Maximum Power Point Tracking Based on Reinforcement Learning in Photovoltaic System

Dingyi Lin, Xingshuo Li, *Member, IEEE*, Shuye Ding
School of Electrical and Automation Engineering, Nanjing Normal University, China
Email: xingshuo.li@njnu.edu.cn

2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES) | 978-1-7281-5672-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/PEDES49360.2020.9379644

***Abstract*— Maximum power point tracking (MPPT) technology is usually used in photovoltaic (PV) systems to extract the maximum power. Although the conventional MPPT techniques are easy to be implemented, they have to tune their control parameters by using trial-and-error method, which is not adaptive to different working conditions. Unlike the conventional MPPT techniques, the reinforcement learning-based MPPT (RL-MPPT) method has advantages of self-learning ability, which is better applicable performance under different weather conditions. To evaluate the RL-MPPT method, the simulations of Standard Test Conditions (STC) and varying irradiance conditions are performed.**

***Index Terms*—Maximum power point tracking, Reinforcement Learning (RL), PV system.**

I. Introduction

Photovoltaic (PV) is considered as one of the most significant sustainable energy sources world-widely. However, how to obtain the maximum available solar energy under different weather conditions is still a challenging problem. Therefore, maximum power point tracking (MPPT) method is used to extract maximum power from the PV system [1], [2].

Although many MPPT methods are proposed, conventional MPPT techniques, such as perturbation and observation (P&O) [3], [4], incremental conductance (INC) [5], hill-climbing (HC) [6], have been widely adopted in practice due to their simple implementation [7]. However, due to the use of fixed step size, tracking accuracy and tracking speed are regarded as two main challenges for these conventional MPPT techniques. By using a large step size, a severe power oscillation may occur when these algorithms converge close to the maximum power point (MPP). By using a small step size, power oscillation is smoothed but at the cost of slow tracking speed.

To handle this thorny obstacle, the modified P&O [8], INC [9], and HC methods [10], have been proposed to solve the tradeoff between tracking accuracy and tracking speed by using the variable step size. The basic principle of these methods is to use a large step size in the transient stage, and a small step in steady-state stage. However, one difficulty is the selection of the scaling factor N for the variable step size. Since the slope of the P-V curve is different under various environment, the simple selection of the scaling factor N fails to minimum power loss in the dynamic tracking process.

Unlike these modified methods, fuzzy logic methods are proposed which can eliminate the power oscillations around the MPP [11]–[13]. However, the effectiveness of these methods relies on expert knowledge and whether the fuzzy parameter has been accurate. Thus these methods need much set-up time, which reflects poor adaptability under different circumstance conditions.

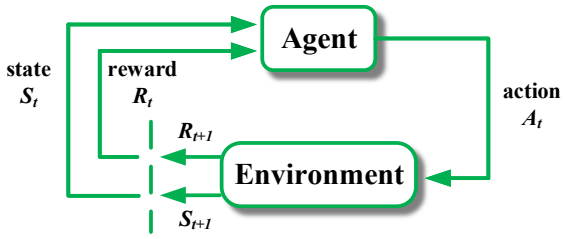

Fig. 1. Interaction between the agent and the environment in MDP.

These aforementioned methods need to tune key parameters to adapt to different weather conditions frequently. In order to settle the issues, this work uses the reinforcement learning (RL) approach to applying in the MPPT. The RL is a machine learning method that action responses are received from the environment to learn behavioral strategies. Thus, the RL-MPPT method has the advantage of achieving the minimum set-up time in different weather conditions. From the recent researches work [14], it has been demonstrated that RL has higher convergence stability and shorter computation time. Further, the RL approach is easy implementation since the controller operating is independent of the PV source's electrical features.

The rest of this paper is arranged as follows. In Section 2, we introduce the concept of RL-MPPT for photovoltaic systems. In Section 3, the results are illustrated in graphs. Meantime, the traditional methods, such as the FLC method and the P&O method, are compared with the RL methods.

II. MPPT based on Reinforcement Learning Method

In order to describe the sequential decision problem, Markov decision process (MDP) can be used as the system framework. In the MDP, a reinforcement learning (RL) method is introduced by [15]. In RL, an object, called agent, actively chooses actions to affect the environment. The environment is an objective that responds passively to the agent. The interaction process between the agent and the environment can be depicted in the MDP (see Fig. 1). The agent is capable of learning to act optimally by experiencing the consequences of actions under the framework.

978-1-7281-5672-9/20/$31.00 ©2020 IEEE

After observing the PV array's environmental conditions, the RL-MPPT determines the perturbation of the working voltage of the PV array (i.e., action), and obtains corresponding rewards through reward function. The positive reward encourages the RL-MPPT to choose better action. The iterations produce a series of positive rewards. Therefore, the agent of RL-MPPT can gradually realize the action selection strategy in the so-called "learning" process. Once the agent understands the strategy, it can automatically adjust the perturbation direction to achieve the operation point to lock at the MPP.

A. State

The state should be sufficiently descriptive and contain all necessary information describing the state of the system and allowing decision-making. On the V-I curve, there is only one optimal point where the maximum power is extracted. In addition, there is only one voltage-current curve for specific solar irradiance. Therefore, a PV source voltage-current pair is sufficient to represent a state, i.e., a working point. The RL-MPPT control method can use two state parameters (V and I) to describe the state of system. The state space is generated as follows:

$$S = \{(V, I)\} \tag{1}$$

where state variable V is discrete from 0 to 25V, with a discrete value of 1V. The state variable I is discrete from 0 to 4A, with a discrete value of 0.1A.

B. Action

The action space A consists of a limited number of actions, which can be implemented to modify the operation point. In the MPPT, the action can be viewed as changing the duty cycle of the DC/DC converter to impact the PV power produced. In a range from 0 to 1, the duty cycle has different optimal values for different circumstance conditions. Although too many actions would improve the accuracy, it would require more larger state space, which increase the difficulty of calculation. Therefore, in order to ensure the validity of the calculation, the action settings should meet the following conditions:

- The action includes positive and negative changes.
- The action needs to contain zero change so that there is no oscillation at the MPP.
- The amplitude of the action should not be too large or too small.

The action space for the controller of the RL-MPPT is composed of five actions and can be expressed by

$$A = \{a \,|-0.05, -0.005, 0, 0.005, 0.05\} \tag{2}$$

where a is the action that determine the change of the duty cycle, and $a = 0$ means that no change is made and previous duty cycle control command is continually used.

C. Reward

When applying RL to solve the MPPT control of photovoltaic arrays, the reward function plays a crucial role. Since a good reward function definition can get the correct feedback for each learning and tracking execution. Meantime it also can improve the efficiency of the learning algorithm.

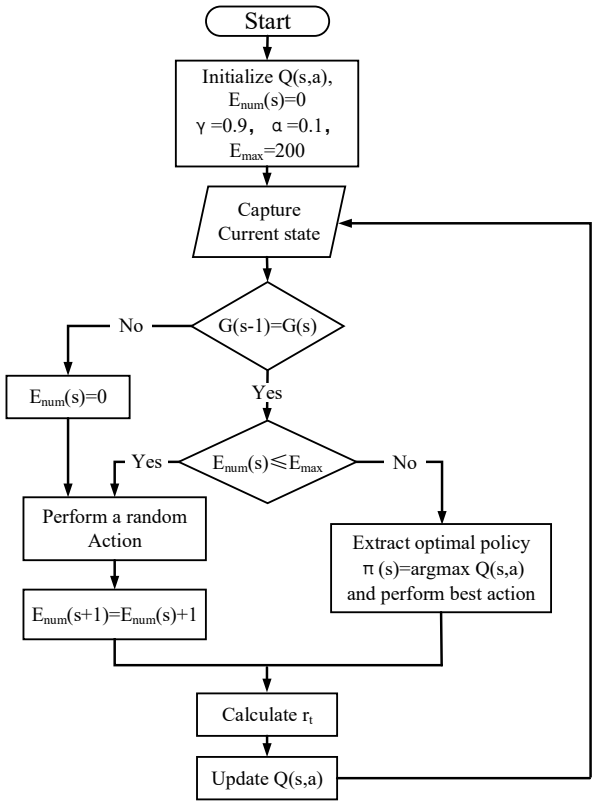

Fig. 2. Q-learning flowchart.

The reward r_t is a direct performance of the interaction between the agent and the environment. Each action makes a state transition, generating a response from the "environment", e.g. reward. Generally, the reward is used to evaluate the performance of the selection action. For the RL-MPPT method, the reward function is defined by

$$r_t = \begin{cases} w_n \frac{\Delta P}{\Delta t}, \frac{\Delta P}{\Delta t} < 0 \\ w_p \frac{\Delta P}{\Delta t}, \frac{\Delta P}{\Delta t} \geq 0 \end{cases} \tag{3}$$

where $\frac{\Delta P}{\Delta t}$ is the discrete change of the power in one cycle, and Δt is the sampling interval. A positive reward will be obtained when the selected action leads to power rise. Negative rewards should certainly represent the punishment for the failure of the agent; that is to say, compared with the zero rewards of tracking photovoltaic array MPP, negative rewards can achieve better learning effects. The reward is established asymmetrically to obviously distinguish between positive and negative states. Thus, weights must differ ($w_p \neq w_n$). In this work, the weights can be set as $w_p = 1.5$, $w_n = 1$.

D. Reinforcement learning algorithm

The used RL method is shown in **Algorithm 1**. As a widely used method in the RL, Q-learning is used in this work. The Q-learning algorithm has an advantage in that it can compare the

978-1-7281-5672-9/20/$31.00 ©2020 IEEE

Algorithm 1: Q-learning algorithm

Input: PV source current I_{pv}, PV source voltage V_{pv};
Output: action a_t;
Initialize: $Q(s,a), a \in A, \forall s \in S; \pi(s); E_{num}(s)$;
Initialize parameters: $\gamma = 0.9$, $\alpha = 0.1$ and $E_{max} = 200$;
repeat
  detect state: s_t;
  if $G(s-1) \neq G(s)$ **then**
    $E_{num}(s)$=0
  end
  if $E_{num}(s) < E_{max}$ **then**
    randomly choose a_t, $a_t \in A$
    $E_{num}(s+1) \leftarrow E_{num}(s)+1$
  else
    extract optimal policy: $\pi(s) = arg\max Q(s_t, a_t)$
  end
  output a_t, then detect next state: s_{t+1};
  calculate the reward r_{t+1} according to eq.(3);
  $Q(s_t,a_t) \leftarrow$
  $Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right]$;
  $s_t \leftarrow s_{t+1}$;
until *end*;

expected effects of all available actions without considering the environmental model, and can deal with random conversions and rewards without any modification. Theoretical research shows that Q-learning can quickly and effectively learn an optimal strategy for any MDP problem with a limited state space and action space [15]. The Q-learning algorithm considers all possible actions when iteratively update, and selects the action with largest value for the next step. The update rules of the Q-learning algorithm are:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right] \quad (4)$$

where α is learning rate , γ is discount factor.

This work applies the Q-learning algorithm to MPPT control, and its flowchart is presented in Fig. 2. Though interacting with its environment, the agent can learn a system's behavior or policy to optimize the system's performance. Thus, it is necessary to introduce the learning policy which divides into two parts: exploration process and exploitation process. In the exploration, the agent randomly chooses an action to explore the state-action space and receive a corresponding reward (i.e., Q-value). The agent explores a finite number of rounds, which should be great enough to make the agent learn enough state-action visit experience. However, it cannot be too great to maintain the computational efficiency of the method. In the exploitation (i.e., greedy policy), the agent picks for every state the action with the highest Q-value to perform the optimal procedure.

The RL-MPPT controller (i.e., the agent) needs to be initialized with α, γ, and maximum number of explorations E_{max}, and Q-table. Then, the agent captures current state s_t. The $G(s)$ represent the solar irradiance. If the solar irradiance changes, the state action space will fails to effectively apply. Therefore it needs to enter the exploration process again. Finally, the agent choose the action based on the learning policy and calculate the reward to update the $Q(s,a)$.

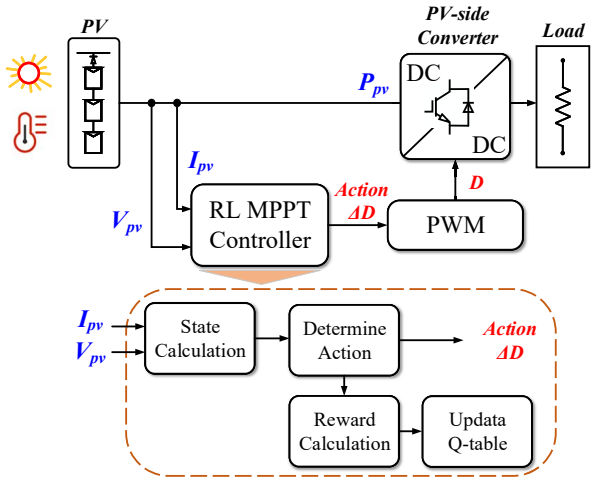

Fig. 3. Simulation layout of RL-MPPT method.

III. Results

In this work, a 60-W PV module (MSX-60W) is selected for simulation. TABLE I shows main parameters of PV module. These parameters are: the current of MPP and voltage of MPP are 3.5A and 17.1V, maximum power of PV module is 60W. The sampling time for the MPPT algorithm, T_p, is set as 0.01 s.

TABLE I
Main Product Parameters of the PV Module MSX-60W

Parameter	Symbol	Value
Maximum power	P_{mpp}	$60W$
Voltage at MPP	V_{mpp}	$17.1V$
Current at MPP	I_{mpp}	$3.5A$
Open-circuit voltage	V_{oc}	$21.1V$
Short-circuit current	I_{sc}	$3.8A$
Temperature coefficient of V_{oc}	K_v	$-80mV/°C$
Temperature coefficient of I_{sc}	K_i	$0.065\%/°C$

In order to evaluate the effectiveness of the application of RL on the MPPT problem, this paper is to verify the RL-MPPT control method power efficiency under Standard Test Conditions (STC) conditions and varying operating conditions. The layout of simulation model is presented in Fig. 3. The simulation model's main components are the PV module, boost converter with the RL-MPPT controller, and load. In this part, the RL-MPPT method is compared with two traditional methods: the FLC method and the P&O method.

The P&O method is one of the most popular methods due to its easy implementation. The principle of the P&O method is to periodically perturb (increase or decrease duty cycle). Then, the instantaneous power $P(k)$ is compared with the previous power $P(k-1)$. When the perturbation increases power, it decides the next perturbation direction as the same as the previous direction. Otherwise, it will turn back the perturbation direction. For comparing under the same test conditions, the perturb time interval T_p of the P&O method is 0.01s.

978-1-7281-5672-9/20/$31.00 ©2020 IEEE

Unlike the P&O method, The FLC method uses "degrees of truth" which can lead to robust and fast tracking speed. However, it is not easy to tune the key parameters to implement in MPPT. Usually, the FLC method is divided into three processes [7], which is demonstrated in Fig. 4. In the process of fuzzification, digital input variables firstly are transformed into equivalent input fuzzy sets. In the process of inference, the fuzzy sets is transformed into the output fuzzy set by using the fuzzification block according to the membership function, which is decided by the expert's knowledge. Finally, these fuzzy sets are transformed into digital output variables in the process of defuzzification. Generally, digital output variables is the duty cycle in the FLC method. However, there are many numerical input variables that can be converted. In this work, the change in error ΔE and the error E, which can be determined by the slope of P-V curve, are used as the digital input variables [7].

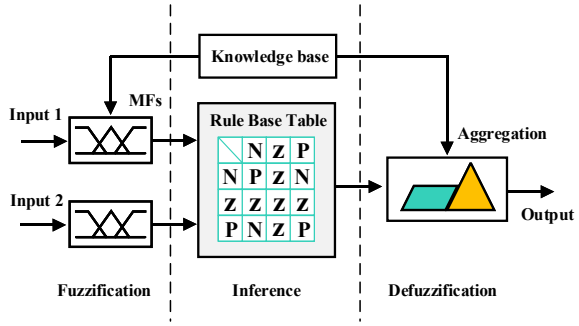

Fig. 4. Structure of the fuzzy logic controller.

A. Testing results based on STC

The RL-MPPT controller and FLC controller are tested under STC, which is 1000 W/m^2 solar irradiance and 25°C temperature.

Fig. 5 (a) shows the output power of RL-MPPT method, which varies randomly in the transient stage. The random power oscillations indicate the agent is learning and exploring. In the exploration process, the agent repeats to randomly choose actions until the maximum exploration number. After the exploration process is ended, the RL-MPPT controller converges to MPP with zero oscillations of power in the steady-state stage, i.e., the exploitation process. Fig. 5 (b) and (c) present the reward and duty cycle of RL-MPPT method, respectively. Since reward weights differ ($w_p \neq w_n$), the positive value is obvious rather than negative value in the random exploration process. The duty cycle is first randomly explored, and then prone to the optimal point.

Fig. 6 (a) shows the output power of the FLC method. Although the FLC method can fast track the MPP under STC, the oscillation of power in the steady-state stage is larger than that of the RL-MPPT method, which leads to power loss.

The output power of the P&O method are shown in Fig. 7 (a). In order to further compare the performance of the P&O method and the RL method, two different fixed step sizes

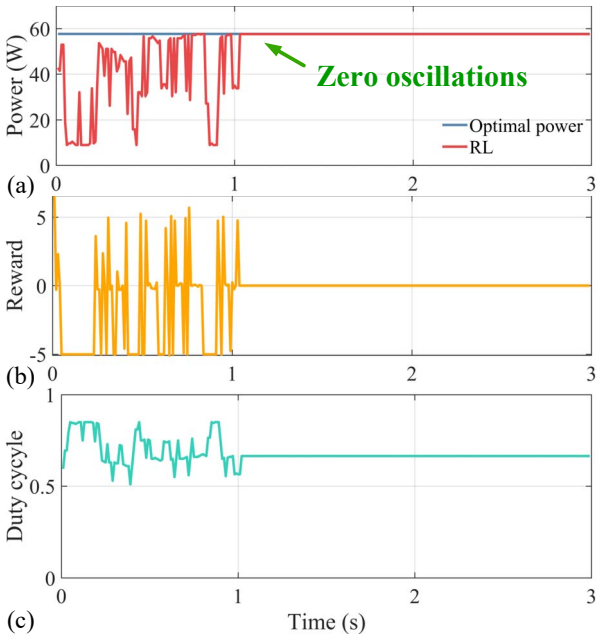

Fig. 5. Simulation results of RL method under STC.

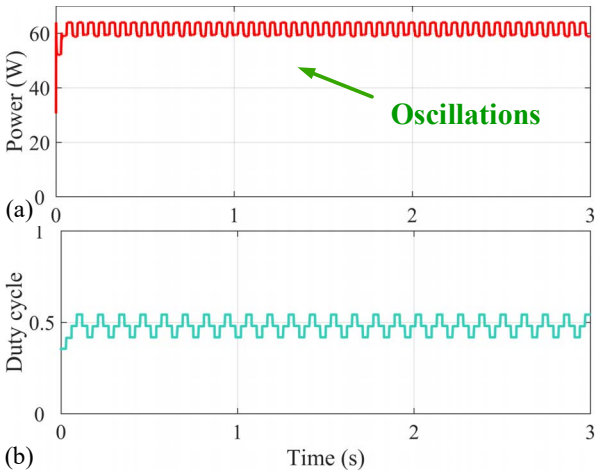

Fig. 6. Simulation results of FLC method under STC.

were selected for simulation. Although using a small step size (step=0.5) can reduce the power oscillation near the MPP, it does so at the expense of tracking speed. In contrast, the large step size (step=2) is capable of fast-tracking speed. However, there is a larger oscillation around the MPP. Therefore, the method needs to the trade-off between the tracking speed and power oscillation. In uniform irradiance conditions, the methods is capable of achieving high efficiency.

B. Testing results based on varying operating conditions

The RL-MPPT controller and FLC controller are tested under varying operating conditions, which is kept on 25°C and the irradiance is varied from 1000 W/m^2 to 500 W/m^2 and then to 800 W/m^2, within 3 s.

978-1-7281-5672-9/20/$31.00 ©2020 IEEE

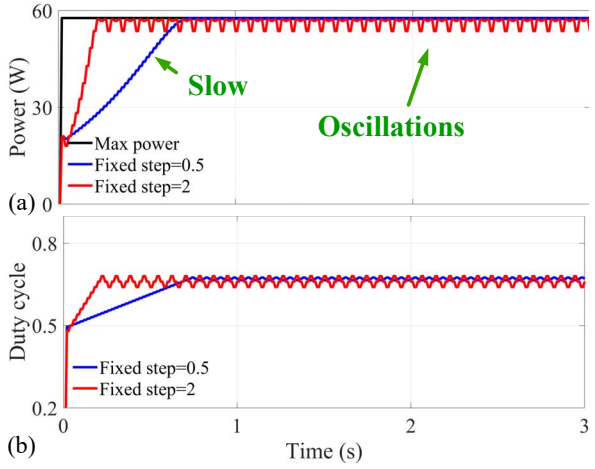

Fig. 7. Simulation results of P&O method under STC.

As shown in Fig. 8, the simulation results of RL-MPPT method show that the algorithm can converge to MPP in different irradiance. In addition, with solar irradiance changes, the RL-MPPT controller needs to explore the new situation and learn how to achieve optimal performance in the presence of continuous disturbances. However, the duration of the exploration process is different since the agent takes a randomly explore strategy. Thus, if the agent frequently chooses the wrong action to explore, it leads to more computationally time and power loss. After convergence and finding the optimal point, there is no oscillation of duty cycle.

Fig. 9 (a) shows the output power of the FLC method under varying operating conditions. In terms of tracking speed and efficiency, comparing the two approaches shows that when new environmental conditions have occurred, the RL-MPPT method is slower than the FLC method since the method learns from the exploration process. However, the FLC method's tracking speed is faster than the RL-MPPT method but produces large oscillations resulting in big power loss. In the view of long term operation conditions, the RL-MPPT method can improve efficiency rather than the FLC method.

Additionally, the steady-state oscillations of the FLC method is different under diverse solar irradiance. As demonstrated in Fig. 9 (a), the power oscillation under 1000 W/m^2 is larger than under both 500 W/m^2 and 800 W/m^2. This can be explained by the fixed fuzzy rules. For specific environmental conditions, one fixed fuzzy rules can show fast convergence and small oscillation, however, it is not suitable for other conditions.

As presented in Fig. 10, the simulation results of the P&O method demonstrate that the methods can fast converge to MPP in varying irradiance by using a large step. However, when it converges to MPP, power oscillating will be occurred, which leads to energy loss. In the steady stage of irradiance, the small step could be in favor of high efficiency. However, the small step is too small to track the MPP under quickly varying irradiance, which leads to power lose. In long-term operation conditions, the RL-MPPT method can improve efficiency rather than the P&O method since RL-MPPT can fast converge to MPP through previous learning experience.

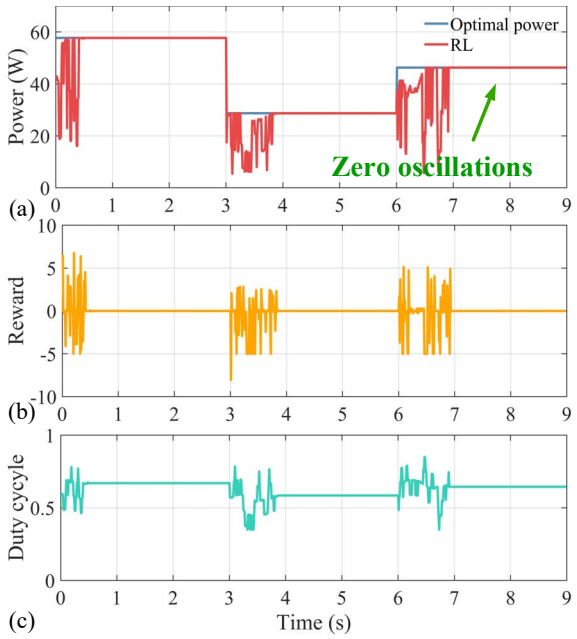

Fig. 8. Simulation results of RL method under varying irradiance.

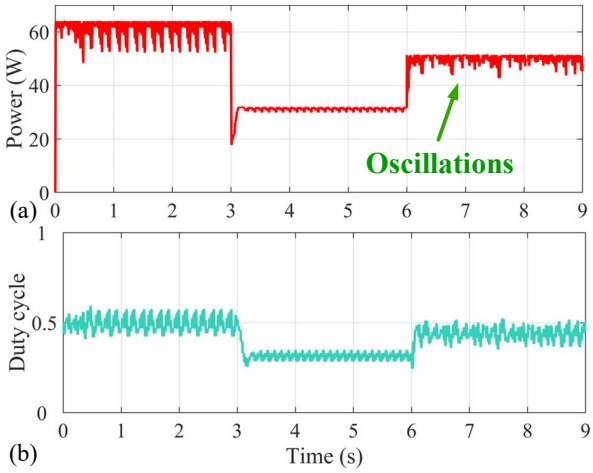

Fig. 9. Simulation results of FLC method under varying irradiance.

IV. Conclusion

Under the MDP framework, this work presents a method that combines reinforcement learning and the MPPT method in PV sources. The RL-MPPT method uses two-state parameters to distinguish whether a working point is nearby MPP. Meantime, the RL method can converge to MPP by interacting with the environment and receiving corresponding rewards. In order to estimate the efficiency and verify the effectiveness, the RL-MPPT method has been implemented and simulated in the STC and varying operating conditions. The results show that there is zero oscillation for the RL method in the steady-state stage. In addition, the RL method is also compared to the FLC

978-1-7281-5672-9/20/$31.00 ©2020 IEEE

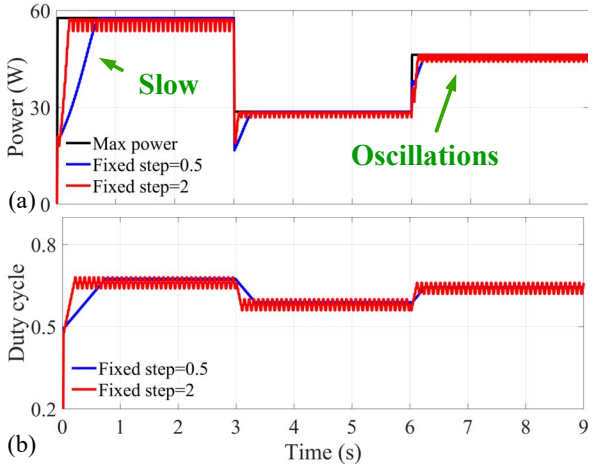

Fig. 10. Simulation results of P&O method under varying irradiance.

method and the P&O method under two scenarios showing better results in terms of efficiency, and perform optimally without any additional set up.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 51977112 and the Graduate Science and Innovation Projects in Jiangsu Province under Grant KYCX19-0808. This work was also supported in part by the Research Start Fund of Nanjing Normal University under Grant 184080H202B232.

REFERENCES

[1] L. Jiang, R. Srivatsan, and D. Maskell, "Computational intelligence techniques for maximum power point tracking in pv systems: A review," *Renewable Sustainable Energy Reviews*, vol. 85, pp. 14–45, 2018.

[2] X. Li, H. Wen, Y. Hu, and L. Jiang, "Drift-free current sensorless mppt algorithm in photovoltaic systems," *Solar Energy*, vol. 177, pp. 118–126, 2019.

[3] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "Optimization of perturb and observe maximum power point tracking method," *IEEE Transactions on Power Electronics*, vol. 20, pp. 963–973, 2005.

[4] M. Elgendy, B. Zahawi, and D. Atkinson, "Assessment of perturb and observe mppt algorithm implementation techniques for pv pumping applications," *IEEE Transactions on Sustainable Energy*, vol. 3, pp. 21–33, 2012.

[5] ——, "Assessment of the incremental conductance maximum power point tracking algorithm," *IEEE Transactions on Sustainable Energy*, vol. 4, pp. 108–117, 2013.

[6] S. B. Kjaer, "Evaluation of the "hill climbing" and the "incremental conductance" maximum power point trackers for photovoltaic power systems," *IEEE Transactions on Energy Conversion*, vol. 27, pp. 922–929, 2012.

[7] X. Li, Q. Wang, H. Wen, and W. Xiao, "Comprehensive studies on operational principles for maximum power point tracking in photovoltaic systems," *IEEE Access*, vol. 7, pp. 121 407–121 420, 2019.

[8] A. Abdelsalam, A. Massoud, S. Ahmed, and P. Enjeti, "High-performance adaptive perturb and observe mppt technique for photovoltaic-based microgrids," *IEEE Transactions on Power Electronics*, vol. 26, pp. 1010–1021, 2011.

[9] Q. Mei, M. Shan, L. Liu, and J. Guerrero, "A novel improved variable step-size incremental-resistance mppt method for pv systems," *IEEE Transactions on Industrial Electronics*, vol. 58, pp. 2427–2434, 2011.

[10] W. Xiao and W. G. Dunford, "A modified adaptive hill climbing mppt method for photovoltaic power systems," *2004 IEEE 35th Annual Power Electronics Specialists Conference (IEEE Cat. No.04CH37551)*, vol. 3, pp. 1957–1963 Vol.3, 2004.

[11] X. Li, H. Wen, L. Jiang, W. Xiao, Y. Du, and C. Zhao, "An improved mppt method for pv system with fast-converging speed and zero oscillation," *IEEE Transactions on Industry Applications*, vol. 52, pp. 5051–5064, 2016.

[12] R. T. D. Rezoug, Mohamed; Chenni, "Fuzzy logic-based perturb and observe algorithm with variable step of a reference voltage for solar permanent magnet synchronous motor drive system fed by direct-connected photovoltaic array," *Energies*, vol. 11, 02 2018.

[13] X. Li, H. Wen, Y. Hu, and L. Jiang, "A novel beta parameter based fuzzy-logic controller for photovoltaic mppt application," *Renewable Energy*, vol. 130, pp. 416–427, 2019.

[14] P. Kofinas, S. Doltsinis, A. I. Dounis, and G. A. Vouros, "A reinforcement learning approach for mppt control method of photovoltaic sources," *Renewable Energy*, vol. 108, no. AUG, pp. 461–473, 2017.

[15] R. Sutton and A. Barto, "Introduction to reinforcement learning," 1998.

978-1-7281-5672-9/20/$31.00 ©2020 IEEE