A. Ortiz, H. Al-Shatri, X. Li, T. Weber and A. Klein, ”Reinforcement Learning for Energy Harvesting Point-to-Point Communications,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016.

Reinforcement Learning for Energy Harvesting Point-to-Point Communications

Andrea Ortiz*, Hussein Al-Shatri*, Xiang Li†, Tobias Weber† and 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
Email: {a.ortiz, h.shatri, a.klein}@nt.tu-darmstadt.de, {xiang.li, tobias.weber}@uni-rostock.de

***Abstract*—Energy harvesting point-to-point communications are considered. The transmitter harvests energy from the environment and stores it in a finite battery. It is assumed that the transmitter has always data to transmit and the harvested energy is used exclusively for data transmission. As in practical scenarios prior knowledge about the energy harvesting process might not be available, we assume that at each time instant only information about the current state of the transmitter is available, i.e., harvested energy, battery level and channel coefficient. We model the scenario as a Markov decision process and we implement reinforcement learning at the transmitter to find a power allocation policy that aims at maximizing the throughput. To overcome the limitations of traditional reinforcement learning algorithms, we apply the concept of function approximation and we propose a set of binary functions to approximate the expected throughput given the state of the transmitter. Numerical results show that the performance of the proposed approach, which requires only causal knowledge of the energy harvesting process and channel coefficients, has only a small degradation compared to the optimum case which requires perfect non-causal knowledge. Additionally, the proposed approach outperforms naïve policies that assume only causal knowledge at the transmitter.**

I. Introduction

Having wireless communication nodes with energy harvesting (EH) capabilities holds the promise of self-sustainability and perpetual operation [1]. The idea behind EH is that the communication nodes can recharge their batteries in an environmentally friendly way using natural energy sources, e.g., solar, thermal, vibrational, chemical, etc. and afterwards use the harvested energy for transmitting data [2]. In addition to the channel fluctuations existing in any wireless communication system, the variable availability of energy inherent to EH communication systems has to be considered. The exact amount of available energy and the precise time when it can be harvested is hard to predict, resulting in new challenges in the design of transmission strategies.

Recent effort has been focused on EH communication systems when non-causal knowledge of the EH process is assumed [3]–[6]. This approach, termed offline, assumes that the energy arrival times and the amounts of harvested energy are completely known at the beginning of the communication. Although this assumption cannot be perfectly fulfilled in reality, it allows the calculation of upper bounds of the performance. In [3], the problem of throughput maximization within a deadline in a point-to-point scenario is considered. Additionally, it is shown that this problem is equivalent to the minimization of the completion time for the transmission of a fixed amount of data. Similarly, in [4], a point-to-point communication scenario with a fading channel is assumed and the corresponding problem of offline throughput maximization within a deadline is addressed. The processing cost at the transmitter in a point-to-point scenario is analyzed in [5] and the effect of inefficient energy storage is studied in [6].

More realistic approaches, termed online, assume only statistical information about the EH process. In [4] and [7]–[9], the point-to-point scenario is investigated. A fading channel is assumed in [4] and the problem of online scheduling for throughput maximization within a deadline is considered. The problem is solved using continuous time stochastic dynamic programming with statistical and causal knowledge of the energy and fading variations. In [7], an on-off mechanism at the transmitter is studied in which for each packet arrival a binary decision of whether to transmit or drop the packet is made. Additionally, the energy arrival is described as a continuous time Markov chain and the statistical distribution of the importance of the messages is assumed to be known. A save-then-transmit protocol that minimizes the system outage probability is proposed in [8]. There, a fixed amount of data is to be transmitted during the duration of a time interval. The energy arrival is modeled as a random variable for each time interval.

All the aforementioned approaches require knowledge of the statistics of the EH process. However, in practical scenarios this knowledge might not be available. Consider, for example, an EH transmitter which collects energy from different sources simultaneously and assume that each source can be switched on or off at random times. In this scenario, the EH process cannot be considered as stationary and consequently, keeping track of its statistics becomes challenging. Another example, in which knowledge of the EH process cannot be obtained, is when information about the exact location where the EH transmitter will operate is unavailable. To overcome these problems, a learning theoretic approach for EH is adopted in [9]. Specifically, the well-known reinforcement learning (RL) algorithm Q-learning is used to maximize the throughput within a deadline. The authors assume that the amount of harvested energy, the channel coefficients and the transmit power in each time instant are taken from a finite discrete set. Moreover, they assume the data arrives in packets and for each data packet the decision of transmit or drop has to be

made. Although this approach requires only causal knowledge of the EH process and the channel fading at the transmitter, its performance is limited by the number of values considered in the discrete sets defined for the harvested energy and the channel coefficients. As stated in [10] and [11], when the size of the sets increases, the number of states in which the transmitter can be also increases and the probability of learning about each of these states is reduced. In other words, the larger the discrete set, the slower the Q-learning algorithm learns the power allocation policy.

In this paper, we consider an EH point-to-point communication scenario in which the transmitter is equipped with a finite battery and no knowledge about the EH process is available. In contrast to [9], where the amount of harvested energy and the channel coefficients are taken from a finite discrete set, we study the more realistic scenario in which the harvested energy, the battery level and the channel coefficients can take any real positive value. As a consequence, in our model the transmitter can be in an infinite number of states. To overcome the limitations of traditional Q-learning, we apply the concept of linear function approximation in RL in order to find a power allocation policy that aims at maximizing the throughput. To achieve this, we propose a set of binary functions to approximate the expected throughput given the state of the transmitter. The proposed RL algorithm is applied at the transmitter and it is able to learn the power allocation policy with only causal knowledge about the EH process and the channel coefficients.

The rest of the paper is organized as follows. In Section II, the system model is presented. The EH power allocation problem is modeled as a Markov decision process in Section III. In Section IV, the RL algorithm used for the continuous valued EH point to point communication scenario is explained. Numerical performance results are presented in Section V and Section VI concludes the paper.

II. System Model

In this paper, a point-to-point communication scenario consisting of two single-antenna nodes is considered. As depicted in Fig. 1, transmitter node N_1 harvests energy from the environment and uses it for transmitting data to receiver node N_2. It is assumed that N_1 has always data available for transmission. As a result, the achievable throughput is only limited by the availability of harvested energy.

As in [3]–[5], it is assumed that the energy is harvested in fixed time instants t_i, where $i = 1, 2, ..., I$ is the index of the EH time instants and I is the total number of EH time instants. This means that at t_i an amount of energy $E_i \in \mathbb{R}^+$ is received by N_1. The maximum amount of energy that can be harvested, termed $E_{\max}$, depends on the energy source that is used. After E_i is harvested, it is stored in a rechargeable finite battery with maximum capacity $B_{\max}$. The battery is assumed to be ideal. Therefore, no energy is lost while storing or retrieving energy from it. As the battery cannot be recharged instantaneously, it is assumed that at t_i the battery only stores the energy which has been harvested until t_{i-1}. Furthermore,

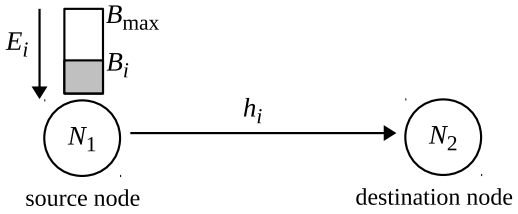

Fig. 1: Point-to-point communication scenario with an EH transmitter node.

it is assumed that at t_1, the node has not yet harvested any energy and the battery is empty. The time interval $\tau_i = t_{i+1} - t_i$ between two consecutive EH time instants t_i and t_{i+1} is assumed to be constant such that $\tau_i = \tau$, $i = 1, 2, ..., I$.

The noise at N_2 is assumed to be independent and identically distributed (i.i.d.) zero mean additive white Gaussian noise (AWGN) with variance σ^2. Additionally, the transmit power p_i is kept constant during each time interval τ [3]. It is assumed that the harvested energy is used solely for the transmission of data to N_2.

In our scenario, only causal information is available at N_1. This means that at t_i, N_1 has knowledge about the current state of the battery $B_i \in \mathbb{R}^+$, the harvested energy E_i, the fading channel coefficient $h_i \in \mathbb{C}$ and the past states. According to the state of N_1 at t_i, it selects p_i and transmits data to N_2. The throughput achieved in one time interval τ is given by

$$R_i = \tau \log_2 \left(1 + \frac{|h_i|^2 p_i}{\sigma^2}\right). \tag{1}$$

As mentioned before, the transmit power can be allocated only after the harvested energy has been stored in the battery. Therefore, the causality condition,

$$\tau p_i \leq B_i \ \ \forall i = 1, ..., I, \tag{2}$$

must be fulfilled by any feasible power allocation solution. Additionally, overflow situations in which part of the harvested energy is wasted because the battery is full, must be avoided. A battery overflow is a suboptimal solution because a higher throughput can always be achieved if a higher p_i is selected. Consequently, the overflow constraint,

$$B_i - \tau p_i + E_i \leq B_{\max}, \tag{3}$$

must also be considered.

III. Markov Decision Process Model for EH

In this section, the EH point-to-point communication scenario is modeled as a Markov decision process (MDP) because it provides a suitable mathematical framework for modeling decision-making situations [9]. The proposed RL algorithm of Section IV provides a solution of the MDP presented here.

As mentioned above, at t_i, N_1 has only causal knowledge about its state. Consequently, since τ is fixed and known, the selection of p_i depends solely on the values of B_i, E_i and h_i. Since the selection of p_i does not depend on the state of the system in previous time instants, the system under

consideration fulfills the Markov property and can be modeled as an MDP [10], [11]. An MDP consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition model $\mathcal{P}$ and a set of rewards $\mathcal{R}$ [11]. At t_i, the corresponding state $S_i \in \mathcal{S}$ is a function of B_i, E_i and h_i. In our model, the battery level, the harvested energy and the channel coefficients can take any value in a continuous range. As a result, the set $\mathcal{S}$ contains an infinite number of possible states given by all the combinations of B_i, E_i and h_i. The set of actions $\mathcal{A}$ corresponds to the values of transmit power that can be selected. In our model, $\mathcal{A}$ is finite and it is given by $\mathcal{A} = \{p_i,\ p_i \in 0 : \delta : B_{\max}\}$, where δ is the step size. The action dependent transition model defines the transition probabilities as $\mathbb{P}\left[S_{i+1} \in \mathcal{U} | S_i, p_i\right]$, where $\mathcal{U}$ is a measurable subset of $\mathcal{S}$ [12]. Finally, the rewards indicate how beneficial the selected p_i is for the corresponding S_i. For each S_i and p_i, we define the reward $R_i \in \mathcal{R}$ as the throughput achieved in the interval τ, which is given by (1). R_i can be calculated at N_1 because it knows h_i and the selected p_i.

Since N_i only has information of its state at t_i, the amount of energy to be harvested in future time instants is unknown. Therefore, it is preferred to achieve a higher throughput in the current t_i over future ones. To take into account this preference, let us define $0 \leq \gamma \leq 1$ as the discount factor of future rewards. Our goal is to select $p_i, \forall i$, in order to maximize the expected throughput which is given by

$$R = \lim_{I \to \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_i\right]. \tag{4}$$

A policy π is a mapping from a given S_i to the p_i that should be selected, i.e. $p_i = \pi(S_i)$, and it corresponds to the solution of an MDP [11]. To measure how good a policy π is from S_i onwards, let us define the so-called value functions. These functions can depend solely on the states, called state-value functions or on the states-actions pairs, called action-value functions [10]. The state-value function V^{π} is the expected reward given that N_1 follows the policy π from state S_i onwards. Similarly, the action-value function Q^{π} is defined as the expected reward starting from state S_i, selecting p_i and following π thereafter [10]. As it would become clear later, the action-value functions play an important role in the RL framework. Following the formulation in [10] it is written as

$$\mathrm{Q}^{\pi}(S_i, p_i) = \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k R_{i+k+1} \middle| S_i, p_i\right\}. \tag{5}$$

The optimal policy π^* is the policy whose state-value function is greater than or equal to any other policy for every state. The corresponding action-value function for the optimal policy π^* is denoted by Q^*. Determining the optimal actions becomes easier when Q^* is known because for each state S_i, any action p_i that maximizes $\mathrm{Q}^*(S_i, p_i)$ is an optimal action. Consequently, any policy formed by the collection of optimal actions is an optimal policy π^*. A fundamental property of the value functions is that they can be written in a recursive manner in what is known as the Bellman equations [10]. This recursive representation facilitates the design of RL algorithms. The general form of the Bellman optimality equation for the action-value function is given in [10] as

$$\mathrm{Q}^*(S_i, p_i) = \sum_{S_k \in \mathcal{S}} P_{S_i, S_k}^{p_i} \left[R_i + \gamma \max_{p_k \in \mathcal{A}} \mathrm{Q}^*(S_k, p_k)\right]. \tag{6}$$

IV. RL for EH Communications

In this section, the concept of linear function approximation [10] is applied in RL to find a power allocation policy that aims at maximizing the throughput in the EH point-to-point scenario. Specifically, we consider an on-policy temporal-difference RL algorithm, termed SARSA, which is based on the estimation of $\mathrm{Q}^{\pi}(S_i, p_i)$ [10]. To handle the infinite number of states, we propose a set of binary functions to approximate $\mathrm{Q}^{\pi}(S_i, p_i)$. For the implementation of the algorithm, the following steps are considered. Firstly, the estimation and update of $\mathrm{Q}^{\pi}(S_i, p_i)$ is presented. Secondly, the ϵ-greedy policy to select p_i, $\forall i$ according to the estimated $\mathrm{Q}^{\pi}(S_i, p_i)$ is defined. Thirdly, the concept of linear function approximation is applied. Fourthly, the set of proposed binary functions are linearly combined to approximate $\mathrm{Q}^{\pi}(S_i, p_i)$ and at last, the resulting algorithm, termed approximated SARSA, is presented.

A. Action-value function update

In this paper, we use the SARSA algorithm due to its favorable convergence properties when linear function approximation is used [10], [13]. In SARSA, given a policy π, $\mathrm{Q}^{\pi}(S_i, p_i)$ is estimated considering the transitions from a state-action pair (S_i, p_i) to another state-action pair (S_{i+1}, p_{i+1}) while obtaining reward R_i. This fact explains the name of the algorithm: State-Action-Reward-State-Action [10]. In other words, when N_1 is in state S_i, it selects p_i following policy π. Afterwards, it obtains a reward R_i and moves to state S_{i+1}. According to the current values of $\mathrm{Q}^{\pi}(S_i, p_i)$ and the policy π, the algorithm selects the next p_{i+1}. At this point, $\mathrm{Q}^{\pi}(S_i, p_i)$ is updated using the gained experience and the current value of $\mathrm{Q}^{\pi}(S_{i+1}, p_{i+1})$. The updating rule for $\mathrm{Q}^{\pi}(S_i, p_i)$ in the SARSA algorithm is given by

$$\begin{aligned}\mathrm{Q}^{\pi}(S_i, p_i) \leftarrow\ & \mathrm{Q}^{\pi}(S_i, p_i)(1 - \alpha_i) \\ & + \alpha_i \left[R_i + \gamma \mathrm{Q}^{\pi}(S_{i+1}, p_{i+1})\right]\end{aligned} \tag{7}$$

[10], where α_i is a small positive fraction which influences the learning rate.

B. ϵ-greedy policy

In the following, the characteristics of the policy π which is followed throughout the learning process are discussed. When the number of states is finite, acting greedily with respect to $\mathrm{Q}^{\pi}(S_i, p_i)$, i.e., given S_i selecting the p_i that achieves the maximum $\mathrm{Q}^{\pi}(S_i, p_i)$, leads to the optimal policy [10]. This is due to the fact that $\mathrm{Q}^{\pi}(S_i, p_i)$ is the expected reward given the state-action pair (S_i, p_i). Therefore, selecting the p_i that maximizes $\mathrm{Q}^{\pi}(S_i, p_i)$ means that we are selecting the p_i that leads to the highest expected reward, which in our case corresponds to the throughput.

It has to be noticed that N_1 can only act greedily with respect to the states it has already encountered and the power values it has already used. Consequently, if N_1 follows the greedy policy, it does not have the opportunity to discover transmit power values that can potentially lead to higher rewards. To ensure that N_1 is able to explore the use of new transmit power values, the ϵ-greedy policy [10] is considered instead. In ϵ-greedy, most of the time N_1 acts greedily, this means

$$\mathbb{P}\left[p_i = \max_{p_k \in \mathcal{A}} \mathrm{Q}^{\pi}(S_i, p_k)\right] = 1 - \epsilon, \quad 0 < \epsilon < 1. \tag{8}$$

However, with a probability ϵ, N_1 will randomly select a transmit power value from the set $\mathcal{A}$. This method provides a trade-off between the exploration of new transmit power values and the exploitation of the known ones [10], [11].

C. Linear function approximation

As mentioned before, the concept of function approximation is used to handle the infinite number of states. When a finite number of states is considered, $\mathrm{Q}^{\pi}(S_i, p_i)$ is a table that assigns values for each state-action pair. However, when the number of states is infinite a table can no longer be constructed. With linear function approximation, $\mathrm{Q}^{\pi}(S_i, p_i)$ is represented by a linear combination of M feature functions $\mathrm{f}_m(S_i, p_i),\ m = 1, ..., M$. Each $\mathrm{f}_m(S_i, p_i)$, maps the state-action pair (S_i, p_i) into a feature value. Let $\mathbf{f} \in \mathbb{R}^{M\times 1}$ be a vector containing the feature values for a given state-action pair and let $\mathbf{w} \in \mathbb{R}^{M\times 1}$ be the vector containing the weights indicating the contribution of each feature. The action-value function approximation is given by

$$\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w}) = \mathbf{f}^{\mathrm{T}}\mathbf{w}. \tag{9}$$

[10]. To ensure that $\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})$ is a good representation of $\mathrm{Q}^{\pi}(S_i, p_i)$, the error between them has to be minimized. This can be done using a gradient descent approach [10]. However, as $\mathrm{Q}^{\pi}(S_i, p_i)$ is still unknown, the gradient descent method is performed using the current reward and the current value of $\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})$ [10].

In approximate SARSA, the updates are not performed on $\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})$ directly, as in the conventional case, but are performed on the weights. At t_i, the vector $\mathbf{w}$ is adjusted in the direction that reduces the error between $\mathrm{Q}^{\pi}(S_i, p_i)$ and $\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})$ following the gradient descent approach. Formally, the update rule for the approximate SARSA algorithm is given by

$$\begin{aligned}\mathbf{w} = \mathbf{w} + \alpha_i \big[R_i + \gamma \hat{\mathrm{Q}}^{\pi}(S_{i+1}, p_{i+1}, \mathbf{w}) \\ - \hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})\big] \nabla_{\mathbf{w}} \hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})\end{aligned} \tag{10}$$

[10]. As linear function approximation is used, the gradient of $\hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w})$ is calculated as

$$\nabla_{\mathbf{w}} \hat{\mathrm{Q}}^{\pi}(S_i, p_i, \mathbf{w}) = \mathbf{f}. \tag{11}$$

D. Feature functions

An important step in the implementation of the approximate SARSA algorithm is the definition of the feature functions. The features should correspond to the natural attributes of the EH problem in order to provide a good model of the effect of possible transmit power values on the state of the transmitter. In our scenario, the most important characteristics are the limited battery at N_1 and the unknown EH process. To apply linear function approximation, we propose a set of $M = 3$ binary functions which take into account the limited battery and the power allocation problem.

As overflow conditions are undesirable, the first feature function $\mathrm{f}_1(S_i, p_i)$ indicates if a given p_i avoids the overflow of the battery. Additionally, it evaluates if the given p_i fulfills the feasibility condition in (2). The binary function assigns "1" if no overflow is caused by the use of p_i in t_i and the feasibility condition is fulfilled. $\mathrm{f}_1(S_i, p_i)$ is written as

$$\mathrm{f}_1(S_i, p_i) = \begin{cases} 1, & \text{if } (B_i + E_i - \tau p_i \leq B_{\max}) \wedge (\tau p_i \leq B_i) \\ 0, & \text{else,} \end{cases} \tag{12}$$

where $\wedge$ represents the logical conjunction operation.

The second feature function $\mathrm{f}_2(S_i, p_i)$ addresses the power allocation problem. From [4], it is known that in the offline case a directional water-filling algorithm can be used to optimally allocate the power. However, as in our scenario the knowledge of future channel coefficients and energy values is unavailable, we propose to use past channel realizations to estimate the mean value of the distribution of the channel gain and to perform water-filling considering the estimated mean value of the channel gain and the current channel realization. For the estimation, the sample mean estimator is used such that at t_i the estimated mean value $\bar{h}_i$ is calculated as

$$\bar{h}_i = \frac{1}{i}\sum_{j=1}^{i} h_j. \tag{13}$$

Although E_i cannot be allocated in t_i, for the water-filling algorithm it is assumed that the available energy is $E_i + B_i$. The reason is that by performing water-filling between $\bar{h}_i$ and h_i, we are assuming that $\bar{h}_i$ approximates the state of the channel in the subsequent time instant and consequently, the available harvested energy has to be considered. The water level υ_i is calculated as

$$\upsilon_i = \frac{1}{2}\left(\frac{B_i}{\tau} + \frac{E_i}{\tau} + \sigma^2\left(\frac{1}{|\bar{h}_i|} + \frac{1}{|h_i|}\right)\right). \tag{14}$$

To ensure that the feasibility condition in (2) is fulfilled, the power allocation value given by the water-filling algorithm is given by

$$p_{i,\mathrm{WF}} = \min\left\{\frac{B_i}{\tau}, \max\left\{0, \upsilon_i - \frac{\sigma^2}{|h_i|}\right\}\right\}. \tag{15}$$

From Section III, we know that $p_i \in \mathcal{A}$. As a result, the calculated $p_{i,\mathrm{WF}}$ has to be rounded such that $p_{i,\mathrm{WF}} \in \mathcal{A}$ also

Algorithm 1 Approximate SARSA for EH

initialize γ, α, ϵ
initialize all the weights to one
observe S_i
select p_i using ϵ-greedy
while N_1 is harvesting energy **do**
  transmit using the selected p_i
  calculate corresponding reward R_i ▷ Eq. (1)
  observe next state S_{i+1}
  select next transmit power p_{i+1} using ϵ-greedy
  update $\mathbf{w}$ ▷ Eq. (10)
  set $S_i = S_{i+1}$
  set $p_i = p_{i+1}$
end while

holds. $\mathrm{f}_2(S_i, p_i)$ is written as

$$\mathrm{f}_2(S_i, p_i) = \begin{cases} 1, & \text{if } \delta \left\lfloor \frac{p_{i,\mathrm{WF}}}{\delta} \right\rfloor = p_i \\ 0, & \text{else,} \end{cases} \tag{16}$$

where $\lfloor x \rfloor$ is the rounding operation to the nearest integer less than or equal to x and δ is the step size used in the definition of the action set $\mathcal{A}$.

The third feature function $\mathrm{f}_3(S_i, p_i)$ handles the case when $E_i \geq B_{\max}$. In such situations, the battery should be depleted to minimize the energy losses due to battery overflow. The function assigns a "1" if the selected p_i is equal to the available power in the battery and it is defined as

$$\mathrm{f}_3(S_i, p_i) = \begin{cases} 1, & \text{if } (E_i \geq B_{\max}) \wedge \left(p_i = \delta \lfloor \frac{B_i}{\tau \delta} \rfloor\right) \\ 0, & \text{else.} \end{cases} \tag{17}$$

E. Approximate SARSA

The approximate SARSA algorithm for EH point-to-point communications is shown in Algorithm 1. Regarding the convergence properties of approximate SARSA, it has been shown in [13] that if α_i satisfies $\sum_i \alpha_i = \infty$ and $\sum_i \alpha_i^2 < \infty$ and the policy is not changed during the learning process, the approximate SARSA algorithm converges to a bounded region with probability one, i.e. it does not diverge. In our case, α_i is selected as $\alpha_i = 1/i$ which fulfills the two conditions. Additionally, throughout the execution of the algorithm, the ϵ-greedy policy is followed.

V. Performance Results

In this section, numerical results for the evaluation of the approximate SARSA algorithm in the EH point-to-point communication scenario are presented. For the simulations, $T = 1000$ independent random channel and energy realizations are generated. It is assumed that each realization corresponds to an episode where N_1 harvests energy from the environment $I = 1000$ times. Moreover, it is assumed that the amount of harvested energy E_i at t_i is taken from a uniform distribution with maximum value $E_{\max}$.

The time interval τ between two consecutive EH time instants is set to one second and the channel between N_1 and N_2 is assumed to be i.i.d. Rayleigh fading with zero mean and unit variance. Additionally, the noise variance is assumed as $\sigma^2 = 1$. The step size δ used in the definition of the action set $\mathcal{A}$ that contains the transmit power values is set to $\delta = 0.01 B_{\max}$. For the approximate SARSA algorithm, the ϵ-greedy policy is used with $\epsilon = 1/i$ and $\gamma = 0.9$. To compare the performance, four additional approaches are considered:

- Offline Optimum [4]: Non-causal information about the EH process is assumed as well as perfect channel state information, as presented in [4].
- Hasty Policy: In this approach, at each t_i, N_1 allocates all the power available in the battery. As a result, overflow conditions are completely avoided.
- Random Policy: In this approach, a set of feasible transmit power values is constructed in each time instant such that (2) is fulfilled. From this set, a transmit power value is randomly selected. It is assumed that all the transmit power values in the set have the same probability of being selected.
- Q-learning [9]: This method is the off-policy temporal-difference RL approach used in [9]. As Q-learning requires finite states, the results are obtained by the discretization of the energy, channel and battery values. For the simulations, the values are discretized using the step size δ defined above.

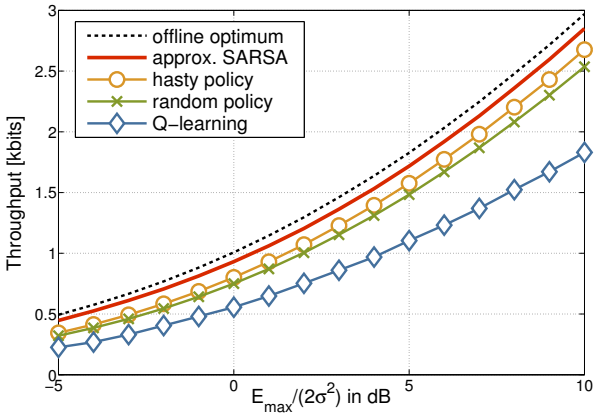

Fig. 2: Average throughput versus $E_{\max}/(2\sigma^2)$.

Fig. 2 shows the average throughput versus different values of $E_{\max}/(2\sigma^2)$. For this simulation, the battery size is set such that $B_{\max} = 2E_{\max}$ for each value of $E_{\max}$. As expected, the performance of all the approaches increases when the amount of harvested energy also increases. The upper bound of the achievable throughput is given by the optimum offline approach which assumes non-causal perfect information regarding the EH process and the channel. The approximate SARSA algorithm is able to overcome this unrealistic requirement at the cost of only 6% of performance reduction when $E_{\max}/(2\sigma^2) = 5$dB. For approximate SARSA, only causal information is assumed at N_1. Similarly, the hasty policy and the random policy assume only causal knowledge. However, since this information is not used for the power allocation, their

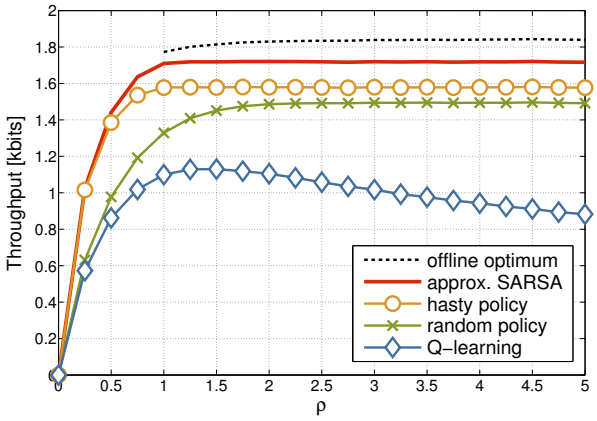

Fig. 3: Average throughput versus the battery size factor ρ. Average $E_{\max}/(2\sigma^2) = 5$dB.

performance is worse compared to our proposed approach. The throughput achieved by approximate SARSA is 9% and 16% higher than the throughput achieved by the hasty and random policy, respectively, when $E_{\max}/(2\sigma^2) = 5$dB. The lowest throughput is achieved by the Q-learning algorithm of [9]. This behavior is explained by the fact that Q-learning requires a finite number of states and to fit it to our system model, discretization is required for the harvested energy, battery and channel coefficients. Additionally, as the number of states increases (depending on how fine the discretization is), the probability of visiting all the states decreases and the learning becomes slower.

Fig. 3 evaluates the effect of the battery size on the throughput achieved by the different approaches for an $E_{\max}/(2\sigma^2) = 5$dB. In this case, the battery size is set to $B_{\max} = \rho E_{\max}$, where ρ is a tunable parameter. When $B_{\max} < E_{\max}$, the offline optimum cannot be calculated because overflow conditions are unavoidable and the problem becomes infeasible. Fig 3 shows that for different battery sizes, the proposed approximate SARSA algorithm performs better than the other approaches. When the battery is small, the performance of the approximate SARSA and the hasty policy is similar because all the harvested energy has to be spent in order to reduced the energy waste due to overflow. However, as the battery size increases, the transmitter conditions in each time instant have to be considered for the power allocation. As in the previous case, the lower throughput of the Q-learning algorithm is explained by the large number of states which reduce the learning speed compared to the approximate SARSA. An interesting result is that when the battery size is large compared to $E_{\max}$, its effect on the performance is reduced. It can be seen that the performance of all the approaches saturates from approximately $\rho = 2$. The reason for this is that as $B_{\max}$ increases, the overflow conditions become less probable.

VI. Conclusions

We have investigated the EH point-to-point communication scenario when only causal information regarding the EH process and channel is available at the transmitter. The scenario is modeled as a Markov decision process. Assuming that the transition probabilities between the states are unknown, RL with linear function approximation is applied at the transmitter in order to find a policy that aims at maximizing the throughput. To achieve this, a set of binary functions has been proposed to approximate the expected throughput in each state. Results show that the proposed approach is able to overcome the requirement of non-causal information with only a small reduction in the performance compared to the optimum offline case. Additionally, it is shown that the proposed approach performs better than naïve approaches that consider only causal information at the transmitter.

Acknowledgement

This work was funded by the LOEWE Priority Program NICER under grant No. III L5-518/81.004.

References

[1] D. Gündüz, K. Stamatiou, N. Michelusi, and M. Zorzi, "Designing intelligent energy harvesting communication systems," *IEEE Commun. Mag.*, vol. 52, no. 1, pp. 210–216, January 2014.

[2] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communication: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, March 2015.

[3] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, March 2012.

[4] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, September 2011.

[5] O. Orhan, D. Gündüz, and E. Erkip, "Throughput maximization for an energy harvesting communication system with processing cost," in *Proc. Inf. Theory Workshop (ITW)*, Lausanne, September 2012, pp. 84–88.

[6] K. Tutuncuoglu and A. Yener, "Optimal power policy for energy harvesting transmitters with inefficient energy storage," in *Proc. Annual Conf. Inform. Sciences Systems (CISS)*, Princeton, March 2012, pp. 1–6.

[7] J. Lei, R. Yates, and L. Greenstein, "A generic model for optimizing single-hop transmission policy of replenishable sensors," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 547–551, February 2009.

[8] S. Luo, R. Zhang, and T. J. Lim, "Optimal save-then-transmit protocol for energy harvesting wireless transmitters," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1196–1207, March 2013.

[9] 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.

[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. A Bradford Book — The MIT Press, 1198.

[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, M. Hirsch, Ed. Prentice Hall, 2010.

[12] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *Proc. 25th Int. Conf. Mach. Learning*, Helsinki, July 2008, pp. 664–671.

[13] G. J. Gordon, "Reinforcement learning with function approximation converges to a region," in *Advances Neural Inform. Process. Syst.* The MIT Press, 2001, pp. 1040–1046.