A. Ortiz, T. Weber and A. Klein, "Resource Allocation in Energy Harvesting Multiple Access Scenarios via Combinatorial Learning," in *Proc. of the IEEE 20th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, July 2019.

Resource Allocation in Energy Harvesting Multiple Access Scenarios via Combinatorial Learning

Andrea Ortiz
Technische Universität Darmstadt
Darmstadt, Germany
a.ortiz@nt.tu-darmstadt.de

Tobias Weber
University of Rostock
Rostock, Germany
tobias.weber@uni-rostock.de

Anja Klein
Technische Universität Darmstadt
Darmstadt, Germany
a.klein@nt.tu-darmstadt.de

***Abstract*— The allocation of K orthogonal resources aiming at maximizing the throughput in an energy harvesting (EH) multiple access scenario is considered. In this setting, the optimal resource allocation (RA) depends on the transmitters' EH and channel fading processes. However, in realistic scenarios, only causal knowledge of these processes is available. We first formulate the offline optimization problem and identify two main challenges, namely, how to exploit causal knowledge to maximize the throughput and how to handle the high dimensionality of the problem. To address these challenges, we propose a novel reinforcement learning (RL) algorithm, termed combinatorial RL (cRL). The name stands for its ability to handle the combinatorial nature of the RA solutions. Exploiting the available causal knowledge, we learn the RA policy aiming at maximizing the throughput. Furthermore, we overcome the curse of dimensionality, typical of combinatorial problems, by splitting the learning task, solving $K+1$ smaller RL problems and using linear function approximation. Through numerical simulations, we show that cRL outperforms known strategies like random and greedy as well as other RL approaches.**

I. Introduction

Energy harvesting (EH) enables wireless communication nodes to collect energy from the environment to recharge their batteries. As a result, the operation of the nodes is not limited by their batteries, but by the hardware's lifetime [1]. However, to efficiently use the harvested energy, two aspects should be considered: a suitable *power allocation* policy at the EH nodes and a suitable *resource allocation* (RA) policy in the network.

Previous work on EH communications, specially in multiple access (MAC) scenarios, has mainly focused on power allocation policies for the EH transmitters. This problem has been tackled following three approaches, i.e., an offline approach in which perfect non-causal knowledge regarding the EH and the channel fading processes is assumed [2], [3], an online approach in which only statistical knowledge of the processes is assumed [4], [5], and a learning approach in which only causal knowledge is assumed [6]–[10]. In the following, we summarize the state of the art of power allocation in EH MAC. Using an offline approach, an EH two-user MAC channel is considered in [2] where a generalized iterative backward water-filling algorithm is proposed to minimize the time required for data transmission. In [3], an iterative water-filling based algorithm is proposed to find the optimal power allocation policy in the EH multi-user MAC channel. The authors of [4] follow an online approach to study a continuous-time power policy for EH MAC. In [5], an EH MAC channel using time division multiple access (TDMA) is considered and the authors investigate the optimal power allocation policy assuming only statistical knowledge. Furthermore, learning approaches have been used to address the power allocation problem in EH point-to-point [6]–[8], two-hop [9] and broadcast scenarios [10], but not yet in MAC. Only few works consider RA in EH MAC. In [11], an online approach is considered to schedule the transmissions according to the transmitters' battery and channel states. Additionally, in [12], the authors model the EH processes using independent two-state Markov chains and formulate the RA problem as a restless multi-armed bandit (MAB) problem.

In this paper, we focus on the allocation of K orthogonal resources in a MAC scenario. In contrast to [11], we consider only causal knowledge, i.e., in a given time interval, only the current amounts of harvested energy and the current channel coefficients are assumed to be known. Moreover, we extend the model in [12], where the nodes are assumed to harvest one energy unit or none, to consider that the harvested energy can take any positive value. We first formulate the offline optimization problem for the MAC scenario and identify two main challenges, namely, finding a RA solution aiming at maximizing the throughput having only causal knowledge of the EH and channel fading processes, and handling the high dimensionality of the problem. The former comes from the consideration of a realistic scenario in which no knowledge about the future is assumed, while the latter comes from the combinatorial nature of the RA solutions and the infinite number of battery and channel states the EH transmitters can experience. To address these challenges, we formulate the RA problem as a reinforcement learning (RL) problem. Specifically, we propose a novel RL algorithm termed combinatorial RL (cRL). The name of our algorithm stands for its ability to handle the combinatorial nature of the RA solutions. cRL is inspired by the so called naive strategy proposed in [13] for MAB. Here, we extend it to a RL setting and combine it with linear function approximation to manage the infinite number of states. The strength of cRL is its ability to split the original RL problem into $K+1$ smaller problems, thus tackling the curse of dimensionality of combinatorial problems. This increases the learning rate and, consequently, the throughput, compared

This work was supported by the German Research Foundation (DFG) within the Collaborative Research Center (CRC) 1053 - MAKI.

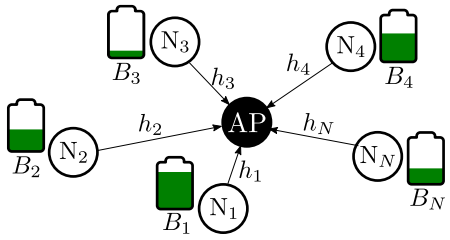

Fig. 1. MAC scenario with EH transmitters.

to traditional RL approaches.

The remainder of the paper is organized as follows. In Sec. II, the system model is presented. The RA problem is formulated in Sec. III and the proposed cRL is explained in Sec. IV. Performance results are presented in Sec. V and Sec. VI concludes the paper.

II. System Model

A MAC scenario consisting of a single access point (AP) and N EH transmitters, as depicted in Fig. 1, is considered. The AP is assumed to be connected to the electrical grid while the EH transmitters, termed N_n with $n = 1, ..., N$, harvest energy and use it to transmit data to the AP. We assume the EH transmitters always have data available for transmission. As a result, the achievable throughput is only limited by the harvested energy and the RA policy.

Time is divided into time slots (TSs) of constant duration τ, and each TS is identified by its index $i = 1, ..., I$, where I is the total number of TSs. Furthermore, within each TS, K orthogonal and distinguishable resources are available for the transmission of data, e.g., a fraction of a TS if TDMA is considered or one sub-carrier in the case of frequency-division multiple access (FDMA). The AP has the task of allocating the K resources aiming at maximizing the throughput.

At the beginning of TS i, an amount of energy $E_{n,i} \in \mathbb{R}^+$, obtained through the EH process, is assumed to be available at N_n. The maximum amount of harvested energy, termed $E_{\max,n}$, depends on the energy source being used. The harvested energy is stored in a rechargeable battery with maximum capacity $B_{\max,n}$. Moreover, in order to transmit the signaling from each N_n to the AP, i.e., battery level and channel coefficient, a constant amount of energy E_n^{sig} is assumed to be spent in each TS. Furthermore, the battery level $B_{n,i} \in \mathbb{R}^+$ is updated at the beginning of each TS as

$$B_{n,i+1} = \min\left(B_{\max,n},\ B_{n,i} + E_{n,i} - \tau \sum_{k=1}^{K} p_{n,i,k} - E_n^{\mathrm{sig}}\right), \tag{1}$$

where $p_{n,i,k}$ is the transmit power used by N_n in TS i over the k^{th} resource. Our goal is to find a resource allocation policy at the AP considering the available causal knowledge.

The transmitters are assumed to be low-power devices with limited processing capabilities. Therefore, a low-complexity greedy power allocation policy is considered, i.e., N_n uses all the energy in its battery for the transmission of data every time that a resource has been allocated to it. In case more than one resource is allocated to N_n in TS i, equal power allocation is considered. Let $\delta_{n,i,k} \in \{0, 1\}$ be a variable that indicates if the k^{th} resource has been allocated to N_n in TS i. The transmit power used by N_n in TS i is calculated as

$$p_{n,i,k} = \begin{cases} \frac{B_{n,i}}{\tau \sum_{k=1}^{K} \delta_{n,i,k}} & \text{if } \sum_{k=1}^{K} \delta_{n,i,k} \geq 1 \\ 0 & \text{else.} \end{cases} \tag{2}$$

The fading channel from each N_n to the AP over the k^{th} resource is described by the channel coefficient $h_{n,i,k} \in \mathbb{C}$ which is assumed to remain constant for one TS. The noise at the AP is independent and identically distributed (i.i.d.) zero mean additive white Gaussian noise with variance σ^2. Furthermore, the throughput

$$R_i = \sum_{n=1}^{N} \sum_{k=1}^{K} \log_2\left(1 + \frac{|h_{n,i,k}|^2 p_{n,i,k}}{\sigma^2}\right) \tag{3}$$

in bits is the amount of data received by the AP in TS i. As only causal knowledge is available, the AP only knows the current battery levels $B_{n,i}$ and the channel coefficients $h_{n,i,k}$ in TS i. Note however that the battery level $B_{n,i}$ comprises the previous amounts of harvested energy, battery levels and amounts of energy used for signaling and data transmission.

III. Problem Formulation

In this section, we formulate the offline optimization problem for the EH MAC scenario. Feasible RA solutions depend on the transmitters EH and channel fading processes. Naturally, only the energy in the batteries can be used for data transmission. Consequently, the energy causality constraint

$$\sum_{i=1}^{M} \sum_{k=1}^{K} \tau p_{n,i,k} + \sum_{i=1}^{M} E_n^{\mathrm{sig}} \leq \sum_{i=1}^{M-1} E_{n,i},\ M = 1, ..., I, \tag{4}$$

has to be fulfilled. Moreover, to avoid overflow situations in which part of the harvested energy is wasted because the battery is full, the constraint

$$\sum_{i=1}^{M} E_{n,i} - \sum_{i=1}^{M} \sum_{k=1}^{K} \tau p_{n,i,k} - \sum_{i=1}^{M} E_n^{\mathrm{sig}} \leq B_{\max,n},\ \forall n, M, \tag{5}$$

is also taken into account. Exclusive allocation is considered, i.e., each resource can be allocated to only one user but multiple resources can be allocated to a single node. As a result, the constraints

$$\sum_{n=1}^{N} \delta_{n,i,k} = 1, \tag{6}$$

$$\sum_{k=1}^{K} \sum_{n=1}^{N} \delta_{n,i,k} = K, \tag{7}$$

must be fullfiled by any feasible RA solution. The offline optimization problem for RA in EH MAC is given by

$$\left(\delta_{n,i,k}^{\mathrm{opt}}\right)_{n,i,k} = \underset{\delta_{n,i,k} \in \{0,1\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_i \tag{8a}$$

$$\text{subject to} \quad (2), (4), (5), (6), (7). \tag{8b}$$

We identify the problem in (8) as a non-linear knapsack problem which is NP-hard. Furthermore, the constraints in (4) and (5) impose a dependency of the RA solution over time. This means, causal knowledge is not sufficient to obtain the optimum solution. Moreover, the dimension of the problem in (8) grows exponentially with K and N. Specifically, in TS i, the number $|\mathcal{A}|$ of feasible RA solutions is bounded by $|\mathcal{A}| = N^K$. To overcome these challenges, we propose cRL, a combinatorial RL algorithm which uses past experience to learn the optimal RA policy and handles the large dimensionality of the problem by separating it into $K+1$ smaller ones.

IV. cRL: Combinatorial Reinforcement Learning

Our algorithm is motivated by the availability of only causal knowledge regarding the EH and channel fading processes. In the following, we first model the problem in (8) as a Markov decision process (MDP). Next, we present the naive strategy proposed in [13] for MAB and extend it to RL problems. We then continue with the application of linear function approximation and explain the action selection strategies.

A. Markov decision process

In our scenario, the TS duration τ is fixed and known. Moreover, the transmitters adopt a greedy power allocation. Consequently, in TS i the RA depends solely on the values of $B_{n,i}$, and $h_{n,i}$. As the previous battery and channel states do not need to be taken into account, the system under consideration fulfils the Markov property and can be modeled as an MDP. This formulation is helpful for the definition of the RL algorithm as will become clear in the following. An MDP is defined by a set $\mathcal{S}$ of states, a set $\mathcal{A}$ of actions, a transition model P and a set $\mathcal{R}$ of rewards [14]. The proposed cRL provides a solution of the MDP presented here. In TS i, the state $S_i \in \mathcal{S}$ corresponds to the battery and channel states of all the transmitters. However, to reduce the number of variables to be considered, we define a pseudo-SNR $\rho_{n,i} = |h_{n,i}|^2 B_{n,i}/\tau$. The higher $\rho_{n,i}$, the more suitable is N_n for the transmission of data in TS i. This is because N_n experiences a good channel, has a large amount of energy stored in its battery, or both. We remark that $\rho_{n,i}$ can take any value in a continuous range. As a result, the set $\mathcal{S}$ contains infinitely many possible states. The set $\mathcal{A}$ contains the RA solutions and in our model, $\mathcal{A}$ is finite but grows exponentially. The transition model P defines the probability of going from state S_i to S_{i+1} after selecting a_i. Finally, the rewards $R_i \in \mathcal{R}$ indicate how beneficial it is to select a_i in S_i and it is given by the throughput in (3).

A policy π, which maps states to actions as $a_i = \pi(S_i)$, provides the solution of an MDP. Furthermore, the policy is evaluated using the so-called action value function $Q^\pi(S_i, a_i)$ which is the expected reward starting in S_i, selecting a_i and following π thereafter [14]. The optimal policy π^* has an action value function Q^* which is greater than or equal to the action value function of any other policy for all $S_i \in \mathcal{S}$ and $a_i \in \mathcal{A}$. Knowing Q^* is important because it leads to the determination of π^*. For each S_i, any a_i that maximizes $Q(S_i, a_i)$ is an optimal action.

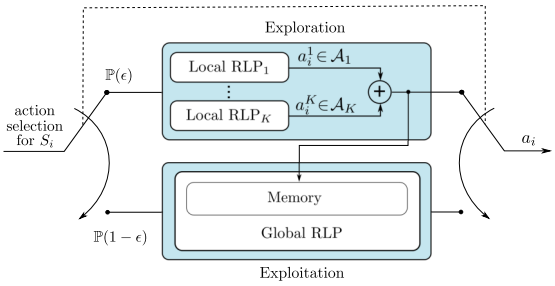

Fig. 2. Schematic of the application of the naive strategy to RL problems.

Moreover, we consider a discount factor $0 \leq \gamma \leq 1$ in order to take into account the preference between achieving a higher throughput in the current TS or achieving it the following TS. Our aim is now to maximize the discount throughput given by $R = \lim_{I\to\infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_i\right]$.

B. Naive Strategy for RL

To tackle the high dimensionality of the problem, we propose the use of the so called naive strategy for MAB [13]. This strategy is based on the idea that the reward distribution can be approximated by the sum of a set of reward functions that depend on only one variable. Here, we extend this idea to the more complex case of RL problems.

In our setting, we can rewrite the reward function in (3) as the sum of the throughput obtained in each of the resources, this means $R_i = \sum_{k=1}^{K} R_i^k$ where

$$R_i^k = \sum_{n=1}^{N} \log_2\left(1 + \frac{|h_{n,i,k}|^2 p_{n,i,k}}{\sigma^2}\right). \tag{9}$$

This decomposition allows us to separate the problem into $K+1$ smaller ones as shown in Fig. 2. K of these problems are termed *local* RL problems (RLP) while the remaining one is termed *global* RLP. As shown in Fig. 2, the dotted line represents that in TS i the action a_i can be selected using the local RLPs or the global RLP. Intuitively, the task of the local RLPs is to efficiently explore the RA solutions while the task of the global RLP is to select, for a given state, the action which is considered the best up to TS i. The action space of the global RLP is initially empty, and it is updated every time that a new RA solution is tried via the local RLPs.

Each local RLP is associated with one resource and its task is to learn how to select one transmitter to which said resource will be allocated. This is motivated by the idea that by learning to maximize each R_i^k, the total R_i is also maximized. Note that the decision for each resource is done simultaneously and independently in each local RLP. As a result, the action set $\mathcal{A}^k$ of the k^{th} local RLP is composed solely by the set of EH transmitters, thus tackling the curse of dimensionality in the original formulation, i.e., $|\mathcal{A}^k| = N$. The collection of the a_i^k selected by each local RLP forms the RA solution a_i.

As mentioned above, when a new RA solution is encountered by the local RLPs, it is stored in the global RLP. This means that when a_i is selected via the global RLP, the action

considered to be the best up to TS i is selected. Therefore, the global RLP does not solve a combinatorial problem, but learns the suitability of the RA solutions that have been tried.

C. Linear Function Approximation

By means of the naive strategy, we are able to deal with the high dimensionality of the action space. However, nothing has yet been done to handle the infinite number of states. For this purpose, we use linear function approximation. The infinite number of states comes from the fact that $B_{n,i}$ and $h_{n,i,k}$ can take any positive value. Furthermore, when $|\mathcal{S}|$ is infinite, the action value function Q^π has also an infinite number of values. In such cases, linear function approximation can be used to represent Q^π as a weighted sum of feature functions [14]. Each feature function maps S_i and a_i onto a feature value. Let $\mathbf{f}$ be a vector formed by all the feature values and let $\mathbf{w}$ be a vector of weights containing the contribution of each feature. Q^π is then approximated as $\mathrm{Q}^\pi(S_i, a_i) \approx \mathbf{f}^{\mathrm{T}}(S_i)\mathbf{w}$ [14]. In this paper, we use tile coding as approximation technique due to its flexibility, computational efficiency and suitaibility for multi-dimensional continuous spaces [14].

D. Action Selection

In every state S_i the action a_i can be selected using the local RLPs or the global RLP. For this purpose, the ϵ-greedy policy is considered. This means, with probability ϵ the local RLPs are used to select the RA solution and with a probability $1-\epsilon$ we make use of the global RLP. The local RLPs learn how to allocate the corresponding resources. As a consequence, in each RLP we face the well known exploration-exploitation dilemma, i.e., whether the corresponding resource is allocated to a transmitter that has not yet used it and can potentially achieve a high throughput or to allocate it to the transmitter that has achieved the highest throughput so far. To handle this tradeoff, we also consider the ϵ-greedy policy at the local RLPs. However, to differentiate it from the previous case, we termed it ϵ_l-greedy policy. In contrast to the local RLPs, no exploration is performed at the global RLP. This is because its task is to learn the suitability of the RA solutions that have been already discovered by the local RLPs. Consequently, a greedy policy is considered. This means, every time a_i is selected via the global RLP, the RA that has achieved the highest throughput, i.e., highest $\hat{Q}(S_i, a_i)$, is selected. The use of the greedy policy enforces the exploitation of the RA solution which is considered the best up to TS i.

E. Combinatorial RL algorithm

Our proposed cRL is composed of $K+1$ RLPs. For each of them, the state-action-reward-state-action (SARSA) algorithm is considered. When linear function approximation is used, the weights $\mathbf{w}$ are adjusted in the direction that reduces the error between Q^π and $\hat{\mathrm{Q}}^\pi$ following the gradient descent approach. The updating rule for the local RLPs is given by [14]

$$\Delta\mathbf{w}^k = \alpha_i\big[R_i^k + \gamma\hat{\mathrm{Q}}^\pi(S_{i+1}, a_{i+1}^k, \mathbf{w}^k) - \hat{\mathrm{Q}}^\pi(S_i, a_i^k, \mathbf{w}^k)\big]\mathbf{f}, \tag{10}$$

where α_i is the learning rate. Similarly, the weights in the global RLP are updated as

$$\Delta\mathbf{w} = \alpha_i\big[R_i + \gamma\hat{\mathrm{Q}}^\pi(S_{i+1}, a_{i+1}, \mathbf{w}) - \hat{\mathrm{Q}}^\pi(S_i, a_i, \mathbf{w})\big]\mathbf{f}. \tag{11}$$

The proposed cRL is summarized in Algorithm 1.

Algorithm 1 Combinatorial RL algorithm

```
initialize parameters, observe S_i and select a random action a_i
for every i = 1, ..., I do
    if in state S_i a new a_i is encountered then
        add it to the global RLP
    end if
    calculate the achieved throughput                          ▷ Eq. (9)
    observe next state S_{i+1} and generate random number z
    if z ≥ ε(i) then                                ▷ Exploit from global RLP
        select next action a_{i+1} with highest Q(S_{i+1}, a_{i+1})
    else                                            ▷ Explore from local RLPs
        for each local RLP do
            select action a_i^k using ε_l-greedy
        end for
    end if
    update the weights in the local RLPs                       ▷ Eq. (10)
    update the weights in the global RLP                       ▷ Eq. (11)
    set S_i = S_{i+1} and a_i = a_{i+1}
end for
```

V. Simulation Results

In this section, numerical results for the evaluation of the proposed cRL are presented. The results are obtained by generating $T = 100$ independent random EH and channel realizations. Each realization is an episode where the transmitters harvest energy $I = 10^4$ times. We consider TDMA, i.e., each resource is a fraction of the TS and all the fractions have the same length. For each N_n, the amounts of harvested energy are taken from a uniform distribution with maximum value $E_{\max}$, where $E_{\max}/(2\tau\sigma^2) = 5\mathrm{dB}$. The time interval τ between two consecutive EH time instants is set to one time unit and the channel coefficients $h_{n,i,k}$ are taken from an i.i.d. Rayleigh fading process with zero mean, unit variance and a path loss exponent of three. Additionally, the noise variance is set to $\sigma^2 = 1$. To perform linear function approximation, each of the N dimensions forming the state space is divided into two tiles and $G = 16$ grids are considered. The learning rate is set to $\alpha = (10G)^{-1}$ for the local and global RLPs, the ϵ and ϵ_l parameters are decreased in each TS and $\gamma = 0.9$. For comparison, we consider three approaches: traditional RL in which a single RL problem using SARSA and linear function approximation is considered, a random strategy where K transmitters are randomly selected and one resource is allocated to each of them, and the greedy strategy where the K transmitters with the stronger channel conditions are selected and one resource is allocated to each of them.

Fig. 3 shows the throughput for different numbers of EH transmitters when $K = 3$ resources are considered. For all the approaches, the throughput increases with the number of transmitters due to the increased diversity, i.e., when more transmitters are considered, there are more possible RA solutions. For $N = 2$, cRL performs similar to the traditional RL and outperforms the random and greedy approaches. However, as the network size increases, the advantages of cRL are better exhibited. By breaking the original RL into $K + 1$ smaller

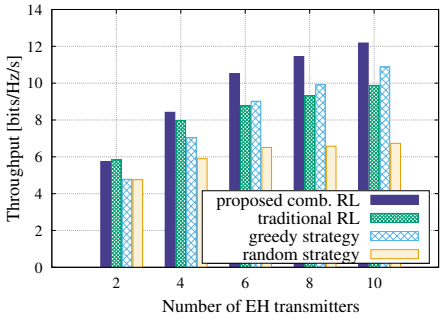

Fig. 3. Throughput vs. number of EH transmitters.

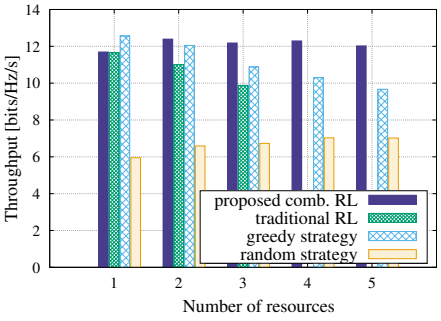

Fig. 4. Throughput vs. number of resources.

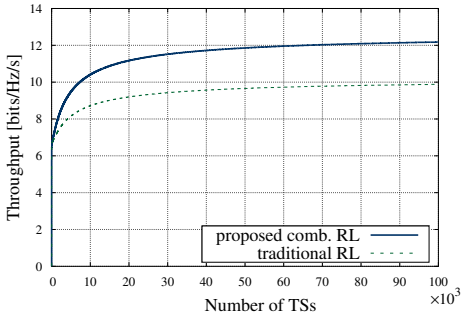

Fig. 5. Throughput vs. number of TSs.

RL problems, cRL is able to handle the larger action spaces and consequently, achieve a higher throughput compared to traditional RL. For $N = 10$, cRL achieves a throughput 23% higher than traditional RL and 12% and 80% higher than the greedy and random strategies, respectively.

The effect of the number of resources on the throughput for $N = 10$ is shown in Fig. 4. cRL achieves on average the same throughput for the different number of resources because it considers the EH and channel fading process of the transmitters, which are the source of the randomness in the system, in the selection of the RA solutions. As the traditional RL approach suffers from the curse of dimensionality, its performance degrades when more resources are considered. Moreover, when the number of resources is larger than three, the action space of the traditional RL approach is so large that a solution cannot be obtained. The greedy strategy performs slightly better than the learning approaches when $K = 1$, because in this case acting greedy is optimal while the learning approaches need to perform exploration in order to learn the RA policy. During exploration, suboptimal RA solutions may be selected which affects the average throughput. However, as the number of available resources increases the performance of the low-complexity approaches decreases. cRL achieves 24% and 71% higher throughput than the greedy and random strategies, respectively, when $N = 10$ and $K = 5$.

The convergence speed of cRL is evaluated in Fig. 5 when $N = 10$ and $K = 3$. From the beginning, cRL achieves a higher throughput compared to the traditional RL. The reason for this is that it explores more efficiently the action space. Additionally, it is designed to cope with the high dimensionality of the problem in both, the state and action space, while the traditional RL only considers the high dimensionality of the state space through the use of linear function approximation.

VI. Conclusions

A MAC scenario with a single AP and multiple EH transmitters was considered. In addition, K orthogonal resources were assumed to be available for the transmission of data and the RA problem was investigated. Our goal was to find a RA policy when only causal knowledge regarding the EH and the channel fading processes is available. To this aim, we formulated the offline throughput maximization problem for this scenario to identify the main challenges to be addressed. As a result, we proposed cRL which exploits the available causal knowledge and tackles the curse of dimensionality by solving $K + 1$ smaller RL problems and leveraging the use of linear function approximation. Through numerical simulations, we showed that cRL outperforms low-complexity strategies like random and greedy as well as traditional learning approaches.

References

[1] S. Ulukus, A. Yener *et al.*, "Energy harvesting wireless communication: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, March 2015.

[2] J. Yang and S. Ulukus, "Optimal packet scheduling in a multiple access channel with energy harvesting transmitters," *J. Commun. and Networks*, vol. 14, no. 2, pp. 140–150, April 2012.

[3] Z. Wang, V. Aggarwal, and X. Wang, "Iterative dynamic water-filling for fading multiple-access channel with energy harvesting," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 382–395, March 2015.

[4] M. B. Khuzani and P. Mitran, "On online energy harvesting in multiple access communication systems," *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1883–1898, January 2014.

[5] J. Liu, H. Dai, and W. Chen, "On throughput maximization of time division multiple access with energy harvesting users," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2457–2470, April 2016.

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

[7] A. Ortiz, H. Al-Shatri *et al.*, "Reinforcement learning for energy harvesting point-to-point communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[8] K. Sugiyama, H. Iirnori, and G. T. F. D. Abreu, "Statistical analysis of EH battery state under noisy energy arrivals," in *Proc. Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, June 2018, pp. 1–5.

[9] A. Ortiz, H. Al-Shatri *et al.*, "Reinforcement learning for energy harvesting decode-and-forward two-hop communications," *IEEE Trans. Green Commun. and Networking*, vol. 1, no. 3, pp. 309–319, Sept. 2017.

[10] A. Ortiz, T. Weber, and A. Klein, "A two-layer reinforcement learning solution for energy harvesting data dissemination scenarios," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Apr. 2018, pp. 6648–6652.

[11] J. Yang and J. Wu, "Online throughput maximization in an energy harvesting multiple access channel with fading," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Hong Kong, June 2015, pp. 2727–2731.

[12] P. Blasco and D. 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.

[13] S. Ontañón, "The combinatorial multi-armed bandit problem and its application to real-time strategy games," in *Proc. AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment AIIDE* , Melbourne, 2013, pp. 2471–2478.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.