Energy 153 (2018) 977–987

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Smart generation control based on multi-agent reinforcement learning with the idea of the time tunnel

Lei Xi [a, *], Jianfeng Chen [a], Yuehua Huang [a], Yanchun Xu [a], Lang Liu [a], Yimin Zhou [b], Yudan Li [a]

[a] College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, 443002, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

ARTICLE INFO

Article history:
Received 5 July 2017
Received in revised form
2 April 2018
Accepted 8 April 2018
Available online 10 April 2018

Keywords:
Automatic generation control
PDWoLF-PHC
Multi-agent
Carbon emission

ABSTRACT

One of the significant solutions for hazy is to reduce carbon emission by introducing renewable energy on a large scale. However, the large-scale integration of new energy will result in stochastic disturbance to power grid. Therefore it becomes a top priority to make new energy compatible with power system. The PDWoLF-PHC(λ) based on the idea of time tunnel is to be proposed in this paper. Optimal strategy could be obtained by adopting the variable learning rate in a variety of complex operating environments, and thence it can deal with stochastic disturbance caused by massive integrations of new energy and distributed energy sources to the power grid, which is difficult for traditional centralized AGC. The proposed algorithm is simulated to be effective according to the improved IEEE standard two-area load-frequency control power system model and the Central China Power Grid model. Compared with the traditional smart ones, the proposed algorithm is characterized with faster convergence and stronger robustness, which makes it able to reduce carbon emission and enhance utilization rate of the new energy.

1. Introduction

In view of the global energy transformation, the use of new energy [1] should be promoted around the world, such as wind power [2], photovoltaic energy [3] and electric vehicles [4] etc. However, the large-scale integration of new energy [5] and the growth of the stochastic load lead to the increasing randomness of power system, so it is urgent to keep the new energy [6] compatible with power system. A kind of smart generation control (SGC) method with distributed structures for new energy is explored in terms of automatic generation control (AGC) [7] to acquire the best control, and then to deal with stochastic disturbance caused by massive integrations of new energy to the power grid, which is difficult for the traditional centralized AGC.

AGC is an important technology that could be used to adjust the power system frequency, active power and guarantee the safety of the power grid. The traditional centralized AGC system will take priority to chase the optimization of its own regional control and lead to a lower degree of synergistic control among all regions [8]. A study from the dispatching center of the China Southern Power Grid shows that if the control performance standard (CPS) [9] of some regions is increased, the control performance of other regions may be degenerated. With the rapid development of smart power grid, the continuous expansion of installed capacity and the incessant integration of new energy and distributed energy, the traditional centralized AGC model is hard to meet the development and operation requirement of the power grid. Therefore, what has become the main trend is to develop a distributed SGC system which can reduce carbon emission (CE) [10], improve the utilization ratio of new energy and adapt the strong random environment as well as multi-area coordination. The biggest difference between SGC and AGC is that the smart method of SGC displaces the original PI control [11] of AGC.

Scholars have successively developed a number of distributed multi-area SGC (MA-SGC) systems [12], in which an intelligent agent with high self-optimization and self-learning abilities is used, and optimal coordination and control of power grid can be effectively realized. Moreover, the agents of the SGC system ultimately achieve the equilibrium through dynamic competition or

* Corresponding author.
E-mail addresses: xilei2014@163.com (L. Xi), chenjf1007@163.com (J. Chen), hyh@ctgu.edu.cn (Y. Huang), xyc7309@163.com (Y. Xu), 2753599289@qq.com (L. Liu), ym.zhou@siat.ac.cn (Y. Zhou), 1193307992@qq.com (Y. Li).

https://doi.org/10.1016/j.energy.2018.04.042
0360-5442/

cooperation in a complex and random condition. According to the information structure and action order of the game taken by the agents, the stochastic game can form a variety of complex equilibrium relations, such as the mixed strategy Nash equilibrium of complete static information, sub-game refined Nash equilibrium of complete dynamic information, Bayes equilibrium under incomplete information static game, refined Bayes equilibrium under incomplete information dynamic game [13]. Thus, many smart algorithms are derived from the equilibrium solution gained on the basis of stochastic game. For example, the dimensionality disaster problem of the CPS instruction distribution process can be solved by the adoption of hierarchical correlated *Q*-learning algorithm in the interconnection grid AGC [14]. Good control effects will be achieved in both the single-agent reinforcement learning and multi-agent reinforcement learning (MARL), under the circumstances that the agents in the centralized AGC system are in small number. Nonetheless, it is difficult to accomplish large-scale and complex tasks due to the independent learning of single-agent reinforcement learning with no consideration about the interrelationships among agents. But MARL can greatly improve the system intelligence by sharing information and experiences and collaborating interactively among the agents [15]. Consequently the decentralized reinforcement learning based on multi-agent has been devoted a lot to research in the field of reinforcement learning.

MARL is applied to AGC earlier in literature [16] and [17] since the agent with MARL can track the decisions of the other agents to coordinate its own behaviour. However, real-time coordination in the multi-area among AGC is not considered. Therefore, a variety of MARL algorithms, such as IGA [18], WoLF-IGA [19], and Exploiter-PHC [20], have been proposed, which are based on the stochastic game theory. Meanwhile the opponent's strategy is observable in IGA and WoLF-IGA. The Nash equilibrium is needed in Exploiter-PHC as prior knowledge and does not converge in the self-game. Thus the above algorithms are limited in the applications. The author also proposed a MARL algorithm DCEQ(λ) in the literature [21], based on correlated equilibrium, to solve the optimal coordination control of SGC, in which some satisfactory results are achieved. Nevertheless, the searching time for the multi-agent equilibrium is the geometric growth if the number of multi-agent increases, which will limit the application of DCEQ(λ) in larger systems. A novel algorithm named win or learn fast policy hill climbing (WoLF-PHC) is proposed in literature [19]. The agent adopts a mixed strategy and only needs to maintain its own Q-value table. Consequently the asynchronous decision-making problem of multi-agents is solved effectively. The author proposed the DWoLF-PHC(λ) algorithm based on WoLF-PHC to solve the stochastic disturbance caused by the massive integration of the new energy and the distributed energy to the power grid effectively in the literature [22]. It also can solve the multi-solution problem when the number of multi-agent increases. However, the agents with the WoLF require a strict knowledge system, which limits the universality of DWoLF-PHC(λ). In the 2*2 game, the agent cannot calculate the win-loss criterion of DWoLF-PHC(λ) accurately, and the convergence rate to the Nash equilibrium is slow. Therefore, it is necessary to explore a new algorithm to calculate the win lose judgement criterion accurately, so as to converge to the Nash equilibrium at a faster speed, and obtain the better control performance. Reference [23] proposed the PDWoLF-PHC algorithm with a new win-lose judgement criterion to solve the problem that the agent could not calculate accurately in the 2*2 game and proved the validity in this algorithm. Nevertheless, the slow convergence speed is still needed to resolve.

In this paper, the Policy Dynamics based Win or Learn Fast Policy Hill-Climbing(λ), namely the PDWoLF-PHC(λ), is put forward by fusing PDWoLF-PHC, state-action-reward-state-action (SARSA) with the idea of time tunnel to obtain the equilibrium solution to MA-SGC, and then address the disturbance problems caused by the integration of new energy, which cannot be solved by traditional centralized AGC. The effectiveness of the algorithm is verified by two examples of the improved IEEE standard two-area load-frequency control (LFC) power system model [24] and the Central China Power Grid model. In contrast to other algorithms, a more powerful learning ability and faster convergence speed exist in PDWoLF-PHC(λ). In addition, it has a significant effect on reducing the CE and enhancing the utilization rate of new energy.

The remaining part of the paper is as follows. The PDWoLF-PHC(λ) is expounded in section 2. A novel MARL-based SGC is designed in section 3. section 4 is the simulation analysis of the algorithm. Some related discussions are provided in section 5. section 6 summarizes the full text.

2. PDWoLF-PHC(λ)

In this paper, the PDWoLF-PHC(λ) which integrates the idea of time tunnel is put forward to obtain the optimal coordination control of MA-SGC system.

2.1. Q(λ)

The PDWoLF-PHC(λ) is based on the *Q*-learning framework. *Q*-learning [25] presented by Watkins in 1989 is a reinforcement learning algorithm, which has a strong self-learning ability, and can obtain the optimum solution through continuous trial and error and environmental interaction. The optimal target value function $V^{\pi^*}(s)$ and strategy $\pi^*(s)$ are as follows

$$V^{\pi^*}(s) = \max_{a \in A} Q(s,a) \tag{1}$$

$$\pi^*(s) = \arg\max_{a \in A} Q(s,a) \tag{2}$$

where A is the set of action.

The multi-step backtracking eligibility trace [26] with time varying can be likened to a hypothesis with the idea of time tunnel. The influence of future control decision has been taken into account, because the time tunnel can be used to solve the problem of time reliability distribution for delay reinforcement learning, and to gain the frequency and the recency information of the algorithm. Thus, the *Q*(λ) merges the value function with the time tunnel. In the iterative process, the frequency of each joint action strategy will be recorded in the time tunnel to update the iterative *Q* value. Time tunnel will allocate the multi-step strengthening information of historical decision-making by means of tracking the status of each state-action-pair, so as to achieve the incentives and penalties of decision-making. The multi-step information updating mechanism of the *Q* function is obtained by the backward evaluation of the time tunnel.

The SARSA(λ) [26] based on idea of time tunnel is chosen as

$$e_{k+1}(s,a) = \begin{cases} \gamma\lambda e_k(s,a) + 1, (s,a) = (s_k, a_k) \\ \gamma\lambda e_k(s,a) \quad otherwise \end{cases} \tag{3}$$

where $e_k(s, a)$ is the kth step time tunnel under status s and action a; γ is the discount factor; and λ is the time tunnel attenuation factor.

The agent calculates the error evaluation of the current value function by the reward value R gained from the current exploration. The formulas are shown as following