Deep Q Network for Wiretap Channel Model with Energy Harvesting

Zhaohui Li(✉) and Weijia Lei

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
lizhaohui40@foxmail.com

Abstract. An energy harvesting wiretap channel model is considered in which the sender is an energy harvesting node. It is assumed that at each time slot only information about the current state of the sending node is available. In order to find an effective power allocation strategy to maximize secrecy rate, we put forward a deep Q network (DQN) scheme. First, we analyze the constraints of the system and the issue of maximizing the secrecy rate. Next, the power allocation problem is formulated as a Markov Decision Process (MDP) with unknown transition probabilities. In order to solve the continuous state space problem that traditional Q learning algorithms cannot handle, we apply neural networks to approximate the value function. Finally, an online joint resource power allocation algorithm based on DQN is presented. Simulation results show that the proposed algorithm can effectively improve the secrecy rate of the model.

Keywords: Energy harvesting · Deep Q network · Online power allocation · Secrecy rate

1 Introduction

With the rapid development of communication industry, the demand for energy supply in communication networks is increasing. Huge energy consumption inevitably produces large amounts of greenhouse gases. Looking for new green energy and using energy reasonably and efficiently has become one of the key issues in the development of communications industry. Energy harvesting node collects energy from environment for information transmission. theoretically, it can work continuously and permanently [1]. Due to the randomness and intermittency of energy harvesting, energy management and power allocation are problems that such nodes need to solve [2].

The energy management model of energy harvesting communication system can be divided into offline management model and online management model according to whether the node knows the information of energy arrival and channel state in advance. The offline management model assumes that data arrival, harvested energy, and channel state of the energy harvesting process are known at the beginning of the communication. Although this assumption inconsistent with reality, it provides a theoretical performance upper bound of energy harvesting communication system. [3–5] studied offline energy

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 312, pp. 433–444, 2020.
https://doi.org/10.1007/978-3-030-41114-5_32

management strategies in different scenarios. [3] studied the optimal offline power allocation strategy for point-to-point energy harvesting communication systems. The transmitter is equipped with an energy harvesting device. A power allocation scheme for maximizing system throughput is given for the single-hop model of direct transmission and the double-hop model forwarded by the relay node. In [4], the energy arrival is known in advance, and the algorithm that finds the optimal transmission policy with respect to the short-term throughput and the minimum transmission completion time is given. In [5], the non-convex power control optimization problem is transformed into a convex optimization problem, and an effective offline algorithm is given. Unlike the specific information of data arrival, harvested energy, and channel state are known in the offline management model, only the statistical information is known of the online management model. [6, 7] researched online energy management strategies. In [6], the issue of dynamical adaptation of transmission rate with regards to the energy arrival process is discussed. In order to optimize system performance, a low complexity transmission power allocation scheme is proposed. In [7], a cooperative communication system consisting of a source node, destination node and relay node with energy harvesting is considered. The optimal joint link selection and power allocation policies for minimizing the average outage probability were obtained through dynamic programming algorithms.

In practical scenario, the system has no prior knowledge about the environment. Therefore, the offline energy management framework is not applicable. These statistics are actually difficult to obtain, so online energy control strategies based on statistical information is also greatly limited. Reinforcement learning is an adaptive online learning that requires no prior knowledge. The agent learns to maximize the reward in the constant interaction with environment. [8, 9] studied the energy control problem in communication systems using reinforcement learning to solve the problem without environmental prior knowledge. In [8], a point-to-point communication system in which the sending node is energy harvesting node is considered. The optimal power control problem is formulated as a reinforcement learning problem. Then, the effects of the parameters of each algorithm are discussed. In [9], a MIMO wireless communication link in which the nodes are equipped with energy harvesters and rechargeable batteries that are continuously charging from a renewable energy source is studied. And a learning approach in order to find the most efficient transmission policy for data communication that maximizes throughput is proposed. It is assumed that the state space is finite, and the channel coefficients are discrete values in [8, 9], but the actual situation is not the case. [10] proposed the DQN algorithm, which uses neural networks to solve the continuous state space problem that cannot be solved in the traditional reinforcement learning algorithm.

The security threat to wireless communication is more serious than wire communication because of its broadcast characteristics and openness. With the development of computing technology, traditional encryption technology has a risk of failure. The security of information transmission from the physical layer came into being. In 1975, Wyner proposed the wiretap channel model [11], which defined secrecy capacity to evaluate the performance of the system's secure transmission. Wyner's research shows that when legal channel is superior to wiretap channel, theoretically secure communication between legitimate users is possible even without any encryption measures. [12] introduces the basic theory of physical layer security and outlines the latest work and future

challenges of physical layer security technology. In [13], the secure communication of energy harvesting Gaussian wiretap channel based on save-then-transmit protocol is studied. Under the condition of limited energy harvested, an optimization algorithm targeting secrecy rate is proposed.

In this paper, an online power allocation algorithm is studied to maximize secrecy rate. And an energy harvesting wiretap channel model composed of three single antenna nodes is considered. Different from [8] and [9], channel coefficients, battery capacity, and harvested energy in the model are all consecutive values. Therefore, the sending node has infinite-state in our model. To solve this problem, we present an online power allocation algorithm based on DQN.

2 System Model

In this paper, the physical layer security transmission problem under the energy harvesting wiretap channel model consisting of three single antenna nodes is considered. As shown in Fig. 1, the sending node contains an energy harvesting device and a rechargeable battery. The energy harvesting device collects energy from the environment and uses it for sending data to destination node B. During transmission, the energy harvested by the sending node changes randomly, so does the wireless channel. In our scenario, node A has no prior knowledge of energy harvested and channel state. In order to maximize the long-term average secrecy rate, the sending node dynamically adjusts the transmission power according to the instantaneous channel state and energy harvested.

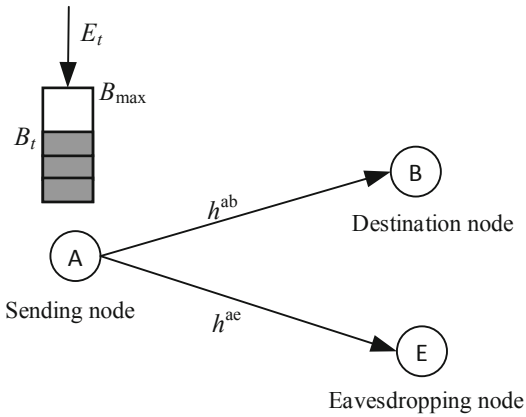

Fig. 1. System model

It is assumed that an amount of energy E_t Joule (J) is harvested at time slot t, and the maximum energy can be harvested is $E_{\max}$ J. The harvested energy is stored in a rechargeable battery with maximum capacity $B_{\max}$ J. There is no energy loss while charging or discharging from the battery. Additionally, the current state of the battery termed B_t J, the transmission power of node A is P_t Watt (W), the channel coefficients of the sending node to the destination node and the sending node to the eavesdropping node

are h_t^{ab} and h_t^{ae} respectively. The transmission power should be less than the maximum discharging power $P_{\max}$ of the battery. Therefore, the constraint in Eq. (1) must be considered.

$$0 \leq P_t \leq P_{\max} \tag{1}$$

At the same time, the transmission power can be allocated only after the harvested energy has been stored in the battery. As a result, causal constraint

$$\tau \cdot P_t \leq B_t \, \forall t = 1, 2, \ldots, T \tag{2}$$

should be satisfied. Where τ is the duration of one time slot. At the beginning of the next time slot, the battery level is

$$B_{t+1} = \min(B_t - \tau \cdot P_t + E_t, B_{\max}) \tag{3}$$

It is assumed that the channel coefficients keep invariant in one time slot. The noise of the legal channel and the wiretap channel is assumed to be independent and identically distributed (i.i.d.) zero mean additive white Gaussian noise with variances σ_{b}^2 and σ_{e}^2 respectively. Consequently, the channel capacity of the legal channel and the wiretap channel are

$$C_t^{\mathrm{b}} = \log_2\left(1 + \frac{P_t \left|h_t^{\mathrm{ab}}\right|^2}{\sigma_{\mathrm{b}}^2}\right) \tag{4}$$

and

$$C_t^{\mathrm{e}} = \log_2\left(1 + \frac{P_t \left|h_t^{\mathrm{ae}}\right|^2}{\sigma_{\mathrm{e}}^2}\right) \tag{5}$$

respectively. According to the relevant theory of physical layer security, when the legal channel is superior to the wiretap channel, the secure transmission of information can be realized even without any secret coding [12]. The system's reachable secrecy rate is defined as the difference between the capacity of legal channel and wiretap channel, as follows.

$$R_{s_t} = \left[C_t^{\mathrm{b}} - C_t^{\mathrm{e}}\right]^+ \tag{6}$$

Where $[a]^+ = \max\{0, a\}$. It can be seen from Eq. (6) that the node A can send message only when $|h_t^{\mathrm{ab}}| > |h_t^{\mathrm{ae}}|$. Therefore, the transmission power P_t is set to 0 when $|h_t^{\mathrm{ae}}| \geq |h_t^{\mathrm{ab}}|$.

To sum up, the goal of power allocation is to maximize the long-term average secrecy rate within the constraints of energy harvesting and battery characteristics, i.e.

$$\begin{aligned} &\max_{\{P_t\}} \lim_{T\to\infty} E\left[\sum_{t=1}^{T} R_{s_t}\right] \\ &\text{s.t.} \quad \text{a) } 0 \leq P_t \leq P_{\max} \\ &\qquad \text{b) } \tau \cdot P_t \leq B_t \\ &\qquad \text{c) } P_t = 0, \left|h_t^{\mathrm{ae}}\right| \geq \left|h_t^{\mathrm{ab}}\right| \end{aligned} \tag{7}$$

3 Reinforcement Learning

3.1 Q Learning

Reinforcement learning [14] is a kind of machine learning that reflects the interaction between agent and environment with states, actions and rewards. The Agent constantly improves strategies in interaction with the environment to maximize the benefits. Reinforcement learning is often modeled as an MDP of quintuples (S, A, T, R, γ). Where S is a set of environmental states. A is a set of actions. T represents the state transition function. R is the reward function. And $\gamma \in [0, 1]$ is the discount factor used to calculate the accumulated reward.

For an MDP, the ultimate goal of reinforcement learning is to find the optimal policy for completing the task. A policy π is a mapping from a given state to the action, i.e. $a_t = \pi(s_t)$. For a given policy π, the accumulated reward is defined as

$$G_t^{\pi} = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{8}$$

In the formula, R_t is the reward obtained at time slot t. Since the sequence of actions in the same state may be different, the accumulated reward for a certain state is not a specific value, but an expected value. We can define a state-action value function to represent the expected value of the accumulated reward for a given strategy, as follows.

$$Q(s_t, a_t) = E_{\pi}\left\{G_t^{\pi} | s = s_t, a = a_t\right\} \tag{9}$$

where E is the mathematical expectation of the strategy. According to the Bellman equation, the accumulated reward is calculated as

$$Q(s_t, a_t) = E_{\pi}\{R_t + \gamma Q(s_{t+1}, a_{t+1}) | s = s_t, a = a_t\} \tag{10}$$

During the learning process, the agent continuously optimizes policy π and finally reaches the optimal policy π^*. The essence of reinforcement learning is to find the best Q function for each state-action pair

$$Q*(s_t, a_t) = E_{\pi}\left\{R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) | s = s_t, a = a_t\right\} \tag{11}$$

Q learning [15] is a reinforcement learning algorithm based on Q function estimation, also known as Temporal Difference (TD) learning algorithm. The rules for updating the Q function are as follows.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right] \tag{12}$$

Where $\alpha \in [0, 1]$ is the learning rate and $R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$ is TD target (estimate of the target value function).

3.2 Deep Q Network

The Q learning algorithm finds the optimal policy by establishing and updating a Q value table. Q learning works well when the state space is small. However, when the state space and the action space are large or continuous, the value function cannot be represented by a table. To solve this problem, an approximation method can be used to estimate $Q(s_t, a_t)$, i.e.

$$Q(s_t, a_t) \approx \tilde{Q}(s_t, a_t; \theta) \tag{13}$$

Where θ is parameter for Approximating the state-action value function. The approximation of value functions can be divided into linear approximation and nonlinear approximation. The use of neural network to approximation value function is a common nonlinear approximation method. The function approximation is a process of supervised learning. The goal of training is

$$\arg\min_{\theta}\Big(Q(s_t, a_t) - \tilde{Q}(s_t, a_t; \theta)\Big)^2 \tag{14}$$

When training neural networks, the training data is required to be independent. But the data obtained at each time slot in reinforcement learning is ordered. Training directly with these data may lead to instability of the neural network. DQN adopts a double neural network with the same structure but different parameters to solve the problem of unstable training caused by the correlation between samples. One for calculating the value function with parameter θ, the other is used to calculate the TD target with parameter θ^{TD}. The parameter θ is updated in every step of learning, while the parameter θ^{TD} is updated every fixed step. Besides, DQN set a reply memory to store sample data for each time slot, randomly extracts data from the memory for learning, which breaks the correlation between experiences and improves the training efficiency of the neural network.

DQN is an algorithm based on Q-learning. During the learning process, the parameters of the neural network are updated by the gradient descent method. Therefore, Eq. (12) is changed into

$$\theta_{t+1} = \theta_t + \alpha\left[R_t + \gamma \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}; \theta^{\mathrm{TD}}\right) - Q(s_t, a_t; \theta_t)\right]\nabla Q(s_t, a_t; \theta_t) \tag{15}$$

Where $R_t + \gamma \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}; \theta^{\mathrm{TD}}\right)$ is TD target.

4 Power Allocation Algorithm Based on DQN

In this section, an online power allocation algorithm based on DQN is designed to maximize the secrecy rate of the model shown in Fig. 1. At time slot t, node A selects the transmission power P_t based on the information of the current battery level, the harvested energy, and the channel coefficients.

4.1 Problem Formulation

Define the following key elements to map the problems to the Q learning model.

(1) **State space:** At time slot t, the system state includes battery level B_t, harvested energy E_t, channel coefficients h_t^{ab} and h_t^{ae}, i.e. $s_t = \left(B_t, E_t, h_t^{\mathrm{ab}}, h_t^{\mathrm{ae}}\right)$.

(2) **Action space:** The action space is a set of transmission power that the sending node can select. In our model, the action set A is a set of values from 0 to $P_{\max}$ by step size δ. Since the transmission power selectable by the sending node is limited by the current battery level and the maximum discharge power of the battery, the action set A_t is set to

$$A_t = \left\{0, \delta, 2\delta, \ldots, \hat{P}_{\max}\right\} \tag{16}$$

where $\hat{P}_{\max} = \begin{cases} P_{\max}, & B_t/\tau > P_{\max} \\ B_t/\tau, & B_t/\tau \leq P_{\max} \end{cases}$.

(3) **Reward function:** In this paper, the reward function is improved to the sum of the immediate reward r_t and penalty functions g_t, as follows.

$$R_t = r_t + \beta g_t \tag{17}$$

where $\beta \in [0, 1]$ is a positive real number used to weigh the additional function. The immediate reward is the corresponding benefit when the sending node selects the transmission power P_t, as show in Eq. (18).

$$r_t = \begin{cases} \log_2\left(1 + \frac{P_t \left|h_t^{\mathrm{ab}}\right|^2}{\sigma_{\mathrm{b}}^2}\right) - \log_2\left(1 + \frac{P_t \left|h_t^{\mathrm{ae}}\right|^2}{\sigma_{\mathrm{e}}^2}\right), & \left|h_t^{\mathrm{ab}}\right| > \left|h_t^{\mathrm{ae}}\right| \\ 0, & \left|h_t^{\mathrm{ab}}\right| \leq \left|h_t^{\mathrm{ae}}\right| \end{cases} \tag{18}$$

Since the limitation of battery level, overflow situations must be avoided. Set an additional function

$$g_t = \begin{cases} -1, & B_{t+1} > B\max \\ 0, & B_{t+1} \leq B\max \end{cases} \tag{19}$$

to punish the actions that cause the battery overflow.

4.2 Exploration and Exploitation

In the process of reinforcement learning, there are two choices: exploration and exploitation. Exploration tries different actions and exploitation selects the currently optimal action. Exploration is aggressive behavior that has the opportunity to discover higher-return actions, but may also adopt actions with lower returns. Exploitation is a conservative behavior, which takes action with the highest current return. In DQN, the ε-greedy strategy is used to trade off between exploration and exploitation, as follows.

$$a_t = \begin{cases} a_{\mathrm{random}}, & \text{Probability } \varepsilon \\ \arg\max\limits_{a_t} Q(s_t, a_t), & \text{Probability } 1 - \varepsilon \end{cases} \tag{20}$$

where ε ($0 \le \varepsilon \le 1$) is a parameter for the compromise of exploration and exploitation. a_{random} represents randomly selected actions. In the early stages of training, the agent explores experience to store in replay memory. As the training progresses, more exploitation is chosen to obtain a higher reward, so the ε gradually decreases. One way to control ε as follows.

$$\varepsilon = \begin{cases} 1, & t \le T_1 \\ 1 - (t - T_1) / T_2, & T_1 < t \le T_1 + T_2 \\ 0, & T_1 + T_2 < t \end{cases} \tag{21}$$

where T_1 is the duration of the exploration phase only, and T_2 is the duration of exploration and exploitation phase. The sending node selects the action of the maximum output Q value of the neural network in the current state When $\varepsilon = 0$. To validate the performance of the proposed algorithm, the duration of $\varepsilon = 0$ is set to T_3.

4.3 Proposed Algorithm

As mentioned before, DQN has a replay memory D with a capacity of N to store the sample data of each time slot. The agent randomly extracts M sample data from the memory to learn. And update the parameters θ to θ^{TD} every C step. The online power allocation algorithm based on DQN is shown in Algorithm 1.

Algorithm 1 DQN-based online joint resource power allocation algorithm

Initialize:

neural network parameter update steps C, replay memory D to capacity N, the Q network with random parameters θ, the target Q network with random parameters θ^{TD}

1. Observe the initial observation
2. **For** $t \le T_1 + T_2 + T_3$ **do**
3. Select the transmit power P_t by (20)
4. Get corresponding reward from (17)
5. Observe the next state s_{t+1}
6. Store (s_t, P_t, R_t, s_{t+1}) in replay memory
7. **If** $t > T_1$ **then**
8. Sample random M of (s_i, P_i, R_i, s_{i+1}) from D and calculate TD target y_i
9. $y_i \quad \begin{cases} R_i, & i = T_1 + T_2 + T_3 \\ R_i + \gamma \max\limits_{a_{i+1}} Q\left(s_{i+1}, P_{i+1}; \theta^{\text{TD}}\right), & i \ne T_1 + T_2 + T_3 \end{cases}$
10. perform a gradient descent step on $\left(y_i - Q\left(s_i, P_i; \theta\right)\right)^2$ with respect to the network parameter θ
11. Every C step reset $\theta^{\text{TD}} = \theta$
12. **End if**
13. $s_t \leftarrow s_{t+1}$
14. **End for**

5 Simulation Results

In this section, numerical results for the evaluation of the proposed algorithm on improving the system security rate are presented. For the simulations, the length of the time slot $\tau = 1$ s. The energy harvested by node A at each time slot follows a uniform distribution between $[0, E_{\max}]$. It is assumed that the channel coefficients h_t^{ab} and h_t^{ae} are taken from i.i.d. Rayleigh fading process with zero mean and unit variance, and keep invariant within a time slot. The noise variance is set to $\sigma_b^2 = \sigma_e^2 = 1$W. The initial battery power is 0 J. The maximum discharge power $P_{\max} = 5$ W. The step size δ of the action set is set to 0.04 $P_{\max}$. In addition, Table 1 provides some relevant parameters used in the simulations.

For comparison, we compare it with greedy policy and random policy. Greedy policy allocations the maximum power available for every time slot that satisfies the communication condition, i.e.

$$P_t = \begin{cases} \min\left(\frac{B_t}{\tau}, P_{\max}\right), & \left|h_t^{ab}\right| > \left|h_t^{ae}\right| \\ 0, & \left|h_t^{ab}\right| \leq \left|h_t^{ae}\right| \end{cases} \tag{22}$$

Table 1. Related simulation parameters

Parameter	Value	Meaning
M	32	Sample size each step
N	20000	Memory size
C	200	Update frequency of TD target neural network
γ	0.9	Discount factor
α	0.01	Learning rate
β	0.5	The arguments used to weigh additional function
T1	2000	Exploration phase
T2	48000	Exploration-exploitation phase
T3	10000	Exploitation phase

The greedy policy maximizes current reward without considering the impact of current decisions on the future. Random policy randomly selects the transmission power within the maximum available power range in the time slot that satisfies $|h_t^{ab}| \geq |h_t^{ae}|$.

In Fig. 2, we compare the performance of the proposed algorithm with greedy policy and random policy. The time average secrecy rate is the average of the secrecy rate of each time slot from the beginning of the simulation to the current time. In this case, the battery capacity is set as $B_{\max} = 15$ J, and the maximum energy of collection $E_{\max} = 1$ J. Results show that the performance of the proposed algorithm is significantly better than the other two algorithms.

Figure 3 shows the curve of the average security rate changing with the maximum $E_{\max}$ of collected energy. In the simulation, $B_{\max} = 12E_{\max}$. It can be seen from the

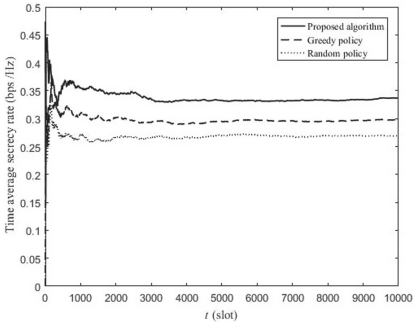

Fig. 2. Time average secrecy rate versus time

figure that the average secrecy rate of all algorithms increases as the $E_{\max}$ increases. This is because, with the harvested energy increases, the more energy the transmission node can use to transmit data, the higher the transmission rate will be.

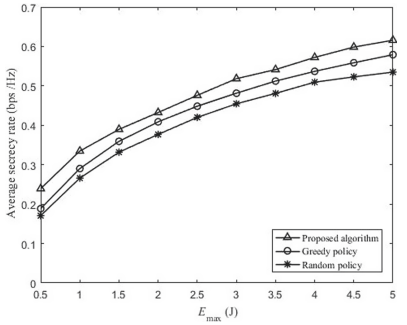

Fig. 3. Average secrecy rate versus $E_{\max}$

Figure 4 shows the effect of the battery buffer size on the performance for $E_{\max} = 1$ J. It can be seen that the performance of the proposed algorithm is close to greedy policy when $B_{\max} < 5$ J. The reason for this is that when the battery capacity is low, it is easy to overflow. In order to avoid this situation, it is reasonable to choose a larger transmission power, so the performance of the proposed algorithm is similar to greedy algorithm. Additionally, the secrecy rate of all algorithms saturates when $B_{\max} > 11$ J. Because the battery has sufficient capacity to buffer the harvested energy for energy dispatching.

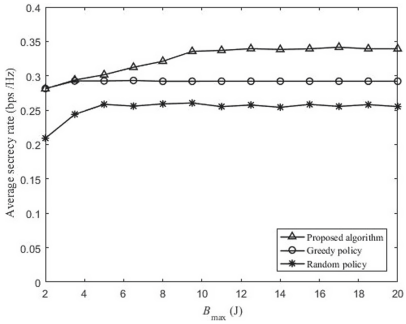

Fig. 4. Average secrecy rate versus $B_{\max}$

6 Conclusion

Based on reinforcement learning, we have studied the maximum secrecy rate of energy harvesting wiretap channel model. The model includes a sending node equipped with energy harvesters, a destination node, and an eavesdropping node. It is assumed that at each time slot only information about the current state of the sending node is available, i.e., battery level, harvested energy, channel state. We analyze the problem of maximizing secrecy rate and model the power allocation problem as an MDP. In order to solve the formulated problem, we use neural networks to estimate the value function. In the end, an online power allocation algorithm based on DQN to improve the secrecy rate is proposed. Simulation results show that the proposed algorithm can effectively optimize energy efficiency.

Acknowledgement. This paper is sponsored by the National Nature Science Foundation of China (61971080, 61471076); Chongqing Basic Research and Frontier Exploration Project (cstc2018jcyjAX0432, cstc2017jcyjAX0204); The Key Project of Science and Technology Research of Chongqing Education Commission (KJZD-K201800603).

References

1. Ku, M., Li, W., Chen, Y., et al.: Advances in energy harvesting communications: past, present, and future challenges. IEEE Commun. Surv. Tutor. **18**(2), 1384–1412 (2016)
2. Ulukus, S., Yener, A., Erkip, E., et al.: Energy harvesting wireless communications: a review of recent advances. IEEE J. Sel. Areas Commun. **33**(3), 360–381 (2015)
3. He, Y., Cheng, X., Peng, W., et al.: A survey of energy harvesting communications: models and offline optimal policies. IEEE Commun. Mag. **53**(6), 79–85 (2015)
4. Tutuncuoglu, K., Yener, A.: Optimum transmission policies for battery limited energy harvesting nodes. IEEE Trans. Wireless Commun. **11**(3), 1180–1189 (2012)
5. Zhou, Q., Yang, Z., Liu, N., et al.: Energy-efficient data transmission with non-FIFO packets with processing cost. IEEE Access **5**, 5158–5170 (2017)
6. Koirala, R., Severi, S., Parajuli, J., et al.: Transmission power optimization for energy harvesting wireless nodes. In: 2015 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, pp. 823–827. IEEE (2015)

7. Mao, Y., Zhang, J., Song, S.H., et al.: Joint link selection and relay power allocation for energy harvesting relaying systems. In: 2014 IEEE Global Communications Conference, Austin, TX, USA, pp. 2568–2573. IEEE (2014)
8. Masadeh, A., Wang, Z., Kamal, A.E.: Reinforcement learning exploration algorithms for energy harvesting communications Systems. In: 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, pp. 1–6. IEEE (2018)
9. Ayatollahi, H., Tapparello, C., Heinzelman, W.: Reinforcement learning in MIMO wireless networks with energy harvesting. In: 2017 IEEE International Conference on Communications (ICC), Paris, France, pp. 1–6. IEEE (2017)
10. Volodymyr, M., Koray, K., David, S., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529 (2015)
11. Wyner, A.D.: The wire-tap channel. Bell Syst. Tech. J. **54**(8), 1355–1387 (1975)
12. Chen, X., Ng, D.W.K., Gerstacker, W.H., et al.: A survey on multiple-antenna techniques for physical layer security. IEEE Commun. Surv. Tutor. **19**(2), 1027–1053 (2017)
13. Xie, X., Zhang, X., Lei, W.: Optimization of secrecy rate for energy harvesting Gaussian wiretap channel. J. Electron. Inf. Technol. **37**(11), 2678–2684 (2015)
14. Jiang, C., Zhang, H., Ren, Y., et al.: Machine learning paradigms for next-generation wireless networks. IEEE Wirel. Commun. **24**(2), 98–105 (2017)
15. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)