A. Ortiz, T. Weber and A. Klein, "A Two-Layer Reinforcement Learning Solution for Energy Harvesting Data Dissemination Scenarios," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, April 2018.

A TWO-LAYER REINFORCEMENT LEARNING SOLUTION FOR ENERGY HARVESTING DATA DISSEMINATION SCENARIOS

Andrea Ortiz[*], *Tobias Weber*[†], *Anja Klein*[*]

[*]Communications Engineering Lab, Technische Universität Darmstadt, Merckstr. 25, 64283 Darmstadt, Germany

[†]Institute of Communications Engineering, University of Rostock, Richard-Wagner-Str. 31, 18119 Rostock, Germany

ABSTRACT

A data dissemination scenario is considered. The transmitter harvests energy from the environment and uses it to transmit individual data to multiple receivers. We consider a realistic scenario in which only causal knowledge regarding the energy harvesting, the channel fading and the data arrival processes is available. Our goal is to find a power allocation policy aiming at maximizing the throughput. We propose a two-layer reinforcement learning algorithm which divides the learning task into two sub-tasks, namely, how much power to use in each time interval and how to split the power among the data to be transmitted. By dividing the task, we increase the learning speed as compared to the standard reinforcement learning algorithms Q-learning and SARSA. Moreover, the proposed algorithm outperforms reference policies that deplete the battery in every time interval.

Index Terms— Data Dissemination, Energy Harvesting, Reinforcement Learning

1. INTRODUCTION

Energy harvesting (EH) communication devices are able to collect ambient energy from the environment to recharge their batteries. This means, EH communication devices are able of self-sustainability and theoretical perpetual operation [1].

We consider a data dissemination scenario with an EH transmitter and multiple receivers. Research effort on EH data dissemination scenarios has mainly focused on offline approaches in which perfect non-causal knowledge regarding the EH, the channel fading and the data arrival processes is available [2–7]. In [2], an EH transmitter with an infinite battery broadcasting individual data packets to two receivers over an additive white Gaussian noise (AWGN) channel is considered. Similarly, in [3] a two-user EH broadcast (BC) channel with a finite battery and fading channels is studied. Authors in [4] and [5] consider an EH transmitter with a fixed number of data packets to be sent to multiple receivers. The goal is to minimize the time required to deliver the data packets. In [6], the total delay in a two-user EH BC channel is minimized and in [7] the effect of an inefficient battery is investigated.

This work was funded by the LOEWE Priority Program NICER.

In [8], a two-user EH BC scenario, in which the amounts of harvested energy are causally known, is studied and the optimal power scheduling policy when the EH process follows a Bernoulli distribution is found.

In real scenarios, the requirement of perfect non-causal knowledge, as in [2–7], or knowledge about the statistics of the processes, as in [8], cannot be fulfilled. However, a learning approach can be considered to overcome this requirement. This is because in learning approaches, more specifically in reinforcement learning (RL), an agent learns how to behave in an unknown environment by interacting with it. This approach has been applied to EH point-to-point scenarios in [9–12], two-hop communication scenarios in [13, 14] and to multiple access channels in [15].

In this paper, an EH data dissemination scenario in which the EH transmitter sends individual data to multiple receivers is considered. Only causal knowledge is assumed to be available. Our goal is to find a power allocation policy that aims at maximizing the amount of data at the receivers. To find the power allocation policy, a two-layer RL algorithm is proposed which divides the learning task into two sub-tasks, namely, how much power to allocate in each time interval and how to split the allocated power among the data to be transmitted. This division is inspired by [5], where the offline optimal transmission policy is found by reducing the BC channel to consider a single receiver at a time and the power is allocated according to the hierarchy of the channel gains. In each time interval, the upper layer of the proposed algorithm learns how much power to allocate in order to avoid battery overflows. The result is fed into the lower layer which learns how to distribute the power for the transmission of the individual data considering the avoidance of data buffer overflows and aiming at maximizing the throughput. By dividing the task, the proposed algorithm achieves a larger learning speed compared to the standard RL algorithms Q-learning and SARSA. Moreover, it outperforms reference low-complexity approaches.

The rest of the paper is organized as follows. In Sec. 2, the system model is presented. In Sec. 3, the power allocation problem in a data dissemination scenario is formulated. The proposed two-layer RL algorithm that aims at maximizing the throughput is explained in Sec. 4. Performance results are presented in Sec. 5 and Sec. 6 concludes the paper.

2. SYSTEM MODEL

A data dissemination scenario consisting of a single-antenna transmitter and K single-antenna receivers is considered. As depicted in Fig. 1, the transmitter N_0 harvests energy from the environment and uses it exclusively for transmitting data to the K receivers $\mathrm{N}_k, k = 1, ..., K$.

A time slotted system using I time intervals is considered with a constant duration τ for each time interval $i, i = 1, ..., I$. At the beginning of time interval i, N_0 receives an amount of energy $E_i \in \mathbb{R}^+$. The maximum amount of energy $E_{\max}$ that can be harvested depends on the energy source. E_i is stored in a rechargeable finite battery with maximum capacity $B_{\max}$. As the harvested energy cannot be instantly stored in the battery, E_i cannot be used in time interval i but earliest in time interval $i+1$. The battery level B_i is measured at the beginning of time interval i. At the beginning of time interval $i = 1$, N_0 has not yet harvested any energy and $B_1 = 0$. The data intended for each N_k is different and depends on a particular data arrival process. In our model, we divide the data buffer of N_0 in K equal size virtual data buffers as shown in Fig. 1. The size of each virtual data buffer in bits is $D_{\max}$. At the beginning of time interval i, $M_{k,i}$ data packets intended for N_k are received and stored in the corresponding virtual data buffer. To simplify the notation, we assume that all incoming data packets have the same size d. The level of the virtual data buffer containing the data intended for N_k is measured at the beginning of time interval i and is denoted by $D_{k,i}$. At the beginning of time interval $i = 1$, $D_{k,1} = 0$.

The fading channel from N_0 to N_k is described by the channel coefficient $h_{k,i} \in \mathbb{C}$. It is assumed that $h_{k,i}$ stays constant for one time interval. The noise at N_k is i.i.d. zero mean AWGN with variance σ^2. Additionally, a bandwidth W is available for the transmission to all receivers. The transmitted signal is the superposition of the data intended for the different receivers. The power values $p_{k,i}$, used for transmitting to N_k in time interval i are kept constant during the time interval. Furthermore, the throughput

$$R_{k,i} = \tau W \log_2 \left(1 + \frac{|h_{k,i}|^2 p_{k,i}}{\sum_{j \neq k; j=1}^{K} |h_{j,i}|^2 p_{j,i} + \sigma^2} \right) \quad (1)$$

in bits, is the amount of data received by N_k in time interval i. Note that in the interference term, $p_{j,i} = 0$ if N_j is not served. Only energy stored in the battery can be allocated. Therefore,

$$\sum_{k=1,...,K} \tau p_{k,i} \leq B_i \ \ \forall i, \quad (2)$$

must be fulfilled. Additionally, to avoid battery overflows in which part of the harvested energy is wasted because the battery is full, the battery overflow condition

$$B_{\max} \geq B_i + E_i - \sum_{k=1,...,K} \tau p_{k,i}, \quad \forall i, \quad (3)$$

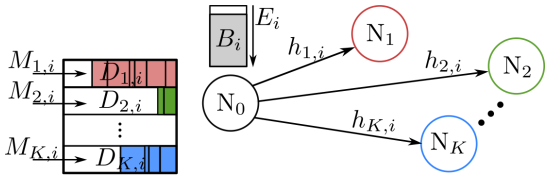

Fig. 1: Data dissemination scenario with an EH transmitter.

is also considered. Only data already stored in the data buffer can be transmitted. Therefore, the data causality condition

$$D_{k,i} \geq R_{k,i} \quad \forall k, i \quad (4)$$

has to be fulfilled. Similar to (3), we define the data buffer overflow condition as

$$D_{\max} \geq D_{k,i} + dM_{k,i} - R_{k,i}, \quad \forall i. \quad (5)$$

3. PROBLEM FORMULATION

Next, we formulate the power allocation problem in an EH data dissemination scenario. With only causal knowledge available, in each time interval i, N_0 decides how much power to allocate for the transmission of the individual data. We model this problem as a Markov decision process (MDP) which consists of a set $\mathcal{S}$ of states, a set $\mathcal{A}$ of actions, a transition model $\mathcal{P}$ and a set $\mathcal{R}$ of rewards [16]. The proposed RL algorithm provides a solution of the MDP presented here.

In time interval i, the state $S_i \in \mathcal{S}$ is a function of E_i, B_i, $h_{k,i}$ and $D_{k,i}, \forall k$. As E_i, B_i and $h_{k,i}$, can take any value in a continuous range, the set $\mathcal{S}$ contains infinitely many possible states. The set $\mathcal{A}$ contains the transmit power tuples $a_i = (p_{1,i}, ..., p_{K,i})$ that can be selected. In our model, $\mathcal{A}$ is finite and each $p_{k,i}$ in a_i is taken from the set $\{0, \delta, 2\delta, ..., B_{\max}\}$, where δ is an arbitrary step size. $\mathcal{P}$ defines the probability of going from S_i to S_{i+i} after performing a_i. Finally, the rewards $r_i \in \mathcal{R}$ indicate how beneficial it is to select a_i in S_i.

The solution of the MDP is given by the policy π which maps states to actions, i.e., $a_i = \pi(S_i)$ [17]. To evaluate π, the so-called action-value function $\mathrm{Q}^\pi(S_i, a_i)$ is used. $\mathrm{Q}^\pi(S_i, a_i)$ is the expected reward starting in S_i, performing a_i and following π thereafter [17]. The policy whose Q^π is greater than or equal to the one for any other policy for every S_i and a_i is an optimal policy π^* and the corresponding optimal action-value function is denoted by Q^*. When Q^* is known, π can be easily determined because for each S_i, any a_i that maximizes $\mathrm{Q}^*(S_i, a_i)$ is an optimal action.

As a consequence of having only causal knowledge, N_0 does not know in advance for how many time intervals it will operate. Similar to [9], we consider a discount factor γ, $0 \leq \gamma \leq 1$ to account for the preference of achieving a higher throughput in the current time interval vs. achieving a higher throughput later on. We aim at maximizing the amount of

transmitted data given by

$$R = \lim_{I\to\infty} \mathbb{E}\left[\sum_{i=1}^{I}\sum_{k=1}^{K} \gamma^i R_{k,i}\right], \quad (6)$$

where $R_{k,i}$ is defined by (1).

4. TWO-LAYER RL ALGORITHM

Here, the proposed two-layer RL algorithm is presented. The two layers are motivated by the fact that the set $\mathcal{A}$ of Sec. 3 grows exponentially with K, i.e., $|\mathcal{A}| = |\{0, \delta, 2\delta, ..., B_{\max}\}|^K$, where $|\cdot|$ is the cardinality of the set. Such a large action set reduces the learning speed and hence the performance since more actions need to be tried to find the optimal policy. Moreover, for large K, only the average channel gain and data buffer levels are relevant to calculate the total power to be used in each time interval. In our two-layer algorithm, each layer solves part of the power allocation problem. In each time interval i, the upper layer decides the total power to be used and the lower layer decides how to distribute it.

4.1. SARSA algorithm

Based on our previous work [10], we use SARSA with linear function approximation in each of the layers. The main idea of SARSA is to build an estimate of Q^π based on the visited states, the performed actions and the obtained rewards. In each time interval i, the actions are taken by following the ϵ-greedy policy on Q^π. This is, in S_i there is a probability $1-\epsilon$ of selecting the action a_i that yields the highest Q^π and a probability ϵ of randomly selecting any action a_i. This method provides a trade-off between the exploitation of known actions and the exploration of new ones [16, 17].

When $\mathcal{S}$ is infinite, linear function approximation can be used to represent Q^π as a weighted sum of feature functions [17]. Each feature function maps S_i and a_i onto a feature value. Let $\mathbf{f}$ be a vector containing all the feature values and let $\mathbf{w}$ be a vector of weights containing the contribution of each feature value. Q^π is then approximated as [17]

$$\mathrm{Q}^\pi(S_i, a_i) \approx \hat{\mathrm{Q}}^\pi(S_i, a_i, \mathbf{w}) = \mathbf{f}^{\mathrm{T}}\mathbf{w}. \quad (7)$$

In each time interval i, $\mathbf{w}$ is adjusted in the direction that reduces the error between Q^π and $\hat{\mathrm{Q}}^\pi$ following the gradient descent approach. Formally, the update rule is given by [17]

$$\Delta\mathbf{w} = \alpha_i\big[r_i + \gamma\hat{\mathrm{Q}}^\pi(S_{i+1}, a_{i+1}, \mathbf{w}) - \hat{\mathrm{Q}}^\pi(S_i, a_i, \mathbf{w})\big]\mathbf{f}, \quad (8)$$

where α_i is the learning rate. Next, we define the set $\mathcal{A}$, the rewards r_i and the feature functions to be used in each layer.

4.2. Upper layer

The upper layer decides on the total transmit power p_i to allocate in each time interval such that battery overflows are avoided. i.e., $a_i = p_i$. In a fading downlink channel, capacity can be achieved if the power is allocated for transmitting to the receiver with the best channel [18]. To find p_i, we reduce the scenario to a point-to-point scenario considering only the receiver with the best channel in time interval i. We denote this best channel as h_i^*. Note that this does not mean that only the receiver with the best channel will be served. It is only used as a reference since it provides an upper bound of the possible performance.

For this layer, we set $\mathcal{A} = \{p_i | p_i \in \{0, \delta, 2\delta, ..., B_{\max}\}\}$ and the reward obtained by selecting p_i as $r_i(p_i) = \log_2(1 + p_i|h_i^*|^2)$. As this layer solves an EH point-to-point communication problem, we use the feature we defined in [10, 13]. $\mathrm{f}_1^{\mathrm{up}}$ indicates if in S_i, the selection of p_i fulfills the conditions in (2) and (3) and it is given by

$$\mathrm{f}_1^{\mathrm{up}}(S_i, p_i) = \begin{cases} 1, & \text{if } (B_i + E_i - \tau p_i \le B_{\max}) \wedge \\ & (\tau p_i \le B_i) \\ 0, & \text{else,} \end{cases} \quad (9)$$

where $\wedge$ is the logical conjunction operation. $\mathrm{f}_2^{\mathrm{up}}$ performs the water-filling (WF) algorithm between h_i^* and the mean of all the past channel gains of all receivers. Let p_i^{WF} be the power calculated with WF. $\mathrm{f}_2^{\mathrm{up}}$ is given by

$$\mathrm{f}_2^{\mathrm{up}}(S_i, p_i) = \begin{cases} 1, & \text{if } \delta \left\lfloor p_i^{\mathrm{WF}}/\delta \right\rfloor = p_i \\ 0, & \text{else,} \end{cases} \quad (10)$$

where $\lfloor\cdot\rfloor$ is the floor function. $\mathrm{f}_3^{\mathrm{up}}$ is activated if $E_i \ge B_{\max}$. In this case, the battery should be depleted to minimize the energy losses due to battery overflow. $\mathrm{f}_3^{\mathrm{up}}$ is written as

$$\mathrm{f}_3^{\mathrm{up}}(S_i, p_i) = \begin{cases} 1, & \text{if } (E_i \ge B_{\max}) \wedge \left(p_i = \delta\lfloor\frac{B_i}{\tau\delta}\rfloor\right) \\ 0, & \text{else.} \end{cases} \quad (11)$$

$\mathrm{f}_4^{\mathrm{up}}$ allocates a larger p_i when a data buffer overflow situation is imminent. Let D_i^* be the highest data buffer level among all $D_{k,i}$ and $\bar{M}_i$ be the average amount of incoming data packets. $\mathrm{f}_4^{\mathrm{up}}$ is given by

$$\mathrm{f}_4^{\mathrm{up}}(S_i, p_i) = \begin{cases} 1, & \text{if } (D_i^* + d\bar{M}_i - r_i(p_i) \le D_{\max}) \\ & \quad \wedge (r_i(p_i) \le D_i^*) \\ 0, & \text{else,} \end{cases} \quad (12)$$

where $r_i(p_i)$ is the reward to be obtained if p_i is selected.

4.3. Lower layer

The task of this layer is to distribute p_i among the individual data to be transmitted aiming at minimizing data buffer overflows and maximizing the throughput. The p_i selected in the upper layer is used as an input. Let $\rho_{k,i}$ be a fraction indicating how much of p_i is assigned to the transmission of data intended for N_k, i.e., $p_{k,i} = \rho_{k,i}p_i$. For this layer,

$\mathcal{A} = \{a_i = (\rho_{1,i}, ..., \rho_{K,i}) | \sum_{k=1}^{K} \rho_{k,i} = 1\}$ and $r_i(a_i) = \sum_{k=1}^{K} R_{k,i}$, with $R_{k,i}$ given by (1). We propose three feature functions based on three different transmission strategies, namely, water-filling (WF), maximum rate (MR) and proportional fairness (PF). $\mathrm{f}_1^{\mathrm{lo}}$ distributes p_i using the WF algorithm. Let a_i^{WF} be the distribution obtained with, then $\mathrm{f}_1^{\mathrm{lo}}$ is defined as

$$\mathrm{f}_1^{\mathrm{lo}}(S_i, a_i) = \begin{cases} 1, & \text{if } a_i = a_i^{\mathrm{WF}} \\ 0, & \text{else.} \end{cases} \tag{13}$$

$\mathrm{f}_2^{\mathrm{lo}}$ is based on MR. It allocates p_i for the transmission to the receiver with the strongest channel. Let j be the index of the receiver with the strongest channel. $\mathrm{f}_2^{\mathrm{lo}}$ is written as

$$\mathrm{f}_2^{\mathrm{lo}}(S_i, a_i) = \begin{cases} 1, & \text{if } a_i \in \mathcal{A} \cap \{a_i | \rho_{j,i} = 1\} \\ 0, & \text{else.} \end{cases} \tag{14}$$

$\mathrm{f}_3^{\mathrm{lo}}$ is based on the PF scheduler in [19]. Let $R'_{k,i}(p_i)$ be the data packets that would be sent if p_i is allocated for the transmission to N_k and let υ and β be tunable parameters that control the fairness. $\mathrm{f}_3^{\mathrm{up}}$ allocates p_i for the transmission to N_j if $j = \mathrm{argmax}_{\forall k} \frac{(R'_{k,i}(p_i) D_{k,i})^{\upsilon}}{\frac{1}{i}\sum_{l=1}^{i} (R_{k,l})^{\beta}}$. For PF, $\upsilon = \beta = 1$ and $\mathrm{f}_3^{\mathrm{lo}}$ is

$$\mathrm{f}_3^{\mathrm{lo}}(S_i, a_i) = \begin{cases} 1, & \text{if } a_i \in \mathcal{A} \cap \{a_i | \rho_{j,i} = 1\} \\ 0, & \text{else.} \end{cases} \tag{15}$$

5. PERFORMANCE RESULTS

For the evaluation of the proposed algorithm, one hundred independent random realizations are generated. Each realization is an episode of $I = 1000$ time intervals. The amounts of harvested energy E_i are taken from a uniform distribution with a maximum value $E_{\max}$. We set the battery capacity $B_{\max} = 2E_{\max}$ and the time interval duration τ to one time unit. The channel between N_0 and N_k is assumed to be i.i.d. Rayleigh fading with zero mean, unit variance and a path loss exponent of 3.5. The noise variances are assumed to be $\sigma^2 = 1$. We set $\delta = 0.02 B_{\max}$, $\gamma = 0.9$ and $\alpha = \epsilon = 1/i$. Moreover, a bandwidth of $W = 1$MHz, and a data packet size of $d = 200$kbits are assumed. The data buffer size is calculated considering a unit channel gain as $D_{\max} = \lfloor W \log_2(1 + B_{\max}) \rfloor$. The incoming data packets are taken from a Poisson distribution with an average amount of five data packets per time interval.

As a comparison, we consider Q-learning, SARSA [10], and the equal power allocation (EPA) and MR policies. For Q-learning, the set $\mathcal{S}$ is discretized and the set $\mathcal{A}$ defined in Sec. 3 is used. For SARSA, we only consider the upper layer explained in Sec. 4.2 and to minimize the interference, the selected power is allocated in each time interval for the transmission to the receiver with best channel conditions. The EPA and MR policies deplete the battery in each time interval. EPA allocates equal amounts of power for the transmission of data and the MR policy spends the energy in the battery for the transmission to the receiver with the best channel conditions.

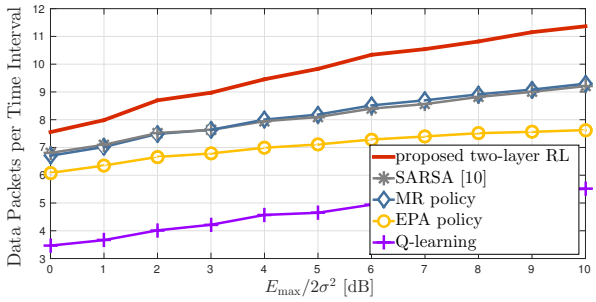

Fig. 2: Average throughput vs. $E_{\max}/(2\sigma^2)$.

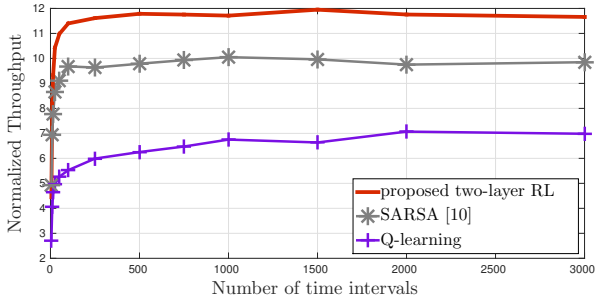

Fig. 3: Average throughput vs. number of time intervals

Fig. 2 shows the average number of transmitted data packets per time interval vs. $E_{\max}/(2\sigma^2)$. As expected, the performance of all approaches increases when $E_{\max}$ increases. The large gain of our proposed approach is due to the consideration of data buffer levels, in addition to the channel conditions, for the power allocation. If only channel conditions are considered, data buffer overflows are not avoided and the achievable throughput is reduced.

The convergence speed of the proposed algorithm is evaluated in Fig. 3 for $E_{\max}/(2\sigma^2) = 10$dB. The figure shows the normalized number of transmitted data packets vs. the number I of time intervals. The number of transmitted data packets is normalized with respect to I. The proposed algorithm converges faster than SARSA and Q-learning and it achieves a better performance. This is because the set $\mathcal{A}$ of each layer is much smaller than for SARSA or Q-learning. The smaller $\mathcal{A}$, the less exploration is required and the faster the learning.

6. CONCLUSIONS

An EH data dissemination scenario with individual data intended for different receivers was investigated. Causal knowledge regarding the EH, channel fading and data arrival processes was assumed. We modelled the power allocation problem as an MDP and we proposed a two-layer RL algorithm to find a power allocation policy that aims at maximizing the throughput. Numerical results show that the proposed algorithm achieves a better performance compared to the standard RL algorithms Q-learning and SARSA.

7. REFERENCES

[1] Sennur Ulukus, Aylin Yener, Elza Erkip, Osvaldo Simeone, Michele Zorzi, Pulkit Grover, and Kaibin Huang, "Energy harvesting wireless communication: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, March 2015.

[2] Hakan Erkal, F Mehmet Ozcelik, and Elif Uysal-Biyikoglu, "Optimal offline broadcast scheduling with an energy harvesting transmitter," *EURASIP J. Wireless Commun. and Networking*, vol. 2013, no. 1, pp. 1–20, July 2013.

[3] Omur Ozel, Jing Yang, and Sennur Ulukus, "Optimal transmission schemes for parallel and fading gaussian broadcast channels with an energy harvesting rechargeable transmitter," *Comput. Comm.*, vol. 36, no. 12, pp. 1360–1372, July 2013.

[4] Mehmet Akif Antepli, Elif Uysal-Biyikoglu, and Hakan Erkal, "Optimal packet scheduling on an energy harvesting broadcast link," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1721–1731, September 2011.

[5] Jing Yang, Omur Ozel, and Sennur Ulukus, "Broadcasting with an energy harvesting rechargeable transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 571–583, February 2012.

[6] Minghan Fu, Ahmed Arafa, Sennur Ulukus, and Wei Chen, "Delay minimal policies in energy harvesting broadcast channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[7] Kaya Tutuncuoglu and Aylin Yener, "Energy harvesting broadcast channel with inefficient energy storage," in *Proc. Asilomar Conf. Signals, Syst. Computers*, Pacific Grove, November 2012, pp. 53–57.

[8] Abdulrahman Baknina and Sennur Ulukus, "Online scheduling for energy harvesting broadcast channels with finite battery," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Barcelona, July 2016, pp. 1–5.

[9] Pol Blasco, Deniz Gündüz, and Mischa Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, April 2013.

[10] Andrea Ortiz, Hussein Al-Shatri, Xiang Li, Tobias Weber, and Anja Klein, "Reinforcement learning for energy harvesting point-to-point communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[11] Yong Xiao, Zhu Han, Dusit Niyato, and Chau Yuen, "Bayesian reinforcement learning for energy harvesting communication systems with uncertainty," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, June 2015, pp. 5398–5403.

[12] Roy Chaoming Hsu, Cheng-Ting Liu, and Wei-Ming Lee, "Reinforcement learning-based dynamic power management for energy harvesting wireless sensor network," in *Next-Generation Applied Intelligence*, Been-Chian Chien, Tzung-Pei Hong, Shyi-Ming Chen, and Moonis Ali, Eds., Berlin, Heidelberg, 2009, pp. 399–408, Springer Berlin Heidelberg.

[13] Andrea Ortiz, Hussein Al-Shatri, Xiang Li, Tobias Weber, and Anja Klein, "A learning based solution for energy harvesting decode-and-forward two-hop communications," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Washington, December 2016, pp. 1–7.

[14] Andrea Ortiz, Hussein Al-Shatri, Xiang Li, Tobias Weber, and Anja, "Reinforcement learning for energy harvesting decode-and-forward two-hop communications," *IEEE Trans. Green Commun. and Networking*, vol. 1, no. 3, pp. 309–319, September 2017.

[15] Pol Blasco and Deniz Gündüz, "Multi-access communications with energy harvesting: A multi-armed bandit model and the optimality of the myopic policy," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 585–597, March 2015.

[16] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 3rd. edition, 2010.

[17] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[18] David Tse and Pramod Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[19] Yaser Barayan and Ivica Kostanic, "Performance evaluation of proportional fairness scheduling in LTE," in *Proc. World Congr. Eng. Comput. Sci. (WCECS)*, San Francisco, Oct. 2013, pp. 1–6.