Int. J. Electrochem. Sci., 15 **(2020)** 9917 – 9932, doi: 10.20964/2020.10.63

International Journal of
ELECTROCHEMICAL
SCIENCE
www.electrochemsci.org

Q-Learning based Maximum Power Point Tracking Control for Microbial Fuel Cell

Li-ping Fan[1,2 *], *Xiang Feng*[1,2]

[1] College of Information Engineering, Shenyang University of Chemical Technology, Shenyang, 110142 China
[2] Key Laboratory of Industry-Environment-Resource Collaborative Control and Optimization Technology of Liaoning Province, Shenyang University of Chemical Technology, Shenyang 110142, China
[*]E-mail: flpsd@163.com

Received: 1 July 2020 / *Accepted:* 6 August 2020 / *Published:* 31 August 2020

Microbial fuel cell (MFC) is a promising technology for wastewater treatment with simultaneous bioenergy production. To improve the power generation efficiency of MFCs, maximum power point tracking control is a good choice. Three kinds of Q-Learning-based maximum power point tracking control scheme based on ε-greedy exploration, Boltzmann exploration and greedy policy are proposed for MFCs. The results show that the maximum power point tracking control based on Q-Learning has better power tracking capabilities than perturbation and observation method. With the introduction of Q-Learning based on greedy policy, the time required for MFC to stabilize at the maximum power point is greatly shortened by setting the action list of Q-Learning reasonably. In this case, the whole process from start-up to stabilization at the maximum power point was 42.9% faster than that of MFC using ε-greedy exploration, and 50% faster than that of MFC using Boltzmann exploration. Q-Learning algorithm based on greedy policy is an effective method to realize MPPT in MFC system.

Keywords: greedy policy; maximum power point tracking; microbial fuel cell; perturbation and observation; Q-Learning

1. INTRODUCTION

The global energy crisis caused by the depletion of fossil fuels is escalating due to the rapid increase in demand. At the same time, the environmental pollution caused by burning fossil fuels is not to be underestimated. Developing renewable energy is an effective way to solve energy crisis and environmental pollution. Microbial fuel cell (MFC) is a promising technology for wastewater treatment with simultaneous bioenergy production [1-3]. The greatest advantage of MFCs is that they convert the chemical energy of pollutants into electrical energy directly and generate almost no pollutants in the power generation process. The application of MFC will solve the problem of environmental pollution and energy crisis for mankind at the same time [4].

Although there have been some reports of MFC implementation of various applications [5], there are still a lot of challenges in the successful application of MFC in real environment. Compared to other alternative energy systems such as solar and wind, MFC is a low power system due to its thermodynamic limitation [6]. Various physical, chemical and biological approaches have been explored to improve electricity generation [7]. For example, through the development and modification of electrode material [8], exchange membrane [9] and cathode catalyst [10], the power generation capacity of MFC can be improved. However, from the user's point of view, once the MFC construction is completed, such methods are often not feasible. In this case, using advanced control method to further improve the performance of MFC is an alternative way [11].

To improve the power generation performance of MFCs, some maximum power point tracking (MPPT) technologies have been proposed in literatures. Among all the conventional MPPT, perturbation and observation (P/O) method is the most widely used algorithm because of its simplicity and robustness. The P/O method works by creating a perturbation and observing its effect [12]. By comparing the previous output power with the current output power, it is straightforward to decide the direction of next step that would increase the power towards MPP [13].

In the last decade, there have been some studies on power control of MFCs using P/O method. P/O method was used in a single chamber air-breathing cathode MFC to optimize energy harvesting, and laboratory tests confirmed that it can provide stable long-term power production [14]. Park et al [15] proposed an integrated control system for solid anolyte MFC that can perform real-time MPPT with P/O algorithm. Study results also demonstrated that P/O based MPPT was able to reduce the start-up time and minimize the internal resistance of MFC, so that the energy loss related with anode and cathode can be reduced [16], and noticeable improvement in MFC performance was observed [17]. Compared with other methods, P/O method is simpler to tune [18]. However, the steady-state fluctuation caused by the P/O algorithm is usually very serious, so it is difficult to make the MFC accurately stable at the maximum power point, and it will deviate from the maximum power point when it is disturbed. In addition, the P/O method may converge to a local optimal power point rather than its actual maximum power point, which violates the original intention of maximum power tracking control [19].

Reinforcement Learning is a machine learning method that uses behaviors received from the environment to learn behavioral strategies. It emerges as a powerful data-driven method for solving complex control problems. Reinforcement Learning technology have been used in some recent studies to solve MPPT problems in some new energy systems such as wind energy conversion systems [20,21], photovoltaic array [22,23] and hybrid electric vehicle [24-26].

This work aims to apply Reinforcement Learning algorithm to MFC system to realize MPPT control. Three kinds of Q-Learning-based MPPT (QLMPPT) were proposed for MFCs and the control effects were compared with those of MFCs with the conventional P/O algorithm and an improved P/O algorithm.

The remainder of the paper was organized as follows. In Section 2, the mathematical formulation of the two-chamber MFC was introduced. A detailed implementation of QLMPPT was proposed in Section 3. Results and discussion were presented in Section 4. Conclusion and future research directions were addressed in Section 5.

2. MODELLING OF MICROBIAL FUEL CELL

Modeling and simulation are effective ways to deeply understand the operation process and verify the effectiveness of optimal control schemes of MFC [27]. Electricity generation in MFCs has been modelled by a few researchers. Pinto et al proposed a two-population model describing the competition of anodophilic and methanogenic microbial populations for a common substrate in a single-chamber MFC [28]. Picioreanu et al proposed a computational model for MFCs based on redox mediators with several populations of suspended and attached biofilm microorganisms and multiple dissolved chemical species [29]. Marcus et al developed a model describing the biofilm as a conductive solid matrix, which is based on that microbe can allow direct transfer of the electrode [30]. All the models discussed above so far are limited to the analysis of a single electrode (anode) of a MFC. It is important to develop coupled models that include the phenomena occurring at both anode and cathode. Esfandyari et al proposed a dynamic model for two-chamber batch MFC with pure culture of Shewanella based on the direct transfer of electron, which was described based on Marcus‘s model [31]. Zeng et al developed a model to simulate both steady and dynamic behavior of a two-chamber continuous MFC by integrating biochemical reactions, Butler–Volmer expressions and mass/charge balances [32]. Zeng's model, which includes key physical quantities such as voltage, power density and fuel concentration, is a comprehensive model for the reaction process of MFCs. therefore, this model was used to describe MFCs in this research.

In the anode compartment, acetate is oxidized under anoxic conditions by the reaction of an eight-electron transfer which can be described as:

$$\left(\mathrm{CH_2O}\right)_2 + 2\mathrm{H_2O} \rightarrow 2\mathrm{CO_2} + 8\mathrm{H^+} + 8\mathrm{e^-} \tag{1}$$

In the cathode compartment, the reduction of dissolved oxygen can be described as:

$$\mathrm{O_2} + 4\mathrm{e^-} + 2\mathrm{H_2O} \rightarrow 4\mathrm{OH^-} \tag{2}$$

The reaction rates of the anode and cathode chamber can be described by Butler-Volmer expression as:

$$r_1 = k_1^0 \exp\left(\frac{\alpha F}{RT}\eta_a\right)\frac{C_{\mathrm{AC}}}{K_{\mathrm{AC}} + C_{\mathrm{AC}}} X \tag{3}$$

$$r_2 = -k_2^0 \frac{C_{\mathrm{O_2}}}{K_{\mathrm{O_2}} + C_{\mathrm{O_2}}} \exp\left[(\beta - 1)\frac{F}{RT}\eta_c\right] \tag{4}$$

where F denotes the Faraday constant; R denotes the gas constant; T denotes the operating temperature; C_{AC} denotes the concentrations of acetate in the anode compartment; X denotes the concentrations of biomass in the anode compartment; $C_{\mathrm{O_2}}$ denotes the Dissolved Oxygen (DO) concentration in the cathode compartment; η_a and η_c are the anodic over potential and the cathodic over potential, respectively; α and β are the charge transfer coefficients of the anodic reaction and the cathodic reaction, respectively; k_1^0 and k_2^0 are the rate constants of the anodic reaction and the cathodic reaction at standard conditions; K_{AC} is the half velocity rate constant for acetate; K_{O_2} is the half velocity rate constant for DO.

Assuming both the anode chamber and the cathode chamber can be regarded as continuous stirring tank reactors (CSTRs), the mass balance equations of the four components (acetate, dissolved CO2, hydrogen ion and biomass) in the anode can be described as:

$$V_{\mathrm{a}}\frac{\mathrm{d}C_{\mathrm{AC}}}{\mathrm{d}t}=Q_a\left(C_{\mathrm{AC}}^{\mathrm{in}}-C_{\mathrm{AC}}\right)-A_{\mathrm{m}}r_1 \tag{5}$$

$$V_{\mathrm{a}}\frac{\mathrm{d}C_{\mathrm{CO_2}}}{\mathrm{d}t}=Q_a\left(C_{\mathrm{CO_2}}^{\mathrm{in}}-C_{\mathrm{CO_2}}\right)+2A_{\mathrm{m}}r_1 \tag{6}$$

$$V_{\mathrm{a}}\frac{\mathrm{d}C_{\mathrm{H}}}{\mathrm{d}t}=Q_{\mathrm{a}}\left(C_{\mathrm{H}}^{\mathrm{in}}-C_{\mathrm{H}}\right)+8A_{\mathrm{m}}r_1 \tag{7}$$

$$V_{\mathrm{a}}\frac{\mathrm{d}X}{\mathrm{d}t}=Q_{\mathrm{a}}\frac{\left(X^{\mathrm{in}}-X\right)}{f_{\mathrm{x}}}+A_{\mathrm{m}}Y_{\mathrm{ac}}r_1-V_{\mathrm{a}}K_{\mathrm{dec}}X \tag{8}$$

The mass balance equations in the cathode can be described as:

$$V_{\mathrm{c}}\frac{\mathrm{d}C_{\mathrm{O_2}}}{\mathrm{d}t}=Q_{\mathrm{c}}\left(C_{\mathrm{O_2}}^{\mathrm{in}}-C_{\mathrm{O_2}}\right)+r_2A_{\mathrm{m}} \tag{9}$$

$$V_{\mathrm{c}}\frac{\mathrm{d}C_{\mathrm{OH}}}{\mathrm{d}t}=Q_c\left(C_{\mathrm{OH}}^{\mathrm{in}}-C_{\mathrm{OH}}\right)-4r_2A_{\mathrm{m}} \tag{10}$$

$$V_{\mathrm{c}}\frac{\mathrm{d}C_{\mathrm{M}}}{\mathrm{d}t}=Q_{\mathrm{c}}\left(C_{\mathrm{M}}^{\mathrm{in}}-C_{\mathrm{M}}\right)+N_{\mathrm{M}}A_{\mathrm{m}} \tag{11}$$

where N_M is the flux of M^+ ions transferred from the anode chamber to cathode chamber through the proton exchange membrane, which can be derived by:

$$N_{\mathrm{M}}=\frac{3600i_{\mathrm{fc}}}{F} \tag{12}$$

The charge balance equations at the anode and cathode can be described as:

$$C_a\frac{\mathrm{d}\eta_a}{\mathrm{d}t}=3600i_{\mathrm{fc}}-8Fr_1 \tag{13}$$

$$C_{\mathrm{c}}\frac{\mathrm{d}\eta_{\mathrm{c}}}{\mathrm{d}t}=-3600i_{\mathrm{fc}}-4Fr_2 \tag{14}$$

In the above equation, the subscripts ‘a’, ‘c’, and ‘in’ stand for the anode, the cathode and the feed flow, respectively; V denotes the volume of the reaction chamber; Q denotes the feed flow rate; A_{m} is the cross-section area of membrane; f_{x} denotes the reciprocal of the wash-out fraction; Y_{ac} denotes the bacterial yield; K_{dec} denotes the decay constant for acetate utilizers; C_{M} denotes the concentration of M^+ ions, and i_{fc} denotes the current density.

Ignoring the ohmic drops in the current collectors and electric connections, the internal resistance of MFC is only related to the membrane and the solution, then the output voltage U_{fc} of MFC can be expressed as:

$$U_{\mathrm{fc}}=U^0-\eta_{\mathrm{a}}+\eta_{\mathrm{c}}-\left(\frac{d^{\mathrm{m}}}{k^{\mathrm{m}}}+\frac{d_{\mathrm{cell}}}{k^{\mathrm{aq}}}\right)i_{\mathrm{fc}} \tag{15}$$

Based on the above mathematical model of MFC, a simulation model of a dual-chamber MFC was established in MATLAB/Simulink, which can be used to simulate the operating state of MFCs under various conditions. The main parameters of the MFC model used in this work are shown in Table 1, which were derived from Zeng's model.

The output power of MFC varies with the reaction conditions. However, many studies have found that no matter how the reaction conditions change, there is always a peak point on the power curve of MFC, and the peak point is the maximum power point. As shown in Figure 1, the maximum power point of MFC which was used in this experiment was about 2.1 mW·m^{-2}. According to the relevant circuit knowledge, this maximum power point can be achieved when the external resistance is equal to the internal resistance. When the internal and external resistances are not equal, MFC will lead to about 50% energy loss [33, 34]. So, MPPT is one of the best techniques to dynamically extract maximum possible power from MFC and decrease energy loss.

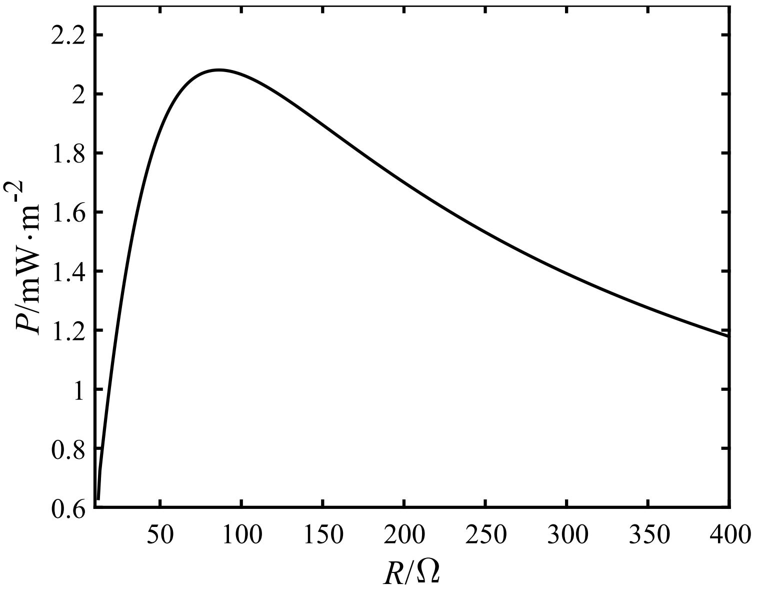

Figure 1. Power density varying with external resistance

Table 1. Parameters of MFC model

Symbol	Description	Unit	Value
F	Faraday’s constant	$C \cdot mol^{-1}$	96485.4
R	Gas constant	$J \cdot mol^{-1} \cdot K^{-1}$	8.3144
T	Temperature	K	303
k^{m}	Electrical conductivity of membrane	$\Omega^{-1} \cdot m^{-1}$	17
d^{m}	Thickness of membrane	m	1.778×10^{-4}
k^{aq}	Electrical conductivity of the aqueous solution	$\Omega^{-1} \cdot m^{-1}$	5
d^{cell}	Distance between anode and cathode in the cell	m	2.2×10^{-2}
C_{a}	Capacitance of anode	$F \cdot m^{-2}$	4×10^{2}
C_{c}	Capacitance of cathode	$F \cdot m^{-2}$	5×10^{2}
V_{a}	Volume of anode compartment	m^{3}	5.5×10^{-5}
V_{c}	Volume of cathode compartment	m^{3}	5.5×10^{-5}
A_{m}	Area of membrane	m^{2}	5×10^{-4}
Y_{ac}	Bacterial yield	Dimensionle	0.05
K_{dec}	Decay constant for acetate utilizers	h^{-1}	8.33×10^{-4}
f_{x}	Reciprocal of wash-out fraction	Dimensionle	10
Q_{a}	Flow rate of fuel feed to anode	$m^{3} \cdot h^{-1}$	2.25×10^{-5}
Q_{c}	Flow rate feeding to cathode compartment	$m^{3} \cdot h^{-1}$	1.11×10^{-3}
C_{AC}^{in}	Concentration of acetate in the influent of anode compartment	$mol \cdot m^{-3}$	1.56
C_{CO2}^{in}	Concentration of CO_2 in the influent of anode compartment	$mol \cdot m^{-3}$	0
X^{in}	Concentration of bacteria in the influent of anode compartment	$mol \cdot m^{-3}$	0
C_{H}^{in}	Concentration of H^{+} in the influent of anode compartment	$mol \cdot m^{-3}$	0
C_{O2}^{in}	Concentration of dissolved O_2 in the influent of cathode	$mol \cdot m^{-3}$	0.3125
C_{M}^{in}	Concentration of M^{+} in the influent of cathode compartment	$mol \cdot m^{-3}$	0
C_{OH}^{in}	Concentration of OH^{-} in the influent of cathode compartment	$mol \cdot m^{-3}$	0
U^{0}	Cell open circuit potential	V	0.77
k_{1}^{0}	Forward rate constant of anode reaction at standard condition	$mol \cdot m^{-2} \cdot h^{-1}$	0.207
k_{2}^{0}	Forward rate constant of cathode reaction at standard condition	$mol \cdot m^{-4} \cdot h^{-1}$	3.288×10^{-5}
K_{AC}	Half velocity rate for acetate	$mol \cdot m^{-2}$	0.592
K_{O_2}	Half velocity rate for dissolved oxygen	$mol \cdot m^{-2}$	0.004
α	Charge transfer coefficient of anode	Dimensionle	0.051
β	Charge transfer coefficient of cathode	Dimensionle	0.063

3. REINFORCEMENT LEARNING-BASED MPPT

3.1 Introduction of Q-Learning

Reinforcement Learning is a machine learning method to understand and automate goal-directed learning and decision making. Reinforcement Learning was designed to infer closed-loop policies for stochastic optimal control problems from a sample of trajectories gathered from interaction with the real system or from simulations [35]. Generally, a basic Reinforcement Learning architecture is composed of two essential elements: an agent and an environment. The optimized controller called as agent is not told which actions a to take, but instead must discover which actions yield the most reward *r* in the future by directly interacting with the current state *s* of the controlled object called as environment. Expectation of total reward is defined by an action-value function *Q*(*s*,*a*), which estimates how good it is for the agent to perform a given action *a* in a given state *s*.

Q-Learning is one of the most popular algorithms that perform Reinforcement Learning, and it is a typical model independent algorithm. The core of the Q-Learning algorithm is a value iteration update of the value function. Under the Q-Learning algorithm, the goal is to achieve the goal state and obtain the highest income. Once the goal state is reached, the final income remains unchanged. The Q-value for each state-action pair is initially chosen by the designer and later, it is updated each time an action is executed and a reward is received, based on the following expression:

$$Q\left(s_k,a_k\right)=Q\left(s_k,a_k\right)+\alpha\left[r_k+\gamma\max_a Q\left(s_{k+1},a\right)-Q\left(s_k,a_k\right)\right] \quad (16)$$

where, r_k is the reward given at time *t*, γ is the discount factor determining the current reward to be received in the future updates, α is the learning rate which is selected by a tradeoff between speed and coverage. At each episode of update, the agent observes the current state s_k and select an action a_t according to the policy. Then the subsequent state s_{t+1} is observed with a reward r_k given by the environment, the current action-value function $Q(s_k, a_k)$ and the maximum value $\max_a Q(s_{k+1},a)$ of the next state s_{k+1} will be used to update the *Q* function according to Equation (16).

The state, which describes the conditions of the system, is critical to the performance of Q-Learning.

3.2 Parameters of Q-Learning

3.2.1. State Space

Only if the state space *S*, action list *A* and reward function $r=f(s_k,\ a_k\ s_{k+1},)$ are all well set, the Q learning algorithm can make MFC track the MPP accurately and quickly. When defining a state space, many aspects need to be considered. Considering that electrical signals are easy to measure, voltage and current of MFC are used as states. However, a continuous state space is usually difficult to handle. On the other hand, too sparse state space may lead to misjudgment of state, resulting in insufficient decision-making capabilities, oscillations between states, and non-optimal policies. Therefore, the voltage and current were normalized by using open circuit voltage and short circuit current. The voltage was discretized into 6 states between 0 to 0.6 V with equal interval. To improve the tracking speed, when the current is less than or equal to 4 mA, all are set to one state, and current which is more than 4 mA and less than 12 mA will be discretized into 9 states with equal interval.

3.2.2. Action List

Boost converter was used to adjust the equivalent resistor of the external resistance to affect the produced power. Therefore, duty cycle increment was selected as actions. Actions varied between 0 and 1 according to the duty cycle. To keep the proposed method efficient, the action list contained 3 actions.

The initialized table $Q(s,a)$ has zero *action-values*. The first action is executed by default. As long as the first action is positive, the output duty cycle will increase all the time without oscillation, so the tracking speed will be accelerated. Positive and negative increment are needed to ensure the algorithm can track in different direction. After finding the MPP, a zero increment is needed to avoid oscillation in states.

Since the power density is low when the MFC is started, a positive increment of duty cycle is required to ensure that the operating point will move towards the MPP, so the first one of the action list is set to a positive value. With an original blank action value table that all $Q(s,a)$ at state s was equal to 0, the algorithm chose the first index of actions, and made a increment of duty cycle. It can make the tracing process move in the same direction with no need to explore all the states with all the actions. Besides, according to Figure.1, there is only one MPP of an MFC, so this setting can save a lot of time without lose the accuracy. Therefore, the actions list was defined as: A = [0.1 -0.1 0].

3.2.3. Reward Function

For each applied action, the system reacts and performs a state transition to generate a response that is monitored as a return to the environment. Rewards are needed to include positive and negative value to ensure that the impact of the action is always proportional to the power change. When power increment is positive, the corresponding reward is positive; on the contrary, when the power increment is negative, the corresponding reward is negative. When the MFC is stable at the MPP, the reward becomes zero because the power change is neither positive nor negative. The reward was defined as:

$$r = \begin{cases} \Delta P * 10 & |\Delta P| > 0.0005\ \mathrm{mW \cdot m^{-2}} \\ 0 & |\Delta P| \le 0.0005\ \mathrm{mW \cdot m^{-2}} \end{cases} \tag{17}$$

Finally, the parameters used to update Equation (16) need to be selected. Here α is set to 0.001, and γ is set to 1. Figure 2 shows the flowchart of the proposed algorithm.

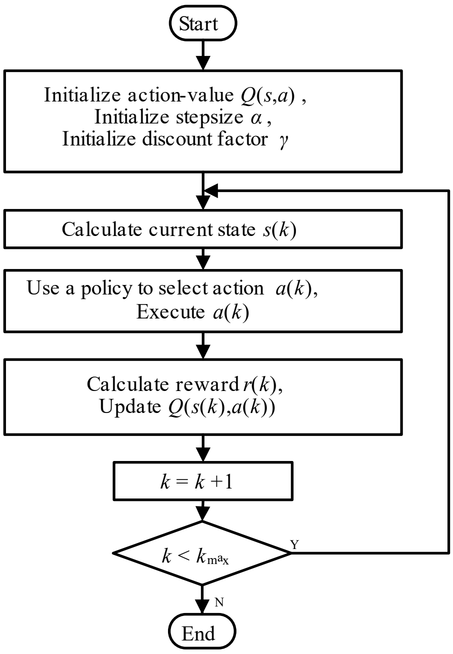
Start
Initialize action-value $Q(s,a)$,
Initialize stepsize α ,
Initialize discount factor γ
Calculate current state $s(k)$
Use a policy to select action $a(k)$,
Execute $a(k)$
Calculate reward $r(k)$,
Update $Q(s(k),a(k))$
$k = k + 1$
$k < k_{max}$
Y
N
End

Figure 2. Flowchart of the proposed algorithm

3.3 Structure of Control System

Based on the proposed Q-Learning algorithm, a maximum power tracking control system of MFC was constructed, as shown in Figure 3.

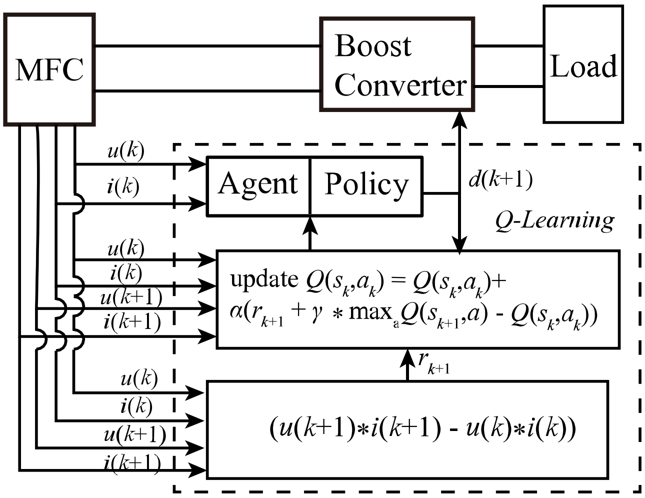
MFC
Boost Converter
Load
u(k)
i(k)
Agent
Policy
d(k+1)
Q-Learning
u(k)
i(k)
u(k+1)
i(k+1)
update $Q(s_k,a_k) = Q(s_k,a_k)+$
$\alpha(r_{k+1} + \gamma * \max_a Q(s_{k+1},a) - Q(s_k,a_k))$
r_{k+1}
u(k)
i(k)
u(k+1)
i(k+1)
$(u(k+1)*i(k+1) - u(k)*i(k))$

Figure 3. Scheme of the QLMPPT Control system of MFC

3.4 Convergence of the Learning Algorithm

The convergence of the Q-Learning algorithm has been proved by Watkins [36]. Here the MFC system was analyzed to illustrate the effectiveness of this algorithm. The MFC is a deterministic system, which means that when the controller executed action $a(k)$ in the state $s(k)$, it would move to the state $s(k+1)$ and receive a reward $r(k)$ from the environment definitely. This whole process would not be random. Therefore, for an MFC with a single maximum power point, when the controller succeeded to explore the maximum power at the first time, assuming that this state was *s*, no matter the controller executed action a_1(a_1=0.1) or action a_2 (a_2= -0.1), a negative reward would be given. Only if the controller executed action a_3 (a_3=0), a zero reward would be given. When the controller explored all the actions in state *s* and then returned to state *s* again, it would only executed action (0) and kept stable at this state *s*, which was also the maximum power.

4. SIMULATION AND VERIFICATION

To evaluate the effects of the proposed QLMPPT control method, simulations of MFC system with several different operating conditions were performed in MATLAB. ε-greedy, greedy policy and Boltzmann methods were respectively used to select action in the Q-Learning.

4.1 MPPT Control Based on P/O and Improved P/O

The proposed Q-Learning based controllers were compared with the conventional MPPT method based on P/O. In conventional P/O, each duty cycle change was fixed and equal to a predetermined offset

Δ*d*. The direction of duty cycle change depends on the sign of power increment. Due to the influence of process dynamics, it is important to wait for a steady state after each duty cycle change. The flowchart of the P/O algorithm was presented in Figure 4.

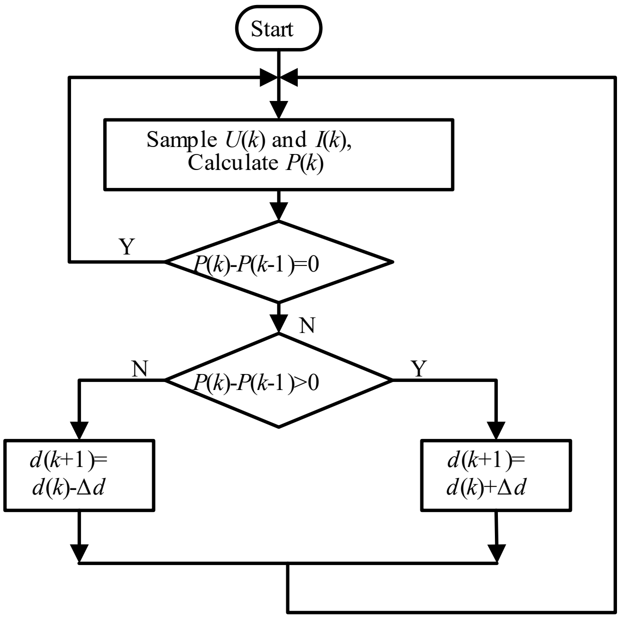

Figure 4. Flowchart of the P/O algorithm

In the simulation experiments, the MFC system was started with a fixed duty cycle of 0.1. After running for 20 hours, MFC completed the startup process and the MPPT algorithm began to work. A load with a resistance of 500 Ω was linked to the Boost converter, and the load changes to 1000 Ω at 800 h. The sampling time of current and voltage was set to 20 hours, and the calculation period of the algorithm was also set to 20 hours.

Figure 5 shows the power density curve of MFC with conventional P/O algorithm. MFC was started up with starting duty cycle, which was set at 0.01. After 50 hours, the P/O algorithm began to work. Due to the large dynamic progress of MFC, the MPP was reached after 100 hours, the steady power density was about 2.1 mW·m^{-2}. Large oscillation can be seen in the curve. This is because a fixed duty cycle and large duty cycle change were used in the P/O algorithm. After sudden change of load resistance at 800 h, the P/O algorithm cannot track the MPP again thus the produced power density oscillated at 1.6 mW·m^{-2}.

In view of the above problems, an improved P/O algorithm was proposed, the flowchart of the improved P/O algorithm was shown in Figure 6 and the power density curve of MFC with the improved P/O algorithm was shown in Figure 7. The setting of the simulation was same as the P/O algorithm until the oscillation occurred. After the oscillation was detected, the step size of the duty cycle was decreased following with a logarithmic function. After 450 hours' back-and-forth settlement, MPP was reached without any oscillation, which means that the improved P/O algorithm can greatly improve the stability of power tracking. However, when the load changed, the improved P/O algorithm made the output power density of MFC stable at another point rather than its actual MPP, which means that the P/O algorithm failed to track the actual maximum power point when the load changed.

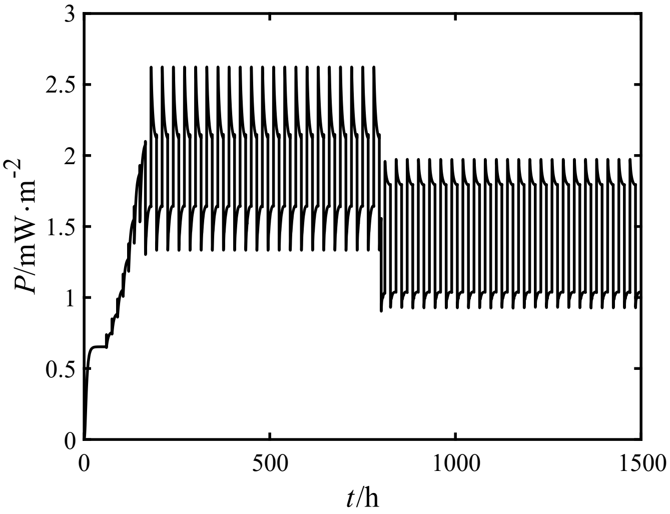

Figure 5. Power density curve of MFC with general P/O algorithm

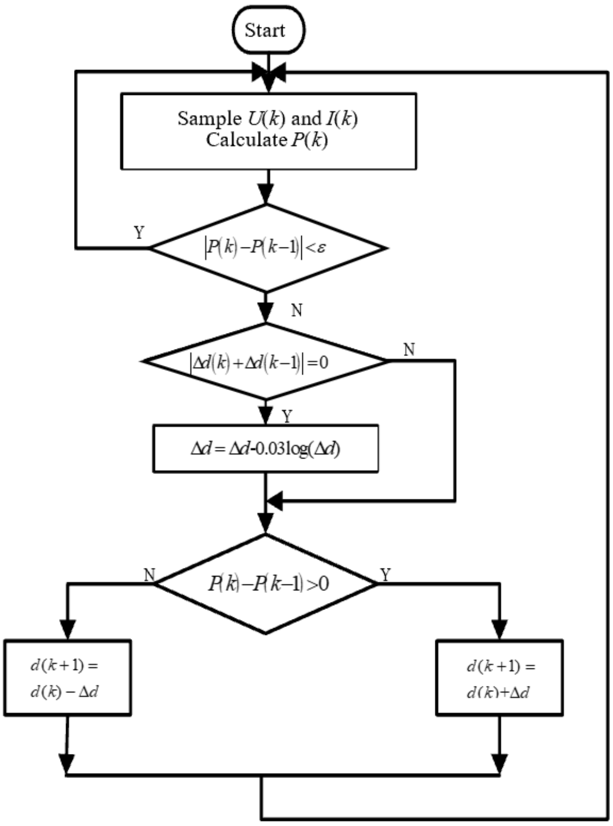

Figure 6. Flowchart of the improved P/O algorithm

4.2 MPPT Control Based on Q-Learning

MPPT based on Q-Learning was then implemented in the MFC system. Besides the greedy policy, two other types of action selection policies based on Boltzmann exploration and ε-greedy exploration were still used in MPPT so as to find the more appropriate approach.

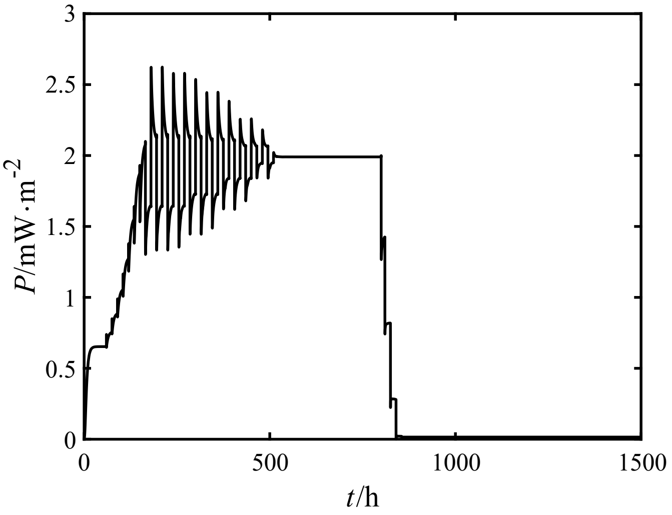

Figure 7. Power density curve of MFC with improved P/O algorithm

4.2.1. MPPT Based on Q-Learning with ε-greedy Exploration

ε-greedy exploration is the most widely used exploration method in reinforcement learning. The power density curve of MFC with Q-Learning based on ε-greedy exploration was shown in Figure 8. In this method, a random number between 0 and 1 was generated. If this number was larger than ε which is also a number in (0,1), the action was selected randomly; otherwise, the action was selected using a greedy policy. In this way, it took about 350 h from starting the MFC to tracking to the MPP and stabilizing. When the load resistance changed at 800 h, it took about 400 h to re-track the MPP. However, the output power density could not keep stable at the MPP but fluctuated near the MPP. The main reason for this result is that a fixed ε (ε=0.9) was used in this method, it still has probability to choose a random action even though the MFC has reached the MPP, so the output power density could not keep stable at the MPP.

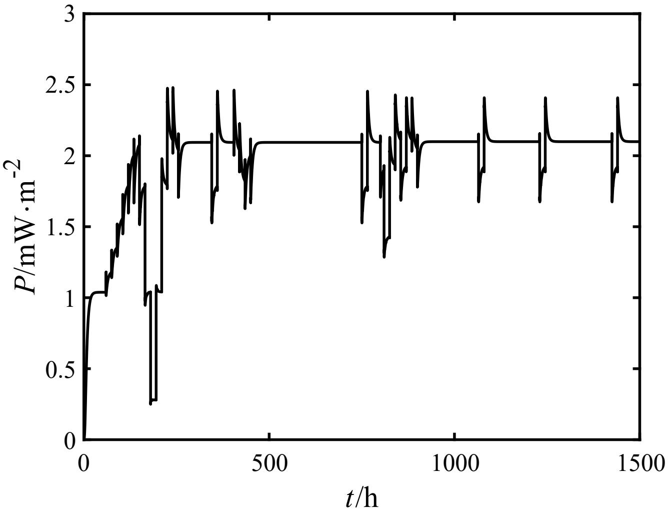

Figure 8. Power density curve of MFC with Q-Learning based on ε-greedy exploration

4.2.2. MPPT Based on Q-Learning with Boltzmann Exploration

The power density curve of MFC with Q-Learning based on Boltzmann exploration was shown in Figure 9. In this algorithm, an action *a* in a state *s* was chosen with a probability *p*(*s*,*a*$_i$) according to the action-value *Q*(*s*, *a*$_i$) as:

$$p(s,a_i)=\frac{e^{Q(s,a_i)/\tau}}{\sum_{a_i}e^{Q(s,a_i)/\tau}} \tag{18}$$

where τ is a positive parameter which controls the randomness of the exploration. A higher τ makes the selection of action more random. Here we set τ to 0.05.

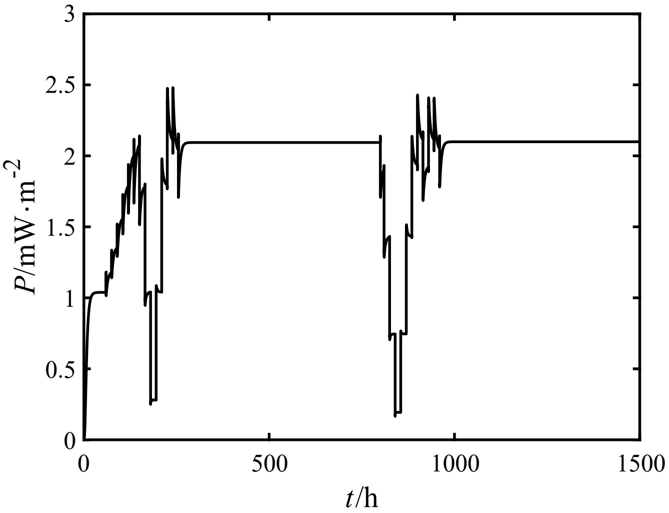

Figure 9. Power density curve of MFC with Q-Learning based on Boltzmann exploration

It can be seen from Figure 9 that the Q-Learning algorithm based on Boltzmann exploration made the MFC stable at the actual MPP after 400 h, which was longer than that of using ε-greedy exploration. When suffer a sudden change of load resistance at 800 h, the method of Q-Learning based on Boltzmann exploration made the MFC re-stabilize at the MPP in about 180 h, which was faster and more stable than that of using the ε-greedy exploration. For this method, to make the tracking faster, a relatively smaller fixed parameter τ should be chosen, while less exploration may make a steady error.

4.2.3. MPPT Based on Q-Learning with greedy policy Exploration

The power density curve of MFC using greedy policy to choose action was shown in Figure 10. After the MFC was started, the algorithm adjusted the duty cycle in one direction as we expected. After exploring all the states with action a_1 (a_1=0.1), the sign of the change in duty cycle inversed and continued to update the state value table with action a_2 (a_2=-0.1). When the MPP was reached, both action a_1 and action a_2 made the reward negative, then action a_3 (a_3=0) was selected and the duty cycle remained constant so that the output power remained at the MPP.

The whole process from startup to stabilization at MPP took about 200 h, which was 42.9% faster than that of MFC with Q-Learning based on ε-greedy exploration, and 50% faster than that of MFC with

Q-Learning based on Boltzmann exploration. When the load resistance changed at 800 h, the power density fluctuated greatly, which was because a sudden change of load resistance caused the algorithm to re-explore; but after a period of adjustment, the output power density could still return to the actual MPP and stabilized. It took about 160 h from the time the load disturbance occurred to the time the power density of MFC was stabilized at the MPP again, which was 11.1% faster than that of MFC with Q-Learning based on Boltzmann exploration. Compared with the method of Q-Learning based on ε-greedy exploration and Q-Learning based on Boltzmann exploration, the method of Q-Learning based on greedy policy has faster response time and better stability.

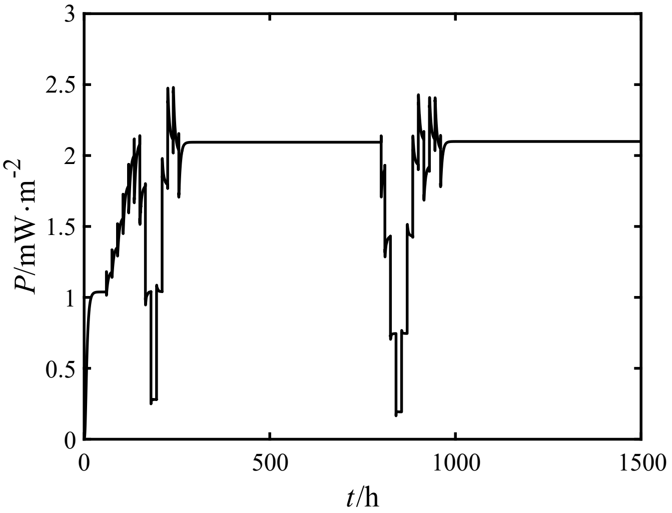

Figure 10. Power density curve of MFC with Q-Learning based on greedy policy

5. CONCLUSION

The general P/O algorithm cannot track the MPP of MFC accurately but make the output power of MFC oscillate greatly near the MPP, and MPPT with general P/O algorithm cannot be retraced to the MPP in case of load change. The improved P/O algorithm can make MFC reach the MPP without oscillation after a period of adjustment, but it still cannot return to the actual MPP when encountering load disturbance. While the reinforcement learning algorithm can solve the problems existing in the P/O algorithm. The method of Q-Learning based on greedy policy or Boltzmann exploration can make the MFC stable at MPP with a short time even though a sudden change of load occurs. Considering the action selection policies of Q-Learning algorithms, greedy policy seems to be more suitable. The Q-Learning algorithm based on the greedy policy can make the output power of MFC track to its MPP and keep stable at a faster speed and can also return to the actual MPP and stabilize at the fastest speed when encountering load disturbance. On the other hand, greedy algorithm does not have an adjustment parameter but the other two methods have, which also makes it more convenient and feasible. Therefore, Q-Learning algorithm based on greedy policy is an effective method to realize MPPT control in MFC system.

ACKNOWLEDGEMENTS

This work was supported by the Chinese-Macedonian Scientific and Technological Cooperation Project of Ministry of Science and Technology of the People's Republic of China (under Grant [2017] 25: 5-5), and the support plan for distinguished professors in Liaoning Province, China (under Grant [2014] 187).

References

1. M. Zhou, H. Wang, D.J. Hassett, T. Gu, *J. Chem. Technol. Biotechnol.*, 88 (2013) 508.
2. G. Palanisamy, H.Y. Jung, T. Sadhasivam, M.D. Kurkuri, S.C. Kim, S.H. Roh, *J. Clean. Prod.*, 221 (2019) 598.
3. A.J. Slate, K.A. Whitehead, D.A.C. Brownson, C.E. Banks, *Renew. Sustain. Energy Rev.*, 101 (2019) 60.
4. L. Fan, J. Zhang, X. Shi, *Int. J. Electrochem. Sci.*, 10 (2015) 737.
5. C. Santoro, C. Arbizzani, B. Erable, I. Ieropoulos, *J. Power Sources*, 356 (2017) 225.
6. H. Wang, J.D. Park, Z.J. Ren, *Environ. Sci. Technol.*, 49 (2015) 3267.
7. Z.H. Tong, H.Q. Yu, W.W. Li, Y.K. Wang, M. Sun, X.W. Liu, G.P. Sheng, *Ecotoxicology*, 24 (2015) 2175.
8. L. Fan, D. Xu, C. Li, S. Xue, *Polish J. Environ. Stud.*, 25 (2016) 2359.
9. L. Fan, J. Shi, T. Gao, *Energies*, 13 (2020) 1383.
10. L. Fan, Y. Zheng, X. Miao, *J. Chem. Eng. Chinese Univ.*, 30 (2016) 491.
11. H.C. Boghani, G. Papaharalabos, I. Michie, K.R. Fradler, R.M. Dinsdale, A.J. Guwy, I. Ieropoulos, J. Greenman, G.C. Premier, *J. Power Sources*, 269 (2014) 363.
12. M. Bahrami, R. Gavagsaz-Ghoachani, M. Zandi, M. Phattanasak, G. Maranzanaa, B. Nahid-Mobarakeh, S. Pierfederici, F. Meibody-Tabar, *Renew. Energy*, 130 (2019) 982.
13. M. Alaraj, M. Radenkovic, J.D. Park, *J. Power Sources*, 342 (2017) 726.
14. A. Adekunle, V. Raghavan, B. Tartakovsky, *Batteries*, 5 (2019) 9.
15. J.D. Park, Z. Ren, *J. Power Sources*, 205 (2012) 151.
16. D. Molognoni, S. Puig, M.D. Balaguer, A. Liberale, A.G. Capodaglio, A. Callegari, J. Colprim, *J. Power Sources*, 269 (2014) 403.
17. J. Coronado, M. Perrier, B. Tartakovsky, *Bioresour. Technol.*, 147 (2013) 65.
18. L. Woodward, M. Perrier, B. Srinivasan, R.P. Pinto, B. Tartakovsky, *AIChE J.*, 56 (2010) 2742.
19. S. Hadji, J.P. Gaubert, F. Krim, *Energies*, 11 (2018) 459.
20. C. Wei, Z. Zhang, W. Qiao, L. Qu, *IEEE Trans. Ind. Electron.*, 62 (2015) 6360.
21. A. Saenz-Aguirre, E. Zulueta, U. Fernandez-Gamiz, J. Lozano, J. Lopez-Guede, *Energies*, 12 (2019) 436.
22. R.C. Hsu, C.T. Liu, W.Y. Chen, H.I. Hsieh, H.L. Wang, *Int. J. Photoenergy*, 2015 (2015) 1.
23. P. Kofinas, S. Doltsinis, A.I. Dounis, G.A. Vouros, *Renew. Energy*, 108 (2017) 461.
24. T. Liu, Y. Zou, D. Liu, F. Sun, *Energies*, 8 (2015) 7243.
25. R. Xiong, J. Cao, Q. Yu, *Appl. Energy*, 211 (2018) 538.
26. J. Wu, H. He, J. Peng, Y. Li, Z. Li, *Appl. Energy*, 222 (2018) 799.
27. S. Kazemi, M. Barazandegan, M. Mohseni, K. Fatih, *Energies*, 9 (2016) 79.
28. R.P. Pinto, B. Srinivasan, M.F. Manuel, B. Tartakovsky, *Bioresour. Technol.*, 101 (2010) 5256.
29. C. Picioreanu, I.M. Head, K.P. Katuri, M.C.M. van Loosdrecht, K. Scott, *Water Res.*, 41 (2007) 2921.
30. A. Kato Marcus, C.I. Torres, B.E. Rittmann, *Biotechnol. Bioeng*., 98 (2007) 1171–1182.
31. M. Esfandyari, M.A. Fanaei, R. Gheshlaghi, M. Akhavan Mahdavi, *Chem. Eng. Res. Des.*, 117 (2017) 34.
32. Y. Zeng, Y.F. Choo, B.H. Kim, P. Wu, *J. Power Sources*, 195 (2010) 79.
33. L.P. Fan, J.J. Li, *Int. J. Circuits, Syst. Signal Process*, 10 (2016) 316.
34. Z. Zhong, H. Huo, X. Zhu, G. Cao, Y. Ren, *J. Power Sources*, 176 (2008) 259.

35. D. Ernst, M. Glavic, F. Capitanescu, L. Wehenkel, *IEEE Trans. Syst. Man, Cybern. Part B*, 39 (2009) 517.
36. C.J.C.H. Watkins, P. Dayan, *Mach. Learn.*, 8 (1992) 279.