A. Ortiz, H. Al-Shatri, X. Li, T. Weber and A. Klein, ”A Learning Based Solution for Energy Harvesting Decode-and-Forward Two-Hop Communications,” in *Proc. IEEE Global. Commun. Conf. (Globecom)*, Washington, USA, December 2016.

A Learning Based Solution for Energy Harvesting Decode-and-Forward Two-Hop 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 (EH) two-hop communications are considered. The transmitter and the relay harvest energy from the environment and use it exclusively for transmitting data. A data arrival process is assumed at the transmitter. At the relay, a finite data buffer is used to store the received data. We consider a realistic scenario in which the EH nodes have only local causal information, i.e., at any time instant, each EH node only knows the current value of its EH process, channel state and data arrival process. Our goal is to find a power allocation policy to maximize the throughput at the receiver. We show that because the EH nodes have local causal information, the two-hop communication problem can be separated into two point-to-point problems. Consequently, independent power allocation problems are solved at each EH node. To find the power allocation policy, reinforcement learning with linear function approximation is applied. Moreover, to perform function approximation two feature functions which consider the data arrival process are introduced. Numerical results show that the proposed approach has only a small degradation as compared to the offline optimum case. Furthermore, we show that with the use of the proposed feature functions a better performance is achieved compared to standard approximation techniques.**

I. Introduction

In recent years, the interest in the design of transmission strategies for energy harvesting (EH) wireless communication networks has increased [1], [2]. EH wireless communications refer to scenarios where the wireless communication nodes have EH capabilities. In contrast to traditional wireless communication nodes, the EH nodes do not rely solely on conventional energy sources to recharge their batteries for transmitting data. EH nodes collect energy from the environment using natural energy sources, e.g., solar, thermal, vibrational, chemical, etc. This results in a reduction of the carbon footprint, higher mobility and self-sustainability [1].

Most of the research effort in EH communications has focused on offline settings in which perfect non-causal knowledge about the EH process is assumed at the nodes [3]–[6]. This assumption is hard to fulfill in real scenarios because the amount of harvested energy at the nodes is time variant and it depends on the energy source that is considered. However, the offline setting provides an upper-bound of the performance of the EH communication networks. The problem of throughput maximization within a deadline in an offline EH point-to-point communication scenario is investigated in [3]. Additionally, the authors show that this problem is equivalent to the minimization of the completion time for the transmission of a fixed amount of data. Offline EH two-hop communication networks are considered in [4]–[6]. In [4], the throughput maximization problem within a deadline is studied and two cases are distinguished, namely a full-duplex and a half-duplex relay. For the case of a full-duplex relay, an optimal transmission scheme is provided. However, in the half-duplex case, a simplified scenario is assumed where a single energy arrival is considered at the transmitter. In [5], the impact of a finite buffer at the relay for the storage of data is investigated. It is assumed that the transmitter harvests energy several times while the relay harvests only once. Furthermore, the authors in [6] formulate a convex problem to find offline transmission policies for multiple parallel relays in the two-hop EH communication scenario.

A more realistic approach is given by the online setting in which non-causal statistical information about the EH process is assumed [7]–[9]. In [7], the EH point-to-point scenario is considered and an on-off mechanism at the transmitter is studied. The authors assume a data arrival process at the transmitter and for each packet, a binary decision of whether to transmit or drop is made. In [8] and [9] dynamic programming is used to solve the throughput maximization problem in the point-to-point and two-hop communication scenarios, respectively.

Despite the fact that online settings do not require perfect knowledge as the offline setting, having knowledge about the statistics of the EH process in advance cannot always be achieved [2]. Moreover, even if the statistical information is available, assuming that the EH process is stationary and does not change with time is a strong assumption, e.g., if different energy sources are considered simultaneously. In emergency scenarios for example, EH wireless communication networks can be used if the communication infrastructure is damaged. In this case, statistical information about the EH sources is not available and the online setting is not applicable. A solution to this problem is proposed in [8] where reinforcement learning (RL) is applied in the EH point-to-point scenario. 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 and apply the well-known RL algorithm Q-learning to maximize the throughput in a fixed period of time. In [10], the RL algorithm SARSA is combined with linear function approximation to overcome the limitations of Q-learning and to improve the performance in a point-to-point communication scenario with only causal information.

In this paper, we consider an EH two-hop communication scenario with a full-duplex decode-and-forward relay. Our goal is to find a power allocation policy at the transmitter and at the relay which aims at maximizing the amount of data at the receiver. Local causal information is assumed to be available at the transmitter and at the relay. This means that at any time instant, the transmitter and the relay have only knowledge about the value of their own EH process, channel state and data arrival process. In general, the power allocation problem for throughput maximization in the two-hop scenario is coupled. However, we show that when the nodes have only causal information about their own process, the problem can be separated into two point-to-point problems. This is due to the fact that the transmitter and the relay do not know the state of each other and therefore, they cannot adapt their power allocation policy to improve the amount of data that reaches the receiver. As a result, independent power allocation problems can be solved at the transmitter and at the relay which aim at maximizing the throughput in each point-to-point scenario. Based on [10], the RL algorithm SARSA with function linear approximation is applied in each point-to-point scenario to find the power allocation policy. Moreover, to perform the linear function approximation, we introduce two new feature functions. These feature functions take into account the data causality constraint given by the data arrival process and avoid data overflow situations caused by the finite data buffer. Furthermore, to evaluate the performance of the proposed feature functions, we implement SARSA with linear function approximation using standard approximation techniques, namely, fixed sparse representation (FSR) and radial basis functions (RBF) [11].

The rest of the paper is organized as follows. In Section II, the system model is introduced. The power allocation problem for throughput maximization in an EH two-hop scenario is presented in Section III. In Section IV, the EH two-hop communication scenario is reformulated as two point-to-point communication problems. In Section V, each point-to-point problem is modeled as a Markov decision process and RL is applied to find power allocation policies. Numerical performance results are presented in Section VI and Section VII concludes the paper.

II. System Model

In this paper, a two-hop EH communication scenario is considered. As depicted in Fig. 1, the scenario consists of three single-antenna nodes. The term N_k, $k \in \{1,2,3\}$, is used to label the nodes. The transmitter node N_1 wants to transmit data to the receiver node N_3. It is assumed that the link between these two nodes is weak. Therefore, the nodes cannot communicate directly. To enable the communication, N_2 acts as a full-duplex decode-and-forward relay which is able to perfectly cancel the self-interference and it forwards the data from N_1 to N_3. A data arrival process is assumed at N_1 from which $R_{0,i}$ bits are received at t_i. It is assumed that N_2 does not have any own data to transmit to the other nodes. The data available for transmission at N_1 is stored in a finite

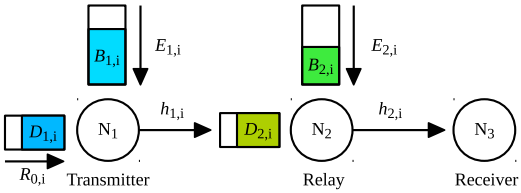


Fig. 1: EH two-hop communication scenario.

data buffer of size $D_{\max,1}$ measured in bits. Moreover, N_2 has a data buffer of size $D_{\max,2}$, where it stores the data received from N_1. As the goal only is to maximize the throughput, it is assumed that the data packets do not have deadlines that need to be fulfilled.

In our scenario, N_1 and N_2 harvest energy from the environment and use this energy exclusively for the transmission of data. As in [3]–[6], it is assumed that the energy is harvested at 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 $E_{l,i} \in \mathbb{R}^+$, $l = \{1, 2\}$ is received by N_l. It has to be noticed that this notation does not mean that at each t_i, both nodes N_1 and N_2 harvest energy. For example, if node N_l does not harvest energy at t_i, then $E_{l,i} = 0$.

The maximum amount of energy that can be harvested at N_l, termed $E_{\max,l}$, depends on the energy source that is used. After $E_{l,i}$ is harvested, it is stored in a rechargeable finite battery with maximum capacity $B_{\max,l}$. Ideal batteries are assumed. Therefore, no energy is lost while storing or retrieving energy. It is assumed that the batteries cannot be recharged instantaneously. Consequently, at t_i the batteries only store the energy which has been harvested until t_{i-1}. Furthermore, it is assumed that at t_1, the nodes have not yet harvested any energy and their batteries are 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 received noise at N_2 and N_3 is assumed to be independent and identically distributed (i.i.d.) zero mean additive white Gaussian noise with variance $\sigma_2^2 = \sigma_3^2 = \sigma^2$. The fading channel coefficient from N_1 to N_2 is termed $h_{1,i} \in \mathbb{C}$ while the fading channel coefficient between N_2 and N_3 is termed $h_{2,i} \in \mathbb{C}$. Further, the transmit power $p_{l,i}$ of N_l is kept constant during the time interval τ from t_i to t_{i+1} [3]. We assume that only local causal information is available at the EH nodes. This means that at t_i, each node N_l has knowledge about the current state of its battery $B_{l,i} \in \mathbb{R}^+$, the harvested energy $E_{l,i}$, the channel state $h_{l,i}$ and the state $D_{l,i} \in \mathbb{R}^+$ of its data buffer. Using this causal information, N_l selects $p_{l,i}$ for the transmission of data during the corresponding time interval.

III. Problem Formulation

In this section, the power allocation problem for throughput maximization is formulated. At t_i, the throughput achieved during one time interval τ is defined as the amount of data that reaches N_3 and it is measured in bits. Since we consider a

decode-and-forward relay and N_1 does not send data directly to N_3, it corresponds to the throughput $R_{2,i}$, i.e, the amount of data received by N_3 from N_2. N_2 only transmits what it has received from N_1. Consequently, $R_{2,i}$ is limited by the throughput $R_{1,i}$ which is the amount of data received at N_2 from N_1. At t_i, $R_{1,i}$ and $R_{2,i}$ are given by

$$R_{l,i} = \tau \log_2 \left(1 + \frac{|h_{l,i}|^2 p_{l,i}}{\sigma^2}\right),\ l = \{1, 2\}. \tag{1}$$

As N_1 and N_2 harvest energy from the environment, the power available for transmission depends on their corresponding EH processes. Moreover, at N_l the transmit power can be allocated only after the harvested energy has been stored in the battery. As a result, the energy causality constraint,

$$\tau p_{l,i} \leq B_{l,i},\ \ l = \{1, 2\}, \tag{2}$$

must be fulfilled. The finite capacity of the battery should be considered in order to avoid overflow situations in which part of the harvested energy is wasted because the battery is full. The energy overflow constraint is given by

$$B_{l,i} - \tau p_{l,i} + E_{l,i} \leq B_{\max,l},\ \ l = \{1, 2\}. \tag{3}$$

As mentioned before, a data arrival process is assumed at N_1 in which $R_{0,i}$ bits are received at each time instant t_i. $R_{0,i}$ is a realization of an independent data arrival process. However, the data arrival process at N_2 depends on the throughput $R_{1,i}$. As N_2 does not have any own information to transmit, it can only transmit the data previously received from N_1, i.e., the data which is already stored in the data buffer. At t_i, the state $D_{l,i}$ of the data buffer at N_l is calculated as

$$D_{l,i} = \sum_{n=1}^{i-1} R_{l-1,n} - \sum_{n=1}^{i-1} R_{l,n},\ \ l = \{1, 2\}. \tag{4}$$

The throughputs $R_{1,i}$ and $R_{2,i}$ are limited by the information causality constraint given by

$$R_{l,i} \leq D_{l,i},\ \ l = \{1, 2\}, \tag{5}$$

which ensures that N_l cannot retransmit data it has not yet received.

The size $D_{\max,l}$ of each data buffer has to be considered to avoid data overflow. When the data buffer is full, the received data cannot be stored and it is discarded. Similar to the energy overflow constraint in (3), N_l has an information overflow constraint

$$D_{l,i} - R_{l,i} + R_{l-1,i} \leq D_{\max,l}. \tag{6}$$

Considering (2), (3), (5) and (6), the power allocation problem for throughput maximization in the EH two-hop communication scenario is written as

$$\left(p_{l,i}^{\text{opt}}\right)_{l,i} = \underset{\{p_{l,i},\, l=\{1,2\}, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \ \sum_{i=1}^{I} R_{2,i} \tag{7a}$$

$$\text{subject to} \quad \sum_{i=1}^{M} \tau p_{l,i} \leq \sum_{i=1}^{M-1} E_{l,i},\ \forall l,\ M = 1, ..., I, \tag{7b}$$

$$\sum_{i=1}^{M} E_{l,i} - \sum_{i=1}^{M} \tau p_{l,i} \leq B_{\max,l},\ \forall l, M, \tag{7c}$$

$$\sum_{i=1}^{M} R_{l,i} \leq \sum_{i=1}^{M-1} R_{l-1,i}, \forall l, M, \tag{7d}$$

$$\sum_{i=1}^{M} R_{l-1,i} - \sum_{i=1}^{M} R_{l,i} \leq D_{\max,l},\ \forall l, M, \tag{7e}$$

$$p_{l,i} \geq 0, \quad \forall l,\ i = 1, ..., I. \tag{7f}$$

Although the problem in (7) is a convex optimization problem it can only be solved if non-causal knowledge about the EH process, the data arrival and channel state is available. In our scenario, it is assumed that the nodes have only local causal information. Therefore, we propose to apply RL at each N_l. The application of RL is discussed in Section V.

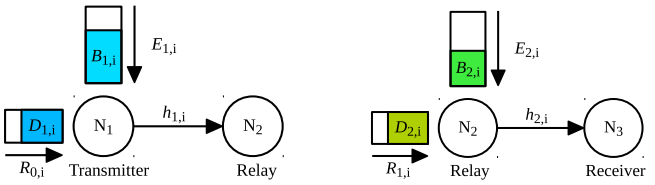
Another consequence of having only causal information is that the nodes do not know in advance for how many EH time intervals I they will operate. Therefore, at t_i it is preferred to achieve a higher throughput in the current interval over future ones. To consider this, the objective function in (7a) is rewritten such as to maximize the expected throughput. Moreover, a discount factor γ, with $0 \leq \gamma \leq 1$, is included to account for the preference of higher throughput values in the current interval. The objective function in (7a) is replaced by the expected throughput given by

$$R = \lim_{I \to \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_{2,i}\right]. \tag{8}$$

IV. Reformulation of the Throughout Maximization Problem

In this section, we show that when only local causal information is available at the transmitter and at the relay, the two-hop communication problem can be seen as two EH point-to-point communication problems, as depicted in Fig. 2. The first problem corresponds to the link $\mathrm{N}_1 \rightarrow \mathrm{N}_2$ between N_1 and N_2 and it is shown in Fig. 2(a). The second one corresponds to the link $\mathrm{N}_2 \rightarrow \mathrm{N}_3$ between N_2 and N_3 and it is illustrated in Fig. 2(b).

The energy harvesting processes of the nodes are independent. Nevertheless, the power allocation problem of N_1 and N_2 described in (7) is coupled because $R_{2,i}$ is limited by the throughput $R_{1,i}$. When only local causal information is available, the problem cannot be solved in a coupled way because the nodes have no information about the power allocation policy of each other, neither the EH process, channel state or data arrival process. As N_1 has no knowledge about the state of the data buffer in N_2, it cannot avoid data overflow



(a) Link $N_1 \rightarrow N_2$ (b) Link $N_2 \rightarrow N_3$

Fig. 2: Reformulation of the two-hop EH communication problem as two point-to-point communication problems

by reducing its transmit power. Therefore, N_1 can allocate its power to maximize the throughput $R_{1,i}$ independently of the state of the data buffer at N_2.

Since at N_l the data arrival process is unknown and only knowledge about the state of its data buffer is available, the data arrival process is treated in the same fashion as the energy arrival process. Consequently, N_l independently allocates its power in order to maximize the throughput $R_{l,i}$. The power allocation problem for throughput maximization in each link $N_1 \rightarrow N_2$ and $N_2 \rightarrow N_3$ is given by

$$p_{l,i}^{\text{opt}} = \underset{\{p_{l,i},\, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \lim_{I\to\infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^i R_{l,i}\right] \tag{9a}$$

$$\text{subject to} \quad \sum_{i=1}^{M} \tau p_{l,i} \leq \sum_{i=1}^{M-1} E_{l,i},\ M = 1, ..., I, \tag{9b}$$

$$\sum_{i=1}^{M} E_{l,i} - \sum_{i=1}^{M} \tau p_{l,i} \leq B_{\max,1},\ \forall M \tag{9c}$$

$$R_{l,i} \leq D_{l,i},\ i = 1, ..., I, \tag{9d}$$

$$D_{l,i} - R_{l,i} + R_{l-1,i} \leq D_{\max},\ \forall i, \tag{9e}$$

$$p_{l,i} \geq 0, \quad \forall i, \tag{9f}$$

for $l = 1$ and $l = 2$, respectively. It has to be noted that at N_2, the data overflow constraint described in (9e) cannot always be fulfilled. This is because N_2 is a full-duplex relay and at t_i, it does not know how much data it will receive from N_1. The throughput $R_{1,i}$ is only known at N_2 at the end of the time interval, i.e. at t_{i+1}. To overcome this problem, we propose the use of an estimate of $R_{1,i}$. This approach is presented in section V-B when the feature functions are discussed.

V. Reinforcement Learning Approach

In this section, we model each point-to-point communication problem as a Markov decision process (MDP) and use a RL approach to find the power allocation policies that aim at maximizing the throughput. Based on our previous work [10], we apply SARSA with linear function approximation. A brief description of the SARSA algorithm and the feature functions used in [10] to approximate the expected throughput are included here for completeness. Additionally, we propose two new feature functions to consider the data arrival processes at the EH nodes.

A. Markov Decision Process Model

For each node N_l, $l = \{1, 2\}$, the MDP consists of a set of states $\mathcal{S}_l$, a set of actions $\mathcal{A}_l$, a transition model $\mathcal{P}_l$ and a set of rewards $\mathcal{R}_l$ [12]. At t_i, the state $S_{l,i} \in \mathcal{S}_l$ of node N_l is a function of $B_{l,i}$, $E_{l,i}$, $h_{l,i}$ and $D_{l,i}$. The battery level, the harvested energy, the channel coefficients and the data buffer state can take any value in a continuous range. As a consequence, the set $\mathcal{S}_l$ contains an infinite number of possible states. For N_l, these states are given by any value of $B_{l,i}$, $E_{l,i}$ and $h_{l,i}$ and $D_{l,i}$.

The set of actions $\mathcal{A}_l$ is composed by all the transmit power values $p_{l,i}$ that each node can select. We consider a finite set given by $\mathcal{A}_l = \{p_{l,i} |\ p_{l,i} \in \{0, \delta_l, 2\delta_l, ..., B_{\max,l}\}\}$, where δ_l is a step size [10]. The action dependent transition model defines the transition probabilities from state $S_{l,i}$ to state $S_{l,i+1}$. Finally, the rewards indicate how beneficial the selected $p_{l,i}$ is for the corresponding $S_{l,i}$ of node N_l. For each pair $S_{l,i}$ and $p_{l,i}$, the reward $R_{l,i} \in \mathcal{R}_l$ is defined as the throughput achieved in one time interval τ and it is calculated as described in (1).

We are interested in finding a power allocation policy at each node N_l to maximize the throughput $R_{l,i}$. A policy π_l is a mapping from a given $S_{l,i}$ to the $p_{l,i}$ that should be selected, i.e. $p_{l,i} = \pi_l(S_{l,i})$, and it corresponds to the solution of an MDP [12]. π_l can be evaluated using the so-called action-value function $Q_l^{\pi}(S_{l,i}, p_{l,i})$ which is defined as the expected reward starting from state $S_{l,i}$, selecting $p_{l,i}$ and following π_l thereafter [13]. The optimal policy π_l^* is the policy whose action-value function is greater than or equal to any other policy for every state. The corresponding action-value function for the optimal policy π_l^* is denoted by Q_l^*. Determining π_l^* is straightforward when Q_l^* is known because for each $S_{l,i}$, any action $p_{l,i}$ that maximizes $Q_l^*(S_{l,i}, p_{l,i})$ is an optimal action.

B. SARSA with Linear Function Approximation

As only local causal information is available at the nodes, the action-value function $Q_l^{\pi_l}$ is unknown. Therefore, SARSA builds an estimate of the action-value function from the states that are visited and the earned rewards. At every t_i, node N_l selects a transmit power value $p_{l,i}$ according to its current state $S_{l,i}$. The selected $p_{l,i}$ leads to a throughput $R_{l,i}$. After the transmission, N_l is in state $S_{l,i+1}$ and for this state a new transmit power value $p_{l,i+1}$ is selected. $Q_l{}^{\pi_l}$ is updated considering $S_{l,i}$, $p_{l,i}$, $R_{l,i}$, $S_{l,i+1}$ and $p_{l,i+1}$.

Linear function approximation is used to represent $Q_l^{\pi_l}$ when the number of states is infinite. The action-value function $Q_l^{\pi_l}$ is approximated using a linear combination of Y feature functions $f_y(S_{l,i}, p_{l,i}),\ y = 1, ..., Y$ which map the state-action pair $(S_{l,i}, p_{l,i})$ into a feature value. The approximate $Q_l^{\pi_l}$, termed $\hat{Q}_l^{\pi_l}$, is calculated as the weighted sum of the features. For a given pair $(S_{l,i}, p_{l,i})$, the feature values are collected in the vector $\mathbf{f}_l \in \mathbb{R}^{Y\times 1}$ and the contribution of each feature is included in the vector of weights $\mathbf{w}_l \in \mathbb{R}^{Y\times 1}$. The action-value function is approximated as

$$Q_l^{\pi_l}(S_{l,i}, p_{l,i}) \approx \hat{Q}_l{}^{\pi_l}(S_{l,i}, p_{l,i}, \mathbf{w}_l) = \mathbf{f}_l^{\mathrm{T}} \mathbf{w}_l, \tag{10}$$

[13]. When SARSA with linear function approximation is applied, the updates are performed on the weights because they control the contribution of each feature function on $\hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i}, p_{l,i})$. At t_i, the vector $\mathbf{w}_l$ is adjusted in the direction that reduces the error between $\mathrm{Q}_l^{\pi_l}(S_{l,i}, p_{l,i})$ and $\hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i}, p_{l,i}, \mathbf{w}_l)$ following the gradient descent approach. Formally, the update rule is given by [13]

$$\Delta \mathbf{w}_l = \alpha_i \big[R_{l,i} + \gamma \hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i+1}, p_{l,i+1}, \mathbf{w}_l) - \hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i}, p_{l,i}, \mathbf{w}_l) \big] \nabla_{\mathbf{w}_l} \hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i}, p_{l,i}, \mathbf{w}_l), \tag{11}$$

where α_i is a small positive fraction which influences the learning rate. Throughout the execution algorithm, the ϵ-greedy policy is followed. In ϵ-greedy, each N_l acts greedily with respect to its action-value function with a probability of $1-\epsilon$, this means

$$\Pr\left[p_{l,i} = \max_{p_{l,k} \in \mathcal{A}_l} \hat{\mathrm{Q}}_l^{\pi_l}(S_{l,i}, p_{l,k}) \right] = 1-\epsilon, \quad 0 < \epsilon < 1. \tag{12}$$

However, with a probability ϵ, N_l will randomly select a transmit power value from the set $\mathcal{A}_l$. This method provides a trade-off between the exploration of new transmit power values and the exploitation of the known ones [12], [13].

For the definition of the feature functions, the natural attributes of the problem should be considered. In our case, these attributes are the EH processes at N_1 and N_2, their finite batteries, their data arrival processes and finite data buffers. In [10], $Y = 3$ binary feature functions were presented for the point-to-point scenario without data arrival. We propose two additional feature functions to consider the data arrival process and the data buffer.

The first feature function $\mathrm{f}_1(S_{l,i}, p_{l,i})$ deals with overflow conditions. It indicates if in state $S_{l,i}$, a given $p_{l,i}$ avoids battery overflow according to (3). Additionally, it evaluates if $p_{l,i}$ fulfills the energy causality constraint of (2). $\mathrm{f}_1(S_{l,i}, p_{l,i})$ is defined in [10] as

$$\mathrm{f}_1(S_{l,i}, p_{l,i}) = \begin{cases} 1, & \text{if } (B_{l,i} + E_{l,i} - \tau p_{l,i} \leq B_{\max,l}) \wedge \\ & \quad (\tau p_{l,i} \leq B_{l,i}) \\ 0, & \text{else,} \end{cases} \tag{13}$$

where $\wedge$ represents the logical conjunction operation.

The second feature function $\mathrm{f}_2(S_{l,i}, p_{l,i})$ addresses the power allocation problem. It uses past channel realizations to estimate the mean value $\bar{h}_{l,i}$ of the channel gain in order to perform water-filling. The water level $\upsilon_{l,i}$ is calculated as

$$\upsilon_{l,i} = \frac{1}{2}\left(\frac{B_{l,i}}{\tau} + \frac{E_{l,i}}{\tau} + \sigma^2 \left(\frac{1}{|\bar{h}_{l,i}|} + \frac{1}{|h_{l,i}|} \right) \right). \tag{14}$$

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

$$p_{l,i}^{\mathrm{WF}} = \min\left\{ \frac{B_{l,i}}{\tau}, \max\left\{ 0, \upsilon_{l,i} - \frac{\sigma^2}{|h_{l,i}|} \right\} \right\}, \tag{15}$$

[10]. As $p_{l,i}$ can only be selected from the discrete set $\mathcal{A}_l$, the calculated $p_{l,i}^{\mathrm{WF}}$ is rounded such that $p_{l,i}^{\mathrm{WF}} \in \mathcal{A}_l$ holds. $\mathrm{f}_2(S_{l,i}, p_{l,i})$ is written in [10] as

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

where $\lfloor x \rfloor$ is the rounding operation to the nearest integer less than or equal to x.

The third feature function $\mathrm{f}_3(S_{l,i}, p_{l,i})$ handles the case when the size of the battery is small compared to the harvested energy, i.e., $E_{l,i} \geq B_{\max,l}$. In this situation, the battery should be depleted to minimize the energy losses due to battery overflow. $\mathrm{f}_3(S_{l,i}, p_{l,i})$ is given in [10] by

$$\mathrm{f}_3(S_{l,i}, p_{l,i}) = \begin{cases} 1, & \text{if } (E_{l,i} \geq B_{\max,l}) \wedge \left(p_{l,i} = \delta \lfloor \frac{B_{l,i}}{\tau\delta} \rfloor \right) \\ 0, & \text{else} \end{cases} \tag{17}$$

As mentioned before, we extend the work in [10] with two additional feature functions. The fourth and fifth feature functions are proposed in order to consider the data arrival process and data buffer at the EH nodes. The information causality constraint is addressed with the fourth feature function. Let us define $R_{l,i}^{(p_{l,i})}$ as the throughput that would be achieved if $p_{l,i}$ is selected. $\mathrm{f}_4(S_{l,i}, p_{l,i})$ indicates if $R_{l,i}^{(p_{l,i})}$ fulfills the constraint in (5) and it is defined as

$$\mathrm{f}_4(S_{l,i}, p_{l,i}) = \begin{cases} 1, & \text{if } R_{2,i}^{(p_{l,i})} \leq D_{l,i} \\ 0, & \text{else.} \end{cases} \tag{18}$$

As discussed in the previous section, data overflow situations cannot be completely avoided at N_2 because at t_i, knowledge about $R_{1,i}$ is not available. To overcome this, we propose that in the case of N_2, the data overflow constraint in (6) is evaluated using the mean value $\bar{R}_{1,i}$ of the previously achieved throughputs, i.e., $\bar{R}_{1,i} = \frac{1}{i-1}\sum_{j=1}^{i-1} R_{1,j}$. Similar to f_4, we consider $R_{l,i}^{(p_{l,i})}$ and use $\mathrm{f}_5(S_{l,i}, p_{l,i})$ to indicate if data overflow situations can be avoided by the selection of a given $p_{l,i}$. $\mathrm{f}_5(S_{1,i}, p_{1,i})$ is given by

$$\mathrm{f}_5(S_{l,i}, p_{l,i}) = \begin{cases} 1, & \text{if } D_{l,i} + \bar{R}_{l-1,i} - R_{l,i}^{(p_{l,i})} \leq D_{\max,l} \\ 0, & \text{else,} \end{cases} \tag{19}$$

where $\bar{R}_{l-1,i} = R_{l-1,i}$ for $l = 1$. As a summary, the approximate SARSA algorithm for each point-to-point scenario is shown in Algorithm 1. For information about the convergence properties of SARSA with linear function approximation, the reader is referred to [14] and [15].

VI. Performance results

In this section, numerical results for the evaluation of the SARSA algorithm in the two-hop communication scenario are presented. As described in previous sections, SARSA with linear function approximation is applied at each node N_l to maximize the throughput at N_3. The results are obtained by generating $T = 1000$ independent random channel and

Algorithm 1 SARSA algorithm [10].

initialize γ, α, ϵ and $\mathbf{w}_l$
observe $S_{l,i}$
select $p_{l,i}$ using ϵ-greedy
while N_l is harvesting energy **do**
  transmit using the selected $p_{l,i}$
  calculate corresponding reward $R_{l,i}$ ▷ Eq. (1)
  observe next state $S_{l,i+1}$
  select next transmit power $p_{l,i+1}$ using ϵ-greedy
  update $\mathbf{w}_l$ ▷ Eq. (11)
  set $S_{l,i} = S_{l,i+1}$ and $p_{l,i} = p_{l,i+1}$
end while

energy realizations. Each realization corresponds to an episode where the nodes harvest energy I times. We are interested in evaluating the throughput when the the data available at the transmitter is not a limiting factor. Therefore, we consider the case in which the transmitter has always data to transmit, i.e. $D_{1,i} = \infty,\ \forall i$.

For each node N_l, the amount of harvested energy $E_{l,i}$ 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 time unit and the channel coefficients $h_{l,i}$ are assumed to be taken from an i.i.d. Rayleigh fading process with zero mean and unit variance. Additionally, the noise variance is set to $\sigma^2 = 1$. For the SARSA algorithm at N_l, the step size δ used in the definition of the action set $\mathcal{A}_l$ is set to $\delta = 0.02B_{\max,l}$. The learning rate α and the ϵ parameter used in the ϵ-greedy policy are reduced in each time instant and are defined as $\alpha = 1/i$ and $\epsilon = 1/i$, respectively. Furthermore, the discount factor γ is selected as $\gamma = 0.9$.

For comparison, we consider the offline optimum and the hasty policy. The offline optimum is obtained by solving the optimization problem of (7) when non-causal information regarding the EH process, the data arrival process and the channel states is available. On the contrary, the hasty policy consists of depleting the battery of N_1 in every time instant. At N_2, the hasty policy tries to deplete the data buffer at each time instant by selecting the maximum power value that fulfills the information causality constraint of (5). Additionally, we implement the SARSA algorithm using two standard approximation techniques, i.e., FSR and RBF [11]. FSR is a low-complexity technique used to represent the continuous states. For N_l, the state $S_{l,i}$ lies in a 4-dimensional space given by $B_{l,i}$, $E_{l,i}$, $h_{l,i}$ and $D_{l,i}$. In FSR, each dimension is split in tiles and a binary feature function is assigned to each tile. A given feature function is equal to one if the corresponding variable is in the tile and zero otherwise [11]. In our implementation, the tiles are generated using the step size δ. In contrast to FSR that uses binary feature functions, RBF works directly in the continuous space. In RBF, each feature function has a Gaussian response that depends on the distance between a given state and the center of the feature [11], [13].

The average throughput performance versus different values of $E_{\max}/(2\sigma^2)$ is shown in Fig. 3. The battery sizes of the

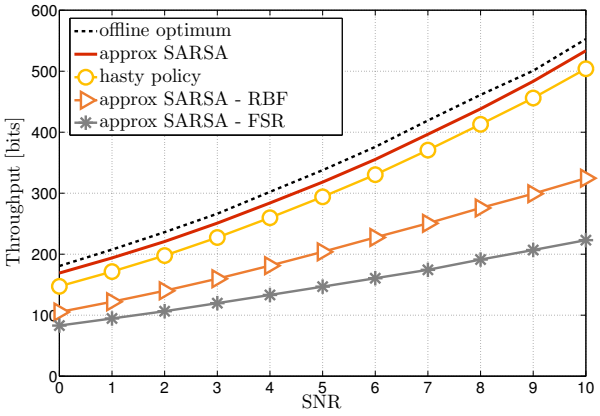


Fig. 3: Average throughput versus $E_{\max}/(2\sigma^2)$.

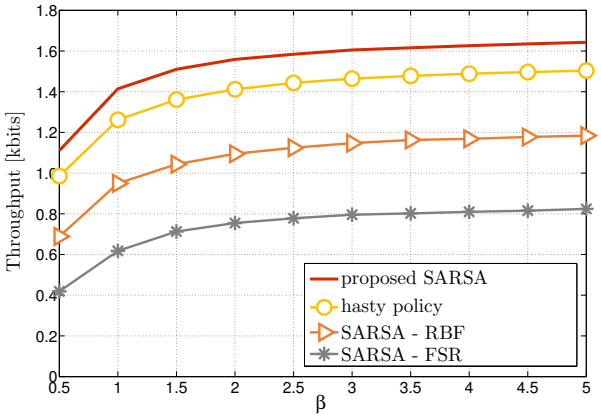


Fig. 4: Average throughput versus data buffer size factor β. Average $E_{\max}/(2\sigma^2) = 5$dB.

nodes are set to $B_{\max,1} = B_{\max,2} = B_{\max} = 2E_{\max}$ and $I = 100$ EH time instants are considered. In this case, we are interested in evaluating the throughput performance when the data buffer at the relay is not limiting the transmission. Therefore, $D_{\max,2}$ is selected as $D_{\max,2} = 5R_{1,i}^{(B_{\max})}$, where $R_{1,i}^{(B_{\max})}$ is the throughput that would be achieved if $|h_{1,i}| = 1$ and $p_{1,i} = B_{\max}/\tau$. As expected, the performance of all the approaches increases when the amount of harvested energy increases. It can be seen that the proposed SARSA algorithm is able to overcome the unrealistic assumption of the offline approach with only 6% performance reduction when $E_{\max}/(2\sigma^2) = 5$dB. As it can be seen in Fig. 5, at $I = 100$ the SARSA algorithm has not yet converged. However, this value was selected to be able to find a numerical solution for the offline optimum. As a consequence, the difference between the hasty policy and the proposed SARSA is only 8%. The low performance of SARSA-FSR and SARSA-RBF is due to the fact that they are general representation techniques that do not consider the characteristics of the problem. Moreover, a large number of feature functions have to be used to approximate all the states which reduces the learning rate.

Fig. 4 shows the effect of the data buffer size on the

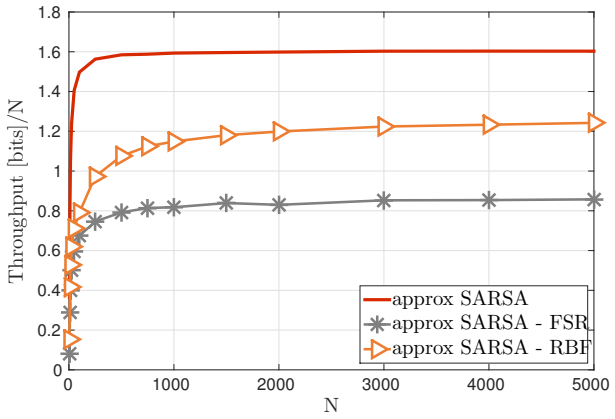


Fig. 5: Average normalized throughput versus the number of EH time instants I. Average $E_{\max}/(2\sigma^2) = 5$dB.

performance for $E_{\max}/(2\sigma^2) = 5$dB. In this case, $I = 1000$ and the buffer size at N_2 is $D_{\max,2} = \beta R_{1,i}^{(B_{\max})}$, where β is a tunable parameter. The offline optimum is not considered because when the data buffer size is small compared to the throughput $R_{1,i}$, data overflow conditions are unavoidable and no feasible solutions can be found for the problem of (7). Results show that the proposed SARSA consistently outperforms the other approaches. For small values of β, the throughput is reduced because not all the data received from N_1 can be stored in the data buffer and it is discarded. When the data buffer size is large compared to $R_{1,i}$, its effect on the performance is reduced. It can be seen that the performance of all the approaches saturates at approximately $\beta = 3$ when the data buffer is big compared to the throughput received from N_1 and the data overflow conditions become less probable.

The convergence speed of the SARSA algorithm is evaluated in Fig. 5 for $E_{\max}/(2\sigma^2) = 5$dB and $\beta = 5$. The figure shows the normalized throughput versus the number I of EH time instants. The throughput is normalized with respect to the number of EH time instants I. The proposed SARSA converges faster than SARSA-FSR and SARSA-RBF and it achieves a higher throughput. The reason for this is that the proposed SARSA uses customized feature functions based on the properties of the problem given by the constraints of (2), (3), (5) and (6). On the contrary, FSR and RBF are general representation techniques that do not consider the characteristics of the problem. Additionally, with the proposed SARSA the number of feature functions used in the approximation is only five. This improves the learning rate compared to FSR and RBF.

VII. Conclusions

A full-duplex decode-and-forward two-hop communication scenario with EH nodes was investigated. A data arrival process was considered at the transmitter and a finite data buffer was assumed at the transmitter and at the relay. Local causal knowledge regarding the EH process, the data arrival process and the channel state was assumed at the transmitter and at the relay. We have shown that the power allocation problem for throughput maximization can be seen as two point-to-point problems when only local causal information is available at the nodes. Each point-to-point problem is modeled as a Markov decision process and the RL algorithm SARSA with linear function approximation is applied. Moreover, for the linear function approximation customized feature functions are proposed to consider the data arrival process at the nodes. Results show that the proposed approach is able to overcome the requirement of non-causal information with only a small reduction in the performance as compared to the optimum offline case. Moreover, it is shown that the use of customized feature functions achieves a better performance than standard approximation techniques

Acknowledgment

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

References

[1] 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.

[2] 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.

[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] D. Gündüz and B. Devillers, "Two-hop communication with energy harvesting," in *Proc. IEEE Int. Workshop on Comput. Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, San Juan, December 2011, pp. 201–204.

[5] B. Varan and A. Yener, "Two-hop networks with energy harvesting: The (non-)impact of buffer size," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Austin, December 2013, pp. 399–402.

[6] O. Orhan and E. Erkip, "Energy harvesting two-hop communication networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2658–2670, December 2015.

[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] 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.

[9] I. Ahmed, A. Ikhlef, R. Schober, and R. K. Mallik, "Power allocation in energy harvesting relay systems," in *Proc. IEEE 75th Veh. Technology Conf. (VTC Spring)*, Yokohama, May 2012, pp. 1–5.

[10] 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, May 2016, accepted.

[11] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 6, no. 4, pp. 375–454, December 2013.

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall, 2010.

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

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

[15] 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.