floor, which is mainly determined by safety constraints .

In a perfect electricity market, any power supplier is a price taker. Microeconomic theory holds the optimal bidding strategy for a supplier is simply to bid marginal cost. When a generator bids other than marginal cost, in an effort to exploit imperfections in the market to increase profits, this behavior is called strategic bidding. If the generator can successfully increase its profits by strategic bidding or by any means other than lowering its costs, it is said to have market power. The new electricity markets are certainly not perfectly competitive, and as a result, a supplier can increase profits through strategic bidding, or in other words, through exercising market power.

D. Optimal power flow algorithm and Smart Market

Since in the electricity transaction, the decision on the winning bidding and a uniform market clearing price (MCP) will significantly influence the market stability, the OPF, optimal power flow algorithm plays a critical role in the clearing process.

The optimal power flow (OPF) problem seeks to optimize certain objective such as power loss and generation cost subject to power flow equations and operational constraints. It is a fundamental problem because it underlies many power system operations and planning applications such as economic dispatch, unit commitment, state estimation, stability and reliability assessment, volt/var control, demand response, etc.

The OPF problem is described as a multi-constrained, non-convex, non-linear problem having an objective function that is non-differentiable. The OPF problem solutions can be obtained using different methods [1-5], they can be categorized into two main groups. The traditional mathematical approach and the heuristic/intelligent methods. Examples of the conventional methods include gradient based methods, newton method [11], linear programming [12], quadratic programming, interior point methods [13] among many others. Examples of the intelligent methods include the genetic algorithm [2-5], particle swarm optimization method [14], teaching-learning based optimization [1], glow worm-based optimization [8], Imperialist competitive algorithm [9], opposition based elitist real genetic algorithm [13], modified cataclysmic genetic algorithm [14] among other such methods.

The smart market algorithm consists of the following basic steps:

1) Convert block offers and bids into corresponding generator capacities and costs.
2) Run an optimal power flow with decommitment option (uopf) to find generator allocations and nodal prices (λ_P).
3) Convert generator allocations and nodal prices into set of cleared offers and bids.
4) Print results.

For step 1, the offers and bids are supplied as two structs, offers and bids, each with fields P for real power and Q for reactive power (optional). Each of these is also a struct with matrix fields qty and prc, where the element in the i_th row and j_th column of qty and prc are the quantity and price, respectively of the j_th block of capacity being offered/bid by the i_th generator. These block offers/bids are converted to the equivalent piecewise linear generator costs and generator capacity.Offer blocks must be in non-decreasing order of price and the offer must correspond to a generator with $0 \leq PMIN < PMAX$. A set of price limits can be specified via the lim struct. Capacity offered above this price is considered to be withheld from the auction and is not included in the cost function produced. Bids must be in non increasing order of price and correspond to a generator with

$PMIN < PMAX \leq 0$. The data specified by a Matpower case file, with the gen and gencost matrices modified according to step 1, are then used to run an OPF.

The OPF solution is used to determine for each offer/bid block, how much was cleared and at what price. These values are returned in co and cb, which have the same structure as offers and bids. The mkt parameter is a struct used to specify a number of things about the market, including the type of auction to use, type of OPF (AC or DC) to use and the price limits.

III. SPOT ELECTRICITY MARKET SIMULATION FRAMEWORK

A. Problem Formulation

The ISO conducts an energy auction for the spot market. The electricity spot market is a wholesale market, and operates each day of the week from 7:00 A.M. until about 1:00 P.M. During this window, electricity is traded for delivery that will start at midnight. The bids are in the form of points of piecewise linear curves on energy and price coordinates. The seller bids the amount of the energy that he or she is willing to sell at a given price or above, and the buyer bids the amount of the energy he or she is willing to buy for a given price or a price lower than it. The unconstrained MCP is determined by the point of intersection of the aggregate demand and supply bid curves.

It is presumed that the external operation conditions that participants knows includes the total generation capacity of each other, the load forecast for the next hour, the past cost curves of competitors. Specially, the load will be predicted by sampling from a Gaussian distribution with mean bounded by the maximum total generation capacity.

The internal operation conditions includes the start-up cost of each of its units and the cost curve is assumed to be known. The cost curve is assumed to be quadratic in the form of:

$$CC^i = a^i + b^i U + c^i U^2 \tag{13}$$

And the start-up cost for an hour t is su^{it} .

The participant first decides in how many parts it wants to bid, say parts. It divides its maximum generation capacity into parts and using the cost curve, finds the marginal generation cost at the higher end of generation for each part. This forms its middle element M of bid set. The higher H and lower L elements of the bid set are obtained from the middle element as: $H_j^1 = 1.1M_j^1$ and $L_j^1 = 0.9M_j^1$ for $j = 1, 2, 3, ..., q$

The bid set consists of a Cartesian product of the three bid elements for each part. For a two-part bid $q = 2$ and three levels (H,M,L) for each part, the bid set consists of nine (3^q) bid-set= $[B_1, B_2, ..., B_{c^q}]$ Where, a bid $B = [0, P_1, P_2, ..., P_q, b, pr_1, pr_2, ..., pr_q]$

The corresponding bid sets for all of the other participants are formed in a similar way. Since target participant cannot predict in how many parts each one of the other participants is going to bid, it takes a single part for each participant. The bid set of every other participant consists of three elements only.

B. Static and Dynamic Environment

Obviously, if the factors influencing clearing price is viewed as state and the bid-set of target power plant is defined as action, the problem can be reformulated as a MDPs.

The state of the system is defined as bids placed by each participant in a particular hour. The bids placed by each participant are selected from their respective bid sets: $s = [s(1), s(2)..., s(n)], s(1) \in bid - set(1), ..., s(n) \in bid - set(n)]$. Where s represents one state and $[s(1), s(2)..., s(n)]$ are its elements.

The framework will be discussed under two different environment conditions, static and dynamic. Only the target plant can adjust strategy and others will bid with marginal cost in the static environment while all plants can adjust their strategy in the dynamic environment. Under static condition, the elements include last round profit, current bid, profit forecast, price-volume forecast vector of the target participant. Under dynamic condition, the elements include last round profit, current bid, profit forecast, price-volume forecast matrix of the all participants. Particularly, according to the real bidding rules of the electricity market, the current spot price should be defined as the key factor the state. The total number of possible states in each hour is the Cartesian product of all the elements of bid sets of every participant. The state which is selected in a particular hour is decided by the bids selected by each participant. The bids for every hour are stochastically selected by each participant, with each bid in the bid set having a finite probability of being selected. The set of probabilities of each bid for every hour is the policy that is being followed by the participant.

Bidder will decide their bidding action, their offers, for the next clearing round in an hour. Approximately, the bidding action should be transferred from the continuous state space into discrete state space, since the continuous price value could be segmented into n classes according to the n bidding volume shaping as price-volume pair pri_i and vol_i , $i \in 1, ..., n$. Suppose s_i is the clearing-price between i_th bidding price pr_i and $(i+1)_th$ bidding price pr_{i+1}(i.e., $pri_i \leq s_i < pri_{i+1}$), the power plant maximum generation output will be the sum of all bidding volumes ($\sum_1^i vol_i$) before the clearing-rice . The action is actually the allocation of the total power generation capability vol_{sum} of each price-volume pair. Bidder needs to decide every vol_i,$i \in 1, ..., n$, ensuring = vol_{sum}.

The reward function is the summation of two parts, the net profit NP and the potential risk R with some coefficient α. NP is the expectation of net profit in all states calculated by subtracting the generation cost including the fixed and variable cost from revenue (NP=Revenue-Cost) according to the clearing result. The revenue is the expectation profit calculated by the multiplying the clearing price with the actual generation volume of plant under different loads in 24 hours, and the cost is calculated by the product between the marginal cost function and the volume respectively. The risk that a power plant is willing to take is calculated by taking the variance of the revenue in 24 hours. The greater the absolute

value of the variance, the higher the risk action that the power plant is taking. Unlike using statistical tools or formulas in the risk measure method, the preference generated by our approach represents the tendency to adopt low profit actions (this can be regarded as high risk actions). Since the coefficient represents the risk preference of the power plant, a positive means the power plant considers the risk of reduced profit as punishment and the risk preference is risk-aversion; while a negative means the power plant considers the risk of reduced profit as stimulation and the risk preference is risk-seeking; and zero means the risk preference is risk-neutral.

There is always a gap between actual bidding action and theoretical optimal bidding action.The potential loss PL is used to evaluate the deviation between the real transaction value after each generator deciding its strategy with risk consideration and the ideal optimal result where participants bid by marginal cost. It is calculated by subtracting the theoretical optimal profit OP from the actual profit AP(PL=AP-OP). It represents the risk of loss that the power plant is willing to take. OP depends on the clearing result when taking the theoretical optimal action a_0 which is learned from solving the maximum profit problem without considering the risk. For comparison, the singular profit SP is defined to measure the transaction result where participants decide bid without considering risk factor. The risk loss RL calculated by subtracting the actual profit AP from the singular profit SP (RL=AP-SP) represents the safety cost that a power plants is willing to take. The risk preference deviation calculated from the risk difference between singular profit and the actual profit is defined as RPD index which is used to evaluate the working power of plants strategy.

C. Application of the ActorCritic Learning Algorithm

In this section, we consider how the DDPG method can be applied to produce the optimal bidding strategy, according to loads, and its corresponding risk, solving the problem that was formulated as an MDP in the previous section.

Under static conditions, the simulations are carried out by the agent of target participant. The target participants agent learns to apply a policy maximizing the profit earned while satisfying risk preference. It assumes that other agents bid with marginal cost as fixed policy. The state elements include last round profit, current bid, profit forecast, price-volume forecast vector of the target participant.

Under dynamic conditions, the simulations are carried out by agent of all participants. The agent of participants assumes that all of the agents are going to bid in such a manner that applying a policy maximizing the profit earned by self participant, while minimizing the others profit constrained by risk preference. The state elements include last round profit, current bid, profit forecast, price-volume forecast matrix of the all participants. Specially, the rule of iteration is the same as under the static conditions except that the result will be a matrix other than 1-D vector.

- BID: The agent starts biding from its marginal cost and selects bids incremental magnitude from its range set according to the probabilities. The load forecast sampling from Gaussian for next hour, together with the risk index selected by private information indicating the risk preference decides the next state for transition.
- MCP: The MCP that would result due to the given bidding by the participants and the allotted generation for each participant are calculated by the smart-market. According to the OPF solution, the clear price-volume result that how much was cleared and at what price for each offer/bid block will be decided.
- PROFIT: The profit for target participant is obtained as:

$$profit^{it} = mcpU^{it} - (a^i + b^i U^{it} + c^i (U^{it})^2) - su^{it} \quad (14)$$

In the first iteration, the agent for participant starts the process by randomly selecting bids incremental based on marginal cost from bid sets. The set of selected bids and its relative MCP results define the next transition state. Equation (19) can be used to calculate the profit for target participant. The profit so obtained, is taken as the step profit for target participant. After finding the final step profit in an episode, the mean profit will be calculated as episode profit. The reward will be calculated as the accumulation of the summation of episode profit and risk part which will be taken as zero firstly. The variance of the episode profit will be recorded as expected risk preference for next round learning.

The TD error is calculated by each agent after every transition using (4). The TD error for specific participant would be equal to $r_(t+1)$ in the first iteration. The policy (i.e., the probabilities with which the agent selects a bid) is updated to reflect the outcome of the first selection. The agents form preference for the state action pair using the most recent reward through (5). The preference reflects the cumulative reward that the agent obtains from a particular state action pair. The probability of that state action pair being selected is calculated using (6). This completes the outer loop of Fig. 2. The state value for the initial state is updated to a better estimate using (8), hence going through the inner loop of Fig. 2. The procedure mentioned before gets repeated in every iteration, as the system makes a transition from a state to next state, each state depicting the bids selected by the participants in a particular hour. The transition from the 24^{th} h is to the 1^{st} h, so the iterations continue in the form of an infinite horizon MDP. The process is stopped typically around 10000 iterations. The result yields probabilities of the selection of each bid during each hour. The expected profit can be found out from the converged values of the bid.

The proposed DDPG algorithm methods under different environment is shown in Fig 3.

IV. Spot Electricity Market Implementation

We apply the method mentioned before to the case file lib/t/t auction case.m, which is a modified version of the 30-bus system that has 9 generators, where the last three have negative PMIN to model the dispatchable loads.

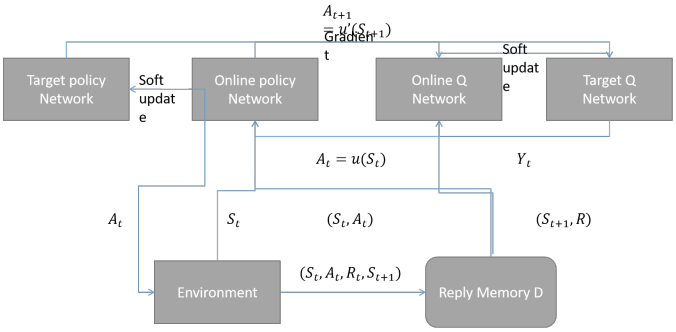

Fig. 3. Framework of DDPG algorithm under different environmentm

A. *Sample System I*

The sample system used in the manual is considered here.

- Three dispatchable loads, bidding three fixed blocks each as shown in Table I.
- Six generators with three blocks of capacity each, initial offering as shown in Table II.
- Load sampling from Gaussian

TABLE I
THREE DISPATCHABLE LOADS

Generator	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	10*$100	10*$100	10*$100
2	10*$100	10*$100	10*$100
3	10*$100	10*$100	10*$100

The six generators are the six participants who submit bids and are allocated portions of the total demand.

TABLE II
SIX DISPATCHABLE GENERATORS

Generator	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	12*$20	24*$50	24*$60
2	12*$20	24*$40	24*$70
3	12*$20	24*$42	24*$80
4	12*$20	24*$44	24*$90
5	12*$20	24*$46	24*$75
6	12*$20	24*$48	24*$60

The cost curves are as

$$CC^1 = 3.0U + 0.03U^2$$
$$CC^2 = 5.5U + 0.01U^2$$
$$CC^3 = 4.8U + 0.015U^2$$
$$CC^4 = 4.5U + 0.02U^2$$
$$CC^5 = 3.5U + 0.04U^2$$
$$CC^6 = 4.0U + 0.03U^2$$

The startup cost for the six participants for the unit supplying upper 25 MW of the cost curve is

$$su^1 = 50$$
$$su^2 = 50$$
$$su^3 = 50$$
$$su^4 = 50$$
$$su^5 = 50$$
$$su^6 = 50$$

The sampling example of 24-h load curve is shown in Fig 4.

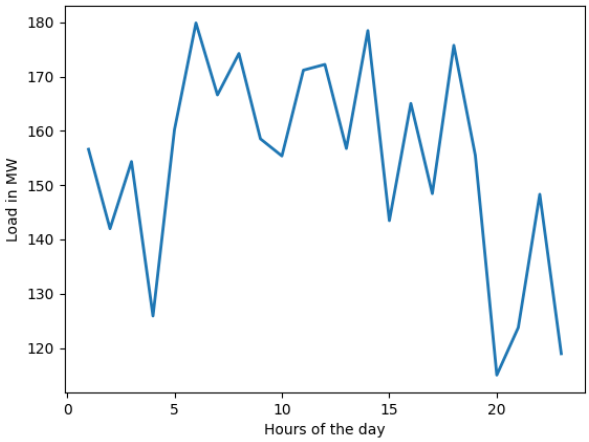

Fig. 4. Framework of DDPG algorithm under different environmentm

The framework will be discussed under two different environment conditions, static and dynamic. Only the target plant can adjust strategy and others will bid with marginal cost in the static environment while all plants can adjust their strategy in the dynamic environment. We will firstly show and analyze the simulation results under static environment where target participant will be the center and the dynamic environment later. The comparison and comprehensive discussion will be given in the end.

B. *Static Environment*

The market simulation under static environment is carried over for the above example. The discount factor γ is 0.8. The parameter of step size α is varied from 5 to 50 according to risk parameter. The result will come to converge earlier with greater risk parameter. The step size parameter for preferences update is 0.02. We have observed that it should be low. Higher values of show apparent early convergence but the solution may not be a feasible one which does not satisfy the boundary conditions. A total of 10 000 iterations were carried out for each experiment. The experiments will be composed of four parts, results without risk consideration and results with risk preference but in three groups different risk index. The experiments results will be shown firstly under different market conditions respectively. And then the generated situation when risk factor matters or not will be compared. Finally. The

TABLE III
PROMOTION PERCENTAGE OF PRICE ON 6 GENERATORS

Generator	1	2	3	4	5	6
Percentage	16.6%	15.0%	15.6%	11.3%	29.3%	39.2%

clear results with risk factor but different risk index will be summarized.

1) Result: The states or bidding strategy having the largest probabilities to be selected without risk consideration by the target participants for the 24 h are shown in Fig 5. It can be seen that target generator prefers to choose a lower price and larger volume pairs as final strategy when others bidding with marginal cost. The transaction feasibility and net revenue is more likely to grow.

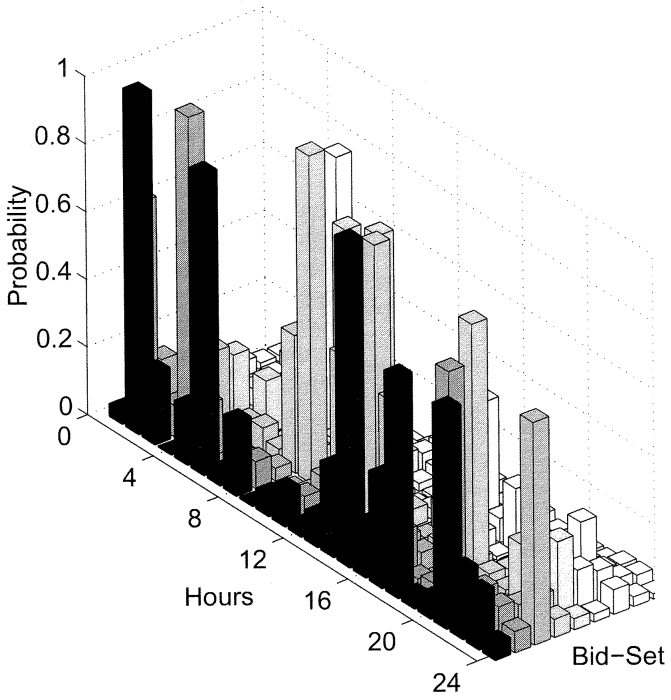

Fig. 5. Framework of DDPG algorithm under different environmentm

The self clear price in 24 hours for the spot market where target generator is 1 to 6 respectively is shown in Fig 6. The original trading price where all generators bid by marginal cost is also shown in Fig with points. From the results, it is obvious that self clear price will increase substantially when target generator take strategy during bidding.

Table III shows the average increasing percentage for each target generator over 24 hours trading. The strategy effect works differently for generators. The promotion is most significant for generator 1, 5, 6 and is weaker for the others. It could be guessed that the strategy effect depends on the generators private attribute including generator physical conditions and risk preference decision.

The Fig 7 shows the self clear volume in 24 hours for the spot market where target generator is 1 to 6 respectively. Fig

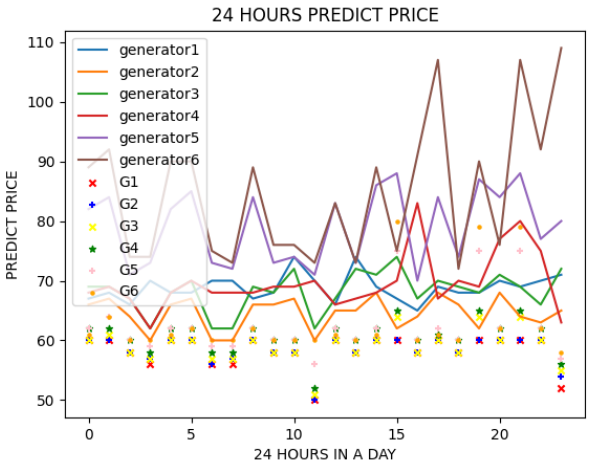

Fig. 6. Framework of DDPG algorithm under different environmentm

together with Fig forms the bid clear pair of trading. The original trading volume where all generators bid by marginal cost is also shown in Fig in points. From the results, it is seen that self clear volume will not absolutely increase when target generator take strategy during bidding. The strategy considers the incremental of overall revenue instead of single factor.

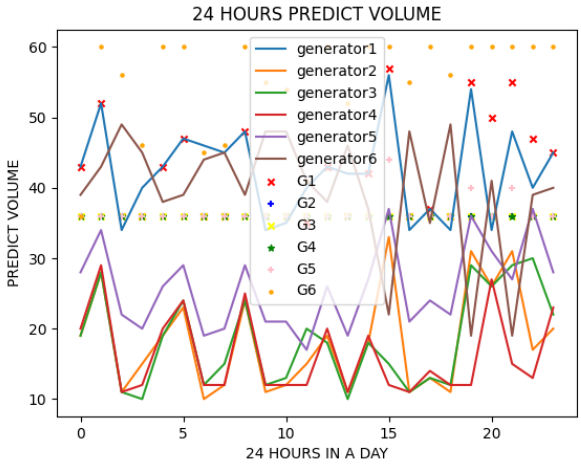

Fig. 7. Framework of DDPG algorithm under different environmentm

Fig 8 summarize the trading revenue (exclude the cost function) in 24 hours for the spot market where target generator is 1 to 6 respectively. The original bidding profit where all generators bid by marginal cost is also shown in Fig in points. From the results, it is obvious that self profit will increase substantially when target generator take strategy during bidding.

Table IV shows the average increasing percentage for each target generator over 24 hours trading. Compared with the promotion percentage of price, direct proportion can be observed.

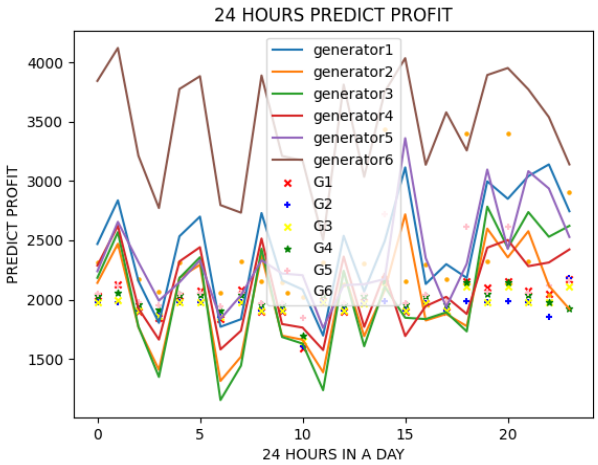

Fig. 8. Framework of DDPG algorithm under different environmentm

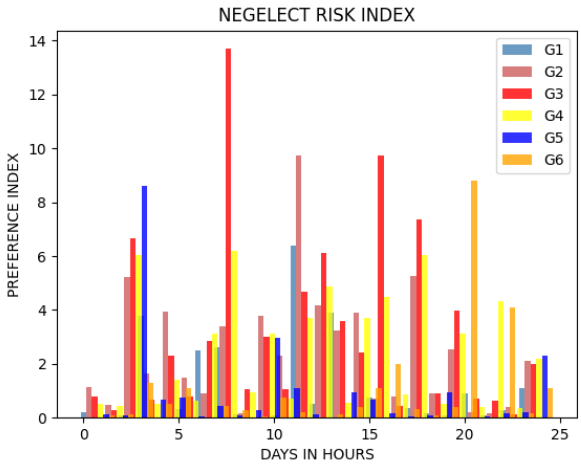

Fig. 9. Framework of DDPG algorithm under different environmentm

TABLE IV
PROMOTION PERCENTAGE OF PROFIT ON 6 GENERATORS

Generator	1	2	3	4	5	6
Percentage	25.0%	3.0%	1.0%	1.5%	50.0%	62.0%

The average rewards for target generators achieved in 24 hours has been shown Table V. Columns of Table presents the averaged values of the immediate rewards taken over all iterations. Maximizing profit is the same as maximizing the state values of the most probable states.

TABLE V
REWARD VALUE OF 6 GENERATORS IN 24 HOURS

Hour	1	2	3	4	5	6	7	8	9	10	11
12	13	14	15	16	17	18	19	20	21	24	
G1	30	29	28	29	32	27	26	31	28	27	27
28	29	30	32	34	25	56	26	27	29	29	
G2	26	25	24	21	23	24	25	25	25	24	24
23	26	27	29	21	22	22	23	21	25	24	
G3	27	28	29	27	28	28	27	26	25	24	23
24	24	23	21	28	26	26	27	24	23	23	
G4	23	22	21	20	19	18	24	23	22	21	20
20	18	20	21	22	24	25	25	25	26	24	
G5	31	32	33	33	31	34	35	35	34	36	31
30	35	36	38	37	36	34	32	32	31	35	
G6	38	40	45	46	42	46	47	42	40	39	42
44	43	43	45	42	41	40	36	39	39	45	

The risk preference value learned from the target generators is shown in Fig 9. Fig 8 and 9 show the relationship between the state values and the risk preference being determined. Generally, generators 2 and 3 assign higher risk preference to their bidding strategy.The immediate profit due to state transition also affects the probabilities.Hence, the agent of participant learns to allot higher probabilities to lower rewards when its bidding strategy is risk adventure. The immediate profit due to state transition is also affected by the private risk decision.

2) Results with 0.5_0.8 risk index: Here we take the same system with six generators, the cost curves, and the maximum generation capacities of each generator which are the same to sample system but taken risk preference into consideration. The results with different risk index decision will also be compared. The daily load curve is taken as earlier, with a total generating capacity of six generators that is just enough to meet the peak system demand. The shape of the load curve is the same as in the earlier system

TABLE VI
STRATEGY EXAMPLE IN RANDOM HOUR

Generator Houes	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	12*$20	24*$45	24*$70
2	12*$20	24*$36	24*$60
3	12*$20	24*$47	24*$50
4	12*$20	24*$51	24*$80
5	12*$20	24*$48	24*$65
6	12*$20	24*$32	24*$60
7	12*$20	24*$45	24*$70
8	12*$20	24*$37	24*$60
9	12*$20	24*$39	24*$50
10	12*$20	24*$48	24*$80
11	12*$20	24*$52	24*$65
12	12*$20	24*$31	24*$71
13	12*$20	24*$55	24*$69
14	12*$20	24*$65	24*$85
15	12*$20	24*$43	24*$73
16	12*$20	24*$51	24*$90
17	12*$20	24*$48	24*$68
18	12*$20	24*$32	24*$69
19	12*$20	24*$45	24*$86
20	12*$20	24*$36	24*$75
21	12*$20	24*$49	24*$88
22	12*$20	24*$42	24*$67
23	12*$20	24*$41	24*$62
24	12*$20	24*$36	24*$49

Taken one of the generators as target example the bidding blocks in 24 hours after risk consideration is shown in Table. The self clear price,volume pairs in 24 hours for the

spot market where target generator is 1 to 6 respectively is shown in Fig when all generators are under low risk index transaction. Generally, the trading price and volume increase more during bidding for those risk adventurers. But they will also have higher chance of broken bidding where the bidding should restart. It can be indicated clearly from the data of rewards since broken bidding will cause penalty in state value. Meanwhile greater fluctuation of the states will be observed for them in 24 hours since they are aggressive to take action even though accompanying with possibility of failing. In contrast, the trading price and volume increase less during bidding for those risk averse. But they will also have less chance of broken bidding and weaker fluctuation of states. For those risk neutral generators, their trading results almost keep the same as the static environment without risk preference.

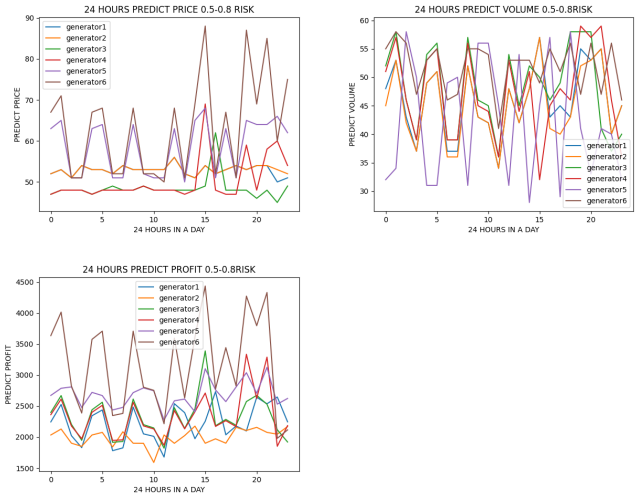

Fig. 10. The self clear results value in 24 hours for the spot market with 0.5_0.8 risk index

Table VII shows the average increasing percentage of profit for each target generator over 24 hours trading. It could be seen that the percentage increase caused by strategy bidding is declined thoroughly comparing with the static environment without risk preference, no matter its risk adventure or averse. But the extent of decline varies depending on its risk preference.

TABLE VII
PROMOTION PERCENTAGE OF PRICE ON 6 GENERATORS WITH RISK

Generator	1	2	3	4	5	6
Percentage	15.0%	0.5%	0.8%	1.2%	42.0%	57.0%

The risk preference value re-learned from the target generators is shown in Fig. It is obvious that for all kinds of risk preference, the result of considering risk is enlarging the variance of profits. Same as above, the extent of enlarging still depends on the attitude to risk. Risk adventure such as generator 2 and 3 is more likely to have a substantial variation while risk averter have less.

The risk preference value re-learned from the target generators is shown in Fig. It is obvious that for all kinds of risk preference, the result of considering risk is enlarging the variance of profits. Same as above, the extent of enlarging still depends on the attitude to risk. Risk adventure such as generator 2 and 3 is more likely to have a substantial variation while risk averter have less.

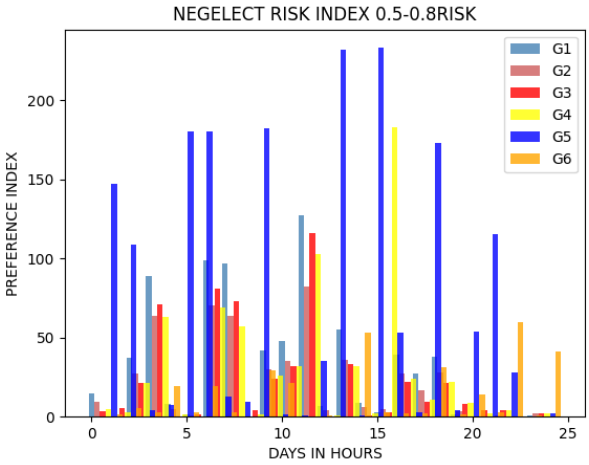

Fig. 11. The risk preference value in 24 hours for the spot market with 0.5_0.8 risk index

3) Results with 0.8_1.2 risk index: The self clear price,volume pairs in 24 hours for the spot market where target generator is 1 to 6 respectively is shown in Fig 12 when all generators are under neural risk index transaction. Similar curve characteristic can be observed compared with generators under high risk index. The main difference focuses on the risk effect plays a more important role during bidding. However since the index is close to 1, the comprehensive market clear results will close to the static conditions without risk preference.

The risk preference value re-learned from the target generators with neural risk index is shown in Fig 13. Obviously, the profit variation grows accompanying with high proportion of risk factor.

4) Results with 1.2_1.5 risk index: Self clear price,volume pairs in 24 hours for the spot market is shown in Fig when all generators are under high risk index transaction. And the roiiiiilmmmmjisk preference value re-learned from the target generators is shown in Fig 14.

V. Dynamic Environment

The same sample system is applied in the market simulation under dynamic conditions. Experiment without risk preference will be firstly carried out and then risk factor is taken into consideration. The discount factor here is 0.8. The step size is 20 and the step size parameter for preferences update is 0.02. A total of 10 000 iterations were carried out for the above problem. In data analysis, the clear results of experiments without risk and risk will be compared. And then, taking

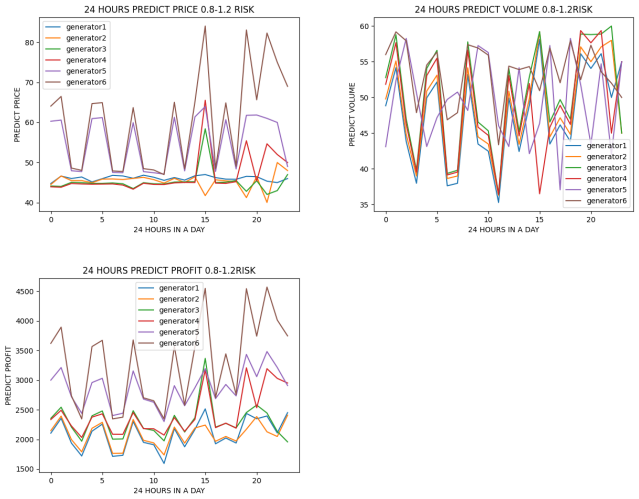

Fig. 12. The self clear results value in 24 hours for the spot market with 0.8_1.2 risk index

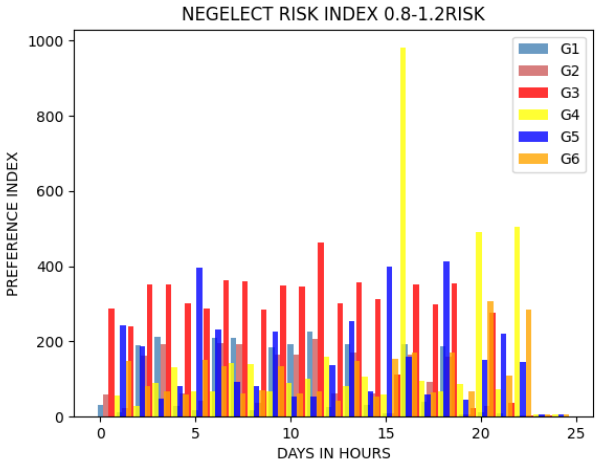

Fig. 13. The risk preference value in 24 hours for the spot market with 0.8_1.2 risk index

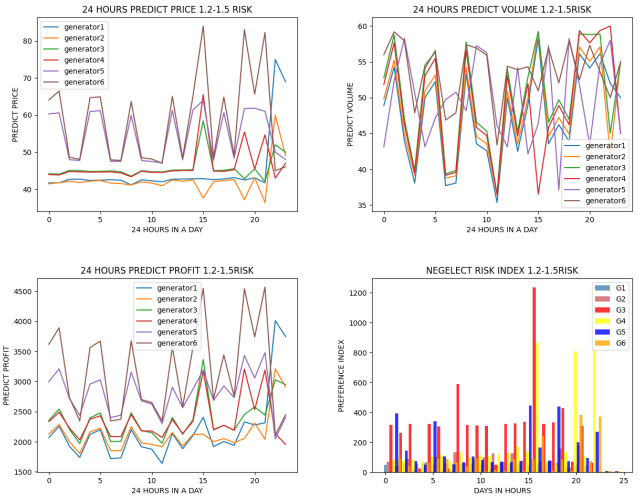

Fig. 14. The self clear results and risk preference value value in 24 hours for the spot market with 1.2_1.5 risk index

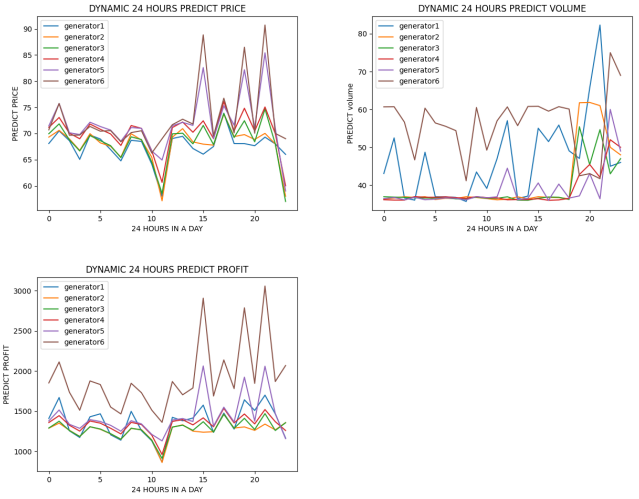

Fig. 15. The self clear results value in 24 hours for the spot market under dynamic conditions

arbitary case in 24 hours as example, the strategy decision will be shown under static and dynamic environment, both without risk, is given. Finally, the states value generated from experiments above will be compared and concluded.

1) Results without risk preference: The self clear price and volume pairs in 24 hours in the dynamic spot market bidding without risk for generator 1 to 6 respectively is shown in Fig 15. The transaction profit after cost subtraction is given in Fig. Obviously, the predict clear price in 24 hours for each generator has almost the same variation trend indicating the dynamic bidding gambling will cause same effect for each participant. But the final transaction situation depends on private characteristic including physical conditions of plants. It can be seen that generator 1, 5 and 6 apparently achieve higher price and larger volume leading to higher net profit. Apart from they have better machine set supporting power generation, their risk tolerance is actually higher which is shown in Fig.16.

2) Results with risk preference: The self clear price and volume pairs in 24 hours in the dynamic spot market bidding with risk for generator 1 to 6 respectively is shown in Fig. Generally the transaction results almost keeps in same proportion among generators as situation without risk. But the numerical value of either clear price or volume decreases respectively for all generators. It is reasonable to achieve such market clear results since every participant will bidding vigilantly trying not to against self risk tolerance.

The re-learned actual risk preference for each generator is shown in Fig 18. It is obvious that the profit variation for participants grows resulting in higher risk comparing with original situation.

3) Strategy comparison example of static and dynamic environment: Table VIII is the original strategy bidding in

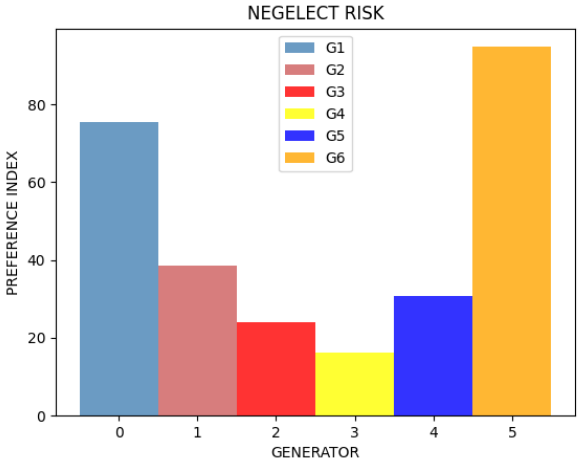

Fig. 16. The risk preference value in 24 hours for the spot market under dynamic conditions

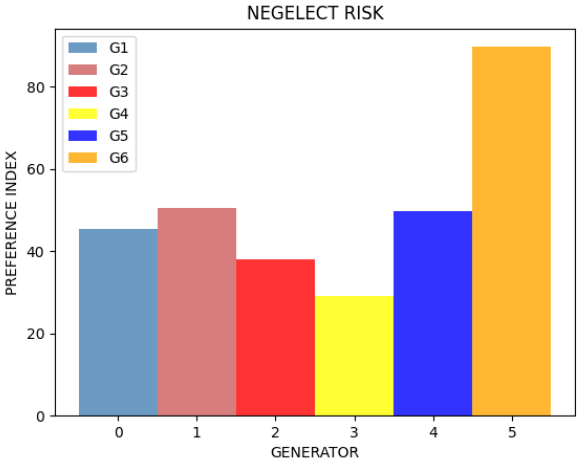

Fig. 18. The risk preference value in 24 hours for the spot market under dynamic conditions with risk

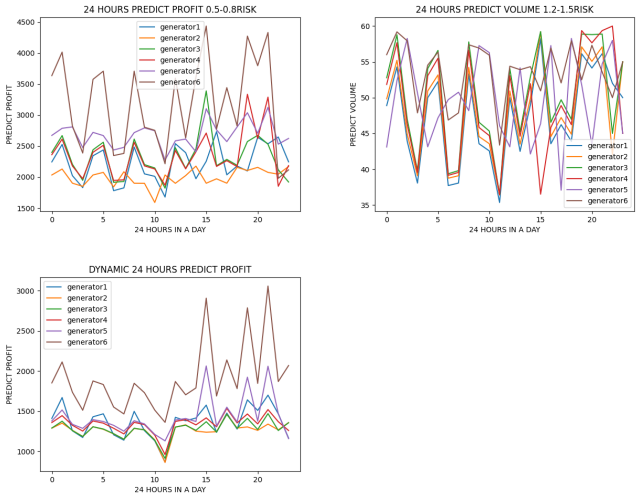

Fig. 17. The self clear results value in 24 hours for the spot market under dynamic conditions with risk

marginal cost. Table IX is the strategy under static environment where only generator 1 is the target participant selected randomly in 24 hours. Table X is the strategy under dynamic environment selected the same hour as in static conditions.

TABLE VIII
STRATEGY BIDDING IN MARGINAL COST

Generator	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	12*$20	24*$45	24*$70
2	12*$20	24*$40	24*$70
3	12*$20	24*$42	24*$80
4	12*$20	24*$44	24*$90
5	12*$20	24*$46	24*$75
6	12*$20	24*$48	24*$60

4) States value comparison example of static and dynamic environment: The original market clear result of bidding by

TABLE IX
STRATEGY UNDER STATIC ENVIRONMENT

Generator	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	12*$20	24*$42	24*$59
2	12*$20	24*$50	24*$68
3	12*$20	24*$36	24*$51
4	12*$20	24*$54	24*$66
5	12*$20	24*$45	24*$62
6	12*$20	24*$40	24*$55

TABLE X
STRATEGY UNDER DYNAMIC ENVIRONMENT

Generator	Block1 MW*$/MWh	Block2 MW*$/MWh	Block3 MW*$/MWh
1	12*$20	24*$50	24*$60
2	12*$20	24*$40	24*$70
3	12*$20	24*$42	24*$80
4	12*$20	24*$44	24*$90
5	12*$20	24*$46	24*$75
6	12*$20	24*$48	24*$60

marginal cost is shown in Fig 19. First two table in Fig 19 is the clear value in an hour for 6 generators under static environment where generator 1 and 2 is the target participant respectively. Last two table in Fig 19 is the clear result under dynamic environment of all generators in the same hour as the static situation.

Through comparison, it can be figured out that whether it is under static conditions or dynamic conditions, the profit will increase for all generators using bidding strategy. Under static conditions, the target generator will be the main increasing point, for generator 1 and 2, 32% and 71% respectively. Meanwhile the other generators bidding in marginal cost have 25% increase in average taking case 1 as example. Under dynamic conditions, generally, there is greater increasing in profit even for the target participant in static situation. Specially, there is less deviation in profit growth among generators.

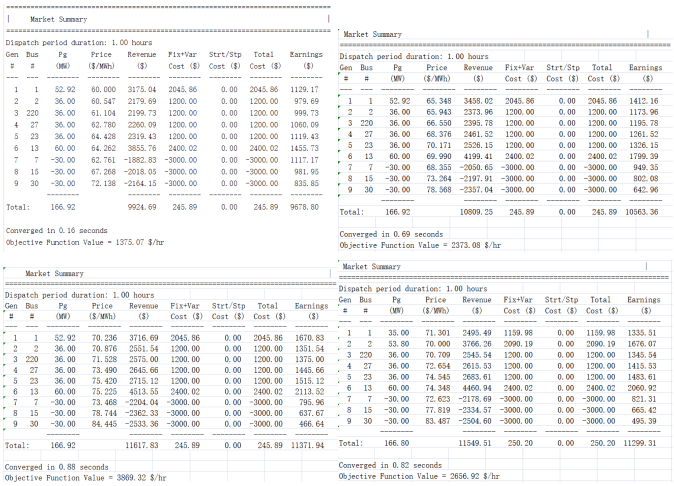

Fig. 19. The self clear results value in 24 hours for the spot market under dynamic conditions with risk

Concrete increasing percentage of profit under both conditions is concluded in Table XI

TABLE XI
PROMOTION PERCENTAGE OF GENERATOR UNDER DIFFERENT CONDITIONS

Generator	Static1:	Static2:	Dynamic:
1	32%	16%	47.9%
2	25.6%	71%	37.9%
3	25.3%	34.6%	37.6%
4	24.5%	33%	36.3%
5	23.8%	32.5%	35.3%
6	30.5%	41.5%	89.2%

VI. CONCLUSIONS

During implementation the market simulation under static environment is illustrated firstly. Under static conditions, the simulations are carried out by the agent of target participant. The target participants agent learns to apply a policy maximizing the profit earned while satisfying risk preference. It assumes that other agents bid with marginal cost as fixed policy. The experiments results shows that: 1) Self clear profit will increase substantially when target generator take strategy during bidding. 2) The promotion from strategy effect depends on the generators private attribute including generator physical conditions and risk preference decision. 3) Risk preference will limit the profit promotion and enlarge the variance of profits. 4)Larger risk index will have larger effects.

The market simulation under dynamic environment is shown later. Under dynamic conditions, the simulations are carried out by agent of all participants. The agent of participants assumes that all of the agents are going to bid in such a manner that applying a policy maximizing the profit earned by self participant, while minimizing the others profit constrained by risk preference. The experiments results shows that: 1) For all generators, self profit will increase and the results are better than under static conditions mostly. 2) The promotion has less deviation comparing to static conditions. 3) Risk preference will also limit the profit promotion and enlarge the variance of profits whose effect is larger than risk consideration under static situation.

Up to now, study on finding optimal bidding strategy are limited in conventional simulation methods, mainly focusing on the day-ahead market. Other problems include market environment analysis deficiency and sampling process of load is not convincing limits the process of introducing fair competition into power industry.

This paper try to solve the problems by modelling the spot market bidding problem as an MDP. The Smart-Market market-clearing system and Gaussian distribution is included in the formulation. Reinforcement learning (RL) methods, the temporal difference technique and actor-critic learning algorithm, are employed. The implementation results shows that under both conditions, the algorithm proposed in paper can devise optimal bidding strategy maximizing the profit with consideration of risk preference.

REFERENCES

[1] R.D.Christie, B.F.Wollenberg and LWangensteen, “Transmission management in the deregulated environment”, *Proceedings of the IEEE*, Vo1.88, No2, February 2000, pp.170-195.

[2] W.Mielczarski, G.Michalik and M.Widjaja, “Bidding strategies in electricity markets”, *Proceedings of the 1999 IEEE PES Power Industry Computer Applications Conference (PICA99)*, 1999, pp.7 I - 76.

[3] M.J.Exelby and N.J.D.Lucas, “Competition in the UK market for electricity generating capacity - a game-theory analysis”, *Energy Policy*, V01.21, N0.4, 1993, pp.348-354

[4] RW.Ferrero, S.M.Shahidehpour and V.C.Ramesh, “Transaction analysis in deregulated power systems using game theory”, *IEEE Transactions on Power Systems*, V01.12, No.3, 1997, pp.1340-1347.

[5] R.J.Green and D.M.Newbery, “Competition in the British electricity market”, *Journal of Political Economy*, Vo1.100, No.5, 1992, pp.929-953.

[6] D.M.Newbery, “Capacity-constrained supply function equilibria: competition and entry in the electricity spot market”, *DAE Working Paper No.9208*, University of Cambridge, 1992.

[7] A.Rudkevich, M.Duckworth, and R.Rosen, “Modelling electricity pricing in a deregulated generation industry: the potential for oligopoly pricing in a poolco”, *Energy Journal*, Vo1.19, No.3, 1998, pp. 19-48.

[8] F.Bolle, “Supply function equilibria and the danger of tacit collusion: the case of spot markets for electricity”, *Energy Economics*, Vo1.14, No.2, April 1992, pp.94-102.

[9] Z.Younes and M.IIic, “Generation strategies for gaming transmission constraints: will the deregulated electric power market be an oligopoly?”, *Decision Support System*, Vo1.24, No.3-4, 1999, pp.207-222.

[10] G. B. Shebl, “Priced based operation in an auction market structure”, *IEEE Trans. Power Syst.*, vol. 11, pp. 17701777, Nov. 1996.

[11] C. Li, A. J. Svoboda, X. Guan, and H. Singh, “Revenue adequate bidding strategies in competitive electricity market”, *IEEE Trans. Power Syst.*, vol. 14, pp. 492497, May 1999.

[12] R. W. Ferrero, S. M. Shahidehpour, and V. C. Ramesh, “Transaction analysis in deregulated power system using game theory”, *IEEE Trans. Power Syst.*, vol. 12, pp. 13401347, Aug. 1997.

[13] V. Krishna and V. C. Ramesh, “Intelligent agent for negotiations in market games, Part 2: Application”, *IEEE Trans. Power Syst.*, vol. 13, pp. 11091114, Aug. 1998.

[14] G. R. Gajjar, S. A. Khaparde, and S. A. Soman, “Modified model for negotiations in market games under deregulated environment”, *in Proc. 11th Nat. Power Syst. Conf.*, Bangalore, India, Dec. 2000.

[15] C. W. Richter Jr., G. B. Shebl, and D. Ashlok, “Comprehensive bidding strategies with genetic programming/finite state automata”, *IEEE Trans. Power Syst.*, vol. 14, pp. 12071212, Nov. 1999.

[16] S. Hao, “A study of basic bidding strategy in clearing pricing auction”, *IEEE Trans. Power Syst.*, vol. 15, pp. 975980, Aug. 2000.

[17] Shoham. Y, Powers. R, and Grenager. T, "If Multi-Agent Learning is the Answer, What is the Question?", *Artificial Intelligence*, Vol 171, No. 3, pp. 365-377, 2007.

[18] Stone. P, and Veloso. M, "Multiagent systems: A survey from a machine learning perspective", *Autonomous Robots*, Vol 8, No. 3, pp. 345383, 2000.

[19] Littman. M. L, "Markov games as a framework for multi-agent reinforcement learning", *Proceeding 11th International Conference on Machine Learning*, New Brunswick, NJ , pp. 157163, 1994.

[20] Littman. M. L, and Szepesvari. C, "A generalized reinforcement-learning model: Convergence and applications", *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, pp. 310318, 1996.

[21] Claus. C, and Boutilier. C, "The dynamics of reinforcement learning in cooperative multiagent systems", *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Orlando, Florida, pp. 746752, 1998.

[22] Kapetanakis. S, and Kudenko, D, "Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems", *Proceedings of the Third Autonomous Agents and Multi-Agent Systems conference*, New York, pp.1258-1259, 2004.

[23] Wang. X, and Sandholm. T, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games", *In Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, MA, MIT Press, 2003.

[24] Littman. M. L, "Friend-or-foe Q-learning in general-sum games", *Proceedings of the Eighteenth International Conference on Machine Learning*, Williamstown, MA, pp. 322-328, 2001.

[25] Hu. J, and Wellman. M, "Nash Q-learning for general-sum stochastic games", *Journal of Machine Learning Research*, No. 4, pp.10391069, 2003.

[26] Greenwald. A, and Hall. K, "Correlated Q-learning", *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, DC, pp. 242249, 2003.

[27] Suematsu. N, and Hayashi. A, "A Multiagent Reinforcement Learning Algorithm Using Extended Optimal Response", *Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, Bologna, pp. 370-377, 2002.

[28] Watkins. C, and Dayan. P, "Technical note: Q-learning", *Machine Learning*, Vol 8, No. 3/4, pp.279292, 1992.

[29] Sutton. R. S, and Barto. A. G, *Reinforcement Learning: An IntroductionMA*, MIT Press, 1998.