Andrea Ortiz, Tobias Weber, Ajna Klein "Multi-Agent Reinforcement Learning for Energy Harvesting Two-Hop Communications with a Partially Observable System State," in *IEEE Transactions on Green Communications and Networking*. 2020

Multi-Agent Reinforcement Learning for Energy Harvesting Two-Hop Communications with a Partially Observable System State

Andrea Ortiz, *Member, IEEE,* Tobias Weber, *Senior Member, IEEE,* and Anja Klein, *Member, IEEE*

***Abstract*—We consider an energy harvesting (EH) transmitter communicating with a receiver through an EH relay. The harvested energy is used for data transmission, including the circuit energy consumption. As in practical scenarios, the system's state, comprised by the harvested energy, battery levels, data buffer levels, and channel gains, is only partially observable by the EH nodes. Moreover, the EH nodes have only outdated knowledge regarding the channel gains for their own transmit channels. Our goal is to find distributed transmission policies aiming at maximizing the throughput. A channel predictor based on a Kalman filter is implemented in each EH node to estimate the current channel gain for its own channel. Furthermore, to overcome the partial observability of the system's state, the EH nodes cooperate with each other to obtain information about their parameters during a signaling phase. We model the problem as a Markov game and propose a multi-agent reinforcement learning algorithm to find the transmission policies. We show the trade-off between the achievable throughput and the signaling required, and provide convergence guarantees for the proposed algorithm. Results show that even when the signaling overhead is taken into account, the proposed algorithm outperforms other approaches that do not consider cooperation.**

***Index Terms*—Two-hop communications, energy harvesting, decode and forward, multi-agent reinforcement learning, linear function approximation.**

I. Introduction

Wireless communication nodes play an important role in many applications of wireless sensor networks such as health monitoring, surveillance or intelligent buildings. However, depending on the specific application, charging or replacing the batteries of the wireless communication nodes can be too expensive or sometimes infeasible [1], e.g., when the nodes are located inside the human body, in remote locations or even inside structures. In order to provide sustainable service or to reduce the operating expenses, energy harvesting (EH) has been considered as a promising energy source for such wireless communication nodes. In EH wireless communication networks, the EH capability of the nodes increases the network lifetime and can lead to perpetual operation because the nodes can use the harvested energy to recharge their batteries [2,3]. However, the benefits of EH are not limited to an increased network lifetime. The fact that the EH nodes can collect energy from natural or man-made sources, e.g., solar, chemical or electromagnetic radiation, helps to reduce greenhouse gas emissions. Furthermore, since the EH nodes can work independently of the power grid, EH wireless communication networks can be deployed in areas that are usually hard to reach. In this paper, we address the problem of how to efficiently use the harvested energy and we tackle the problem from a communications perspective, i.e., we discuss how to efficiently transmit data using the harvested energy as the only energy source.

In an EH scenario, the communication range depends on the amount of harvested energy at the EH transmitter. This amount of harvested energy varies according to the energy source that is considered. For example, for energy harvesting based on electromagnetic radiation, the power density is in the order of fractions of $\mathrm{nW/cm^2}$, and for solar energy, it is in the order of hundreds of $\mathrm{mW/cm^2}$. To increase the limited communication range in an EH communication scenario, relaying techniques can be considered since they are cost effective solutions for increasing the coverage, throughput and robustness of wireless networks [4,5]. By using relaying techniques, the communication between a transmitter and a receiver which are located far apart can be achieved by introducing one or more intermediate relays for reducing the communication range of each hop. The reduction of the communication range implies a reduction of the amount of energy required for data transmission in each hop. We focus on the case where only a single EH relay is used to assist the communication between an EH transmitter and a receiver, i.e., EH two-hop communications. This scenario is the essential building block of more complicated EH multi-hop communication networks and exhibits all important challenges that need to be addressed when using relaying techniques, i.e., the design of transmission policies for the EH transmitter and the EH relay considering the amount of energy that is available to each of them. Our goal is to design transmission policies aiming at an efficient use of the harvested energy at the transmitter and at the relay in order to maximize the throughput. This problem is equivalent to the minimization of the time required to transmit a given amount of data [6].

A. Related Work

For EH two-hop scenarios, offline approaches have been the major direction of state-of-the-art research [6]–[12]. Offline approaches assume that perfect non-causal knowledge about the system dynamics is available. This means, all the amounts of energy to be harvested, the amounts of incoming data and the channel gains to be experienced are perfectly known before the data transmission starts. In [7], the throughput maximization problem within a deadline is studied and two cases are

This work has been performed in the context of the LOEWE Center emergenCITY.

distinguished, namely, a full-duplex and a half-duplex relay. For the case of a full-duplex relay, the optimal transmission strategy is provided. However, in the half-duplex case, the optimal transmission strategy is only found for a simplified scenario in which a single energy arrival is considered at the transmitter. This scenario is extended in [6], where two energy arrivals at the transmitter node and the relay station are considered. For this case, the authors derive transmission policies to maximize the data transmitted to the receiver within a deadline. An amplify-and-forward relay is considered in [8] where two relaying protocols are proposed based on time-switching and power-splitting architectures. The throughput maximization problem when the transmitter harvests energy multiple times and the decode-and-forward relay has only one energy arrival is investigated in [11]. A similar scenario is considered in [12]. However, in [12], the impact of a finite data buffer at the relay is investigated. Multiple parallel relays in a decode-and-forward EH two-hop scenario are investigated in [9,10], where the authors formulate a convex optimization problem to find the optimal offline transmission policy that maximizes the throughput. In [13], an EH two-scenario with a full-duplex amplify-and-forward relay is considered and the authors propose a two-phase protocol for efficient energy transfer and information relaying.

In [14]–[16], simultaneous wireless information and power transfer in a two-hop scenario with multiple relays is considered. In [14], the authors assume randomly located relays and analyze the performance of the system considering the impact of the number of relays. In [15], the concept of distributed space-time coding is applied to multiple relays which assist the communication between the transmitter and the receiver, and the authors in [16] aim at minimizing the transmission time and propose a harvest-then-decode-and-forward algorithm at the relays. Energy cooperation is introduced in [17] for the EH relay, two-way and multiple access channels in order to find offline energy management policies that maximize the throughput. (R2.1) In [18], the throughput maximization problem in an EH multi-hop scenario with full-duplex relays is considered.

In [19]–[21], online approaches for EH two-hop scenarios are considered. In this case, only statistical knowledge of the system dynamics is assumed. In [19], a half-duplex amplify-and-forward relay in an EH two-hop scenario is studied. The authors assume statistical knowledge about the EH process and find the transmission policy using discrete dynamic programming. A similar scenario is considered in [20,21], where a power allocation policy aiming at maximizing the long-time average throughput is found using Lyapunov optimization techniques. In [22], an EH multi-hop scenario with full-duplex EH relays is investigated. Assuming Bernoulli distributed EH processes, the authors design power control policies based on the retransmission-index following an online approach.

In real scenarios, perfect non-causal knowledge or statistical knowledge of the system dynamics is usually not available, especially if non-stationary EH, data arrival and channel fading processes are considered [23]. In such cases, learning techniques, specifically reinforcement learning (RL), can be exploited to find transmission policies that aim at maximizing a given objective, e.g., the throughput. Learning techniques, although promising for EH scenarios, have hardly been used so far in EH two-hop scenarios [23]–[26]. In [24], a learning approach for an EH two-hop scenario is considered where the authors optimize the average delay of the packets sent by the source in a scenario with multiple half-duplex EH relays. In our previous work [23,25], a blind approach is considered, and the two-hop communications scenario is separated into two independent point-to-point scenarios. In this paper, we overcome the limitation on the performance imposed by this blind approach and propose the introduction of a signaling phase in which the transmitter and the relay cooperate with each other to observe the system state and to improve the achievable throughput while taking into account both, the energy required for transmission and the energy consumed by the circuit of each of the nodes. A different problem is considered in [26], where the authors optimize the relay operation mode to maximize the throughput in a two-hop scenario with an EH relay and a non-EH transmitter.

B. Contributions

We focus our work on EH two-hop communications. We consider a realistic scenario in which the state of the system is only partially and causally observable to the EH nodes. This means, the proposed approach does not require previous knowledge about the statistics of the EH, data arrival or channel fading processes. In each time interval, each EH node only knows its own previous and current states. The state of a node consists of the values of its own parameters, i.e., the amount of incoming energy, the battery level, the data buffer level and the channel gain for its own transmit channel. Additionally, we investigate the case when only outdated channel state information is available. To this aim, we leverage the use of a channel predictor based on a Kalman filter in each EH node in order to obtain a current estimate of the channel gain. Furthermore, inspired by the information exchange in wireless sensor networks [27], we propose a signaling phase in which the EH nodes share information about their current parameters in order to overcome the partial observability of the system's state. We are interested in a distributed solution where each EH node finds its own transmission policy taking into account its observation of the system's state and the knowledge obtained during the signaling phase. Considering that the problem consists of two agents, the transmitter and the relay, who should make simultaneous decisions to achieve a common goal, i.e., decide on the transmit powers in order to maximize the throughput, we model this scenario as a Markov game. This is because Markov games provide a framework to include multiple decision making agents with interacting or competing goals [28]. Additionally, to find the distributed transmission policies at the transmitter and at the relay, we propose a cooperative multi-agent RL algorithm termed cooperative SARSA. The use of RL is motivated by the fact that complete non-causal knowledge is unavailable. As a consequence, standard optimization techniques cannot be used. To validate our proposed cooperative SARSA, we derive convergence guarantees for the case when the EH nodes

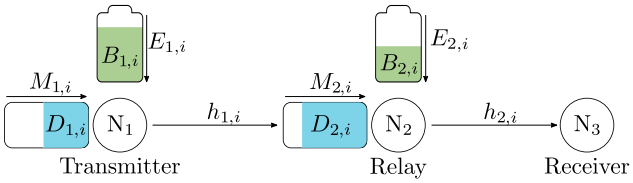


Fig. 1. Two-hop communication scenario with an EH transmitter and an EH relay.

are able to observe the system's state, i.e., when the channel prediction and the transmission of the signaling are successful. By means of a computational complexity analysis, we show that the leading factor of the complexity of the proposed cooperative SARSA increases only linearly with the number of transmit power values the nodes can select. Moreover, by numerical results we show that the performance of the proposed algorithm has only a small degradation compared to the offline case which requires complete non-causal knowledge of the system's state. Furthermore, we show that even when the overhead caused by the signaling phase is taken into account, cooperative SARSA outperforms other approaches that do not consider cooperation among the EH nodes, and therefore do not require a signaling phase.

The rest of the paper is organized as follows. In Sec. II, the system model is presented. In Sec. III, the power allocation problem for throughput maximization in the EH two-hop communication scenario is formulated. The proposed cooperative SARSA algorithm is described in Sec. IV. Convergence guarantees and a computational complexity analysis of the proposed algorithm are presented in Sec. IV-H and Sec. IV-H, respectively. Numerical performance results are presented in Sec. VI and Sec. VII concludes the paper.

II. System Model

In this section, we describe the EH two-hop communication scenario. A summary of all the considered parameters is given in Table I. The scenario consists of three single-antenna nodes N_1, N_2, and N_3, as depicted in Figure 1, where the EH transmitter N_1 wants to transmit data to the non-EH receiver N_3. It is assumed that the link between N_1 and N_3 is weak and the nodes cannot communicate directly. Therefore, N_2 acts as an EH decode-and-forward relay in order to enable the communication between N_1 and N_3.

In our scenario, N_1 and N_2 harvest energy from the environment and use it for data transmission. An amount of harvested energy, denoted by $E_{1,i}$ and $E_{2,i}$, is received by N_1 and N_2, respectively, at the end of time interval i, $i = 1, ..., I$. The harvested energy is stored in batteries with finite capacities given by $B_{\max,1}$ and $B_{\max,2}$ for N_1 and N_2, respectively. Furthermore, the battery levels $B_{1,i}$ and $B_{2,i}$ are measured at the beginning of time interval i. For simplicity, the energy $E_{1,i}^{\mathrm{Circ}}$ consumed by the circuit at N_1 is assumed to be constant for all the time intervals, i.e., $E_{1,i}^{\mathrm{Circ}} = E_1^{\mathrm{Circ}}$, $\forall i$. Similarly, for N_2, $E_{2,i}^{\mathrm{Circ}} = E_2^{\mathrm{Circ}}$, $\forall i$. However, the model can be extended to consider a variable decoding cost at N_2, as in [29].

A data arrival process is assumed at N_1 in which an amount $M_{1,i}$ of incoming data is arriving at N_1 at the end of each time interval i and it is stored in a finite buffer with capacity $D_{\max,1}$. The data buffer level $D_{1,i}$ is measured at the beginning of time interval i and indicates the amount of data available for transmission. In the considered EH two-hop scenario, the communication between N_1 and N_3 is as follows. In each time interval i, N_1 selects a transmit power $p_{1,i}^{\mathrm{Tx}}$ to transmit data to N_2 for a duration $\Delta\tau$ of the time interval, i.e., an amount $E_{1,i}^{\mathrm{Tx}} = \Delta\tau p_{1,i}^{\mathrm{Tx}}$ of energy is used for data transmission. The value of the prelog factor Δ depends on the relay's transmission mode and it is defined as $\Delta = 1$ if N_2 operates in full-duplex mode and $\Delta = 0.5$ if it operates in half-duplex mode.[1] This definition accounts for the fact that when the relay operates in full-duplex mode, the total duration of the time interval is used for the transmission from N_1 to N_2 and from N_2 to N_3. On the contrary, when half-duplex is considered, we assume that one half of the time interval is reserved for the transmission from N_1 to N_2 and the other half is used for the transmission from N_2 to N_3. Note that the operation mode of N_2 is selected at the beginning of the first time interval $i = 1$ and cannot be changed throughout the operation, i.e., Δ is fixed for all time intervals. The throughput $R_{1,i}^{\mathrm{DF}}$ is the amount of data received at N_2 in time interval i. When there is sufficient data in the data buffer of N_1, $R_{1,i}^{\mathrm{DF}}$ is approximated using Shannon's capacity formula as

$$R_{1,i}^{\mathrm{DF}} = \Delta W \tau \log_2 \left(1 + \frac{g_{1,i} p_{1,i}^{\mathrm{Tx}}}{\sigma_2^2}\right), \tag{1}$$

where W denotes the available bandwidth, $g_{1,i}$ is the channel gain for the link between N_1 and N_2 and σ_2^2 is the noise power at N_2. Otherwise, $R_{1,i}^{\mathrm{DF}}$ is limited by the amount of data stored in the data buffer. Additionally, note that for full-duplex it is assumed that the relay is able to perfectly cancel the self-interference caused by its transmission. The battery level at N_1 is updated at the beginning of each time interval as

$$B_{1,i+1} = \min\left\{B_{\max,1},\ B_{1,i} - \Delta\tau p_{1,i}^{\mathrm{Tx}} + E_{1,i} - E_1^{\mathrm{Circ}}\right\}. \tag{2}$$

Similarly, the data buffer level at N_1 is updated at the beginning of each time interval as

$$D_{1,i+1} = \min\left\{D_{\max,1},\ D_{1,i} - R_{1,i}^{\mathrm{DF}} + M_{1,i}\right\}. \tag{3}$$

The EH relay N_2 only forwards the data from N_1 to N_3 and it does not have any own data to transmit to the receiver. Therefore, the data arrival process at N_2 depends solely on the data transmitted by N_1. This means that the amount $M_{2,i}$ of incoming data at N_2, which arrives at the end of time interval i, corresponds to the throughput $R_{1,i}^{\mathrm{DF}}$, i.e., $M_{2,i} = R_{1,i}^{\mathrm{DF}}$. The received $M_{2,i}$ is stored in a finite data buffer with size $D_{\max,2}$ and the data buffer level $D_{2,i}$ is measured at the beginning of each time interval i. Similar to the previous case, N_2 selects a transmit power $p_{2,i}^{\mathrm{Tx}}$ to use for the transmission of data to N_3 for a duration $\Delta\tau$ of the time interval. The throughput $R_{2,i}^{\mathrm{DF}}$ is the amount of data received at N_3, measured in bits. In case there is enough data available for transmission, $R_{2,i}^{\mathrm{DF}}$

[1] Note that the value of E_2^{Circ} depends on the operation mode of the relay N_2. In case N_2 is a full-duplex relay, E_2^{Circ} additionally includes the energy required for self-interference cancellation.

TABLE I. Parameters associated to the EH two-hop communication scenario.

	Parameter	Description
General	i	Index of the time interval
	I	Total number of time intervals
	N_1	EH transmitter node
	N_2	EH relay node
	N_3	Non-EH receiver node
	τ	Time interval duration
	Δ	Prelog factor depending on the relay's transmission mode
Energy	$B_{n,i}$	Battery level of EH node N_n, measured at the beginning of time interval i
	$B_{\max,n}$	Battery capacity of EH node N_n
	$E_{n,i}$	Amount of harvested energy, received at the end of time interval i, by EH node N_n
	$E_{n,i}^{\mathrm{Circ}}$	Amount of energy consumed by the circuit of EH node N_n in time interval i
	$E_{n,i}^{\mathrm{Tx}}$	Energy of the signal transmitted by EH node N_n in time interval i
	$E_{\max,n}$	Maximum amount of energy that can be harvested by EH node N_n
	$p_{n,i}^{\mathrm{Tx}}$	Transmit power used by EH node N_n in time interval i
Data	$D_{\max,n}$	Data buffer size of EH node N_n
	$D_{n,i}$	Data buffer level of EH node N_n, measured at the beginning of time interval i
	$M_{n,i}$	Amount of incoming data, arriving at the end of time interval i, at EH node N_n
	$R_{n,i}^{\mathrm{DF}}$	Amount of data transmitted from N_n to N_{n+1} in time interval i
Channel	$g_{n,i}$	Channel gain of the link between N_n and N_{n+1}
	W	Bandwidth
	σ_n^2	Noise power at N_n

is approximated using Shannon's capacity formula as

$$R_{2,i}^{\mathrm{DF}} = \Delta W \tau \log_2 \left(1 + \frac{g_{2,i} p_{2,i}^{\mathrm{Tx}}}{\sigma_3^2}\right), \tag{4}$$

where $g_{2,i}$ is the channel gain for the link between N_2 and N_3 and σ_3^2 is the noise power at N_3. Otherwise, $R_{2,i}^{\mathrm{DF}}$ is limited by the amount of data available in the data buffer. As done for N_1, the battery level and the data buffer level at N_2 are updated using (2) and (3), respectively, by replacing the index $n = 1$ by $n = 2$. Additionally, N_3 is assumed to be connected to a fixed power supply and it is always available to receive the transmitted data.

It is assumed that the transmitter side channel state information is only causally known and could be outdated. This means that at the beginning of time interval i, only the channel gains up to time interval $i-1$ are known at the transmitter and at the relay. Furthermore, it is assumed that the EH transmitter does not know the channel gains associated to the link between the EH relay and the receiver.

III. Problem Formulation

In this section, the power allocation problem for the EH two-hop scenario with a decode-and-forward relay is formulated. Our goal is to find a transmission policy at N_1 and at N_2 that maximizes the throughput, i.e., the amount of data transmitted to N_3. Considering the system model of Sec. II, the power allocation problem is written as

$$\left(p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{n,i} = \underset{\{p_{n,i}^{\mathrm{Tx}},\, n=\{1,2\}, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_{2,i}^{\mathrm{DF}} \tag{5a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \Delta\tau p_{n,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{n,i}, \tag{5b}$$

$$\sum_{i=1}^{J} E_{n,i} - \sum_{i=1}^{J} \Delta\tau p_{n,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq B_{\max,n}, \tag{5c}$$

$$\sum_{i=1}^{J} R_{n,i}^{\mathrm{DF}} \leq \sum_{i=1}^{J-1} M_{n,i}, \tag{5d}$$

$$\sum_{i=1}^{J} M_{n,i} - \sum_{i=1}^{J} R_{n,i}^{\mathrm{DF}} \leq D_{\max,n}, \tag{5e}$$

$$p_{n,i}^{\mathrm{Tx}} \geq 0, \tag{5f}$$

$$n = 1, 2,\ i = 1, \ldots, I,\ J = 1, \ldots, I, \tag{5g}$$

where $R_{1,i}^{\mathrm{DF}}$ and $R_{2,i}^{\mathrm{DF}}$ are defined in (1) and (4), respectively, (5b) is the energy causality constraint that ensures that only the energy stored in the battery can be used, (5c) is the battery overflow constraint, (5d) is the data causality constraint that ensures that only data already stored in the data buffers can be transmitted and (5e) is the data buffer overflow constraint for N_1 and N_2, respectively. By examining the problem in (5), it can be seen that perfect non-causal knowledge of the system's state for all time intervals $i = 1, \ldots, I$ is required to find the optimal solution. The amount of data to be transmitted by N_2 depends on its own EH, data arrival and channel fading processes as well as the ones associated to N_1. Moreover, N_1 should adapt its transmission based on the EH and channel fading processes associated to N_2 to avoid data buffer overflow situations. As a result, the state of each EH node affect the power allocation policy of the other.

IV. Multi-agent RL for EH Two-hop Communications with Partially Observable System State

A. Cooperation in multi-agent RL

As mentioned before, both N_1 and N_2 have only causal, and possibly outdated, knowledge regarding their own state. While $E_{n,i}$, $B_{n,i}$ and $D_{n,i}$, $n = \{1, 2\}$, are causally known by the corresponding node N_n, only outdated channel state information is available at N_1 and N_2. This means, at time interval i, we know the values of $E_{n,j}$, $B_{n,j}$ and $D_{n,j}$, $\forall j \leq i$, whereas for the channel gains $g_{n,j}$, only the values up to time interval $i-1$ are assumed to be known. As a result, the only outdated parameters are the channel gains. However,

knowledge about the system's state is required at both nodes in order to achieve optimum performance. To this aim, in this section we propose a cooperative multi-agent learning approach, termed cooperative SARSA, to find power allocation policies at N_1 and at N_2 that aim at maximizing the amount of data transmitted to N_3. Note that in addition to the challenge posed by the partial observability of the system's state, the nodes might not be able to observe the decision made by the other node before making their own, e.g., if a full-duplex relay is considered. For this reason, in the following we focus on a full-duplex decode-and-forward relay. Note that the same approach can be used for a half-duplex relay. The only difference is that the nodes will not make simultaneous decisions.

Our proposed cooperative SARSA includes mechanisms to overcome the limitation that N_1 and N_2 are only able to partially observe the system's state. Specifically, we consider that the channel state information might be outdated and use a channel predictor based on a Kalman filter in each EH node in order to obtain a current estimate of the channel gain. Furthermore, we propose a signaling phase in which the EH nodes cooperate which each other by exchanging information about their current state. Based on their knowledge of their own state and the knowledge they have obtained during the signaling phase, N_1 and N_2 find their own transmission policies.

B. Markov game for multi-agent learning

In this section, we model the power allocation problem in the EH two-hop communication scenario as a Markov game. This model is motivated by the fact that in the cooperative SARSA approach, N_1 and N_2 decide on the transmit power to use based on the system state, i.e., the value of the parameters associated to both of them. Such decision-making situations, in which more than one agent is involved, can be modeled as a Markov game. Markov games are a generalization of Markov decision processes (MDPs) to the case when multiple agents, which make decisions based on observations of a common environment, are considered [28].

A Markov game of n players is defined by the tuple $\langle \mathcal{S}, \mathcal{A}_1, ..., \mathcal{A}_n, \mathsf{P}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle$. The set $\mathcal{S}$ corresponds to all the possible states in which the system can be, the sets $\mathcal{A}_1, ..., \mathcal{A}_n$ contain the actions of each player, P is the transition model and $\mathcal{R}_1, ..., \mathcal{R}_n$ are the reward functions for each player [30]. In our case, the players are N_1 and N_2. Therefore, $n = 2$ is considered. Each state $S_i \in \mathcal{S}$ corresponds to the system state and it is defined as the tuple $\langle E_{1,i}, E_{2,i}, B_{1,i}, B_{2,i}, D_{1,i}, D_{2,i}, g_{1,i}, g_{2,i} \rangle$. Note that the set $\mathcal{S}$ comprises an infinite number of states S_i because the parameters can take values in a continuous range. The sets $\mathcal{A}_n$ of actions are formed by the possible transmit power values $p_{n,i}^{\mathrm{Tx}}$ that can be selected. As in practical settings [31], we define $\mathcal{A}_1$ and $\mathcal{A}_2$ for N_1 and N_2, respectively, as finite sets given by $p_{n,i}^{\mathrm{Tx}} \in \mathcal{A}_n = \{0, \delta, 2\delta, ..., B_{\max,n}\}$, where δ is the step size. The transition model P is defined as $\mathsf{P} : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathcal{S}$ and it specifies that, given state S_i, the system reaches state S_{i+1} after the EH nodes have selected $p_{1,i}^{\mathrm{Tx}} \in \mathcal{A}_1$ and $p_{2,i}^{\mathrm{Tx}} \in \mathcal{A}_2$, i.e., $S_{i+1} = \mathsf{P}(S_i, p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}})$. The reward function $\mathcal{R}_n$ gives the immediate reward obtained by N_n when $p_{n,i}^{\mathrm{Tx}}$ is selected while being in state S_i. In our case, the nodes aim at maximizing the throughput, i.e., the amount of data received by N_3. Consequently, N_1 and N_2 share the same objective, thus $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}$. In each time interval, the reward is calculated using (4).

Similar to MDPs, in the Markov game formulation we need to find the transmission policies π_n, $l \in \{1, 2\}$ for N_1 and N_2 which correspond to the transmit powers to be used for data transmission in each time interval. Each π_n, is a mapping from a given system state S_i to the action $p_{n,i}^{\mathrm{Tx}}$ that should be selected, i.e. $p_{n,i}^{\mathrm{Tx}} = \pi_n(S_i)$, and it is evaluated using the so-called action-value function $\mathrm{Q}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ [32]. However, as N_n has only causal knowledge about the system state, it does not know how much energy will be harvested, how much data will arrive or what the channel gain will be in future time intervals. We consider this uncertainty by defining the discount factor of future rewards γ, $0 \leq \gamma \leq 1$, which quantifies the preference of achieving a larger throughput in the current time interval over future ones. Our goal is to select $p_{n,i}^{\mathrm{Tx}}$, $\forall n, i$, in order to maximize the expected throughput

$$R^{\mathrm{DF}} = \lim_{I \to \infty} \mathbb{E} \left[\sum_{i=1}^{I} \gamma^{i-1} R_{2,i}^{\mathrm{DF}} \right]. \tag{6}$$

C. Action-value function update

The proposed cooperative SARSA algorithm is based on the RL algorithm SARSA [32]. Therefore, to facilitate its description, in this section we first consider the single-agent case by assuming that an ideal central entity has, in each time interval, perfect knowledge about S_i and uses RL to find the combined policy $\Pi = (\pi_1, \pi_2)$. Next in this section, we describe the case when the two EH nodes are considered.

The policy Π can be evaluated using the action-value function $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$, with $P_i^{\mathrm{Tx}} = (p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}})$. However, this action-value function cannot be calculated before the data transmission starts because only causal knowledge is available at the nodes and the statistics of the EH, data arrival and channel fading processes are unknown. As a result, the RL algorithm builds an estimate of the action-value function Q^{Π} using SARSA as

$$\begin{aligned} \mathrm{Q}_{i+1}^{\Pi}(S_i, P_i^{\mathrm{Tx}}) = & \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})(1 - \zeta_i) \\ & + \zeta_i \left[R_i^{\mathrm{DF}} \gamma \mathrm{Q}_i^{\Pi}(S_{i+1}, P_{i+1}^{\mathrm{Tx}}) \right] \end{aligned} \tag{7}$$

[32], where ζ_i is a small positive fraction which influences the learning rate.

In our scenario, the nodes have a common objective, which is to maximize the expected throughput given in (6), and in every time interval they make independent decisions that aim at achieving this objective taking into account the system state. However, as the nodes do not know in advance the transmit power which will be selected by the other node, they cannot build an estimate of the centralized action-value function $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$. Consequently, instead of the action-value function $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$, in the proposed cooperative SARSA algorithm, each node builds an estimate of its own

action-value function $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$, which is termed the local action-value function. In order to guarantee the convergence of the proposed learning approach, the local action-value function $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is designed such that it is a projection of the centralized $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ onto the corresponding state-action space $(S_i, p_{n,i}^{\mathrm{Tx}})$. For this purpose, the EH nodes will only update their current estimate of $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ when the value of the update is larger than the current one. This ensures that the local action-value policy is only updated when higher rewards are achieved. The relation between $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ and $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ and its effect on the convergence guarantees of cooperative SARSA is presented in detail in Sec. IV-H. Furthermore, the proposed updating rule for $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is given by

$$\begin{aligned}\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max \big\{ & \mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}), \\ & (1-\zeta_i)\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) + \\ & \zeta_i \left[R_i^{\mathrm{DF}} + \gamma \mathrm{q}_{n,i}^{\pi_n}(S_{i+1}, p_{n,i+1}^{\mathrm{Tx}}) \right] \big\}. \end{aligned} \tag{8}$$

D. Linear function approximation

The update of the action-value function, presented in Sec. IV-C, does not take into account the fact that in our scenario, the number of states is infinite. Therefore, in this section we exploit the use of linear function approximation for the representation of the action-value function when an infinite number of states are considered. With linear function approximation, $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is approximated as the linear combination of a set of F feature functions. Each feature function $\mathrm{f}_f(S_i, p_{n,i}^{\mathrm{Tx}})$, $f = 1, ..., F$, maps the state-action pair $(S_i, p_{n,i}^{\mathrm{Tx}})$ onto a feature value. Moreover, for a given pair $(S_i, p_{n,i}^{\mathrm{Tx}})$, the feature values are collected in the vector $\mathbf{f}_n \in \mathbb{R}^{F\times 1}$ and the contribution of each feature is included in the vector of weights $\mathbf{w}_n \in \mathbb{R}^{F\times 1}$. Using linear function approximation, the local action-value function q_n is approximated as

$$\hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}, \mathbf{w}_n) = \mathbf{f}_n^{\mathrm{T}} \mathbf{w}_n \approx \mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}). \tag{9}$$

When SARSA with linear function approximation is applied, the updates of the local action-value function $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ are performed on the weights $\mathbf{w}_n$ because they control the contribution of each feature function on $\hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}, \mathbf{w}_n)$. In every time interval, the vector $\mathbf{w}_n$ is adjusted in the direction that reduces the error between $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ and $\hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}, \mathbf{w}_n)$, following the gradient descent approach presented in [32]. Considering the update for $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ given in (8), we propose to update $\mathbf{w}_n$ as

$$\mathbf{w}_{n,i+1} = \mathbf{w}_{n,i} + \max\left\{0,\, \zeta_i \left[R_i^{\mathrm{DF}} + \gamma\, \mathbf{f}_n^{\mathrm{T}} \mathbf{w}_{n,i} - \mathbf{f}_n^{\mathrm{T}} \mathbf{w}_{n,i} \right] \mathbf{f}_n \right\}. \tag{10}$$

E. Partially observable states

In this section, we describe the mechanisms proposed to overcome the fact that the EH nodes are only able to partially observe the system state. Specifically, we describe the channel predictor based on a Kalman filter which is used by every EH node N_n to estimate its own channel coefficients $h_{n,i}$, with $g_{n,i} = |h_{n,i}|^2$, when only outdated channel state information is available, and the signaling phase in which N_1 and N_2 exchange the current values of their own parameters in order to be able to observe the system state.

Channel predictor: To obtain channel state information at the receiver, a known symbol $x_{n,i}$ is assumed to be transmitted from N_n to N_{n+1}. The received signal $y_{n+1,i}$ at N_{n+1} in the low-pass domain is $y_{n+1,i} = x_{n,i} h_{n,i} + w_{n+1,i}$, where $w_{n+1,i}$ accounts for the receiver noise and interference, and has variance σ^2. This received signal $y_{n+1,i}$ is used by N_{n+1} to determine the channel coefficient $h_{n,i}$. However, in order to have channel state information at N_1 side, it is assumed that N_{n+1} feeds back the channel coefficients to N_n. Since these channel coefficients might be outdated, channel prediction can be exploited at N_1 to determine an estimate of $h_{n,i}$. For this purpose, the past channel coefficients $h_{n,j}$, $j < i$, which have been fed back by N_{n+1} are used.

The magnitude $|h_{n,i}|$ of the channel coefficient $h_{n,i}$ is assumed to follow a Rayleigh distribution and the Jakes' model [33] is used to model the autocorrelation function ACF of the channel coefficients [34,35] as

$$\mathrm{ACF} = \mathrm{J}_0(2\pi f_{D,\max} \tau), \tag{11}$$

where J_0 is the zero$^{\text{th}}$ order Bessel function of the first kind and $f_{D,\max}$ is the maximum Doppler frequency. As extensively reported in literature [34]–[36], for the channel prediction at each N_n, the dynamics of the channel coefficient are modeled as an autoregressive process with order o and parameters $c_{n,1}, ..., c_{n,o}, \psi_n$. Specifically, $h_{n,i}$ is modeled as

$$h_{n,i} = -\sum_{j=1}^{o} c_{n,j} h_{n,i-j} + \psi_n z_{n,i}, \tag{12}$$

where $z_{n,i}$ is additive white Gaussian noise. The parameters $c_{n,1}, ..., c_{n,o}, \psi_n$ are calculated at N_n by means of solving the Yule-Walker equation considering the ACF in (11). Considering $y_{n,i}$ and (12), the state-space model for $h_{n,i}$ can be built. For this purpose, let us define the vectors $\mathbf{h}_{n,i} = [h_{n,i}, h_{n,i-1}, ..., h_{n,i-o}]^{\mathrm{T}}$, $\mathbf{a}_n = [\psi_n, 0, ..., 0]$ and $\mathbf{x}_{n,i} = [x_{n,i}, 0, ..., 0]$ such that

$$\mathbf{h}_{n,i} = \mathbf{C}_n \mathbf{h}_{n,i-1} + \mathbf{a}_n \mathbf{v}_{n,i}, \tag{13}$$

$$y_{n+1,i} = \mathbf{x}_{n,i} \mathbf{h}_{n,i} + w_{n+1,i} \tag{14}$$

where $\mathbf{v}_{n,i}$ is white Gaussian noise and

$$\mathbf{C}_n = \begin{pmatrix} -c_{n,1} & -c_{n,2} & \cdots & -c_{n,o} \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{15}$$

Considering (14), each N_n can estimate its own channel coefficient in time interval i using the Kalman filter described in Algorithm 1. The algorithm is initialized by considering that no past channel coefficients are available, i.e., $\mathbf{h}_n = \mathbf{0}_o$, where $\mathbf{0}_o$ is a vector of length o full of zeros. Note that in Algorithm 1, $\mathbf{I}_o$ represents the identity matrix of size o and $\mathbf{a}_n^{\mathrm{H}}$ is the conjugate transpose of vector $\mathbf{a}_n$. Furthermore, the estimate $\hat{h}_{n,i}$ of the channel coefficient of N_n in time interval

Algorithm 1 Kalman filter based channel predictor

1: initialize $\mathbf{h}_{n,1} = \mathbf{0}_o$ and set $\mathbf{M}_{n,1} = \mathbf{I}_o$
2: **for** every time interval $i = 1, ..., I$ **do**
3: set $\mathbf{M}_{n,i} = \mathbf{C}_n\mathbf{M}_{n,i-1}\mathbf{C}_n^{\mathrm{H}} + \mathbf{a}_n\mathbf{a}_n^{\mathrm{H}}$
4: set $\Upsilon = \mathbf{x}_{n,i}\mathbf{M}_{n,i}\mathbf{x}_{n,i}^{\mathrm{H}} + \sigma^2$
5: calculate the Kalman gain $\mathbf{k}_{n,i} = \mathbf{M}_{n,i}\mathbf{x}_{n,i}^{\mathrm{H}}/\Upsilon$
6: update $\mathbf{h}_{n,i} = \mathbf{C}_n\mathbf{h}_{n,i-1} + (y_{n,i} - \mathbf{x}_{n,i}\mathbf{C}_{n,i}\mathbf{h}_{n,i-1})\mathbf{k}_{n,i}$
7: update $\mathbf{M}_{n,i} = (\mathbf{I}_o - \mathbf{k}_{n,i}\mathbf{x}_{n,i})\mathbf{M}_{n,i}$
8: obtain $\hat{h}_{n,i} = [1, 0, ..., 0]\mathbf{h}_{n,i}$
9: **end for**

i is given by $\hat{h}_{n,i} = [1, 0, ..., 0]\mathbf{h}_{n,i}$.

Signaling: The purpose of the signaling phase is to allow the nodes to exchange the value of their current parameters in order to observe the current system state S_i. Thus, we consider a transmission scheme which consists of a signaling phase and a data transmission phase. During the signaling phase of duration τ^{Sig}, N_1 transmits $\langle E_{1,i}, B_{1,i}, D_{1,i}\rangle$ and N_2 transmits $\langle E_{2,i}, B_{2,i}, \hat{g}_{2,i}, D_{2,i}\rangle$, where $\hat{g}_{n,i} = |\hat{h}_{n,i}|^2$, for $n = 1, 2$. Note that N_1 does not transmit $\hat{g}_{1,i}$ because $h_{1,i}$, and consequently $g_{1,i}$, are already known at N_2. During the data transmission phase of duration $\tau^{\mathrm{Data}} = \tau - \tau^{\mathrm{Sig}}$, the EH nodes transmit the data stored in their data buffers. To facilitate the coordination among the nodes, we keep τ^{Sig} fixed and in each time interval i, calculate the power $p_{n,i}^{\mathrm{Sig}}$ required for the transmission of the signaling. In the following, we describe how to compute $p_{n,i}^{\mathrm{Sig}}$.

Let $u_{n,i}$ be a variable that represents any parameter associated to N_n, i.e., $u_{n,i} \in \{E_{n,i}, B_{n,i}, \hat{g}_{n,i}, D_{n,i}\}$. Then, the number $Z_{u_{n,i}}$ of bits required for the transmission of each $u_{n,i}$ depends on the type of quantizer that is used. For simplicity, we consider a uniform quantizer. Consequently, $Z_{u_{n,i}}$ depends on the tolerable quantization error $e_{\mathrm{quant},u_{n,i}}$, the maximum value $V_{\mathrm{max},u_{n,i}}$ and the minimum value $V_{\mathrm{min},u_{n,i}}$ each $u_{n,i}$ can take. The number $Z_{u_{n,i}}$ of bits is calculated as

$$Z_{u_{n,i}} = \left\lceil \log_2 \left(\frac{V_{\mathrm{max},u_{n,i}} - V_{\mathrm{min},u_{n,i}}}{e_{\mathrm{quant},u_{n,i}}} \right) - 1 \right\rceil, \tag{16}$$

where $\lceil \cdot \rceil$ is the rounding operation to the next integer value greater than or equal to the evaluated number. Since $V_{\mathrm{max},u_{n,i}}$ and $V_{\mathrm{min},u_{n,i}}$ are assumed to be fixed for each $u_{n,i}$, the number of bits required for signaling is constant for all the time intervals and it is given by $Z_n = \sum_{\forall u_{n,i}} Z_{u_{n,i}}$. Given Z_n, the power $p_{n,i}^{\mathrm{Sig}}$ required to transmit the signaling from N_n to N_m is

$$p_{n,i}^{\mathrm{Sig}} = \frac{\sigma^2}{g_{n,i}} \left(2^{\frac{Z_n}{W\tau^{\mathrm{Sig}}}} - 1 \right). \tag{17}$$

It should be noted that the amount of energy $\tau^{\mathrm{Sig}} p_{n,i}^{\mathrm{Sig}}$ used by each node for the transmission during the signaling phase is deducted from the battery level $B_{n,i}$ and the rest is available for data transmission. Moreover, if for any of the EH nodes the energy in the battery is lower than the value required to send the signaling and the tolerable quantization error is fixed, then the number of parameters sent during the signaling phase is reduced.[2] The order in which this reduction is done is given by the impact each parameter has on the feature functions described in Sec. IV-F and the approximation of the action-value function. First, the transmission of $E_{n,i}$ is skipped. If the energy in the battery is not sufficient, then the transmission of $D_{n,i}$ is skipped as well. Finally, if the energy is still not sufficient, also the transmission of $B_{n,i}$ is skipped. When N_n cannot transmit the signaling, N_m, $m \in \{1, 2\}$, $m \neq n$, assumes that N_n has harvested an amount of energy equal to its own, i.e., $E_{n,i} = E_{m,i}$, and that the signaling was not sent because the battery level of N_n is zero, i.e., $B_{n,i} = 0$. Additionally, since there is no knowledge about the channel gain, it is assumed that $\hat{g}_{n,i} = \hat{g}_{n,i-1}$. For the data buffer level of node N_n, it is assumed that $D_{n,i} = \max\{0, D_{n,i-1} - R_{n,i-1}^{\mathrm{DF}}\}$, where $R_{n,i-1}^{\mathrm{DF}}$ is the number of bits transmitted by N_n in time interval $i - 1$.

[2] Another approach to deal with cases when the energy in the battery is not enough to send the signaling, is to decrease the quality of the quantization.

F. Feature functions

The feature functions used for the linear function approximation exploit the characteristics of the offline solution for the problem in (5). They are defined considering the EH, data arrival and channel fading processes at the EH nodes, as well as the finite size of the batteries and data buffers. For the proposed cooperative SARSA, we consider $F = 6$ binary feature functions. The first four feature functions are defined in our previous work [23,25], and we reproduce them here for readability. The first feature function $\mathrm{f}_1(S_i, p_{n,i}^{\mathrm{Tx}})$ takes into account the energy causality and battery overflow constraints in (5b) and (5c), respectively. It indicates if a given $p_{n,i}^{\mathrm{Tx}}$ avoids the overflow of the battery. Additionally, it evaluates if the given $p_{n,i}^{\mathrm{Tx}}$ fulfills the energy causality constraint.

$$\mathrm{f}_1(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } (B'_{n,i} \leq B_{\max,1}) \wedge \\ & \quad (\tau p_{n,i}^{\mathrm{Tx}} + E_n^{\mathrm{Circ}} \leq B_{n,i}) \\ 0, & \text{else,} \end{cases} \tag{18}$$

where $\wedge$ represents the logical conjunction operation and

$$B'_{n,i} = B_{n,i} + E_{n,i} - \tau p_{n,i}^{\mathrm{Tx}} - E_n^{\mathrm{Circ}}. \tag{19}$$

The second feature function $\mathrm{f}_2(S_i, p_{n,i}^{\mathrm{Tx}})$ addresses the power allocation problem by leveraging a water-filling approach that considers the current channel gain and the mean value $\bar{g}_{n,i}$ of the past channel gains. The use of water-filling is motivated by the water-filling-like characteristic of the offline approach in the EH single hop scenario [37]. As described in [23,25], the water level $\nu_{n,i}$ is calculated as

$$\nu_{n,i} = \frac{1}{2} \left(\frac{B_{n,i} - E_n^{\mathrm{Circ}}}{\tau^{\mathrm{Data}}} + \frac{E_{n,i}}{\tau^{\mathrm{Data}}} + \sigma^2 \left(\frac{1}{\bar{g}_{n,i}} + \frac{1}{g_{n,i}} \right) \right), \tag{20}$$

and the power $p_{n,i}^{\mathrm{Tx}}$ given by the water-filling solution is given by

$$p_{n,i}^{\mathrm{WF}} = \min \left\{ \frac{B_{1,i} - E_1^{\mathrm{Circ}}}{\tau^{\mathrm{Data}}}, \max \left\{ 0, \nu_i - \frac{\sigma^2}{g_{1,i}} \right\} \right\}. \tag{21}$$

As $p_{n,i}^{\mathrm{tx}}$ has to be selected from the set $\mathcal{A}_n$, the second feature function $\mathrm{f}_2(S_i, p_{1,i}^{\mathrm{Tx}})$ is written as

$$\mathrm{f}_2(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \delta \left\lfloor \frac{p_{n,i}^{\mathrm{WF}}}{\delta} \right\rfloor = p_{n,i}^{\mathrm{Tx}} \\ 0, & \text{else,} \end{cases} \tag{22}$$

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 $\mathrm{f}_3(S_i, p_{n,i}^{\mathrm{Tx}})$ handles the case when $E_{n,i} \geq B_{\max,n}$. From (5c), it is clear that in such situations, battery overflow is unavoidable. Therefore, the battery should be depleted in order to minimize the energy losses due to battery overflow. $\mathrm{f}_3(S_i, p_{n,i}^{\mathrm{Tx}})$ is given by

$$\mathrm{f}_3(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } (E_{n,i} \geq B_{\max,n}) \wedge \\ & \left(p_{n,i}^{\mathrm{Tx}} = \delta \lfloor \frac{B_{n,i} - E_n^{\mathrm{Circ}}}{\tau \delta} \rfloor\right) \\ 0, & \text{else.} \end{cases}$$

The fourth feature function $\mathrm{f}_4(S_i, p_{n,i}^{\mathrm{Tx}})$ addresses the data causality and data buffer overflow constraints in (5d) and (5e), respectively. For its definition, let $R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})}$ be the throughput that would be achieved if $p_{n,i}^{\mathrm{Tx}}$ is selected. Then, $\mathrm{f}_4(S_i, p_{n,i}^{\mathrm{Tx}})$ indicates if $R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})}$ fulfils both, the data causality and the data buffer overflow constraints. $\mathrm{f}_4(S_i, p_{n,i}^{\mathrm{Tx}})$ is defined as

$$\mathrm{f}_4(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \left(R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})} \leq D_{n,i}\right) \wedge \\ & \left(D_{n,i} + M_{n,i} - R_{1,i}^{(p_{n,i}^{\mathrm{Tx}})} \leq D_{\max,n}\right) \\ 0, & \text{else.} \end{cases}$$

Additionally, we propose two new feature functions to take into account the knowledge obtained during the signaling phase. Similar to $\mathrm{f}_4(S_i, p_{n,i}^{\mathrm{Tx}})$, these feature functions consider the constraints in (5d) and (5e). The fifth feature function $\mathrm{f}_5(S_i, p_{n,i}^{\mathrm{Tx}})$ takes the available information N_n has about N_m, $n, m \in \{1, 2\}$, $n \neq m$ into consideration and uses it to avoid data buffer overflows at N_2. We focus on the data buffer overflow of N_2 because the data buffer level $D_{2,i}$ depends on the throughput of N_1 and N_2. On the contrary, $D_{n,1}$ depends only on the throughput of N_1 and its data arrival process which we cannot control. Each N_n determines an estimate of the power $\bar{p}_{m,i}^{\mathrm{Tx}}$ to be selected by the other node N_m, $n \neq m$ using the water-filling procedure in (20)-(22). With $\bar{p}_{m,i}^{\mathrm{Tx}}$, the corresponding throughput $R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})}$ is calculated and it is compared to the data buffer level $D_{m,i}$. If $R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})} > D_{m,i}$, then $\bar{p}_{m,i}^{\mathrm{Tx}}$ is scaled down to the minimum power value $\bar{p}_{m,i}^{\mathrm{Tx}} \in \mathcal{A}_m$ that can be used to deplete the data buffer at N_m. The feature function is then defined for $n = 1$ as

$$\mathrm{f}_5(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \left(R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})} + D_{2,i} - R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})} \leq D_{\max,2}\right) \\ & \quad \wedge \left(R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})} + D_{2,i} - R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})} \geq 0\right), \\ & \quad n = \{1, 2\}, n \neq m \\ 0, & \text{else.} \end{cases}$$

In the case $n = 2$, the indices n and m should be interchanged.

The sixth feature function $\mathrm{f}_6(S_i, p_{n,i}^{\mathrm{Tx}})$ aims at the depletion of the data buffers as a preventive measure against data buffer overflows. With this feature function, we push for the selection of higher power values that will reduce the probability of data buffer overflow situations. $\mathrm{f}_6(S_i, p_{n,i}^{\mathrm{Tx}})$ is defined as

$$\mathrm{f}_6(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } p_{n,i}^{\mathrm{Tx}} = \underset{\bar{p}_{n,i}^{\mathrm{Tx}} \in \mathcal{A}_n}{\operatorname{argmin}} \left\{ D_{n,i} - R_{n,i}^{(\bar{p}_{n,i}^{\mathrm{Tx}})} \right\} \\ 0, & \text{else.} \end{cases} \tag{23}$$

Algorithm 2 Cooperative SARSA

1: initialize γ, ζ, ϵ and $\mathbf{w}_n$
2: predict own channel coefficient ▷ Sec. IV-E
3: exchange parameters and observe state S_i ▷ Sec. IV-E
4: select $p_{n,i}^{\mathrm{Tx}}$ using the ϵ-greedy policy ▷ Eq. 24
5: **for** every time interval $i = 1, ..., I$ **do**
6: transmit using the selected $p_{n,i}^{\mathrm{Tx}}$
7: calculate corresponding reward $R_{2,i}^{\mathrm{DF}}$ ▷ Eq. (4)
8: predict own channel coefficient ▷ Sec. IV-E
9: exchange parameters and observe state S_{i+1} ▷ Sec. IV-E
10: select next $p_{n,i+1}^{\mathrm{Tx}}$ using the ϵ-greedy policy ▷ Eq. (24)
11: update $\mathbf{w}_n$ ▷ Eq. (10)
12: set $S_i = S_{i+1}$ and $p_{n,i}^{\mathrm{Tx}} = p_{n,i+1}^{\mathrm{Tx}}$
13: **end for**

G. Action selection policy

To select $p_{n,i}^{\mathrm{Tx}}$, each node follows the ϵ-greedy policy [32], i.e., with probability $1-\epsilon$, node N_n selects the transmit power $p_{n,i}^{\mathrm{Tx}}$ that maximizes $\hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ for a given state S_i. This means,

$$\Pr\left[p_{n,i}^{\mathrm{Tx}} = \max_{p_{n,i}^{\mathrm{Tx}} \in \mathcal{A}_n} \hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})\right] = 1-\epsilon, \quad 0 < \epsilon < 1. \tag{24}$$

Furthermore, with probability ϵ, N_n will randomly select a transmit power value from the set $\mathcal{A}_n$. This method provides a trade-off between the exploration of new transmit power values and the exploitation of the known ones [32].

H. Cooperative SARSA algorithm

The proposed cooperative SARSA algorithm is summarized in Algorithm 2. Note that this algorithm is run at both, N_1 and N_2. First, each N_n initializes the values for the discount factor γ, the learning rate ζ, and the probability ϵ (line 1). Then, the EH node predicts its own channel coefficient (line 2) and exchanges its parameters $E_{n,i}$, $B_{n,i}$, $D_{n,i}$, $g_{n,i}$ during τ^{Sig} in order to observe S_i (line 3). According to S_i and using the $\epsilon-$greedy policy, the node selects its own $p_{n,i}^{\mathrm{Tx}}$ (line 4). After the data transmission phase, the node calculates the obtained reward (line 7), predicts its own next channel coefficient (line 8), and exchanges its updated parameters during the next signaling phase in order to observe the next state S_{i+1} (line 9). Each node selects the new $p_{n,i+1}^{\mathrm{Tx}}$ using the $\epsilon-$greedy policy and updates its weights $\mathbf{w}_n$ (lines 10-11). The same procedure is repeated in every time interval for as long as N_1 and N_2 are operative.

V. Analysis of Cooperative SARSA

A. Convergence guarantees

In this section, we provide convergence guarantees for the proposed cooperative SARSA algorithm for the case when the EH nodes are able to perfectly observe the current system

state, i.e., when the signaling is successfully sent. Furthermore, as the EH, data arrival and channel fading processes might be non-stationary, we consider a constant learning rate ζ_i to ensure that the new obtained rewards are considered in the learning process given by the update of (8) [32]. Inspired by the work of [38], we show that the local action-value function $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is a projection of the centralized action-value function $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ onto the corresponding state-action space $(S_i, p_{n,i}^{\mathrm{Tx}})$. This means, the use of the local action-value function $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ leads to the selection of the transmit power that maximizes the throughput, i.e., the one that would be selected if the centralized action-value function $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ were available.

Proposition 1. Consider an n-player Markov game, which is defined by the tuple $\langle \mathcal{S}, \mathcal{A}_1, ..., \mathcal{A}_n, \mathcal{T}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle$ and where the nodes have the same reward function $\mathcal{R}_1 = ... = \mathcal{R}_n = \mathcal{R}$, $\mathcal{R} \geq 0$. For this game $\mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ and $\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ are the values of the centralized and local action-value function in time interval i, respectively. Moreover the values of $\mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ and $\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ are updated in each time interval using (7) and (8), respectively, and by considering $\zeta_i = 1$. Let $P_i^{(l)}$ bet the l^{th} element in P_i^{Tx} which corresponds to the action of player n in time interval i according to the centralized policy Π. Then, for such Markov game, the equality

$$\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max_{\substack{P_i^{\mathrm{Tx}}=(p_{1,i}^{\mathrm{Tx}},\ldots,p_{n,i}^{\mathrm{Tx}}) \\ P_i^{(l)}=p_{n,i}^{\mathrm{Tx}}}} \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}), \tag{25}$$

holds for any player n, any S_i, and any individual action $p_{n,i}^{\mathrm{Tx}}$ in time interval i.

Proof. As in [38], the proof is done by induction on i. At $i = 1$, no reward has been obtained. Therefore, Q^{Π} and $\mathrm{q}_n^{\pi_n}$ are zero for every state $S_1 \in \mathcal{S}$ and $p_{n,1}^{\mathrm{Tx}} \in \mathcal{A}_n$, $n \in \{1, ..., n\}$ and (25) holds. For arbitrary i, (25) holds for any pair $(S_j, p_{m,j}^{\mathrm{Tx}})$, $S_j \neq S_i$, $p_{m,j}^{\mathrm{Tx}} \neq p_{n,i}^{\mathrm{Tx}}$ and $n \neq m$, because the updates in (7) and (8) are only performed on the particular pair $(S_i, p_{n,i}^{\mathrm{Tx}})$. Now, to prove (25) for the pair $(S_i, p_{n,i}^{\mathrm{Tx}})$, we include the right side of (25) in the update of $\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ in (8) as

$$\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max \left\{ \begin{array}{l} \max\limits_{\substack{P_i^{\mathrm{Tx}} \\ P_i^{(l)}=p_{n,i}^{\mathrm{Tx}}}} \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}), \\ \\ R_i + \gamma \max\limits_{P_{i+1}^{\mathrm{Tx}}} \mathrm{Q}_i^{\Pi}(S_{i+1}, P_{i+1}^{\mathrm{Tx}}) \end{array} \right\}. \tag{26}$$

By considering the equality $\max\{\mathrm{f}(x) + a\} = a + \max\{\mathrm{f}(x)\}$, (26) can be rewritten as shown in (26). From (7), it is clear that the second term on the right side of (26) corresponds to the centralized action-value function $\mathrm{Q}_{i+1}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$. Therefore, assuming enough exploration has already been made such that P_{i+1}^{Tx} is selected by acting greedily with respect to Q_i^{Π}, we can rewrite (26) as in (27). Now, by expanding the term on the right side of (27), we obtain the expression in (28). The first term on the right side of (28) is equal to $\mathrm{Q}_{i+1}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ because for $P_j^{\mathrm{Tx}} \neq P_i^{\mathrm{Tx}}$ there is no update. The second term is always smaller than or equal to $\mathrm{Q}_{i+1}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ because, as the rewards are always greater than or equal to zero, $\mathrm{Q}^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ is monotonically increasing. With this in mind, $\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is then written as in (29). □

B. Computational complexity analysis

In this section, we evaluate the computational complexity of one iteration of the proposed cooperative SARSA algorithm. For this purpose, we use the $O(\cdot)$ notation. By examining Algorithm 2, it is clear that the most computationally demanding tasks are the estimation of the channel coefficients (Lines 2 and 7), the selection of the transmit power $p_{n,i}^{\mathrm{Tx}}$ (Lines 3 and 8) and the update of $\mathbf{w}_n$ (Line 9). The complexity of the Kalman-filter based channel estimator scales as $O(o^3)$ [39], where o is the order of the filter. Furthermore, for the selection of $p_{n,i}^{\mathrm{Tx}}$, the ϵ-greeedy policy is considered. In this case, the highest complexity is due to the calculation of $\mathrm{q}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ for all the possible actions and the selection of the $p_{n,i}^{\mathrm{Tx}}$ that leads to the maximum $\mathrm{q}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$. The computational complexity for the calculation of $\mathrm{q}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is $O(|\mathcal{A}|F)$ while the selection of the maximum value scales as $O(|\mathcal{A}|)$. Lastly, the update of $\mathbf{w}_n$ using (10) has a complexity of $O(F^2)$. As in our model o is fixed, the computational complexity of one iteration of the algorithm scales linearly with $|\mathcal{A}|$ and polynomially with the number of feature functions F as $O(2|\mathcal{A}|F + F^2)$. In our proposed cooperative SARSA, $F = 6$ and usually $|A| >> F$, e.g., $|A| \approx 100$ when a step size $\delta = 2\%$ is considered. This means, the leading factor in the computational complexity of the proposed cooperative SARSA is $|\mathcal{A}|$. The extra factor $2F$ in the expression of the complexity, which is caused by the use of the linear function approximation, is the price to be paid for the improvement in the performance compared to reference schemes. An additional advantage of the iterative nature of our proposed cooperative SARSA is that it reduces the memory requirements on the system compared to traditional learning approaches. Note that even though a continuous state is considered, the use of linear function approximation causes that only the vector of weights needs to be stored in addition to the vector of features used to describe the state in time interval i.

VI. Performance evaluation

In this section, we present numerical results for the evaluation of the proposed cooperative SARSA. For the simulations, the parameters listed in Table II are considered, unless it is otherwise specified.

It is assumed that $E_{n,i}$ at time interval i is taken from a uniform distribution with maximum value $E_{\text{max,n}}$. We consider solar energy as our EH source with an average power density $\rho = 10\text{mW/cm}^2$ and an EH panel size $\Omega = 16\text{cm}^2$ [3]. Consequently, $E_{\max,n} = 2\rho\Omega\tau$.

We define the average signal to noise ratio (SNR), denoted by Γ, as the ratio between the average power of the received signal and the noise at the receiver as $\Gamma = \frac{\rho\Omega\bar{g}_n}{\sigma^2} = 5\text{dB}$, where $\bar{g}_n$ is the average channel gain on the link between N_n and N_{n+1}. The channel coefficients are modeled as complex

$$\mathrm{q}^{\pi_n}_{n,i+1}(S_i,p^{\mathrm{Tx}}_{n,i}) = \max\left\{\max_{\substack{P_i^{\mathrm{Tx}}\\ P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}}} \mathrm{Q}^{\Pi}_i(S_i,P_i^{\mathrm{Tx}}), \max_{P_{i+1}^{\mathrm{Tx}}}\left\{R_i+\gamma \mathrm{Q}^{\Pi}_i(S_{i+1},P_{i+1}^{\mathrm{Tx}})\right\}\right\}. \tag{26}$$

$$\mathrm{q}^{\pi_n}_{n,i+1}(S_i,p^{\mathrm{Tx}}_{n,i}) = \max\left\{\max_{\substack{P_i^{\mathrm{Tx}}\\ P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}}} \mathrm{Q}^{\Pi}_i(S_i,P_i^{\mathrm{Tx}}), \mathrm{Q}^{\Pi}_{i+1}(S_i,P_i^{\mathrm{Tx}})\right\}. \tag{27}$$

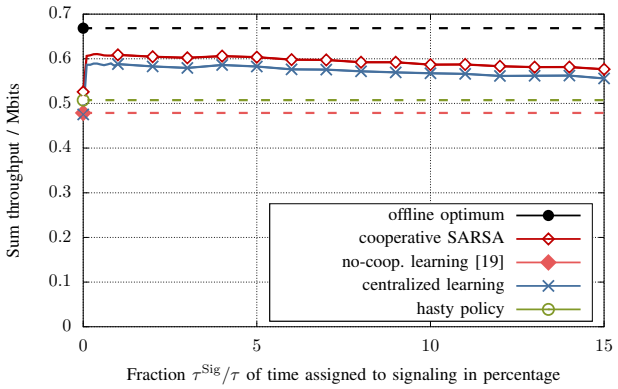
$$\begin{aligned}\mathrm{q}^{\pi_n}_{n,i+1}(S_i,p^{\mathrm{Tx}}_{n,i}) =& \max\Big\{\Big\{\mathrm{Q}^{\Pi}_i(S_i,P_i^{\mathrm{Tx}}) \mid P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}, P_j^{\mathrm{Tx}}\neq P_i^{\mathrm{Tx}}\Big\}\cup\\ &\Big\{\mathrm{Q}^{\Pi}_i(S_i,P_i^{\mathrm{Tx}}) \mid P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}, P_j^{\mathrm{Tx}}= P_i^{\mathrm{Tx}}\Big\}\cup\big\{\mathrm{Q}^{\Pi}_{i+1}(S_i,P_i^{\mathrm{Tx}})\big\}\Big\}.\end{aligned} \tag{28}$$

$$\begin{aligned}\mathrm{q}^{\pi_n}_{n,i+1}(S_i,p^{\mathrm{Tx}}_{n,i}) =& \max\Big\{\Big\{\mathrm{Q}^{\Pi}_{i+1}(S_i,P_i^{\mathrm{Tx}}) \mid P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}, P_j^{\mathrm{Tx}}\neq P_i^{\mathrm{Tx}}\Big\}\cup\big\{\mathrm{Q}^{\Pi}_{i+1}(S_i,P_i^{\mathrm{Tx}})\big\}\Big\}\\ =& \max_{\substack{P_i^{\mathrm{Tx}}=(p^{\mathrm{Tx}}_{1,i},\ldots,p^{\mathrm{Tx}}_{n,i})\\ P_i^{(l)}=p^{\mathrm{Tx}}_{n,i}}} \mathrm{Q}^{\Pi}_{i+1}(S_i,P_i^{\mathrm{Tx}}).\end{aligned} \tag{29}$$

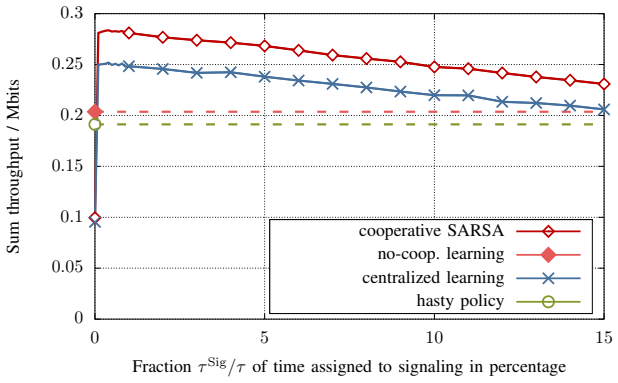
Gaussian processes using the model described in [40]. To compare the performance of the proposed cooperative SARSA, we consider the following reference schemes:

- Offline optimum: It assumes that a central entity has perfect non-causal knowledge of the EH, data arrival and channel fading processes and solves the optimization problem in (5).
- No-Cooperation Learning [25]: This approaches assumes the nodes have only causal knowledge of their own states. No cooperation between the nodes is exploited and each node aims at maximizing its own throughput.
- Centralized Learning: Using the signaling phase to observe the system state, a centralized RL problem is considered in which N_2 decides jointly on the transmit powers of N_1 and N_2. Note that this approach also considers the use of Kalman filter based channel estimators at the nodes in order to obtain an estimate of the current channel coefficients.
- Hasty policy: This approach depletes the battery of N_1 in each time interval to transmit the maximum possible amount of data to N_2. At N_2, the policy aims at depleting the data buffer by selecting the maximum transmit power value that fulfills the data causality constraint.

In Figures 2(a) and 2(b), we compare the average sum throughput, i.e., the amount of data received by N_3, measured in bits, for different values of the fraction τ^{Sig}/τ of the duration of the time interval assigned for the signaling phase, considering an infinitely full data buffer at N_1. In this case, we have reduced the number of time intervals to $I = 100$ in order to be able to calculate the offline optimum as a reference for the case when $E_n^{\mathrm{Circ}} = 0$. Moreover, the offline optimum, no-cooperation learning and hasty policy approaches are depicted with dashed lines because they do not consider a signaling phase and use the complete duration τ of the time interval for the transmission of data. Consequently, they are only defined for the value $\tau^{\mathrm{Sig}}/\tau = 0$. Figure 2(a) considers that $E_n^{\mathrm{Circ}} = 0$ and as expected, the largest throughput is achieved by the offline optimum approach which provides the upper bound of the performance assuming perfect non-causal knowledge



(a) $E_n^{\mathrm{Circ}} = 0$



(b) $E_n^{\mathrm{Circ}} = 1\mathrm{mJ}$

Fig. 2. Sum throughput versus fraction of time τ^{Sig}/τ assigned to signaling.

TABLE II. Simulation set-up.

	Parameter	Value	Description
General	$e_{\mathrm{quant},u_{n,i}}$	1%	Quantization error
	I	1000	Number of time intervals
	T	1000	Number of realizations
	o	2	Order of the autoregressive process
	τ	10ms	Time interval duration
	τ^{Sig}	0.1ms	Signaling phase duration
Energy	$B_{\max,n}$	$\varsigma E_{\max,n}$	Battery capacity of EH node N_n
	E_n^{Circ}	1mJ	Energy consumed by the circuit of EH node N_n
	ρ	10mW/cm^2	Power density of the EH source
	ς	5	Battery size factor for EH nodes N_n
	Ω	16cm^2	Size of EH panel
Data	d	10 kbit	Packet size
	$D_{\max,1}$	50kbits	Data buffer size of EH node N_1
	$D_{\max,2}$	$W\tau\log_2(1+\Gamma)$	Data buffer size of EH node N_2
	λ	10	Average number of packets arriving per time interval
Channel	f_0	2.4 GHz	Carrier frequency
	W	1 MHz	Bandwidth
	α	3	Path loss exponent
	Γ	5dB	Average SNR per link
Learning	γ	0.9	Discount factor
	δ	2%	Step size
	ϵ	$1/i$	Exploration probability
	ζ	$1/i$	Learning rate

of the system dynamics. The achieved throughput of the cooperative SARSA and the centralized learning depends on the time assigned for the signaling. For $\tau^{\mathrm{Sig}}/\tau < 15\%$, the cooperative SARSA outperforms the other approaches which also consider only causal knowledge. The reason for this improvement is that by including the signaling phase, N_1 and N_2 overcome the partial observability of the system state and are able to learn a transmission policy that adapts to the battery levels, data buffer levels and channel gains of both nodes. Moreover, the cooperative SARSA outperforms the centralized approach because in a distributed solution, a smaller action space needs to be considered, which increases the learning speed. In Figure 2(a), the largest throughput of the cooperative SARSA is achieved at approximately $\tau^{\mathrm{Sig}}/\tau = 0.3\%$. For $\tau^{\mathrm{Sig}}/\tau < 0.3\%$, the throughput is reduced because, as shown in (17), the relation between τ^{Sig} and $p_{n,i}^{\mathrm{Sig}}$ required to transmit the signaling is not linear and the smaller τ^{Sig}, the over-proportionally larger $p_{n,i}^{\mathrm{Sig}}$. As $p_{n,i}^{\mathrm{Sig}}$ increases, the probability of not having enough energy in the battery to fulfill this requirement increases. Consequently, the nodes do not have enough energy to transmit during the signaling phase and to exchange their causal knowledge. When τ^{Sig}/τ increases to values beyond 0.3%, the achieved throughput slowly decreases. Even though for increasing values of τ^{Sig}/τ, the EH nodes have a longer signaling phase to exchange their causal knowledge, and can therefore use less power for the transmission of the signaling and save energy for data transmission, less time is left for the transmission of data. As a result, the power required to transmit a certain amount of data increases.

In Figure 2(b), the energy E_n^{Circ} consumed by the circuit is considered. In this case, the offline optimum is not included because for such scenario, the feasibility cannot be guaranteed. When $E_n^{\mathrm{Circ}} \neq 0$, the throughput of all the approaches is reduced because less energy is available for data transmission compared with the case when $E_n^{\mathrm{Circ}} = 0$. Note that all the learning approaches outperform the hasty policy. This is because they take into account the energy consumed by the circuit when allocating the power. However, as the cooperative SARSA and the centralized learning approaches are able to overcome the partial observability of the system state, their corresponding achieved throughput is higher compared to the one achieved by the other schemes. Specifically, for $\tau^{\mathrm{Sig}}/\tau = 1\%$, the cooperative SARSA approach achieves a throughput which is 17% larger than for the centralized approach, 42% larger than for the no-cooperation learning approach and 51% larger than for the hasty policy.

The number of data buffer overflows at N_2 versus the data buffer size of the EH relay N_2 is shown in Figure 3. To evaluate different values of the data buffer size at N_2, we consider the data buffer size factor β and calculate $D_{\max,2} = W\tau\log_2(1+\beta\Gamma)$ and an infinitely full data buffer at N_1. Note that the result of the offline optimum is omitted because the feasibility of the optimization problem cannot be guaranteed for all the considered data buffer sizes. It can be seen that, as the data buffer size increases, the number of data buffer overflows is reduced for all the approaches, as expected. For $\beta = 1$, the cooperative SARSA approach has 22% less data buffer overflows than the centralized learning approach, 44% less than the no-cooperation learning approach and 43% less than the hasty policy. The better performance of the cooperative SARSA results from the fact that by exchanging the causal knowledge during the signaling phase, N_1 knows the data buffer level of N_2 and can limit the amount of transmitted data when the data buffer of N_2 is almost full. It should be noted that although the cooperative SARSA is able to significantly reduce the number of data buffer overflows, it cannot reduce it to zero. This is because non-causal knowledge would be required to adapt the transmission policy according to the amounts of energy that will be harvested as well as the future channel gains.

Figure 4 shows the impact of the data arrival process at N_1. For this simulation, we consider that the data arrival process at N_1 consists of an average number λ of data packets arriving in each time interval i. We assume that the number of packets arriving is taken from a Poisson distribution with

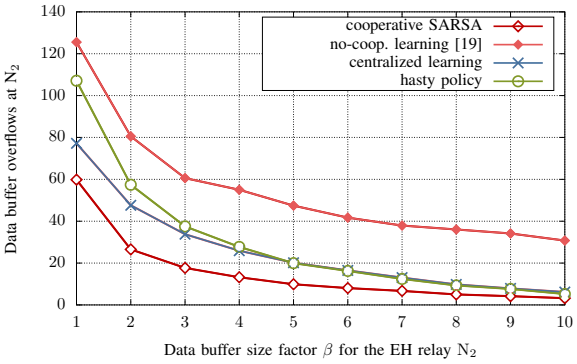


Fig. 3. Number of data buffer overflows at N_2 versus the data buffer size factor β.

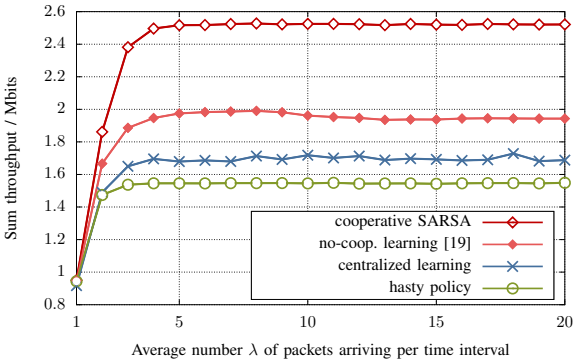


Fig. 4. Sum throughput versus the average number λ of incoming data packets.

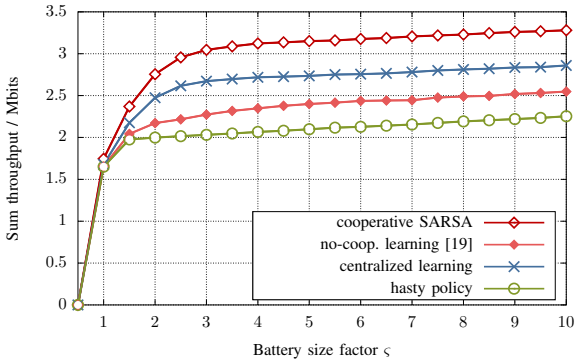


Fig. 5. Sum throughput versus the battery size factor ς.

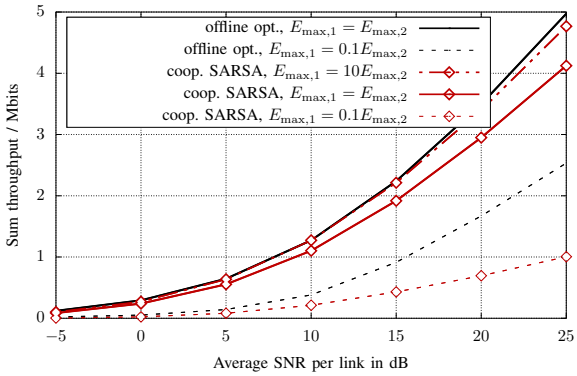


Fig. 6. Sum throughput versus average SNR per link.

parameter λ, and consider a packet size of 10kbit. The offline optimum policy is not considered because the feasibility of the optimization problem depends on each particular realization of the data arrival process. In Figure 4, it can be seen that for $\lambda = 1$, all the approaches achieve almost the same performance. This is because for $\lambda = 1$, the data buffer is almost empty all the time. Therefore, data buffer overflows are unlikely and the data packets received by N_1 can be retransmitted by N_2 to N_3. As the number of data packets received per time interval increases, the cooperative SARSA outperforms the reference approaches because it prevents data buffer overflows at N_2, as previously observed in Figure 3. In this case, the performance of the centralized learning is further decreased because the consideration of the state of the data buffer at N_1 increases the dimensions of the state-action space and reduces the learning speed. As a result, the centralized approach ends up in a local maximum.

The impact of the battery size on the achieved throughput is evaluated in Figure 5. As expected, the cooperative SARSA approach outperforms the reference schemes when $B_{\max,n} > E_{\max,n}$, i.e., $\varsigma > 1$. For $\varsigma = 5$, it is able to achieve a throughput 30% higher than the no-cooperation learning approach. Moreover, its performance is 13% and 47% higher than for the centralized approach and for the hasty policy, respectively.

In Figure 6, we compare the performance of the offline optimum policy and the cooperative SARSA as a function of the average SNR per link, i.e., from N_1 to N_2 and from N_2 to N_3. Note that the no-cooperation learning approach is not considered because, as it can be observed in the previous results, the cooperative SARSA approach consistently outperforms it. To be able to calculate the throughput achieved by the offline optimum, $I = 100$ time intervals and $E^{\mathrm{Circ}} = 0$ are considered. We additionally evaluate the effect of the maximum amount of energy which N_1 and N_2 can harvest. For this purpose, we consider three different cases, i.e., $E_{\max,2} = 10E_{\max,1}$, $E_{\max,2} = E_{\max,1}$ and $E_{\max,2} = 0.1E_{\max,1}$. For the first case, i.e. $E_{\max,2} = 10E_{\max,1}$, the offline optimum policy cannot be applied because battery overflows cannot be avoided at N_2 when it harvests much more energy than N_1. This is due to the fact that N_2 has more energy available in its battery than what is needed to retransmit the data it receives from N_1. To allow battery overflows at N_2, a different optimization problem would need to be considered. In all the three cases, the throughput increases when the average SNR increases. The largest throughput is achieved by the cooperative SARSA for the case when $E_{\max,2} = 10E_{\max,1}$ and this throughput is close to the offline optimum performance for $E_{\max,2} = E_{\max,1}$. This is because harvesting more energy at N_2 cannot lead to a larger throughput if the amount of harvested energy is not increased

at N_1. The throughput is limited by the amount of data N_1 can transmit which in turn is limited by the amount of energy N_1 harvests, which for the two cases, $E_{\max,2} = 10E_{\max,1}$ and $E_{\max,2} = E_{\max,1}$, is in a similar order of magnitude. For $E_{\max,2} = E_{\max,1}$, the performance of the cooperative SARSA is reduced compared to the case when $E_{\max,2} = 10E_{\max,1}$. This is because there is less energy available at N_2. As a result, in each time interval, N_2 allocates less energy for data transmission. For the case when $E_{\max,2} = 0.1E_{\max,1}$, the performance of the cooperative SARSA is close to the performance of the offline optimum policy in the low SNR regime, i.e., SNR $<$ 10dB. This is due to the fact that in this case, N_2 is the bottleneck because it harvests on average much less energy than N_1. Both approaches, the offline optimum policy and the cooperative SARSA, limit the amount of data N_1 transmits while aiming at maximizing the throughput in each time interval.

Finally, in Figure 7, we evaluate the convergence of the proposed learning approaches. For this purpose, we compare the average throughput per time interval versus the number I of time intervals. In addition to the cooperative SARSA, the centralized approach and the no-cooperation learning approach, we evaluate the performance of the proposed feature functions by implementing the cooperative SARSA using two standard approximation techniques, namely, fixed sparse representation (FSR) and radial basis functions (RBF) [41]. Both, FSR and RBF are low-complexity techniques used to represent the continuous states. For each N_n, $n \in \{1, 2\}$, the state S_i, observed after the signaling phase, lies in an 8-dimensional space given by the parameters $E_{n,i}$, $B_{n,i}$, $g_{n,i}$ and $D_{n,i}$ of both nodes. 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 [41]. In our implementation, the tiles are generated by quantizing each dimension using the step size δ used in the definition of the action spaces $\mathcal{A}_n$. In RBF, each feature function has a Gaussian shape that depends on the distance between a given state and the center of the feature [32,41]. In contrast to FSR, in RBF a given state is represented by more than one feature function. In Figure 7, it can be seen that the cooperative SARSA, the centralized approach and the no-cooperation learning approach converge at approximately the same number of iterations. This is due to the fact that the three approaches are based on the SARSA update. However, since the cooperative SARSA considers the full cooperation among the EH nodes to exchange their causal knowledge, it can achieve a larger throughput. Note that the number of feature functions required by a learning approach impacts the performance. This is due to the fact that by increasing the number of feature functions used to represent the state space, a larger amount of weights have to be learned. Consequently, the cooperative SARSA approach outperforms FSR and RBF because they require a larger number of feature functions compared to the cooperative SARSA which only needs six.

To summarize the simulation results, it can be seen that with a proper selection of τ^{Sig}, the cooperative SARSA, which considers cooperation between the EH nodes, outperforms other approaches which also only consider causal knowledge but without cooperation between the nodes. This means that reserving a fraction of time for the exchange of signaling among the nodes is more beneficial than assuming no cooperation at all, even though the time dedicated to data transmission is reduced in order to include the signaling phase.Furthermore, the cooperative SARSA reduces the number of data buffer overflows at N_2 as compared to the other approaches. This implies a reduction in the number of required retransmissions.

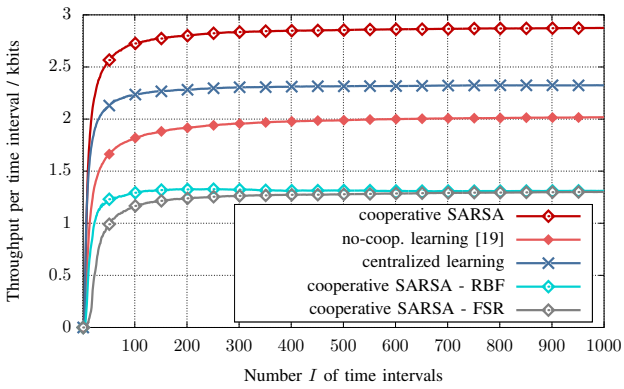


Fig. 7. Throughput per time interval versus the number I of time intervals.

VII. Conclusion

We have investigated an EH two-hop communication scenario where only partial causal knowledge regarding the EH processes, the data arrival processes and the channel fading processes was assumed at the EH transmitter and at the EH relay. We considered the case when a signaling phase is available in each time interval. This signaling phase is used by the EH nodes to cooperate with each other by exchanging their own causal knowledge. After the signaling phase, the EH nodes exploit the obtained knowledge to find transmission policies which adapt to the battery levels, data buffer levels and channel gains of the EH nodes and which aim at maximizing the throughput. We modeled the problem as a Markov game and proposed a multi-agent RL algorithm to find the transmission policies at the transmitter and at the relay. Furthermore, we have provided convergence guarantees for the proposed algorithm. Through several simulation results we have shown that a larger throughput can be achieved when cooperation among the EH nodes is considered, compared to the case when no cooperation is assumed even after the signaling overhead is subtracted from the number of bits transmitted. Moreover, we have shown the trade-off between the duration of the signaling phase and the performance of the proposed algorithm and we have shown that the number of data buffer overflows is reduced when our proposed algorithm is considered. The distributed nature of our proposed algorithm makes it suitable for more complex relay networks, e.g., multi-hop networks.

REFERENCES

[1] W. Dargie and C. Poellabauer, *Fundamentals of Wireless Sensor Networks: Theory and Practice*. John Wiley & Sons, 2010.

[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] M.-L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 2, pp. 1384–1412, November 2016.

[4] D. Gündüz, A. Yener, A. Goldsmith, and H. V. Poor, "The multi-way relay channel," *IEEE Trans. Inform. Theory*, vol. 59, no. 1, pp. 51–63, January 2013.

[5] E. Yilmaz, R. Zakhour, D. Gesbert, and R. Knopp, "Multi-pair two-way relay channel with multiple antenna relay station," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Cape Town, May 2010, pp. 1–5.

[6] O. Orhan and E. Erkip, "Optimal transmission policies for energy harvesting two-hop networks," in *Proc. Annual Conf. Inform. Sciences Syst. (CISS)*, Princeton, March 2012, pp. 1–6.

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

[8] A. A. Nasir, X. Zhou, S. Durrani, and R. A. Kennedy, "Relaying protocols for wireless energy harvesting and information processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3622–3636, July 2013.

[9] O. Orhan and E. Erkip, "Throughput maximization for energy harvesting two-hop networks," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Istanbul, July 2013, pp. 1596–1600.

[10] ——, "Energy harvesting two-hop communication networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2658–2670, December 2015.

[11] Y. Luo, J. Zhang, and K. B. Letaief, "Optimal scheduling and power allocation for two-hop energy harvesting communication systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4729–4741, September 2013.

[12] B. Varan and A. Yener, "Two-hop networks with energy harvesting: The (non-)impact of buffer size," in *Proc. IEEE Global Conf. Signal Inform. Process. (GlobalSIP)*, Austin, December 2013, pp. 399–408.

[13] Y. Zeng and R. Zhang, "Full-duplex wireless-powered relay with self-energy recycling," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 201–204, April 2015.

[14] A. Zanella, A. Bazzi, and B. M. Masini, "Analysis of cooperative systems with wireless power transfer and randomly located relays," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, June 2015, pp. 1–6.

[15] Y. Liu, "Wireless information and power transfer for multirelay-assisted cooperative communication," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 784–787, April 2016.

[16] L. Tang, X. Zhang, and X. Wang, "Joint data and energy transmission in a two-hop network with multiple relays," *IEEE Commun. Lett.*, vol. 18, no. 11, pp. 2015–2018, September 2014.

[17] B. Gurakan, O. Ozel, J. Yang, and S. Ulukus, "Energy cooperation in energy harvesting communications," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4884–4898, December 2013.

[18] M. Rezaee, M. Mirmohseni, V. Aggarwal, and M. R. Aref, "Optimal transmission policies for multi-hop energy harvesting systems," *IEEE Trans. Green Commun. and Networking*, vol. 2, no. 3, pp. 751–763, March 2018.

[19] A. Minasian, S. ShahbazPanahi, and R. S. Adve, "Energy harvesting cooperative communication systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6118–6131, November 2014.

[20] F. Amirnavaei and M. Dong, "Online power control for cooperative relaying with energy harvesting," in *Proc. Asilomar Conf. Signals, Syst. Computers*, Pacific Grove, November 2015, pp. 817–822.

[21] M. Dong, W. Li, and F. Amirnavaei, "Online joint power control for two-hop wireless relay networks with energy harvesting," *IEEE Trans. Signal Proces.*, vol. 66, no. 2, pp. 463–478, February 2018.

[22] M. K. Sharma and C. R. Murthy, "Distributed power control for multi-hop energy harvesting links with retransmission," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4064–4078, June 2018.

[23] 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, December 2016, pp. 1–7.

[24] V. Hakami and M. Dehghan, "Distributed power control for delay optimization in energy harvesting cooperative relay networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4742–4755, September 2016.

[25] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting decode-and-forward two-hop communications," *IEEE Trans. Green Commun. and Networking*, vol. 1, no. 3, pp. 309–319, September 2017.

[26] J. Gong, X. Chen, and M. Xia, "Transmission optimization for hybrid half/full-duplex relay with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3046–3058, February 2018.

[27] V. Rajendran, K. Obraczka, and J. J. Garcia-Luna-Aceves, "Energy-efficient, collision-free medium access control for wireless sensor networks," *Wireless Networks*, vol. 12, no. 1, pp. 63–78, February 2006.

[28] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. Int. Conf. Machine Learning*, New Brunswick, July 1994, pp. 157–163.

[29] A. Arafa and S. Ulukus, "Optimal policies for wireless networks with energy harvesting transmitters and receivers: Effects of decoding costs," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2611–2625, December 2015.

[30] M. L. Littman, "Value-function reinforcement learning in Markov games," *J. Cognitive Syst. Research*, vol. 2, no. 1, pp. 55–66, October 2001.

[31] N. Instruments, "National Instruments Specification USRP-2954," May 2017. [Online]. Available: http://www.ni.com/pdf/manuals/375725c.pdf

[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2018.

[33] W. C. Jakes, *Microwave Mobile Communications*. Wiley-IEEE Press, 1974.

[34] B. Y. Shikur and T. Weber, "Channel prediction using an adaptive Kalman filter," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Ilmenau, March 2015, pp. 1–7.

[35] W. Chen and R. Zhang, "Kalman-filter channel estimator for OFDM systems in time and frequency-selective fading environment," in *Proc. IEEE Int. Conf. Acoust,, Speech, Signal Process. (ICASSP)*, Montreal, May 2004, pp. 377–380.

[36] M. McGuire and M. Sima, "Low-order Kalman filters for channel estimation," in *Proc. IEEE Pacific Rim Conf. Commun., Computers and Signal Process. (PACRIM)*, Victoria, August 2005, pp. 1–4.

[37] 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.

[38] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. Int. Conf. Machine Learning*, Stanford, June 2000, pp. 535–542.

[39] F. Daum, "Nonlinear Filters: Beyond the Kalman Filter," *IEEE Aerospace and Electronic Syst. Mag.*, vol. 20, no. 8, pp. 57–69, August 2005.

[40] A. Kühne, "Analysis of hybrid adaptive/non-adaptive multi-user ofdma systems with imperfect channel knowledge," Ph.D. dissertation, Technische Universität Darmstadt, Darmstadt, April 2011.

[41] 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," *Found. and Trends in Mach. Learning*, vol. 6, no. 4, pp. 375–454, December 2013.

Dr. Andrea Ortiz (S'14 M'20) received the Master degree in Information and Communication Engineering and Dr.-Ing. (Ph.D.) degree in Electrical Engineering from Technische Universität Darmstadt, Darmstadt, Germany. Currently she is post-doctoral researcher at the Communications Engineering Lab, Technische Universität Darmstadt, Germany. Her research interests include reinforcement learning for wireless communications, signal processing for wireless communications and energy harvesting communications.

Tobias Weber received the Dipl.-Ing. degree in electrical engineering, and the Ph.D. and Habilitation degrees from the University of Kaiserslautern, Kaiserslautern, Germany, in 1996, 1999, and 2003, respectively. From 1996 to 2005, he was a Member of the Staff of the Research Group for RF Communications, University of Kaiserslautern. From 1996 to 1999, he was active in the development of a hardware demonstrator for a 3rd generation mobile radio system, where his work focused on future signal processing concepts. In 2005, he became a Professor of Microwave Technology with the University of Rostock, Rostock, Germany. His research interests include future mobile radio systems, OFDM mobile radio systems, MIMO techniques, and localization techniques. He is a member of Verband Deutscher Elektrotechniker—Informationstechnische Gesellschaft (VDE/ITG) and a senior member of IEEE.

Prof. Dr.-Ing. Anja Klein (M'96) received the Diploma and Dr.-Ing. (Ph.D.) degrees in electrical engineering from the University of Kaiserslautern, Germany, in 1991 and 1996, respectively. In 1996, she joined Siemens AG, Mobile Networks Division, Munich and Berlin. She was active in the standardization of third generation mobile radio in ETSI and in 3GPP, for instance leading the 3GPP RAN1 TDD group. She was director of a development department and a systems engineering department. In 2004, she joined the Technische Universität Darmstadt, Germany, as full professor, heading the Communications Engineering Laboratory. Her main research interests are in mobile radio, including interference management, cross-layer design, relaying and multi-hop, computation offloading, smart caching and energy harvesting. Dr. Klein has authored over 280 peer-reviewed papers and has contributed to 12 books. She is inventor and co-inventor of more than 45 patents in the field of mobile communications. In 1999, she was named the Inventor of the Year by Siemens AG. She is a member of Verband Deutscher Elektrotechniker - Informationstechnische Gesellschaft (VDE-ITG).