Multi-base Station Energy Cooperation Based on Nash Q-Learning Algorithm

Yabo Lv, Baogang Li, Wei Zhao(✉), Dandan Guo, and Yuanbin Yao

Department of Electronics and Communication Engineering,
North China Electric Power University, Baoding 071000, Hebei, China
yabolv@163.com, baogangli@ncepu.edu.cn,
andyzhaoster@163.com, guodanstyle@163.com,
hdyaoyuanbin@outlook.com

Abstract. In view of the current energy problems of communication base station, a multi-base station energy cooperation strategy is proposed to reduce the energy consumption of power grid, which is introducing renewable energy and energy cooperation between the base station based on the Nash-Q learning algorithm. We analyze the packet rate and throughput of the system under the proposed approach. The simulation results show that the proposed algorithm can enhances the adaptability to the changing environment, effectively improve the system capacity.

Keywords: Multi-agent reinforcement learning · Nash equilibrium Q-learning · Energy harvesting

1 Introduction

Recently, with the arrival of the fifth generation (5G) and the rapid development of the cellular network [1, 2], the number of users and the corresponding traffic have greatly increased. Therefore, the energy consumption in cellular networks has also increased significantly. According to the statistics, the cellular network consumes more than 0.5% of the global energy supply [3], the figure will increase as users' demand grows. In some mountainous areas, grasslands and other special areas, communication base stations are not usually directly connected to the grid, so it is necessary to introduce the wind power, solar power or other renewable energy in the base station. Because of the instability and uneven distribution of renewable energy, it is difficult for a single base station to achieve the optimal utilization of energy, and energy cooperation between the base stations can solve the above problems.

The optimal energy allocation based on the off-line algorithm, assuming the non-causal information of the energy and data are known ahead at the transmitter [4]. The other is the study based on the online algorithm, assuming that the transmitter can not know the statistical information ahead [5]. In [6], a distributed reinforcement learning algorithm is proposed to solve the problem of energy cooperation between multi-base stations. In [7], based on a complete model, which is unrealistic, it is difficult or even impossible to obtain such a priori knowledge in reality. [8] studies the single-step TD algorithm only modifies the estimate of the neighbor state, leading to

K. Long et al. (Eds.): 5GWN 2017, LNICST 211, pp. 60–68, 2018.
https://doi.org/10.1007/978-3-319-72823-0_7

algorithm convergence is too slow. Maximize energy efficiency under power constraints in each sub-channel is considered [9–12].

In this paper, we consider a wireless communication system equipped with an energy harvesting device and a limited-capacity rechargeable battery,the base station can maintain system operation by harvesting renewable energy. Because of the uneven distribution of energy, a single base station can not meet the requirements, to solve this problem, we start with general-sum stochastic games, combining with reinforcement learning, propose an on-line algorithm. and apply proposed algorithm to energy coordination in wireless communication systems. The simulation results show the superiority of the proposed algorithm. We compare the data rate under the presence of energy cooperation, and obtain the desired experimental results.

The remainder of the paper is organized as follows. Section 2 describes the system model, while Sect. 3 presents the optimization algorithm, and Sect. 3 also details how the problem is solved using proposed algorithm. Simulation results are presented in Sect. 4. Finally, the conclusion is given in Sect. 5.

2 System Model

We consider a wireless communication system equipped with an energy harvesting device and a rechargeable battery with limited storage capacity, assuming energy and data packets arrive at each time slot (TS), the channel conditions being constant during each TS, and changes from one TS to the next TS. We believe that the packet transmission has a strict transmission delay constraints, that is, the data packets must be sent or be dropped before the next TS arrivals, and the arrival of data and energy in each TS follows a first-order discrete-time Markov model. System model is shown in Fig. 1. There are energy and data package arriving at the transmitter i at TS t, the energy which in battery can cooperate with other base station (BS) through power line or radio frequency, user 1 and user 2 connect BS1 and BS2 respectively.

Where $D_i(t)$ is the data packet arriving at the BS i in the TS t, wherein the data packet satisfies $D_i(t) \in D = \{\mathrm{d}_1, \mathrm{d}_2 \ldots \mathrm{d}_\mathrm{N}\}$, N is the number of elements in D, and d_i represents the type of the packet; $p_d(d_j, d_k)$ is the probability that the packet changes from state d_j to d_k. $\varepsilon_\mathrm{i}(t)$ is the energy harvested by the BS i in TS t, denoted as $\varepsilon_\mathrm{i}(t) \in E^H = \{e_1, e_2 \ldots e_N\}$, and N represents the number of elements in E^H, e is the energy harvested in each TS, $p_e(e_j, e_k)$ is the probability that the harvested energy from e_j to e_k in the next TS. $H_i(t)$ is the channel state of the BS i in TS t, denoted by $H_i(t) \in \mathrm{H} = \{h_1, h_2 \ldots h_N\}$, and $p_h(h_j, h_k)$ is the probability that the next slot channel state is converted from h_j to h_k. $f_1(t)$ is the energy that BS1 transmits to BS2 in TS t, $f_2(t)$ is the energy transmitted by BS2 to BS1 in slot t, K indicates energy transfer efficiency, with $0 < K < 1$, The battery capacity is $B_{\max}$, At any time, the battery power to meet the $0 \le B_i(t) \le B_{\max}$, when the battery is full, the harvested energy is no longer stored in the battery.

At the beginning of the TS t, the transmitter can obtain the channel state $H_i(t)$ and packet $D_i(t)$. According to the Shannon formula, the energy needed to send the packet $E_i^T(t)$ can be calculated, and the packet can be successfully transmitted by the energy

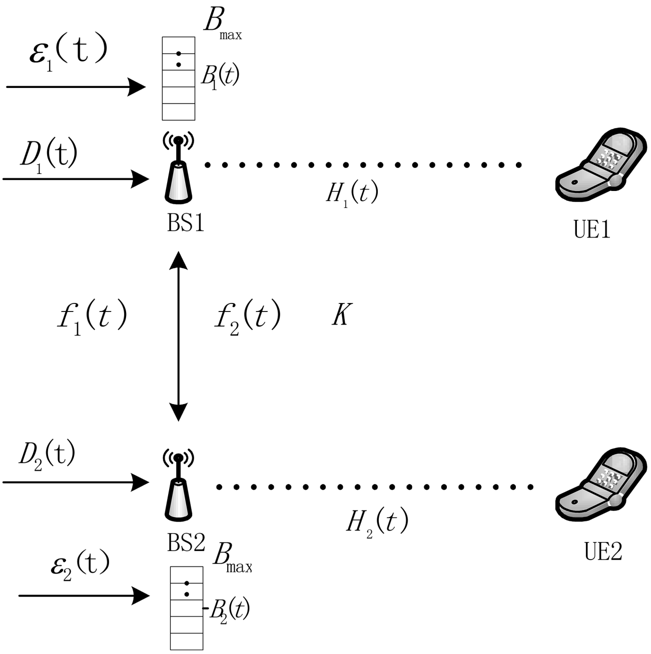

Fig. 1. System model

consumption. According to the causality of energy harvesting, It can be seen that the energy harvested in TS t can only be used in its subsequent time slot, then the next slot battery energy meets:

$$B_1(t+1) = \min\{B_1(t) + \varepsilon_1(t) - f_1(t) + kf_2(t) - a_1(t)E_1^T(t), B_{\max}\} \tag{1}$$

$$a_1(t)E_1^T(t) \le B_1(t) \tag{2}$$

$$B_2(t+1) = \min\{B_2(t) + \varepsilon_2(t) - f_2(t) + kf_1(t) - a_2(t)E_2^T(t), B_{\max}\} \tag{3}$$

$$a_2(t)E_2^T(t) \le B_2(t) \tag{4}$$

The objective of this paper is to maximize the average transmission data rate of the system, which is given by:

$$\bar{r}(t) = \max \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=0}^{T} a_i D_i \tag{5}$$

S.t. (1) (2) (3) and (4).

3 Optimization Algorithm Analysis

3.1 Basic Q-Learning Algorithm

The Q-learning algorithm was first proposed by C. Watkins in his PhD thesis. The algorithm can only be assumed to be a Markov decision process model based on the underlying system. The system does not need to know other priori information, and the algorithm can converge to the optimal strategy by learning to enhance the discount return value. The iterative calculation formula is

$$Q(s_t, a_t) = Q(s_t, a_t) + l(r(s_t, a_t) + \gamma \ \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \tag{6}$$

Where (s_t, a_t) is the state-action pair of MDP at time t, s_{t+1} is the state at time $t+1$, $r(s_t, a_t)$ is the return at time t, and $l > 0$ is the learning factor. When certain conditions are satisfied, the algorithm converges to the optimal solution [13].

From Eq. (6), it can be seen that the update of Q-learning-valued function is carried out in an iterative way. The value function is the expectation of the reward discount after the action a_t selected according to a certain policy under the state s_t. There are two ways to achieve Q learning: One method is to use lookup table, the other is the use of neural network. This paper uses lookup table method. In the initialization phase, all states and actions are initialized to discrete quantities, and the value functions are learned according to algorithmic flow and constraints.

3.2 Nash Q Learning Algorithm

Hu and others in 1998 proposed the Nash-Q algorithm, which extends the multi-agent learning to complete information non-cooperative general and stochastic game with incompletely antagonistic interests [14].

First, we give the definition of Nash equilibrium: Nash equilibrium is a joint strategy, in this state, each participant's strategy for other participants are excellent. In a random game process, Nash equilibrium is a n-tuple stratergy $(\pi_*^1, \cdots, \pi_*^n)$, such that for all $s \in S, i = 1, \cdots, n$, there are:

$$v^i(s, \pi_*^1, \cdots, \pi_*^n) \geq v^i(s, \pi_*^1, \cdots, \pi_*^{i-1}, \pi_*^i, \pi_*^{i+1}, \cdots, \pi_*^n) \tag{7}$$

There $\pi_i \in \Pi_i$, Π_i for the agent i available strategies.

In the Nash Q-learning algorithm, the Nash equilibria are used to define the value functions. In this case, the agents can observe each other to obtain information such as the action taken and the reward they get, etc., and to update their own valued functions. The value function of other agents is modeled. In a game with n players, the Q value of all agents in the same state forms a countermeasure form, $Q_t^1(s), \ldots, Q_t^n(s)$, and the value function update formula of Nash Q-learning algorithm is:

$$Q_{t+1}^{i}(s_t^i, a_t^1 \ldots a_t^n) = (1-\alpha_t)Q_t^i(s_t^i, a_t^1 \ldots a_t^n) + \alpha_t\left[r_t^i + \beta NashQ_t^i(s')\right] \tag{8}$$

S.t. $N\text{ash}Q_t^i(s') = \pi^1(s') \ldots \pi^n(s') \cdot Q_t^i(s')$

$\pi^1(s') \ldots \pi^n(s')$ is the Nash equilibrium solution of Q value in the state s'. Indicates that the participant i selects the winning function of the Nash equilibrium solution under state s'.

At the same time, the agent needs to update the value function of other agents by the following equation:

$$Q_{t+1}^{j}(s_t^j, a_t^1 \ldots a_t^n) = (1-\alpha_t)Q_t^j(s_t^j, a_t^1 \ldots a_t^n) + \alpha_t\left[r_t^j + \beta NashQ_t^j(s')\right] \tag{9}$$

S.t. $N\text{ash}Q_t^j(s') = \pi^1(s') \ldots \pi^n(s') \cdot Q_t^j(s'), \quad i \neq j$

Given the Nash Q-learning algorithm steps, the learning process of the agent i can be described as follows:

Initialization:

Set initial time index $t \leftarrow 0$, initial state $s_{\mathrm{t}}^i = s_0^i$.

For all $s_t^i \in S$, $\mathrm{a}_t^i \in A$, $i = 1, 2, \cdots, n$, $t = 0, 1, \cdots$, the initialized value functions $Q_t^i(s_t^i, a_t^1 \ldots a_t^n) = 0$, $Q_t^j(s_t^j, a_t^1 \ldots a_t^n) = 0, j = 1, 2 \cdots n$, and $j \neq i$.

Repeat the following steps until the condition is met:

(a) observe the current state s_t^i, according to the rules and learning process to get $Q_t^i(s_t^i, a_t^1 \ldots a_t^n)$ and $Q_t^j(s_t^j, a_t^1 \ldots a_t^n)$, according to the greedy strategy to select action a_t^i;

(b) Observe the joint reward $r_t^1 \ldots r_t^n$ and joint action $a_t^1 \ldots a_t^n$ in the current state and update the value functions of themselves and other agents according to (8) and (9).

(c) Let $t = t + 1$, observe the next state s'.

In this section, we consider two base stations, and each base station can be regarded as an agent with learning ability. Agent i calculates $\pi^1(s')\pi^2(s')$ for the stage game $(Q_t^1(s')Q_t^2(s'))$, and update the Q-value according to (8) and (9). the state set consists of four parts: the harvested energy, arrival data, channel state, and battery capacity. At TS t, the state is represented by $S_t = \, < \varepsilon(t), D(t), H(t), B(t) >$, according to the following simulation parameters set, the system is divided into 48 discrete state. And the action set is represented by $A_t = \, < a(t), f(t) >$, where $a(t) = \, < 0, 1 >$ is whether to send data packets, $f(t) = \, < 0, 2 >$ Which indicates the energy of cooperation between the two BS, whether the action can be performed is governed by Eqs. (2) and (4).

4 Simulation Results Analysis

In this paper, simulation parameters similar to the paper [5] are used. Assuming the length of each TS is $\Delta_{TS} = 10\,\text{ms}$, the time that the transmitter transmits data is $\Delta_{Tx} = 5\,\text{ms}$, the available bandwidth is $W = 2\,\text{MHz}$, the noise power spectral density of Gaussian channel is $N_0 = 10^{-20.4}$ W/Hz, The basic energy unit is 2.5 μj, assuming that the transmitter in time slot t, the available energy unit is $\varepsilon_i(t) = \{0, 5\}\mu j$, the packet size is $D_i(t) = \{300, 600\}$ bit/s, the channel state is $H_i(t) = \{1.655 \times 10^{-13}$, $3.311 \times 10^{-13}\}$; set to harvested energy in each TS, the arrival of the packet, the channel state are random.

In order to ensure the reliable transmission of data, it is necessary to calculate the energy required to transmit data in each state. From the channel capacity formula (Shannon formula) under the Gaussian channel, can be described as

$$D_i(t) = W\Delta_{Tx} \log_2(1 + \frac{H_i(t)P}{WN_0}) \tag{10}$$

The channel capacity may be approximately equal to

$$D_n \approx \frac{\Delta_{Tx} H_n P}{\log(2)N_0} \tag{11}$$

Where is the energy required to transmit the data in the TS, so we get the energy required to reliably send a packet is:

$$E_i^T = f_e(D_i(t), H_i(t)) = \frac{D_i(t)\log(2)N_0}{H_i(t)} \tag{12}$$

In the simulation, taking into account the causal nature of energy harvesting, the renewable energy harvested by the base station is used in the following TS, the system depends on the current battery remaining capacity to determine whether to send data packets and co-energy, the data package will be sent successfully if the energy in battery is enough in current TS, otherwise the package will be dropped.

Figure 2 Shows the relationship between the number of learning iterations and the transmission data rate. It can be seen that the system throughput increases with the number of learning times. When the number of iterative times reaches 600, the data rate reaches 77 kb/s, and the learning data rate is no longer significantly improved and stable. The learning process has been gradually completed. The blue curve is the data rate under the greedy strategy. If the greedy strategy is adopted, only the local optimal solution is adopted. The throughput of the system is 52 kb/s, its throughput is much lower than the Nash Q learning algorithm. It can be seen that the Nash Q learning algorithm can effectively improve the data rate of the system.

The impact of battery capacity on the system throughout shown in Fig. 3. With the battery capacity increases, the system throughout gradually increased, because the energy harvesting probability is a fixed value, in a continuous period of time, the system will harvest more renewable energy, increased battery capacity can store more available electrical energy, which can send more data during no energy harvesting time, so an appropriate increase in battery capacity will help improve the system throughout. From the curve, we can see that as the battery capacity increases, the data rate increases slowly, the growth rate decreased significantly when the rate increase to 80 kb/s, because the system throughout is not only determined by the battery capacity, but also by the packet size, capture Energy probability, channel state and other factors.

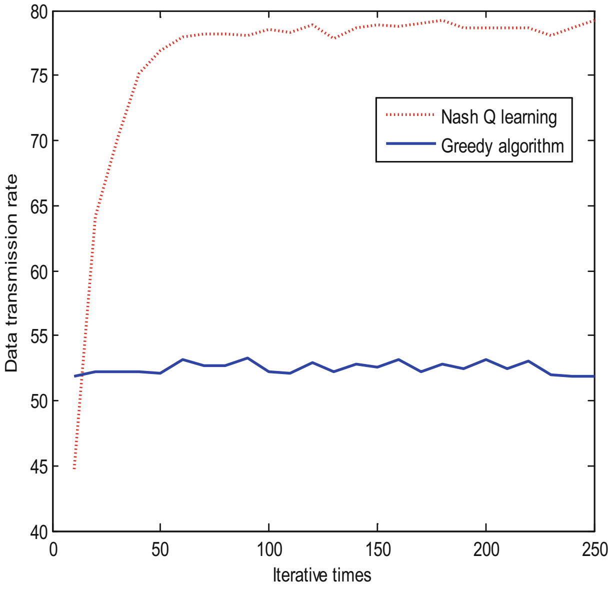

Fig. 2. Influence of iteration times on rate

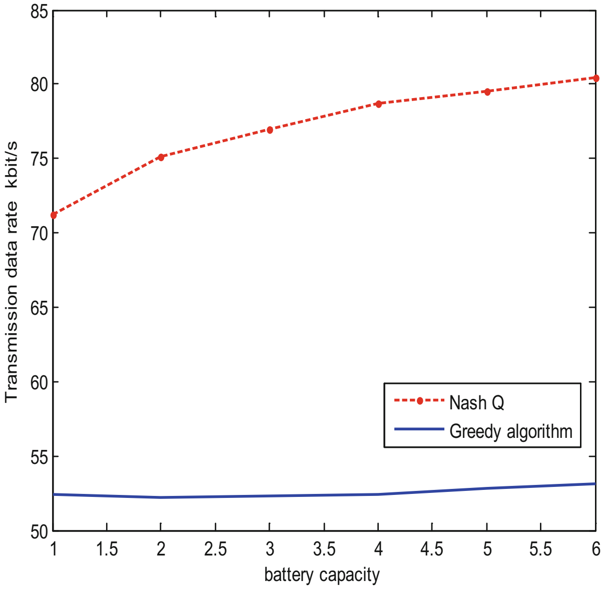

Fig. 3. Battery capacity on the rate of impact

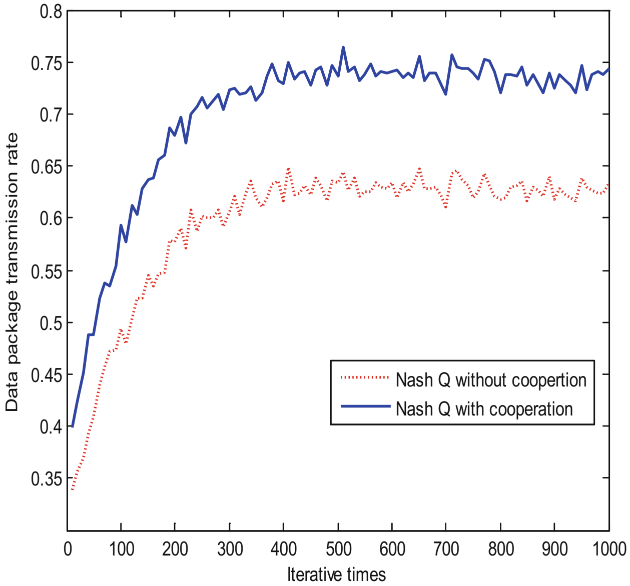

Fig. 4. Effect of energy cooperation on transmit rate under Q-Learning

Figure 4 shows the impact of energy cooperation on the successful transmission of data packets to the base station. From the simulation results, we can see that when the energy cooperation is not carried out between the base stations, the data packets transmission rate of the system is as high as 60%. When the energy cooperation is carried out, The two base stations can share the excess energy, the current state of excess energy sharing to another energy-poor base stations, the data packet rate can be up to 73%. From the comparison we can conclusion that the cooperation energy have a significance on improving the data transmission rate.

5 Conclusion

In order to reduce the energy consumed by the communication base station and increase the flexibility of the deployment of the base station, more and more researchers consider the use of renewable energy to the base station power supply. This paper have studied the energy allocation and the energy cooperation between off-grid base stations. Based on the knowledge of game theory, we have proposed an online energy management method. The simulation results of Nash Q learning algorithm have shown that the information rate of the system can be improved effectively with the agent learning process. It has been shown that, multi-base station energy cooperation method is superior to single base station communication system. For the off-grid connected base station rechargeable battery capacity, we can conclusion that the appropriate increase in battery capacity can increase the system speed.

The following research will be carried out from the following three aspects: The reinforcement learning algorithm is applied to solve the energy optimization problem in the coexistence scenario of grid and renewable energy. Improvement of reinforcement learning algorithm, the emphasis is on improving the robustness and convergence speed of the algorithm, reducing the complexity of multi-agent learning and exploring the method of solving the dimension problem. Consider the co-optimization of energy efficiency and spectral efficiency.

References

1. Zhang, H., Dong, Y., Cheng, J., Hossain, M.J., Leung, V.C.M.: Fronthauling for 5G LTE-U ultra dense cloud small cell networks. IEEE Wirel. Commun. **23**(6), 48–53 (2016)
2. Zhang, H., Liu, N., Chu, X., Long, K., Aghvami, A., Leung, V.: Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. IEEE Commun. Mag. **55**(8), 138–145 (2017)
3. Tombaz, S., Vastberg, A., Zander, J.: Energy-and cost-efficient ultrahigh-capacity wireless accross. IEEE Wirel. Commun. **18**(5), 18–24 (2011)
4. Gong, J., Zhou, S., Niu, Z.: Optimal power allocation for energy harvesting and power grid coexisting wireless communication systems. IEEE Trans. Commun. **61**(7), 3040–3049 (2013)
5. Pol Blasco, D., Dohler, M.: A learning theoretic approach to energy harvesting communication system optimization. IEEE Trans. Wirel. Commun. **12**(4), 1872–1882 (2013)
6. Lin, W.T., Lai, I.W., Lee, C.H.: Distributed energy cooperation for energy harvesting nodes using reinforcement learning. In: 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, pp. 1584–1588 (2015)
7. Udenze, A., McDonald-Msier, K.: Discrete time markov based optimisation for dynamic control of transmitter power in wireless sensor networks. In: IET Proceedings of the 3rd UK Embedded Forum, Newcastle, UK, 2–3 April (2007)
8. Alhajry, M., Alvi, F., Ahmed, M.: TD(λ) and q-learning based ludo players. In: IEEE Conference on Computation Intelligence and Games (CIG), pp. 83–90 (2012)
9. Li, W., Zhang, H., Zheng, W., Su, T., Wen, X.: Energy-efficient power allocation with dual-utility in two-tier OFDMA femtocell. In: Proceedings of IEEE Globecom (2012)
10. Liu, H., Zheng, W., Zhang, H., Zhang, Z., Wen, X.: An iterative two-step algorithm for energy efficient resource allocation in multi-Cell OFDMA networks. In: Proceedings of IEEE WCNC (2013)
11. Ma, W., Zheng, W., Zhang, H., Wen, X.: MOS-driven energy efficient power allocation for wireless video communications. In: Proceedings of IEEE Globecom (2012)
12. Zhang, Z., Zhang, H., Lu, Z., Zhao, Z., Wen, X.: Energy-efficient resource optimization in OFDMA-based dense femtocell networks. In: Proceedings of IEEE ICT (2013)
13. Dayan, W.P.: Q-Learning. Mach. Learn. **38**(2), 362–399 (2002)
14. Hu, J.L., Wellmam, M.P.: Multiagent reinforcement learning: theoretical framework and an algorithm. In: 15th International Conference on Machine Learning, Washington, pp. 242–249 (1999)