The application of Deep Reinforcement Learning in Coordinated Control of Nuclear Reactors

Jing Li¹, Yanyang Liu¹, Xianguo Qing¹, Kai Xiao¹, Ying Zhang¹, Pengcheng Yang¹, Yue (Marco) Yang²

¹ Science and Technology on Reactor System Design Technology Laboratory, Nuclear Power Institute of China, Chengdu, 610213, China
² Department of Engineering, The University of Sydney, Sydney, 2006, Australia

*WinkOvO2021@yeah.net

Abstract. The nuclear reactor control system plays a crucial role in the operation of nuclear power plants. The coordinated control of power control and steam generator level control has become one of the most important control problems in these systems. In this paper, we propose a mathematical model of the coordinated control system, and then transform it into a reinforcement learning model and develop a deep reinforcement learning control algorithm so-called DDPG algorithm to solve the problem. Through simulation experiments, our proposed algorithm has shown an extremely remarkable control performance.

1. Introduction

The nuclear power plant control system automatically maintains the relevant parameters of the nuclear power plant within the limits specified in the operating conditions and can automatically adapt the operating state of the nuclear power plant, so as to match the energy generated by the nuclear reactor with the load demand [1]. The design of nuclear power control system can be complicated. The characteristics of the controlled object in the nuclear power plant are multivariable, nonlinear, and strong coupling. As a consequence, it is hard to formulate the thermal process and involved characteristics by an accurate and concise mathematical model. The traditional model-based control method is not perfectly suitable for the above problem.

In recent research of the nuclear power plant, the widely used method is single-input, single-output, closed-loop, and cascade loop PID control [2-3]. Although the PID-based method can capture the operation characteristics in nuclear power plants, the improvement is still required considering its response speed, overshoot and other indicators. On the other hand, the traditional PID-based method is often designed for a single subsystem with a single controlled variable, the coupling and cooperation among multiple systems cannot be tackled by this method. Therefore, the PID-based method cannot be applied to the coordinated control of nuclear power plant systems.

To address the above challenges, the dynamic scheduling problem of the nuclear reactor coordinated control system can be modelled as a stochastic sequential decision problem, which can be solved by reinforcement learning. Reinforcement learning is a classical and critical method in control theory and machine learning, that focuses on how agents take actions in the environment to obtain the maximum cumulative return [4]. Unlike traditional methods modelling the control system in advance, Model-free reinforcement learning does not depend on any prior knowledge. Since most of the control parameters in the control system studied in this paper are continuous variables, there will be a
dimension disaster problem by the traditional reinforcement learning method (e.g., Q-learning). To this end, the deep reinforcement learning algorithm is further introduced to model and solve the coordinated control problem of the nuclear reactor.

In this paper, we aim to develop an intelligent control method for the small nuclear reactor. According to the deep reinforcement learning algorithm, the collaborative control of reactor power regulation system and steam generator liquid level regulation system is studied in this work. These two systems can be modelled as the environment in reinforcement learning and Deep Deterministic Policy Gradient (DDPG) is further introduced to solve the problem. Through extensive simulation experiments, the proposed method demonstrates a remarkable performance compared to several benchmarks.

2. Control system modelling

2.1. Description of nuclear reaction coordinated control system

The complete reactor control system mainly includes power control system, steam generator liquid level control system, main pump control system, pressurizer pressure control system, valve control system, etc. This paper makes further design and research on the reactor power control system and steam generator liquid level control system.

In the reactor power control system, the signals to be considered include nuclear power $N_0$, demand power $N$, measured value of average temperature of primary coolant $T_{\text{avg}}(t)$, set value of average temperature of primary coolant $T_{ref}$, feed-water flow of steam generator $F_w$ and control rod drive signal $Rv$. The control rod drive signal is used to drive the action of the control rod. The action of the control rod should not only adjust the reactor power, but also adjust the average coolant temperature. Therefore, the purpose of the reactor power control system is to make the reactor power deviation $\Delta N$ and the average coolant temperature deviation $\Delta T$ close to zero. The control principle of reactor power control system based on classical cybernetics is shown in Figure 1:

![Figure 1. The control principle of reactor power control system.](image)

In the steam generator liquid level control system, the signals to be considered include the measured value of steam generator liquid level $L$, liquid level setting value $L_s$, steam generator feed water flow $F_w$, steam flow $F_s$ and feed-water regulating valve opening signal $V$. The opening signal of feed water regulating valve is used to control the action of feed water regulating valve. The purpose of controlling the action of feed water regulating valve is to make the feed water flow deviation of steam generator $\Delta F_w$ and the water level deviation of steam generator $\Delta L$ tend to zero. The control principle of steam generator liquid level control system based on classical cybernetics is shown in Figure 2:
2.2. Mathematical description of control problems

The optimization objective of this paper is to minimize the deviation signals of reactor power control system and steam generator level control system. Each deviation signal can be expressed as:

\[ \Delta T = T_{\text{ref}} - T_{\text{avg}} \]  

\[ \Delta N = K_T + F_W - N \]  

\[ \Delta L = L_S - L \]  

\[ \Delta F_W = K_L + F_S - F_W \]  

Among them, \( \Delta T \), \( \Delta N \), \( \Delta L \) and \( \Delta F_W \) are the average temperature deviation, power deviation, steam generator water level deviation and feed-water flow deviation respectively. The reactor coordinated control system changes the reactivity in the reactor through the control rod drive signal, and changes the feed-water flow and steam generator liquid level through the feed-water regulating valve opening signal, so as to eliminate the deviation between the measured feedback signal in the system and the secondary circuit load reference signal, so as to restore the primary and secondary circuits to the state of energy balance. The objective function to be selected in this paper is:

\[ F = \min f(\Delta T, \Delta N, \Delta L, \Delta F_W) \]  

In this paper, we focus on the control optimization problem under discrete-time steps. A complete time control interval can be discretized into \( \{1, 2, \ldots, M\} \). In each time step of control decision, we can write the objective function in (5) above as:

\[ f_t(\Delta T, \Delta N, \Delta L, \Delta F_W) = e_T(t) + e_N(t) + e_L(t) + e_{F_W}(t) \]  

Among them, \( \Delta T \), \( \Delta N \), \( \Delta L \) and \( \Delta F_W \) are the average temperature deviation, power deviation, steam generator water level deviation and feed-water flow deviation at time respectively.

Therefore, the optimization objective of this problem is to optimize the system deviation under the whole control time window:

\[ F = \min \sum_t f_t(\Delta T, \Delta N, \Delta L, \Delta F_W) \]  

3. Reinforcement learning modelling

3.1. Reinforcement learning modelling

Using the advantage that reinforcement learning is very suitable for solving the optimization decision problem with uncertain factors, this paper models and solves the control decision of reactor control system in discrete time window. Here, we first need to transform the mathematical description constructed in the previous section into a reinforcement learning framework.
The basic components of reinforcement learning include state set $S$ representing environment, action set $A$ representing agent action and reward $r$ for agent. Based on the current state $s \in S$, the agent selects an action $a \in A$ to act on the environment. After the environment accepts the action, it changes and generates a reward signal $r$ to feed back to the agent. The agent then continuously selects the next action according to the signal and the current state of the environment. In this paper, the reinforcement learning model is described as follows:

1. **State $s$**: The observed states of nuclear reactor coordinated control system include nuclear power $N(t)$, average temperature of primary coolant $T_{avg}(t)$, feed-water flow $F_w(t)$, steam flow $F_s(t)$, steam generator liquid level $L(t)$, etc.

   

   $s(t) = \{N(t), T_{avg}(t), F_w(t), F_s(t), L(t)\}$  \hspace{1cm} (8)

2. **Action $a$**: In decision time step $t$, the action of the coordinated control system can be represented by control rod drive signal $Rv(t)$ and feed-water regulating valve opening signal $V(t)$,

   

   $a(t) = \{Rv(t), V(t)\}$  \hspace{1cm} (9)

3. **Reward $r$**: In this paper, the problem of minimizing the total deviation of the system is transformed into the classical reward maximization form of reinforcement learning. Therefore, the reward obtained by the agent in period 1 is expressed as (where $1/1000$ is the scaling of the objective function):

   

   $r_1(s(t), a(t)) = -\frac{1}{1000} \left[ e_T(t) + e_S(t) + e_Z(t) + e_{fw}(t) \right]$  \hspace{1cm} (10)

4. **Action value function $Q(s, a)$**: When a certain state $s(t)$ in the reactor coordinated control system is determined, the advantages and disadvantages of system action $a(t)$ can be evaluated using action value function $Q(s, a)$:

   

   $Q(s(t), a(t)) = E(\sum_{k=0}^{M} \gamma^k \cdot r(s(t+k), a(t+k)))$  \hspace{1cm} (11)

   Among them, $E(\cdot)$ is the expectation function, $\gamma \in [0,1]$, $\gamma$ is the discount factor, which represents the impact proportion of the reward in the cumulative reward at a certain time in the future.

5. **Optimal strategy $\pi^*$**: The objective of the reactor coordinated control system is to find the optimal strategy to maximize the action value function:

   

   $\pi^* \leftarrow \arg \min_s Q(s, a)$  \hspace{1cm} (12)

### 3.2. Solution algorithm based on deep reinforcement learning

The traditional reinforcement learning method performs well in the problems of small-scale discrete actions and state space. However, when the number of states and actions increases exponentially, the dimensional disaster occurs [5]. The traditional reinforcement learning method cannot effectively deal with this situation. In the reactor coordinated control system studied in this paper, the state space and action space are continuous variables. Therefore, this paper uses deep neural network (DNN) [6] to approximate the function of reinforcement learning. The specific selection of the algorithm is DDPG algorithm based on actor critical, which estimates the optimal strategy function through deep neural
network, it cannot only avoid dimension disaster, but also save the information of the whole action domain.

The principle of reactor coordinated control scheme based on DDPG algorithm is shown in Figure 3. For DDPG algorithm, the input of policy network is state vector $s(t)$ as shown in formula (7), and the output is action vector $a(t)$ as shown in formula (8); The input of the value network is state vector $s(t)$ and action vector $a(t)$, and the output is action value function $Q(s(t), a(t))$. DDPG algorithm uses the empirical playback mechanism in deep Q network. It enters the storage of experience playback by storing the agent's experience $e_t = \{s(t), a(t), r(t), s(t+1)\}$ in each period. During training, min-batch (size M) experience samples will be extracted from the experience pool each time, and the network parameters will be updated based on the gradient rules. Then, DDPG algorithm trains the value network and policy network respectively to obtain the optimal parameters in the network.

![Figure 3. The framework of DDPG algorithm in Coordinated control of nuclear reactors.](image)

For the value network, we optimize the parameters by minimizing the loss function $L(\theta^Q)$:

$$L(\theta^Q) = E\left(y_t - Q(s(t), a(t)|\theta^Q)\right)^2$$  \hspace{1cm} (13)

Among them, $y_t$ is the target Q value

$$y_t = r(t) + \gamma Q'(s(t+1), \pi'(s(t+1)|\theta^\pi)|\theta^Q)$$  \hspace{1cm} (14)

We update the value network parameters according to the gradient rules:

$$\theta^Q \leftarrow \theta^Q - \mu Q \nabla_{\theta^Q} L(\theta^Q)$$  \hspace{1cm} (15)
For the policy network, the gradient information \( \nabla_{\theta} Q(s(t), a(t) | \theta^\pi) \) provided by it is used as the direction of action improvement. According to the deterministic policy gradient, the parameters of the policy network are updated as follows:

\[
\theta^\pi \leftarrow \theta^\pi - \mu \nabla_{\theta} L(\theta^\pi) \quad (16)
\]

Parameters \( \theta^\pi \) and \( \theta^\pi' \) of the target network use soft update technology to improve the stability of learning:

\[
\theta^\pi \leftarrow \tau \theta^0 + (1 - \tau) \theta^\pi \quad (17)
\]

\[
\theta^\pi' \leftarrow \tau \theta^\pi + (1 - \tau) \theta^\pi' \quad (18)
\]

Among them, \( \tau \) is the soft update parameter, \( \tau \ll 1 \).

4. Numerical experiment

In order to evaluate the effectiveness of the proposed reactor coordinated control system based on deep reinforcement learning DDPG algorithm, this paper takes the small reactor as the control object, including point reactor neutron dynamics model, reactor and primary loop main equipment, steam generator and its secondary side boundary, and uses the same transient simulation model as the engineering design to build the basic simulation environment.

The control parameters of the reactor coordinated control system are shown in Table 1. We mainly conduct the control simulation test under the working condition of 100% - 85% FP linear load reduction, in which the change rate of linear load reduction is selected as 1.33% FP / s.

| Deviation signal     | Fluctuation range of steady state |
|----------------------|----------------------------------|
| Power deviation (%)  | ≤±0.6                            |
| Temperature deviation (°C) | ≤±0.5                           |
| Liquid level deviation (%) | ≤±0.3                         |
| Feed water flow deviation (%) | ≤±0.1                         |

4.1. The training of reinforcement learning

Before applying the deep reinforcement learning network to the reactor coordinated control system, the parameters of deep reinforcement learning are trained through historical data to obtain the deep reinforcement learning network. In order to better evaluate the performance of the DDPG algorithm, we introduce deep Q network learning (DQN) as a reference algorithm for further comparison.

In the value network of DDPG, we set up two hidden layers, each layer has 200 neurons, and the activation function of the hidden layer is ReLU. In DDPG’s policy network, we also set two hidden layers, and the activation function is Tanh. Other training parameters of DDPG are shown in Table 2. For the DQN network, the input is a 5-Dimensional state vector \( s(t) \) and the output is the Q value of the state-action pair. In this paper, \( R_V(t) \) and \( V(t) \) are discretized according to the step size of 0.1. thus the discretized action space is obtained. DQN network has two hidden layers, each layer has 200 neurons, and the activation function of the hidden layer is ReLU. In order to show the convergence performance of the proposed method, Figure 4 shows the average reward value curve every 60 minutes during agent training. The algorithm converges after about 40 minutes of training, and the optimal control strategy is obtained. It can be observed that because the agent is not familiar with the environment at first, the reward value obtained by the agent after executing the scheduling decision is small. With the continuation of the training process, agents constantly interact with the environment.
and gain experience. Therefore, the overall trend of reward value is to gradually increase and finally converge. This shows that the agent has learned the control strategy to minimize the deviation signal cost.

Table 2. Parameters of DDPG method.

| Parameter                  | Value |
|----------------------------|-------|
| Experience playback scale  | 5000  |
| Experience playback batch  | 32    |
| Discount factor $\gamma$   | 0.9   |
| Network learning rate      | 0.01  |
| Soft update parameter $\tau$ | 0.05  |

Figure 4. The reward function in the training of DDPG and DQN.

4.2. Control effect

In order to verify the effectiveness of the reactor coordinated control system based on DDPG algorithm proposed in this paper, we compare the scheduling method based on DDPG algorithm with the scheduling method based on DQN [7] and the scheduling method based on random strategy control. In the random control strategy, the controlled parameters $R_v(t)$ and $V(t)$ are randomly taken within their value ranges. Table 3 shows the simulation data of the three methods. Among them, the system deviation based on random strategy is 174.3, the system deviation based on DQN strategy is 43.1, and the system deviation of our proposed DDPG strategy is 17.3. Further, we draw the system deviation of the reactor coordinated control system under the DDPG strategy in Figure 5. We can observe that the DDPG algorithm can basically control the four system deviations within the range of nearly 0 after the system runs for about 40 minutes, showing a superior control effect. Therefore, the scheduling method based on DDPG algorithm proposed in this paper is more suitable to solve the control problem of the reactor coordinated control system.
Table 3. Evaluation of different methods

| Strategy | System deviation $F$ |
|----------|----------------------|
| Random   | 174.3                |
| DQN      | 43.1                 |
| DDPG     | 17.3                 |

![Figure 5. The systemic deviation controlled by DDPG.](image)

5. Conclusion

Based on the control problem of reactor coordinated control system, we propose a mathematical model to minimize the system deviation. Considering the uncertainty of the system and the requirements of dynamic real-time optimization, the mathematical model is further transformed into a reinforcement learning control model. Then we use DDPG deep neural network reinforcement learning algorithm to realize interactive learning and problem solving of reactor coordinated control system. In the experimental part, we fully compare the control effects of different deep reinforcement learning algorithms and control schemes formed under random strategies and verify the superior control effect and performance of DDPG algorithm.

References

[1] Zhang, Y., Jiang, L., Cai, Q. (2019) Research on sensitivity analysis of passive system for nuclear power plants based on HFR. Nuclear Power Engineering, 40: 150-154.
[2] Jiang, Q., Zeng, W. (2020) Application of T-S fuzzy switching controller in core power control. Nuclear Power Engineering, 41: 126-130.
[3] Teng, L., Wang, S., Wang, X. (2020) Design of high pressure water jet decontamination device with pressure and flow synchronous control function. Nuclear Power Engineering, 41: 153-157.
[4] Sutton, R.S, Barto, A.G. (1998) Introduction to Reinforcement Learning. MIT Press, Cambridge.
[5] Yang, T., Zhao, L., Li, W., et al. (2020) Reinforcement learning in sustainable energy and electric systems: A survey. Annual Reviews in Control, 49: 145-163.
[6] LeCun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. Nature, 521: 436-444.
[7] Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015) Human-level control through deep reinforcement learning. Nature, 518: 529-533.