Journal of Southeast University (English Edition) Vol. 36, No. 4, pp. 393 - 398 Dec. 2020 ISSN 1003—7985

Q-learning-based energy transmission scheduling over a fading channel

Wang Zhiwei[1] Wang Junbo[2] Yang Fan[2] Lin Min[3]

([1] School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China)
([2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([3] School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: To solve the problem of energy transmission in the Internet of Things (IoTs), an energy transmission schedule over a Rayleigh fading channel in the energy harvesting system (EHS) with a dedicated energy source (ES) is considered. According to the channel state information (CSI) and the battery state, the charging duration of the battery is determined to jointly minimize the energy consumption of ES, the battery's deficit charges and overcharges during energy transmission. Then, the joint optimization problem is formulated using the weighted sum method. Using the ideas from the Q-learning algorithm, a Q-learning-based energy scheduling algorithm is proposed to solve this problem. Then, the Q-learning-based energy scheduling algorithm is compared with a constant strategy and an on-demand dynamic strategy in energy consumption, the battery's deficit charges and the battery's overcharges. The simulation results show that the proposed Q-learning-based energy scheduling algorithm can effectively improve the system stability in terms of the battery's deficit charges and overcharges.

Key words: energy harvesting; channel state information; Q-learning; transmission scheduling

DOI: 10.3969/j.issn.1003 - 7985.2020.04.004

With the rapid development of the IoTs, energy harvesting has been regarded as a favorable supplement to drive the numerous sensors in the emerging IoT[1]. Due to several key advantages such as being pollution free, having a long lifetime, and energy self-sustainability, the energy harvesting systems (EHSs) are competitive in a wide spectrum of applications[2].

The EHS generally consists of an antenna either separating or shared with data communications, an energy harvesting device (EHD) converting the RF signal from energy sources (ESs) to power, and a battery that stores the harvested energy[3]. According to different ESs, the RF-based energy harvesting system can be classified into two categories: EHS with ambient ESs and EHS with a dedicated ES[3].

Recent research of the EHS mainly focuses on how to effectively utilize energy from ambient or dedicated ESs[4-6]. In Ref. [4], an energy neutrality theorem for EHN was proposed and it was proved that perpetual operation can be achieved by maintaining the energy neutrality of EHN. Then, an adaptive duty cycle (ADC) control method was further proposed in order to assign the duty cycle online to achieve the perpetual operation of EHN. In Ref. [5], a reinforcement learning-based energy management scheme was proposed to achieve the sustainable operation of EHN. In Ref. [6], a fuzzy Q-leaning-based power management scheme was proposed for EHN under energy neutrality criteria. To achieve the sustainable operation of EHN, the duty cycle is decided from the fuzzy inference system for the EHN. In fact, all the research managed to adjust power in the EHS with ambient ESs to maximize the utilization of the harvested energy. However, due to the lack of the contact between the ESs and EHDs, the energy transmission period in the EHS with ambient ESs are more uncontrollable and unstable. However, in the EHS with a dedicated ES, the progress of energy transmission can be scheduled effectively due to the dedicated ES which is installed to power the EHDs. Hence, some research began to focus on the EHS with a dedicated ES. In Ref. [3], a two-step dual tunnel energy requesting (DTER) strategy was proposed to minimize the energy consumption at both the EHD and the ES on timely data transmission. However, these existing strategies did not consider the exhaustion or overflow of the battery's energy during the transmission. Hence, this paper will concentrate on the online energy management strategies to improve system stability in terms of the battery's deficit charges and overcharges.

In this paper, a Q-learning-based energy transmission scheduling algorithm is proposed to improve the EHS with a dedicated ES. Based on the basic theories of the Q-learning algorithm[7], an energy transmission scheduling algorithm is used to decrease energy consumption through adjusting transmitted energy. By using the energy scheduling scheme in this paper, the EHS can adjust the transmitted energy of ES timely and effectively to change

Received 2020-06-02, **Revised** 2020-11-03.

Biographies: Wang Zhiwei (1990—), male, doctor; Wang Junbo (corresponding author), male, doctor, professor, jbwang@seu.edu.cn.

Foundation item: The National Natural Science Foundation of China (No. 51608115).

Citation: Wang Zhiwei, Wang Junbo, Yang Fan, et al. Q-learning-based energy transmission scheduling over a fading channel[J]. Journal of Southeast University (English Edition), 2020, 36(4): 393 - 398. DOI: 10.3969/j.issn.1003 - 7985.2020.04.004.

the energy consumption. First, the system model of the EHS is presented in detail. Then, a multi-objective optimization problem is formulated to improve system performance in terms of the battery's deficit charges and overcharges. Next, a Q-learning-based scheduling algorithm is proposed for the optimization problem. Finally, the simulation results and conclusions are presented, respectively.

1 System Model

Consider an RF-based EHS, where the EHD requests and harvests energy from the ES, as shown in Fig. 1. The harvested energy stored in the EHD's battery is consumed to send out data. Moreover, the system time is assumed to be equally divided into N time slots and $T_n(1\leqslant n\leqslant N)$, the duration of time slot n is constant and selected to be less than the channel coherence time. Therefore, the channel states remain invariant over each time slot but vary across successive time slots. Assume that the fading of the wireless channel follows a correlated Rayleigh fading channel model[8]. Using the ellipsoidal approximation, the CSI can be deterministically modeled as[9]

$$g_n = h_n 10^{-v_n/10} \tag{1}$$

where v_n is the uncertain parameter and θ denotes the uncertainty bound which is a non-negative constant; g_n and h_n denote the actual and estimated channel gains at time slot n, respectively.

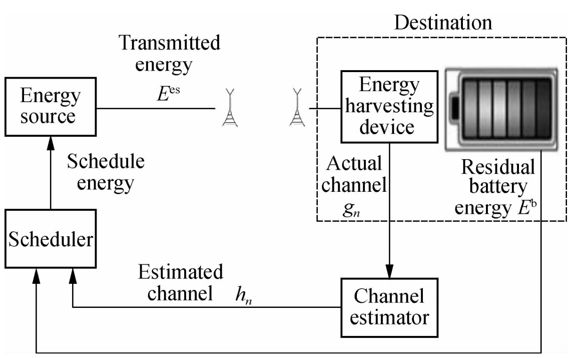

Fig. 1 The energy-harvesting system

To charge the EHD, we assume that the ES transmits energy at a constant power p^{es}. First, before the EHD receives energy at time slot n, the voltage and energy stored in the battery are denoted as v_n and E_n^{b}, respectively. The transmission duration of energy at time slot n is T_n^{es}. Then, after energy harvesting, the voltage of the battery will increase by ΔV_n. Mathematically, the charging function can be deduced as[10]

$$V_n = V_{\mathrm{m}}(1-\mathrm{e}^{-t'/(RC)}) \tag{2}$$

$$V_n + \Delta V_n = V_{\mathrm{m}}(1-\mathrm{e}^{-(t'+T_n^{\mathrm{es}})/(RC)}) \tag{3}$$

$$E_n^{\mathrm{b}} = \frac{1}{2}CV_n^2 \tag{4}$$

$$E_n^{\mathrm{b}} + E_n^{\mathrm{r}} = \frac{1}{2}C(V_n+\Delta V)^2 \tag{5}$$

where t' is the time consumed during charging the voltage of the battery from 0 to V_n with V_{m} volts of voltage; V_{m} is the maximum voltage that the battery can approach. R and C are the resistance and capacitance of the charging circuit in EHD, respectively. Eq. (2) represents that the battery needs to spend time t' on voltage changing from 0 to V_n and Eq. (3) represents that the voltage changes from V_n to $V_n+\Delta V_n$ after energy harvest at time slot n. Eq. (4) and Eq. (5) reflect the relationship between the battery's voltage and stored energy. Using Eq. (2) to Eq. (5), the charge duration can be derived as

$$T_n^{\mathrm{es}} = RC\ln\left[\frac{(2E_n^{\mathrm{eh}})^{1/2}-(2E_n^{\mathrm{b}})^{1/2}}{(2E_{\mathrm{m}}^{\mathrm{eh}})^{1/2}-(2E_n^{\mathrm{b}}+2E_n^{\mathrm{re}})^{1/2}}\right] \tag{6}$$

where $E_{\mathrm{m}}^{\mathrm{eh}}$ represents the maximal energy that the battery can store. Eq. (6) shows that the energy consumption is affected by the amount of the expected received energy E_n^{re} and the residual energy in the battery E_n^{b}. Considering that the bad channel states can significantly reduce the efficiency of the battery charge, it is assumed that if the channel state at time slot n is bad, the ES will not send energy to the destination node. Hence, Eq. (6) can be further improved as

$$T_n^{\mathrm{es}} = \begin{cases} RC\ln\left[\dfrac{(2E_{\mathrm{m}}^{\mathrm{eh}})^{1/2}-(2E_n^{\mathrm{b}})^{1/2}}{(2E_n^{\mathrm{eh}})^{1/2}-(2E_n^{\mathrm{b}}+2E_n^{\mathrm{re}})^{1/2}}\right] & g_n p^{\mathrm{es}} \geqslant p^{\mathrm{th}} \\ 0 & g_n p^{\mathrm{es}} < p^{\mathrm{th}} \end{cases} \tag{7}$$

where p^{th} denotes the charge power of a battery.

At time slot n, it is assumed that the EHD sends data at transmitted power p_n^{eh}. Then, the residual energy at time slot $n+1$ can be determined as

$$E_{n+1}^{\mathrm{b}} = E_n^{\mathrm{r}} + E_n^{\mathrm{b}} - p_n^{\mathrm{eh}}T \tag{8}$$

where T is the value of the working period of EHD T_n; E_n^{r} is the real received energy of the battery, and its relationship with energy transmitted by ES is

$$E_n^{\mathrm{r}} = \eta p^{\mathrm{es}} T_n^{\mathrm{es}} |g_n|^2 \tag{9}$$

where η represents the conversion efficiency of a battery.

2 Problem Formulation

To efficiently make use of scarce transmission resources, multiple objectives should be considered simultaneously. The primary objective is to save energy consumption. In a practical system, most of the energy is consumed for wireless transmission. Therefore, the primary objective becomes how to save transmission energy. According to Eq. (6), the consumed energy is mainly affected by E_n^{r} and E_n^{b}. Hence, adjusting it properly can significantly reduce the consumed energy. Obviously, the primary objective can be described mathematically by minimizing Eq. (7). After each charge, the EHD can stop working due to the low residual energy of the battery. To prevent this situation, the residual energy of the battery at each time slot should be no less than the minimum energy that ensures normal working. Therefore, the condition of the battery's energy exhaustion at time slot n can be described as

$$E_n^{\mathrm{b}} < \upsilon E_{\mathrm{m}}^{\mathrm{eh}} \tag{10}$$

where υ represents the minimum capacity percentage of the battery that can keep EHD normally. Meanwhile, due to the limitation of the storage size, the overflow of the battery's energy will occur when the received energy is too large. Therefore, how to avoid overcharges of the battery should be taken into account as well. The condition of the battery's overcharge at time slot n can be described as

$$E_n^{\mathrm{r}} + E_n^{\mathrm{b}} - p_n^{\mathrm{eh}} T > E_{\mathrm{m}}^{\mathrm{eh}} \tag{11}$$

In most cases, it is unlikely that the three objectives can simultaneously be optimized by the same solution. Therefore, some tradeoff between the above three objectives is needed to ensure satisfactory system performance. The most well-known tradeoff method is the weighted sum method[11]. Accordingly, the multi-objective optimization problem can be converted into the following minimization problem,

$$\min_{T_n^{\mathrm{es}} \leqslant T} E\Big[\sum_{n=0}^{N-1} p^{\mathrm{es}} T_n^{\mathrm{es}} + \tau I(E_n^{\mathrm{r}} + E_n^{\mathrm{b}} - p_n^{\mathrm{eh}} T_n > E_m^{\mathrm{eh}}) + \mu I(E_n^{\mathrm{b}} < \upsilon E_{\mathrm{m}}^{\mathrm{eh}})\Big] \tag{12}$$

where $E(\cdot)$ is the expectation operator; $I(\cdot)$ is an indicator function and is used to show the occurrence of overcharges or deficit charges; τ and μ are two small positive constants, which are used to adjust the weight of deficit charges and overcharges of the battery during the optimization.

3 Online Scheduling Algorithm

In this section, optimization problem Eq. (12) is transformed into a reinforcement learning problem. The Q-learning algorithm first creates a Q-table which records the Q-value of all the combinations of states and actions. Then, through training, the Q-value converges and the EHS can choose the best T_n^{es} at every state according to the maximum or minimum value in the Q-table. The elements in Eq. (12) can be mapped into Q-learning elements as follows.

3.1 State

Channel state and residual battery energy are continuous variables, which should be converted into discrete and finite. Therefore, we divided the ranges of the continuous variable into several intervals. If different variables are located in the same interval, they are regarded the same. To distinguish these intervals, we use continuous natural numbers to label them and these numbers can be regarded as different states.

In the proposed scheduling scheme, the channel states are assumed to be discrete and finite. Without loss of generality, the range of the estimated channel gain can be divided into D states. The states can be defined as

$$\varphi_d = \begin{cases} [0, \omega_1) & d = 1 \\ [\omega_{d-1}, \omega_d) & d = 2, 3, \dots, D-1 \\ [\omega_{d-1}, +\infty) & d = D \end{cases} \tag{13}$$

where $0 < \omega_1 < \omega_2 < \dots < \omega_{D-1}$. Therefore, at time slot n, the channel state can be determined as

$$H_n = \arg_{d \in \{1, 2, \dots, D-1\}} \{h_n \in \varphi_d\} \tag{14}$$

Similarly, the residual battery energy, which is also assumed to be discrete and finite, can be divided into E states as follows:

$$\varphi_e = \begin{cases} [0, \sigma_1) & e = 1 \\ [\sigma_{e-1}, \sigma_e) & e = 2, 3, \dots, E-1 \\ [\sigma_e, E_{\mathrm{m}}^{\mathrm{eh}}) & e = E \end{cases} \tag{15}$$

where $0 < \sigma_1 < \sigma_2 < \dots < \sigma_{e-1} < E_{\mathrm{m}}^{\mathrm{eh}}$. At time slot n, the residual energy state can be determined as

$$E_n = \arg_{e \in \{1, 2, 3, \dots, E\}} \{E_n^{\mathrm{b}} \in \varphi_e\} \tag{16}$$

Using the residual energy and channel states, the current composite state of the system is defined in a vector as

$$\boldsymbol{S}_n = \{H_n, E_n\} \triangleq \{1, 2, 3, \dots, D\} \times \{1, 2, 3, \dots, E\} \tag{17}$$

Eq. (17) represents that every state can be mapped into the only combination of H_n and E_n.

3.2 Action

Obviously, the charging duration T_n^{es} can be viewed as an action. Assume that the actions are discrete and divided into N levels. Therefore, the set of available actions can be given as

$$A_n = \left\{\frac{j}{N-1} T, j = 0, 1, 2, \dots, N-1\right\} \tag{18}$$

3.3 Cost

In the optimization problem Eq. (12), the objective is to save energy consumption, avoid overflow of a battery's energy and prevent a battery from draining. Therefore, the total cost is determined as

$$\boldsymbol{c}(\boldsymbol{S}_n, T_n^{\mathrm{es}}) = p^{\mathrm{es}} T_n^{\mathrm{es}} + \tau I(E_n^{\mathrm{r}} + E_n^{\mathrm{b}} - p_n^{\mathrm{eh}} T > E_{\mathrm{m}}^{\mathrm{eh}}) + \mu I(E_{n+1}^{\mathrm{b}} < \nu E_{\mathrm{m}}^{\mathrm{eh}}) \tag{19}$$

As different circumstances have different QoS requirements, by adjusting μ and τ, the reward function is generic enough to satisfy different requirements in real systems.

3.4 Action selection

Using the states, actions and cost functions defined above, the received energy at time slot n can be selected by

$$T_n^{\mathrm{es}} = \arg \min_{T_n^{\mathrm{es}} \in A_n} \boldsymbol{Q}(\boldsymbol{S}_n, T_n^{\mathrm{es}}) \tag{20}$$

where $\boldsymbol{Q}(\boldsymbol{S}_n, T_n^{\mathrm{es}})$ is the action-state value associated with tuple $[(\boldsymbol{S}_n, T_n^{\mathrm{es}})]$. In matrix $\boldsymbol{Q}$, the value of an arbitrary element is equal to the summation of its cost and the minimum discounted value of $\boldsymbol{Q}$ over all possible actions in the next state.

After selecting the proper action, the next state of battery energy E_{n+1} can be determined by Eq. (8) and Eq. (16). Also, the next channel state H_{n+1} can be obtained by Eq. (14). Hence, combined with the information of E_{n+1} and H_{n+1}, the next state $\boldsymbol{S}_{n+1}$ is determined as well. Accordingly, matrix $\boldsymbol{Q}$ will be updated as

$$\boldsymbol{Q}(\boldsymbol{S}_n, T_n^{\mathrm{es}}) = \boldsymbol{Q}(\boldsymbol{S}_n, T_n^{\mathrm{es}}) + \alpha[\boldsymbol{c}(\boldsymbol{S}_n, T_n^{\mathrm{es}}) + \gamma \min_{T_{n+1}^{\mathrm{es}}} \boldsymbol{Q}(\boldsymbol{S}_{n+1}, T_{n+1}^{\mathrm{es}}) - \boldsymbol{Q}(\boldsymbol{S}_n, T_n^{\mathrm{es}})] \tag{21}$$

where α is the time-varying learning rate parameter; γ is the discount factor. The detailed procedures of the algorithm are shown in Algorithm 1.

Algorithm 1 The Q-learning-based scheduling algorithm

Step 1 Initialization.

Step 2 If rand() $< \varepsilon$, randomly select an action from A_n. Else, select an action using Eq. (19).

Step 3 Calculate the cost using Eq. (18) and then determine next state $\boldsymbol{S}_{n+1}$.

Step 4 Update $\boldsymbol{Q}$ by Eq. (20).

Step 5 $n = n + 1$, then go to step 2.

4 Simulation and Results

Under the same simulation environments, the proposed algorithm is compared with the constant strategy algorithm and the on-demand dynamic strategy algorithm[3] in terms of the battery's deficit charges, the battery's overcharges and the total consumed energy. The proposed algorithm and the reference algorithms are, respectively, deployed at most 100 times in one trial, and the trial is repeated 1 000 times. In other words, the ES transmits energy to EHD in each trial, which will not stop unless the battery's energy is exhausted or transmission is carried out more than 100 times. After trials are completed, the data from simulations will be collected to analyze the performance of the algorithms.

4.1 Simulation settings

The constant energy transmitting power of energy source node $p^{\mathrm{es}} = 10$ W. The capacitance C and resistance R of EHD are 1 kΩ and 2 nF, respectively. The maximum capacity of the battery in energy harvesting device $E_{\mathrm{m}}^{\mathrm{eh}}$ is 2 nJ. To maintain the normal function of EHD, the minimum capacity percentage of a battery υ is 5%. The constant parameters in Eq. (12) are 0.4 and 0.2. Considering that at every time slot, the ES cannot obtain information about energy consumption during the EHD's data transmission, the power used in sending the data by an energy harvesting device can be assumed to obey a uniform distribution, and the maximum power is set to be $p_{\mathrm{m}}^{\mathrm{eh}} = 1$ mW. Then, the proposed algorithm and the reference algorithms can be simulated to compare their performance.

4.2 Performance comparison

For comparison purpose, the reference algorithms are described as follows.

1) The constant strategy algorithm. In this strategy, the EHD transmits constant energy. When the residual energy of a battery is not greater than half of $E_{\mathrm{m}}^{\mathrm{eh}}$, the EHD will request a replenishment and charge the battery to 95% of the maximum capacity according to Eq. (7) and Eq. (9).

2) The on-demand dynamic strategy algorithm. To avoid the overflow of arriving energy when the residual energy of a battery is too high, the ES adjusts its transmission energy adaptively based on the status of battery storage. The higher the occupancy of the storage, the greater the applied transmission energy. The occupancies of battery storage and the corresponding transmission energy are $E_n^{\mathrm{b}} = \left[\frac{1}{20}, \frac{2}{20}, \ldots, \frac{19}{20}\right] \times E_{\mathrm{m}}^{\mathrm{eh}}$ and $E_n^{\mathrm{es}} = \left[\frac{19}{20}, \frac{18}{20}, \ldots, \frac{1}{20}\right]$. With this strategy, the EHD is scheduled to request adequate energy for the next data transmission.

Fig. 2 shows the performance comparison between the proposed Q-learning algorithm and reference algorithms. In Fig. 2(a), it is noted that the Q-learning algorithm achieves an excellent performance in terms of the battery's deficit charges. As the reference algorithms do not consider the effect of the battery's deficit charges, the battery's energy cannot be prevented from becoming exhausted during trials and the occurrence of the battery's deficit charges increases with the trials' continuation. In Fig. 2(b), the preference algorithms outperform the Q-learning algorithm slightly in the overcharges. The reason is that both the constant strategy and on-demand strategy algorithms have considered the restriction of overcharges so that the overflow of the battery's energy never occurs during trials. In Fig. 2(c), both the reference algorithms consume less energy than the Q-learning algorithm, but this consequence is based on the degradation in its performance of the battery's deficit charges. To sum up, although the Q-learning algorithm seems to consume more energy than the reference algorithms, it actually provides better system stability during the energy transmission period.

For the Q-learning algorithm, the size of action space can be an important factor that influences algorithm performance. To verify how action space size affects algorithm performance, the simulations of the Q-learning algorithm with different action space sizes are executed under the same simulation environment. The results are shown in Fig. 3.

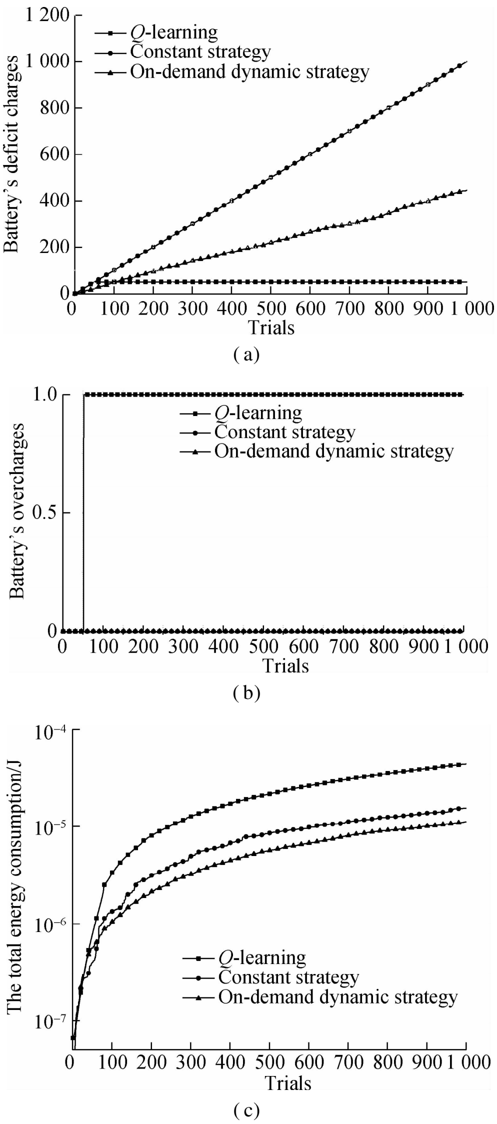

Fig. 2 The performance comparison between the proposed Q-learning algorithms and the reference algorithms. (a) The battery's deficit charges; (b) The battery's overcharges; (c) The total consumed energy

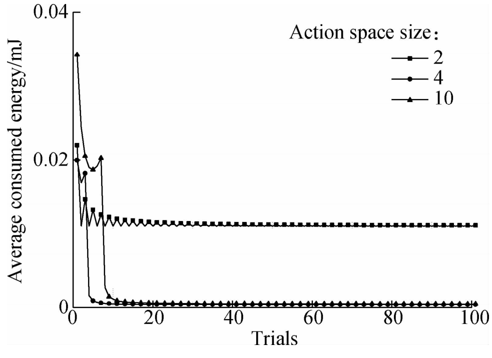

Fig. 3 The averaged energy consumption of the Q-learning algorithms with different sizes of action space

Assume that the size of state space is kept at 10 during simulations. It can be seen that a large action space will result in longer convergence time[12], which is also demonstrated in Fig. 3. Through the accumulated information of multiple iterations, the information of CSI will be obtained. In other words, the Q-learning algorithm spends time in learning before the first 20 trials. In the practical, for the first 20 trials, the system is in the progress of learning, and thus the derived results are not optimal. After the first 20 trials of learning, the system can grasp the best strategy of all the states and the averaged energy consumption of the ES converges to a constant value. In addition, the action space never becomes as large as possible. If the action space is large enough to obtain the optimal averaged energy consumption, a larger action space will only extend the convergence time without reducing energy consumption.

5 Conclusions

1) The proposed Q-learning algorithm can solve the proposed issue and achieves acceptable system performance over different Rayleigh fading channels in terms of energy consumption, a battery's deficit charges and overcharges.

2) Compared with the two reference algorithms, the Q-learning algorithm shows a significant advantage in avoiding a battery's energy from becoming exhausted. From the practical view, it is worthwhile to sacrifice performance in energy consumption in exchange for better system stability.

3) The size of action space can affect the Q-learning algorithm's performance. A small action space causes a shorter convergence time, but cannot converge to the optimal solution. In fact, the Q-learning algorithm with a larger action space can effectively reduce energy consumption during a long time energy transmission.

References

[1] Adila A S, Husam A, Husi G. Towards the self-powered internet of things (IoT) by energy harvesting: Trends and technologies for green IoT[C]// 2018 2nd International Symposium on Small-scale Intelligent Manufacturing Systems (SIMS). Cavan, Ireland, 2018: 1-5.

[2] Kamalinejad P, Mahapatra C, Sheng Z, et al. Wireless energy harvesting for the internet of things[J]. *IEEE Communications Magazine*, 2015, **53**(6): 102-108. DOI: 10.1109/MCOM.2015.7120024.

[3] Luo Y, Pu L N, Zhao Y X, et al. DTER: Optimal two-step dual tunnel energy requesting for RF-based energy harvesting system[J]. *IEEE Internet of Things Journal*, 2018, **5**(4): 2768-2780. DOI: 10.1109/jiot.2018.2813429.

[4] Kansal A, Hsu J, Zahedi S, et al. Power management in energy harvesting sensor networks[J]. *ACM Transactions on Embedded Computing Systems*, 2007, **6**(4): 32. DOI:

10.1145/1274858.1274870.
[5] Hsu R C, Liu C T, Wang H L. A reinforcement learning-based ToD provisioning dynamic power management for sustainable operation of energy harvesting wireless sensor node[J]. *IEEE Transactions on Emerging Topics in Computing*, 2014, **2**(2): 181 − 191. DOI: 10.1109/tetc.2014.2316518.
[6] Hsu R C, Lin T. A fuzzy Q-learning based power management for energy harvest wireless sensor node[C]// 2018 *Int Conf HPCS*. Orléans, France, 2018: 957 − 961.
[7] Mastronarde N, van der Schaar M. Joint physical-layer and system-level power management for delay-sensitive wireless communications[J]. *IEEE Transactions on Mobile Computing*, 2013, **12**(4): 694 − 709. DOI: 10.1109/tmc.2012.36.
[8] Ertel R B, Reed J H. Generation of two equal power correlated Rayleigh fading envelopes[J]. *IEEE Communications Letters*, 1998, **2**(10): 276 − 278. DOI: 10.1109/4234.725222.
[9] Wang J B, Feng M, Song X Y, et al. Imperfect CSI based joint bit loading and power allocation for deadline constrained transmission[J]. *IEEE Communications Letters*, 2013, **17**(5): 826 − 829. DOI: 10.1109/lcomm.2013.031313.122583.
[10] Luo Y, Pu L, Zhao Y, et al. Optimal energy requesting strategy for rf-based energy harvesting wireless communications[C]// *IEEE Conf Comput Commun.* Atlanta, GA, USA, 2017: 1 − 9.
[11] Liu G P, Whidborne J F, Yang J B, et al. Multiobjective optimisation and control[J]. *Wetlands Ecology & Management*, 2003, **17**(2): 157 − 164. DOI: 10.1007/s11273-008-9090-x.
[12] Mitchell T M, Carbonell J G, Michalski R S. *Machine learning*[M]. Boston, MA, USA: Springer, 1986: 39 − 42.

衰落信道下基于 Q 学习的能量调度方案

王志伟[1] 王俊波[2] 杨 凡[2] 林 敏[3]

([1]东南大学网络空间安全学院,南京 210096)
([2]东南大学信息与工程学院,南京 210096)
([3]南京邮电大学理学院,南京 210003)

摘要:研究了处于瑞利衰落信道下,具有一个固定能量源的能量收集系统对能量传输进行调度的问题.根据信道信息和电池的剩余电量状态,确定电池的充电时长使得能量传输过程中能量源的能量消耗、电池的耗尽次数以及电池电量溢出的次数尽可能最小.接着再使用加权和的方式来表示该优化问题.利用 Q 学习的思想,提出了一种基于 Q 学习的能量调度方案来解决此问题.通过将基于 Q 学习的能量传输调度方案与 2 种离线传输策略(静态策略和按需分配的动态策略)在能量消耗、电池电量耗尽次数以及电池电量溢出等方面进行比较,分析该算法的优势与不足.仿真结果表明,基于 Q 学习的能量传输调度方案有效地抑制了电池电量耗尽和电池电量溢出的发生,从而提高了系统的稳定性.

关键词:能量收集;信道状态信息;Q 学习;传输调度
中图分类号:U491