Optimization and Learning Approaches for Energy Harvesting Wireless Communication Systems

Vom Fachbereich 18
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von
M.Sc. Andrea Patricia Ortiz Jimenez
geboren am 18.03.1986 in Barranquilla-Kolumbien

Referent:	Prof. Dr.-Ing. Anja Klein
Korreferent:	Dr. Deniz Gündüz
Tag der Einreichung:	06.08.2019
Tag der mündlichen Prüfung:	31.10.2019

D 17
Darmstädter Dissertation

The work of Andrea Patricia Ortiz Jimenez has been funded by the *LOEWE initiative (Hessen-Germany)* within the priority program *Networked Infrastructureless Cooperation for Emergency Response (NICER)* as part of project P05-B. (`https://www.nicer.tu-darmstadt.de`)

Ortiz Jimenez, Andrea Patricia : Optimization and Learning Approaches for Energy Harvesting Wireless Communication Systems
Darmstadt, Technische Universität Darmstadt,
Jahr der Veröffentlichung der Dissertation auf TUprints: 2019
URN: urn:nbn:de:tuda-tuprints-93006
Tag der mündlichen Prüfung: 31.10.2019

Acknowledgments

With these lines I would like to thank all the people who, with their advice, encouragement, critique, help and support, made this thesis possible.

I would like to thank Prof. Dr.-Ing. Anja Klein for giving me the opportunity to join her group to pursue this research. I am extremely grateful for her mentoring, her interest on my development as a scientist and all the support, specially after Santi's arrival. Moreover, I am thankful for her keen insight and constructive criticism which have been an uninterrupted source of personal and professional growth.

Next, I would like to thank Prof. Dr.-Ing. habil. Tobias Weber from Universität Rostock for all the fruitful discussions and valuable feedback regarding the different scenarios considered in this thesis. Special thanks to Dr. Deniz Gündüz from Imperial College in London for agreeing to be the second referee of this thesis.

I am grateful for the great work atmosphere of the KT group. Thanks to Lioba Fischer for her support regarding all the administrative issues that arose during the last years. Many thanks to my colleagues, the former and the new ones, for making this a really fun journey. Mousie, Sabrina, Fabian, Daniel, Alex, Alexey, Mahdi and Hussein, thanks for all the good times, the technical discussions and non-technical conversations. I enjoyed them a lot. To the new ones, Tobias, Killian, Bernd, Laszlon, Weskley and Jaime, thanks for the support during the last months of preparing the thesis and defense.

A nivel personal, quiero dar primero gracias a Dios por la vida y salud que me ha dado. A nuestros amigos de este lado del charco, Alfonso Josefo, Angela María, Jhoncito, Astridcita, Mercedes, Zeeshan y Zoey, gracias por todo el apoyo que nos han brindado a Dani, a Santi y a mi. Sin duda, tenerlos a nuestro lado ha hecho que el estar tan lejos de la casa sea mas llevadero. A mis amigos en Colombia, Pedro y Primo gracias por su amistad. A pesar de la distancia saben que tienen un lugar especial en mi corazón. A mi tío Rodo y a mi tía Mayo, gracias por su apoyo para poder venir a estudiar a Alemania. A Daniela, gracias por creer en mi, eres la mejor hermana del mundo mundial. A mi papá y a mi mamá, gracias por apoyarme siempre y quererme. Ustedes son la base de lo que soy y cada logro mío es también de ustedes. Finalmente, gracias a mis amados Dani y Santi por todo. Gracias por los talk-shows, la paciencia, el tiempo y la comprensión. Sin ustedes esta tesis no hubiera sido posible. Es tanto mía como de ustedes.

Abstract

Emerging technologies such as Internet of Things (IoT) and Industry 4.0 are now possible thanks to the advances in wireless sensor networks. In such applications, the wireless communication nodes play a key role because they provide the connection between different sensors as well as the communication to the outside world. In general, these wireless communication nodes are battery operated. However, depending on the specific application, charging or replacing the batteries can be too expensive or even infeasible, e.g., when the nodes are located in remote locations or inside structures. Therefore, in order to provide sustainable service and to reduce the operation expenses, energy harvesting (EH) has been considered as a promising technology in which the nodes collect energy from the environment using natural or man-made energy sources such as solar or electromagnetic radiation. The idea behind EH is that the wireless communication nodes can recharge their batteries while in idle mode or while transmitting data to neighboring nodes. As a result, the lifetime of the wireless communication network is not limited by the availability of energy.

The consideration of EH brings new challenges in the design of transmission policies. This is because in addition to the fluctuating channel conditions and data arrival processes, the variability of the amount of energy available for the communication should be taken into account. Moreover, the three processes, EH, data arrival and channel fading, should be jointly considered in order to achieve optimum performance. In this context, this dissertation contributes to the research on EH wireless communication networks by considering power allocation and resource allocation problems in four different scenarios, namely, EH point-to-point, EH two-hop, EH broadcast and EH multiple access, which are the fundamental constituents of more complicated networks. Specifically, we determine the optimal allocation policies and the corresponding upper bounds of the achievable performance by considering offline approaches in which non-causal knowledge regarding system dynamics, i.e., the EH, data arrival and channel fading processes, is assumed. Furthermore, we overcome this unrealistic assumption by developing novel learning approaches, based on reinforcement learning, under the practical assumption that only causal knowledge of the system dynamics is available.

First, we focus on the EH point-to-point scenario where an EH transmitter sends data to a receiver. For this scenario, we formulate the power allocation problem for throughput maximization considering not only the transmit power, but also the energy consumed by the circuit. Adopting an offline approach, we characterize the optimum power allocation policy and exploit this analysis in the development of a learning approach. Specifically, we develop a novel learning algorithm which considers a realistic

EH point-to-point scenario, i.e., only causal knowledge of the system dynamics is assumed to be available. For the proposed learning algorithm, we exploit linear function approximation to cope with the infinite number of values the harvested energy, the incoming data and the channel coefficients can take. In particular, we propose four feature functions which are inspired by the characteristics of the problem and the insights gained from the offline approach. Through numerical simulations, we show that the proposed learning approach achieves a performance close to the offline optimum without the requirement of non-causal knowledge of the system dynamics. Moreover, it can achieve a performance up to 50% higher than the performance of reference learning schemes such as Q-learning, which do not exploit the characteristics of the problem.

Secondly, we investigate an EH two-hop scenario in which an EH transmitter communicates with a receiver via an EH relay. For this purpose, we consider the main relaying strategies, namely, decode-and-forward and amplify-and-forward. Furthermore, we consider both, the transmit power and the energy consumed by the circuit in each of the EH nodes. For the EH decode-and-forward relay, we formulate the power allocation problem for throughput maximization and consider an offline approach to find the optimum power allocation policy. We show that the optimal power allocation policies of both nodes, transmitter and relay, depend on each other. Additionally, following a learning approach, we investigate a more realistic scenario in which the EH transmitter and the EH decode-and-forward relay have only partial and causal knowledge about the system dynamics, i.e., each node has only causal knowledge about the EH, data arrival and channel fading processes associated to it. To this aim, two novel learning algorithms are proposed which take into account whether or not the EH nodes cooperate with each other to improve their learning processes. For the cooperative case, we propose the inclusion of a signaling phase in which the EH nodes exchange their current parameters. Through numerical simulations, we show that by providing the nodes with a complete view of the system state in a signaling phase, a performance gain of up to 40% can be achieved compared to the case when no cooperation is considered. Following a similar procedure, we investigate the EH two-hop scenario with an EH amplify-and-forward relay. We show that the resulting power allocation problem for throughput maximization is non-convex. Consequently, we propose an offline approach based on a branch-and-bound algorithm tailored to the EH two-hop scenario to find the optimal power allocation policy. Additionally, a centralized learning algorithm is proposed for the realistic case in which only causal knowledge of the system dynamics is available. The proposed learning approach exploits the fact that, with an amplify-and-forward relay, the communication between the transmitter and the receiver depends on a single effective channel, which is composed of the link between the transmitter and the relay, the relay gain and the channel from the relay to the receiver. By means of

numerical simulations, we show that the proposed learning algorithm achieves a performance up to two times higher than the performance achieved by reference schemes. Additionally, the extension of the proposed approaches to EH multi-hop scenarios is discussed.

Thirdly, an EH broadcast scenario in which an EH transmitter sends individual data to multiple receivers is studied. We show that the power allocation problem for throughput maximization in this scenario leads to a non-convex problem when an arbitrary number of receivers is considered. However, following an offline approach we find the optimal power allocation policy for the special case when two receivers are considered. Furthermore, inspired by the offline approach for two users, a novel learning approach which does not pose any restriction on the number of receiver nodes is developed. The proposed learning approach is a two-stage learning algorithm which separates the learning task into two subtasks: determining how much power to use in each time interval and deciding how to split this selected power for the transmission of the individual data intended for each receiver. Through numerical simulations, we show that the separation of tasks leads to a performance up to 40% higher than the one achieved by standard learning techniques, specially for large numbers of receivers.

Finally, an EH multiple access scenario is considered in which multiple EH transmitters communicate with a single receiver using multiple orthogonal resources. In this case, the focus is on the formulation of the resource allocation problem considering the EH processes at the different transmitters. We show that the resulting resource allocation problem falls into the category of non-linear knapsack problems which are known to be NP-hard. Therefore, we propose an offline approach based on dynamic programming to find the optimal solution. Furthermore, by exploiting the characteristics of the scenario, a novel learning approach is proposed which breaks the original resource allocation problem into smaller subproblems. As a result, it is able to handle the exponential growth of the space of possible solutions when the network size increases. Through numerical simulations, we show that in contrast to conventional reinforcement learning algorithms, the proposed learning approach is able to find the resource allocation policy that aims at maximizing the throughput when the network size is large. Furthermore, it achieves a performance up to 25% higher than the performance of the greedy policy that allocates the resources to the users with the best channel conditions.

Additionally, in order to carry out a full assessment of the proposed learning algorithms, we provide convergence guarantees and a computational complexity analysis for all the developed learning approaches in the four considered scenarios.

Kurzfassung

Fortschritte im Bereich drahtloser Sensornetze haben die Entwicklung von Technologien wie dem Internet der Dinge (engl. Internet of Things, IoT) und der Industrie 4.0 ermöglicht. Eine Schlüsselrolle in solchen Anwendungen spielen die drahtlosen Kommunikationsknoten, da sie Verbindungen zwischen verschiedenen Sensoren sowie die Kommunikation nach außen ermöglichen. Typischerweise sind die drahtlosen Kommunikationsknoten batteriebetriebene Geräte. Je nach Anwendung kann das Laden oder Ersetzen der Akkus jedoch zu teuer oder sogar nicht möglich sein, etwa wenn sich die Knoten an abgelegenen Orten befinden oder fest verbaut sind. Energy Harvesting (EH) gilt als eine vielversprechende Technologie, um in solchen Fällen einen dauerhaften Dienst zu erbringen und die Betriebskosten zu senken, indem die Kommunikationsknoten Energie aus natürlichen oder künstlichen Energiequellen in ihrer Umgebung, wie Sonnenstrahlung oder elektromagnetischer Strahlung, sammeln. EH beruht auf der Idee, dass die drahtlosen Kommunikationsknoten ihre Batterien nicht nur dann aufladen können, während sie auf das Ankommen neuer Daten warten, sondern auch während sie Daten an benachbarte Knoten übertragen. Infolgedessen ist die Lebensdauer des drahtlosen Kommunikationsnetzes nicht durch die Verfügbarkeit von Energie begrenzt.

Die Berücksichtigung von EH bringt neue Herausforderungen bei der Gestaltung drahtloser Übertragungsstrategien mit sich. Grund dafür ist, dass neben schwankenden Kanalbedingungen und Datenankunftsprozessen auch die Variabilität der für die Kommunikation verfügbaren Energiemenge erwogen werden muss. Darüber hinaus sollten die Prozesse des EHs, der Datenankunft und des Kanalfadings gemeinsam betrachtet werden, um eine optimale Performanz zu erzielen. Die vorliegende Dissertation trägt zur Erforschung drahtloser EH-Kommunikationsnetze bei, indem sie die Probleme der Leistungsverteilung und der Ressourcenallokation in vier verschiedenen Szenarien betrachtet, welche die grundlegenden Kommunikationsmuster in drahtlosen Netzwerken darstellen. Diese sind die Punkt-zu-Punkt-, die Zwei-Hop-, die Broadcast- und die Vielfachzugriff-Kommunikation. Konkret ermitteln wir optimale Allokationsstrategien und entsprechende obere Schranken an die erreichbare Performanz mithilfe von Offline-Ansätzen, die auf der für praktische Anwendungen unrealistischen Annahme nichtkausaler Kenntnis der Systemdynamik, d.h. der EH-, Datenankunfts- und Kanalfadingprozesse, basieren. Zudem schlagen wir neuartige Lernansätze basierend auf Methoden des bestärkenden Lernens vor, welche auf der praxistauglichen Annahme beruhen, dass nur kausale Kenntnis der Systemdynamik verfügbar ist.

Im EH-Punkt-zu-Punkt-Szenario sendet ein EH-Sender Daten an einen Empfänger. Für dieses Szenario wird das Problem der Leistungsverteilung zur Durchsatzmaximierung formuliert, unter Berücksichtigung sowohl der Sendeleistung, als auch des Energieverbrauchs der Schaltung. Mithilfe eines Offline-Ansatzes charakterisieren wir die optimale Strategie der Leistungsverteilung und nutzen diese Analyse zur Entwicklung eines Lernansatzes. Wir entwickeln einen neuartigen Lernalgorithmus, der ein realistisches EH-Punkt-zu-Punkt-Szenario berücksichtigt, in welchem nur kausale Kenntnis der Systemdynamik vorausgesetzt wird. Um die unendliche Anzahl an Werten zu bewältigen, die die gewonnene Energie, die eingehenden Daten und die Kanalkoeffizienten annehmen können, nutzt der vorgeschlagene Lernalgorithmus eine lineare Approximation. Insbesondere schlagen wir vier Merkmals-Funktionen vor, die sich aus den Eigenschaften des Problems und den Erkenntnissen aus dem Offline-Ansatz ableiten lassen. Mittels numerischer Simulationen zeigen wir, dass der vorgeschlagene Lernansatz eine Performanz nahe dem Offline-Optimum erreicht, ohne dass nicht-kausale Kenntnis der Systemdynamik erforderlich ist. Darüber hinaus kann der Algorithmus eine bis zu 50% höhere Performanz erzielen als Lernalgorithmen aus der Literatur, welche die spezifischen Eigenschaften des Problems nicht ausnutzen, wie etwa Q-Learning.

Im betrachteten EH-Zwei-Hop-Szenario kommuniziert ein EH-Sender über ein EH-Relais mit einem Empfänger, wobei entweder Decode-And-Forward oder Amplify-And-Forward als Relaisstrategie verwendet wird. Wir berücksichtigen sowohl die Sendeleistung als auch den Energieverbrauch der Schaltung in jedem der EH-Knoten. Für das EH-Decode-and-Forward-Relais formulieren wir das Problem der Leistungsverteilung zur Durchsatzmaximierung und betrachten einen Offline-Ansatz, um die optimale Leistungsverteilungsstrategie zu finden. Wir zeigen, dass die optimalen Strategien für die Leistungsverteilung an beiden Knoten, Sender und Relais, voneinander abhängen. Darüber hinaus untersuchen wir mithilfe eines Lernansatzes ein realistischeres Szenario, in welchem der EH-Sender und das EH-Decode-and-Forward-Relais nur partielle und kausale Kenntnis der Systemdynamik haben, d.h. jeder Knoten verfügt nur über kausale Kenntnis der EH-, Datenankunfts- und Kanalfadingprozesse. Zu diesem Zweck werden zwei neue Lernalgorithmen vorgeschlagen, die berücksichtigen, ob die EH-Knoten miteinander kooperieren, um ihre Lernprozesse zu verbessern, oder nicht. Im Falle der Kooperation schlagen wir den Einsatz einer Signalisierungsphase vor, in der sich die EH-Knoten über ihre aktuellen Parameter austauschen. Mittels numerischer Simulationen zeigen wir, dass das Bereitstellen eines vollständigen Überblicks über den Systemzustand an den Knoten mithilfe einer Signalisierungsphase einen Performanzgewinn von bis zu 40% ermöglicht, verglichen mit dem Fall, in dem keine Kooperation in Betracht gezogen wird. Basierend auf einem ähnlichen Verfahren untersuchen wir das EH-Zwei-Hop-Szenario mit einem EH-Amplify-And-Forward-Relais. Wir zeigen, dass das daraus resultierende Problem der Leistungsverteilung zur Durchsatzmaximierung

nicht konvex ist. Um die optimale Leistungsverteilungsstrategie zu finden, schlagen wir daher einen Offline-Ansatz vor, der auf einem Branch-and-Bound-Algorithmus basiert. Zusätzlich wird ein zentralisierter Lernalgorithmus für den realistischen Fall vorgeschlagen, in dem nur kausale Kenntnis der Systemdynamik vorhanden ist. Der vorgeschlagene Lernansatz basiert auf der Tatsache, dass die Kommunikation zwischen Sender und Empfänger mit einem Amplify-And-Forward-Relais von einem einzigen effektiven Kanal abhängt, der sich aus der Verbindung zwischen dem Sender und dem Relais, der Relaisverstärkung und dem Kanal vom Relais zum Empfänger zusammensetzt. Anhand numerischer Simulationen zeigen wir, dass der vorgeschlagene Lernalgorithmus eine Performanz erreicht, die bis zu zweimal höher ist als die Performanz von Referenzansätzen. Zusätzlich zeigen wir, wie die vorgeschlagenen Ansätze auf EH-Multi-Hop-Szenarien erweitert werden können.

Im EH-Broadcast-Szenario sendet ein EH-Sender individuelle Daten an mehrere Empfänger. Wir zeigen, dass das Problem der Leistungsverteilung zur Durchsatzmaximierung in diesem Szenario zu einem nicht-konvexen Problem führt, wenn eine beliebige Anzahl von Empfängern berücksichtigt wird. Basierend auf einem Offline-Ansatz finden wir jedoch die optimale Leistungsverteilungsstrategie für den Sonderfall von zwei Empfängern. Inspiriert durch den Offline-Ansatz für zwei Empfänger wird ein neuartiger Lernansatz entwickelt, der für eine beliebige Zahl an Empfängerknoten geeignet ist. Der vorgeschlagene Lernalgorithmus ist zweistufig und unterteilt die Lernaufgabe in zwei Teilaufgaben: Einerseits, zu bestimmen, wie viel Energie in jedem Zeitintervall verbraucht werden soll, und andererseits, zu entscheiden, wie die gewählte Energiemenge zur Übertragung individueller Daten an die verschiedenen Empfänger aufgeteilt werden soll. Mittels numerischer Simulationen zeigen wir, dass die Unterteilung der Lernaufgabe zu einer um bis zu 40% höheren Performanz führt als die von Standard-Lerntechniken, insbesondere für eine große Anzahl von Empfängern.

Im EH-Vielfachzugriff-Szenario kommunizieren mehrere EH-Sender mit einem einzigen Empfänger über mehrere orthogonale Ressourcen. In diesem Fall liegt der Fokus auf der Formulierung des Ressourcenallokationsproblems unter Berücksichtigung der EH-Prozesse an den verschiedenen Sendern. Wir zeigen, dass das daraus resultierende Ressourcenallokationsproblem in die Kategorie der nichtlinearen Rucksackprobleme fällt, welche NP-schwer zu lösen sind. Um die optimale Lösung zu finden, schlagen wir daher einen Offline-Ansatz vor, der auf dynamischer Programmierung basiert. Unter Ausnutzung der Eigenschaften des Szenarios wird ein neuartiger Lernansatz vorgeschlagen, der das ursprüngliche Problem der Ressourcenallokation in kleinere Teilprobleme zerlegt. Dieses Vorgehen ermöglicht es, das exponentielle Wachstum des Lösungsraums bei zunehmender Netzwerkgröße zu bewältigen. Anhand numerischer Simulationen zeigen wir, dass der vorgeschlagene Lernansatz in großen Netzwerken, im Gegensatz zu herkömmlichen Lernalgorithmen auf Basis des bestärkenden Lernens, jene Ressourcenallokationsstrategie findet, die darauf abzielt, den Durchsatz zu maximieren. Desweiteren erreicht der vorgeschlagene Lernansatz eine bis zu 25% höhere Performanz als die sogenannte gierige Strategie, welche die Ressourcen den Nutzern mit den besten Kanalbedingungen zuweist.

Um die vorgeschlagenen Lernalgorithmen umfassend bewerten zu können, leiten wir Konvergenzgarantien her und analysieren die Komplexität aller entwickelter Lernansätze in den vier betrachteten Szenarien.

Contents

Chapter 1

Introduction

1.1. Energy harvesting communications

Wireless sensor networks are formed by the collection of a large number of sensor nodes which are, in general, low-cost and low-power devices consisting of sensing, data processing, and communication components [ASSC02]. Thanks to the research effort in this area, wireless sensor networks have become essential in many different applications like environmental monitoring, traffic control networks, health monitoring, surveillance and object tracking [SZ16]. Moreover, they are a key enabling technique for emerging technologies such as Internet of Things (IoT) [AIM10] and Industry 4.0 [LLW$^+$17]. In many of these applications, the wireless communication nodes play an important role because they provide the connection between different sensors in the network as well as the connection to the outside world. However, depending on the specific application, charging or replacing the batteries of the wireless communication nodes can be too expensive or sometimes infeasible [DP10], 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 technology for such wireless communication nodes.

As depicted in Fig. 1.1, the idea behind EH is that the wireless communication nodes can recharge their batteries in an environmentally friendly way using natural or man-made energy sources, e.g., solar, thermal, vibrational, chemical, or electromagnetic radiation [UYE$^+$15, KLCL16]. Furthermore, the harvesting process is performed continuously during the operation of the wireless communication nodes, which translates in self-sustainability and theoretically perpetual operation of the nodes. However, it should be noted that the benefits of EH are not limited to an increased network lifetime. The fact that the EH nodes can collect energy from their environment reduces the carbon footprint and increases the mobility of the nodes [UYE$^+$15].

In addition to the channel fluctuations and stochastic data arrivals existing in any wireless communication system, the variable availability of energy inherent to EH communication systems has to be taken into account. When EH is considered, the energy available for transmission cannot be treated as a constant, as usually done in

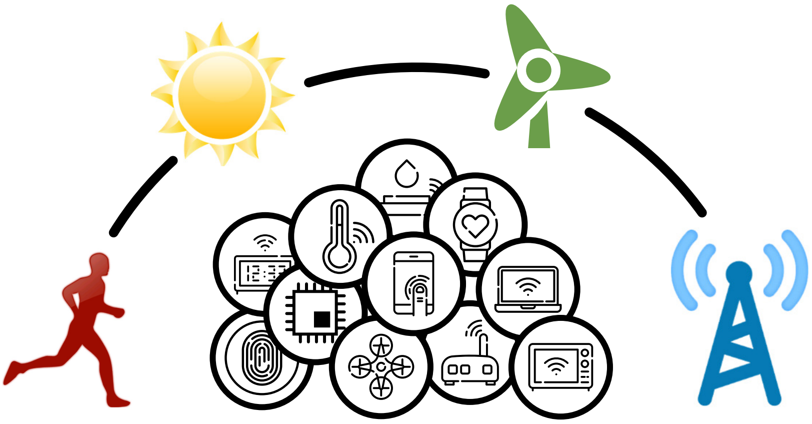


Figure 1.1. Example of different types of wireless sensor nodes and EH sources.

traditional communication systems. Moreover, the exact amount of available energy and the precise time when it can be harvested are hard to predict, which brings new challenges in the design of transmission strategies. In this thesis, we are particularly interested in finding transmission strategies that make an efficient use of the harvested energy in order to maximize the throughput in the system.

The most basic EH communication system is the point-to-point scenario, in which a single EH transmitter wants to communicate with a single receiver. This scenario, although basic, illustrates the fundamental dilemma faced by wireless communication nodes with EH capabilities, i.e., how to allocate the harvested energy in order to maximize the amount of data transmitted to the receiver, while at the same time avoiding battery overflow situations in which part of the harvested energy is wasted because the battery capacity has been reached. In addition to the EH process, this power allocation problem should also consider the remaining random processes in the system, namely, the data arrival and channel fading processes.

Naturally, the communication range in an EH communication system 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 EH based on electromagnetic radiation, the power density is in the order of fractions of nW/cm^2, and for solar energy, it is in the order of hundreds of mW/cm^2 [KLCL16]. In order to increase the limited communication range of an EH point-to-point communication system, relaying techniques can be considered since they are cost effective solutions for increasing the coverage, throughput and robustness of wireless networks [GYGP13, YZGK10]. 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. Such reduction of the communication range implies a reduction of the amount of energy required for data transmission in each hop. However, the consideration of EH relays in EH scenarios entails the joint design of transmission strategies for the relay and the transmitter [DLF18, OASL$^+$16b]. This requirement comes from the coupling between the data transmissions of the transmitter and the relay, i.e., the relay cannot retransmit data that has not yet been received from the transmitter. Moreover, the transmitter should consider the EH and channel fading processes associated to the relay in order to adapt its own transmission and avoid data buffer overflows at the relay. Therefore, in order to maximize the throughput and avoid wasting energy due to battery overflows, the EH, data arrival and channel fading processes of both, the transmitter and the relay, have to be considered.

The benefits of EH can be applied to systems beyond the single transmitter and single receiver case, i.e., to broadcast and multiple access scenarios. In wireless sensor networks, these two topologies are of paramount importance as they address two basic problems: on the one hand, how does a node disseminate data to multiple nodes (broadcast), and on the other hand, how does a node collect data from multiple nodes (multiple access). Nevertheless, these scenarios bring additional challenges in the design of the transmission strategies because the complexity of the problem increases with the number of nodes considered [GSMZ14]. For EH broadcast scenarios, in which a single EH transmitter wants to communicate with multiple receivers, the additional challenge is given by the need to consider the different channel fading processes associated to the links to the receiver nodes [YU12a]. Furthermore, if individual data is assumed to be intended for each receiver, multiple data arrival processes have to be taken into account in order to maximize the throughput. In the case of EH multiple access scenarios, multiple EH transmitters send data to a single receiver using multiple, and possibly orthogonal, resources. These orthogonal resources could correspond, for example, to a fraction of time if time-division multiple access (TDMA) is considered or one sub-carrier in the case of frequency- division multiple access (FDMA). As a consequence, in addition to the power allocation problem of the previous scenarios, the resource allocation problem needs to be solved in EH multiple access scenarios.

Regardless of the scenario being considered, the design of transmission strategies for EH communication systems depends on the amount of knowledge available about the random processes in the system, i.e., the EH, the data arrival and the channel fading processes. In the literature, three categories are distinguished, namely, offline, online and learning approaches [UYE$^+$15, GSMZ14]. The offline approaches assume that complete non-causal knowledge regarding the random processes is available [GSMZ14]. This means, the EH nodes know in advance, and before the data transmission starts, how much energy will be harvested in each time instant, how much data will arrive at the data buffer and what channel state will be experienced. Although this assumption cannot be fulfilled in real applications, it permits the definition of optimization problems that lead to the derivation of performance bounds for EH systems. A more relaxed assumption is considered by the online approaches, where only statistical knowledge is assumed to be available in advance [GSMZ14]. In these approaches, the exact amounts of harvested energy, the battery and data buffer levels, as well as the channel coefficients are not known. However, the probability distributions of the EH, data arrival and channel fading processes are assumed to be causally known. Within online approaches, dynamic programming strategies can be exploited to find transmission policies that maximize the throughput in the system [BG15]. However, in real scenarios perfect non-causal knowledge or statistical knowledge of the random processes is usually not available, especially if non-stationary EH, data arrival and channel fading processes are considered [OASL$^+$16b]. In such cases, where no knowledge is assumed, learning approaches can be used to find transmission strategies for EH systems. In learning approaches, more specifically in reinforcement learning (RL), an agent learns how to behave in an unknown environment by interacting with it [SB18]. In the case of EH communications, the agent can be the EH transmitter and the environment comprises the unknown random processes, i.e., the EH, data arrival and channel fading processes. The transmitter learns how much power to use for the transmission by making decisions and evaluating the response, for example, by evaluating the achieved throughput.

In this thesis, we investigate the design of transmission strategies for EH communication systems. Following offline and learning approaches, we consider the four different scenarios depicted in Figure 1.2, i.e., point-to-point, two-hop, broadcast and multiple access, which are the main building blocks of more complicated networks. In Figure 1.2, the battery symbols indicate which nodes are harvesting energy from the environment in each of the considered scenarios. Furthermore, the battery represents the battery size, i.e., the amount of energy that can be stored, and the green areas represent the amount of available energy. The receiver nodes do not harvest energy and are assumed to be connected to a continuous power supply.

1.2. State-of-the-art

1.2.1. Introduction

This section presents a review of the state-of-the-art with regard to the EH communication scenarios investigated in this thesis. First, we consider the EH point-to-point communication scenario which consists of a single EH transmitter and a single receiver.

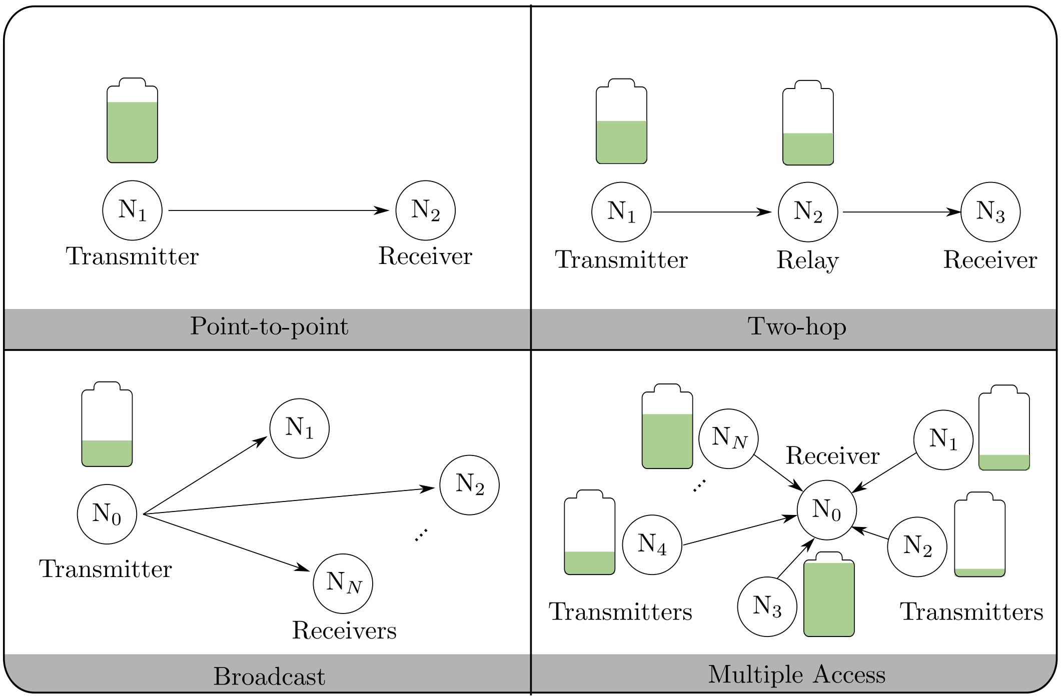


Figure 1.2. Four scenarios considered in the thesis.

Next, we review the literature on EH two-hop communication scenario. In this case, two EH nodes are considered, namely, the EH transmitter and the EH relay. Afterwards, the state-of-the-art considering an EH broadcast scenario is presented. The broadcast scenario is composed of a single EH transmitter that sends data to multiple receivers. Finally, the works considering EH multiple access scenario are summarized. In the EH multiple access scenario, multiple EH transmitters communicate with a single receiver. For each of these scenarios, the presented literature considers the use of offline, online and learning approaches.

1.2.2. Point-to-point scenario

Offline approaches for EH point-to-point communications have been investigated in [TY12c, OTY$^+$11, YU12b, LOAS$^+$17, OGE13, OGE12, TY12b]. Specifically, in [TY12c] it is shown that the power allocation problem for throughput maximization within a deadline is equivalent to the minimization of the completion time given that a fixed amount of data needs to be transmitted. A similar scenario is investigated in [OTY$^+$11], where the authors consider a fading channel between the transmitter and the receiver, and a modified water-filling algorithm is proposed to maximize the throughput within a deadline. The optimal packet scheduling problem is considered in [YU12b], where the authors derive the optimal policy for two cases, namely, when the packets to be transmitted are available at the transmitter and when a data arrival process is considered. In [LOAS$^+$17], the case when each data packet to be sent has an individual deadline is studied. In this paper, to which the author of this thesis has contributed, the optimal transmission strategy for the delay-constrained throughput maximization problem as well as for the delay-constrained energy minimization problem is found. The authors of [OGE13] study the minimization of the distortion for an EH transmitter communicating over a fading channel, assuming that each received message has to be reconstructed at the destination within a certain deadline. The energy cost of transmission and processing at the transmitter in an EH point-to-point scenario is investigated in [OGE12] and the effect of inefficient energy storage on the achievable throughput is studied in [TY12b].

Online approaches for the EH point-to-point scenario are considered in [OTY$^+$11, LYG09, LZL13a, BGD13]. A fading channel is assumed in [OTY$^+$11] and the problem of online scheduling for throughput maximization within a deadline is considered. Furthermore, assuming statistical and causal knowledge of the energy and fading variations, the authors propose the use of continuous time stochastic dynamic programming in order to find the corresponding transmission strategy. A similar scenario is considered in [LYG09], where an on-off mechanism at the transmitter is proposed, i.e., for each packet arrival, a binary decision of whether to transmit or drop the packet is made. In this case, the energy arrival is described as a continuous time Markov chain and the statistical distribution of the importance of the messages is assumed to be known. The minimization of the system outage probability is studied in [LZL13a]. To this aim, the authors assume that in one time interval, a fixed amount of data is transmitted, model the energy arrival as a random variable and propose a save-then-transmit protocol. In [BGD13], the authors model the throughput maximization problem as a Markov decision process and propose a transmission strategy based on the policy iteration algorithm.

Learning approaches have been applied to EH point-to-point scenarios in [BGD13, GGV16, XHNY15, SKN17]. In [BGD13], the well-known Q-learning algorithm is used to maximize the throughput within a deadline. The authors assume that the amount of harvested energy, the channel coefficients and the transmit power in each time instant are taken from a finite and discrete set. Moreover, they assume that the data arrives in packets and for each data packet, the decision of transmit or drop has to be made. In [GGV16], the authors use online convex optimization to derive online algorithms to learn the transmission policy from previous observations. Authors in [XHNY15] use Bayesian RL at the EH transmitter in order to learn the statistics of EH and channel fading processes, and the probability distribution of the achievable throughput. Finally,

Table 1.1. Summary of the state-of-the-art of offline and learning approaches for the power allocation problem in the EH point-to-point communication scenario

		Finite battery	Circuit energy	Infinite data	Data arrival and finite data buffer	Fading channel	Continuous sets
Offline	[TY12c]	✓	-	✓	-	-	✓
	[OTY^{+}11]	✓	-	✓	-	✓	✓
	[YU12b]	-	-	✓	-	-	✓
	[LOAS^{+}17]	✓	-	-	✓	-	✓
	[OGE13]	-	-	-	-	✓	✓
	[OGE12]	✓	✓	✓	-	✓	✓
	[TY12b]	✓	-	✓	-	-	✓
	Our work	✓	✓	✓	✓	✓	✓
Learning	[BGD13]	✓	-	-	✓	✓	-
	[GGV16]	-	-	✓	-	-	✓
	[XHNY15]	✓	-	-	✓	✓	-
	[SKN17]	✓	-	-	-	-	-
	Our work	✓	✓	✓	✓	✓	✓

the authors of [SKN17] exploit weather forecast data to enhance the performance of the RL algorithm at the EH transmitter, assuming solar energy as the EH source.

Table 1.1 summarizes the state-of-the-art of offline and learning approaches for the EH point-to-point scenario. In the table, the categories are given by the considered assumptions regarding the EH, data arrival and channel fading processes. For the EH process, the categories correspond to whether or not a finite battery is studied and whether or not the energy consumed by the circuit is taken into account. For the data arrival process, two categories are considered, namely, whether infinite data is available at the transmitter or if a data arrival process with a finite data buffer is assumed. For the channel fading process, we indicate whether or not a fading channel is assumed between the EH transmitter and the receiver. Additionally, we indicate whether or not the fact that the amounts of energy, battery levels, amounts of incoming data and channel coefficients take values in a continuous range is taken into account. This consideration has implications on the design of learning algorithms, as it will become clear throughout this thesis. When one of these assumptions is taken into account by one of the reference works listed in the table, the corresponding cell is marked with the check mark symbol ✓. Additionally, we indicate the assumptions considered in this thesis regarding the EH, data arrival and channel fading process in the context of offline and learning approaches for EH point-to-point communication scenarios.

1.2.3. Two-hop scenario

For EH two-hop scenarios, offline approaches have been the major direction of state-of-the-art research [GD11, OE12, OE13, OE15, LZL13b, VY13]. In [GD11], 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, 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 [OE12], 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. 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 [LZL13b]. A similar scenario is considered in [VY13]. However, in [VY13], 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 [OE13, OE15], where the authors formulate a convex optimization problem to find the optimal offline transmission policy that maximizes the throughput. In [ZBM15, Liu16, TZW14], simultaneous wireless information and power transfer in a two-hop scenario with multiple relays is considered. In [ZBM15], the authors assume randomly located relays and analyze the performance of the system considering the impact of the number of relays. In [Liu16], 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 [TZW14] aim at minimizing the transmission time and propose a harvest-then-decode-and-forward algorithm at the relays.

In [MSA14] and [AD15], online approaches for EH two-hop scenarios are considered. In [MSA14], 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 [AD15], where a power allocation policy aiming at maximizing the long time average throughput is found using Lyapunov optimization techniques.

Learning techniques, although promising for EH scenarios, have hardly been exploited so far to find transmission policies for EH two-hop scenarios. In [HD16], 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.

Table 1.2. Summary of the state-of-the-art of offline and learning approaches for the power allocation problem in the EH two-hop communication scenario

		EH relay	Finite batteries	Circuit energy	Infinite data at transmitter	Data arrival at transmitter	Finite data buffer at relay	Fading channel	Decode-and-forward	Amplify-and-forward	Full-duplex	Half-duplex
Offline	[GD11]	✓	-	-	✓	-	-	-	✓	-	✓	✓
	[OE12]	✓	-	-	✓	-	-	-	✓	-	-	✓
	[OE13]	✓	-	-	✓	-	-	-	✓	-	-	✓
	[OE15]	✓	-	-	✓	-	✓	-	✓	-	-	✓
	[LZL13b]	-	-	-	✓	-	-	-	✓	-	-	✓
	[VY13]	-	-	-	✓	-	✓	-	✓	-	-	✓
	[ZBM15]	✓	-	-	✓	-	-	✓	✓	-	-	✓
	[Liu16]	✓	-	-	✓	-	-	✓	✓	-	-	✓
	[TZW14]	✓	✓	-	-	-	✓	-	✓	-	-	✓
	Our work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Learning	[HD16]	✓	✓	-	-	✓	✓	✓	-	✓	-	✓
	Our work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

We summarize the state-of-the-art of offline and learning approaches for the EH two-hop communication scenario in Table 1.2. Compared to the EH point-to-point scenario of Section 1.2.2, more categories are taken into account. Specifically, for the EH process we consider three categories, namely, whether or not the relay is harvesting energy, whether or not finite batteries are assumed and whether or not the energy consumed by the circuit is taken into account. For the data arrival process we distinguished whether infinite data or a data arrival process is considered at the transmitter, regardless of the size of the data buffer. For the relay, we do not make this differentiation because in all the reference works, as well as in this thesis, it is assumed that the relay only retransmits what it receives from the transmitter and does not have any own data to send. Nevertheless, we indicate whether or not the relay is equipped with a finite data buffer. For the channel, we distinguished whether or not fading channels are assumed. Additionally, for the EH two-hop communication scenario we have included categories corresponding to the two main relaying techniques, namely, decode-and-forward and amplify-and-forward, as well as categories corresponding to the relay transmission modes, i.e., full-duplex and half-duplex.

1.2.4. Broadcast scenario

Research effort on EH broadcast scenarios has primarily focused on offline approaches [EOUB13, OYU13, AUBE11, YOU12, FAUC16, TY12a]. In [EOUB13], an EH transmitter with an infinite battery broadcasting individual data packets to two receivers over an additive white Gaussian noise (AWGN) channel is considered. For this scenario, the authors show the structural properties of the optimal solution and prove its uniqueness. Similarly, in [OYU13] a two-user EH broadcast scenario is studied. However, in this case the authors consider the effect of a finite battery and fading channels. The total delay in a two-user EH broadcast scenario is minimized in [FAUC16]. For this case, the authors report that in the optimal policy, both users may not be served simultaneously all the time, and that gaps in the data transmission, in which none of the receivers is served, might occur. Furthermore, in [TY12a], the effect of an inefficient battery in a two-user EH broadcast scenario is studied. Authors in [AUBE11] and [YOU12] consider an EH transmitter with a fixed number of data packets to be sent to multiple receivers. In both cases, the goal is to find a power allocation policy that minimizes the time required to transmit the data intended for all the different receivers. In [YU12a] it is shown that the optimal total transmit power sequence has the same structure as in the point-to-point scenario. Moreover, the authors propose an algorithm to find the optimal policy based on the reduction of the broadcast scenario to a point-to-point scenario.

Table 1.3. Summary of the state-of-the-art of offline and learning approaches for the power allocation problem in the EH broadcast communication scenario

		Two users	Arbitrary number of users	Finite battery	Circuit energy	Individual data	Infinite data	Data arrival and finite data buffer	Fading channel
Offline	[EOUB13]	✓	-	-	-	✓	-	-	-
	[OYU13]	✓	-	✓	-	✓	✓	-	✓
	[AUBE11]	✓	-	-	-	✓	-	-	-
	[YOU12]	✓	-	-	-	✓	-	-	-
	[FAUC16]	✓	-	-	-	✓	-	-	-
	[TY12a]	✓	-	✓	-	-	✓	-	-
	Our work	✓	-	✓	✓	✓	✓	✓	✓
Learning	State of the art	-	-	-	-	-	-	-	-
	Our work	✓	✓	✓	✓	✓	✓	✓	✓

Using an online approach, a two-user EH broadcast scenario, in which the amounts of harvested energy are causally known, is studied in [BU16]. The authors consider AWGN channels and find the optimal online power allocation policy when the EH process follows a Bernoulli distribution. For any other distribution, a sub-optimal transmission strategy is proposed.

Learning techniques, although promising for EH scenarios, have not yet been used to find transmission policies for EH broadcast scenarios when only causal knowledge regarding the EH, data arrival and channel fading processes is available.

The state-of-the-art of offline and learning approaches for the EH broadcast communication scenario is summarized in Table 1.3. As in the previous scenarios, the categories included in the table correspond to the considered assumptions regarding the EH, data arrival and channel fading processes. Initially, we distinguish the number of receivers considered in the scenario. For the EH process, we differentiate whether or not a finite battery is assumed and whether or not the energy consumed by the circuit is taken into account. For the data arrival process, we first indicate whether or not individual data is intended for each receiver. Additionally, we differentiate two cases regarding the data arrival process, i.e., whether infinite data is available at the transmitter or if a data arrival process with a finite data buffer is assumed. Regarding the channel fading process, we indicate whether or not a fading channel between the transmitter and each of the receivers is assumed.

1.2.5. Multiple access scenario

Previous work on EH multiple access scenarios, has primarily focused on finding power allocation policies for the EH transmitters using offline approaches [YU12a, GKU16, ZHC$^+$15, WAW15, JE15]. An EH two-user multiple access senario is considered in [YU12a] where the authors propose a generalized iterative backward water-filling algorithm to minimize the time required for data transmission. A similar scenario is considered in [GKU16, ZHC$^+$15], where the EH transmitters are able to cooperate with each other. In [GKU16], the authors find the optimum power allocation policy assuming the EH transmitters are able to overhear each other's transmitted signals and can cooperate by forming common messages. In [ZHC$^+$15], a wired rate-limited channel is assumed to be available for the communication between the transmitters. The two-user scenario is extended in [WAW15], where multiple users are considered and an iterative water-filling based algorithm is proposed to find the optimal power allocation policy. In [JE15], the authors characterize the stability region when two bursty EH users are randomly accessing the channel to a common receiver. For this scenario, the authors take into account the effects of multi-packet reception capabilities at the receiver. Note however, that the resource allocation problem for throughput maximization in the EH multiple access scenario has not yet been studied.

Online approaches for power allocation in EH multiple access scenarios are investigated in [AD16, KM14, LDC16]. In [AD16], the authors use Lyapunov optimization techniques to find the power allocation policy aiming at maximizing the long-term time-average transmission rate considering finite batteries at the EH transmitters. The authors of [KM14] follow an online approach to study a continuous-time power policy for EH multiple access scenarios. To this aim, the battery is modeled as a compound Poisson dam and the cases of infinite and finite batteries are analyzed. In [LDC16], an EH multiple access channel using TDMA is considered and the authors investigate the optimal power allocation policy assuming only statistical knowledge regarding the EH processes of all the users. The resource allocation problem in EH multiple access scenarios is investigated in [YW15]. Assuming that the EH processes at the transmitters can be modeled as independent Bernoulli processes, the authors consider an online approach to schedule the transmissions according to the instantaneous battery and channel states of the transmitters.

Learning approaches for EH multiple access scenarios have been considered in [BG15]. The authors model the EH processes using independent two-state Markov chains, i.e., the transmitters can harvest either one energy unit or none, and formulate the resource allocation problem as a restless multi-armed bandit (MAB) problem.

Table 1.4. Summary of the state-of-the-art of offline and learning approaches for the resource allocation problem in the EH multiple access communication scenario

		Arbitrary number of users	Finite battery	Circuit energy	Infinite data	Fading channel	Continuous sets
Offline	State of the art	-	-	-	-	-	-
	Our work	✓	✓	✓	✓	✓	✓
Learning	[BG15]	✓	✓	-	-	✓	-
	Our work	✓	✓	✓	✓	✓	✓

Table 1.4 summarizes the state-of-the-art of offline and learning approaches for the resource allocation problem in the EH multiple access communication scenario. In the table, we initially indicate whether or not an arbitrary number of users is considered. Regarding the EH process, we distinguish the cases when a finite battery is considered and when the circuit energy is taken into account. For the data arrival, we indicate whether or not infinite data is assumed at the transmitters. For the channel, we show whether or not a fading channel is considered. Furthermore, as in the EH point-to-point case of Table 1.1, we indicate whether or not the fact that the amounts of energy and channel gains can take any value in a continuous range is considered.

1.3. Open issues

In this section, the open questions resulting from the review of the state-of-the-art are summarized.

As discussed in the previous sections, finding the offline optimal power allocation policy in EH scenarios requires complete non-causal knowledge regarding the EH, the data arrival and the channel fading processes. However, in real applications this non-causal knowledge is not available. Consequently, approaches that can cope with this limitation need to be developed. In particular, the requirement of perfect non-causal knowledge can be overcome if learning approaches, specifically RL, are considered. Naturally, the application of RL to EH communications opens a set of questions that depend on the particular scenario being considered. In this thesis, we investigate four different scenarios which are the main building blocks of larger networks, i.e., point-to-point, two-hop, broadcast and multiple access scenarios.

First, a point-to-point communication scenario is considered in which an EH transmitter sends data to a receiver. In this case, the following questions arise:

1. How can RL be used to develop an algorithm which finds the power allocation policy in an EH point-to-point scenario? How can the energy consumed by the circuit be taken into account?
2. How to deal with the fact that the amount of harvested energy, the battery levels, the amount of incoming data, the data buffer level and the channel gains can take any value in a continuous range?
3. Can convergence guarantees be provided for the learning approach?
4. What is the computational complexity of the learning approach?

Second, we consider a two-hop scenario in which an EH transmitter sends data to a receiver through an EH relay. For this scenario, two types of relay are considered, namely, an EH decode-and-forward and an EH amplify-and-forward relay. We initially focus on an EH decode-and-forward relay and investigate offline and learning approaches which lead to the following open questions:

5. Are the power allocation problems of the transmitter and relay coupled?
6. In a learning approach, how to deal with the fact that the EH nodes only have partial knowledge about the system state, i.e., they only know their own amounts of harvested energy, battery levels, data buffer levels and channel gains? Can cooperation among the EH nodes be exploited in order to increase the achieved throughput?
7. Can convergence guarantees be provided for the learning approach?
8. What is the computational complexity of the learning approach?

Next, we consider an EH amplify-and-forward relay. In this case, neither offline approaches nor learning approaches have been considered so far in the literature. Consequently, the following questions arise:

9. How to formulate an optimization problem to find the optimal power allocation in an EH two-hop scenario with an amplify-and-forward relay?

10. How can the resulting optimization problem be efficiently solved?
11. How can RL be used to develop a learning algorithm that finds the power allocation policies for the transmitter and relay in the EH two-hop scenario with an amplify-and-forward relay?

Third, we consider a broadcast scenario with an EH transmitter which sends individual data to multiple receivers. In this case, we are interested in the power allocation problem for the transmission of the data intended for the different receivers. To this aim, the following open questions are considered:

12. How to develop a learning approach to find the power allocation policy in the EH broadcast scenario when only causal knowledge of the system dynamics is available?
13. Can convergence guarantees be provided for the learning approach?
14. What is the computational complexity of the learning approach?

Fourth, a multiple access scenario is investigated in which multiple EH transmitters want to communicate with a single receiver. For this scenario, we focus on the allocation of multiple orthogonal resources and address the following open questions:

15. How to model the resource allocation problem considering that only causal knowledge regarding the EH, data arrival and channel fading processes is available?
16. How to design an RL algorithm that handles the combinatorial nature of the resource allocation problem?
17. Can convergence guarantees be provided for the learning approach?
18. What is the computational complexity of the learning approach?

1.4. Thesis overview and contributions

In this section, an overview of the thesis and a summary of the main contributions addressing the open questions introduced in Section 1.3 are presented. Additionally, the contents of each chapter are briefly described.

In Chapter 2, the system model is presented. Specifically, the energy harvesting and the energy consumption models for the EH nodes are provided. Additionally, the channel and data arrival models considered for all the scenarios investigated in this thesis are explained. Furthermore, an introduction to Markov decision processes is provided which includes the definition of the value functions and the concept of linear function approximation.

In Chapter 3, the power allocation problem in an EH point-to-point scenario is considered. Using the existing results in the literature, we first formulate the offline optimization problem in order to use it as a benchmark. We then propose a learning approach for the more realistic case when only causal knowledge about the system dynamics, i.e., the EH, data arrival and channel fading processes[1], is assumed. This chapter addresses open questions 1 to 4 which lead to the following contributions:

1. An RL algorithm, which leverages linear function approximation and the state-action-reward-state-action (SARSA) update, is proposed to find the power allocation policy at the EH transmitter which aims at maximizing the throughput.

2. A set of feature functions that exploit the characteristics of the offline solution are proposed in order to perform linear function approximation and handle the fact that the amounts of harvested energy, battery levels, data buffer levels and channel gains are taken from a continuous set.

3. By exploiting results from the RL literature, we show that the convergence of the proposed learning approach to a bounded region depends on the selection of the learning rate parameter.

4. By means of a computational complexity analysis, we show that the complexity of the proposed learning approach increases only linearly with the number of transmit power values the transmitter can select.

In Chapter 4, the power allocation problem for throughput maximization in EH two-hop communications is investigated. Initially, we consider an EH decode-and-forward relay and study offline as well as learning approaches. The following contributions give answer to open questions 5-8:

[1] Throughout this thesis, the term "system dynamics" refers to the EH, data arrival and channel fading processes of the considered scenario. Both expressions, i.e., system dynamics and EH, data arrival and channel fading processes, are used interchangeably.

5. Following an offline approach, we show that the power allocation problems of the EH transmitter and the EH relay are coupled. This means, in order to find the optimal power allocation policy, the EH, data arrival and channel fading processes associated to both nodes should be jointly considered.

6. As only partial causal knowledge of the system state is available at the EH nodes, we propose two learning approaches that consider different levels of cooperation. In the first case, we assume the transmitter and the relay do not have any knowledge about the battery level, data buffer level or channel gain associated to the other node. As a result, we propose to separate the power allocation problem into two EH point-to-point communication problems. The resulting learning approach, termed independent SARSA, solves independent power allocation problems at the transmitter and at the relay and aims at maximizing the throughput in each point-to-point link. In the second case, mechanisms to overcome the partial observability of the system state and increase the throughput are proposed. Specifically, the use of a channel predictor based on a Kalman filter to estimate the channel gains and the inclusion of a signaling phase in which the transmitter and receiver exchange their current battery levels, data buffer levels and channel gains, are proposed. The resulting learning approach, termed cooperative SARSA, is a multi-agent RL algorithm in which the nodes cooperate with each other to maximize the throughput in the system.

7. For the two proposed learning approaches, convergence guarantees are provided. In the case of independent SARSA, we show that the learning approach corresponds to two independent instances of an EH point-to-point scenario. Therefore, the same convergence guarantees apply. For the cooperative SARSA algorithm, we show that the local action-value function of the transmitter and the receiver, which represents the expected throughput given a certain system state and transmit power, is a projection of the centralized action-value function obtained when the system state is perfectly known by a central entity.

8. For the proposed independent SARSA algorithm, we show that the computational complexity increases only linearly with the number of transmit power values that can be selected, as in the EH point-to-point case. For the cooperative SARSA algorithm, we demonstrate that the computational complexity depends linearly on the product of the number of features functions considered in the linear function approximation and the number of transmit power values that can be selected. This means, the extra complexity incurred by the cooperative SARSA algorithm compared to the independent SARSA algorithm is the price to be paid for the improvement in the performance.

Afterwards, we consider an EH amplify-and-forward relay and, as in the previous case, investigate offline and learning approaches. In particular, offline approaches considering an EH amplify-and-forward relay have not yet been considered in the literature. As a consequence, in this chapter we address open questions 9 and 10 through the following contributions:

9. We show that the consideration of an EH amplify-and-forward relay results in a non-convex optimization problem. Therefore, we reformulate the original optimization problem as the difference between two concave functions which fits in a class of global optimization techniques known as difference of convex functions (D.C.) programming problems.
10. A branch-and-bound algorithm is tailored to fit the EH constraints in the two-hop scenario with a half-duplex amplify-and-forward relay. We show that in order to facilitate the branching process, the representation of the feasible region has to be adapted. Furthermore, we reduce the complexity in the calculation of the lower and upper bounds by relaxing the D.C. programming problem into a convex problem with a linear objective function.

After considering the offline approach, we investigate learning approaches for this scenario. Specifically, we answer open question 11 through the following contribution:

11. We show that in an EH two-hop scenario with an amplify-and-forward relay, the communication between the transmitter and the receiver cannot be separated as in the decode-and-forward case, but has to be considered as a single link with an effective channel that depends on the channel from the transmitter to the relay, the relay gain and the channel from the relay to the receiver. As a result, a centralized learning algorithm based on the EH point-to-point scenario is proposed.

In Chapter 5, an EH broadcast scenario is investigated. Using existing results from the literature, we first present the offline optimum solution of the problem and then propose a learning approach to find the power allocation policy that aims at maximizing the throughput. Open questions 12 and 14 are addressed in the following contributions:

12. Considering that the power allocation problem in the EH broadcast scenario entails the selection of the total power to use in each time interval and its distribution for the transmission of the data intended for the different receivers, we

propose a two-stage RL algorithm that divides the learning task into two smaller sub-tasks. This division increases the learning speed and the performance because each sub-task addresses a different problem, i.e., how much power to allocate in each time interval and how to split the allocated power among the data to be transmitted to each receiver.

13. We show that each stage in the proposed learning approach is independent of the other. Therefore, the convergence is evaluated for each of them. We show that the convergence of each stage to a bounded region depends only on the selection of the learning rate parameter.

14. Through a computational complexity analysis, we show that the complexity of the proposed learning approach is determined by the second stage which decides on how to split the available power. This means, it depends linearly on the number of possible splitting solutions that are considered.

In Chapter 6, we investigate the allocation of multiple orthogonal resources in a multiple access scenario with a single receiver and multiple EH transmitters. We initially present the offline optimum solution and continue with the learning approach. The following contributions address open questions 15 and 18:

15. Taking into account the combinatorial nature of the resource allocation solutions, the offline optimization problem for the considered scenario is formulated. The resulting problem is identified as a non-linear knapsack problem which is known to be NP-hard. An offline approach based on dynamic programming is proposed to find the optimum resource allocation policy.

16. An RL algorithm termed combinatorial SARSA is proposed. The name of the algorithm stands for its ability to handle the combinatorial nature of the resource allocation solutions by breaking the original problem into smaller subproblems, thus tackling the curse of dimensionality in the search of resource allocation solutions and leading to a high throughput.

17. We show that, similar to the previous cases, the use of linear function approximation together with the SARSA update results in the fact that the convergence of each of the learning subproblems to a bounded region depends only on the selection of the learning rate parameter.

18. Through a computational complexity analysis, we show that the complexity of the proposed learning approach depends linearly on the minimum between the number of resource allocation solutions and the number of solutions that can be

stored in the memory. Therefore, the complexity can be bounded by the amount of memory that is allocated.

In addition to the aforementioned contributions, in Chapters 3-6, the performances of the proposed learning approaches are analyzed and compared to offline approaches, standard RL algorithms and low-complexity heuristics through numerical simulations.

Finally, in Chapter 7 the main conclusions of the thesis and a brief outlook for future work are presented.

Chapter 2

System model and Markov decision process

2.1. Introduction

In this chapter, the system model considered in this thesis as well as an introduction on Markov decision processes are presented. The system model comprises the model of the system dynamics, i.e., the EH, data arrival and channel fading processes associated to the EH nodes. As these models are associated to each EH node and are, therefore, independent of the considered scenario, we present a general description which applies to the four scenarios investigated in this thesis, i.e., EH point-to-point, EH two-hop, EH broadcast and EH multiple access. Additionally, we introduce the framework of Markov decision processes which is a mathematical tool suitable for the modeling of decision making situations. As it will become clear throughout this thesis, Markov decision processes play an important role in RL because they facilitate the modeling of the learning problem and the subsequent design of learning algorithms [SB18].

The chapter is organized as follows. First, the system model is presented in Section 2.2 and Markov decision processes are introduced in Section 2.3. The author of this thesis has used this system model in previous publications [OASL$^+$15, OASL$^+$16b, OASL$^+$16a, OASL$^+$17, OWK18, OWK19]

2.2. System model

2.2.1. The energy harvesting node

In the study of wireless communications considering EH nodes, the random processes associated to it should be taken into account. As depicted in Figure 2.1, these random processes correspond to the EH, data arrival and channel fading processes, shown in green, blue and yellow, respectively. Moreover, they are associated to every EH node, regardless of whether it is an EH transmitter or an EH relay. In the following sections, we present a detailed description of these three random processes associated to an EH node. Note that although the descriptions focus on one EH node, the models apply to all the EH nodes considered in the four different scenarios investigated in this thesis.

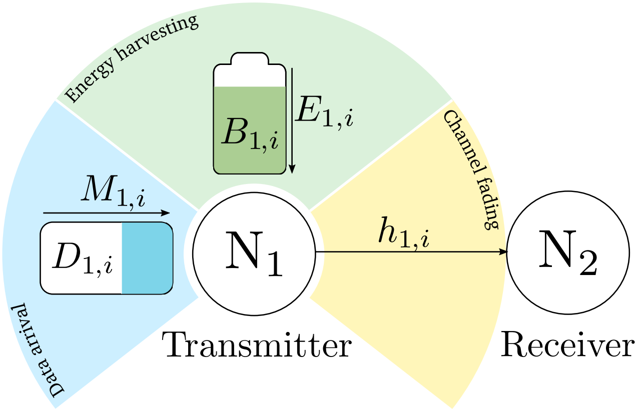


Figure 2.1. Diagram of the random processes associated to an EH node.

The next sections are organized as follows. First, the EH model is described in Section 2.2.2. This model includes the characteristics of the EH process as well as the corresponding constraints. Next, in Section 2.2.3, the data arrival model is presented and last, the considered channel model, which includes the characteristic of the channel as well as the definition of the channel capacity, is explained in Section 2.2.4.

2.2.2. Energy harvesting model

In this section, the EH model is described. This model includes both, the EH as well as the energy consumption model. A summary of all the parameters associated to the EH model, which will be introduced throughout this section, is provided in Table 2.1.

In our model, the nodes are denoted by N_n, $n = 1, 2, ..., N$. The transmitters and relays are termed EH nodes while the receiver nodes are termed non-EH nodes. The EH nodes harvest energy from the environment and use this energy for the transmission of data. The non-EH nodes do not harvest energy and are assumed to be connected to a fixed power supply. As extensively done in the literature [TY12c, OTY+11, OGE12], we consider a discrete time model divided in I time intervals. The time duration τ_i between two consecutive EH time intervals i and $i+1$ is assumed to be constant such that $\tau_i = \tau$, $i = 1, ..., I$. At the end of each time interval i, an amount of energy $E_{n,i} \in \mathbb{R}^+$ is received by the EH node N_n. The amount of energy $E_{n,i}$ may also take the value $E_{n,i} = 0$ to include the case when N_n does not harvest energy in time interval i. Furthermore, the maximum amount of energy that can be harvested at N_n, termed $E_{\max,n}$, as well as the probability distribution of the energy harvesting process depend on the energy source that is considered, e.g., solar, thermal, chemical, vibrational, etc.

Table 2.1. Summary of the parameters associated to the EH model.

Parameter	Description
$B_{\max,n}$	Battery capacity of EH node N_n
$B_{n,i}$	Battery level of EH node N_n, measured at the beginning of time interval i
$E_{n,i}$	Amount of harvested energy, received at the end of time interval i, by EH node N_n
$E_{\max,n}$	Maximum amount of energy that can be harvested 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
$p_{n,i}^{\mathrm{Circ}}$	Power consumed by the circuit of EH node N_n in time interval i
$p_{n,i}^{\mathrm{Tx}}$	Transmit power used by EH node N_n in time interval i
τ	Time interval duration

The harvested energy $E_{n,i}$ is stored in a rechargeable battery with maximum capacity $B_{\max,n}$ and it is assumed that no energy is lost in the process of storing or retrieving energy from the batteries. Moreover, the battery level $B_{n,i}$ is always measured at the beginning of each time interval i. As the battery cannot be recharged instantaneously, it is assumed that at the beginning of time interval i, the battery only stores the energy which has been harvested in the previous time intervals j, $j \leq i-1$. Additionally, we consider that at the beginning of time interval $i=1$, the nodes have not yet harvested any energy and their batteries are empty, i.e., $B_{n,1}=0$.

As mentioned before, it is assumed that the harvested energy is used for data transmission. Our energy consumption model includes both, the circuit energy $E_{n,i}^{\mathrm{Circ}}$ and the transmit energy $E_{n,i}^{\mathrm{Tx}}$ used in each time interval i. $E_{n,i}^{\mathrm{Circ}}$ corresponds to the energy required by all the modules that process the signal to be transmitted, e.g., base-band signal processing unit, digital-to-analog converter, etc. $E_{n,i}^{\mathrm{Tx}}$ is the energy of the transmitted signal. For simplicity, we assume that the energy consumed when the EH nodes are in sleep mode is much smaller than the energy consumed while transmitting and can be neglected [XZ14]. Furthermore, the power $p_{n,i}^{\mathrm{Circ}}$ consumed by the circuit and the transmit power $p_{n,i}^{\mathrm{Tx}}$ used for data transmission in time interval i are defined as

$$p_{n,i}^{\mathrm{Circ}} = \frac{E_{n,i}^{\mathrm{Circ}}}{\tau}, \tag{2.1}$$

and

$$p_{n,i}^{\mathrm{Tx}} = \frac{E_{n,i}^{\mathrm{Tx}}}{\tau}, \tag{2.2}$$

respectively. Moreover, $p_{n,i}^{\mathrm{Circ}}$ and $p_{n,i}^{\mathrm{Tx}}$ are assumed to be constant for the duration τ of one time interval. The value of the power consumed by the circuit depends on the considered hardware while the value of the transmit power is adjusted in each time interval in order to maximize a utility function, e.g., the throughput.

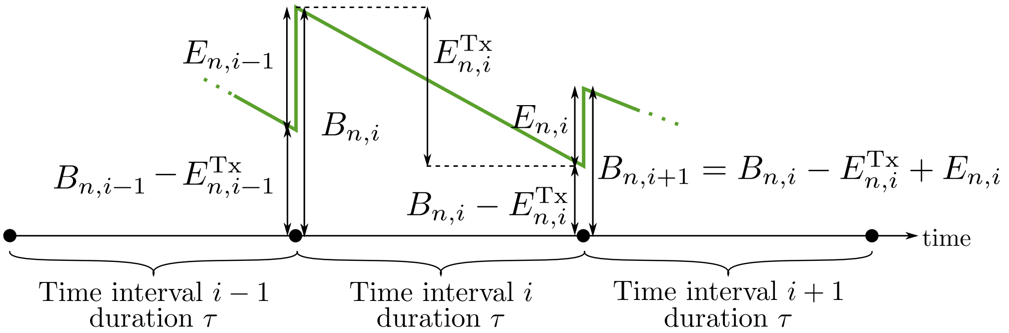

$E_{n,i-1}$
$E_{n,i}^{\mathrm{Tx}}$
$B_{n,i}$
$E_{n,i}$
$B_{n,i-1} - E_{n,i-1}^{\mathrm{Tx}}$
$B_{n,i} - E_{n,i}^{\mathrm{Tx}}$
$B_{n,i+1} = B_{n,i} - E_{n,i}^{\mathrm{Tx}} + E_{n,i}$
time
Time interval $i-1$ duration τ
Time interval i duration τ
Time interval $i+1$ duration τ

Figure 2.2. Diagram of the EH model assuming $E^{\mathrm{Circ}} = 0$.

The EH model is depicted in Figure 2.2 where three time intervals $i-1$, i and $i+1$ are considered. To simplify the figure, we assume that $E_{n,i}^{\mathrm{Circ}} = 0$. The green line represents the amount of energy in the battery while the black arrows indicate the values of different parameters. On the left side, the figure shows the battery level $B_{n,i}$ measured at the beginning of time interval i. $B_{n,i}$ is composed of the amount $E_{n,i-1}$ of harvested energy received at the end of the previous time interval $i-1$ and the remaining amount $B_{n,i-1} - E_{n,i-1}^{\mathrm{Tx}}$ of energy in the battery after the data transmission in time interval $i-1$. In time interval i, an amount $E_{n,i}^{\mathrm{Tx}}$ of energy is consumed due to the transmission of data. The new battery level $B_{n,i+1}$ is then determined at the beginning of time interval $i+1$ considering the amount $E_{n,i}$ of harvested energy received at the end of time interval i and the amount $B_{n,i} - E_{n,i}^{\mathrm{Tx}}$ of energy remaining in the battery after the data transmission in time interval i. In general, the battery level is calculated as

$$B_{n,i+1} = \max\left\{B_{\max,n},\, B_{n,i} - E_{n,i}^{\mathrm{Tx}} + E_{n,i} - E_{n}^{\mathrm{Circ}}\right\}. \tag{2.3}$$

Only the energy already stored in the battery can be used for data transmission. As a result, the energy causality constraint

$$\tau(p_{n,i}^{\mathrm{Circ}} + p_{n,i}^{\mathrm{Tx}}) \leq B_{n,i}, \tag{2.4}$$

has to be fulfilled by any feasible power allocation solution. Moreover, for the selection of the transmit power $p_{n,i}^{\mathrm{Tx}}$, the finite capacities of the batteries have to be considered. Specifically, battery overflow situations, in which part of the harvested energy is lost because the batteries are full, should be avoided as they are suboptimal. A battery overflow is a suboptimal solution because a higher throughput can always be achieved if a higher $p_{n,i}^{\mathrm{Tx}}$ is selected. The battery overflow constraint is given by

$$B_{n,i} - \tau p_{n,i}^{\mathrm{Tx}} + E_{n,i} - E_{n,i}^{\mathrm{Circ}} \leq B_{\max,n}. \tag{2.5}$$

Table 2.2. Summary of the parameters associated to the data arrival model.

Parameter	Description
$D_{\max,n}$	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
$R_{n,i}$	Throughput of EH node N_n in time interval i

2.2.3. Data arrival model

In this section, the data arrival model considered throughout this thesis is presented. This model includes both, the data arrival and the transmission of data. Specifically, two cases are distinguished:

- Infinitely full data buffer
- Finite data buffer

These two cases are motivated by the fact that in the considered scenarios, the achievable throughput is limited by the EH, data arrival and channel fading processes. Therefore, it is interesting to evaluate the performance, on the one hand, when the achievable throughput is only limited by the availability of energy, i.e., the data buffer is infinitely full, and on the other hand, when the data arrival process also plays a role. The consideration of an infinitely full data buffer allow us to determine the maximum achievable throughput, while the consideration of a data arrival process gives a more realistic view of the performance. All the parameters associated to the data arrival model presented in this section are summarized in Table 2.2.

In the case of an infinitely full data buffer, it is assumed that the EH node N_n is equipped with a data buffer of infinite size $D_{\max,n}$, and that it has an infinite amount of data to transmit, i.e, it is fully backlogged. This situation is modeled as a data buffer whose data buffer level $D_{n,i}$ is infinite for all the time intervals $i = 1, ..., I$.

When a finite buffer is considered, the data available for transmission at the EH node N_n is the result of its own data arrival process whose probability distribution depends on the considered application. For example, the incoming data could be measurements gathered by the sensors or data forwarded by another communication node. In this thesis, we focus on the transmission strategies that aim at maximizing the throughput. Therefore, it is assumed that the data to be transmitted does not have deadlines that need to be fulfilled. In our model, it is assumed that at the end of each time interval

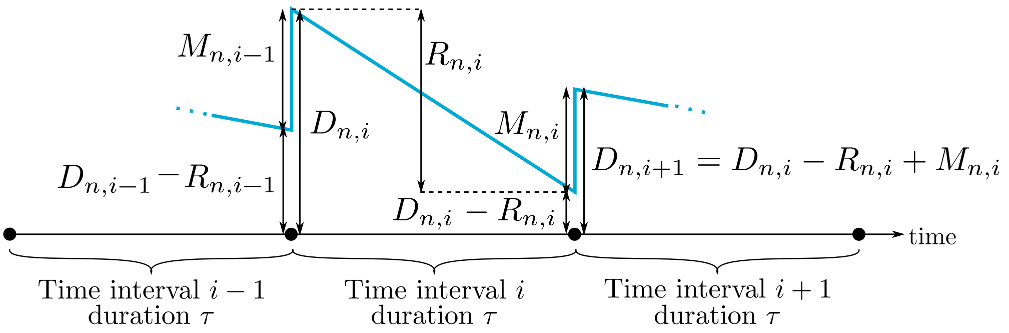

$M_{n,i-1}$
$R_{n,i}$
$D_{n,i}$
$M_{n,i}$
$D_{n,i-1} - R_{n,i-1}$
$D_{n,i+1} = D_{n,i} - R_{n,i} + M_{n,i}$
$D_{n,i} - R_{n,i}$
time
Time interval $i-1$ duration τ
Time interval i duration τ
Time interval $i+1$ duration τ

Figure 2.3. Diagram of the data arrival model when finite data buffers are considered.

i, an amount $M_{n,i}$ of incoming data, measured in bits, is arriving at EH node N_n. $M_{n,i} = 0$ represents the case when there is no new data arriving at N_n at the end of time interval i. The incoming data is stored in a finite data buffer with size $D_{\max,n}$, also measured in bits. Moreover, the data buffer level $D_{n,i}$ is measured at the beginning of time interval i. Similar to the EH model, it is assumed that at the beginning of each time interval i, only the data received in the previous time intervals j, $j \leq i-1$, is stored in the data buffer. Additionally, we consider that at the beginning of time interval $i = 1$, no data has yet arrived to the nodes and the data buffers are empty, i.e., $D_{n,1} = 0$. The throughput of EH node N_n in time interval i is defined as the amount of data transmitted by N_n in time interval i and it is denoted by $R_{n,i}$, measured in bits.

The data arrival model when finite data buffers are considered is depicted in Figure 2.3. In the figure, three time intervals, $i-1$, i and $i+1$, are considered. The blue line represents the amount of data stored in the data buffer and the black arrows represent the different parameters listed in Table 2.2. On the left side of the figure, we show the data buffer level $D_{n,i}$ measured at the beginning of time interval i. The value of $D_{n,i}$ depends on the amount $M_{n,i-1}$ of incoming data arriving at the end of time interval $i-1$ and the remaining amount $D_{n,i-1} - R_{n,i-1}$ of data in the data buffer after the data transmission in time interval $i-1$. In time interval i, an amount $R_{n,i}$ of data which corresponds to the achieved throughput is retrieved from the data buffer. As a result, the new data buffer level $D_{n,i+1}$ is determined at the beginning of time interval $i+1$ considering the amount $M_{n,i}$ of incoming data arriving at the end of time interval i and the remaining amount $D_{n,i} - R_{n,i}$ of data in the data buffer. In general, the data buffer level is calculated as

$$D_{n,i+1} = \{D_{\max,n},\, D_{n,i} - R_{n,i} + M_{n,i}\}\,. \tag{2.6}$$

Naturally, only data already stored in the data buffer can be transmitted. Therefore, the data causality constraint

$$R_{n,i} \leq D_{n,i} \tag{2.7}$$

has to be fulfilled in every time interval. Moreover, in order not to lose data, data buffer overflows should be avoided. Similar to the battery overflow situations, data buffer overflows are cases in which the incoming data is lost because the size of the data buffer has been reached. It should be noted that, in contrast to the battery overflow situations, data buffer overflows cannot always be avoided. This is because the transmission of data depends on the available energy. Therefore, for a given EH profile, the harvested energy might not be enough to transmit all the data that arrives. Nevertheless, we aim at reducing the number of data buffer overflows in order to maximize the throughput. Therefore, we define the data buffer overflow condition in an analogous way to the battery overflow constraint in (2.5) as

$$D_{n,i} - R_{n,i} + M_{n,i} \leq D_{\max,n}. \tag{2.8}$$

In addition to the throughput $R_{n,i}$, we define the sum throughput R_n as the amount of data transmitted by N_n over a given time horizon composed of I time intervals, which is calculated as

$$R_n = \sum_{i=1}^{I} R_{n,i}. \tag{2.9}$$

2.2.4. Channel model

In this section, the considered channel model is described. Note that throughout this thesis, the system is considered in the equivalent baseband.

In this thesis, the nodes are assumed to be equipped with a single antenna which has a gain $G_n = 1$. Furthermore, for the transmission of data from N_n to N_{n+1} in time interval i, a transmit power $p_{n,i}^{\mathrm{Tx}}$ is selected by N_n. Depending on the distance r between N_n and N_{n+1}, the power $p_{n+1,i}^{\mathrm{Rx}}$ of the received signal at N_{n+1} is attenuated according to the path loss. When free space transmission is considered, the path loss PL is calculated as

$$\mathrm{PL} = \left(\frac{c_0}{4 f_0 \pi r}\right)^2 \tag{2.10}$$

where c_0 is the speed of light and f_0 is the carrier frequency of the transmitted signal. However, when other environments are considered, e.g., urban areas, the received power $p_{n+1,i}^{\mathrm{Rx}}$ decreases with a higher power α of the distance r as

$$p_{n+1,i}^{\mathrm{Rx}} \propto \frac{1}{r^{\alpha}}, \tag{2.11}$$

where α is termed the path loss exponent [Skl17].

Table 2.3. Summary of the parameters associated to the channel model.

Parameter	Description
c_0	Speed of light
f_0	Carrier frequency of the transmitted signal
G_n	Antenna gain of node N_n
$g_{n,i}$	Channel gain of the link associated to node N_n in time interval i
$h_{n,i}$	Channel coefficient of the link associated to node N_n in time interval i
$p_{n,i}^{\mathrm{Tx}}$	Transmit power used by node N_n in time interval i
$p_{n,i}^{\mathrm{Rx}}$	Power of the received signal at node N_n in time interval i
r	Distance between two nodes
W	Bandwidth
α	Path loss exponent
σ_n^2	Noise variance of node N_n

In addition to the path loss, the received signal at N_{n+1} is affected by fast fading. Fast fading is a consequence of multipath propagation, i.e., the non-coherent superposition of multiple signals at the receiver due to reflection, diffraction and scattering, of the transmitted signal [Skl17]. Assuming a sufficiently large number of multipath components, the channel between two nodes N_n and N_{n+1} is described by the channel coefficient $h'_{n,i} \in \mathbb{C}$. This complex channel coefficient $h'_{n,i}$, which does not consider the distance law in (2.10), is written as

$$h'_{n,i} = X + jY, \tag{2.12}$$

where X and Y are two independent and identically distributed (i.i.d.) zero mean Gaussian random variables with equal variance [NM93]. Therefore, the channel coefficient $h_{n,i}$ which considers both, pathloss and fast fading, is modeled as

$$h_{n,i} = \sqrt{\mathrm{PL}} h'_{n,i}. \tag{2.13}$$

Moreover, a block fading model is assumed in which the channel coefficients $h_{n,i}$ stay constant for the duration τ of one time interval. Additionally, the channel gain $g_{n,i}$ is defined as

$$g_{n,i} = |h_{n,i}|^2. \tag{2.14}$$

The noise at node N_n is assumed to be i.i.d. zero mean additive white Gaussian noise (AWGN) and the interference is treated as noise. The resulting noise variance is denoted by σ_n^2. Additionally, a bandwidth W is assumed to be available for the transmission of data and the throughput in one time interval is approximated using Shannon's capacity formula since it provides the upper bound of the achievable throughput. For this purpose, the signal $x_{n,i}$ transmitted by N_n in time interval i is assumed to be zero mean circularly symmetric complex Gaussian distributed [NM93]. A summary of all the parameters introduced in this section is provided in Table 2.3.

2.3. Markov decision process

2.3.1. MDPs for EH scenarios

In the scenarios considered in this thesis, the energy is harvested over time. As a result, the EH nodes are faced with the problem of how to efficiently use the available resources, e.g., the harvested energy, in order to maximize the performance. Such decision-making situations can be modeled using Markov decision processes (MDPs), specially when learning approaches are considered. This is because MDPs are a suitable mathematical tool for modeling problems in which sequential decisions need to be made [BGD13].

In this section, we present a brief and formal description of MDPs. First, in Section 2.3.2 we introduce the finite MDP including the key elements of its structure such as actions, states, rewards and policies. Then, in Section 2.3.3 the state-value and action-value functions are defined. These functions are useful to evaluate the suitability of the policies that provide a solution for the MDP. Last, in Section 2.3.4, we extend the definition of MDP to the infinite case and describe how linear function approximation can be used to overcome the challenge that an infinite MDP conveys.

2.3.2. Finite MDP

An MDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathsf{P}, \mathcal{R} \rangle$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, P corresponds to a transition model and $\mathcal{R}$ denotes a set of rewards [RN10]. The MDP is said to be finite if the sizes of the state, action and rewards sets, as well as the transition model, are finite [SB18]. Furthermore, the MDP is defined with respect to the decision making agent or, in learning approaches, the learner. As a result, the set $\mathcal{S}$ contains all the possible states S the agent can experience. In general, the states are determined by the environment in which the agent is situated. For example, in the EH point-to-point scenario, the agent could be the EH transmitter and the states correspond to the amounts of harvested energy, the battery levels and the channel gains. The actions $a \in \mathcal{A}$ determine how the agent interacts with the environment. Following the previous example, the actions could correspond to the transmit power values that can be selected by the transmitter. The action dependent transition model P defines the transition probabilities $P^{a_i}_{S_i,S_{i+1}} = \mathbb{P}\left[S_{i+1}|S_i, a_i\right]$ of going from state S_i to the next state S_{i+1} as a consequence of selecting action a_i in time interval i. Note that the probabilities $P^{a_i}_{S_i,S_{i+1}}$, completely characterize the dynamics of the environment.

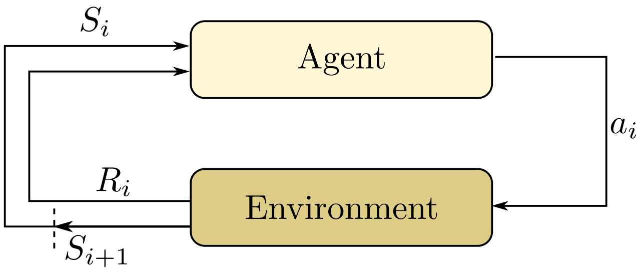


Figure 2.4. Schematic of the interaction between the agent and the environment in an MDP [SB18].

Finally, the rewards $R_i \in \mathcal{R}$ indicate how beneficial it is for the agent to select action a_i when it experiences state S_i.

In most of the applications, the agent aims at maximizing the sum of the rewards obtained over a certain number I of time intervals. In other words, it maximizes

$$R = \sum_{i=1}^{I} R_i. \tag{2.15}$$

However, in some cases, the agent can continue to make decisions and obtain rewards for an undetermined amount of time, e.g., an EH transmitter harvests energy and transmits data for as long as it is operative. For such cases, the discount factor γ is considered in order to account for the preference of achieving a higher throughput in the current time interval versus achieving a higher throughput later on. In other words, the discount factor allows us to determine, in the current time interval, the value of the future rewards. As a result, the aim is to maximize the discounted sum of rewards defined as

$$R := \lim_{I \to \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^{i-1} R_i\right] \tag{2.16}$$

[SB18], where $\mathbb{E}\left[\cdot\right]$ denotes the expected value. When $\gamma \to 0$, the agent aims at maximizing the reward in each time interval. On the contrary, as γ approaches one, the future rewards are taken more into account.

The interaction between the agent and the environment is depicted in Figure 2.4. In a given time interval i, the environment is in state $S_i \in \mathcal{S}$. The agent observes this current state and based on it, selects the action $a_i \in \mathcal{A}$ which produces a transition in the environment to a new state $S_{i+1} \in \mathcal{S}$. Additionally, due to the selected action a_i, the agent receives a reward R_i which is a feedback of the effect of the selected action. The cycle continues as long as the agent is operative.

2.3.3. Value functions

The solution of an MDP is given by a policy [RN10]. This policy, termed π, is a mapping from a given state S_i to the probability of selecting action a_i. In other words, $\pi(a_i|S_i)$ is the probability of selecting action a_i when state S_i is encountered. Nevertheless, if the policy π is deterministic, then $\pi(S_i)$ indicates the action a_i that should be selected, i.e., $a_i = \pi(S_i)$ [SB18].

Many different policies can provide a solution for a given MDP. However, not all of them yield optimum behavior. Therefore, in order to measure how good a policy π is, the so-called value functions can be used. Specifically, the state-value function, termed $\mathrm{V}^{\pi}(S_i)$, is defined as the expected reward considering a certain initial state S_i and following policy π afterwards. Formally, $\mathrm{V}^{\pi}(S_i)$ is defined as

$$\mathrm{V}^{\pi}(S_i) := \mathbb{E}_{\pi}\left\{\sum_{j=0}^{\infty}\gamma^{j}R_{i+j+1}\middle| S_i\right\} \tag{2.17}$$

[SB18]. Similarly, the action-value function, termed $\mathrm{Q}^{\pi}(S_i, a_i)$, is defined as the expected reward considering that in a certain initial state S_i, action a_i is selected and policy π is followed thereafter. Formally, $\mathrm{Q}^{\pi}(S_i, a_i)$ is defined as

$$\mathrm{Q}^{\pi}(S_i, a_i) := \mathbb{E}_{\pi}\left\{\sum_{j=0}^{\infty}\gamma^{j}R_{i+j+1}\middle| S_i, a_i\right\} \tag{2.18}$$

[SB18]. Note that both value functions indicate the expected reward of a given policy π and both can be used to evaluate it. The difference between them is what they consider as a starting point, namely, a state S_i for the state-value function or a state-action pair (S_i, a_i) for the action-value function. Furthermore, the decision of which value-function to use for the evaluation of the policy depends on the considered problem, e.g., in dynamic programing the use of $\mathrm{V}^{\pi}(S_i)$ is favored [Bel54] while the consideration of $\mathrm{Q}^{\pi}(S_i, a_i)$ facilitates the design of learning algorithms [SB18]. This is because the suitability of the value functions depends on the availability of a perfect model of the environment, i.e., perfect knowledge about all the transition probabilities $P^{a_i}_{S_i,S_{i+1}}$ and possible rewards $R_i(S_i, a_i)$[1]. In particular, when such model of the environment is assumed to be available, e.g., in dynamic programing, the state-value function $\mathrm{V}^{\pi}(S_i)$ is sufficient to determine the policy. This is due to the fact that the transition probabilities and the rewards can be used to determine which one of the possible actions leads to the best combination of current reward and next state. However, if such a model is

[1] Note that here we have written $R_i(S_i, a_i)$ instead of R_i to emphasize the fact that the reward depends on the state and the action selected.

not available, e.g., when learning approaches are considered, the state-value function is not sufficient to determine the policy. In this case, the value of each action needs to be established and therefore, the action-value function is the most suitable alternative.

A fundamental property of the value functions is that they can be written in a recursive manner in what is known as the Bellman equations [SB18]. For the state-value function, the Bellman equation is given by

$$\mathrm{V}^{\pi}(S_i) := \sum_{a_i \in \mathcal{A}} \pi(a_i|S_i) \sum_{\forall S_j \in \mathcal{S}} P^{a_i}_{S_i,S_j} \left[R_i + \gamma \mathrm{V}^{\pi}(S_j)\right] \tag{2.19}$$

[SB18]. Similarly, the Bellman equation for the action-value function is

$$\mathrm{Q}^{\pi}(S_i, a_i) := \sum_{S_j \in \mathcal{S}} P^{a_i}_{S_i,S_j} \left[R_i + \gamma \sum_{a_j \in \mathcal{A}} \pi(a_j|S_j) \mathrm{Q}^{\pi}(S_j, a_j)\right] \tag{2.20}$$

[SB18]. As it will become clear throughout this dissertation, this recursive representation facilitates the evaluation of the policies and the design of RL algorithms. Moreover, note that the value functions can be calculated from each other as

$$\mathrm{V}^{\pi}(S_i) = \sum_{a_i \in \mathcal{A}} \pi(a_i|S_i) \mathrm{Q}^{\pi}(S_i, a_i) \tag{2.21}$$

and

$$\mathrm{Q}^{\pi}(S_i, a_i) = \sum_{S_j \in \mathcal{S}} P^{a_i}_{S_i,S_j} \left[R_i + \gamma \mathrm{V}^{\pi}(S_j)\right]. \tag{2.22}$$

Considering the value functions, the optimal policy π^* is the policy for which the outcome of its respective value functions is greater than or equal to the outcome of any other policy for every state and action. This means that the state-value function of the optimal policy, denoted by V^*, is larger than or equal to the state-value function of any other policy for every state S_i. This is because $\mathrm{V}^{\pi}(S_i)$ indicates the expected reward to be achieved when starting in the considered state S_i and following the policy π afterwards, i.e., it considers the future transitions and future rewards. Similarly, the action-value function of the optimal policy, denoted by Q^*, is larger than or equal to the action-value function of any other policy for every possible state-action pair (S_i, a_i). The Bellman equations defined in (2.19) and (2.20) can be defined for the optimal state-value function and the optimal action-value function as

$$\mathrm{V}^*(S_i) := \max_{a_i} \sum_{\forall S_j \in \mathcal{S}} P^{a_i}_{S_i,S_j} \left[R_i + \gamma \mathrm{V}^*(S_j)\right] \tag{2.23}$$

[SB18] and

$$\mathrm{Q}^*(S_i, a_i) := \sum_{S_j \in \mathcal{S}} P^{a_i}_{S_i,S_j} \left[R_i + \gamma \max_{a_j \in \mathcal{A}} \mathrm{Q}^*(S_j, a_j)\right] \tag{2.24}$$

[SB18], respectively. Furthermore, by considering the optimal policy π^*, the relation between the optimal state-value function V^* and the optimal action-value function Q^* is given by

$$\mathrm{V}^*(S_i) = \max_{a_i \in \mathcal{A}} \mathrm{Q}^*(S_i, a_i) \tag{2.25}$$

[SB18]. Note that as explained before, determining the optimal actions becomes easier when Q^* is known because for each state S_i, any action a_i that maximizes $\mathrm{Q}^*(S_i, a_i)$ is an optimal action. Consequently, any policy formed by the collection of optimal actions is an optimal policy π^*.

2.3.4. Infinite MDP and linear function approximation

In the previous sections it was assumed that the state, action and reward sets have a finite size. Therefore, the value functions could be seen as tables that store the expected reward for every state or state-action pair. However, in the EH scenario the set $\mathcal{S}$ is, in general, infinite. This is because the states depend on the amounts of harvested energy, the battery levels, the data buffer levels and the channel coefficients, all of which can take values in a continuous range. As a consequence of having an infinite number of states, a table containing the values of the value function for each of the infinitely many possible states cannot be constructed. Therefore, suitable and computationally feasible techniques to represent the value functions V^* and Q^* are necessary. To this aim, the concept of function approximation is exploited in which the original value function is approximated using a computationally tractable representation.

Function approximation is an instance of supervised learning used to approximate a certain function given a set of training samples and it has been extensively studied [BBSE10, GWT$^+$13, SB18, Rip96]. In the context of the value functions, function approximation methods like linear function approximation, artificial neural networks, multi-variate regression or decision trees can be explored for the representation of the value functions given the infinite possible states. In this dissertation, we focus on linear function approximation because it allows the derivation of convergence guarantees for the proposed learning algorithms.

With linear function approximation, the value functions are represented as a linear combination of F feature functions, as depicted in Figure 2.5 where three linear feature functions are considered in the approximation. Depending on the value function being considered, state-value function or action-value function, the feature function will map the state or the state-action pair onto a feature value. In other words, for the approximation of the state-value function $\mathrm{V}^\pi(S_i)$, the feature function $\mathrm{f}_f^{\mathrm{V}}(S_i),\ f = 1, ..., F^{\mathrm{V}}$,

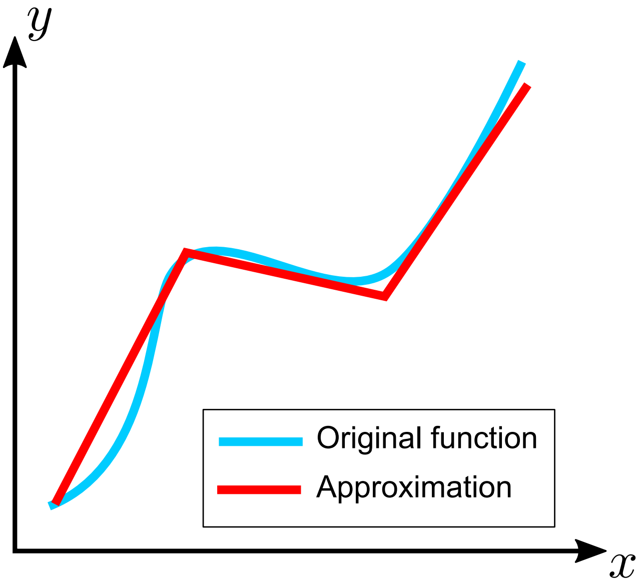


Figure 2.5. Example of the use of linear function approximation.

maps the current state onto a feature value. On the contrary, when the action-value function $\mathrm{Q}^{\pi}(S_i, a_i)$ is considered, the feature function $\mathrm{f}_f^{\mathrm{Q}}(S_i, a_i)$, $f = 1, ..., F^{\mathrm{V}}$, maps the current state-action pair onto a feature value. Note that we have added the superscripts V and Q to differentiate these two cases. Furthermore, the collection of the feature values, i.e., the value each feature function takes given a certain state or state-action pair, can be written as the vectors $\mathbf{f}^{\mathrm{V}} \in \mathbb{R}^{F^{\mathrm{V}} \times 1}$ and $\mathbf{f}^{\mathrm{Q}} \in \mathbb{R}^{F^{\mathrm{Q}} \times 1}$ for V^{π} and Q^{π}, respectively.

The contribution of each feature to the value of the value functions is taken into account via a vector of weights. For the state-value function, the vector is termed $\mathbf{w}^{\mathrm{V}} \in \mathbb{R}^{F^{\mathrm{V}} \times 1}$ and for the action-value function, the vector of weights is denoted by $\mathbf{w}^{\mathrm{Q}} \in \mathbb{R}^{F^{\mathrm{Q}} \times 1}$. With the previous definitions, the state-value and action-value functions are approximated as

$$\mathrm{V}^{\pi}(S_i) \approx \hat{\mathrm{V}}^{\pi}(S_i, \mathbf{w}^{\mathrm{V}}) = \left(\mathbf{f}^{\mathrm{V}}\right)^{\mathrm{T}} \mathbf{w}^{\mathrm{V}}, \tag{2.26}$$

and

$$\mathrm{Q}^{\pi}(S_i, a_i) \approx \hat{\mathrm{Q}}^{\pi}(S_i, a_i, \mathbf{w}^{\mathrm{Q}}) = \left(\mathbf{f}^{\mathrm{Q}}\right)^{\mathrm{T}} \mathbf{w}^{\mathrm{W}} \tag{2.27}$$

[SB18], where $(\mathbf{x})^{\mathrm{T}}$ denotes the transpose of vector $\mathbf{x}$. Additionally, note that in general, the true action-value function cannot be perfectly represented, but only approximated [SB18].

Chapter 3

Energy harvesting point-to-point scenario

3.1. Introduction

In this chapter, offline and learning approaches which find the power allocation policy that aims at maximizing the throughput in an EH point-to-point communication scenario are investigated.

The chapter is organized as follows. In Section 3.2, the considered scenario and the corresponding system assumptions are described. In Section 3.3, the throughput maximization problem for the considered scenario is formulated. The offline optimum solution derived in [OGE12] and [OTY$^+$11] for the throughput maximization in an EH point-to-point scenario is explained in Section 3.4. In Section 3.5, the proposed learning approach is described. This description comprises the modeling of the problem as a Markov decision process, the use of linear function approximation, the proposed feature functions, convergence guarantees and the computational complexity analysis of the proposed algorithm. In Section 3.6, numerical simulation results are presented in which the proposed approach is compared to the offline optimum solution, the reference learning scheme in [BGD13], a hasty policy that depletes the battery in each time interval and the random power allocation policy. Finally, Section 3.7 concludes the chapter.

Parts of this Chapter 3 have been originally published by the author in [OASL$^+$16a] and [OASL$^+$16b].

3.2. Scenario description and assumptions

In this section, the description of the considered EH point-to-point communication scenario is presented. Moreover, based on the system model described in Section 2.2, the corresponding assumptions are introduced. A summary of all the parameters associated to the EH point-to-point communication scenario is presented in Table 3.1.

Table 3.1. Parameters associated to the EH point-to-point communication scenario.

	Parameter	Description
General	i	Index of the time interval
	I	Total number of time intervals
	N_1	EH transmitter node
	N_2	non-EH receiver node
	τ	Time interval duration
Energy	$B_{1,i}$	Battery level of EH node N_1, measured at the beginning of time interval i
	$B_{\max,1}$	Battery capacity of EH node N_1
	$E_{1,i}$	Amount of harvested energy, received at the end of time interval i, by EH node N_1
	$E_{1,i}^{\mathrm{Circ}}$	Amount of energy consumed by the circuit of EH node N_1 in time interval i
	$E_{1,i}^{\mathrm{Tx}}$	Energy of the signal transmitted by EH node N_1 in time interval i
	$E_{\max,1}$	Maximum amount of energy that can be harvested by EH node N_1
	$p_{1,i}^{\mathrm{Tx}}$	Transmit power used by EH node N_1 in time interval i
Data	$D_{\max,1}$	Data buffer size of EH node N_1
	$D_{1,i}$	Data buffer level of EH node N_1, measured at the beginning of time interval i
	$M_{1,i}$	Amount of incoming data, arriving at the end of time interval i, at EH node N_1
	$R_{1,i}$	Amount of data transmitted from EH node N_1 to non-EH node N_2 in time interval i
Channel	$g_{1,i}$	Channel gain of the link between N_1 and N_2
	$h_{1,i}$	Channel coefficient of the link between N_1 and N_2
	W	Bandwidth
	σ_2^2	Noise power at N_2

The EH point-to-point communication scenario consists of two single-antenna nodes that communicate in a single direction. As depicted in Figure 3.1, the transmitter N_1 is an EH node which harvests energy from the environment and uses it for transmitting data to the receiver N_2. An amount of harvested energy, denoted by $E_{1,i}$, is received at the end of every time interval i, $i = 1, ..., I$, and it is stored in a battery with maximum capacity $B_{\max,1}$. The battery level $B_{1,i}$ is measured at the beginning of each time interval i and indicates the amount of energy available in the battery of N_1. Furthermore, the maximum amount of energy that can be harvested is termed $E_{\max,1}$. In every time interval i, the transmitter N_1 uses a transmit power $p_{1,i}^{\mathrm{Tx}}$ for the duration τ of the time interval. As a result, an amount of energy $E_{1,i}^{\mathrm{Tx}} = \tau p_{1,i}^{\mathrm{Tx}}$ is used for data transmission. Moreover, the energy $E_{1,i}^{\mathrm{Circ}}$ consumed by the circuit at N_1 when it transmits data is assumed to be constant for all the time intervals i. This means, $E_{1,i}^{\mathrm{Circ}} = E_{1}^{\mathrm{Circ}}\ \forall i$. Considering (2.3), the battery level is calculated as

$$B_{1,i+1} = \min\left\{B_{\max,1},\ B_{1,i} - E_{1,i}^{\mathrm{Tx}} + E_{1,i} - E_{1}^{\mathrm{Circ}}\right\}. \tag{3.1}$$

Additionally, the receiver node N_2 is assumed to be connected to a fixed power supply, e.g., the electrical grid, and it is always available to receive the transmitted data.

As described in Section 2.2.3, we consider two cases regarding the data arrival model, namely, when N_1 has an infinitely full data buffer and when N_1 is equipped with a finite data buffer. In the first case, the data buffer size $D_{\max,1}$ is assumed to be infinite. Moreover, it is assumed that the buffer level $D_{i,1}$ is also infinite for all the time intervals.

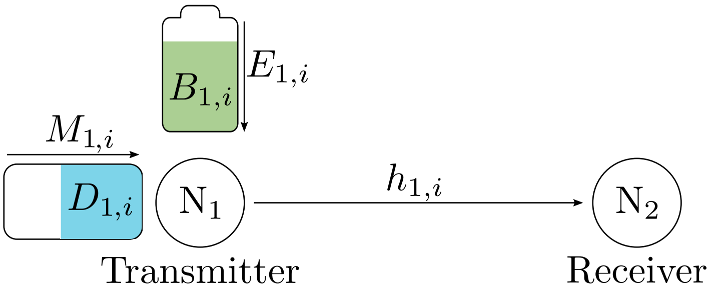


Figure 3.1. Point-to-point communication scenario with an EH transmitter node.

In the second case, an amount $M_{1,i}$ of incoming data arrives at N_1 at the end of each time interval i and it is stored in a finite data buffer with size $D_{\max,1}$. The data buffer level $D_{1,i}$ is measured at the beginning of time interval i and indicates the amount of data available at N_1 for the transmission to N_2. Moreover, the throughput $R_{1,i}$ denotes the amount of data transmitted to N_2 in time interval i. Considering (2.6), the data buffer level is updated as

$$D_{1,i+1} = \min \{D_{\max,1},\ D_{1,i} - R_{1,i} + M_{1,i}\} . \tag{3.2}$$

In case there is enough data in the data buffer, the throughput $R_{1,i}$ is approximated using Shannon's capacity formula as

$$R_{1,i} = W\tau \log_2 \left(1 + \frac{g_{1,i} p_{1,i}^{\mathrm{Tx}}}{\sigma_2^2}\right) . \tag{3.3}$$

where W is the available bandwidth, $g_{1,i} = |h_{1,i}|^2$ is the channel gain of the link between N_1 and N_2, and σ_2^2 is the noise power at N_2. Otherwise, the throughput $R_{1,i}$ is limited by the amount of data stored in the data buffer.

Additionally, transmitter side channel state information is assumed to be available. Depending on the considered approach, i.e., offline or learning, this channel state information is assumed to be causally or non-causally known. In the causal case, only the current and past channel gains are assumed to be known at the transmitter in each time interval i. This means, the transmitter knows the channel gains for all time intervals $j = 1, ..., i$. In the non-causal case, the channel gains of all the time intervals, $i = 1, ..., I$ are assumed to be known at the transmitter at the beginning of the data transmission, i.e., at the beginning of time interval $i = 1$.

3.3. Problem formulation

In this section, the power allocation problem for throughput maximization in the EH point-to-point communication scenario is formulated.

For the considered scenario, our goal is to find a transmission policy at node N_1 that maximizes the throughput, which is defined as the total amount of data transmitted to node N_2, while considering the energy causality and battery overflow constraints described in Section 2.2.2, as well as the data causality and data overflow constraints introduced in Section 2.2.3.

The power allocation problem for throughput maximization is given by

$$\left(p_{1,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{1,i} = \underset{\{p_{1,i}^{\mathrm{Tx}}, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_{1,i} \tag{3.4a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \tau p_{1,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_1^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{1,i},\ J = 1, \ldots, I, \tag{3.4b}$$

$$\sum_{i=1}^{J} E_{1,i} - \sum_{i=1}^{J} \tau p_{1,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_1^{\mathrm{Circ}} \leq B_{\mathrm{max},1},\ \forall J, \tag{3.4c}$$

$$\sum_{i=1}^{J} R_{1,i} \leq \sum_{i=1}^{J-1} M_{1,i},\ \forall J, \tag{3.4d}$$

$$\sum_{i=1}^{J} M_{1,i} - \sum_{i=1}^{J} R_{1,j} \leq D_{\mathrm{max},1},\ \forall J \tag{3.4e}$$

$$p_{1,i}^{\mathrm{Tx}} \geq 0, \quad i = 1, \ldots, I. \tag{3.4f}$$

where $R_{1,i}$ is given in (3.3). The constraint in (3.4b) is derived from the energy causality constraint in (2.4), (3.4c) comes from the battery overflow constraint in (2.5), (3.4d) is determined from the data causality constraint in (2.7) and (3.4e) is derived from the data overflow constraint in (2.8). Moreover, note that when an infinitely full data buffer is considered at N_1, the constraints in (3.4d) and (3.4e) are not taken into account.

3.4. Offline approach

In this section, the offline approach for the EH point-to-point scenario is described. The approach is based on the work of [OGE12] and [OTY+11]. However, here we have extended it to our model in which a constant time slot duration τ, a data arrival process, and a block fading channel model are assumed. In the offline approaches, it is assumed that the EH, data arrival and channel fading processes are perfectly and non-causally known. This means, before the transmission starts, the transmitter knows the amounts of energy that will be harvested, the amount of data that will be received for transmission and how the channel gains will vary. Such assumption, although unrealistic, allows us to find the upper bound of the performance.

By taking a closer look at the problem in (3.4), it can be observed that feasibility cannot be guaranteed when a data arrival process is considered. This is because the data overflow constraint in (3.4e) might not be fulfilled when the amount of harvested energy is not sufficient to deplete the data buffer. As a result, for the analysis of the throughput maximization problem in the EH point-to-point scenario and the derivation of an offline approach, an infinitely full data buffer is assumed. This means, the constraints in (3.4d) and (3.4e) are not considered and the resulting optimization problem is written as

$$\left(p_{1,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{1,i} = \underset{\{p_{1,i}^{\mathrm{Tx}},\, i=1,\dots,I\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_{1,i} \tag{3.5a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \tau p_{1,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_{1}^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{1,i},\ J = 1, \dots, I, \tag{3.5b}$$

$$\sum_{i=1}^{J} E_{1,i} - \sum_{i=1}^{J} \tau p_{1,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_{1}^{\mathrm{Circ}} \leq B_{\max,1},\ \forall J, \tag{3.5c}$$

$$p_{1,i}^{\mathrm{Tx}} \geq 0, \quad i = 1, \dots, I. \tag{3.5d}$$

The problem in (3.5) is a convex optimization problem. This is because the objective function (3.5a) is a concave function and the constraints in (3.5b)-(3.5d) are linear functions of $p_{1,i}^{\mathrm{Tx}}$. As a result, the Lagrangian function of (3.5) can be written as

$$\begin{aligned}\mathfrak{L} = &\sum_{i=1}^{I} R_{1,i} \\ &- \sum_{i=1}^{I} \mu_i \left(\sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_{1}^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{1,j} \right) \\ &- \sum_{i=1}^{I} \omega_i \left(\sum_{j=1}^{i} E_{1,j} - \sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_{1}^{\mathrm{Circ}} - B_{\max,1} \right) \\ &+ \sum_{i=1}^{I} \upsilon_i p_{1,i}^{\mathrm{Tx}},\end{aligned} \tag{3.6}$$

where μ_i, ω_i and υ_i are Lagrange multipliers.

Moreover, the corresponding Karush-Kuhn-Tucker (KKT) conditions, which are necessary conditions for a global optimum, are given by

$$\frac{\partial \mathfrak{L}}{\partial p_{1,i}^{\mathrm{Tx}}} = \frac{\tau W g_{1,i}}{(\ln 2)\left(\sigma^2 + g_{1,i} p_{1,i}^{\mathrm{Tx}}\right)} - \tau \sum_{j=i}^{I} \left(\mu_j - \omega_j\right) + \upsilon_i = 0, \tag{3.7}$$

$$\mu_i \left(\sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_{1}^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{1,j} \right) = 0, \tag{3.8}$$

$$\omega_i \left(\sum_{j=1}^{i} E_{1,j} - \sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_1^{\mathrm{Circ}} - B_{\max,1} \right) = 0, \tag{3.9}$$

$$\upsilon_i p_{1,i}^{\mathrm{Tx}} = 0, \tag{3.10}$$

$$\sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_1^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{1,j} \leq 0, \tag{3.11}$$

$$\sum_{j=1}^{i} E_{1,j} - \sum_{j=1}^{i} \tau p_{1,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_1^{\mathrm{Circ}} - B_{\max,1} \leq 0, \tag{3.12}$$

$$-p_{1,i}^{\mathrm{Tx}} \leq 0, \tag{3.13}$$

$$\mu_i \geq 0,\ \omega_i \geq 0,\ \upsilon_i \geq 0, \tag{3.14}$$

for all $i = 1, ..., I$.

From the KKT conditions it is clear that when $p_{1,i}^{\mathrm{Tx}^{\mathrm{opt}}} > 0$, $\upsilon_i = 0$ due to (3.10). Consequently, by considering (3.7), the optimal power allocation in time interval i can be calculated as

$$p_{1,i}^{\mathrm{Tx}^{\mathrm{opt}}} = \nu_i - \frac{\sigma^2}{g_{1,i}}, \quad \text{if } p_{1,i}^{\mathrm{Tx}^{\mathrm{opt}}} > \upsilon_i^{\mathrm{opt}}, \tag{3.15}$$

where ν_i can be interpreted as the water level given by

$$\nu_i = \frac{W}{(\ln 2) \sum\limits_{j=i}^{I} (\mu_j - \omega_j)}. \tag{3.16}$$

Note that in this context, the water filling interpretation is done over time, i.e., the allocation of power in the different time intervals, and not over multiple channels as it is usually done.

From (3.15) and (3.16), it is clear that the water level term ν_i implies that the selection of the transmit power to be used in time interval i depends on the future use of the harvested energy. Moreover, from the complementary slackness conditions in (3.8) and (3.9), we know that μ_i and ω_i cannot be simultaneously bigger than zero. This is because $\mu_i > 0$ holds whenever the battery is depleted, i.e., when all the harvested energy has been used. In that case, ω_i must be equal to zero so that (3.9) holds. Similarly, by examining (3.9), we can deduce that $\omega_i > 0$ holds only when the battery is full. In such a case, μ_i must be equal to zero for (3.8) to hold.

By further analyzing the problem in (3.5) and the corresponding KKT conditions in (3.7)-(3.14), the following observations can be extracted:

Proposition 3.1. *In the optimal power allocation policy, the transmit power* $p_{1,i}^{\mathrm{Tx}}$ *is constant during one time interval [YU12b].*

Proof. The proof is detailed in [YU12b]. However, we summarize it here for completeness. Assume that the transmitter changes the transmit power during one time interval, i.e., transmit power ${p'}^{\mathrm{Tx}}_{1,i}$ is used during a fraction ϱ of the time interval and a transmit power ${p''}^{\mathrm{Tx}}_{1,i}$ is used for the remaining part of the time interval, $\tau - \varrho$. The throughput $R'_{1,i}$ achieved in this case is given

$$R'_{1,i} = W\varrho \log_1 \left(1 + \frac{g_{1,i}}{\sigma^2 {p'}^{\mathrm{Tx}}_{1,i}}\right) + W(\tau - \varrho) \log_1 \left(1 + \frac{g_{1,i} {p''}^{\mathrm{Tx}}_{1,i}}{\sigma^2}\right). \tag{3.17}$$

However, due to the concavity of the throughput function in (3.3), it is easy to check that a larger throughput can be achieved if the constant transmit power

$$\bar{p}^{\mathrm{Tx}}_{1,i} = \frac{\varrho {p'}^{\mathrm{Tx}}_{1,i} + (\tau - \varrho){p''}^{\mathrm{Tx}}_{1,i}}{\tau}, \tag{3.18}$$

is used during the total duration of the time interval. □

Proposition 3.2. *In the optimal power allocation policy, the transmit power increases monotonically over time when the battery capacity is infinite and an infinitely full data buffer is considered [OTY$^+$11].*

Proof. The proof is detailed in [OTY$^+$11]. However, as in the previous case, we summarize it here for completeness. A battery of infinite capacity, i.e., $B_{\max,1} = \infty$, implies that $\omega_i = 0$ for all the time intervals due to the complementary slackness condition in (3.9). As a consequence, the water level fulfills the condition $\nu_{i+1} \geq \nu_i$. This is because $\mu_i \geq 0$ for all the time intervals. Moreover, as the battery capacity is infinite, battery overflow situations cannot occur and energy can always be saved for future use. □

To find the power allocation in each time interval, in [OGE12] the authors proposed the Directional Backward Glue Pouring algorithm. The idea behind this approach is to allocate the available energy starting from the last interval in which energy is harvested and continue the allocation backwards until the first time interval is reached. Using the Glue Pouring algorithm in [YMZM08], the authors calculate the water level for each time interval and find the optimal transmit power using (3.15). The same procedure is repeated until the first time interval is reached. The term directional stems for the fact that, due to the energy causality constraint, energy can only be shared in one direction, i.e., the energy harvested at the beginning of time interval i, can only be shared among the subsequent time intervals and not among the previous ones.

3.5. Learning approach

3.5.1. Markov decision process

In this section, we model the power allocation problem for throughput maximization in the EH point-to-point communication scenario as an MDP. As it will become clear in the next Section 3.5.2, this MDP model facilitates the design of the proposed learning approach.

In learning approaches, we consider a realistic scenario in which only causal knowledge of the system dynamics, i.e., EH, data arrival and channel fading processes, is assumed. This means, at the beginning of time interval i, N_1 only knows its current and past amounts of harvested energy $E_{1,j}$, battery levels $B_{1,j}$, data buffer levels $D_{1,j}$, and channel gains $g_{1,j}$, where $j \leq i$. Note that the battery level $B_{1,i}$, in time interval i, summarizes the history of how the harvested energy has been used up to time interval i. Similarly, the data buffer level $D_{1,i}$ summarizes how much data have been received and transmitted. Therefore, taking into account that τ is fixed and known, the selection of the transmit power $p_{1,i}^{\mathrm{Tx}}$ depends solely on the values of $E_{1,i}$, $B_{1,i}$, $D_{1,i}$ and $g_{1,i}$, i.e., the selection of $p_{1,i}^{\mathrm{Tx}}$ in time interval i does not depend on the previous values of $E_{1,i}$, $B_{1,j}$, $D_{1,j}$ and $g_{1,j}$, $j < i$. As a result, the system under consideration fulfils the Markov property and can be modeled as an MDP [RN10, SB18].

As described in Section 2.3, an MDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathsf{P}, \mathcal{R} \rangle$. At time interval i, the corresponding state $S_i \in \mathcal{S}$ is a function of the amount of harvested energy $E_{1,i}$, the battery level $B_{1,i}$, the data buffer level $D_{1,i}$ and the channel gain $g_{1,i}$. In our model, $E_{1,i}$, $B_{1,i}$, $D_{1,i}$ and $g_{1,i}$ can take values in a continuous range. As a result, the set $\mathcal{S}$ contains an infinite number of possible states given by all the combinations of their values. The set $\mathcal{A}$ of actions corresponds to the values of the transmit power that can be selected. As in practical settings [Ins17], $\mathcal{A}$ is assumed to be a finite set given by $\mathcal{A} = \left\{p_{1,i}^{\mathrm{Tx}},\ p_{1,i}^{\mathrm{Tx}} \in \{0, \delta, ..., B_{\max,1}/\tau\}\right\}$, where δ is the step size. The action dependent transition model P, described in Section 2.3, defines the transition probabilities as $P_{S_i,S_{i+1}}^{p_{1,i}^{\mathrm{Tx}}} = \mathbb{P}\left[S_{i+1}|S_i, p_{1,i}^{\mathrm{Tx}}\right]$. However, as only causal knowledge is available, the transition model P is not known. Finally, the rewards indicate how beneficial the selected transmit power $p_{1,i}^{\mathrm{Tx}}$ is for the corresponding state S_i. For each state S_i and each transmit power $p_{1,i}^{\mathrm{Tx}}$, we define the reward $R_{1,i} \in \mathcal{R}$ as the throughput achieved in time interval i, which is given by (3.3). The reward $R_{1,i}$ can be calculated at N_1 because the channel gain $g_{1,i}$ as well as the selected transmit power $p_{1,i}^{\mathrm{Tx}}$ are known at the transmitter.

As only causal knowledge regarding the EH, data arrival and channel fading processes is assumed, the amounts of energy which will be harvested by N_1 in future time intervals is unknown. Due to this uncertainty, it might be preferred to achieve a higher throughput in the current time interval over future ones. To take into account this preference, the discount factor $0 \leq \gamma \leq 1$ is considered. As explained in Section 2.3, when $\gamma \rightarrow 0$, the transmitter aims at maximizing the throughput only in the current time interval. On the contrary, when $\gamma \rightarrow 1$, the throughputs achieved in future time intervals are taken more into consideration. Our goal is to select $p_{1,i}^{\mathrm{Tx}}$, $\forall i$, in order to maximize the expected throughput which is given by

$$R = \lim_{I \rightarrow \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^{i-1} R_{1,i}\right]. \tag{3.19}$$

Note that in (3.19), we have considered that $I \rightarrow \infty$. This is because the number I of time intervals in which the EH transmitter N_1 will be operative is not known in advance. Consequently, as done in [BGD13], γ can be also interpreted as the probability of the EH transmitter being operative. This means that N_1 has a probability $1-\gamma$ of terminating its transmission in any time interval.

The solution of an MDP is given by a policy [RN10]. In our scenario, this policy, termed π, is a mapping from a given state S_i to the transmit power $p_{1,i}^{\mathrm{Tx}}$ that should be selected, i.e. $p_{1,i}^{\mathrm{Tx}} = \pi(S_i)$. Furthermore, since the transition model P is not known, the action-value function $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$, defined in (2.20), is used to measure the suitability of a policy π for the solution of the power allocation problem.

3.5.2. Approximate SARSA

3.5.2.1. RL for the EH point-to-point scenario

In this section, the proposed RL algorithm used to perform power allocation in the EH point-to-point scenario is described. To find the transmission policy π, we consider an on-policy temporal-difference RL algorithm, termed SARSA [SB18], which is based on the estimation of $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$. Furthermore, to handle the infinite number of states, we combine it with linear function approximation. The selection of SARSA is based on its favourable convergence properties when linear function approximation is used, compared to other well known RL algorithms such as Q-learning [SB18, Gor01]. In the following, the details of the algorithm are presented. First, the estimation and update of the action-value function $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ is presented. Then, the use of linear function approximation is described and the proposed feature functions are introduced.

Afterwards, the action selection policy is defined and the proposed algorithm, termed approximate SARSA, is presented. Finally, the convergence properties and computational complexity of the proposed algorithm are discussed.

3.5.2.2. Action-value function update

In SARSA, given a policy π, the action-value function $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ is estimated considering the transitions from a state-action pair $(S_i, p_{1,i}^{\mathrm{Tx}})$ to another state-action pair $(S_{i+1}, p_{i+1}^{\mathrm{Tx}})$ while obtaining reward $R_{1,i}$. This fact explains the name of the algorithm: State-Action-Reward-State-Action (SARSA) [SB18]. In other words, when N_1 is in state S_i, it selects $p_{1,i}^{\mathrm{Tx}}$ following policy π. Afterwards, it obtains a reward $R_{1,i}$ and moves to state S_{i+1}. According to the current values of $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ and the policy π, the algorithm selects the next $p_{1,i+1}^{\mathrm{Tx}}$. At this point, $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ is updated using the gained experience and the current value of $\mathrm{Q}^{\pi}(S_{i+1}, p_{1,i+1}^{\mathrm{Tx}})$. The updating rule for $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ in the SARSA algorithm is given by

$$\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}) \leftarrow \mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})(1-\zeta_i) + \zeta_i \left[R_{1,i} + \gamma \mathrm{Q}^{\pi}(S_{i+1}, p_{1,i+1}^{\mathrm{Tx}})\right], \tag{3.20}$$

where ζ_i is a small positive fraction which influences the learning rate [SB18].

3.5.2.3. Linear function approximation

As discussed in Section 2.3.4, linear function approximation is used to represent the action-value function $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$ as a linear combination of F feature functions $\mathrm{f}_f(S_i, p_{1,i}^{\mathrm{Tx}})$, $f = 1, ..., F$, where each $\mathrm{f}_f(S_i, p_{1,i}^{\mathrm{Tx}})$, maps the state-action pair $(S_i, p_{1,i}^{\mathrm{Tx}})$ into a feature value. Note that in order to simplify the notation, here we have dropped the superscript Q considered in Section 2.3.4 because only the action-value function Q^{π} is being considered.

Let $\mathbf{f} \in \mathbb{R}^{F\times 1}$ be the vector containing the collection of feature values for a given state-action pair and let $\mathbf{w} \in \mathbb{R}^{F\times 1}$ be the vector containing the weights indicating the contribution of each feature. Then, the approximated action-value function $\hat{\mathrm{Q}}^{\pi}(S_i, p_i^{\mathrm{Tx}}, \mathbf{w})$ is written as

$$\hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w}) = \mathbf{f}^{\mathrm{T}}\mathbf{w} \approx \mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}). \tag{3.21}$$

To ensure that $\hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w})$ is a good representation of $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$, the error between them has to be minimized. This can be done using a stochastic gradient descent approach in which in each time interval, the vector $\mathbf{w}$ of weights is updated in the direction that most reduces the error [SB18], i.e., in time interval i, $\mathbf{w}$ is updated as

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \frac{1}{2}\zeta\nabla_{\mathbf{w}} \left(\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}) - \hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w})\right)^2. \tag{3.22}$$

Taking into account that linear function approximation is used, the gradient of $\hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w})$ can be calculated from (3.21) as

$$\nabla_{\mathbf{w}}\hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w}) = \mathbf{f}. \tag{3.23}$$

As a result, (3.22) can be written as

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \zeta_i \left(\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}) - \hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w})\right)\mathbf{f}. \tag{3.24}$$

Moreover, based on the SARSA update in 3.20, the weights $\mathbf{w}$ are updated as

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \zeta_i \left(R_{1,i} + \gamma\hat{\mathrm{Q}}^{\pi}(S_{i+1}, p_{1,i+1}^{\mathrm{Tx}}, \mathbf{w}) - \hat{\mathrm{Q}}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}}, \mathbf{w})\right)\mathbf{f}. \tag{3.25}$$

3.5.2.4. Feature functions

An important step in the implementation of linear function approximation is the definition of the feature functions. These features functions should correspond to the natural attributes of the EH problem in order to provide a good model of the effect of possible transmit power values on the state of the transmitter [SB18]. In our scenario, the most important characteristics are the unknown EH, data arrival and channel fading processes as well as the limited battery and data buffer at N_1. Based on the results found for the offline approach, we propose a set of $F = 4$ binary feature functions which take into account the limited battery, the limited data buffer and the power allocation problem.

From (3.4b), it is clear that battery overflow situations are suboptimal and therefore, should be avoided. Consequently, the first feature function $\mathrm{f}_1(S_i, p_{1,i}^{\mathrm{Tx}})$ indicates if a given $p_{1,i}^{\mathrm{Tx}}$ avoids the overflow of the battery. Additionally, it evaluates if the given $p_{1,i}^{\mathrm{Tx}}$ fulfills the energy causality constraint in (3.4b). The binary function assigns "1" if no overflow is caused by the use of $p_{1,i}^{\mathrm{Tx}}$ in time interval i while fulfilling the energy causality constraint. The first feature function $\mathrm{f}_1(S_i, p_{1,i}^{\mathrm{Tx}})$ is written as

$$\mathrm{f}_1(S_i, p_{1,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } (B_{1,i} + E_{1,i} - \tau p_{1,i}^{\mathrm{Tx}} - E_1^{\mathrm{Circ}} \leq B_{\mathrm{max},1}) \wedge (\tau p_{1,i}^{\mathrm{Tx}} + E_1^{\mathrm{Circ}} \leq B_{1,i}) \\ 0, & \text{else,} \end{cases} \tag{3.26}$$

where $\wedge$ represents the logical conjunction operation.

The second feature function $\mathrm{f}_2(S_i, p_{1,i}^{\mathrm{Tx}})$ addresses the power allocation problem. From Section 3.4 it is known that, in the offline case, a directional backward glue pouring algorithm can be used to optimally allocate the power. However, as in our scenario the knowledge of future channel gains and energy values is unavailable, we propose to use past channel realizations to estimate the mean value of the distribution of the channel gain and to perform glue pouring considering the estimated mean value of the channel gain and the current channel realization. For the estimation, the sample mean estimator is used such that at time interval i, the estimated mean value of the channel gain $\bar{g}_{1,i}$ is calculated as

$$\bar{g}_{1,i} = \frac{1}{i} \sum_{j=1}^{i} g_{1,j}. \tag{3.27}$$

Although $E_{1,i}$ cannot be allocated in time interval i, for the glue pouring algorithm it is assumed that the available energy is $E_{1,i} + B_{1,i}$. The reason is that by performing glue pouring between $\bar{g}_{1,i}$ and $g_{1,i}$, we assume that $\bar{g}_{1,i}$ approximates the state of the channel in the subsequent time interval and consequently, the available harvested energy has to be considered. The water level ν_i is calculated as

$$\nu_i = \frac{1}{2}\left(\frac{B_{1,i} - E_1^{\mathrm{Circ}}}{\tau} + \frac{E_{1,i}}{\tau} + \sigma^2\left(\frac{1}{\bar{g}_{1,i}} + \frac{1}{g_{1,i}}\right)\right). \tag{3.28}$$

To ensure that the constraints in (3.4) are fulfilled, the power allocation value given by the glue pouring algorithm is given by

$$p_{1,i}^{\mathrm{GP}} = \min\left\{\frac{B_{1,i} - E_1^{\mathrm{Circ}}}{\tau}, \max\left\{0, \nu_i - \frac{\sigma^2}{g_{1,i}}\right\}\right\}. \tag{3.29}$$

From Section 3.5.1, we know that $p_{1,i}^{\mathrm{Tx}} \in \mathcal{A}$. As a result, the calculated $p_{1,i}^{\mathrm{GP}}$ has to be rounded such that $p_{1,i}^{\mathrm{GP}} \in \mathcal{A}$ also holds. Consequently, the second feature function $\mathrm{f}_2(S_i, p_{1,i}^{\mathrm{Tx}})$ is written as

$$\mathrm{f}_2(S_i, p_{1,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \delta \left\lfloor \frac{p_{1,i}^{\mathrm{GP}}}{\delta} \right\rfloor = p_{1,i}^{\mathrm{Tx}} \\ 0, & \text{else,} \end{cases} \tag{3.30}$$

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 $\mathrm{f}_3(S_i, p_{1,i}^{\mathrm{Tx}})$ handles the case when $E_{1,i} \geq B_{\mathrm{max},1}$. In such situations, battery overflow situation are unavoidable. Therefore, the battery should be depleted in order to minimize the energy losses due to battery overflow. The function assigns a “1” if the selected $p_{1,i}^{\mathrm{Tx}}$ is equal to the available power in the battery and it is defined as

$$\mathrm{f}_3(S_i, p_{1,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } (E_{1,i} \geq B_{\mathrm{max},1}) \wedge \left(p_{1,i}^{\mathrm{Tx}} = \delta \lfloor \frac{B_{1,i} - E_1^{\mathrm{Circ}}}{\tau\delta} \rfloor \right) \\ 0, & \text{else.} \end{cases} \tag{3.31}$$

The fourth feature function $\mathrm{f}_4(S_i, p_{1,i}^{\mathrm{Tx}})$ addresses the data causality and data buffer overflow constraints. For its definition, let $R_{1,i}^{(p_{1,i}^{\mathrm{Tx}})}$ be the throughput that would be achieved if $p_{1,i}^{\mathrm{Tx}}$ is selected. Then, $\mathrm{f}_4(S_i, p_{1,i}^{\mathrm{Tx}})$ indicates if $R_{1,i}^{(p_{1,i}^{\mathrm{Tx}})}$ fulfils both, the data causality constraint in (3.4d) and the data buffer overflow constraint in (3.4e), by assigning a “1” when the throughput $R_{1,i}^{(p_{1,i}^{\mathrm{Tx}})}$ is smaller than or equal to the current data buffer level and no data buffer overflow is caused. f_4 is defined as

$$\mathrm{f}_4(S_i, p_{1,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \left(R_{1,i}^{(p_{1,i}^{\mathrm{Tx}})} \leq D_{1,i}\right) \wedge \left(D_{1,i} + M_{1,i} - R_{1,i}^{(p_{1,i}^{\mathrm{Tx}})} \leq D_{\mathrm{max},1}\right) \\ 0, & \text{else.} \end{cases} \tag{3.32}$$

3.5.2.5. Action selection policy

In this section, we describe how the power to be used in each time interval is selected. For this purpose, the characteristics of the action selection policy π, which is followed throughout the learning process, are discussed.

When the number of states is finite and Q^{π} is known, acting greedily with respect to $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$, i.e., given S_i selecting the $p_{1,i}^{\mathrm{Tx}}$ that achieves the maximum $\mathrm{Q}^{\pi}(S_i, p_{1,i}^{\mathrm{Tx}})$, leads to the optimal policy [SB18]. This is due to the fact that $\mathrm{Q}^{\pi}(S_i, p_i^{\mathrm{Tx}})$ is the expected reward given the state-action pair $(S_i, p_{1,i}^{\mathrm{Tx}})$. Therefore, selecting the $p_{1,i}^{\mathrm{Tx}}$ that maximizes $\mathrm{Q}^{\pi}(S_i, p_i^{\mathrm{Tx}})$ means that we are selecting the $p_{1,i}^{\mathrm{Tx}}$ that leads to the highest expected reward, which in our case corresponds to the highest throughput. However, note that N_1 can only act greedily with respect to the states it has already encountered and the power values it has already used. As a consequence, if N_1 follows the greedy policy, it does not have the opportunity to discover other transmit power values that can potentially lead to higher throughput. To ensure that N_1 is able to explore new transmit power values, the ϵ-greedy policy is considered instead [SB18]. In ϵ-greedy, N_1 acts greedily with a probability (1-ϵ), this means

$$\mathbb{P}\left[p_{1,i}^{\mathrm{Tx}} = \max_{p_{1,j}^{\mathrm{Tx}} \in \mathcal{A}} \mathrm{Q}^{\pi}(S_i, p_{1,j}^{\mathrm{Tx}})\right] = 1 - \epsilon, \quad 0 < \epsilon < 1. \tag{3.33}$$

Algorithm 3.1 Approximate SARSA

```
 1: initialize $\gamma, \zeta_i, \epsilon$
 2: initialize all the weights $\mathbf{w}$ to one
 3: observe $S_i$
 4: select $p_{1,i}^{\mathrm{Tx}}$ randomly
 5: while N$_1$ is harvesting energy do
 6:     transmit using the selected $p_{1,i}^{\mathrm{Tx}}$
 7:     calculate corresponding reward $R_{1,i}$                    ▷ Eq. (3.3)
 8:     observe next state $S_{i+1}$
 9:     select next transmit power $p_{i+1}^{\mathrm{Tx}}$ using $\epsilon$-greedy
10:     update $\mathbf{w}$                                             ▷ Eq. (3.25)
11:     set $S_i = S_{i+1}$
12:     set $p_{1,i}^{\mathrm{Tx}} = p_{1,i+1}^{\mathrm{Tx}}$
13: end while
```

However, with a probability ϵ, N$_1$ will randomly select a transmit power value from the set $\mathcal{A}$, where all the transmit power values have the same probability of being selected. This method provides a trade-off between the exploration of new transmit power values and the exploitation of the known ones.

3.5.2.6. Approximate SARSA algorithm

The proposed approximate SARSA algorithm is summarized as Algorithm 3.1. At the beginning of the execution, the learning parameters are initialized (line 1). These parameters correspond to the discount factor γ, the learning rate ζ and the exploration probability ϵ. Furthermore, the weights used in the linear function approximation are initialized to one (line 2). Afterwards, the initial state S_i of the transmitter is observed (line 3) and a transmit power value $p_{1,i}^{\mathrm{Tx}} \in \mathcal{A}$ is randomly selected. Note that as no power value has been tried before, there is no prior experience that can be exploited. The selected transmit power is then used to transmit data to N$_2$ (line 6) and the corresponding reward is calculated using (3.3) (line 7). After the transmission, the new state S_{i+1} of the transmitter is observed (line 8). This state corresponds to the new battery level, the amount of harvested energy and the channel state. Next, the new transmit power value $p_{1,i+1}^{\mathrm{Tx}}$ is selected using the ϵ-greedy policy (line 9). The learning weights $\mathbf{w}$ are then updated using (3.24) and by considering the transition from S_i to S_{i+1}, the selected transmit power values $p_{1,i}^{\mathrm{Tx}}$ and $p_{1,i+1}^{\mathrm{Tx}}$ as wells as the obtained reward $R_{1,i}$ (line 10). Next, the current values of S_i and $p_{1,i}^{\mathrm{Tx}}$ are updated (lines 11-12) and the same procedure described above is repeated in every time interval as long as the transmitter is operational.

3.5.2.7. Convergence guarantees

Regarding the convergence properties of approximate SARSA, it has been shown in [Gor01] that if the policy π is not changed during the learning process and the learning rate ζ_i satisfies

$$\sum_{i=1}^{I} \zeta_i = \infty \tag{3.34}$$

and

$$\sum_{i=1}^{I} \zeta_i^2 < \infty, \tag{3.35}$$

then the SARSA algorithm combined with linear function approximation converges to a bounded region with probability one, i.e. it does not diverge. In our case, ζ_i is selected as $\zeta_i = 1/i$ which fulfills the two conditions in (3.34) and (3.35). Additionally, throughout the execution of the algorithm, the ϵ-greedy policy is followed.

3.5.2.8. Computational complexity analysis

In this section, we evaluate the computational complexity of one iteration of the proposed algorithm. As the number of feature functions used in the linear function approximation is fixed, we evaluate the complexity with respect to the size $A = |\mathcal{A}|$ of the action space, i.e., the number of possible transmit power values available. For this purpose, we analyze what are the most computationally demanding tasks in the proposed algorithm. Moreover, to describe the computational complexity we consider the O-notation, which is used to characterize the limiting behavior of a function by giving an asymptotic upper bound of its growth rate [CLRS09].

From Algorithm 3.1, it can be seen that the complexity of lines 1-3 does not grow with A, therefore the computational complexity of each of this lines is $O(1)$. The selection of the first transmit power, line 4, has a complexity that grows as $O(1)$ because only a random number, that corresponds to the index of the possible actions, needs to be generated. The calculation of the reward function in line 7 has also a complexity that grows as $O(1)$. This is because in each iteration, only the selected transmit power is considered to determine $R_{1,i}$. The complexity of the selection of the transmit power using the ϵ-greedy policy, in line 9, grows as $O(A)$ because when exploring, each element has to be evaluated once in order to find the maximum. Furthermore, the update of the weights in line 10 has a complexity that grows as $O(1)$ because only the selected transmit power is considered in the update. From this analysis, we can determine that the complexity of the proposed approximate SARSA algorithm grows only linearly with the size of the action space, i.e., $O(A)$.

Table 3.2. Simulation set-up.

	Parameter	Value	Description
General	I	1000	Number of time intervals
	T	1000	Number of realizations
	τ	10ms	Time interval duration
Energy	$B_{\max,1}$	$2E_{\max,1}$	Battery capacity of EH node N_1
	E_1^{Circ}	1mJ	Energy consumed by the circuit of EH node N_1
	ρ	10mW/cm^2	Power density of the EH source
	Ω	16cm^2	Size of EH panel
Data	d	1 kbit	Packet size (finite data buffer case)
	$D_{\max,1}$	∞	Data buffer size of EH node N_1 (infinitely full data buffer case)
	$D_{\max,1}$	50kbits	Data buffer size of EH node N_1 (finite data buffer case)
	λ	10	Average number of packets arriving per time interval (finite data buffer case)
Channel	f_0	2.4 GHz	Carrier frequency
	W	1 MHz	Bandwidth
	α	3	Path loss exponent
	Γ	5dB	Average SNR
Learning	γ	0.9	Discount factor
	δ	2%	Step size
	ϵ	$1/i$	Exploration probability
	ζ	$1/i$	Learning rate

3.6. Performance evaluation

In this section, numerical results for the evaluation of the offline approach and the proposed approximate SARSA algorithm in the EH point-to-point communication scenario are presented. A summary of all the variables considered for the simulations is presented in Table 3.2.

For the simulations, $T = 1000$ independent random energy, data and channel realizations are generated. It is assumed that each realization corresponds to an episode where N_1 harvests energy from the environment $I = 1000$ times. Moreover, as commonly done in the literature [LZL13a, TVY13, AINS13], it is assumed that the amount of harvested energy $E_{1,i}$ at time interval i is taken from a uniform distribution with maximum value $E_{\max,1}$ and it is stored in a finite battery with capacity $B_{\max,1} = 2E_{\max,1}$. We consider solar energy as our EH source with an average power density $\rho = 10\mathrm{mW/cm}^2$ and an EH panel size $\Omega = 16\mathrm{cm}^2$ [KLCL16]. Consequently, $E_{\max,1} = 2\rho\Omega\tau$. Furthermore, it is assumed that the energy consumed by the circuit when data is transmitted is $E^{\mathrm{Circ}} = 1\mathrm{mJ}$ [XZ14].

As in 5G systems, where the frame length is 10ms [3GP17], the time interval duration τ between two consecutive EH time instants is set to 10ms and the channel between N_1 and N_2 is assumed to be i.i.d. Rayleigh fading. Furthermore, a path loss exponent of three is considered and a bandwidth of $W = 1\text{MHz}$ is assumed to be available for the communication, which takes place over the unlicensed frequency $f_0 = 2.4$ GHz [ECC19]. 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

$$\Gamma = \frac{\rho \Omega \bar{g}_1}{\sigma^2} = 5\text{dB}, \tag{3.36}$$

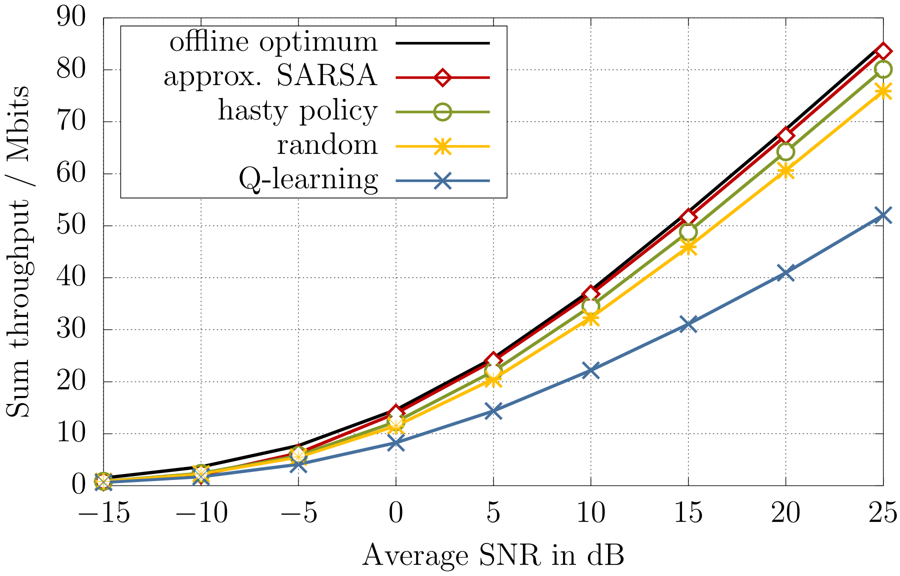
where $\bar{g}_1$ is the average channel gain on the link between the EH transmitter and the receiver. Moreover, to guarantee the feasibility of the offline approach, an infinitely full data buffer is assumed for this case, unless it is otherwise specified, i.e., $D_{\text{max},1} = \infty$, $D_{1,i} = \infty \; \forall i$.

The step size δ used in the definition of the action set $\mathcal{A}$ that contains the transmit power values is set to $\delta = 2\%$. Moreover, in order to guarantee the convergence of the approximate SARSA algorithm, the learning rate parameter is $\zeta_i = 1/i$. Additionally, the ϵ-greedy policy is used with $\epsilon = 1/i$, and in order to take into account the future rewards, a discount factor $\gamma = 0.9$ is selected [BGD13]. To compare the performance of the offline approach and our proposed approximate SARSA algorithm, three approaches are considered as reference schemes:

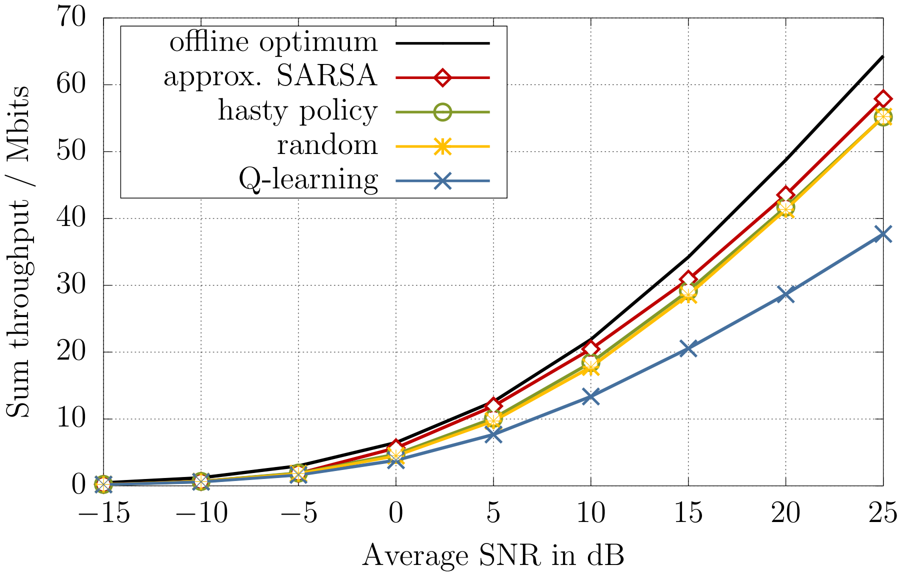
- Hasty Policy: In this approach, the battery is depleted in each time interval i. This means N_1 allocates all the power available in the battery regardless of the data buffer level or the state of the channel. As a result, battery overflow conditions are completely avoided.
- Random Policy: In this approach, the set $\mathcal{A}$ of transmit power values defined for the approximate SARSA is used. In each time interval, and based on the battery level, the subset of feasible transmit power values is determined such that (3.4b) is fulfilled. From this subset of $\mathcal{A}$, a transmit power value is randomly selected. It is assumed that all the transmit power values in the set have the same probability of being selected.
- Q-learning [BGD13]: This method is the off-policy temporal-difference RL approach used in [BGD13]. Note that Q-learning requires finite states. Therefore, in order to have a fair comparison, the results are obtained by the discretization of the energy, battery and channel values. For the simulations, the values are discretized using the step size δ.

Figure 3.2(a) shows the achieved sum throughput, i.e., the sum of the throughputs achieved over all time intervals, versus the average SNR. In this case, we assume that the energy consumed by the circuit is negligible, i.e., $E^{\text{Circ}} = 0$, and that the data buffer is infinitely full. As expected, the performance of all the approaches increases when the SNR increases. The upper bound of the achievable throughput is given by the optimum offline approach which assumes non-causal perfect knowledge regarding the EH and channel fading processes. The approximate SARSA algorithm is able to overcome this unrealistic requirement at the cost of only 2% of performance reduction when an SNR of 5dB is considered. For approximate SARSA, only causal knowledge is assumed at N_1. Similarly, the hasty policy and the random policy also assume only causal knowledge. However, since this information is not exploited for the power allocation, their performance is worse compared to our proposed approach. The throughput achieved by approximate SARSA is 9% and 17% higher than the throughput achieved by the hasty and random policy, respectively, for an average SNR of 5dB. Moreover, the lowest throughput is achieved by the Q-learning algorithm of [BGD13]. This behavior is explained by the fact that Q-learning requires a finite number of states and to fit it to our system model, discretization is required for the harvested energy, battery and channel gains. Additionally, as the number of states increases (depending on how fine or coarse the discretization is), the probability of visiting all the states decreases and the learning becomes slower.

The impact of the energy E^{Circ} consumed by the circuit is evaluated in Figure 3.2(b). As in the previous case, we show the achieved sum throughput versus the average SNR. In this case, to guarantee the feasibility of the offline optimization problem, we increase the power density of the EH source to $\rho = 50\text{mW/cm}^2$ while maintaining the same average SNR as in the previous case. As expected, the throughput achieved by all the approaches decreases compared to Figure 3.2(a) because part of the harvested energy is consumed by the circuitry at the transmitter. It can be seen that the proposed approximate SARSA algorithm maintains a performance close to the offline optimum. However, the gap between the offline optimum and the proposed approximate SARSA has increased compared to Figure 3.2(a), especially in the high SNR regime. For an average SNR of 5dB, the performance of the approximate SARSA is approximately 5% below the optimum and for an SNR of 20dB, it is 10 % below. This difference is caused by the fact that by incorporating the energy consumed by the circuit, the power allocation problem becomes more complex. This means, the transmitter should save the energy in the battery for the best channel conditions in order to achieve a throughput that compensates the energy consumption of the circuit. Nevertheless, the performance of the other reference approaches keeps the same trend as in Figure 3.2(a). For an SNR of 5dB, the approximate SARSA algorithm achieves a throughput that is 17%, 23% and 54% higher than for the hasty policy, the random approach and Q-learning, respectively.



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



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

Figure 3.2. Sum throughput versus average SNR.

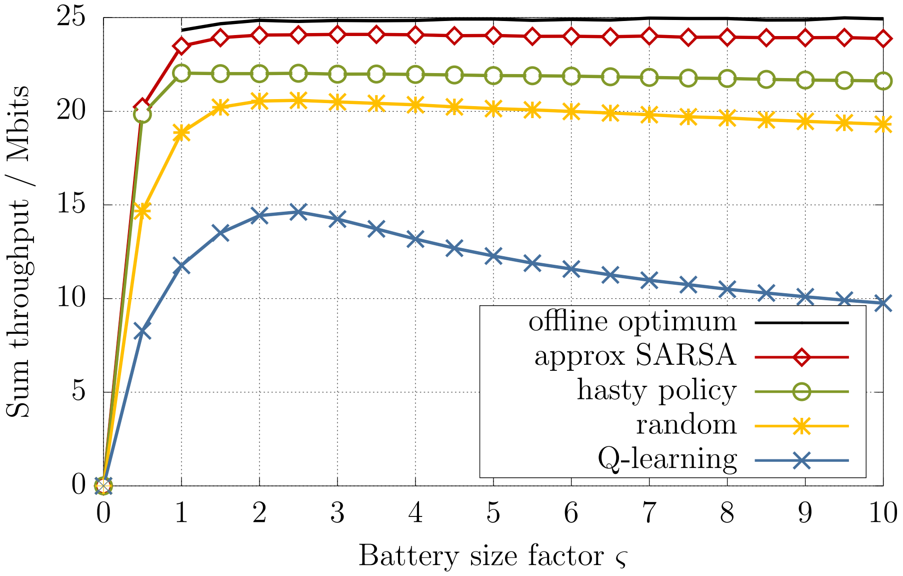


Figure 3.3. Sum throughput versus battery size factor ς for an average SNR of 5dB.

Figure 3.3 shows the effect of the battery size on the throughput achieved by the different approaches for an SNR of 5dB. For this simulation, the battery size is set to $B_{\max,1} = \varsigma E_{\max,1}$, where ς is a tunable parameter and $E^{\text{Circ}} = 0$. When $B_{\max,1} < E_{\max,1}$, the offline optimum cannot be calculated because overflow conditions are unavoidable. Thus, the optimization problem becomes infeasible. Consequently, the curve of the offline optimum starts only at $\varsigma = 1$. Results show that the approximate SARSA outperforms the other approaches for the complete range of battery sizes. When the battery is small, the performance of the approximate SARSA and the hasty policy is similar because all the harvested energy has to be spent in order to reduced the energy waste due to overflow. However, as the battery size increases, the transmitter conditions, i.e., channel gains and battery level, in each time interval have to be considered for the power allocation. As in the previous case, the lower throughput of the Q-learning algorithm is explained by the large number of states which reduces the learning speed compared to the approximate SARSA. An interesting result is that when the battery size is large compared to $E_{\max}$, its effect on the performance is reduced. It can be seen that the performance of the approximate SARSA saturates from approximately $\varsigma = 2$. The reason for this is that as $B_{\max}$ increases, the overflow conditions become less probable. Nevertheless, note that for larger values of the battery size, all the reference schemes tend to decrease their performance and a slight degradation can also be observed for approximate SARSA when $\varsigma > 9$. This is because the set of possible transmit power values depends on the battery size, but the number of

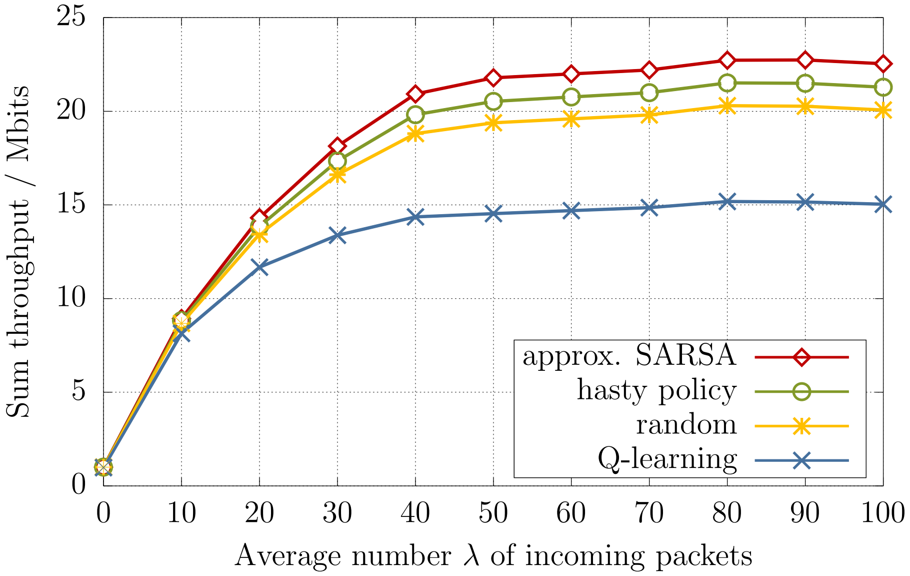


Figure 3.4. Sum throughput versus the average number λ of incoming data packets.

transmit power values that can be selected is kept constant by the use of the same step size $\delta = 2\%$. This means, for larger battery sizes, the transmit power values vary, e.g., the maximum transmit power increases, but always the same amount of power values are considered. Therefore, as the nominal difference between two consecutive transmit power values increases with larger battery sizes, the possibility to adapt to the optimal transmit power value that should be used, is reduced.

The impact of the data arrival process on the sum throughput is evaluated in Figure 3.4. Considering the throughput achieved in the previous figures, where an infinitely full data buffer was adopted, we assume a finite data buffer with size $D_{\max,1} = 50$kbits. Additionally, we consider a data arrival process consisting of the arrival of a certain number of data packets of size d in each time interval. The packets arrive following a Poisson distribution with parameter λ. Specifically, we evaluate the sum throughput for different values of the average number λ of incoming data packets. A packet size $d = 1$kbit is assumed and we set $E_1^{\text{Circ}} = 0$. Note that for this simulation we have not considered the offline optimum because the feasibility of the power allocation problem cannot be guaranteed. This is due to the fact that data buffer overflow situations might not be avoided if the harvested energy is not enough to transmit all the incoming data. An effect that is more noticeable when λ is large. The results show that when λ is small, the sum throughput of all the approaches is constrained by the availability of data. As λ increases, the sum throughput also increases until it saturates around $\lambda = 50$ packets, which perfectly matches the data buffer size. This means that for $\lambda > 50$, data buffer

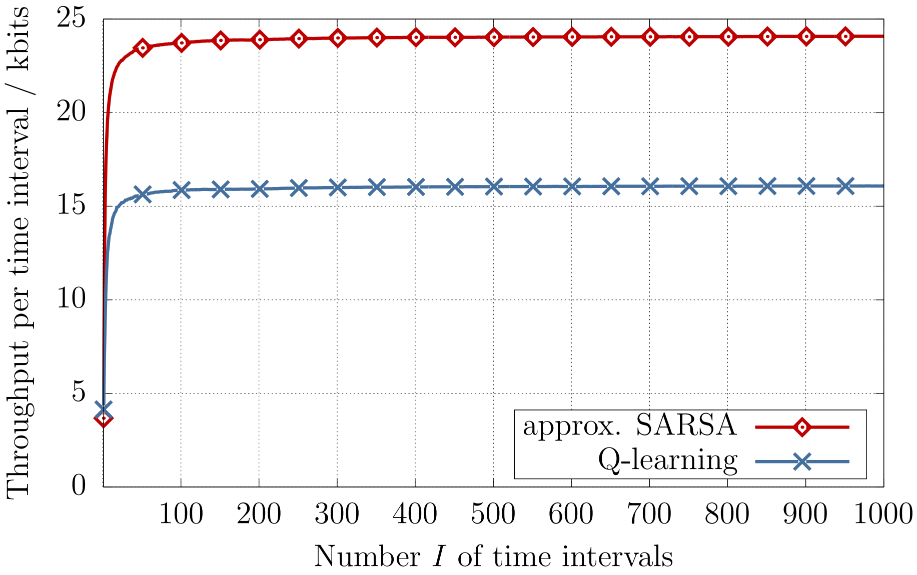


Figure 3.5. Throughput per time interval versus the number of time intervals.

overflow situations are unavoidable. As in the previous cases, approximate SARSA outperforms the reference schemes since it adapts the transmission to efficiently use the harvested energy. In particular, for $\lambda = 50$ packets, it achieves a sum throughput that is 6%, 12%, and 50% larger than the throughput achieved by the hasty policy, the random approach and Q-learning, respectively.

The convergence speed of the two learning algorithms, i.e., approximate SARSA and Q-learning, is presented in Figure 3.5. In the figure, the average throughput per time interval versus the number I of time intervals is depicted. For a fair comparison, we have used the same exploration probability ϵ for approximate SARSA and Q-learning, which is decreased in each time interval as $\epsilon = 1/i$. It is shown that both algorithms converge approximately at the same time, however approximate SARSA is able to identify the transmission policy that leads to a higher throughput. In contrast to approximate SARSA, in which the representation of Q^{π} is done via linear function approximation, Q-learning discretizes the state space and uses a tabular representation of the action value function Q^{π}. This tabular representation has an entry for every combination of the possible battery levels, amounts of harvested energy and channel gains. Consequently, Q-learning requires a much larger exploration phase in order to estimate all the possible values of Q^{π}. Although it is barely noticeable, the throughput of Q-learning slowly increases when a larger number I of time intervals is considered. This means that Q-learning requires a careful parametrization in order to balance the exploration and exploitation trade-off in such a large state space. By exploiting the properties of the problem in the definition of the feature functions, the proposed approximate SARSA algorithm has a more efficient representation of the state-action space and it is able to generalize in similar situations, e.g., within fewer trials it is able to identify that overflow situations should be avoided, thus learning to select the correct power values quicker.

3.7. Conclusions

In this chapter, we have investigated offline and learning approaches for the power allocation problem for throughput maximization in the EH point-to-point communication scenario.

Assuming perfect non-causal knowledge of the system dynamics, an offline approach is presented in order to find the upper bound of the performance. By analyzing the KKT conditions, it is observed that although the resulting optimization problem is a convex optimization problem, a closed-form solution of the power to be used in each time interval cannot be obtained. This is because, in the optimal policy, the power to be allocated in time interval i depends on the Lagrange multipliers associated to the energy causality constraints and battery overflow constraints in the future time intervals. Nevertheless, by extending results from the literature, a characterization of the optimal policy is provided. In particular, it is shown that in the optimal policy, the allocated power should be constant for the duration of one time interval and that when an infinite battery is considered, the transmit power increases monotonically over time, i.e., it never decreases.

Based on the analysis performed in the offline approach, a learning approach is proposed to find the power allocation policy that aims at maximizing the throughput when only casual knowledge of the system dynamics is available. The proposed approach, termed approximate SARSA, is based on the RL algorithm SARSA. We have combined it with linear function approximation to handle the fact that the amounts of harvested energy, the battery levels, the data buffer levels, and the channel gains can be taken from a continuous range. To perform linear function approximation, four feature functions are proposed which exploit the characteristics of the EH point-to-point communication scenario. Furthermore, we show that by the appropriate selection of the learning rate parameter, the convergence of the algorithm to a bounded region can be guaranteed. Moreover, by means of a computational complexity analysis, we show that the complexity of the proposed approximate SARSA increases only linearly with the number of transmit power values that can be selected by the transmitter. By numerical simulations, we have shown that the proposed approximate SARSA significantly outperforms the reference schemes found in the literature.

Chapter 4
Energy harvesting two-hop scenario

4.1. Introduction

In this chapter, an EH two-hop communication scenario, in which an EH transmitter communicates with a receiver via an EH relay, is investigated. In contrast to the EH point-to-point scenario, which consists of only one EH transmitter, two EH nodes have to be considered in the two-hop scenario, i.e., the EH transmitter and the EH relay. This means, the EH, data arrival and channel fading processes associated to both of them should be taken into account in the power allocation problem. For this scenario, two relay types are considered, namely, a decode-and-forward and an amplify-and-forward relay. Furthermore, in order to find the power allocation policies that aim at maximizing the throughput in this setting, offline and learning approaches are studied.

The chapter is organized as follows. In Section 4.2, an EH two-hop scenario with a decode-and-forward relay is considered. For this case, we first present the corresponding system assumptions and formulate the power allocation problem for throughput maximization. Afterwards, we show that the offline approach leads to the solution of a convex optimization problem. Next, we propose two learning approaches, termed independent and cooperative SARSA, which are motivated by the fact that the transmitter and the relay have only causal knowledge about their own parameters. For each of these approaches, we discuss their convergence guarantees and analyze their computational complexity. Finally, through several numerical simulations we evaluate the performance of the proposed approaches. In the following Section 4.3, an amplify-and-forward relay is considered. Similar to the previous case, we first describe the scenario and the corresponding system assumptions. Next, we formulate the throughput maximization problem and show that the resulting problem is non-convex. To overcome this challenge, we propose an offline approach based on the reformulation of the original problem as the difference of two concave functions. Afterwards, we show how a centralized learning approach can be used to find the power allocation policy that aims at maximizing the throughput when only causal knowledge is available. Then, we evaluate the performance of the proposed schemes through several numerical simulations. Last, in Section 4.4, we discuss how the proposed learning approaches can be extended to EH multi-hop relaying scenarios.

Parts of this Chapter 4 have been published by the author in [OASL$^+$15], [OASL$^+$16a] and [OASL$^+$17].

Table 4.1. Parameters associated to the EH two-hop communication scenario with a decode-and-forward relay.

	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}
	$h_{n,i}$	Channel coefficient of the link between N_n and N_{n+1}
	W	Bandwidth
	σ_n^2	Noise power at N_n

4.2. Decode-and-forward relay

4.2.1. Scenario description and assumptions

In this section, we describe the EH two-hop communication scenario when a decode-and-forward EH relay is considered. A summary of all the considered parameters is given in Table 4.1. Specifically, the scenario consists of three single-antenna nodes N_1, N_2, and N_3, as depicted in Figure 4.1, 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 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. The energy $E_{1,i}^{\mathrm{Circ}}$ consumed by the circuit at N_1 is

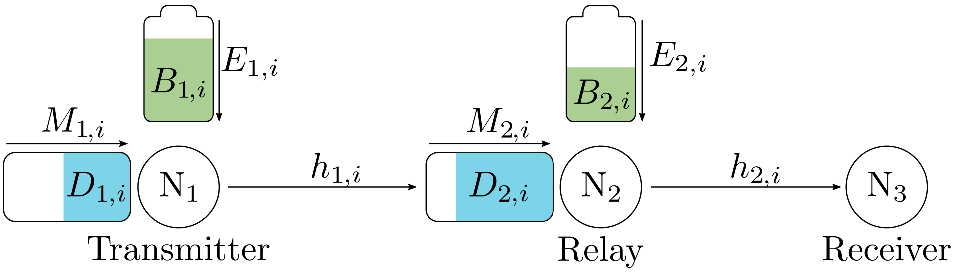


Figure 4.1. Two-hop communication scenario with an EH transmitter and an EH decode-and-forward relay.

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, $E_{2,i}^{\mathrm{Circ}} = E_2^{\mathrm{Circ}}$, $\forall i$.

As described in Section 2.2.3, two cases are distinguished for the data arrival model at N_1. In the first case, an infinitely full data buffer is considered. This means that the data buffer size $D_{\mathrm{max},1}$ is infinite and the data buffer level $D_{1,i}$ is also infinite for all the time intervals. In the second case, 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 $D_{\mathrm{max},1}$. 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

$$\Delta = \begin{cases} 1, & \text{if } \mathrm{N}_2 \text{ operates in full-duplex mode} \\ 1/2, & \text{if } \mathrm{N}_2 \text{ operates in half-duplex mode.} \end{cases} \tag{4.1}$$

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. 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

$$R_{1,i}^{\mathrm{DF}} = \Delta W \tau \log_2 \left(1 + \frac{g_{1,i} p_{1,i}^{\mathrm{Tx}}}{\sigma_2^2} \right), \tag{4.2}$$

where W denotes the available bandwidth, $g_{1,i} = |h_{1,i}|^2$ 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. Considering (2.3), the battery level at N_1 is updated at the beginning of each time interval as

$$B_{1,i+1} = \min\left\{B_{\max,1},\ B_{1,i} - \Delta\tau p_{1,i}^{\mathrm{Tx}} + E_{1,i} - E_1^{\mathrm{Circ}}\right\}. \tag{4.3}$$

Similarly, considering (2.6), the data buffer level at N_1 is updated at the beginning of each time interval as

$$D_{1,i+1} = \min\left\{D_{\max,1},\ D_{1,i} - R_{1,i}^{\mathrm{DF}} + M_{1,i}\right\}. \tag{4.4}$$

Regardless of the data arrival model considered at N_1, 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, $R_{2,i}^{\mathrm{DF}}$ is approximated using Shannon's capacity formula as

$$R_{2,i}^{\mathrm{DF}} = \Delta W\tau \log_2\left(1 + \frac{g_{2,i}p_{2,i}^{\mathrm{Tx}}}{\sigma_3^2}\right), \tag{4.5}$$

where $g_{2,i} = |h_{2,i}|^2$ 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 (4.3) and (4.4), 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.

Transmitter side channel state information is assumed to be available at N_1 and N_2, i.e., each EH node has knowledge about the channel gains associated to its own links. For the offline approach, it is assumed that this channel state information is non-causally known. Moreover, in this case it is assumed that both EH nodes know also the channel gains associated to the other node. On the contrary, for the learning approaches, it is realistically assumed that the 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.

4.2.2. 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 Section 2.2, and the scenario description of Section 4.2.1, the power allocation problem is written as

$$\left(p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{n,i} = \underset{\{p_{n,i}^{\mathrm{Tx}}, n=\{1,2\}, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_{2,i}^{\mathrm{DF}} \tag{4.6a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \Delta\tau p_{n,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{n,i}, \;\; n=1,2, \;\; J=1,\ldots,I, \tag{4.6b}$$

$$\sum_{i=1}^{J} E_{n,i} - \sum_{i=1}^{J} \Delta\tau p_{n,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq B_{\max,n}, \; n=1,2, \;\; J=1,\ldots,I, \tag{4.6c}$$

$$\sum_{i=1}^{J} R_{n,i}^{\mathrm{DF}} \leq \sum_{i=1}^{J-1} M_{n,i}, n=1,2, \;\; J=1,\ldots,I, \tag{4.6d}$$

$$\sum_{i=1}^{J} M_{n,i} - \sum_{i=1}^{J} R_{n,i}^{\mathrm{DF}} \leq D_{\max,n}, \; n=1,2, \;\; J=1,\ldots,I, \tag{4.6e}$$

$$p_{n,i}^{\mathrm{Tx}} \geq 0, \quad n=1,2, \, i=1,\ldots,I, \tag{4.6f}$$

where $R_{1,i}^{\mathrm{DF}}$ and $R_{2,i}^{\mathrm{DF}}$ are defined in (4.2) and (4.5), respectively, (4.6b) is the energy causality constraint, (4.6c) is the battery overflow constraint, (4.6d) is the data causality constraint and (4.6e) is the data buffer overflow constraint for N_1 and N_2, respectively. Note that when an infinitely full data buffer is considered at N_1, the respective constraints in (4.6d) and (4.6e) for $n=1$ are not taken into account. By examining the problem in (4.6), it can be seen that 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 in order to avoid data buffer overflow situations. As a result, the dynamics of each EH node affect the power allocation policy of the other. This means that, if the aim is to find the optimal solution, the power allocation problem for throughput maximization in the EH two-hop scenario cannot be treated as two parallel point-to-point scenarios due to this interdependency.

4.2.3. Offline approach

In this section, an offline approach is presented in order to find the optimal power allocation policy at N_1 and N_2 when perfect non-causal knowledge of the system dynamics is available at both nodes. This approach is based on the work in [OE15] in which a centralized optimization problem for the EH two-hop scenario with a half-duplex decode-and-forward relay, is solved. Here we have extended it to consider the case of a full-duplex relay and to include the energy consumed by the circuit.

The objective function in (4.6a) is a concave function of $p_{2,i}^{\mathrm{Tx}}$. Furthermore, (4.6b), (4.6c) and (4.6f) are linear functions of $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$, the constraint in (4.6d) is a concave function of $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$, and (4.6e) is a convex function of $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$. As a consequence, (4.6) is a convex optimization problem and its corresponding Lagrangian function can be written as

$$\begin{aligned}
\mathfrak{L} = &\sum_{i=1}^{I} R_{2,i}^{\mathrm{DF}} \\
&- \sum_{n=1}^{2}\sum_{i=1}^{I} \mu_{n,i}\left(\sum_{j=1}^{i}\Delta\tau p_{n,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{n,j}\right) \\
&- \sum_{n=1}^{2}\sum_{i=1}^{I} \omega_{n,i}\left(\sum_{j=1}^{i} E_{n,j} - \sum_{j=1}^{i}\Delta\tau p_{n,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - B_{\max,n}\right) \\
&- \sum_{n=1}^{2}\sum_{i=1}^{I} \kappa_{n,i}\left(\sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - \sum_{j=1}^{i-1} M_{n,j}\right) \\
&- \sum_{n=1}^{2}\sum_{i=1}^{I} \xi_{n,i}\left(\sum_{j=1}^{i} M_{n,j} - \sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - D_{\max,n}\right) \\
&+ \sum_{n=1}^{2}\sum_{i=1}^{I} \upsilon_{n,i} p_{n,i}^{\mathrm{Tx}},
\end{aligned} \tag{4.7}$$

where $\mu_{n,i}$, $\omega_{n,i}$, $\kappa_{n,i}$, $\xi_{n,i}$ and υ_i are the Lagrange multipliers associated with the energy causality constraint, battery overflow constraint, data causality constraint, data buffer overflow constraint and the power value, respectively.

The corresponding KKT conditions, which are necessary conditions for a global optimum, are given by

$$\frac{\partial \mathfrak{L}}{\partial p_{1,i}^{\mathrm{Tx}}} = \frac{\Delta\tau W g_{1,i}}{(\ln 2)\left(\sigma^2 + g_{1,i} p_{1,i}^{\mathrm{Tx}}\right)} \left(\sum_{j=i+1}^{I} \kappa_{2,j} - \sum_{j=i}^{I} (\kappa_{1,j} - \xi_{1,j}) \right) \\ -\Delta\tau \sum_{j=i}^{I} \left(\mu_{1,j} - \omega_{1,j}\right) + \upsilon_{1,i} = 0, \tag{4.8}$$

$$\frac{\partial \mathfrak{L}}{\partial p_{2,i}^{\mathrm{Tx}}} = \frac{\Delta\tau W g_{2,i}}{(\ln 2)\left(\sigma^2 + g_{2,i} p_{2,i}^{\mathrm{Tx}}\right)} \left(1 - \sum_{j=i}^{I} (\kappa_{2,j} - \xi_{2,j}) \right) \\ -\Delta\tau \sum_{j=i}^{I} \left(\mu_{2,j} - \omega_{2,j}\right) + \upsilon_{2,i} = 0, \tag{4.9}$$

$$\mu_{n,i} \left(\sum_{j=1}^{i} \Delta\tau p_{n,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{n,j} \right) = 0, \tag{4.10}$$

$$\omega_{n,i} \left(\sum_{j=1}^{i} E_{n,j} - \sum_{j=1}^{i} \Delta\tau p_{n,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - B_{\max,n} \right) = 0, \tag{4.11}$$

$$\kappa_{n,i} \left(\sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - \sum_{j=1}^{i-1} M_{n,j} \right) = 0, \tag{4.12}$$

$$\xi_{n,i} \left(\sum_{j=1}^{i} M_{n,j} - \sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - D_{\max,n} \right) = 0, \tag{4.13}$$

$$\upsilon_{n,i} p_{n,i}^{\mathrm{Tx}} = 0, \tag{4.14}$$

$$\sum_{j=1}^{i} \Delta\tau p_{n,j}^{\mathrm{Tx}} + \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - \sum_{j=1}^{i-1} E_{n,j} \leq 0, \tag{4.15}$$

$$\sum_{j=1}^{i} E_{n,j} - \sum_{j=1}^{i} \Delta\tau p_{n,j}^{\mathrm{Tx}} - \sum_{j=1}^{i} E_n^{\mathrm{Circ}} - B_{\max,n} \leq 0, \tag{4.16}$$

$$\sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - \sum_{j=1}^{i-1} M_{n,j} \leq 0, \tag{4.17}$$

$$\sum_{j=1}^{i} M_{n,j} - \sum_{j=1}^{i} R_{n,j}^{\mathrm{DF}} - D_{\max,n} \leq 0, \tag{4.18}$$

$$-p_{n,i}^{\mathrm{Tx}} \leq 0, \tag{4.19}$$

$$\mu_{n,i} \geq 0,\ \omega_{n,i} \geq 0,\ \kappa_{n,i} \geq 0,\ \xi_{n,i} \geq 0,\ \upsilon_{n,i} \geq 0. \tag{4.20}$$

From the slackness condition in (4.14), it is clear that when $p_{n,1}^{\mathrm{Tx}} > 0$, $\upsilon_{n,i}$ must be equal to zero. Consequently, from (4.8), (4.9) and (4.14), the optimal power allocations in time interval i for N_1 and N_2 can be expressed as

$$p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}} = \nu_{n,i} - \frac{\sigma^2}{g_{n,i}}, \ n = \{1,2\}, \ p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}} > \upsilon_{n,i}^{\mathrm{opt}}. \tag{4.21}$$

Considering the well-known water-filling algorithm for power allocation over multiple channels [TV05], $\nu_{n,i}$ is interpreted as the water level in time interval i which is calculated as

$$\nu_{1,i} = \frac{W\left(\sum\limits_{j=i+1}^{I} \kappa_{2,j} - \sum\limits_{j=i}^{I} (\kappa_{1,j} - \xi_{1,j})\right)}{\sum\limits_{j=i}^{I} (\mu_{1,j} - \omega_{1,j})}, \tag{4.22}$$

and

$$\nu_{2,i} = \frac{W\left(1 - \sum\limits_{j=i}^{I} (\kappa_{2,j} - \xi_{2,j})\right)}{\sum\limits_{j=i}^{I} (\mu_{2,j} - \omega_{2,j})}, \tag{4.23}$$

for N_1 and N_2, respectively.

Note that in the EH two-hop scenario, the water levels of N_1 and N_2 do not necessarily increase monotonically with time when the batteries have infinite capacity. This means, the power values might increase or decrease over time. This is because in addition to the EH processes, the water levels $\nu_{n,i}$ vary according to the data arrival process at N_1 and the data buffer levels of both nodes. From (4.22) it is clear that, in contrast to the EH point-to-point scenario, even when the data buffers have infinite capacity, i.e., the Lagrange multiplier $\xi_{n,i} = 0$ in (4.13), the water levels $\nu_{n,i}$ might increase or decrease depending on the value of $\kappa_{n,i}$ which is the Lagrange multiplier associated with the data causality constraint in (4.12). Assuming batteries and data buffers with infinite capacities, the relationship between the data buffers of N_1 and N_2 can be summarized in the following proposition.

Proposition 4.1. *In the optimal policy for the case when the batteries and data buffers of the transmitter* N_1 *and the relay* N_2 *have infinite capacity, if the data buffer of* N_1 *is depleted at time interval i, i.e., all the data is transmitted to the relay, then the data buffer at* N_2 *has to be depleted a least once in the following intervals* j, $i+1 \leq j \leq I$.

Proof. If the battery and the data buffers have infinite capacity, then from (4.11) and (4.13), it follows that the Lagrange multipliers $\omega_{n,i}$ and $\xi_{n,i}$ must be equal to zero for all the time intervals. Furthermore, in order to deplete the data buffer of N_1 in time interval i, the transmit power $p_{n,i}^{\mathrm{Tx}}$ must be strictly greater than zero. Therefore, for (4.22) to hold, the data buffer at N_2 should be depleted at least once in the following intervals, i.e., $\kappa_{2,j} > 0$, for at least one time interval j, with $i+1 \leq j \leq I$. □

Although a closed-form solution, which depends only on the system parameters $E_{n.i}$, $B_{n.i}$, $g_{n.i}$ $D_{n.i}$, $B_{\max,n}$ and $D_{\max,n}$, cannot be obtained for the power $p_{n,i}^{\mathrm{Tx}}$ to use in each time interval, standard convex optimization algorithms can be used to find a numerical solution for the power allocation that maximizes the throughput in the EH two-hop communication scenario.

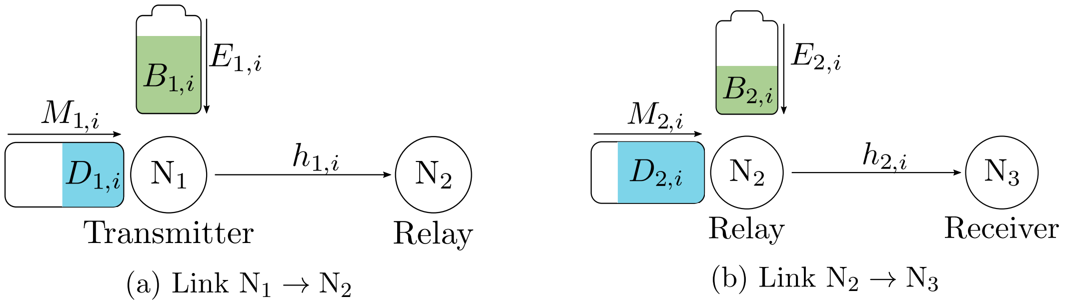
4.2.4. Learning approach: Independent SARSA

From the analysis presented in the offline approach in Section 4.2.3, it is clear that the power allocation policies of the transmitter N_1 and the relay N_2 depend on the dynamics of the system, i.e., the EH, data arrival and channel fading processes associated to both nodes. This means that in order to find the optimum power allocation policy at each node, these processes have to be jointly taken into account. However, N_1 and N_2 have only knowledge about their own processes. Moreover, in learning approaches this knowledge is only causally obtained and can be outdated, e.g., when only outdated channel state information is available. Therefore, in the EH two-hop communication scenario, two types of states are considered:

- State of a node: It results from the causal knowledge of the EH, data arrival and channel fading processes associated to one EH node N_n. It consists of the values of the parameters $E_{n,i}$, $B_{n,i}$, $D_{n,i}$, and $g_{n,i}$ associated to N_n in time interval i.
- State of the system: It results from the causal knowledge of the EH, data arrival and channel fading processes associated to both EH nodes N_1 and N_2. It is composed by the state of both nodes.

Considering that N_1 and N_2 are able to observe their own state but can only partially observe the system state, two learning approaches are proposed, namely, independent and cooperative SARSA. In this section, we present independent SARSA, a learning algorithm in which N_1 and N_2 make independent decisions regarding the transmit powers to use in each time interval without considering the state of the other node. The cooperative SARSA algorithm, which exploits cooperation between the two nodes in order to observe the system state, is described in Section 4.2.5.

The proposed independent SARSA is motivated by the fact that, in addition to the challenge posed by the partial observability of the system state, the nodes might not be able to observe the decision made by the other node before making their own, e.g., if a



(a) Link $N_1 \rightarrow N_2$

(b) Link $N_2 \rightarrow N_3$

Figure 4.2. Reformulation of the two-hop EH communication problem as two point-to-point communication problems.

full-duplex relay is considered. Consequently, as the EH nodes neither have information about the power allocation policy nor the state of the other node, the power allocation problem cannot be jointly solved. We propose to solve independent power allocation problems at N_1 and N_2 which aim at maximizing the throughput of each link. The idea behind the approach is to separate the EH two-hop communication scenario into two EH point-to-point communication scenarios, as depicted in Figure 4.2. The first scenario corresponds to the link $N_1 \rightarrow N_2$ between N_1 and N_2 and it is shown in Figure 4.2(a). The second one corresponds to the link $N_2 \rightarrow N_3$ between N_2 and N_3 and it is illustrated in Figure 4.2(b). In the following, we focus on a full-duplex decode-and-forward relay because, as mentioned before, it brings the additional challenge that the nodes cannot observe the decisions made by the other before making their own. However, note that the same procedure can be applied to a half-duplex relay. The only difference is that the nodes will not make simultaneous decisions.

The independent SARSA algorithm falls into the category of multi-agent reinforcement learning because two learning agents, the transmitter and the relay, are considered. However, note that the nodes act independently to maximize their own throughput and do not explicitly consider the other node in their learning model. As a result, for both, N_1 and N_2, the problem reduces to a point-to-point communication problem and the proposed approximated SARSA algorithm, described in Section 3.5, is applied to each of them. Moreover, the corresponding convergence guarantees and computational complexity analysis apply. In summary, the independent SARSA approach is a distributed multi-agent learning algorithm in which the goal of each node is to select the transmit power $p_{n,i}^{\mathrm{Tx}}$ aiming at maximizing its own throughput, regardless of the decision of the other node, i.e., no cooperation between the nodes is considered. In the independent SARSA approach, N_1 maximizes the amount of data transmitted to N_2, and it is the task of N_2 to maximize the amount of data that will finally reach N_3.

4.2.5. Learning approach: Cooperative SARSA

4.2.5.1. Cooperation in multi-agent RL

As mentioned before, both N_1 and N_2 have only causal, and possibly outdated, knowledge regarding their own EH, data arrival and channel fading processes. However, knowledge about the dynamics of the system is required at both nodes in order to achieve optimum performance. To this aim, in this section we propose a cooperative learning approach, termed cooperative SARSA, to find power allocation policies at the transmitter and at the relay that aim at maximizing the amount of data transmitted to the receiver. The scenario is depicted in Figure 4.1 and as in the previous section, we focus on a full-duplex decode-and-forward relay because it has the additional challenge that the nodes cannot observe the decisions made by the other node before making their own. However, note that the approach can be applied to a half-duplex decode-and-forward relay. In contrast to the independent SARSA approach of Section 4.2.4, here we propose mechanisms to overcome the limitation that the transmitter and the relay are only able to partially observe the system 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 with each other by exchanging information about the value of their current parameters.

The proposed cooperative SARSA is a distributed solution in which the nodes cooperate with each other during the signaling phase. Based on their knowledge of their own state and the knowledge they have obtained about the state of the other node during the signaling phase, N_1 and N_2 find their own transmission policies. However, since both, N_1 and N_2, are deciding on their own transmit power values, the problem can no longer be modeled as an MDP. This is because MDPs consider only one decision-making agent. Therefore, in order to take into account both nodes, we first model this scenario as a Markov game in Section 4.2.5.2. In Section 4.2.5.3, the proposed update rule for the estimation of the action-value function in the cooperative SARSA algorithm is presented. Afterwards, in Sections 4.2.5.4 and 4.2.5.5, we describe the use of linear function approximation and present the proposed mechanisms to overcome the partial observability of the system state, respectively. The proposed feature functions used in the linear function approximation are defined in Section 4.2.5.6, the action selection policy is explained in Section 4.2.5.7 and a summary of the proposed SARSA algorithm is presented in Section 4.2.5.8. Additionally, to validate our proposed algorithm, we derive convergence guarantees based on RL in Section 4.2.5.9. These guarantees are obtained by assuming that the EH nodes are able to observe the system state, i.e., when the channel prediction and the transmission of the signaling are successful, and a constant learning rate is used. Finally, in Section 4.2.5.10, we present a computational complexity analysis of the proposed approach.

4.2.5.2. 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 contrast to the independent SARSA approach, where N_1 and N_2 only consider the value of their own parameters, 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 MDPs to the case when multiple agents, which make decisions based on observations of a common environment, are considered [Lit94].

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 [Lit01]. In our case, the players are the transmitter and the relay. 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 [Ins17], 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 (4.5).

Similar to MDPs, in the Markov game formulation we need to find the transmission policies for N_1 and N_2 which correspond to the transmit powers to be used for data transmission in each time interval. Each transmission policy π_n, $n \in \{1, 2\}$, 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 action-value function $\mathrm{Q}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$. Nevertheless, 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

$$R^{\mathrm{DF}} = \lim_{I \to \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^{i-1} R_{2,i}^{\mathrm{DF}}\right]. \tag{4.24}$$

4.2.5.3. Action-value function update

The proposed cooperative learning algorithm is based on the RL algorithm SARSA [SB18]. 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

$$\mathrm{Q}_{i+1}^{\Pi}(S_i, P_i^{\mathrm{Tx}}) = \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})(1 - \zeta_i) + \zeta_i \left[R_i^{\mathrm{DF}} + \gamma \mathrm{Q}_i^{\Pi}(S_{i+1}, P_{i+1}^{\mathrm{Tx}})\right], \tag{4.25}$$

where ζ_i is a small positive fraction which influences the learning rate [SB18].

In our scenario, the nodes have a common objective, which is to maximize the expected throughput given in (4.24), 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 Section 4.2.5.9. Furthermore, the proposed updating rule for $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ is given by

$$\begin{aligned}\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max\big\{&\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}),\\ &(1-\zeta_i)\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) + \zeta_i\left[R_i^{\mathrm{DF}} + \gamma \mathrm{q}_{n,i}^{\pi_n}(S_{i+1}, p_{n,i+1}^{\mathrm{Tx}})\right]\big\}.\end{aligned} \quad (4.26)$$

4.2.5.4. Linear function approximation

The update of the action-value function, presented in Section 4.2.5.3, 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 $\mathbf{w}_n \in \mathbb{R}^{F\times 1}$. As described in Section 3.5.2.3, the action-value function is approximated as

$$\hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}, \mathbf{w}_n) = \mathbf{f}_n^{\mathrm{T}}\mathbf{w}_n \approx \mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}). \quad (4.27)$$

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 [SB18]. Considering the update for $\mathrm{q}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ given in (4.26), we propose to update $\mathbf{w}_n$ as

$$\mathbf{w}_{n,i+1} = \mathbf{w}_{n,i} + \max\left\{0,\, \zeta_i\left[R_i^{\mathrm{DF}} + \gamma\, \mathbf{f}_n^{\mathrm{T}}\mathbf{w}_{n,i} - \mathbf{f}_n^{\mathrm{T}}\mathbf{w}_{n,i}\right]\mathbf{f}_n\right\}. \quad (4.28)$$

4.2.5.5. 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 $h_{n,i}$ 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

$$y_{n+1,i} = x_{n,i} h_{n,i} + w_{n+1,i}, \tag{4.29}$$

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 the transmitter 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 the transmitter 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.

As described in Section 2.2.4, the magnitude $|h_{n,i}|$ of the channel coefficient $h_{n,i}$ is assumed to follow a Rayleigh distribution and the Jakes' model [Jak74] is used to model the autocorrelation function ACF of the channel coefficients [SW15, CZ04] as

$$\mathrm{ACF} = \mathrm{J}_0(2\pi f_{D,\max}\tau), \tag{4.30}$$

where J_0 is the zero$^{\mathrm{th}}$ order Bessel function of the first kind and $f_{D,\max}$ is the maximum Doppler frequency. As extensively reported in literature [SW15, CZ04, MS05], for the channel prediction at each N_n, the dynamics of the channel coefficient are modeled as an autoregressive process whose order and parameters are denoted by o and $c_{n,1}, ..., c_{n,o}, \psi_n$, respectively. Specifically, $h_{n,i}$ is modeled as

$$h_{n,i} = -\sum_{j=1}^{o} c_{n,j} h_{n,i-j} + \psi_n z_{n,i}, \tag{4.31}$$

where $z_{n,i}$ is AWGN. 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 (4.30). From (4.29) and (4.31), 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

$$\mathbf{h}_{n,i} = \mathbf{C}_n \mathbf{h}_{n,i-1} + \mathbf{a}_n \mathbf{v}_{n,i}, \tag{4.32}$$

$$y_{n+1,i} = \mathbf{x}_{n,i} \mathbf{h}_{n,i} + w_{n+1,i} \tag{4.33}$$

Algorithm 4.1 Kalman filter based channel predictor

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

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

$$\mathbf{C}_n = \begin{pmatrix} -c_{n,1} & -c_{n,2} & \cdots & -c_{n,o} \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{4.34}$$

Considering (4.33), each N_n can estimate its own channel coefficient in time interval i using the Kalman filter described in Algorithm 4.1 which 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 4.1, $\mathbf{I}_o$ 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 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, $u_{n,i}$ depends on the tolerable quantization error $e_{\mathrm{quant},u_{n,i}}$, the maximum value $V_{\max,u_{n,i}}$ and the minimum value $V_{\min,u_{n,i}}$ each $u_{n,i}$ can take. The number $Z_{u_{n,i}}$ of bits is calculated as

$$Z_{u_{n,i}} = \left\lceil \log_2 \left(\frac{V_{\max,u_{n,i}} - V_{\min,u_{n,i}}}{e_{\mathrm{quant},u_{n,i}}} \right) - 1 \right\rceil , \tag{4.35}$$

where $\lceil \cdot \rceil$ is the rounding operation to the next integer value greater than or equal to the evaluated number. Since $V_{\max,u_{n,i}}$ and $V_{\min,u_{n,i}}$ 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

$$Z_n = \sum_{\forall u_{n,i}} Z_{u_{n,i}} . \tag{4.36}$$

Given Z_n, the power $p_{n,i}^{\mathrm{Sig}}$ required to transmit the signaling from N_n to N_m is

$$p_{n,i}^{\mathrm{Sig}} = \frac{\sigma^2}{g_{n,i}} \left(2^{\frac{Z_n}{W\tau^{\mathrm{Sig}}}} - 1 \right) . \tag{4.37}$$

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. The order in which this reduction is done is given by the impact each parameter has on the feature functions described in Section 4.2.5.6. 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 $i - 1$.

4.2.5.6. Feature functions

The feature functions used for the linear function approximation are defined based on the EH processes at the EH nodes, the finite size of the batteries, the data arrival processes, the finite size of the data buffers and the channel fading processes. For the proposed cooperative SARSA, we consider $F = 6$ binary feature functions. The first four feature functions were defined in (3.26), (3.30), (3.31) and (3.32). Here, we propose two new feature functions to take into account the knowledge obtained during the signaling phase. In the following, we describe the proposed feature functions $\mathrm{f}_5(S_i, p_{n,i}^{\mathrm{Tx}})$ and $\mathrm{f}_6(S_i, p_{n,i}^{\mathrm{Tx}})$.

For the cooperative SARSA approach, 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 it is used 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, the data buffer level at N_1 depends only on the throughput of N_1 and its data arrival process which we cannot control. For this purpose, 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 (3.28)-(3.30). 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

$$\mathrm{f}_5(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } \left(R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})} + D_{2,i} - R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})} \leq D_{\max,2}\right) \wedge \\ & \quad \left(R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})} + D_{2,i} - R_{m,i}^{(\bar{p}_{m,i}^{\mathrm{Tx}})} \geq 0\right), \; n = \{1, 2\}, n \neq m \\ 0, & \text{else.} \end{cases} \tag{4.38}$$

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 and it is defined as

$$\mathrm{f}_6(S_i, p_{n,i}^{\mathrm{Tx}}) = \begin{cases} 1, & \text{if } p_{n,i}^{\mathrm{Tx}} = \underset{\bar{p}_{n,i}^{\mathrm{Tx}} \in \mathcal{A}_n}{\operatorname{argmin}} \left\{ D_{n,i} - R_{n,i}^{(\bar{p}_{n,i}^{\mathrm{Tx}})} \right\} \\ 0, & \text{else.} \end{cases} \tag{4.39}$$

4.2.5.7. Action selection policy

To select $p_{n,i}^{\mathrm{Tx}}$, each node follows the ϵ-greedy policy [SB18], i.e., with probability $1 - \epsilon$, node N_n selects the transmit power $p_{n,i}^{\mathrm{Tx}}$ that maximizes the action-value function for a given state S_i. This means that

$$\Pr\left[p_{n,i}^{\mathrm{Tx}} = \max_{p_{n,i}^{\mathrm{Tx}} \in \mathcal{A}_n} \hat{\mathrm{q}}_n^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) \right] = 1 - \epsilon, \quad 0 < \epsilon < 1. \tag{4.40}$$

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 [SB18, RN10].

Algorithm 4.2 Cooperative SARSA

```
1:  initialize $\gamma, \zeta, \epsilon$ and $\mathbf{w}_n$
2:  predict own channel coefficient                              ▷ Section 4.2.5.5
3:  exchange parameters and observe state $S_i$                  ▷ Section 4.2.5.5
4:  select $p_{n,i}^{\text{Tx}}$ using the $\epsilon$-greedy policy                   ▷ Eq. 4.40
5:  for every time interval $i = 1, ..., I$ do
6:      transmit using the selected $p_{n,i}^{\text{Tx}}$
7:      calculate corresponding reward $R_{2,i}^{\text{DF}}$                      ▷ Eq. (4.5)
8:      predict own channel coefficient                          ▷ Section 4.2.5.5
9:      exchange parameters and observe state $S_{i+1}$          ▷ Section 4.2.5.5
10:     select next $p_{n,i+1}^{\text{Tx}}$ using the $\epsilon$-greedy policy            ▷ Eq. (4.40)
11:     update $\mathbf{w}_n$                                          ▷ Eq. (4.28)
12:     set $S_i = S_{i+1}$ and $p_{n,i}^{\text{Tx}} = p_{n,i+1}^{\text{Tx}}$
13: end for
```

4.2.5.8. Cooperative SARSA algorithm

The proposed cooperative SARSA algorithm is summarized in Algorithm 4.2. Note that this algorithm is run at both, the transmitter and the relay. 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 $p_{n,i}^{\text{Tx}}$ (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 $p_{n,i+1}^{\text{Tx}}$ 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 the transmitter and the relay are operative.

4.2.5.9. 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 (4.26) [SB18]. Inspired by the work of [LR00], we show that the local action-value function $\text{q}_n^{\pi_n}(S_i, p_{n,i}^{\text{Tx}})$ is a projection of the centralized action-value function $\text{Q}^{\Pi}(S_i, P_i^{\text{Tx}})$ onto the corresponding state-action space $(S_i, p_{n,i}^{\text{Tx}})$. This means, the use of the local action-value function $\text{q}_n^{\pi_n}(S_i, p_{n,i}^{\text{Tx}})$ 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 4.2. *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* $Q_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ *and* $q_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ *are the values of the centralized and local action-value function in time interval i, respectively. Moreover the values of* $Q_i^{\Pi}(S_i, P_i^{\mathrm{Tx}})$ *and* $q_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ *are updated in each time interval using (4.25) and (4.26), 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*

$$q_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max_{\substack{P_i^{\mathrm{Tx}} = (p_{1,i}^{\mathrm{Tx}}, ..., p_{n,i}^{\mathrm{Tx}}) \\ P_i^{(l)} = p_{n,i}^{\mathrm{Tx}}}} Q_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}), \tag{4.41}$$

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

Proof. As in [LR00], 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 (4.41) holds. For arbitrary i, (4.41) 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 (4.25) and (4.26) are only performed on the particular pair $(S_i, p_{n,i}^{\mathrm{Tx}})$. Now, to prove (4.41) for the pair $(S_i, p_{n,i}^{\mathrm{Tx}})$, we include the right side of (4.41) in the update of $\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}})$ in (4.26) as

$$\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max \left\{ \max_{\substack{P_i^{\mathrm{Tx}} \\ P_i^{(l)} = p_{n,i}^{\mathrm{Tx}}}} \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}),\ R_i + \gamma \max_{P_{i+1}^{\mathrm{Tx}}} \mathrm{Q}_i^{\Pi}(S_{i+1}, P_{i+1}^{\mathrm{Tx}}) \right\}. \tag{4.42}$$

By considering the equality $\max\{\mathrm{f}(x) + a\} = a + \max\{\mathrm{f}(x)\}$, (4.42) can be rewritten as

$$\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max \left\{ \max_{\substack{P_i^{\mathrm{Tx}} \\ P_i^{(l)} = p_{n,i}^{\mathrm{Tx}}}} \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}),\ \max_{P_{i+1}^{\mathrm{Tx}}} \Big\{ R_i + \gamma \mathrm{Q}_i^{\Pi}(S_{i+1}, P_{i+1}^{\mathrm{Tx}}) \Big\} \right\}. \tag{4.43}$$

From (4.25), it is clear that the second term on the right side of (4.43) 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 (4.43) as

$$\mathrm{q}_{n,i+1}^{\pi_n}(S_i, p_{n,i}^{\mathrm{Tx}}) = \max \left\{ \max_{\substack{P_i^{\mathrm{Tx}} \\ P_i^{(l)} = p_{n,i}^{\mathrm{Tx}}}} \mathrm{Q}_i^{\Pi}(S_i, P_i^{\mathrm{Tx}}),\ \mathrm{Q}_{i+1}^{\Pi}(S_i, P_i) \right\}. \tag{4.44}$$

By expanding the term on the right side of (4.44), we obtain

$$\begin{aligned}\mathrm{q}^{\pi_n}_{n,i+1}(S_i, p^{\mathrm{Tx}}_{n,i}) =&\max\Big\{\Big\{\mathrm{Q}^{\Pi}_i(S_i, P^{\mathrm{Tx}}_i) \mid P^{(l)}_i = p^{\mathrm{Tx}}_{n,i}, P^{\mathrm{Tx}}_j \neq P^{\mathrm{Tx}}_i\Big\} \cup \\ &\Big\{\mathrm{Q}^{\Pi}_i(S_i, P^{\mathrm{Tx}}_i) \mid P^{(l)}_i = p^{\mathrm{Tx}}_{n,i}, P^{\mathrm{Tx}}_j = P^{\mathrm{Tx}}_i\Big\} \cup \big\{\mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i)\big\}\Big\}.\end{aligned} \tag{4.45}$$

The first term on the right side of (4.45) is equal to $\mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i)$ because for $P^{\mathrm{Tx}}_j \neq P^{\mathrm{Tx}}_i$ there is no update. The second term is always smaller than or equal to $\mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i)$ because, as the rewards are always greater than or equal to zero, $\mathrm{Q}^{\Pi}(S_i, P^{\mathrm{Tx}}_i)$ is monotonically increasing. $\mathrm{q}^{\pi_n}_{n,i}(S_i, p^{\mathrm{Tx}}_{n,i})$ is then written as

$$\begin{aligned}\mathrm{q}^{\pi_n}_{n,i+1}(S_i, p^{\mathrm{Tx}}_{n,i}) =&\max\Big\{\Big\{\mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i) \mid P^{(l)}_i = p^{\mathrm{Tx}}_{n,i}, P^{\mathrm{Tx}}_j \neq P^{\mathrm{Tx}}_i\Big\} \cup \big\{\mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i)\big\}\Big\} \\ =&\max_{\substack{P^{\mathrm{Tx}}_i=(p^{\mathrm{Tx}}_{1,i},\ldots,p^{\mathrm{Tx}}_{n,i}) \\ P^{(l)}_i=p^{\mathrm{Tx}}_{n,i}}} \mathrm{Q}^{\Pi}_{i+1}(S_i, P^{\mathrm{Tx}}_i).\end{aligned} \tag{4.46}$$

□

4.2.5.10. 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 as in Section 3.5.2.8. By examining Algorithm 4.2, 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 $p^{\mathrm{Tx}}_{n,i}$ (Lines 3 and 8) and the update of $\mathbf{w}_n$ (Line 9). The complexity of the Kalman-filter based channel estimator scales as $O(o^3)$ [Dau05], where o is the order of the filter. Furthermore, for the selection of $p^{\mathrm{Tx}}_{n,i}$, the ϵ-greeedy policy is considered. In this case, the highest complexity is due to the calculation of $\mathrm{q}^{\pi_n}(S_i, p^{\mathrm{Tx}}_{n,i})$ for all the possible actions and the selection of the $p^{\mathrm{Tx}}_{n,i}$ that leads to the maximum $\mathrm{q}^{\pi_n}(S_i, p^{\mathrm{Tx}}_{n,i})$. The computational complexity for the calculation of $\mathrm{q}^{\pi_n}(S_i, p^{\mathrm{Tx}}_{n,i})$ 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 (4.28) 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.

4.2.6. Performance evaluation

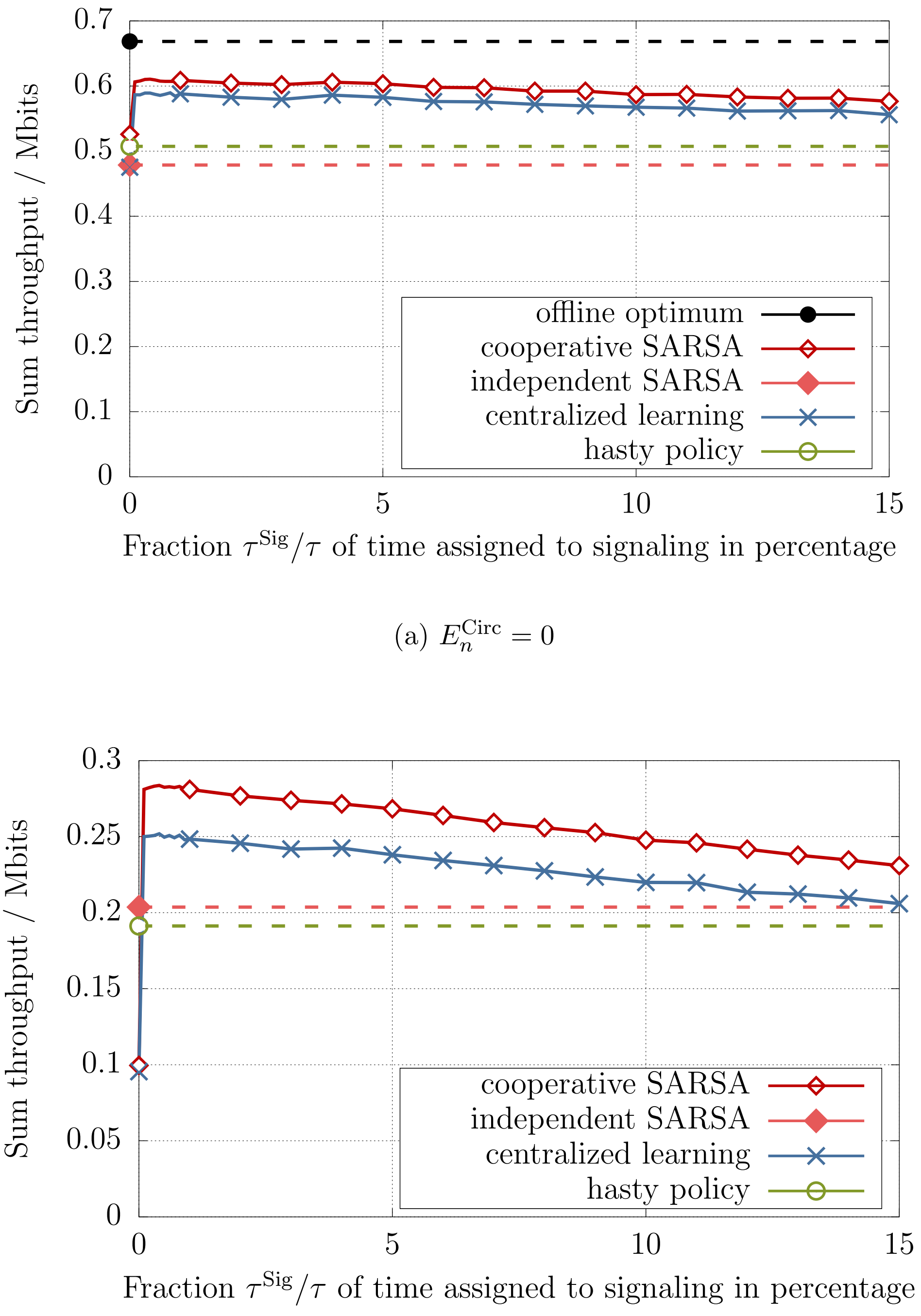
In this section, we present numerical results for the evaluation of the proposed offline and learning approaches in the EH two-hop communication scenario with a decode-and-forward relay. For the simulations, the parameters listed in Table 4.2 are considered, unless it is otherwise specified. In addition to the parameters described in Section 3.6, which we do not describe here for brevity, a finite data buffer with size $D_{\max,2}$ is assumed at N_2. $D_{\max,2}$ is calculated according to the average SNR Γ of the link between N_1 and N_2 as $D_{\max,2} = W\tau \log_2(1 + \Gamma)$. Furthermore, for the cooperative SARSA, a signaling phase of duration $\tau^{\mathrm{Sig}} = 0.1\mathrm{ms}$ is assumed, an autoregressive process of order $o = 2$ is used for the channel prediction [MS05], and a quantization error $e_{\mathrm{quant},u_{n,i}} = 1\%$ is considered for the transmission of the parameters during the signaling phase.

To compare the performance of the offline approach and the two proposed learning approaches, we consider the following reference schemes:

- 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. For simplicity, the resources required by N_2 to signal the transmit power to be used by N_1 are not taken into account.

- 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 4.3(a) and 4.3(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



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

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

Figure 4.3. Sum throughput versus fraction of time τ^{Sig}/τ assigned to signaling.

Table 4.2. Simulation set-up.

	Parameter	Value	Description
General	$e_{\text{quant},u_{n,i}}$	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 (finite data buffer case)
	$D_{\max,1}$	∞	Data buffer size of EH node N_1 (infinitely full data buffer case)
	$D_{\max,1}$	50kbits	Data buffer size of EH node N_1 (finite data buffer case)
	$D_{\max,2}$	$W\tau\log_2(1+\Gamma)$	Data buffer size of EH node N_2
	β	1	Data buffer size factor for EH ndoe N_2
	λ	10	Average number of packets arriving per time interval (finite data buffer case)
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

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 $E_n^{\text{Circ}} = 0$. Moreover, the offline optimum, independent SARSA 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 $\tau^{\text{Sig}}/\tau = 0$. Figure 4.3(a) considers that $E_n^{\text{Circ}} = 0$ 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 of the system dynamics. The achieved throughput of the cooperative SARSA and the centralized learning depends on the time assigned for the signaling. For $\tau^{\text{Sig}}/\tau < 15\%$, 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, i.e., the centralized approach requires a larger number of iterations to learn the optimal power allocation policies. In Figure 4.3(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 (4.37), 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 4.3(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 the independent SARSA approach outperforms the hasty policy. This is because both learning approaches 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 independent SARSA 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 4.4. To evaluate different values of the data buffer size at N_2, we consider the data buffer size factor β and calculate $D_{\max,2} = W\tau \log_2(1+\beta\Gamma)$. 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 independent learning approach and 43% less than the hasty policy. The better performance of the cooperative SARSA results from the fact that by exchanging the

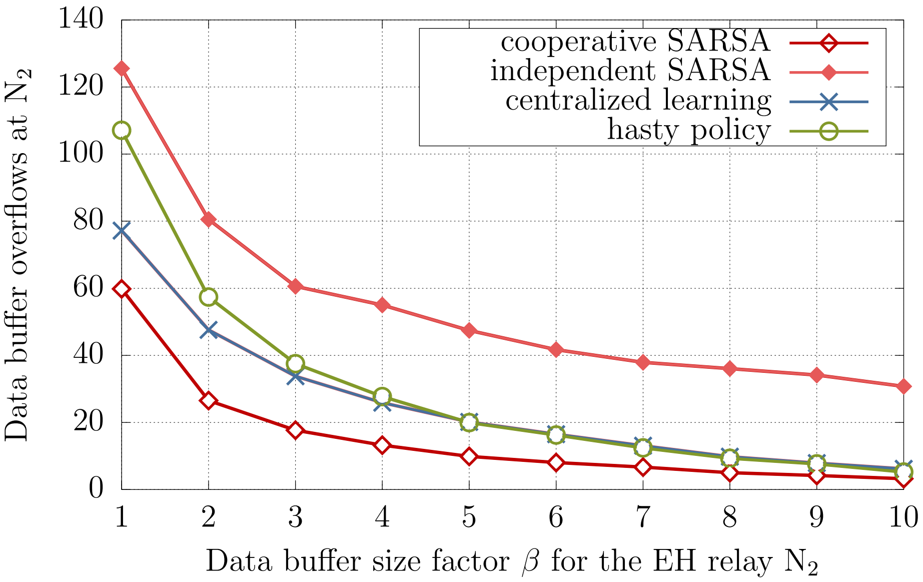


Figure 4.4. Number of data buffer overflows at N_2 versus the data buffer size factor β.

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 4.5 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 parameter λ. Moreover, we 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 4.5, 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 centralized approach, the independent SARSA approach and the hasty policy because it prevents data buffer overflows at N_2, as previously observed in Figure 4.4. In this case, the performance of the centralized learning is further decreased because the consideration of the state of

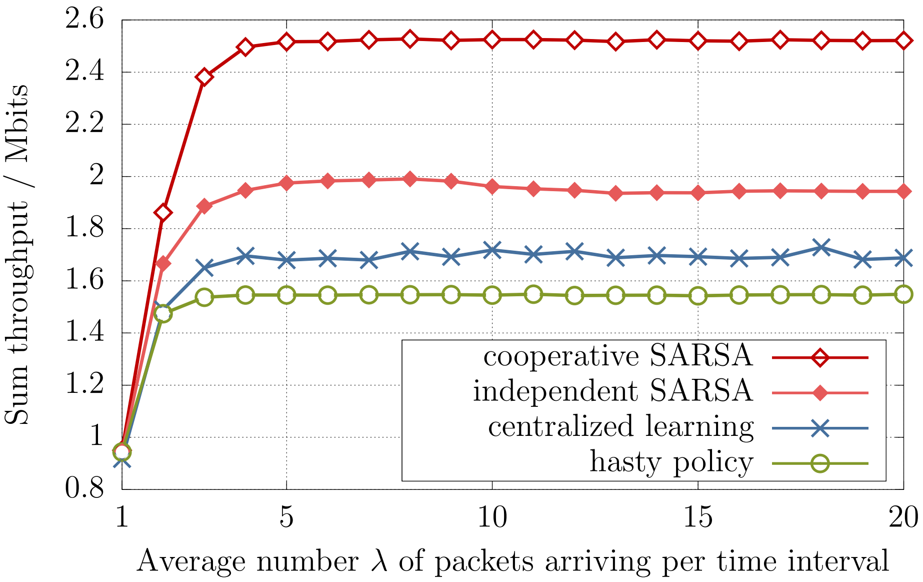


Figure 4.5. Sum throughput versus the average number λ of incoming data packets.

the data buffer at N_1 increases the dimensions of the state-action space and reduces the learning speed.

The impact of the battery size on the achieved throughput is evaluated in Figure 4.6. 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 independent SARSA approach. Moreover, its performance is 13% and 47% higher than for the centralized approach and for the hasty policy, respectively.

In Figure 4.7, 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 independent SARSA 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

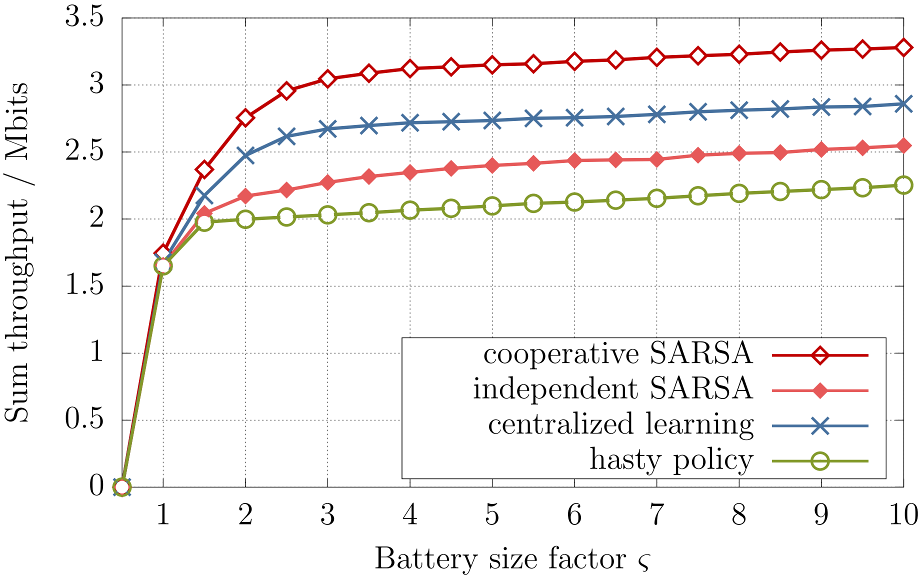


Figure 4.6. Sum throughput versus battery size factor ς.

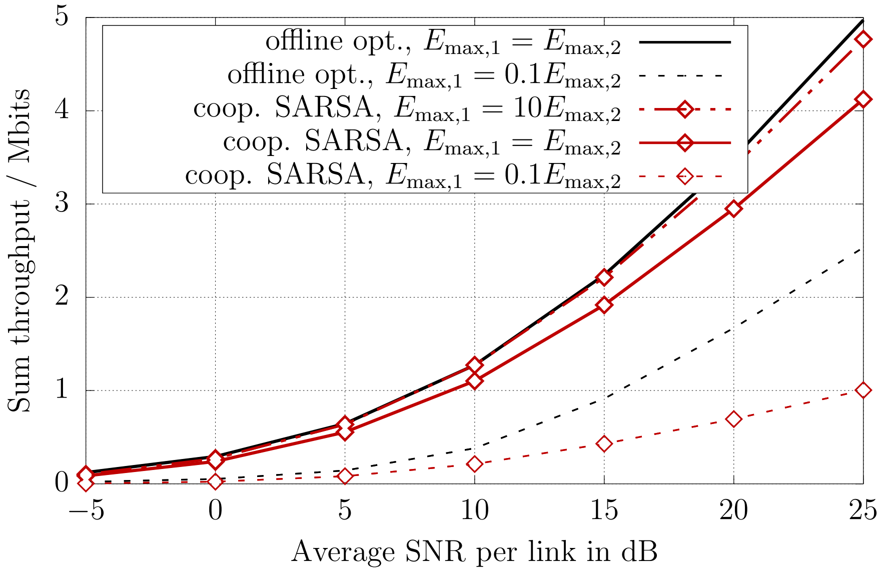


Figure 4.7. Sum throughput versus average SNR per link.

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 4.8, 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 independent SARSA 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) [GWT$^+$13]. 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 [GWT$^+$13]. 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 [SB18, GWT$^+$13]. In contrast to FSR, in RBF a given state is represented by more than one feature function. In Figure 4.8, it can be seen that the cooperative SARSA, the centralized approach and the independent SARSA 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,

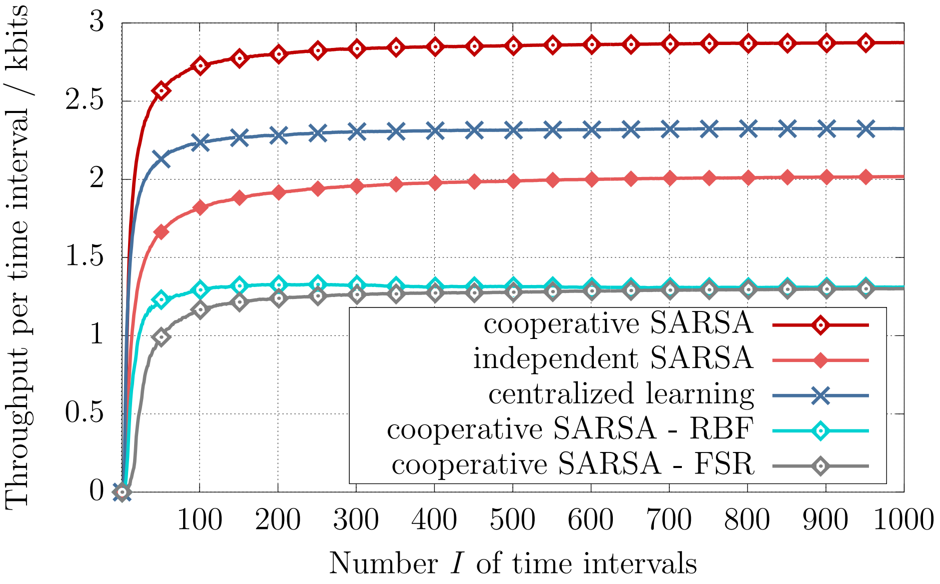


Figure 4.8. Throughput per time interval versus the number I of time intervals.

it can achieve a larger throughput. Furthermore, note that the number of feature functions required by a learning approach affects 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 has 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. When the nodes cooperate with each other, a higher throughput can be achieved. 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.

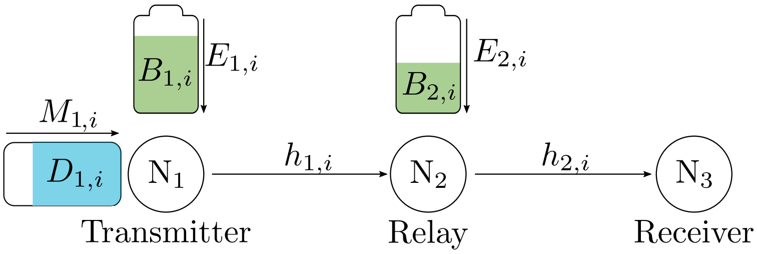


Figure 4.9. Two-hop communication scenario with an EH transmitter and an EH amplify-and-forward relay.

4.3. Amplify-and-forward relay

4.3.1. Scenario description and assumptions

In this section, we present the scenario description and the system assumptions when an amplify-and-forward relay is considered.

The EH two-hop communication scenario with an amplify-and-forward relay is depicted in Figure 4.9. Similar to the decode-and-forward case of Section 4.2.1, the scenario consists of three single-antenna nodes, in which the EH transmitter N_1 wants to transmit data to the non-EH receiver N_3 via the EH amplify-and-forward relay N_2. For brevity, we do not explain again the parameters which are common to the decode-and-forward case, but focus on the description of how the communication takes place when an amplify-and-forward relay is considered. We kindly refer the reader to Table 4.1 for a summary of the considered parameters.

For the communication between N_1 and N_3 only half-duplex transmission is considered. This is because the feasibility of achieving full-duplex transmission with an amplify-and-forward relay depends on the SNR regime in which the relay operates [RWWZ09], i.e., full-duplex transmission is only feasible in the high SNR regime. Therefore, taking into account that EH communications are mainly operating in the low SNR regime, we focus on a half-duplex amplify-and-forward relay, i.e., $\Delta = 1/2$. Specifically, the communication between N_1 and N_3 is as follows. In every time interval i, N_1 selects a transmit power $p_{1,i}^{\mathrm{Tx}}$ for the transmission of a signal $x_{1,i}$ to N_2 which contains the data intended for N_3. Considering the channel coefficient $h_{1,i}$ for the link between N_1 and N_2, and the noise $w_{2,i}$ at N_2, the received signal $y_{2,i}$ is written as

$$y_{2,i} = x_{1,i}h_{1,i} + w_{2,i}. \tag{4.47}$$

In contrast to the decode-and-forward relay, $y_{2,i}$ is not decoded at N_2 but only amplified. This means, no data buffer is considered at N_2 and, consequently, $D_{\max,2} = 0$. For the communication between N_2 and N_3, let $p_{2,i}^{\text{Rx}} = \mathbb{E}\{|y_{2,i}|^2\} = g_{1,i}p_{1,i}^{\text{Tx}} + \sigma_2^2$ be the power of the received signal at N_2, with σ_2^2 being the noise power at N_2, and let $\theta_i \in \mathbb{C}$ be the amplification factor at the relay which must fulfill the power constraint on the transmit power $p_{2,i}^{\text{Tx}}$ of N_2 given by

$$|\theta_i|^2(p_{1,i}^{\text{Tx}}g_{1,i} + \sigma_2^2) = p_{2,i}^{\text{Tx}}. \tag{4.48}$$

Note that $p_{2,i}^{\text{Tx}}$ is selected in each time interval i and depends on the amount of energy available in the battery. Then, considering the channel coefficient $h_{2,i}$ for the link between N_2 and N_3, and the noise $w_{3,i}$ at N_3, the received signal $y_{3,i}$ at N_3 can be written as

$$y_{3,i} = \theta_i x_{1,i}h_{1,i}h_{2,i} + \theta_i w_{2,i}h_{2,i} + w_{3,i}. \tag{4.49}$$

We assume that the noise power at N_3 is σ_3^2. Furthermore, we consider $\sigma_2^2 = \sigma_3^2 = \sigma^2$. As a result, from (4.48) and (4.49), the SNR $\Gamma_{3,i}$ of the received signal at N_3 in time interval i is written as

$$\Gamma_{3,i} = \frac{g_{1,i}g_{2,i}p_{1,i}^{\text{Tx}}p_{2,i}^{\text{Tx}}}{\sigma^2\left(g_{1,i}p_{1,i} + g_{2,i}p_{2,i} + \sigma^2\right)}. \tag{4.50}$$

Assuming a bandwidth W is available for the communication and enough data is available in the data buffer of N_1, the corresponding achieved throughput R_i^{AF}, which is the amount of data transmitted in one time interval, is approximated using Shannon's capacity formula as

$$R_i^{\text{AF}} = \frac{W\tau}{2}\log_2\left(1 + \Gamma_{3,i}\right), \tag{4.51}$$

where the factor $1/2$ comes from the half-duplex nature of the communication. Note that if the amount of data in the data buffer is not enough, the throughput is limited by the data buffer level of N_1. Furthermore, the battery levels at N_n, $n = \{1, 2\}$ are updated using (4.3) Additionally, considering (2.6), the data buffer level $D_{1,i}$ at N_1 is updated as

$$D_{1,i+1} = \min\left\{D_{\max,1},\, D_{1,i} - R_i^{\text{AF}} + M_{1,i}\right\}. \tag{4.52}$$

As in the decode-and-forward relay case, it is assumed that transmitter side channel state information is available at the transmitter and at the relay. For the offline approach, it is assumed that this channel state information is non-causally known at N_1 and N_2. On the contrary for the learning approach, it is realistically assumed that the channel state information is only causally known.

4.3.2. Problem formulation

In this section, the power allocation problem for the EH two-hop scenario with an amplify-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 Section 2.2, and the scenario description of Section 4.3.1, the power allocation problem is written as

$$\left(p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{n,i} = \underset{\{p_{n,i}^{\mathrm{Tx}},\, n=\{1,2\}, i=\{1,\ldots,I\}\}}{\operatorname{argmax}} \sum_{i=1}^{I} R_i^{\mathrm{AF}} \tag{4.53a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \frac{1}{2}\tau p_{n,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{n,i}, \;\; n=1,2, \;\; J=1,...,I, \tag{4.53b}$$

$$\sum_{i=1}^{J} E_{n,i} - \sum_{i=1}^{J} \frac{1}{2}\tau p_{n,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq B_{\max,n}, \; n=1,2, \;\; J=1,...,I, \tag{4.53c}$$

$$\sum_{i=1}^{J} R_i^{\mathrm{AF}} \leq \sum_{i=1}^{J-1} M_{1,i}, \;\; J=1,...,I, \tag{4.53d}$$

$$\sum_{i=1}^{J} M_{1,i} - \sum_{i=1}^{J} R_i^{\mathrm{AF}} \leq D_{\max,1}, \;\; J=1,...,I, \tag{4.53e}$$

$$p_{n,i}^{\mathrm{Tx}} \geq 0, \quad n=1,2,, \; i=1,...,I, \tag{4.53f}$$

where R_i^{AF} is defined in (4.51), (4.53b) is the energy causality constraint and (4.53c) is the battery overflow constraint for N_1 and N_2, respectively,. Moreover, (4.53d) is the data causality constraint and (4.53e) is the data buffer overflow constraint for N_1. Note that when an infinitely full data buffer is considered at N_1, the constraints (4.53d) and (4.53e) are not taken into account.

The objective function in (4.53a) is non-convex with respect to the optimization variables. As a result, (4.53) is a non-convex optimization problem and a closed-form solution cannot be obtained. Nevertheless, using basic properties of logarithms, the objective function can be rewritten as the difference of two concave functions. Consequently, the optimization problem of (4.53) is reformulated as a difference of concave functions (D.C.) programming problem for which an offline approach can be derived [HPT00]. Applying the quotient and product properties of logarithms, (4.53a) is rewritten as

$$R^{\mathrm{AF}}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right) = \mathrm{f}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right) - \mathrm{g}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right), \tag{4.54}$$

where $\mathrm{f}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right)$ and $\mathrm{g}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right)$ are two concave functions given by

$$\mathrm{f}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right) = \frac{1}{2}\sum_{i=1}^{I} W\tau\left[\log_2(g_{1,i}p_{1,i}^{\mathrm{Tx}} + \sigma^2) + \log_2(g_{2,i}p_{2,i}^{\mathrm{Tx}} + \sigma^2)\right], \tag{4.55}$$

and

$$\mathrm{g}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right) = \frac{1}{2}\sum_{i=1}^{I} W\tau \log_2(g_{1,i}p_{1,i} + g_{2,i}p_{2,i} + \sigma^2). \tag{4.56}$$

Using (4.55) and (4.56), problem (4.53) is reformulated as

$$\left(p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}}\right)_{n,i} = \underset{\{p_{n,i}^{\mathrm{Tx}},\, n=\{1,2\}, i=\{1,...,I\}\}}{\operatorname{argmax}} \left(\mathrm{f}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right) - \mathrm{g}\left(p_{1,i}^{\mathrm{Tx}}, p_{2,i}^{\mathrm{Tx}}\right)\right) \tag{4.57a}$$

$$\text{subject to} \quad \sum_{i=1}^{J}\frac{1}{2}\tau p_{n,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{n,i},\ n=1,2,\ J=1,...,I, \tag{4.57b}$$

$$\sum_{i=1}^{J} E_{n,i} - \sum_{i=1}^{J}\frac{1}{2}\tau p_{n,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_n^{\mathrm{Circ}} \leq B_{\max,n},\ n=1,2,\ J=1,...,I, \tag{4.57c}$$

$$\sum_{i=1}^{J} R_i^{\mathrm{AF}} \leq \sum_{i=1}^{J-1} M_{1,i},\ J=1,...,I, \tag{4.57d}$$

$$\sum_{i=1}^{J} M_{1,i} - \sum_{i=1}^{J} R_i^{\mathrm{AF}} \leq D_{\max,1},\ J=1,...,I, \tag{4.57e}$$

$$p_{n,i}^{\mathrm{Tx}} \geq 0, \quad n=1,2,,\ i=1,...,I, \tag{4.57f}$$

4.3.3. Offline approach

4.3.3.1. Branch-and-bound algorithm

The optimization problem in (4.57) is still a non-convex optimization problem. However, due to its formulation as the difference between two concave functions, an offline approach to find the optimal power allocation policy can be developed. Specifically, we propose a branch-and-bound algorithm to find the power allocation at N_1 and N_2 that maximizes the throughput. The proposed algorithm is inspired by [HPT00], where branch-and-bound algorithms for canonical D.C. programming problems are discussed.

In general, branch-and-bound is an iterative algorithm which works as follows. A recurrent partitioning of the feasible region (branching) is performed and in each iteration, one partition is considered. For the partition, the corresponding lower and upper bounds of the objective function are calculated (bounding) and based on these bounds, decision rules are applied to determine if the partition should be further divided or not. The algorithm stops when there are no more partitions to examine. Due to the complexity of the offline approach, we simplify the problem in (4.57) by considering an infinitely full data buffer at N_1, infinite battery capacities at N_1 and N_2, and that no energy is consumed by the circuits, i.e., $D_{\max,1} = \infty$, $D_{1,i} = \infty$, $\forall i$, $B_{\max,1} = B_{\max,2} = \infty$ and $E_1^{\mathrm{Circ}} = E_2^{\mathrm{Circ}} = 0$. Furthermore, to facilitate the notation and the description of the offline approach, we write the transmit power values $p_{n,i}^{\mathrm{Tx}}$ and the amounts $E_{n,i}$ of energy in vector form. Such notation aggregates the values taken by the parameters in the different time intervals. For this purpose, let the vector $\mathbf{p} \in \mathbb{R}^{2I\times 1}$ contain the power allocation of both, the transmitter and the relay such that

$$\mathbf{p} = \left(p_{1,1}^{\mathrm{Tx}}, ..., p_{1,I}^{\mathrm{Tx}}, p_{2,1}^{\mathrm{Tx}}, ..., p_{2,I}^{\mathrm{Tx}}\right)^{\mathrm{T}}. \tag{4.58}$$

Moreover, let the vector $\mathbf{e} \in \mathbb{R}^{2I\times 1}$ contain the amounts $E_{1,i}$ and $E_{2,i}$ of harvested energy for N_1 and N_2, as

$$\mathbf{e} = \left(E_{1,1}, E_{1,1} + E_{1,2}, ..., E_{1,1} + ... + E_{1,I}, E_{2,1}, E_{2,1} + E_{2,2}, ..., E_{2,1} + ... + E_{2,I}\right)^{\mathrm{T}}, \tag{4.59}$$

and let the matrix $\mathbf{T} \in \mathbb{R}^{2I\times 2I}$ be defined as

$$\mathbf{T} = \tau \begin{pmatrix} \mathbf{L}_I & \mathbf{0}_{I\times I} \\ \mathbf{0}_{I\times I} & \mathbf{L}_I \end{pmatrix}, \tag{4.60}$$

where $\mathbf{L}_I$ is an $I \times I$ lower triangular matrix of ones and $\mathbf{0}_{I\times I}$ is an $I \times I$ matrix of zeros. Considering (4.58), (4.59) and (4.60), the problem in (4.57) reduces to

$$\mathbf{p}^{\mathrm{opt}} = \underset{\mathbf{p}}{\operatorname{argmax}} \left(\mathrm{f}(\mathbf{p}) - \mathrm{g}(\mathbf{p})\right) \tag{4.61a}$$

$$\text{subject to} \quad \mathbf{T}\mathbf{p} \leq \mathbf{e}, \tag{4.61b}$$

$$\mathbf{p} \geq \mathbf{0}_{2I}, \tag{4.61c}$$

where $\mathbf{0}_{2I}$ is a vector of zeros of length $2I$.

4.3.3.2. Partitioning the feasible region

According to [HPT00], to facilitate the branching, an initial simplex is constructed from the feasible region. An m-simplex is a polytope which is the convex hull of its $m + 1$ affinely independent vertices [HPT00]. Depending on the decision rules, this initial simplex is partitioned using bisection in each iteration. The use of bisection ensures that the resulting partitions are simplices as well. However, the feasible region described by (4.61b) and (4.61c) does not fulfill the definition of a simplex because the available power in each time interval depends on the previous power allocations. In the considered scenario, two nodes harvest energy independently during I time intervals. Therefore, for each node, I different power values are calculated. This means, the dimension of the problem is $2I$ and the feasible region is a $2I$-dimensional polytope. Consequently, to construct a simplex, non-feasible power values must be considered in addition to the feasible region.

The initial simplex must include the complete feasible region. Hence, we propose to create the initial simplex based on the maximum power values that can be allocated to the nodes. Remember that if a node saves all the harvested energy and transmits only during the last interval, the maximum power that can be allocated to it is calculated using (4.61b) as $p_{n,i}^{\mathrm{Tx}} = \frac{1}{\tau}\sum_{i=1}^{I} E_{n,i}$, for $n = \{1, 2\}$. Therefore, a simplex whose vertices are given by the sum of the maximum power values of all the EH nodes is guaranteed to include the complete feasible region. In other words, the $2I + 1$ vertices $\mathbf{v}_j \in \mathbb{R}^{2I\times 1}$, with $j = 0, ..., 2I$, of the initial simplex are calculated as

$$\mathbf{v}_j = \begin{cases} \mathbf{0}_{2I\times 1} & \text{if } j = 0, \\ [v_{j,1}, v_{j,2}, ..., v_{j,2I}]^{\mathrm{T}} & \text{if } j = 1, ..., 2I, \end{cases} \tag{4.62}$$

where $v_{j,k},\ k = 1, ..., 2I$ are the elements in the vertex vector $\mathbf{v}_j$ which are calculated as

$$v_{j,k} = \begin{cases} \frac{1}{\tau}\sum\limits_{i=1}^{I}(E_{1,i} + E_{2,i}) & j = k, \\ 0 & j \neq k. \end{cases} \tag{4.63}$$

To illustrate the feasible region, let us consider the simplest case of $I = 1$. From the constraint of (4.61b), the maximum power values for N_1 and N_2 are given by $\frac{E_{1,1}}{\tau}$ and $\frac{E_{2,1}}{\tau}$, respectively. Similarly, from (4.61c), the minimum power value is zero for both nodes. The resulting feasible region corresponds to a rectangle as shown in Figure 4.10. The required initial simplex is calculated using (4.62) and (4.63). The result is the triangle shown in Figure 4.10 which contains the complete feasible region and some non-feasible power values.

4.3.3.3. Lower and upper bounds

In this section, we describe the calculation of the lower and upper bounds of the objective function. As mentioned in the previous section, the branch-and-bound algorithm works in an iterative fashion. In each iteration, a partition of the initial simplex is considered and the corresponding lower and upper bounds are calculated. For this

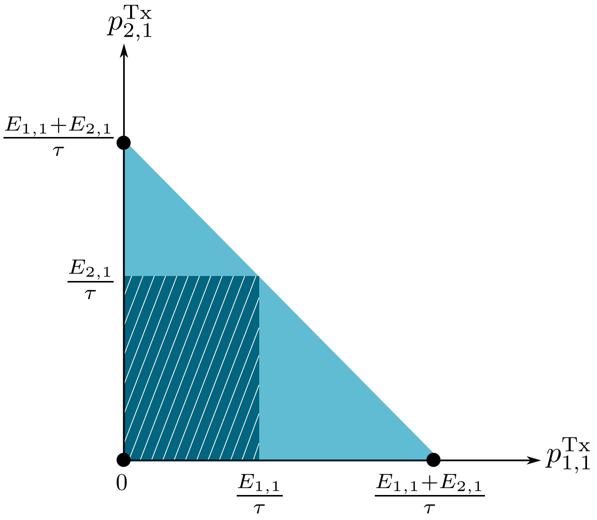


Figure 4.10. Example of the feasible region and the initial simplex in a scenario where $I = 1$.

purpose, decision rules are applied to determine if the considered partition should be further divided.

In the reference work of [HPT00, ASW12], the bounds are calculated by relaxing the D.C. problem into a linear problem. However, in this approach the number of constraints increases linearly with the number of iterations. Therefore, to reduce the complexity in the calculation of the bounds, we propose to linearize only the objective function (4.61a). As a result, the optimization problem in (4.61) is relaxed into a convex problem. As described in [HPT00], to linearize the objective function, an artificial variable φ is included in (4.61). Moreover, a property of simplices is used to rewrite the power variables as a function of the vertices of the considered simplex. Given that any point in a simplex can be uniquely represented as a weighted sum of the vertices [HPT00], any vector $\mathbf{p}$ in the considered simplex can be written as

$$\mathbf{p} = \sum_{j=0}^{2I} \vartheta_j \mathbf{v}_j, \tag{4.64}$$

where $\vartheta_j, j = 0, ..., 2N$ are the weights which satisfy the condition that $\sum_{j=0}^{2I} \vartheta_j = 1$. Taking into account that $\mathrm{g}(\mathbf{p})$ is a concave function, its lower bound is found considering (4.64) as

$$\sum_{j=0}^{2I} \vartheta_j \mathrm{g}(\mathbf{v}_j) \leq \mathrm{g}\left(\sum_{j=0}^{2I} \vartheta_j \mathbf{v}_j\right), \tag{4.65}$$

where the equality is met at the vertices. To include the artificial variable φ, the constraint

$$\varphi - \mathrm{f}\left(\sum_{j=0}^{2I} \vartheta_j \mathbf{v}_j\right) \leq 0, \tag{4.66}$$

has to be fulfilled. Considering (4.64), (4.65) and (4.66) the problem in (4.61) is relaxed into the convex problem

$$(\varphi^{\mathrm{opt}}, \vartheta_0^{\mathrm{opt}}, ..., \vartheta_{2I}^{\mathrm{opt}}) = \underset{\varphi,\vartheta_0,...,\vartheta_{IN}}{\mathrm{argmax}} \left(\varphi - \sum_{j=0}^{2I} \vartheta_j \mathrm{g}(\mathbf{v}_j)\right) \tag{4.67a}$$

$$\text{subject to} \quad \varphi - \mathrm{f}\left(\sum_{j=0}^{2I} \vartheta_j \mathbf{v}_j\right) \leq 0, \tag{4.67b}$$

$$\mathbf{T}\sum_{j=0}^{2I} \vartheta_j \mathbf{v}_j \leq \mathbf{e}, \tag{4.67c}$$

$$\sum_{j=0}^{2I} \vartheta_j = 1, \tag{4.67d}$$

$$0 \leq \vartheta_j \leq 1,\ j = 0, ..., 2I, \tag{4.67e}$$

where the new optimization variables are the weighting factors ϑ_j and φ.

The solution of (4.67) leads to the calculation of the upper bound UB as

$$\mathrm{UB} = \varphi^{\mathrm{opt}} - \sum_{j=0}^{2I} \vartheta_j^{\mathrm{opt}} \mathrm{g}(\mathbf{v}_j). \tag{4.68}$$

However, note that UB is a non-achievable throughput value because it is obtained by linearizing the original objective. In other words, the objective function in (4.67) is an outer approximation of (4.61a).

The lower bound LB is calculated by applying the throughput function of (4.61a) to the obtained power vector as

$$\mathrm{LB} = \mathrm{f}\left(\sum_{j=0}^{2I} \vartheta_j^{\mathrm{opt}} \mathbf{v}_j\right) - \mathrm{g}\left(\sum_{j=0}^{2I} \vartheta_j^{\mathrm{opt}} \mathbf{v}_j\right). \tag{4.69}$$

In contrast to the upper bound, the lower bound LB is an achievable throughput value. Furthermore, note that in each iteration of the algorithm, the lower and upper bounds are calculated for the considered simplex. The largest LB among all the simplices leads to the maximum throughput.

4.3.3.4. Decision rules

In this section, the decision rules used to determine if the considered simplex should be further partitioned, are presented.

Our goal is to maximize the throughput. Therefore, considering that LB is an achievable throughput value, the highest lower bound, termed LB^{best}, leads to the maximum throughput. In every iteration of the algorithm, the value of LB^{best} is only updated if for a given simplex, the calculated LB is higher than the current LB^{best}. Moreover, as the initial simplex includes non-feasible power values, it is possible that simplices obtained during branching lie in a non-feasible region and consequently, lead to non-feasible solutions. In the algorithm, these solutions are ignored and the corresponding simplices are not further partitioned. This means, the decision rules presented in the following apply only to feasible solutions of (4.67).

Decision rule 1. If $\text{UB} < \text{LB}^{\text{best}}$, the considered simplex is not further partitioned because the current LB^{best} exceeds the corresponding UB of the simplex. This means, the power vector which leads to the maximum throughput cannot be in the region determined by the considered simplex.

Decision rule 2. If $\text{UB} - \text{LB}^{\text{best}} > \varepsilon$, where ε is the desired tolerance, the considered simplex is partitioned because it may contain a power allocation that leads to the maximum throughput.

Decision rule 3. If $0 \leq \text{UB} - \text{LB}^{\text{best}} \leq \varepsilon$, the considered simplex contains a local maximum given by LB^{best}. If no other simplex leads to a higher lower bound, then the current LB^{best} is considered as the maximum throughput.

4.3.3.5. Summary of the algorithm

The proposed branch-and-bound algorithm, used to find the power allocation at N_1 and N_2 that leads to the maximum throughput in the EH two-hop scenario with an amplify-and-forward relay, is summarized in Algorithm 4.3.

As described in the previous sections, we first determine the initial simplex based on the harvested energy of each node using (4.62) and (4.63) (line 1). Furthermore, as no simplex has yet been evaluated, LB^{best} is set to zero (line 2). Then, for every simplex, the corresponding upper bound is calculated using (4.68) (line 4) and the decision rules

Algorithm 4.3 Branch-and-bound algorithm

```
 1: create the initial simplex                              ▷ Eq. (4.62) and (4.63)
 2: set $\mathrm{LB}^{\mathrm{best}} = 0$
 3: while there are simplices to be inspected do
 4:     select a simplex and calculate UB                   ▷ Eq. (4.67) and (4.68)
 5:     calculate the corresponding $\mathbf{p}$                         ▷ Eq. (4.64)
 6:     if a feasible solution is found then
 7:         calculate the corresponding LB                  ▷ Eq. (4.69)
 8:         if $\mathrm{LB} > \mathrm{LB}^{\mathrm{best}}$ then
 9:             update $\mathrm{LB}^{\mathrm{best}}$ and the corresponding $\mathbf{p}^{\mathrm{best}}$
10:         end if
11:         if $\mathrm{UB} - \mathrm{LB}^{\mathrm{best}} > \varepsilon$ then
12:             partition the simplex using bisection
13:         end if
14:     end if
15: end while
16: return $R^{\mathrm{AF}} = \mathrm{LB}^{\mathrm{best}}$ and $\mathbf{p}^{\mathrm{opt}} = \mathbf{p}^{\mathrm{best}}$
```

described in Section 4.3.3.4 are considered in order to determine if the current lower bound yields a higher throughput than the current $\mathrm{LB}^{\mathrm{best}}$ (line 8), and whether the current simplex should be further divided or not (line 10). When there are no more simplices to inspect, the maximum achievable throughput R^{AF} is set equal to $\mathrm{LB}^{\mathrm{best}}$ and the corresponding power vector $\mathbf{p}$ is the optimal power allocation for N_1 and N_2.

4.3.4. Learning approach

In an EH two-hop scenario with an amplify-and-forward relay, the relay N_2 transmits an amplified version of the signal received from the transmitter N_1. Consequently, the communication between the transmitter N_1 and the receiver N_3 cannot be separated as in the decode-and-forward case, but has to be considered as a single link with an effective channel that depends on the channel from N_1 to the N_2, the relay gain and the channel from N_2 to N_3. For this reason, in this section, a centralized learning algorithm is proposed.

The proposed centralized learning approach is based on the algorithm presented in Section 3.5 for the EH point-to-point case. This is due to the fact that, for the learning approach, the two-hop scenario with an amplify-and-forward relay reduces to a point-to-point communication scenario because the two links (from N_1 to N_2 and from N_2 to N_3) are viewed as a single effective channel between N_1 and the N_3. However, note that in contrast to Chapter 3, signaling between the transmitter and the relay is required such that the system state is fully observable.

Algorithm 4.4 Centralized SARSA

```
 1: initialize $\gamma, \zeta, \epsilon, \mathbf{w}$
 2: estimate channel coefficients
 3: receive signaling from $\mathrm{N}_1$ and observe $S_i$                ▷ Section 4.2.5.5
 4: select $p_{1,1}^{\mathrm{Tx}}$ and $p_{2,1}^{\mathrm{Tx}}$ randomly
 5: while $\mathrm{N}_1$ and $\mathrm{N}_2$ are operative do
 6:     transmit using the selected $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$
 7:     calculate corresponding reward $R_i^{\mathrm{AF}}$                 ▷ Eq. (4.51)
 8:     estimate channel coefficients
 9:     receive signaling from $\mathrm{N}_1$ and observe next state $S_{i+1}$
10:     select next transmit power values $p_{1,i+1}^{\mathrm{Tx}}$ and $p_{2,i+1}^{\mathrm{Tx}}$ using $\epsilon$-greedy
11:     update $\mathbf{w}$                                               ▷ Eq. (3.24)
12:     set $S_i = S_{i+1}$
13:     set $p_{1,i}^{\mathrm{Tx}} = p_{1,i+1}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}} = p_{2,i+1}^{\mathrm{Tx}}$
14: end while
```

A summary of the proposed algorithm is presented in Algorithm 4.4. It is assumed that the relay N_2 is the central entity. As a result, in each time interval i, the transmitter N_1 signals its own parameters, i.e., $E_{1,i}$, $B_{1,i}$, and $D_{1,i}$ to N_2 such that N_2 can decide, using Algorithm 4.4, the power values $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$ that should be used. Note that the same procedure would apply if the transmitter N_1 were the node in charge of the learning. In such a case, N_2 would send its own parameters to N_1 in order to provide it with a view of the current system state. The centralized learning algorithm works as follows. First, the learning parameters are initialized (line 1). Next, the nodes estimate their own channel coefficients, and the transmitter N_1 sends its current parameters to the relay such that N_2 can observe the system state S_i (lines 2 and 3). For the channel estimation and the transmission of the signaling, the procedures described in Section 4.2.5.5 are followed. Afterwards, the first transmit power values $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$ are randomly selected by the relay (line 4). Note that in this scenario, the action space has quadratically increased compared to the point-to-point case because it contains all the possible permutations (with repetitions) of the power values. Next, the selected $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$ are used for the transmission of data and the corresponding achieved throughput R_1^{AF} is observed (lines 6 and 7). After the reward is observed, the nodes estimate their new channel coefficients and the transmitter N_1 sends its current parameters to N_2 (lines 8 and 9). The new transmit power values $p_{1,i+1}^{\mathrm{Tx}}$ and $p_{2,i+1}^{\mathrm{Tx}}$ are selected using the ϵ-greedy policy (line 10) and by considering S_i, $p_{1,i}^{\mathrm{Tx}}$, $p_{2,i}^{\mathrm{Tx}}$, R_i^{AF}, S_{i+1}, $p_{1,i+1}^{\mathrm{Tx}}$ and $p_{2,i+1}^{\mathrm{Tx}}$ the weights $\mathbf{w}$ are updated (line 12). The same procedure described above is repeated in each time interval in which the transmitter and the relay remain operative.

As the approach described in Algorithm 4.4 is based on the approximate SARSA algorithm of Chapter 3, the convergence guarantees described in Section 3.5.2.7 and computational complexity analysis of Section 3.5.2.8 apply.

Table 4.3. Simulation set-up.

Parameter	Value	Description
I	3	Number of time intervals
T	100	Number of realizations
τ	1s	Time interval duration
$B_{\max,n}$	∞	Battery capacity of EH node N_n
E_n^{Circ}	0mJ	Energy consumed by the circuit of EH node N_n
$D_{\max,1}$	∞	Data buffer size of EH node N_1
$g_{n,i}$	1	Channel gain of the link between N_n and N_{n+1}
W	1 Hz	Bandwidth

4.3.5. Performance evaluation

4.3.5.1. Offline approach

In this section, numerical results for the evaluation of the proposed offline and learning approaches are presented. Due to the complexity of the branch-and-bound algorithm, the same simulation set-up cannot be considered for the offline and the learning approaches, i.e., while a small number of intervals I should be considered in the offline approach because the complexity of the proposed branch-and-bound algorithm increases with I, the learning approach requires a large I in order to learn the power allocation policy. As a consequence, we separate the performance evaluation between the offline and learning approaches in Section 4.3.5.1 and Section 4.3.5.2, respectively.

For the offline optimum, it is assumed that the amount of harvested energy $E_{n,i}$ is taken from a uniform distribution with maximum value $E_{\max}$ and each realization is assumed to be known non-causally. The offline approach assumes infinite battery capacities at N_1 and N_2, an infinitely full data buffer at N_1 and no circuit energy consumption, $B_{\max,n} = \infty$, $D_{\max,n} = \infty$ and $E_n^{\text{Circ}} = 0$. Moreover, $T = 100$ random energy realizations consisting of $I = 3$ time intervals are considered and the time interval duration τ is assumed to be one second. Additionally, the channel gains are assumed to be one for all the time intervals, i.e., $g_{n,i} = 1, \forall i$ and a bandwidth $W = 1$Hz is considered. A summary of all the parameters is given in Table 4.3.

As the proposed branch-and-bound algorithm calculates the optimal power allocation, it provides the upper bound of the performance. In Figure 4.11, we evaluate the effect of the EH processes of N_1 and N_2 on the performance. For the comparison, we evaluate the sum throughput versus the amount of harvested energy $E_{\max}$ considering four different cases:

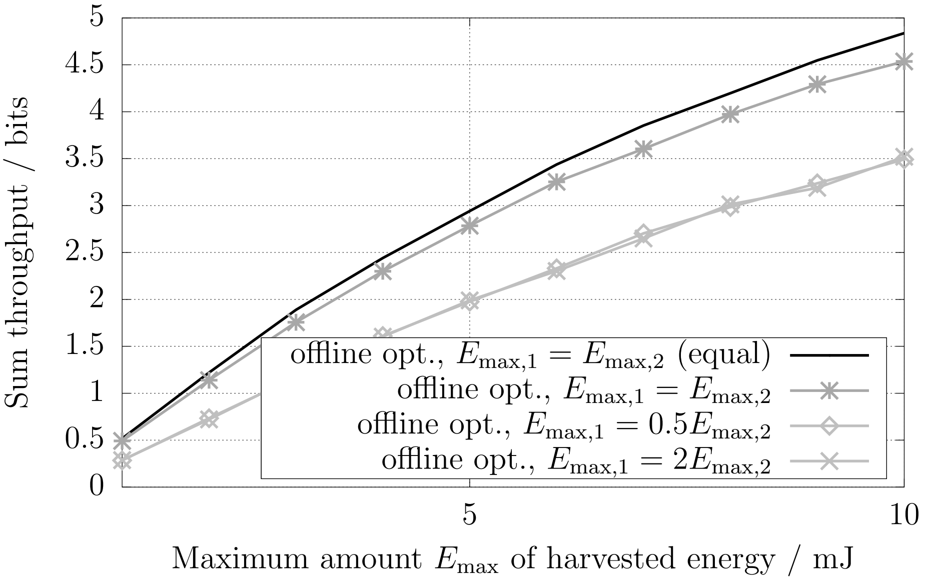


Figure 4.11. Sum throughput versus maximum harvested energy for an EH two-hop scenario and $I = 3$ time intervals.

- Equal energy: It is assumed that the EH processes of N_1 and N_2 are exactly the same. Consequently, the amount of harvested energy in each time instant is equal for both nodes and $E_{\max,1} = E_{\max,2} = E_{\max}$.

- Equal mean: In this case, the amounts of harvested energy $E_{1,i}$ and $E_{2,i}$ are independent, uniformly distributed random variables with maximum values $E_{\max,1} = E_{\max,2} = E_{\max}$.

- Double mean - relay: In this case, the amounts of harvested energy $E_{1,i}$ and $E_{2,i}$ are independent, uniformly distributed random variables with maximum values, $E_{\max,1} = 0.5E_{\max,2}$ and $E_{\max,2} = E_{\max}$.

- Double mean - transmitter: In this case, the amounts of harvested energy $E_{1,i}$ and $E_{2,i}$ are independent, uniformly distributed random variables with maximum values, $E_{\max,1} = 2E_{\max,2}$ and $E_{\max,2} = 0.5E_{\max}$.

The results show that the maximum throughput is achieved when the EH processes of the two nodes are equal. This is because the two-hop communication channel can be seen as a single effective channel whose capacity in each time interval depends on $p_{1,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}}$ simultaneously. Therefore, the throughput is maximized when the available energies at the nodes are equal. When only the mean values of the two EH processes is equal, the throughput is reduced compared to the initial case because in each realization, one of the nodes is limited compared to the other. The maximum reduction is observed when the mean values are not equal. Here, a larger reduction on the throughput is expected compared to the other two cases because the total amount of harvested energy of the two nodes is less than in the previous two cases. Interestingly, the throughput achieved when $E_{\max,1} = 0.5E_{\max,2}$ and $E_{\max,2} = E_{\max}$ is on average equal to the throughput achieved when $E_{\max,1} = 2E_{\max,2}$ and $E_{\max,2} = 0.5E_{\max}$. This means that the reduction in the throughput due to energy limitation does not depend on which EH node is limited, but on the difference between the maximum energy values $E_{\max,1}$ and $E_{\max,2}$ of N_1 and N_2, respectively.

4.3.5.2. Learning approach

In this section, we evaluate the performance of the proposed centralized SARSA algorithm. As a reference, we consider the hasty policy and the random power allocation policy, described in Section 4.2.6, which also assume only causal knowledge about the EH and channel fading processes. Additionally, we consider a two-hop scenario with a full-duplex decode-and-forward relay and the cooperative SARSA approach described in Section 4.2.5. Similar to the decode-and-forward case of Section 4.2.6, we assume that the amounts of harvested energy $E_{n,i}$, $n \in \{1, 2\}$ are taken from a uniform distribution with maximum value $E_{\max,n}$. Furthermore, the channel coefficients are modeled as complex Gaussian processes using the model described in [Küh11], and the variables listed in Table 4.2 are considered unless it is otherwise specified.

In Figure 4.12, the achieved sum throughput versus the fraction τ^{Sig}/τ of time assigned for the signaling is shown when an infinitely full data buffer is considered at N_1. The proposed centralized SARSA approach outperforms the reference schemes that consider an amplify-and-forward relay and causal knowledge regarding the EH, data arrival and channel fading processes. For a fraction $\tau^{\mathrm{Sig}}/\tau = 1\%$, the throughput achieved by the centralized SARSA is 2.5 and 2.7 times higher than the one achieved by the hasty and random policies, respectively. As expected, the highest throughput is achieved when a full-duplex decode-and-forward relay is considered. This is because, in a decode-and-forward scheme there is no noise amplification since the received signal from the transmitter is decoded at the relay. On the contrary, in the case of an amplify-and-forward relay, the received signal and the noise are both amplified. Moreover, the considered amplify-and-forward relay operates in half-duplex mode, thus further reducing the achievable throughput by one half.

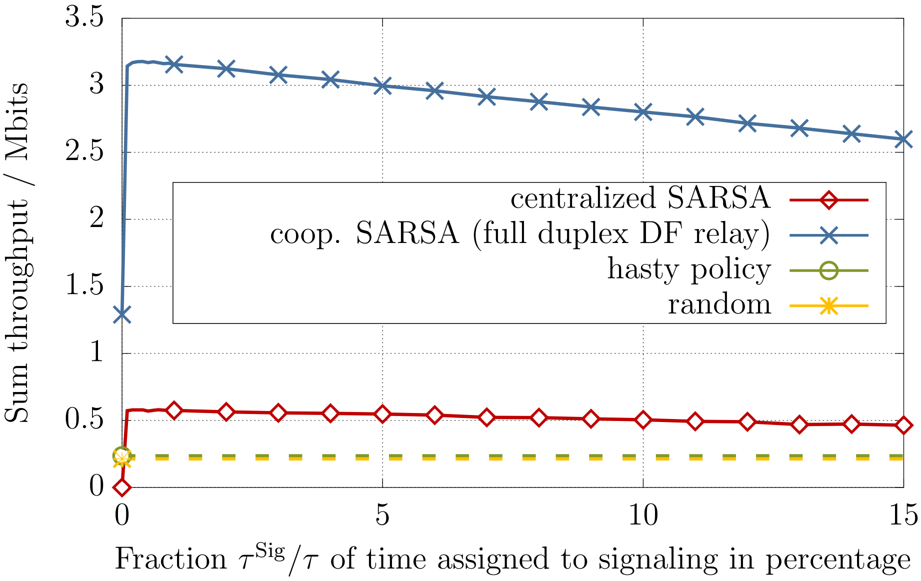


Figure 4.12. Sum throughput versus fraction of time τ^{Sig}/τ assigned to signaling.

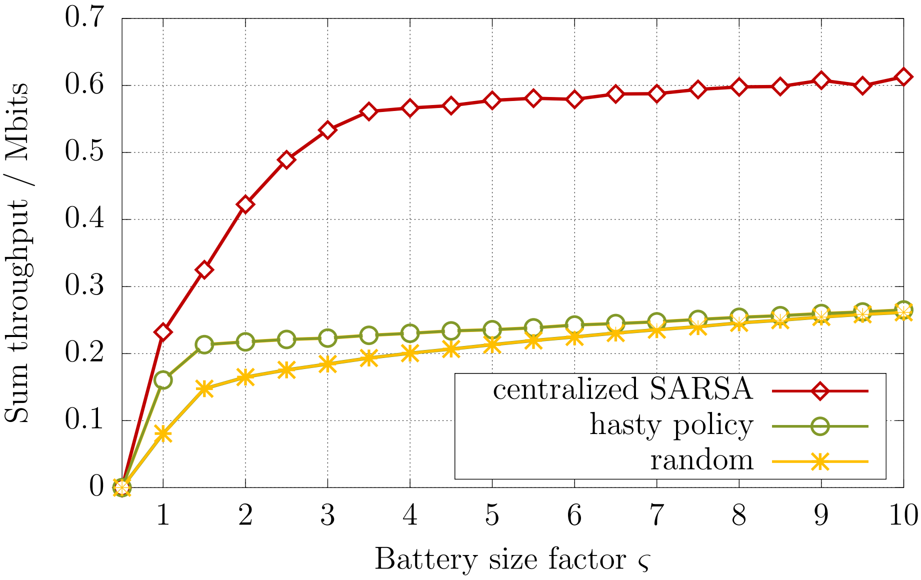


Figure 4.13. Sum throughput versus battery size.

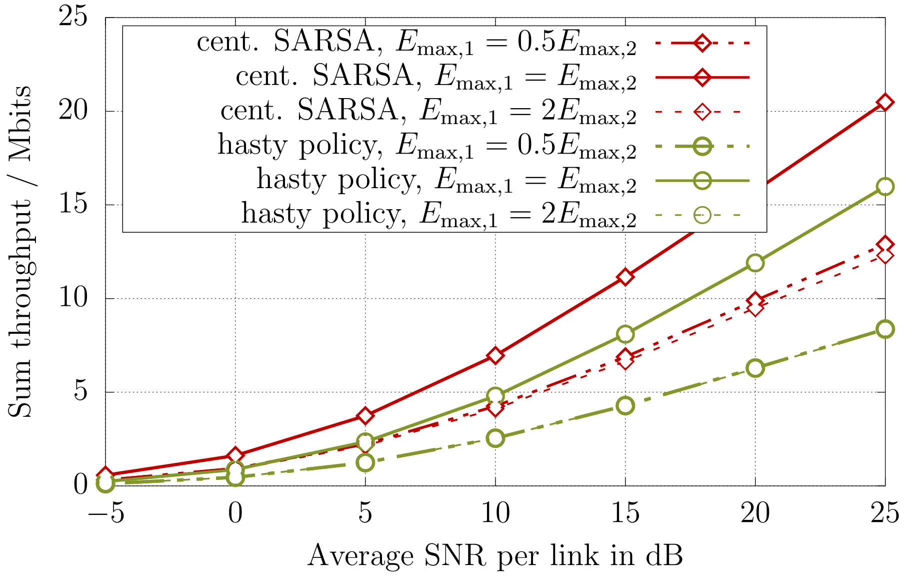


Figure 4.14. Sum throughput versus average SNR per link.

The impact of the battery capacity on the sum throughput is depicted in Figure 4.13. Here, it can be seen that the achieved throughput increases with the battery capacity. Furthermore, the centralized SARSA algorithm outperforms the benchmark schemes throughout the complete considered battery capacity range. For $\varsigma = 5$, centralized SARSA achieves a throughput 2.4 and 2.6 larger than the hasty policy and the random approach, respectively. The reason for this performance gain is that the centralized SARSA is able to adapt the power allocation policy according to the amounts of harvested energy and the channel gains. Interestingly, when the battery capacity is large, the performance of the random allocation reaches the performance of the hasty policy. This is because as ς increases, the probability of storing energy up to the capacity of the battery or the probability of having battery overflow situations reduces. As a result, the battery size becomes less relevant in the power allocation.

In Figure 4.14, we evaluate the achieved throughput as a function of the average SNR per link when $E_n^{\text{Circ}} = 0$. Note that it is assumed that each link, i.e., the link between N_1 and N_2, and the link between N_2 and N_3, have on average the same SNR. Additionally, we consider different ratios between the amounts of harvested energy at the transmitter and at the relay. Specifically, we consider that the maximum amount of harvested energy is fixed and given by $E_{\max} = \rho\Omega\tau = 16\text{mJ}$, and three cases are distinguished:

- $E_{\max,1} = E_{\max,2} = E_{\max}$.

- $E_{\max,1} = 0.5E_{\max,2}$ and $E_{\max,2} = E_{\max}$.
- $E_{\max,1} = 2E_{\max,2}$ and $E_{\max,2} = 0.5E_{\max}$

Similar to the result obtained for the offline approach, we observe in Figure 4.14 that the maximum throughput is achieved when the nodes harvest, on average, the same amount of energy. When one of the nodes is constrained, the overall achieved throughput is reduced because less energy is available for the transmission. Furthermore, the throughput achieved when one of the EH nodes harvests less energy is, on average, the same, regardless of which node harvests more energy. This behavior can also be observed for the hasty policy.

4.4. Extension to EH multi-hop relaying scenarios

In this section, the extension of the proposed approaches to EH multi-hop relaying scenarios is discussed.

The EH multi-hop communication scenario, consisting of a single EH transmitter which wants to transmit data to a single receiver using multiple intermediate EH relays in a multi-hop fashion, can be addressed using the proposed learning algorithms. Assuming only local causal knowledge at the EH nodes and decode-and-forward relays, it is straightforward to extend the independent SARSA approach proposed for the two-hop scenario to the multi-hop case. As each EH node has only local causal knowledge, data overflow situations in the next node cannot be fully avoided. As described in Section 4.2.4, each node aims at maximizing the amount of data it can transmit. To find the transmission policy, each node solves an independent point-to-point communication problem using the independent SARSA approach described in Section 4.2.4. Note that the cooperative SARSA algorithm can also be considered for the multi-hop scenario. However, in this case the required signaling increases with the number of hops. When amplify-and-forward relays are considered, the problem cannot be separated, as explained in Section 4.3.4. As a consequence, the proposed centralized SARSA can be exploited to find the power allocation policies that aim at maximizing the throughput. However, as in the case of cooperative SARSA, the amount of required signaling increases with the number of hops.

The proposed approaches can also be considered for an EH multi-node multi-hop communication scenario with multiple transmitter and receiver pairs. In contrast to the previous case, this scenario considers multiple transmitter and receiver pairs communicating using multiple intermediate relays. To apply the proposed learning approaches, the reward function given in (4.2) has to be modified according to the particular goal being considered. For instance, given the limited amount of energy in the relays, if the goal is to guarantee that each receiver is able to receive data from its corresponding transmitter at least one time, fairness has to be taken into account in the definition of the reward function. This can be done, for example, by considering a weighted throughput as the reward function where different weights are assigned to the different achieved throughputs, i.e., the amounts of data transmitted by each transmitter.

4.5. Conclusions

In this chapter, we have investigated the EH two-hop communication scenario considering two different types of relays, namely, a decode-and-forward relay and an amplify-and-forward relay. For each of them, offline and learning approaches are studied in order to find power allocation policies that maximize the throughput.

For the case when a decode-and-forward relay is considered, we have followed an offline approach in which perfect non-causal knowledge of the EH, data arrival and channel fading processes is assumed and have formulated the power allocation problem for throughput maximization. We have shown that the resulting problem is a convex optimization problem and have used the KKT conditions to characterize it. From the analysis of the KKT conditions, we have found that the optimal power to be used in each time interval i depends on the exact values of the Lagrange multipliers associated to the energy causality, battery overflow, data causality and data buffer overflow constraints of future time intervals j, $j > i$. Consequently, a closed-form solution of the power to be used in time interval i cannot be obtained. Furthermore, we have shown that the power allocation policies of the transmitter and the relay depend on each other and should be jointly considered in order to achieve optimum performance. Additionally, a more realistic scenario has been considered in which each EH node has only causal, and possibly outdated, knowledge about the EH, data arrival and channel fading processes associated to it. Following a learning approach, we have proposed two learning algorithms to find power allocation policies that aim at maximizing the throughput in this setting. These algorithms are motivated by the fact that, in the optimal power allocation policy, knowledge about the system dynamics, i.e., EH, data arrival and channel fading processes associated to both nodes, is required. As a result, the algorithms exploit different levels of cooperation among the nodes in order to learn the power allocation policy. Specifically, we have shown that when a signaling phase is introduced, in which the nodes cooperate with each other to exchange their current parameters, a higher performance can be achieved. Furthermore, we have provided convergence guarantees for the two proposed algorithms and by means of a computational complexity analysis, we have shown that the computational complexity of the proposed approaches increases only linearly with the number of possible transmit power values the EH nodes can select.

Similar to the previous case, offline and learning approaches have been investigated for the EH two-hop scenario considering an amplify-and-forward relay. Initially, assuming perfect non-causal knowledge of the system dynamics, the power allocation problem for throughput maximization has been formulated. We have shown that the resulting problem is non-convex. However, by exploiting basic properties of logarithms we have been able to reformulate it as a D.C. programming problem. Moreover, based on this reformulation, a branch-and-bound algorithm to find the optimal power allocation policy that maximizes the throughput has been proposed. Following a learning approach, we have proposed a centralized algorithm that takes into account the fact that the communication between the transmitter and the receiver cannot be separated, as in the decode-and-forward case, but should be considered as an effective channel which includes the link between the transmitter and the relay, the relay gain and the link between the relay and the receiver. The proposed centralized algorithm assumes that one of the EH nodes, either the transmitter or the relay, decides on the transmit power to be used by both EH nodes. For this purpose, a signaling phase is considered in which the node in charge of the learning task obtains the parameters, i.e., amount of harvested energy, battery level, data buffer level and channel gain, associated to the other node. Through numerical simulations, we have shown that the proposed centralized learning algorithm outperforms the reference approaches.

Additionally, in this chapter we have discussed how the proposed learning approaches can be extended to consider multi-hop relaying scenarios. Specifically, in an EH multi-hop scenario with a single transmitter and a single receiver and in an EH multi-node multi-hop scenario with multiple transmitter and receiver pairs.

Chapter 5

Energy harvesting broadcast scenario

5.1. Introduction

In this chapter, an EH broadcast scenario is considered and both, offline and learning approaches are investigated in order to find the power allocation policy at the transmitter that aims at maximizing the throughput.

The chapter is organized as follows. In Section 5.2, the EH broadcast scenario is introduced and the corresponding system assumptions are described. Next, in Section 5.3 the power allocation problem for throughput maximization is formulated. The resulting optimization problem is non-convex when more than two receivers are considered. Therefore, for the offline approach the special case of the two-user EH broadcast channel is studied. Specifically, in Section 5.4, we extend the offline approach proposed in [OYU13] in order to consider the energy consumed by the circuit and the individual data arrival processes. Next, in Section 5.5, we propose a learning algorithm, termed two-stage SARSA, to find the power allocation that aims at maximizing the throughput when only causal knowledge about the EH, data arrival and channel fading processes is available. The proposed learning approach is applicable to scenarios with an arbitrary number of receivers. Finally, in Section 5.6, the performance of the proposed two-stage SARSA algorithm is evaluated by means of numerical simulations.

Parts of this Chapter have been published by the author of this dissertation in [OWK18].

5.2. Scenario description and assumptions

An EH broadcast scenario consisting of a single-antenna EH transmitter and N single-antenna non-EH receivers is considered. A summary of all the parameters associated to this scenario, which are described in the following, is given in Table 5.1.

As depicted in Figure 5.1, the EH transmitter N_0 harvests energy from the environment and uses it to transmit data to the N non-EH receivers N_n, $n = 1, ..., N$. Specifically,

Table 5.1. Parameters associated to the EH broadcast communication scenario.

	Parameter	Description
General	i	Index of the time interval
	I	Total number of time intervals
	N	Number of non-EH receivers
	N_0	EH transmitter node
	N_n	n^{th} non-EH receiver, $n = 1, ..., N$
	τ	Time interval duration
Energy	$B_{0,i}$	Battery level of EH node N_0, measured at the beginning of time interval i
	$B_{\max,0}$	Battery capacity of EH node N_0
	$E_{0,i}$	Amount of harvested energy, received at the end of time interval i, by EH node N_0
	$E_{0,i}^{\mathrm{Circ}}$	Amount of energy consumed by the circuit of EH node N_0 in time interval i
	$E_{0,i}^{\mathrm{Tx}}$	Energy of the signal transmitted by EH node N_0 in time interval i
	$E_{\max,0}$	Maximum amount of energy that can be harvested by EH node N_0
	$p_{0,i}^{\mathrm{Tx}}$	Total transmit power used by EH node N_0 in time interval i
	$p_{n,i}^{\mathrm{Tx}}$	Transmit power assigned for the transmission of the signal intended for N_n
Data	$D_{\max,0}$	Data buffer size of EH node N_0
	$D_{\max,n}$	Size of the virtual data buffer containing the data intended for non-EH node N_n
	$D_{n,i}$	Level of the virtual data buffer containing the data intended for non-EH node N_n, measured at the beginning of time interval i
	$M_{n,i}$	Amount of incoming data intended for non-EH node N_n, arriving at the end of time interval i, at EH node N_0
	R_i^{BC}	Total amount of data transmitted in time interval i
	$R_{n,i}^{\mathrm{BC}}$	Amount of data transmitted to non-EH node N_n in time interval i
Channel	$g_{n,i}$	Channel gain of the link between N_0 and N_n
	$h_{n,i}$	Channel coefficient of the link between N_0 and N_n
	W	Bandwidth
	σ_n^2	Noise power at N_n

an amount $E_{0,i}$ of harvested energy is received at the end of each time interval i, $i = 1, ..., I$, and it is stored in a battery with finite capacity $B_{\max,0}$. The battery level $B_{0,i}$ is measured at the beginning of time interval i and indicates the amount of available energy. The maximum amount of harvested energy is denoted by $E_{\max,0}$ and the energy $E_{0,i}^{\mathrm{Circ}}$ consumed by the circuit at N_0 is assumed to be constant for all the time intervals, such that $E_{0,i}^{\mathrm{Circ}} = E_0^{\mathrm{Circ}}$, $\forall i$. In every time interval i, N_0 decides on the transmit power $p_{0,i}^{\mathrm{Tx}}$ to use for the duration τ of the time interval. As a result, an amount $E_{0,i}^{\mathrm{Tx}} = \tau p_{0,i}^{\mathrm{Tx}}$ of energy is used for the transmission of data to the receivers. Considering (2.3), the battery level is updated as

$$B_{0,i+1} = \min\left\{B_{\max,0}, B_{0,i} - E_{0,i}^{\mathrm{Tx}} + E_{0,i} - E_0^{\mathrm{Circ}}\right\}. \tag{5.1}$$

The non-EH receiver nodes N_n are assumed to be connected to a fixed power supply. Therefore, they have always enough energy to receive the transmitted data from N_0.

The data intended for each N_n is different and depends on a receiver-specific data arrival process. In our model, we consider a data buffer at N_0 with size $D_{\max,0}$ and divide it into N equal-size virtual data buffers as shown in Figure 5.1. The size of each virtual data buffer is $D_{\max,n} = D_{\max,0}/N$, measured in bits. As described in Section

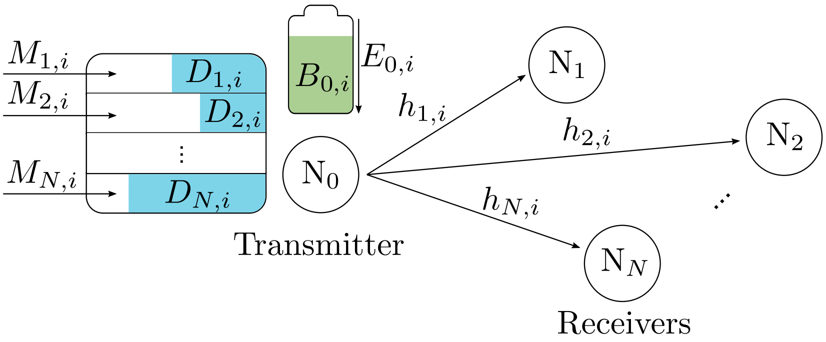


Figure 5.1. Data dissemination scenario with an EH transmitter.

2.2.3, two cases regarding the data arrival model are considered. In the first case, an infinitely full data buffer is assumed. This means that the data buffer sizes $D_{\max,0}$ and $D_{\max,n}$, as well as the buffer levels $D_{n,i}$ are infinite for $n = 1, ..., N$ and $i = 1, ..., I$. In the second case, an amount $M_{n,i}$ of incoming data intended for N_n arrives at N_0 at the end of every time interval i and it is stored in the corresponding virtual data buffer with size $D_{\max,n}$. The data buffer level $D_{n,i}$ is measured at the beginning of time interval i and indicates the amount of data available for transmission to N_n. Furthermore, the throughput $R_{n,i}^{\mathrm{BC}}$ is the amount of data transmitted to N_n in time interval i. Considering (2.6), the data buffer level of each virtual data buffer is updated in each time interval as

$$D_{n,i+1} = \min\left\{D_{\max,n},\, D_{n,i} - R_{n,i}^{\mathrm{BC}} + M_{n,i}\right\}. \tag{5.2}$$

Taking into account the data intended for each receiver, the communication between N_0 and the N receivers in the considered broadcast scenario is as follows. N_0 uses superposition coding [CT06, TV05] for the encoding of the data intended for each N_n in a signal x_i using an i.i.d. Gaussian code spread on the entire bandwidth W. In each time interval i, a transmit power $p_{0,i}^{\mathrm{Tx}}$ is used by N_0 for the transmission of x_i. Moreover, let $p_{n,i}^{\mathrm{Tx}}$ be a fraction of the power $p_{0,i}^{\mathrm{Tx}}$ such that

$$\sum_{n=1}^{N} p_{n,i}^{\mathrm{Tx}} = p_{0,i}^{\mathrm{Tx}} \tag{5.3}$$

holds. Considering the channel coefficients $h_{n,i}$ for the link between the transmitter N_0 and the receiver N_n, the received signal $y_{n,i}$ at node N_n in time interval i is given by

$$y_{n,i} = h_{n,i} x_i + w_{n,i}, \tag{5.4}$$

where $w_{n,i}$ is the receiver noise at N_n in time interval i with noise power σ_n^2. When there is enough data available in the data buffer, the throughput $R_{n,i}^{\mathrm{BC}}$ is approximated using Shannon's capacity formula as

$$R_{n,i}^{\mathrm{BC}} = \tau W \log_2 \left(1 + \frac{g_{n,i} p_{n,i}^{\mathrm{Tx}}}{\sum\limits_{m \neq n; m=1}^{N} g_{m,i} p_{m,i}^{\mathrm{Tx}} + \sigma_n^2} \right) \tag{5.5}$$

measured in bits. Otherwise, the throughput is limited by the corresponding data buffer level. Note that in (5.5), in the interference term, $p_{m,i}^{\mathrm{Tx}} = 0$ if N_m is not served during time interval i. Moreover, the total throughput achieved in time interval i is denoted by R_i^{BC} and it is calculated as

$$R_i^{\mathrm{BC}} = \sum_{n=1}^{N} R_{n,i}^{\mathrm{BC}}. \tag{5.6}$$

As in the previous chapters, transmitter side channel state information is assumed to be non-causally known at N_0 in offline approaches and causally known at N_0 when learning approaches are considered.

5.3. Problem formulation

In this section, the power allocation problem for throughput maximization in the EH broadcast scenario is formulated. Our goal is to find a transmission policy at the transmitter N_0 that aims at maximizing the total throughput in (5.6), i.e., the total amount of data transmitted to the non-EH nodes N_n, $n = 1, ..., N$, considering the energy causality, battery overflow, data causality and data buffer overflow constraints defined in (2.4), (2.5), (2.7) and (2.8), respectively. For this purpose, the transmit powers $p_{n,i}^{\mathrm{Tx}}$ to use in each time interval i for the transmission of the individual data need to be determined.

Taking into account that the receivers might be served with different preferences, let ϕ_n be a weighting factor proportional to the priority associated to the receiver node N_n with $\sum_{n=1}^{N} \phi_n = 1$. These priorities are assumed to be fixed for all the time intervals. Therefore, for given weights ϕ_n, the power allocation problem for throughput maximization in the EH broadcast scenario is given by

$$\left(p_{n,i}^{\mathrm{Tx}^{\mathrm{opt}}} \right)_{n,i} = \underset{\{p_{n,i}^{\mathrm{Tx}}, n=\{1,...,N\}, i=\{1,...,I\}\}}{\operatorname{argmax}} \sum_{i=1}^{I} \sum_{n=1}^{N} \phi_n R_{n,i}^{\mathrm{BC}} \tag{5.7a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \tau p_{0,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_0^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{0,i}, \; J = 1, ..., I, \tag{5.7b}$$

$$\sum_{i=1}^{J} E_{0,i} - \sum_{i=1}^{J} \tau p_{0,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_0^{\mathrm{Circ}} \leq B_{\max,0},\ J = 1, ..., I, \tag{5.7c}$$

$$\sum_{i=1}^{J} R_{n,i}^{\mathrm{BC}} \leq \sum_{i=1}^{J-1} M_{n,i},\ n = 1, .., N,\ J = 1, ..., I, \tag{5.7d}$$

$$\sum_{i=1}^{J} M_{n,i} - \sum_{i=1}^{J} R_{n,i}^{\mathrm{BC}} \leq D_{\max,n},\ n = 1, .., N,\ J = 1, ..., I, \tag{5.7e}$$

$$\sum_{n=1}^{N} p_{n,i}^{\mathrm{Tx}} = p_{0,i}^{\mathrm{Tx}},\ i = 1, ..., I, \tag{5.7f}$$

$$p_{m,i}^{\mathrm{Tx}} \geq 0,\ m = 0, ..., N,\ i = 1, ..., I, \tag{5.7g}$$

where $R_{n,i}^{\mathrm{BC}}$ is given in (5.5), (5.7b) corresponds to the energy causality constraint derived from (2.4), (5.7c) is the battery overflow constraint derived from (2.5), (5.7d) is obtained considering the data causality constraint in (2.7), and (5.7e) takes into account the data buffer overflow constraint in (2.8). Note that when an infinitely full data buffer is considered, the constraints (5.7d) and (5.7e) are not taken into account.

5.4. Offline approach

The offline approach presented in the following is based on the work of [OYU13]. We have extended it such that the individual data arrival processes and the energy E_0^{Circ}, consumed by the circuit at N_0, are considered.

The optimization problem in (5.7) is non-convex with respect to the optimization variables $p_{n,i}^{\mathrm{Tx}}$ because the objective function (5.7a) is a non-convex function of $p_{n,i}^{\mathrm{Tx}}$. In this section, we consider the special case when $N = 2$ receivers are considered. For this case, the objective function can be written as a function of the total power $p_{0,i}^{\mathrm{Tx}}$ used in time interval i and a power sharing parameter η_i,

$$0 \leq \eta_i \leq 1, \tag{5.8}$$

such that $p_{1,i}^{\mathrm{Tx}} = \eta_i p_{0,i}^{\mathrm{Tx}}$ and $p_{2,i}^{\mathrm{Tx}} = (1 - \eta_i) p_{0,i}^{\mathrm{Tx}}$. Such formulation allows us to overcome the non-convexity of (5.7) by optimizing η_i and p_0^{Tx} separately.

In the offline approach, perfect non-causal knowledge regarding the EH, data arrival and channel fading processes is assumed. Consequently, as the receivers N_n know the channel conditions of the other nodes, they can reduce the interference by successively decoding the signals intended for the receivers with degraded channel conditions and subtracting them from the received signal, before decoding their own. This technique, known as successive interference cancellation (SIC) [CT06, TV05], leads to the optimal throughput.

Let $\tilde{g}_{n,i} = \sigma_n^2/g_{n,i}$, for $n = \{1, 2\}$. Then, the amount of data received at N_1 and N_2 in time interval i can be calculated as

$$R_{1,i}^{\mathrm{BC}} = \tau W \log_2\left(1 + \frac{\eta_1 p_{0,i}^{\mathrm{Tx}}}{(1-\eta_i)p_{0,i}^{\mathrm{Tx}}\mathbf{1}(\tilde{g}_{1,i} > \tilde{g}_{2,i}) + \tilde{g}_{1,i}}\right) \tag{5.9}$$

and

$$R_{2,i}^{\mathrm{BC}} = \tau W \log_2\left(1 + \frac{(1-\eta_i)p_{0,i}^{\mathrm{Tx}}}{\eta_i p_{0,i}^{\mathrm{Tx}}\mathbf{1}(\tilde{g}_{1,i>}\tilde{g}_{2,i}) + \tilde{g}_{2,i}}\right), \tag{5.10}$$

respectively, where $\mathbf{1}(x)$ is the indicator function that takes the value of one when the condition x is true and is zero otherwise.

Considering (5.9) and (5.10), the optimization problem in (5.7) for $N = 2$ can be reformulated as

$$\left(p_{0,i}^{\mathrm{Tx}^{\mathrm{opt}}}, \eta_i^{\mathrm{opt}}\right)_i = \underset{\{p_{0,i}^{\mathrm{Tx}}, \eta_i,\, i=\{1,\dots,I\}\}}{\operatorname{argmax}} \left(\sum_{i=1}^{I} \phi_1 R_{1,i}^{\mathrm{BC}} + \phi_2 R_{2,i}^{\mathrm{BC}}\right) \tag{5.11a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \tau p_{0,i}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_0^{\mathrm{Circ}} \leq \sum_{i=1}^{J-1} E_{0,i},\ J = 1, \dots, I, \tag{5.11b}$$

$$\sum_{i=1}^{J} E_{0,i} - \sum_{i=1}^{J} \tau p_{0,i}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_0^{\mathrm{Circ}} \leq B_{\max,0},\ J = 1, \dots, I, \tag{5.11c}$$

$$\sum_{i=1}^{J} R_{n,i}^{\mathrm{BC}} \leq \sum_{i=1}^{J-1} M_{n,i},\ n = 1, 2,\ J = 1, \dots, I, \tag{5.11d}$$

$$\sum_{i=1}^{J} M_{n,i} - \sum_{i=1}^{J} R_{n,i}^{\mathrm{BC}} \leq D_{\max,0},\ n = 1, 2,\ J = 1, \dots, I, \tag{5.11e}$$

$$p_{0,i}^{\mathrm{Tx}} \geq 0,\ i = 1, \dots, I, \tag{5.11f}$$

$$0 \leq \eta_i \leq 1,\ i = 1, \dots, I. \tag{5.11g}$$

As the objective function in (5.11a) contains the product of the two optimization variables $p_{0,i}^{\mathrm{Tx}}$ and η_i, the problem in (5.11) is still a non-convex optimization problem. However, when only one optimization variable is considered, the convexity of the problem can be ensured because (5.11a) is concave with respect to the considered variable. As described in [OYU13], the optimal power allocation policy is found by taking advantage of this property, i.e., first the power sharing parameter η_i is optimized and then, the result is used to optimize $p_{0,i}^{\mathrm{Tx}}$.

The optimal power sharing parameter η_i^{opt} is found by solving

$$\eta_i^{\text{opt}} = \underset{\{0 \leq \eta_i \leq 1\}}{\operatorname{argmax}} \left(\phi_1 R_{1,i}^{\text{BC}} + \phi_2 R_{2,i}^{\text{BC}}, \right) \tag{5.12}$$

for given weights ϕ_1, ϕ_2.

Let $\phi = \frac{\phi_2}{\phi_1}$ and assume $\tilde{g}_{1,i} < \tilde{g}_{2,i}$. Then, for $1 \leq \phi \leq \frac{\tilde{g}_{2,i}}{\tilde{g}_{1,i}}$, the optimal η_i^{opt} is given by

$$\eta_i^{\text{opt}} = \begin{cases} 1, & 0 \leq p_{0,i}^{\text{Tx}} \leq \frac{\phi \tilde{g}_{1,i} - \tilde{g}_{2,i}}{1-\phi} \\ \frac{1}{p_{0,i}^{\text{Tx}}} \frac{\phi \tilde{g}_{1,i} - \tilde{g}_{2,i}}{1-\phi}, & p_{0,i}^{\text{Tx}} > \frac{\phi \tilde{g}_{1,i} - \tilde{g}_{2,i}}{1-\phi} \end{cases} \tag{5.13}$$

[OYU13]. Note that $\eta_i^{\text{opt}} = 0$ when $\phi \geq \frac{\tilde{g}_{2,i}}{\tilde{g}_{1,i}}$ and $\eta_i^{\text{opt}} = 1$ when $\phi \leq 1$.

In every time interval i, η_i is fully determined by the channel gains $g_{n,i}$, the noise powers σ_n^2 and the weighting factors ϕ_n for the priorities, as defined in (6.4). Consequently, to find the power allocation policy, it remains to solve (5.11) for $p_{0,1}^{\text{Tx}}$. In [LG01] and [OYU13], it is shown that the objective function (5.11a) is a monotically increasing concave function of $p_{0,i}^{\text{Tx}}$. As a result, for a given set of power sharing parameters η_i, the optimization problem in (5.11) has a unique solution. Furthermore, the optimal power allocation policy is found by considering the Directional Backward Glue Pouring algorithm of [OGE12] described in Section 3.4.

5.5. Learning approach

5.5.1. A two-stage approach

In this section, a learning approach is proposed in order to find the power allocation policy that aims at maximizing the throughput in the EH broadcast scenario when an arbitrary number of receivers is considered. The proposed approach is motivated by the offline approach of the previous section. Specifically, we propose a learning algorithm, termed two-stage SARSA, which divides the learning task into two subtasks, namely, how much power to allocate in each time interval i and how to split the allocated power among the data to be transmitted to the different receivers. By dividing the task, smaller RL problems need to be addressed in each stage, which in turn facilitates the identification of power allocation solutions that lead to a higher throughput. As a result, the proposed two-stage SARSA achieves a higher performance compared to standard RL algorithms like SARSA. The section is organized as follows. First, an MDP formulation of the power allocation problem is presented in Section 5.5.2. Next, in Section 5.5.3, the description of the proposed two-stage SARSA algorithm is presented.

5.5.2. Markov decision process

As explained in Section 2.3, the power allocation problem in an EH transmitter can be modeled as an MDP. MDPs are defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathsf{P}, \mathcal{R} \rangle$. For the EH broadcast scenario, the state $S_i \in \mathcal{S}$ is a function of the transmitter's parameters $E_{0,i}$, $B_{0,i}$, $D_{n,i}$ and $g_{n,i}$, for all $n = 1, ..., N$. As all of these parameters can take values in a continuous range, the set $\mathcal{S}$ contains infinitely many possible states. The action set $\mathcal{A}$ contains the transmit power tuples $a_i = \langle p_{1,i}^{\text{Tx}}, ..., p_{N,i}^{\text{Tx}} \rangle$ that can be selected. As in practical scenarios, we consider that only a finite set of transmit power values can be selected [Ins17]. Therefore, in our model $\mathcal{A}$ is finite and it is defined as $\mathcal{A} = \{a_i = \langle p_{n,i}^{\text{Tx}} \rangle_{n=1,...,N} | p_{n,i}^{\text{Tx}} \in \{0, \delta, 2\delta, ..., B_{\max,0}\}\}$, where δ is the step size. P is the transition model which defines the probability of going from S_i to S_{i+i} after performing a_i. However, as only causal knowledge regarding the system dynamics is assumed to be available, this transition model P is unknown. Finally, the rewards $R_i^{\text{BC}} \in \mathcal{R}$ indicate how beneficial it is to select a_i when the transmitter is in state S_i.

As a consequence of having only causal knowledge of the system dynamics, N_0 does not know in advance for how many time intervals it will remain operative. Similar to [BGD13], we consider a discount factor γ, $0 \leq \gamma \leq 1$ to account for the preference of achieving a higher throughput in the current time interval versus achieving a higher throughput later on. For this purpose, we aim at maximizing the amount of transmitted data given by

$$R^{\text{BC}} = \lim_{I \to \infty} \mathbb{E} \left[\sum_{i=1}^{I} \sum_{n=1}^{N} \gamma^{i-1} R_{n,i}^{\text{BC}} \right]. \tag{5.14}$$

The solution of the MDP is given by the policy π which maps states to actions, i.e., $a_i = \pi(S_i)$ [SB18]. As in the previous chapters, the action-value function $\text{Q}^{\pi}(S_i, a_i)$, described in detail in Section 2.3.3, is used to evaluate the suitability of a policy π for the solution of the power allocation problem.

5.5.3. Two-stage SARSA

5.5.3.1. Upper and lower stages

From the MDP formulation of the EH broadcast scenario, we observe that each action $a_i \in \mathcal{A}$ corresponds to a possible combination of the power values that can be assigned for the transmission of the data intended for the different receivers. As a consequence,

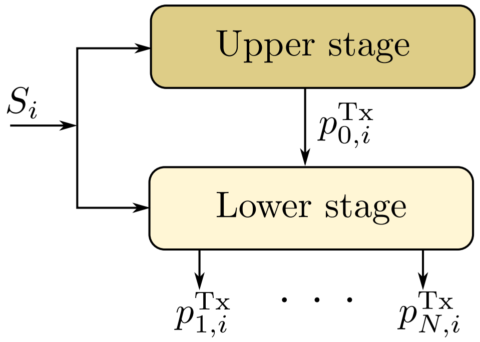


Figure 5.2. Schematic of the two-stage approach.

the size A of the action set $\mathcal{A}$ grows exponentially with the number of receivers N, as $A = |\mathcal{A}| = |\{0, \delta, 2\delta, ..., B_{\max}\}|^N$. Such a large action set reduces the learning speed and hence the performance since more actions need to be tried to find the optimal policy. Therefore, to overcome this challenge, the proposed two-stage SARSA algorithm separates the learning task into two stages: an upper stage which decides on the total power to be used and a lower stage that decides how to distribute it. This separation is motivated by the offline approach described in Section 5.4 which uses a similar strategy.

As depicted in Figure 5.2, considering the state S_i, the upper stage decides on the total transmit power $p_{0,i}^{\mathrm{Tx}}$ to allocate in each time interval i such that battery overflow is avoided. This selected power is then fed into the lower stage which decides on how to distribute it, i.e., it selects $p_{n,i}^{\mathrm{Tx}}$ for $n = 1, ..., N$. Similar to the previous chapters, we use SARSA with linear function approximation in each of the stages and define feature functions tailored to the task to be solved. Moreover, for the linear function approximation, independent weights are considered in each stage, i.e., $\mathbf{w}^{\mathrm{up}}$ is used to approximate the action-value function in the upper stage while the weight vector $\mathbf{w}^{\mathrm{low}}$ is considered in the approximation of the action-value function in the lower stage. In the following, the proposed two stages of our two-stage SARSA algorithm are described.

Upper stage In a fading downlink channel, capacity can be achieved if the power is allocated to the transmission to the receiver with the best channel [TV05]. Consequently, in the selection of $p_{0,i}^{\mathrm{Tx}}$, we reduce the broadcast scenario to a point-to-point scenario in which only the receiver with the best channel conditions in time interval i is considered. We denote this best channel as $g_i^* = \max\{g_{1,i}, ..., g_{N,i}\}$. Note that this does not mean that only the receiver with the best channel will be served, but rather that the channel g_i^* is used as a reference since it provides an upper bound of the maximum possible performance that can be achieved. Since in this stage only one receiver is considered, the action set $\mathcal{A}^{\mathrm{up}}$ is defined as $\mathcal{A}^{\mathrm{up}} = \{p_{0,i}^{\mathrm{Tx}} | p_{0,i}^{\mathrm{Tx}} \in \{0, \delta, 2\delta, ..., B_{\max,0}\}\}$.

Furthermore, the reward obtained by selecting $p_{0,i}^{\mathrm{Tx}}$ is defined as

$$R_i^{\mathrm{up}} = \tau W \log_2 \left(1 + \frac{p_{0,i}^{\mathrm{Tx}} g_i^*}{\sigma^2} \right). \tag{5.15}$$

Taking into account that this stage solves an EH point-to-point communication problem, for the linear function approximation we use the feature functions defined in Section 3.5.2.4 which consider the energy causality, the battery overflow, the data causality and the data buffer overflow constraints.

Lower stage Given the selected $p_{0,i}^{\mathrm{Tx}}$, the task of the lower stage is to distribute the power among the individual data to be transmitted to the different receivers while aiming at minimizing data buffer overflows and maximizing the throughput. In contrast to the upper stage, in which the action set $\mathcal{A}^{\mathrm{up}}$ is discrete, the number of actions in the lower stage is infinite. This is due to the fact that the selected transmit power $p_{0,i}^{\mathrm{Tx}}$ can be split in infinitely many ways while fulfilling the condition in (5.3). To overcome this challenge and have an action set $\mathcal{A}^{\mathrm{low}}$ which is independent of the selected $p_{0,i}^{\mathrm{Tx}}$, let $0 \leq \eta_{n,i} \leq 1$, with $\sum_{n=1}^{N} \eta_{n,i} = 1$, indicate what fraction of $p_{0,i}^{\mathrm{Tx}}$ is assigned to the transmission of data intended for N_n, i.e., $p_{n,i}^{\mathrm{Tx}} = \eta_{n,i} p_{0,i}^{\mathrm{Tx}}$. Then, by further constraining η_i to take values from a finite set, the number of combinations, and thus the size of the action set $\mathcal{A}^{\mathrm{low}}$ of the lower stage, becomes finite. In particular, every action $\mathbf{a}_i^{\mathrm{low}} \in \mathcal{A}^{\mathrm{low}}$ is a vector of size N containing the values $\eta_{n,i}$ selected for each user, i.e., $\mathbf{a}_i^{\mathrm{low}} = (\eta_{1,i}, ..., \eta_{N,i})$, where each $\eta_{n,i}$ is taken from the set $\mathcal{J} = \{0, \delta, ..., 1\}$ and δ is a step size. The reward $R_{n,i}^{\mathrm{low}}$ considered in the lower stage corresponds to the achieved throughput given the selected action, and it is calculated as

$$R_i^{\mathrm{low}} = \sum_{n=1}^{N} R_{n,i}^{\mathrm{BC}}, \tag{5.16}$$

where $R_{n,i}^{\mathrm{BC}}$ is given by (5.5).

As in the upper stage, linear function approximation is considered to handle the infinite number of states. For this purpose, we propose three feature functions based on three different transmission strategies, namely, water-filling, maximum rate and proportional fairness. The first feature function $\mathrm{f}_1^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}})$ distributes the power $p_{0,i}^{\mathrm{Tx}}$ using the water-filling algorithm considering only the links to the receivers whose corresponding virtual data buffer levels fulfill the condition $D_{n,i} > 0$. Let the vector $\mathbf{a}_i^{\mathrm{WF}}$ contain the transmit powers $p_{n,i}^{\mathrm{Tx}}$ obtained with water-filling, then

$$\mathrm{f}_1^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}}) = \begin{cases} 1, & \text{if } p_{0,i}^{\mathrm{Tx}} \mathbf{a}_i^{\mathrm{low}} = \mathbf{a}_i^{\mathrm{WF}} \\ 0, & \text{else.} \end{cases} \tag{5.17}$$

The second feature function $\mathrm{f}_2^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}})$ is based on the maximum rate approach, in which $p_{0,i}^{\mathrm{Tx}}$ is used for the transmission to the receiver with the strongest channel. Let m be the index of the receiver with the strongest channel in time interval i, then

$$\mathrm{f}_2^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}}) = \begin{cases} 1, & \text{if } \mathbf{a}_i^{\mathrm{low}} \in \mathcal{A}^{\mathrm{low}} \cap \{\mathbf{a}_i^{\mathrm{low}} | \eta_{m,i} = 1\} \\ 0, & \text{else.} \end{cases} \tag{5.18}$$

The third feature function $\mathrm{f}_3^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}})$ is based on the proportional fairness scheduler in [BK13]. In this case, let $R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})}$ be the throughput that would be achieved if the power $p_{n,i}^{\mathrm{Tx}} = \eta_{n,i} p_{0,i}^{\mathrm{Tx}}$ is allocated for the transmission to receiver N_n. Then, in time interval i, $\mathrm{f}_3^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}})$ allocates the total power $p_{0,i}^{\mathrm{Tx}}$ for the transmission to the receiver N_m that satisfies

$$m = \underset{n=1,\ldots,N}{\mathrm{argmax}} \frac{\min\left\{R_{n,i}^{(p_{n,i}^{\mathrm{Tx}})}, D_{n,i}\right\}}{\frac{1}{i}\sum\limits_{j=1}^{i} R_{n,j}^{\mathrm{BC}}}. \tag{5.19}$$

By considering (5.19), we define

$$\mathrm{f}_3^{\mathrm{low}}(S_i, \mathbf{a}_i^{\mathrm{low}}) = \begin{cases} 1, & \text{if } \mathbf{a}_i^{\mathrm{low}} \in \mathcal{A}^{\mathrm{low}} \cap \{\mathbf{a}_i^{\mathrm{low}} | \eta_{m,i} = 1\} \\ 0, & \text{else.} \end{cases} \tag{5.20}$$

5.5.3.2. Action-value functions update

In both stages, upper and lower, we use SARSA to estimate the corresponding action-value functions Q^{π}. To differentiate the two, we denote as Q^{up} and $\mathrm{Q}^{\mathrm{low}}$ the action-value function of the upper and lower stage, respectively. Furthermore, in both stages we combine SARSA with linear function approximation to handle the infinite number of states. As described in Section 3.5.2.3, when SARSA with linear function approximation is used, the estimation of the action-value function is done by updating the vector of weights which contains the contribution of each feature function. This update is done based on the states that are encountered, the actions that are selected and the obtained rewards. Specifically, for the upper stage, the vector $\mathbf{w}^{\mathrm{up}}$ of weights is considered and the action-value function $\mathrm{Q}^{\mathrm{up}}(S_i, p_{0,i}^{\mathrm{Tx}})$ is approximated as

$$\mathrm{Q}^{\mathrm{up}}(S_i, p_{0,i}^{\mathrm{Tx}}) \approx \hat{\mathrm{Q}}^{\mathrm{up}}(S_i, p_{0,i}^{\mathrm{Tx}}, \mathbf{w}^{\mathrm{up}}) = (\mathbf{f}^{\mathrm{up}})^{\mathrm{T}} \mathbf{w}^{\mathrm{up}}, \tag{5.21}$$

where $\mathbf{f}^{\mathrm{up}}$ is the vector containing the values of the feature functions in the upper stage for state S_i. Furthermore, the weights $\mathbf{w}^{\mathrm{up}}$ are updated in the direction that reduces the error between Q^{up} and $\hat{\mathrm{Q}}^{\mathrm{up}}$ following the gradient descent approach as

$$\mathbf{w}_{i+1}^{\mathrm{up}} = \mathbf{w}_i^{\mathrm{up}} + \zeta_i \big[R_i^{\mathrm{up}} + \gamma \hat{\mathrm{Q}}^{\mathrm{up}}(S_{i+1}, p_{0,i+1}^{\mathrm{Tx}}, \mathbf{w}^{\mathrm{up}}) - \hat{\mathrm{Q}}^{\mathrm{up}}(S_i, p_{0,i}^{\mathrm{Tx}}, \mathbf{w}^{\mathrm{up}})\big] \mathbf{f}^{\mathrm{up}}, \tag{5.22}$$

where ζ_i is the learning rate.

Similarly, for the lower stage, the vector $\mathbf{w}^{\text{low}}$ of weights and the vector $\mathbf{f}^{\text{low}}$, containing the values of the feature functions in the lower stage for state S_i, are considered and the action-value function $\text{Q}^{\text{low}}(S_i, \mathbf{a}_i^{\text{low}})$ is approximated as

$$\text{Q}^{\text{low}}(S_i, \mathbf{a}_i^{\text{low}}) \approx \hat{\text{Q}}^{\text{low}}(S_i, \mathbf{a}_i^{\text{low}}, \mathbf{w}^{\text{low}}) = \left(\mathbf{f}^{\text{low}}\right)^{\text{T}} \mathbf{w}^{\text{low}}. \tag{5.23}$$

Moreover, the weights $\mathbf{w}^{\text{low}}$ are updated as

$$\mathbf{w}_{i+1}^{\text{low}} = \mathbf{w}_i^{\text{low}} + \zeta_i \big[R_i^{\text{low}} + \gamma \hat{\text{Q}}^{\text{low}}(S_{i+1}, \mathbf{a}_{i+1}^{\text{low}}, \mathbf{w}^{\text{up}}) - \hat{\text{Q}}^{\text{low}}(S_i, \mathbf{a}_i^{\text{low}}, \mathbf{w}^{\text{low}})\big] \mathbf{f}^{\text{low}}. \tag{5.24}$$

5.5.3.3. Two-stage SARSA algorithm

A summary of the proposed two-stage SARSA algorithm is presented in Algorithm 5.1. As in the previous approaches, the algorithm starts by initializing the learning parameters ζ and ϵ as well as the weight vectors $\mathbf{w}^{\text{up}}$ and $\mathbf{w}^{\text{low}}$ (line 1). After observing the initial state (line 2), the total transmit power $p_{0,i}^{\text{Tx}}$ is randomly selected in the upper stage (line 3). Next, using the selected $p_{0,i}^{\text{Tx}}$, the lower stage randomly selects the fractions $\eta_{n,i}$ for all the nodes $n = 1, ..., N$ and determines the power values $p_{n,i}^{\text{Tx}}$ (lines 4-5). Afterwards, for every time interval i in which the transmitter is operative, the achieved rewards in each stage are calculated using (5.15) and (5.16) and the resulting new state is observed (lines 7-9). The next transmit power values $p_{0,i+1}^{\text{Tx}}$ and $p_{n,i+1}^{\text{Tx}}$ are selected in the upper and lower stage, respectively, using the ϵ-greedy policy (lines 10-12) and the weight vectors $\mathbf{w}_{\text{up}}$ and $\mathbf{w}_{\text{low}}$ are updated using (5.22) and (5.24) (line 13). The procedure described above is then repeated in all time intervals, until the transmitter is no longer operative.

5.5.3.4. Convergence guarantees

Since the learning process in each stage is independent of the other, the convergence of the proposed two-stage SARSA algorithm can be evaluated separately for each stage. As described in Section 5.5.3.1, the upper stage only considers, in each time interval, the receiver with the best channel condition in order to select the total power to use. Consequently, the problem reduces to a point-to-point scenario for which the convergence into a bounded region is guaranteed if the learning rate parameter ζ_i fulfills the conditions in (3.34) and (3.35), and the policy π is used throughout the execution of the algorithm. Furthermore, these conditions must be also satisfied in

Algorithm 5.1 Two-stage SARSA algorithm

```
initialize $\zeta, \epsilon$, $\mathbf{w}^{\text{up}}$ and $\mathbf{w}^{\text{low}}$
observe state $S_i$
randomly select $p_{0,i}^{\text{Tx}}$ in the upper stage
randomly select $\eta_{n,i}$ for every node in the lower stage
calculate $p_{n,i}^{\text{Tx}} = \eta_{n,i} p_{0,i}^{\text{Tx}}$, $n = 1, ..., N$
for every time interval $i = 1, ..., I$ do
    transmit using the selected $p_{n,i}^{\text{Tx}}$
    calculate the reward for both stages                                  ▷ Eq. (5.15), (5.16)
    observe state $S_{i+1}$
    in the upper stage, select next $p_{0,i+1}^{\text{Tx}}$ using the $\epsilon$-greedy policy
    in the lower stage, select next $\eta_{n,i+1}$ for every node using the $\epsilon$-greedy policy
    calculate $p_{n,i}^{\text{Tx}} = \eta_{n,i} p_{0,i}^{\text{Tx}}$, $\forall n$
    update $\mathbf{w}^{\text{up}}$ and $\mathbf{w}^{\text{low}}$                                          ▷ Eq. (5.22), (5.24)
    set $S_i = S_{i+1}$, $p_{0,i}^{\text{Tx}} = p_{0,i+1}^{\text{Tx}}$, and $p_{n,i}^{\text{Tx}} = p_{n,i+1}^{\text{Tx}}$, $n = 1, ..., N$
end for
```

the lower stage to guarantee the converge of the learning process. This is because in the lower stage, linear function approximation is also considered to handle the infinite number of states. In our implementation, both stages consider the ϵ-greedy policy and the learning rate is set to $\zeta_i = 1/i$. As a result, both stages converge to a bounded region.

5.5.3.5. Computational complexity analysis

In this section, we evaluate the complexity of the proposed two-stage SARSA algorithm with respect to the size $A^{\text{up}} = |\mathcal{A}^{\text{up}}|$ and $A^{\text{low}} = |\mathcal{A}^{\text{low}}|$ of the action sets of the upper and lower stages, respectively. From Algorithm 5.1, it is clear that the most computationally demanding tasks correspond to the selection of the action using the ϵ-greedy policy. This is because the ϵ-greedy policy involves finding, for a given state, the action that leads to the maximum value of the estimated action-value function. Specifically, the complexity of the upper stage grows as $O(A^{\text{up}})$, while the complexity of the lower stage grows as $O(A^{\text{low}})$. Note that from the definition of the action set in the upper stage, in Section 5.5.3.1, the size A^{up} does not depend on the number N of receivers, but only on the step size parameter δ that is considered. On the contrary, the size A^{low} grows exponentially with N. This is due to the fact that the set $\mathcal{A}^{\text{low}}$ is formed by all the possible permutations of the $\eta_{n,i}$ values that can be selected for each user. As in general $A^{\text{low}} \gg A^{\text{up}}$, the leading order of the complexity of the proposed two-stage SARSA grows linearly with the size A^{low} of the action space in the lower stage as $O(A^{\text{low}})$.

Table 5.2. Simulation set-up.

	Parameter	Value	Description
General	I	2000	Number of time intervals
	N	3	Number of receivers
	T	2000	Number of realizations
	τ	10ms	Time interval duration
	ϕ_n	$1/N$	Weighting factor proportional to the priority of N_n
Energy	$B_{\max,0}$	$10E_{\max,0}$	Battery capacity
	E^{Circ}	1mJ	Energy consumed by the circuit
	ρ	50mW/cm^2	Power density EH source
	Ω	16cm^2	Size of EH panel
Data	d	1 kbit	Packet size (finite data buffer case)
	$D_{\max,0}$	∞	Data buffer size of N_0 (infinitely full data buffer case)
	$D_{\max,0}$	$W\tau\log_2(1+\beta\tilde{\Gamma})$	Data buffer size of N_0 (finite data buffer case)
	β	1	Data buffer size factor (finite data buffer case)
	λ	10	Average number of packets arriving in time interval i (finite data buffer case)
Channel	f_0	2.4 GHz	Carrier frequency
	$r_{\max}$	50m	Coverage radius
	W	1 MHz	Bandwidth
	α	3	Path loss exponent
Learning	$\mathcal{J}$	$\{0, 0.2, 0.4, 0.6, 0.8, 1\}$	Set of possible values for $\eta_{n,i}$
	γ	0.9	Discount factor
	δ	2%	Step size
	ϵ	$1/i$	Exploration probability
	ζ	$1/i$	Learning rate

5.6. Performance evaluation

In this section, numerical results for the evaluation of the offline approach and the proposed two-stage SARSA algorithm are presented. A summary of all the variables considered for the simulations is given in Table 5.2.

In addition to the parameters introduced in Section 3.6, which we do not describe here again for brevity, we consider $N = 3$ receivers, which are assumed to be uniformly distributed around the transmitter in a radius $r_{\max} = 50$m. Taking into account that this scenario includes more nodes than in the previous chapters, $T = 2000$ independent random energy, data and channel realizations are generated, where each realization corresponds to an episode where N_0 harvests energy from the environment $I = 2000$ times. The step size δ used in the definition of the action set $\mathcal{A}^{\mathrm{up}}$ of the upper stage is set to $\delta = 2\%$. For the lower stage, the parameters $\eta_{n,i}$ are assumed to be taken from the set $\mathcal{J} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Moreover, the data arrival process for the data intended for N_n consists of a random number of data packets arriving in each time interval and following a Poisson distribution with mean value λ. In addition, the packets are assumed to be of equal size $d = 1$kbit. The incoming data packets are stored in the corresponding finite virtual data buffer. The size of the data buffer of N_0 is $D_{\mathrm{max},0} = W\tau \log_2(1+\beta\bar{\Gamma})$, where β is the data buffer size factor and $\bar{\Gamma}$ is the average SNR considering all the receivers.

To compare the performance of our proposed two-stage SARSA algorithm, three different approaches are considered as references:

- SARSA: This approach only considers the upper stage explained in Section 5.5.3.1. To minimize the interference, the selected power is allocated in each time interval i for the transmission to the receiver with best channel conditions.
- Maximum rate: In this approach, the battery is depleted in each time interval i and all the power is used to serve the receiver with the best channel condition.
- Equal power allocation: In this approach, the battery is depleted in each time interval i and the power is evenly distributed for the transmission of data to all the N receivers.

The sum throughput versus the power density ρ of the EH source is shown in Figure 5.3. In this case, we have reduced the number of time intervals to $I = 100$ and consider $N = 2$ receivers in order to be able to calculate the offline optimum as a reference. Furthermore, an infinitely full data buffer is assumed at N_0. As expected, the largest throughput is achieved by the offline approach which assumes perfect non-causal knowledge of the system dynamics. The maximum rate and SARSA approaches outperform the proposed two-stage SARSA algorithm. This is because they only transmit to the user with the best channel condition in every time interval. As the data buffer is infinitely full, this is the optimal approach for the considered scenario, as explained in section 5.4. However, our two-stage SARSA algorithm is designed for the most general case in which a data arrival process is associated to each of the users. Therefore, it needs to learn that, as the data buffers are infinitely full, the data buffer levels $D_{n,i}$ are not longer relevant for the power allocation. For comparison, we have depicted with a dashed line the performance achieved by a modified two-stage SARSA algorithm which does not consider the data buffer levels $D_{n,i}$ as a deciding factor among the different receivers. It can be seen that the performance is close to the one achieved by SARSA and the maximum rate approach. The difference between the two learning approaches, i.e., modified two-stage SARSA and SARSA, and the maximum rate approach is due to the fact that only $N = 100$ time intervals are considered for the simulation. As it is shown in Figure 5.7, when such a small number of time intervals is considered, the learning approaches have not yet converged.

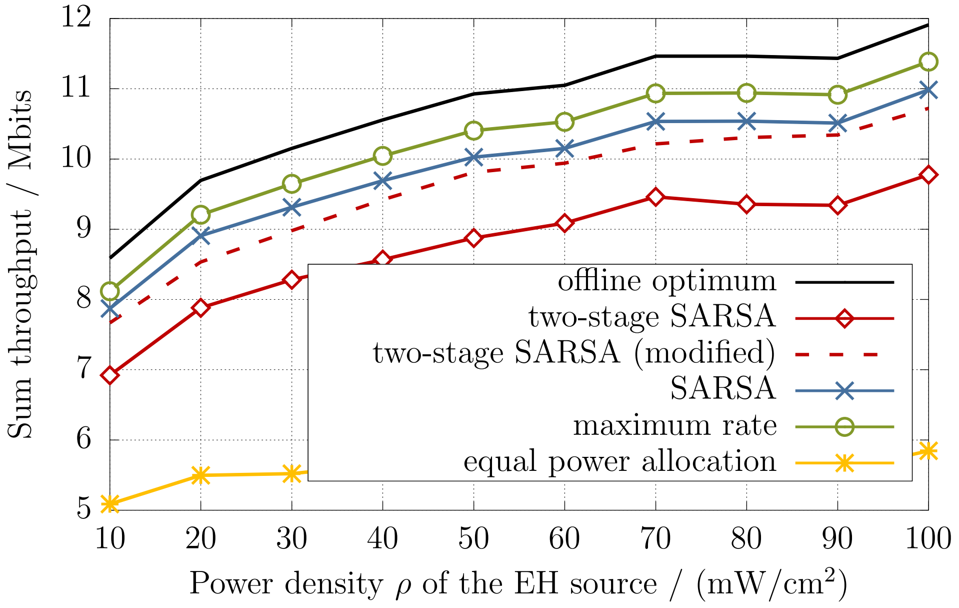


Figure 5.3. Sum throughput versus the power density ρ of the EH source for $N = 2$.

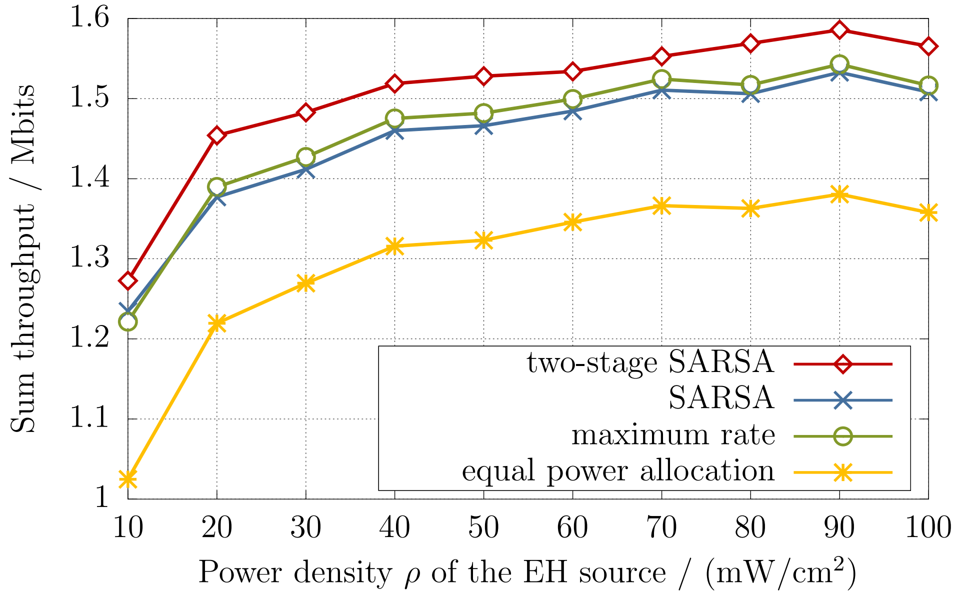


Figure 5.4. Sum throughput versus the power density ρ of the EH source for $N = 2$.

In Figure 5.4, we show the sum throughput when a finite data buffer is assumed in the same scenario considered in Figure 5.3, i.e., $N = 2$ users and $I = 100$ intervals. The offline optimum is not included because the feasibility of the problem cannot be guaranteed when data arrival processes are considered. For all the schemes, the throughput increases with increasing values of the power density ρ of the EH source because more energy is available for the transmission of data and a larger average SNR can be achieved. Note that the achieved throughput is lower compared to Figure 5.3, because it is limited by the data arrival processes. When a finite data buffer is considered, the proposed two-stage SARSA algorithm outperforms all the reference schemes. This is because, in addition to the battery level and channel conditions, it takes into account the levels of the virtual data buffers in the power allocation. If only channel conditions are considered, like in the case of SARSA and maximum rate, data buffer overflows are not avoided and the achievable throughput is reduced. For a power density of 50 mW/cm^2, the proposed two-stage SARSA algorithm achieves a throughput 3%, 4% and 15% higher than the SARSA, maximum rate and equal power allocation algorithms, respectively. When only $N = 2$ receivers are considered, there is little room for improvement. As a result, the throughput obtained by the proposed two-stage SARSA is only modestly increased in comparison to the SARSA and maximum rate approaches. However, as it will become evident in the next figures, larger gains can be achieved when larger network sizes are considered.

Figure 5.5 shows the impact of the number N of receivers on the average throughput per time interval achieved in the system when a finite data buffer is assumed. In this case, the offline optimum is not considered because, as explained in Section 5.4, the resulting throughput maximization problem when more than two receivers are considered is non-convex and a unique solution cannot be found. The achieved throughput increases with the number N of users but saturates after a certain network size is reached. This saturation point depends on the considered scheme. Specifically, the proposed two-stage SARSA algorithm achieves a higher throughput compared to the reference schemes, i.e., approximately 1.3 times higher than the SARSA and maximum rate approaches and two times higher than the equal power allocation scheme, for $N = 10$. This is because two-stage SARSA is able to efficiently use the harvested energy to transmit data to the different receivers by considering not only their channel conditions but also the data buffer levels. The SARSA and maximum rate approaches achieve approximately the same throughput because in each time interval, both schemes transmit data only to the receiver with the best channel conditions. Note that the throughput is lower than for the two-stage SARSA because the power allocation decision is based solely on the channel condition and does not consider the state of the data buffers. Moreover, the throughput achieved by the equal power allocation scheme saturates at $N = 2$

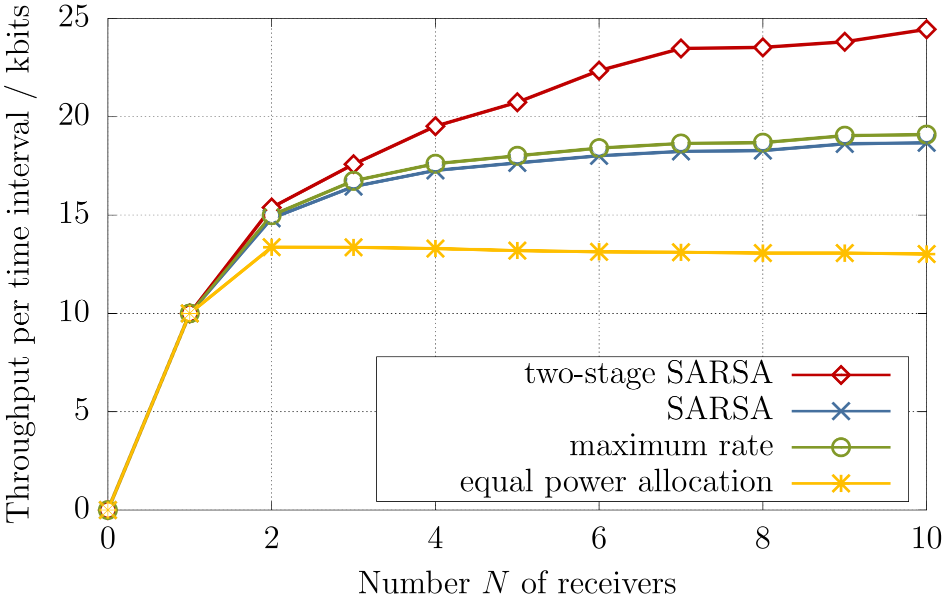


Figure 5.5. Throughput per time interval versus the number N of receivers.

users and slowly decreases when the network size increases. This is due to the fact that the available harvested energy is equally distributed to the transmission to all the receivers. Therefore, as more receivers are considered, less energy is available for the transmission to the receivers with better channel conditions.

In Figure 5.6, we show the average throughput per receiver per time interval versus the number N of receivers. Specifically, we consider five different numbers of receivers and show the mean value and the standard deviation of the achieved throughput per receiver. It can be observed that although the two-stage SARSA approach does not focus on fairness, it has the smallest variation among all the considered approaches. Furthermore, as the different data arrival processes have the same mean value, the reduced variation of two-stage SARSA means that two-stage SARSA enables the transmitter to send data to more receivers compared to the reference schemes. Moreover, for $N = 10$ receivers, the two-stage SARSA approach achieves a higher average throughput per receiver compared to the other schemes. As in the previous cases, the throughput achieved by the SARSA and maximum rate approaches is approximately the same, while the lowest throughput corresponds to the equal power allocation scheme.

Finally, in Figure 5.7, the convergence of the two learning approaches, i.e., two-stage SARSA and SARSA, is depicted. It can be seen that the convergence rate of both approaches is equal because they are both based on the SARSA update. However,

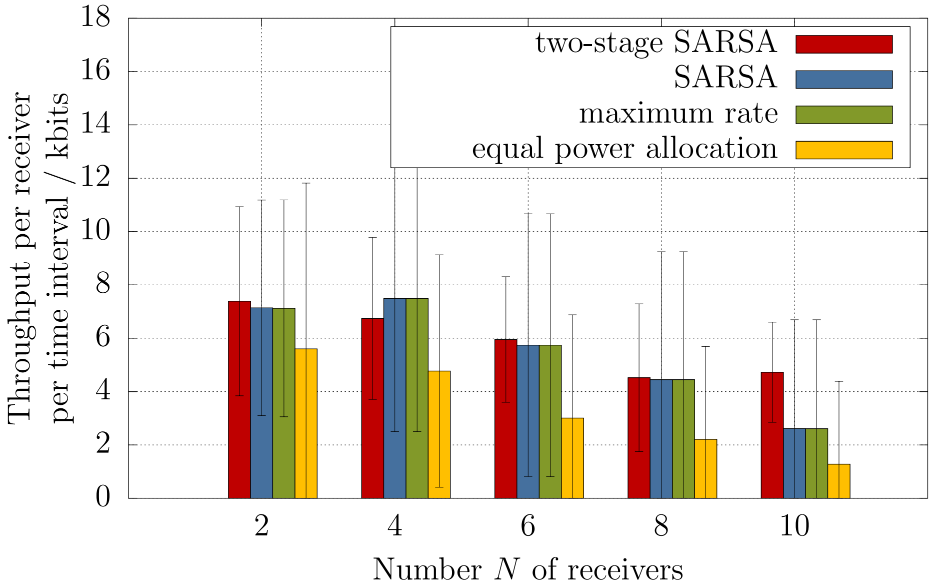


Figure 5.6. Throughput per receiver and time interval versus the number N of receivers.

note that by splitting the learning task into two sub-task, the proposed two-stage SARSA is able to identify power allocation solutions that lead to a higher performance even for a small number of iterations.

5.7. Conclusions

In this chapter, we have investigated offline and learning approaches for the power allocation problem for throughput maximization in the EH broadcast scenario. This power allocation problem entails the selection of the transmit power values to use for the transmission of data to each receiver while fulfilling the energy causality and energy overflow constraints at the transmitter.

Initially, we formulate the optimization problem to find the power allocation policy that maximizes the throughput. We show that the resulting problem is non-convex for the general case in which an arbitrary number N of receivers is considered. Based on results from the literature, an offline approach is proposed for the case when $N = 2$ receivers are assumed. For this set-up, it is shown that the original power allocation problem, which finds the transmit power values to use for the transmission of data to the two receivers in each time interval, can be reformulated to ensure convexity.

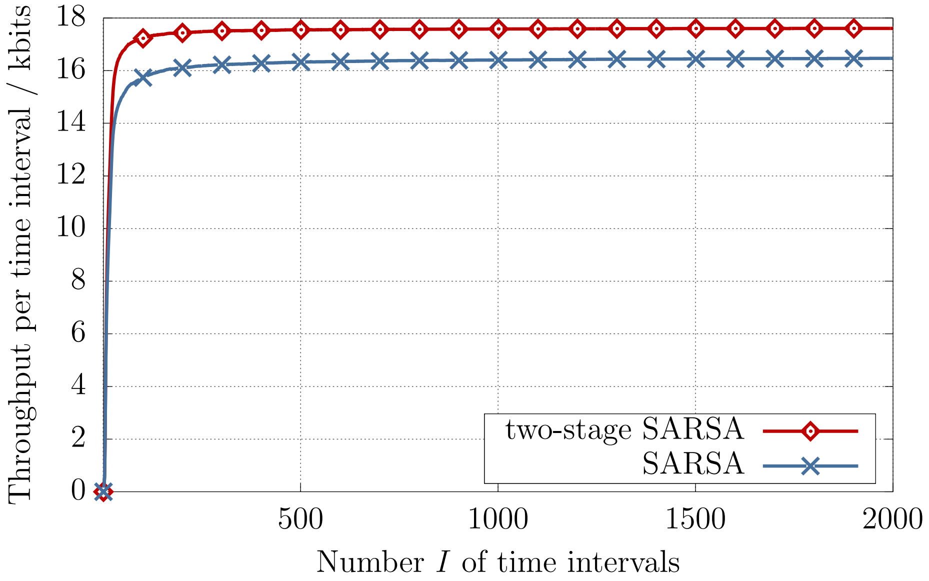


Figure 5.7. Throughput per time interval versus the number I of time intervals.

Specifically, the problem is rewritten such that it depends on the total power to be used in each time interval and a power splitting parameter which determines how the transmit power is split for the transmission of data to the two receivers. Furthermore, it is shown that in every time interval, the power splitting parameter is fully determined by the channel gains, the noise power and the priority assigned to each receiver. As a result, it can be optimized independently of the total transmit power. Moreover, considering that the selection of the total transmit power resembles an EH point-to-point scenario, the directional backward glue pouring algorithm described in Chapter 3 is used to find the optimal power allocation policy.

Additionally, a learning approach is proposed to find the power allocation policy that aims at maximizing the throughput when only causal knowledge regarding the EH, data arrival and channel fading processes is available and an arbitrary number of receivers is considered. The proposed learning algorithm is motivated by the offline approach in which the power allocation problem for the two receiver case is separated into choosing the total transmit power to use in every time interval and the selection of the corresponding power splitting parameter. Specifically, our learning algorithm is composed of two stages which solve independent learning tasks, namely, the selection of the total power to be used in each time interval and the distribution of this total power among the different transmissions. In both stages, linear function approximation is considered to handle the infinite number of states and customized feature functions are proposed. Furthermore, we show that the convergence of the learning process in each stage depends on the selection of the learning rate parameter. Moreover, by means of a computational complexity analysis we show that the task of learning how to distribute the power is more computationally demanding than the selection of the total transmit power to be used. Additionally, through numerical simulations we show that with the proposed learning approach a gain of up to 40 % can be obtained compared to standard learning approaches.

Chapter 6

Energy harvesting multiple access scenario

6.1. Introduction

In this chapter, an EH multiple access scenario is considered. In contrast to the previous chapters, we shift our focus from the power allocation problem and investigate the resource allocation problem in the EH multiple access scenario. For this purpose, we consider that K orthogonal and distinguishable resources are available for the communication and address the problem of how to efficiently allocate them to the multiple EH transmitters considering their own EH process as well as the channel fading processes associated to them. Moreover, for this scenario, offline and learning approaches are investigated.

The chapter is organized as follows. In Section 6.2 the EH multiple access scenario is described and the corresponding system assumptions are presented. In Section 6.3, the resource allocation problem for throughput maximization is formulated. Due to the complexity of the resource allocation problem, only the case when the EH transmitters have an infinitely full data buffer is considered. The resulting problem is identified as a non-linear knapsack problem which has a combinatorial nature. Therefore, an offline approach based on dynamic programming is investigated in Section 6.4 to find the optimal solution. A learning approach is proposed in Section 6.5 for the case when only causal knowledge regarding the EH and channel fading processes is available. Specifically, a novel RL algorithm, termed combinatorial SARSA, is proposed. In Section 6.6, the performance of the proposed approaches is evaluated through several numerical simulations. In Section 6.7 we discuss how the proposed approaches can be extended to consider finite data buffer at the EH transmitters. Finally, Section 6.8 concludes the chapter.

Parts of this chapter have been published by the author of this dissertation in [OWK19].

6.2. Scenario description and assumptions

In this section, the EH multiple access communication scenario is described. Furthermore, considering the system model presented in Section 2.2, the system assumptions

Table 6.1. Parameters associated to the EH multiple access communication scenario.

	Parameter	Description
General	i	Index of the time interval
	I	Total number of time intervals
	k	Index of the resources
	K	Number of available resources
	N	Total number of EH transmitters
	N_n	n$^{\text{th}}$ EH transmitter node, $n = 1, ..., N$
	N_0	non-EH receiver node
	τ	Time interval duration
	τ^{Sig}	Time duration required for the signaling
	$\chi_{n,i,k}$	Binary variable indicating if resource k is allocated to N_n in time interval i
	$\mathbf{X}_i$	Resource allocation solution in time interval i
Energy	$B_{n,i}$	Battery level of EH node N_n, measured at the beginning of time interval i
	$B_{\mathrm{max},n}$	Battery capacity of EH node N_n
	$E_{\mathrm{max},n}$	Maximum amount of energy that can be harvested by 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^{Sig}	Amount of energy used by EH node N_n for signaling to N_0 in time interval i
	$E_{n,i}^{\mathrm{Tx}}$	Transmit energy used by EH node N_n in time interval i
	$p_{n,i,k}^{\mathrm{Tx}}$	Transmit power used by EH node N_n within resource k in time interval i
Data	$D_{\mathrm{max},n}$	Data buffer size of EH node N_n
	$D_{n,i}$	Data buffer level of EH node N_n
	R_i^{MAC}	Total amount of data received by N_0 in time interval i
Channel	$g_{n,i,k}$	Channel gain of the link between N_0 and N_0 when resource k is considered
	$h_{n,i,k}$	Channel coefficient of the link between N_n and N_0 when resource k is considered
	W	Bandwidth
	σ_0^2	Noise power at N_0

are introduced. All the related parameters, which will be described in the following, are summarized in Table 6.1.

As depicted in Figure 6.1, the EH multiple access scenario consists of N single-antenna EH transmitters, termed N_n, with $n = 1, ..., N$, and one single-antenna non-EH receiver, termed N_0. The transmitters N_n harvest energy from the environment and use it to send data to N_0. Specifically, an amount of harvested energy, denoted by $E_{n,i}$, is received by N_n at the end of every time interval i, $i = 1, ..., I$ and it is stored in the corresponding finite battery with capacity $B_{\mathrm{max},n}$. The battery level $B_{n,i}$ is measured at the beginning of each time interval i and indicates the amount of energy available in the battery of N_n. The energy $E_{n,i}^{\mathrm{Circ}}$ consumed by the circuit at N_n is assumed to be constant for all the time intervals such that $E_{n,i}^{\mathrm{Circ}} = E_n^{\mathrm{Circ}}$, $\forall i$. Additionally, the non-EH receiver N_0 is assumed to be connected to a fixed power supply and therefore, it is always available to receive the transmitted data.

In contrast to the previous chapters, only infinitely full data buffers are considered at the EH transmitters N_n. This is because, the resource allocation problem is more complex than the power allocation problem investigated in the previous chapters. There-

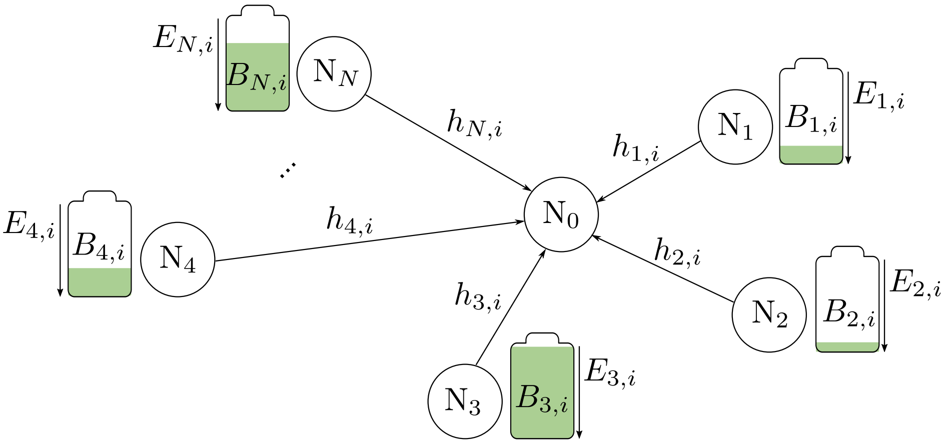


Figure 6.1. EH multiple access scenario with EH transmitters.

fore, we focus on the impact of EH and investigate the case when the achievable throughput is only limited by the availability of harvested energy and the resource allocation policy. Nevertheless, in Section 6.7, we describe how the proposed approaches can be extended to consider the finite data buffer case. When infinitely full data buffers are considered, the data buffer size $D_{\max,n}$ of each N_n is assumed to be infinite. Moreover, it is assumed that the data buffer level $D_{n,i}$ is also infinite for all the time intervals.

In our model, it is assumed that within each time interval i, there are K orthogonal and distinguishable resources available for the transmission of data, e.g., a fraction of a time interval if TDMA is considered or one sub-carrier in the case of FDMA. Therefore, we consider that N_0 has the task of allocating the available K orthogonal resources to the EH transmitters in each time interval i while aiming at maximizing the amount of received data. For this purpose, exclusive allocation is considered, i.e., each resource k, $k = 1, ..., K$ can be allocated to only one transmitter N_n at a time but multiple resources can be allocated to a single transmitter. Our goal is to find a resource allocation policy at N_0 considering the limited amount of resources, the EH processes at the transmitters, and the channel fading processes associated to their channels.

In order to find the resource allocation policy, it is assumed that N_0 has knowledge about the EH and channel fading processes associated to all the transmitters N_n. In offline approaches, this knowledge is assumed to be non-causally available. This means that at the beginning of the data transmission, N_0 knows all the channel gains $g_{n,i,k} = |h_{n,i,k}|^2$ associated to the channels to every N_n for all the time intervals $i = 1, ..., I$ in every resource $k = 1, ..., K$, as well as the amounts of energy $E_{n,i}$, $\forall n, i$ that will be harvested. On the contrary, in learning approaches this knowledge about the EH and channel fading processes is only causally available. Considering receiver side channel state information, we assume that in every time interval i the receiver N_0 knows the current and past channel gains $g_{n,j,k}$, $j = 1, ..., i$ associated to the channels to every transmitter N_n in every resource k. Furthermore, it is assumed that in each time interval i a constant amount E_n^{Sig} of energy is used by each N_n in order to signal its current battery level $B_{n,i}$ to N_0. Based on the results obtained in Chapter 4[1], it is assumed that the time duration τ^{Sig} required for the signaling is much smaller than the time interval duration τ. Thus, we assume that the time duration τ is used for data transmission.

When resource k is granted to N_n in time interval i, the transmit power $p_{n,i,k}^{\mathrm{Tx}}$ to use within this assigned resource needs to be determined. However, as the assigned resources are orthogonal, this power allocation problem does not depend on the power allocation policy of the other transmitters. Consequently, the power allocation approaches described in Chapter 3 can be applied in every transmitter N_n separately. As our focus is the resource allocation problem in the EH multiple access scenario, we consider a simple hasty power allocation policy in which each node N_n uses all the energy in its battery for the transmission of data every time that a resource has been granted to it. In case more than one resource is allocated to N_n in time interval i, equal power allocation is considered. Let $\chi_{n,i,k} \in \{0, 1\}$ be a binary variable that indicates if the resource k has been granted to N_n in time interval i. A resource allocation solution for time interval i, termed $\mathbf{X}_i$, is a matrix formed by the collection of the $\chi_{n,i,k}$ values for all the N EH transmitters and all the K resources in time interval i such that

$$\mathbf{X}_i = \begin{pmatrix} \chi_{1,i,1} & \chi_{2,i,1} & \cdots & \chi_{N,i,1} \\ \chi_{1,i,2} & \chi_{2,i,2} & \cdots & \chi_{N,i,2} \\ \vdots & \vdots & \ddots & \vdots \\ \chi_{1,i,K} & \chi_{2,i,K} & \cdots & \chi_{N,i,K} \end{pmatrix} \tag{6.1}$$

Considering $\mathbf{X}_i$, the transmit power $p_{n,i,k}^{\mathrm{Tx}}$ is calculated as

$$p_{n,i,k}^{\mathrm{Tx}} = \begin{cases} \frac{B_{n,i} - E_n^{\mathrm{Circ}} - E_n^{\mathrm{Sig}}}{\tau \sum\limits_{k=1}^{K} \chi_{n,i,k}} & \text{if} \sum\limits_{k=1}^{K} \chi_{n,i,k} \geq 1 \\ 0 & \text{else.} \end{cases} \tag{6.2}$$

Note that if at least one resource is granted to N_n in time interval i, the amount of energy $E_{n,i}^{\mathrm{Tx}}$ used for data transmission in the time interval is given by

$$E_{n,i}^{\mathrm{Tx}} = B_{n,i} - E_n^{\mathrm{Circ}} - E_n^{\mathrm{Sig}}. \tag{6.3}$$

[1] Simulation results in Section 4.2.6 show that when τ^{Sig} is approximately one thousandth of τ, a larger throughput can be achieved compared to the cases when longer τ^{Sig} are considered.

Otherwise, it is zero. Considering (2.3), the battery level of each N_n is calculated as

$$B_{n,i+1} = \begin{cases} \min\{B_{\max,n},\ E_{n,i}\}, & \text{if } \sum\limits_{k=1}^{K} \chi_{n,i,k} \geq 1 \\ \min\{B_{\max,n},\ B_{n,i} + E_{n,i}\}, & \text{else.} \end{cases} \tag{6.4}$$

The total throughput R_i^{MAC} in the system is the total amount of data received by N_0 in time interval i and it is measured in bits. R_i^{MAC} is approximated using Shannon's capacity formula as

$$R_i^{\mathrm{MAC}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \tau W \log_2 \left(1 + \frac{g_{n,i,k} p_{n,i,k}^{\mathrm{Tx}}}{\sigma_0^2} \right), \tag{6.5}$$

where W is the available bandwidth and σ_0^2 is the noise power at N_0.

6.3. Problem formulation

In this section, we formulate the resource allocation problem for throughput maximization in the EH multiple access scenario.

Since the feasibility of the resource allocation solutions depends on the EH and channel fading processes associated to the different transmitters, the energy causality and battery overflow constraints, given in (2.4) and (2.5), should be taken into account. Furthermore, as exclusive allocation is considered, not every possible permutation of $\chi_{n,i,k}$, $\forall n, k$, is a suitable solution of the problem. Thus, the resource allocation problem in the EH multiple access scenario is written as

$$\left(\chi_{n,i,k}^{\mathrm{opt}}\right)_{n,i,k} = \underset{\substack{\chi_{n,i,k} \in \{0,1\},\, n=\{1,\ldots,N\}, \\ i=\{1,\ldots,I\},\, k=\{1,\ldots,K\}}}{\operatorname{argmax}} \sum_{i=1}^{I} R_i^{\mathrm{MAC}} \tag{6.6a}$$

$$\text{subject to} \quad \sum_{i=1}^{J} \sum_{k=1}^{K} \tau p_{n,i,k}^{\mathrm{Tx}} + \sum_{i=1}^{J} E_n^{\mathrm{Circ}} + \sum_{i=1}^{J} E_n^{\mathrm{Sig}} \leq \sum_{i=1}^{J-1} E_{n,i},\ J = 1, \ldots, I, \tag{6.6b}$$

$$\sum_{i=1}^{J} E_{n,i} - \sum_{i=1}^{J} \sum_{k=1}^{K} \tau p_{n,i,k}^{\mathrm{Tx}} - \sum_{i=1}^{J} E_n^{\mathrm{Circ}} - \sum_{i=1}^{J} E_n^{\mathrm{Sig}} \leq B_{\max,n},\ \forall n, J, \tag{6.6c}$$

$$p_{n,i,k}^{\mathrm{Tx}} = \begin{cases} \dfrac{B_{n,i} - E_n^{\mathrm{Circ}} - E_n^{\mathrm{Sig}}}{\tau \sum\limits_{k=1}^{K} \chi_{n,i,k}} & \text{if } \sum\limits_{k=1}^{K} \chi_{n,i,k} \geq 1 \\ 0 & \text{else,} \end{cases} \tag{6.6d}$$

$$\sum_{n=1}^{N} \chi_{n,i,k} = 1, \tag{6.6e}$$

$$\sum_{k=1}^{K} \sum_{n=1}^{N} \chi_{n,i,k} = K. \tag{6.6f}$$

where R_i^{MAC} is calculated as in (6.5).

6.4. Offline approach

6.4.1. Dynamic programming for EH multiple access scenarios

The problem in (6.6) can be categorized as a non-linear knapsack problem which is NP-hard [BS02]. The dimension of the problem grows exponentially with the number K of resources and the number N of transmitters. Furthermore, the constraints in (6.6b) and (6.6c) impose a dependency of the resource allocation solution over time.

For such problems, dynamic programming (DP) can be exploited to find the optimal solution [KPP04]. The idea behind DP is the so-called principle of optimality defined by Bellman in [Bel54] as *"Any optimal policy has the property that, whatever the current state and decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the current decision."* Considering the optimization problem in (6.6), the principle of optimality determines that the optimal solution can be constructed in a piecewise fashion, i.e., by first solving a subproblem, which considers only a subset of the variables, and then, by iteratively increasing and checking the optimality of the resulting subproblems until the solution of the original problem is found.

In this section, we leverage the DP algorithm Policy Iteration in order to find the optimal resource allocation policy in the EH multiple access scenario. For this purpose, we first model the scenario as an MDP in Section 6.4.2 and then use this formulation to describe the Policy Iteration algorithm in Section 6.4.3.

6.4.2. Markov decision process

In our scenario, the time interval duration τ is fixed and known. Moreover, as discussed in Section 6.2, each transmitter adopts a hasty power allocation policy. Consequently, in time interval i, the resource allocation depends solely on the amount of energy available for transmission and the channel conditions of the transmitters, i.e., the values of $B_{n,i}$, and $g_{n,i,k}$. As the previous battery levels $B_{n,j}$ and channel gains $g_{n,j,k}$, $j < i$, do not need to be taken into account, the system under consideration fulfills the Markov property and can be modeled as an MDP.

In time interval i, the state $S_i \in \mathcal{S}$ is determined by the battery levels $B_{n,i}$ and the channel gains $g_{n,i,k}$ associated to all the transmitters N_n in every resource k. However, to reduce the number of variables to be considered, a pseudo-SNR $\tilde{\Gamma}_{n,i,k}$, given by

$$\tilde{\Gamma}_{n,i,k} = \frac{g_{n,i,k} B_{n,i}}{\tau \sigma_0^2}, \tag{6.7}$$

is defined for every transmitter N_n in every avaiable resource k. The larger $\tilde{\Gamma}_{n,i,k}$, the more suitable is N_n for the transmission of data in time interval i using resource k. This is because N_n experiences a good channel, has a large amount of energy stored in its battery, or both. Furthermore, note that $\tilde{\Gamma}_{n,i,k}$ can take any value in a continuous range. As a result, the set $\mathcal{S}$ contains infinitely many possible states. The set $\mathcal{A}$ contains all the possible resource allocation solutions $\mathbf{X}_i$, defined in (6.1), which can be selected in every time interval i. Taking into account that $\mathbf{X}_i$ is a permutation of all the values $\chi_{n,i,k}$ can take, the set $\mathcal{A}$ is finite but its size grows exponentially with the number of transmitters as

$$|\mathcal{A}| = N^K. \tag{6.8}$$

The transition model P defines all the probabilities $P_{S_i,S_{i+1}}^{\mathbf{X}_i} = \mathbb{P}\left[S_{i+1}|S_i, \mathbf{X}_i\right]$ of going from state S_i to state S_{i+1} after selecting $\mathbf{X}_i \in \mathcal{A}$ in time interval i. Finally, the rewards $R_i^{\mathrm{MAC}} \in \mathcal{R}$ indicate how beneficial it is to select $\mathbf{X}_i$ in S_i and it is given by the throughput R_i^{MAC} defined in (6.5). Additionally, in order to account for the preference of achieving a higher throughput in the current time interval versus achieving a higher throughput later on, we consider a discount factor $0 \leq \gamma \leq 1$. Our aim is now to maximize the discounted throughput given by

$$R^{\mathrm{MAC}} = \lim_{I \to \infty} \mathbb{E}\left[\sum_{i=1}^{I} \gamma^{i-1} R_i^{\mathrm{MAC}}\right]. \tag{6.9}$$

The solution of the MDP is given by the resource allocation policy π which is a map from the states S_i to the action $\mathbf{X}_i$ that should be selected, i.e., $\mathbf{X}_i = \pi(S_i)$ [SB18].

6.4.3. Policy Iteration

The MDP formulated in Section 6.4.2 is an infinite MDP. However, to facilitate the implementation of the policy iteration algorithm, in this section we assume that the

Algorithm 6.1 Policy iteration [SB18]

```
 1: initialize $\mathrm{V}^{\pi}$, $\pi$ and $\varepsilon$
 2: repeat                                                        ▷ policy evaluation
 3:     $\Theta = 0$
 4:     for each state $S_i \in \mathcal{U}$ do
 5:         set $\mathbf{X}_i = \pi(S_i)$
 6:         set $v = \mathrm{V}^{\pi}(S_i)$
 7:         $\mathrm{V}^{\pi}(S_i) = \sum_{S_{i+1} \in \mathcal{U}} P^{\mathbf{X_i}}_{S_i, S_{i+1}} \left[ R^{\mathrm{MAC}}_i(S_i, \mathbf{X}_i) + \gamma \mathrm{V}^{\pi}(S_{i+1}) \right]$
 8:         set $\Theta = \max\{\Theta, |v - \mathrm{V}^{\pi}(S_i)|\}$
 9:     end for
10: until $\Theta < \varepsilon$
11: set policyStable = true                                       ▷ policy improvement
12: for each state $S_i \in \mathcal{U}$ do
13:     $\mathbf{X}_{\mathrm{old}} = \pi(S_i)$
14:     $\pi(S_i) = \mathrm{argmax}_{\mathbf{X}_i} \sum_{S_{i+1} \in \mathcal{U}} P^{\mathbf{X}_i}_{S_i, S_{i+1}} \left[ R^{\mathrm{MAC}}_i(S_i, \mathbf{X}_i) + \gamma \mathrm{V}^{\pi}(S_{i+1}) \right]$     ▷ Eq. 6.11
15:     if $\mathbf{X}_{\mathrm{old}} \neq \pi(S_i)$ then
16:         policyStable = false
17:     end if
18: end for
19: if policyStable = true then                                   ▷ stopping criteria
20:     return $\mathrm{V}^{*} \approx \mathrm{V}^{\pi}$, $\pi^{*} \approx \pi$
21: else
22:     go to 2
23: end if
```

state space is discretized such that only a finite subset $\mathcal{U}$ of possible states, with $\mathcal{U} \subset \mathcal{S}$, is considered. This means that the amounts $E_{n,i}$ of harvested energy and the channel gains $g_{n,i,k}$ are taken from the finite sets $\mathcal{E}$ and $\mathcal{H}$, respectively.

As its name suggests it, policy iteration is an iterative approach to find the optimal policy in an MDP. It is composed of two stages, namely, policy evaluation and policy improvement. The policy evaluation stage computes the state-value function V^{π} for any resource allocation policy π, while the policy improvement stage produces a new and improved policy π' by making it greedy with respect to the state-value function V^{π} of the original policy π [SB18, Ber07]. Note that in contrast to the learning approaches presented in the previous chapters, the policy iteration considers the state-value function V^{π} instead of the action-value function Q^{π}. This is because the policy iteration algorithm evaluates the policy as a whole and not based on the individual actions given a certain state.

The policy iteration algorithm is presented in Algorithm 6.1. Initially, an arbitrary policy π is selected and the values of the state-value function V^{π} are initialized, e.g., $\mathrm{V}^{\pi}(S_i) = 0$, $\forall S_i \in \mathcal{U}$ (line 1). Then, the policy evaluation stage is considered in order to compute the state-value function V^{π} for the current resource allocation policy π. When the transition probabilities $P^{\mathbf{X}_i}_{S_i, S_{i+1}}$ are known, the state-value function V^{π} can be calculated using the Bellman equation defined in (2.19), which we repeat it here for readability

$$\mathrm{V}^{\pi}(S_i) = \sum_{\mathbf{X}_i \in \mathcal{A}} \pi(\mathbf{X}_i|S_i) \sum_{S_{i+1} \in \mathcal{U}} P^{\mathbf{X}_i}_{S_i,S_{i+1}} \left[R^{\mathrm{MAC}}_i(S_i, \mathbf{X}_i) + \gamma \mathrm{V}^{\pi}(S_{i+1})\right] \tag{6.10}$$

[SB18], where $\pi(\mathbf{X}_i|S_i)$ is the probability of selecting action $\mathbf{X}_i$ in state S_i when policy π is followed. As such formulation implies the solution of $|\mathcal{U}|$ linear equations, an iterative approach is considered instead. For this purpose, an accuracy parameter ε is used as a stopping criteria for the estimation of V^{π} (lines 2-10). Specifically, the state-value function V^{π} is estimated by iteratively solving the Bellman equation for the state-value function in (6.10). Once the desired accuracy is obtained, the algorithm enters the policy improvement stage where a new and improved policy π' is obtained. The idea behind the policy improvement stage is to modify the policy π in a greedy fashion by forcing it to select, in every state S_i, the action $\mathbf{X}_i$ that yields the maximum value of the state-value function V^{π} (lines 12-18), i.e.,

$$\pi'(S_i) = \operatorname*{argmax}_{\mathbf{X}_i} \sum_{S_{i+1} \in \mathcal{U}} P^{\mathbf{X}_i}_{S_i,S_{i+1}} \left[R^{\mathrm{MAC}}_i(S_i, \mathbf{X}_i) + \gamma \mathrm{V}^{\pi}(S_{i+1})\right]. \tag{6.11}$$

Note that this improvement is done only once for each state. Next, the resulting policy π' is compared to the current policy π (lines 19-23). If there are no changes in the policy, then it is considered stable and the policy iteration algorithm terminates. On the contrary, if π' differs from π, the algorithm returns to the policy evaluation stage for a new iteration.

6.5. Learning approach

6.5.1. The combinatorial RL problem

In this section, we propose a learning approach to find the resource allocation policy that aims at maximizing the throughput in the EH multiple access scenario when only causal knowledge regarding the EH and channel fading processes is available.

As described in the previous section, the dimension of the problem in (6.6) grows exponentially with the number K of resources and the number N of EH transmitters. Therefore, the main challenge to be addressed by the learning approach is how to manage the high dimensionality of the problem while still considering an infinite set of states. For this purpose, we formulate the resource allocation problem as an RL problem. Specifically, we propose a novel RL algorithm termed combinatorial SARSA.

The name of the algorithm stands for its ability to handle the combinatorial nature of the resource allocation solutions. Our approach is inspired by the so called naive strategy proposed in [Ont13] for MAB. Here, we extend it to an RL setting and combine it with linear function approximation to manage the infinite number of states. The strength of our algorithm is its ability to split the original RL problem into $K+1$ smaller problems, thus tackling the curse of dimensionality of combinatorial problems. This increases the learning speed and, consequently, the throughput, compared to traditional RL approaches.

In the following and based on the MDP formulation of Section 6.4.2, we first introduce the naive strategy proposed in [Ont13] for MAB and extend it to RL problems in Section 6.5.2. Next, we continue with the application of linear function approximation in Section 6.5.3 and explain the action selection strategy in Section 6.5.5. A summary of the proposed combinatorial SARSA is presented in 6.5.6. Moreover, convergence guarantees and a computational complexity analysis of the proposed algorithm are presented in Sections 6.5.7 and 6.5.8, respectively.

6.5.2. Naive Strategy for RL

The naive strategy was initially proposed for combinatorial MAB problems in [Ont13]. It is based on the idea that the reward distribution, which depends on the combination of multiple variables, can be approximated by the sum of a set of reward functions that depend on only one variable at a time. Here, we extend this idea to the more complex case of RL problems.

In our setting, the reward function, defined in (6.5), is the throughput R_i^{MAC} achieved in one time interval given a resource allocation solution $\mathbf{X}_i$. However, note that (6.5) can also be written as the sum of the throughputs $R_{k,i}^{\mathrm{MAC}}$ achieved by the allocation of resource k, $k=1,...K$, as

$$R_i^{\mathrm{MAC}} = \sum_{k=1}^{K} R_{k,i}^{\mathrm{MAC}}, \tag{6.12}$$

where $R_{k,i}^{\mathrm{MAC}}$ is calculated as

$$R_{k,i}^{\mathrm{MAC}} = \sum_{n=1}^{N} \tau W \log_2 \left(1 + \frac{g_{n,i,k} p_{n,i,k}^{\mathrm{Tx}}}{\sigma_0^2}\right). \tag{6.13}$$

This reformulation allows us to see the resource allocation problem from a new angle, i.e., instead of jointly considering how the K resources can be distributed among the

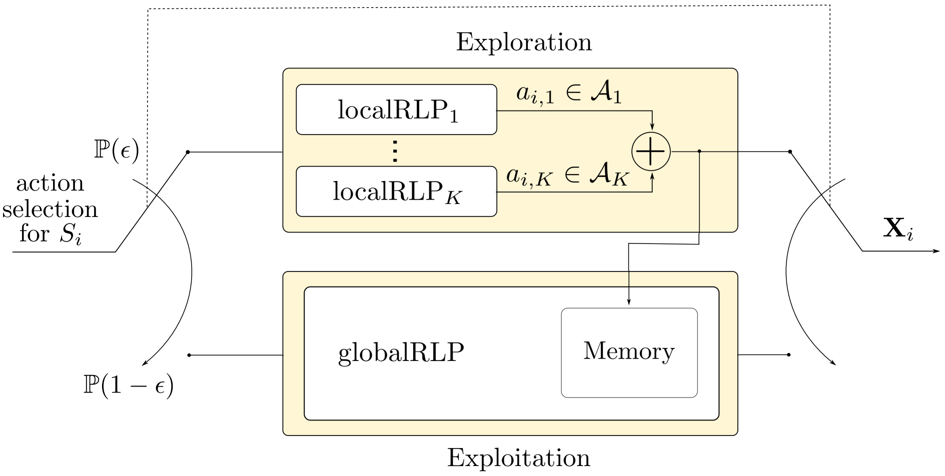


Figure 6.2. Schematic of the application of the naive strategy to RL problems.

N users such that the overall throughput is maximized, the throughput maximization problem is considered for one resource at a time. As a result, the original problem of finding the combination of $\chi_{n,i,k}$ that aims at maximizing the throughput is separated into $K+1$ smaller problems, as shown in Figure 6.2.

Following the terminology in [Ont13], K of these problems are termed local RL problems (localRLP) while the remaining one is termed global RL problem (globalRLP). Specifically, each localRLP is associated with one resource k and its task is to learn how to select one transmitter N_n to which said resource will be allocated. The motivation behind this idea is that by learning to maximize each $R_{k,i}^{\mathrm{MAC}}$, the total R_i^{MAC} is also maximized. Moreover, note that, as shown in Figure 6.2, the decisions for all the K resources k, $k=1,...,K$ are simultaneously and independently done in each localRLP. As a result, the action set $\mathcal{A}_k$ of the k^{th} localRLP is composed solely of the set of indices associated to the EH transmitters, i.e., $\mathcal{A}_k = \{1,...,N\}$, $\forall k$, thus tackling the curse of dimensionality in the original formulation because $|\mathcal{A}_k| = N$. The actions $a_{i,k} \in \mathcal{A}_k$ indicate to which EH transmitter N_n resource k is granted in time interval i. By setting the corresponding $\chi_{n,i,k}$ to one, the resource allocation solution $\mathbf{X}_i$ is determined by the collection of the K actions $a_{i,k}$.

While the localRLPs focus on the individual resources, the globalRLP has the task of evaluating the effect of the composite resource allocation solution $\mathbf{X}_i$ on the achieved throughput. Intuitively, the task of the localRLPs is to efficiently explore the resource allocation solutions while the task of the globalRLP is to select, for a given state S_i, the $\mathbf{X}_i$ which is considered the best up to time interval i. Note that the action space of the globalRLP is initially empty, and it is updated every time that a new resource allocation solution $\mathbf{X}_i$ is tried via the localRLPs. This means that when a new resource allocation solution is encountered by the localRLPs, it is stored in the globalRLP. Therefore, the globalRLP does not solve a combinatorial problem, but learns the suitability of the resource allocation solutions that have been tried.

6.5.3. Linear Function Approximation

By means of the naive strategy, explained in Section 6.5.2, we are able to deal with the high dimensionality of the action space. In this section, we focus on the use of linear function approximation to handle the infinite number of states in the definition of the action-value function Q^{π} of each of the $K+1$ RL problems.

As in the previous chapters, the infinite number of states comes from the fact that $B_{n,i}$ and $g_{n,i,k}$ can take any positive value in a continuous range. As a result, the action-value function Q^{π} has also an infinite number of values. Therefore, to be able to compute Q^{π} for every state, it is approximated as a weighted sum of feature functions such that

$$\hat{\mathrm{Q}}^{\pi}(S_i, a_i) = \mathbf{f}^{\mathrm{T}}(S_i, a_i)\mathbf{w} \approx \mathrm{Q}^{\pi}(S_i, a_i) \tag{6.14}$$

[SB18]. Note that in (6.14), we have used a_i to denote the action in a general manner. However, a_i corresponds to $a_{i,k}$ if a localRLP is considered and to $\mathbf{X}_i$ for the globalRLP. In each of the K+1 RL problems, the SARSA update in (3.25) is considered in the estimation of the action-value function Q^{π}. As described in Chapter 3, when linear function approximation is used, the weights $\mathbf{w}$ are adjusted in the direction that reduces the error between Q^{π} and $\hat{\mathrm{Q}}^{\pi}$ following the gradient descent approach. Hence, the updating rule for the localRLPs is given by

$$\Delta\mathbf{w}_k = \zeta_i\big[R_{k,i}^{\mathrm{MAC}} + \gamma\hat{\mathrm{Q}}^{\pi}(S_{i+1}, a_{i+1,k}, \mathbf{w}_k) - \hat{\mathrm{Q}}^{\pi}(S_i, a_{i,k}, \mathbf{w}_k)\big]\mathbf{f}, \tag{6.15}$$

where $\mathbf{w}_k$ denotes the weights of localRLP k, ζ_i is the learning rate and $R_{k,i}^{\mathrm{MAC}}$ is defined in (6.13). Similarly, the weights $\mathbf{w}$ in the globalRLP are updated as

$$\Delta\mathbf{w} = \zeta_i\big[R_{i}^{\mathrm{MAC}} + \gamma\hat{\mathrm{Q}}^{\pi}(S_{i+1}, \mathbf{X}_{i+1}, \mathbf{w}) - \hat{\mathrm{Q}}^{\pi}(S_i, \mathbf{X}_i, \mathbf{w})\big]\mathbf{f}. \tag{6.16}$$

6.5.4. Feature functions

In this section, we describe how the state-action space is represented by the feature functions. Specifically, we use tile coding as approximation technique due to its flexibility, computational efficiency and suitaibility for multi-dimensional continuous spaces

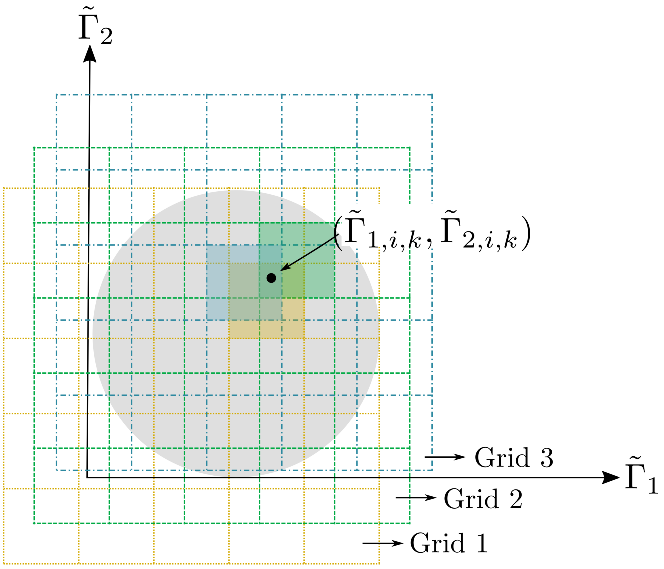


Figure 6.3. Example of tile coding in a two-dimensional state space.

[SB18]. In tile coding, the state-action space is partitioned into a grid of tiles, and multiple, overlapping, and shifted grids are used. Furthermore, each feature corresponds to a tile in a grid. As a result, the number of features is given by the product of the number t of tiles and the number G of grids. Given the set of grids, a point in the state-action space, i.e., the state-action pair $(S_i, \mathbf{X}_i)$, is described by the collection of features that are activated. In other words, the point $(S_i, \mathbf{X}_i)$ is characterized by the collection of tiles that contain it. Consequently, the corresponding action-value function $\mathrm{Q}^{\pi}(S_i, \mathbf{X}_i)$ is approximated as the weighted sum of the activated features.

Figure 6.3 shows an example of the application of tile coding when $N = 2$ EH transmitters, a single action and a single resource k are considered. The axes in Figure 6.3 correspond to the pseudo-SNR $\tilde{\Gamma}_n$ of each of the transmitters. Therefore, the state-action space has only two dimensions. In the figure, the state-action space is arbitrarily plotted as a gray circle which represents all the values the pseudo-SNR $\tilde{\Gamma}_{n,i,k}$ can take. The black dot represents a given state $S_i = (\tilde{\Gamma}_{1,i,k}, \tilde{\Gamma}_{2,i,k})$. In the example, $G = 3$ grids are considered, each of them containing $t = 25$ tiles. As a consequence, $\mathrm{Q}^{\pi}(S_i, \mathbf{X}_i)$ is approximated as the weighted sum of the three features blue, green and yellow that are activated (one feature in each of the grids). The number G of considered grids and the number t of tiles per grid determine the resolution in the approximation. The larger the number G of grids and the number t of tiles per grid, the more accurate the approximation of the state-action space. However, it should be noted that, as the resolution increases, so does the complexity of the approximation.

6.5.5. Action Selection

As illustrated by the dotted line in Figure 6.2, in every time interval i, the resource allocation solution $\mathbf{X}_i$ can be selected using the localRLPs or the globalRLP. Therefore, to determine which path to follow, i.e., select $\mathbf{X}_i$ using the localRLPs or the globalRLP, the ϵ-greedy policy is considered. This means, with probability ϵ, the localRLPs are used to select the resource allocation solution and with a probability $1-\epsilon$, we make use of the globalRLP.

As mentioned in Section 6.5.2, the action set of the globalRLP is initially empty. This means that the exploration of new possible resource allocation solutions $\mathbf{X}_i$ is done solely by the localRLPs. Furthermore, each of the K localRLPs faces the well known exploration-exploitation dilemma, i.e., whether to allocate the corresponding resource k to a transmitter N_n that has not yet used it and can potentially achieve a high throughput, or to allocate it to the transmitter that has achieved the highest throughput so far. To handle this tradeoff, we also consider the ϵ-greedy policy at each of the K localRLPs. However, to differentiate it from the previous case, we term it ϵ_{local}-greedy policy.

Since the task of the globalRLP is to learn the suitability of the resource allocation solutions that have been already tried, it considers a greedy policy. This means, every time $\mathbf{X}_i$ is selected via the globalRLP, the resource allocation $\mathbf{X}_i$ that leads to the highest value $\hat{\mathrm{Q}}^{\pi}(S_i, \mathbf{X}_i)$ of the action-value function for the considered state S_i, is selected. The use of the greedy policy enforces the exploitation of the resource allocation solution $\mathbf{X}_i$ which is considered the best up to time interval i.

6.5.6. Combinatorial SARSA algorithm

The proposed combinatorial SARSA algorithm is summarized in Algorithm 6.2. At the beginning, the learning parameters, which correspond to the discount factor γ, the learning rate ζ, the explorations probabilities ϵ and ϵ_{local}, and the weights $\mathbf{w}_k$ and $\mathbf{w}$, are initialized and the first state S_1 is observed (lines 1-3). As no action has been selected, the globalRLP does not contain any resource allocation solution $\mathbf{X}_i$ in its memory. Consequently, the first resource allocation solution $\mathbf{X}_i$ is randomly chosen through the localRLPs (line 4-7). Then, for the current state S_i, the selected action $\mathbf{X}_i$ is stored in the globalRLP (line 9-11). Afterwards, the available resources are allocated according to $\mathbf{X}_i$ and the achieved throughput is calculated (lines 12-13). After the transmission, the new state S_{i+1} is observed (line 14). Furthermore, in order to select the new action,

Algorithm 6.2 Combinatorial SARSA

```
 1: initialize $\gamma$, $\zeta$, $\epsilon$, and $\epsilon_{\text{local}}$
 2: initialize weights in the localRLPs and the globalRLP
 3: observe $S_1$
 4: for each localRLP do
 5:     randomly select action $a_{k,i}$
 6: end for
 7: collect the selected $a_{k,i}$ in the action $\mathbf{X}_i$
 8: for every $i = 1, ..., I$ do
 9:     if in state $S_1$ a new $\mathbf{X}_i$ is encountered then
10:         add it to the globalRLP
11:     end if
12:     allocate the resources according to $\mathbf{X}_i$
13:     calculate the achieved throughput $R_i^{\text{MAC}}$                ▷ Eq. (6.5)
14:     observe next state $S_{i+1}$
15:     generate random number $z$
16:     if $z \geq \epsilon(i)$ then                            ▷ select $\mathbf{X}_{i+1}$ from globalRLP
17:         select next action $\mathbf{X}_{i+1}$ with highest $\hat{\text{Q}}(S_{i+1}, \mathbf{X}_{i+1})$
18:     else                                           ▷ select $\mathbf{X}_{i+1}$ from the localRLPs
19:         for each localRLP do
20:             select action $a_{i+1,k}$ using $\epsilon_{\text{local}}$-greedy
21:         end for
22:         collect the selected $a_{i+1,k}$ in the action $\mathbf{X}_{i+1}$
23:     end if
24:     update the weights in the localRLPs                       ▷ Eq. (6.15)
25:     update the weights in the globalRLP                       ▷ Eq. (6.16)
26:     set $S_i = S_{i+1}$ and $\mathbf{X}_i = \mathbf{X}_{i+1}$
27: end for
```

a random number $0 \leq z \leq 1$ is generated to decide whether to use the localRLPs or the globalRLP (lines 15-16). In case the globalRLP is chosen, the action that yields the maximum value of the action-value function $\hat{\text{Q}}^{\pi}$ for S_{i+1}, is selected (lines 16-17). Moreover, in case the localRLPs are used, the ϵ_{local}-greedy policy is considered in each localRLP and the action $\mathbf{X}_{i+1}$ is obtained by collecting the actions $a_{k,i}$ selected by each of them (lines 18-23). Considering S_i, $\mathbf{X}_i$, R_i^{MAC}, S_{i+1} and $\mathbf{X}_{i+1}$, the weights in all the RL problems are updated using the SARSA updates in (6.15) and (6.16) (lines 24-25). At last, the values of the current state and action are updated (line 26) and the same procedure described above is repeated for as long as the receiver is operative.

6.5.7. Convergence guarantees

As explained in the previous sections, the proposed combinatorial SARSA algorithm is composed of K localRLPs and one globalRLP. Furthermore, each of them uses linear function approximation and the SARSA update in their corresponding learning processes as described in (6.15) and (6.16), respectively. As a result, the convergence of each of these learning processes is determined by the the selection of the learning rate parameters ζ as described in Section 3.5 and the considered policy. This is, if ζ satisfies the constraints in (3.34) and (3.35), and the same policy is followed throughout the execution of the algorithm, each learning process convergences to a bounded region with probability one [Gor01].

6.5.8. Computational complexity analysis

In this section, we analyze the computational complexity of the proposed combinatorial SARSA algorithm. As in the previous chapters, we evaluate the computational complexity with respect to the sizes of the action spaces, which in turn depend on the number N of EH transmitters and the number K of available resources. From Algorithm 6.2, it is clear that the most computationally demanding operations are the selection of the actions in the localRLPs and globalRLP via exploitation using the $\epsilon-$greedy and greedy policies, respectively. This is due to the fact that exploitation requires the selection of the action that leads to the maximum value of the estimated action-value function $\hat{\mathrm{Q}}^{\pi}$ for the given state. Specifically, the complexity grows as $O(N)$ for the localRLPs because the size of the action space is given by the number of EH transmitters. For the globalRLP, the size of the action space is not fixed, but increases every time that a new resource allocation solution is found via the localRLPs. Therefore, the complexity increases linearly with the minimum between the number of solutions that can be stored, i.e., the memory, and the total number of feasible resource allocation solutions. As in general the number of EH transmitters is much smaller than the number of actions available in the globalRLP, the computational complexity of the proposed combinatorial SARSA increases linearly with the size of the action space of the globalRLP.

6.6. Performance evaluation

In this section, we present numerical results to evaluate the performance of the proposed offline and learning approaches. For the simulations, the parameters listed in Table 6.2 are considered unless it is otherwise specified.

In addition to the parameters introduced in the previous chapters, which we do not describe here again for brevity, we consider TDMA, i.e., each resource k is a fraction of the time interval and all the fractions have the same length. The results are obtained by generating $T = 100$ independent random EH and channel realizations. Each

Table 6.2. Simulation set-up.

	Parameter	Value	Description
General		TDMA	Access method
	I	10000	Number of time intervals
	K	3	Number of resources
	N	10	Number of EH transmitters
	T	100	Number of realizations
	τ	1s	Time interval duration
Energy	$B_{\max,n}$	$2E_{\max,n}$	Battery capacity of EH transmitter N_n
	E_n^{Circ}	1mJ	Energy consumed by the circuit in EH transmitter N_n
	ρ	10mW/cm^2	Power density EH source
	Ω	16cm^2	Size of EH panel
Channel	f_0	2.4 GHz	Carrier frequency
	$r_{\max}$	50m	Coverage radius
	W	1 MHz	Bandwidth
	α	3	Path loss exponent
	$\Gamma_{\max}$	5 dB	Maximum possible SNR for the link between N_n and N_0
Learning	G	16	Number of grids for tile coding
	t	2	Number of tiles per grid
	γ	0.9	Discount factor
	δ	2%	Step size
	ϵ	$I/(I+4i)$	Exploration probability
	$\epsilon_{\mathrm{local}}$	$I/(I+10i)$	Exploration probability in the localRLPs
	ζ	$10G^{-1}$	Learning rate

realization is an episode where the transmitters harvest energy $I = 10^4$ times. As in the previous chapters, the amounts $E_{n,i}$ of harvested energy are taken from a uniform distribution with maximum value $E_{\max}$. We assume that the EH transmitters are uniformly distributed in a radius $r_{\max} = 50$m around N_0. Furthermore, the maximum SNR $\Gamma_{\max}$ for the link between any N_n and N_0 is set to $\Gamma_{\max} = 5$dB. To perform linear function approximation, each of the N dimensions forming the state space is divided into $t = 2$ tiles and $G = 16$ grids are considered. The following reference approaches are considered for the performance comparison:

- SARSA: A single RL problem using SARSA and linear function approximation is considered. In this case, the action space includes all the possible RA solutions, i.e., $|\mathcal{A}| = N^K$. As in the proposed combinatorial SARSA, tile coding is used as approximation technique.
- Greedy policy: In this approach, the K EH transmitters with the best channel conditions in each time interval are selected for transmission and one resource is allocated to each of them.
- Random: In this approach, K EH transmitters are randomly selected and one resource is allocated to each of them. Here, all the transmitters have the same probability of being selected.

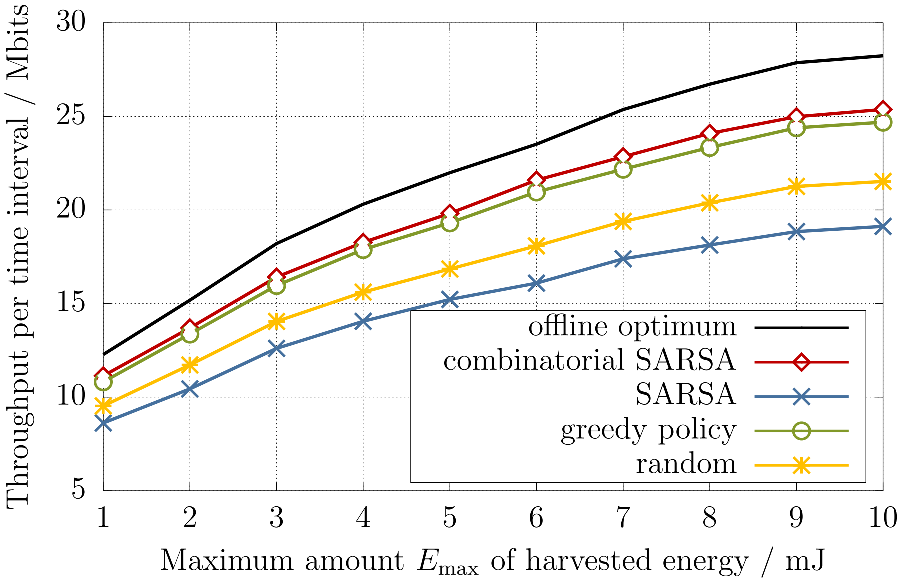


Figure 6.4. Throughput per time interval versus the maximum amount $E_{\max}$ of harvested energy.

The average throughput per time interval versus the maximum amount $E_{\max}$ of energy that can be harvested is shown in Figure 6.4. For this Figure, a simplified scenario is considered in order to be able to compute the offline optimum. In this case, $N = 3$ EH transmitters and $K = 2$ resources are assumed. Moreover, the channel is assumed to be constant for all the time intervals such that $g_{n,i,k} = 1$, $\forall n, i, k$. The maximum amount $E_{\max,\mathrm{n}}$ of harvested energy of node N_n is randomly selected from the interval $[0, E_{\max}]$. Additionally, the amounts $E_{n,i}$ of harvested energy of N_n are taken from the set $\mathcal{E} = \{0, E_{\max,\mathrm{n}}/2, E_{\max,\mathrm{n}}\}$. As expected, the throughput achieved by all the approaches increases with the amount $E_{\max}$ of harvested energy. This is because more energy is available for the transmission of data in each time interval. The highest throughput is achieved by the offline optimum at the cost of non-causal knowledge of the system dynamics. The proposed combinatorial SARSA outperforms all the reference approaches. For $E_{\max} = 5$ it achieves a performance 3%, 17% and 30% higher than SARSA, greedy policy and random approach, respectively. As the scenario has been simplified to consider a small number of EH transmitters and resources, the performance of the two learning approaches is similar. However, as shown in Figures 6.5 and 6.6, the benefits of combinatorial SARSA are better exhibited when larger values of N and K are considered.

Figure 6.5 shows the average throughput per time interval for different numbers of EH transmitters when $K = 3$ resources are considered. For all the approaches, the

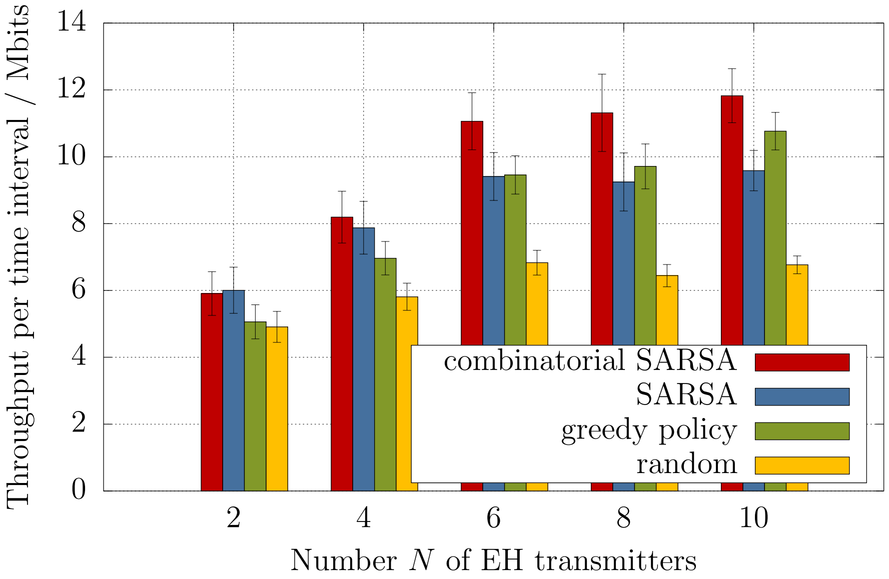


Figure 6.5. Throughput per time interval versus the number N of EH transmitters for $K = 3$ resources.

throughput increases with the number of transmitters due to the increased diversity, i.e., when more transmitters are considered, there are more possible resource allocation solutions. For $N = 2$, the proposed combinatorial SARSA performs similar to the traditional SARSA algorithm and outperforms the greedy and random approaches. However, as the network size increases, the advantages of the combinatorial SARSA are better exhibited. By breaking the original RL problem into $K + 1$ smaller RL problems, our proposed approach is able to handle the larger action spaces and consequently, achieve a higher throughput compared to SARSA. Specifically, for $N = 10$, combinatorial SARSA achieves a throughput 23% higher than SARSA and 12% and 80% higher than the greedy and random strategies, respectively.

Figure 6.6 shows the effect of the number of resources on the average throughput per time interval for $N = 10$ transmitters. It can be seen that the combinatorial SARSA algorithm achieves roughly the same throughput for the different numbers of resources. This is due to the fact that it considers the EH and channel fading process of the transmitters, which are the source of the randomness in the system, in the selection of the resource allocation solutions. As mentioned before, the SARSA approach suffers from the curse of dimensionality. Therefore, its performance degrades when more resources are considered. As a matter of fact, when the number of resources is larger than three, the action space of the SARSA approach is so large that a solution cannot be obtained in reasonable computational time. Note that the greedy strategy performs

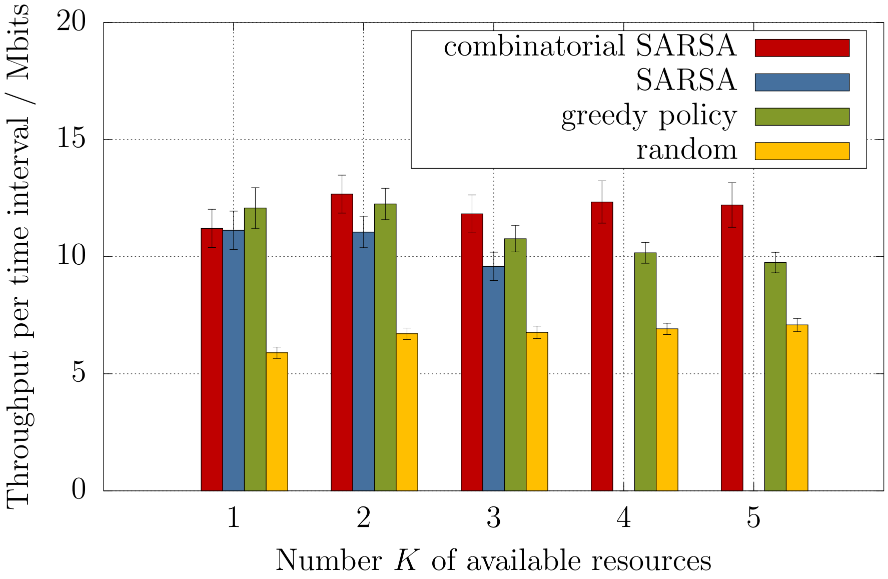


Figure 6.6. Throughput per time interval versus the number K of resources for $N = 10$ transmitters.

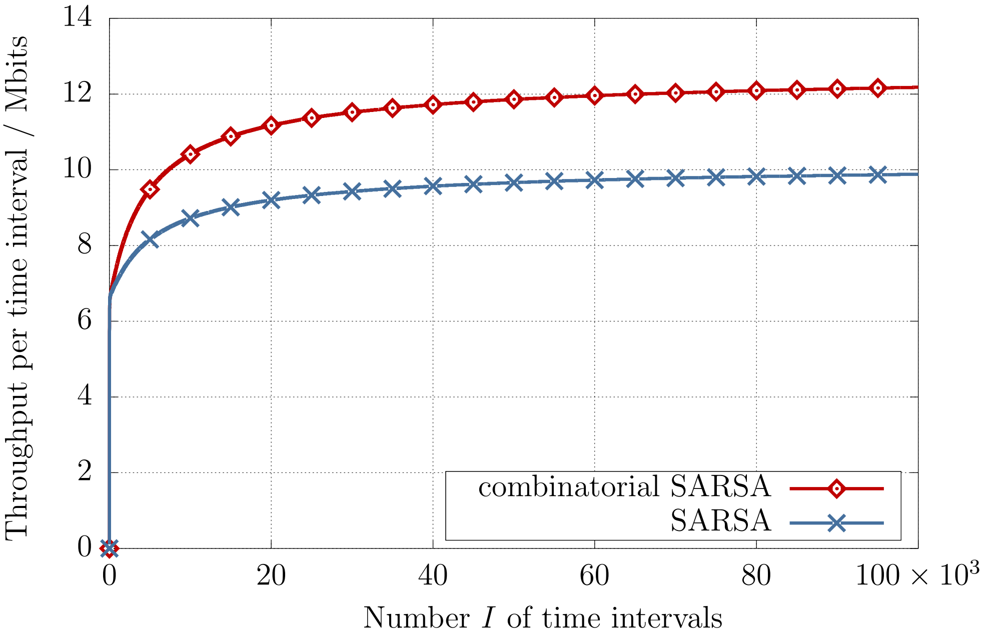


Figure 6.7. Throughput per time interval versus the number I of time intervals.

slightly better than the learning approaches when $K = 1$. This is because in this case, acting greedily is optimal. However, the learning approaches need to perform exploration in order to learn the resource allocations policy. During exploration, suboptimal resource allocation solutions may be selected which affects the average throughput. Nevertheless, as the number of available resources increases, the performance of the low-complexity approaches decreases. Specifically, when $K = 5$ resources are considered, combinatorial SARSA achieves 24% and 71% higher throughput than the greedy and random strategies.

The convergence speed of the combinatorial SARSA is evaluated in Figure 6.7 in a scenario with $N = 10$ EH transmitters and $K = 3$ available resources. From the beginning, combinatorial SARSA achieves a higher throughput compared to the SARSA. The reason for this is that it explores more efficiently the action space. Additionally, it is designed to cope with the high dimensionality of the problem in both, the state and action space, while SARSA only tackles the high dimensionality of the state space through the use of linear function approximation. Note that due to the complexity of the resource allocation problem, both approaches need a large number of iterations to converge.

6.7. Extension to the finite data buffer case

In this section, we discuss how the proposed offline and learning approaches can be extended in order to consider a finite data buffer, and the corresponding data arrival process, in each of the EH transmitters.

In addition to the EH and channel fading processes, N_0 must consider the data arrival process of the EH transmitters in order to find the optimal resource allocation solution. Due to the characteristics of the proposed approaches, this consideration only requires the inclusion of the data buffer level $D_{n,i}$ of each N_n in the definition of the state S_i. To reduce the number of considered variables, we propose to use $D_{n,i}$ as a weighting factor of the pseudo-SNR $\tilde{\Gamma}_{n,i,k}$ defined in (6.7), such that it is calculated as

$$\tilde{\Gamma}_{n,i,k} = \frac{D_{n,i}}{D_{\max,n}} \frac{g_{n,i,k} B_{n,i}}{\tau \sigma_0^2}. \tag{6.17}$$

By doing so, the suitability of the allocation of resource k to node N_n increases when the data buffer level is high and it is reduced when it does not have data to transmit.

As shown in (6.17), the inclusion of a finite data buffer in our model is straightforward. However, it increases the complexity of the proposed approaches because it enlarges the number of possible states S_i that can be encountered. For the offline approach, it also increases the number of transition probabilities $P_{S_i,S_{i+1}}^{\mathbf{X}_i}$ to be considered in the transition model P. To illustrate the growth of $\mathcal{S}$ and P, let us consider the case in which the set $\mathcal{S}$ of states is finite and let $\mathcal{E}$, $\mathcal{H}$ and $\mathcal{D}$ be the set of possible amounts of harvested energy, channel coefficients and data buffer levels, respectively. When finite data buffers are considered at the EH transmitters, the number $|\mathcal{S}|$ of possible states is given by

$$|\mathcal{S}| = \left(|\mathcal{E}||\mathcal{H}||\mathcal{D}|\right)^N, \tag{6.18}$$

which is $|\mathcal{D}|^N$ times larger than the case when infinitely full data buffers are assumed. Similarly, the number P of transition probabilities to be calculated in the offline approach is given by

$$P = |\mathcal{A}| \left(|\mathcal{E}||\mathcal{H}||\mathcal{D}|\right)^{2N}, \tag{6.19}$$

where $|\mathcal{A}|$ is the number of resource allocation solutions. This means, $|\mathcal{D}|^{2N}$ times more transition probabilities need to be computed compared to the case of infinitely full data buffers.

6.8. Conclusions

In this chapter, we have investigated offline and learning approaches for the resource allocation problem for throughput maximization in the EH multiple access communication scenario.

We have formulated the resource allocation problem and identified it as a non-linear knapsack problem which is NP-hard. In order to find the optimal solution, we have proposed an offline approach based on dynamic programming. Specifically, we have modeled the problem as an MDP and leveraged the policy iteration algorithm to find the optimal resource allocation solution assuming perfect non-causal knowledge of the system dynamics.

Additionally, we have investigated the case when only causal knowledge of the EH and channel fading processes is available. In this case, the main challenge to address is the fact that the number of resource allocation solutions increases exponentially with the number N of EH transmitters and the number K of available resources. For learning approaches, this growth of the action space translates in a reduced learning speed. To overcome this challenge, a learning algorithm, termed combinatorial SARSA, has been proposed. The proposed combinatorial SARSA handles the combinatorial nature of the resource allocation problem by breaking it into $K + 1$ smaller problems. This separation can be done by considering the characteristics of the problem, specifically, by exploiting the fact that the total throughput achieved in the system is the sum of the throughputs achieved in each available resource. As a result, a separate learning problem is formulated for each resource k in order to find, in time interval i, the EH transmitter N_n to whom said resource should be granted. Through numerical simulations we have shown that the proposed combinatorial SARSA outperforms reference schemes, including traditional learning techniques like SARSA.

Chapter 7

Conclusions

7.1. Summary

In this thesis, EH communications have been investigated. We have addressed power and resource allocation problems for throughput maximization in four different scenarios which are the main building blocks of more complicated networks. Specifically, we have considered the EH point-to-point, EH two-hop, EH broadcast and EH multiple access scenarios. Moreover, for the two-hop case, the use of decode-and-forward as well as amplify-and-forward relays in both, full-duplex and half-duplex mode, have been studied. Taking into account that the design of optimal transmission policies requires perfect non-causal knowledge of the dynamics of the system, i.e., the EH, channel fading and data arrival processes, offline approaches have been exploited in the four scenarios in order to find the upper bound of the performance in each of them. Furthermore, based on the results found using the offline approaches, we proposed novel learning algorithms that overcome the requirement of non-causal knowledge of the system dynamics and achieve a performance close to the optimum. Through extensive numerical simulations, it has been shown that the proposed learning algorithms achieve higher throughput than the reference algorithms found in the literature.

In Chapter 1, EH communications are introduced and an overview of the current state-of-the-art is presented. Based on this review of the literature, the open issues are identified and formulated. Additionally, the main contributions of this dissertation are summarized and an overview of the thesis is provided.

In Chapter 2, the system model, which comprises the EH, data arrival and channel fading models, is described. Furthermore, an introduction to Markov decision processes, value functions and linear function approximation is provided.

In Chapter 3, EH point-to-point communications are investigated. For this purpose, the considered scenario, which consists of a single EH transmitter and a single receiver, and the corresponding system assumptions are introduced. Furthermore, the power allocation problem for throughput maximization is formulated. Two cases regarding the availability of data at the EH transmitter have been considered. Specifically, we have investigated the case when infinite data is available at the transmitter and the case when the data to be transmitted is the result of a data arrival process. The resulting optimization problem is a convex problem. However, we have shown that a closed-form solution for the power allocation policy cannot be found because in every time interval, the optimal power allocation depends on the powers to be allocated in future time intervals. Despite this fact, we have characterized the offline optimal solution by means of the derivation of the KKT conditions. From the analysis of the KKT conditions and based on results from the literature, it is shown that in the optimal power allocation policy, the transmit power should be kept constant during one time interval. Moreover, it is shown that when an infinite battery is considered and an infinite amount of data is available at the transmitter, the transmit power increases monotonically over time. Exploiting the results found in the offline approach, a learning approach is proposed in order to overcome the requirement of non-causal knowledge regarding the EH, data arrival and channel fading processes. The proposed learning approach, termed approximated SARSA, models the throughput maximization problem as a Markov decision process and uses this model to learn the optimal power allocation policy by deciding, in each time interval, the transmit power to use and evaluating the resulting throughput. Additionally, we have proposed the use of linear function approximation in order to handle the infinite values the amounts of energy and data, and the channel coefficients can take. For the linear function approximation, a set of four original feature functions, which are based on the insights gained through the characterization of the offline optimum solution, are proposed. Additionally, exploiting results found in the RL literature, convergence guarantees for the proposed learning approach are provided. It has been demonstrated that the complexity of the proposed approximate SARSA algorithm grows only linearly with the number of possible transmit power values. Finally, through numerical simulations, it is shown that the proposed learning approach has a performance close to the offline optimum and outperforms reference schemes found in the literature. Specifically, a performance up to 50% higher than for Q-learning is achieved.

In Chapter 4, the EH two-hop communication scenario is studied. Two types of relays are considered, namely, a decode-and-forward and an amplify-and-forward relay. For these two cases, the scenario and corresponding system assumptions are described and offline and learning approaches are investigated. Similar to the point-to-point case, we have shown that the use of a decode-and-forward relay in the two-hop scenario leads to a convex throughput maximization problem. However, a closed-form solution cannot be found. Nonetheless, by analysing the corresponding KKT conditions, the dependency between the power allocation policies at the transmitter and at the relay has been established. Specifically, it is shown that in an offline setting where the batteries and data buffers have infinite capacities, if the data buffer of the transmitter is depleted at any given time interval i, i.e., all the data is transmitted to the relay, then the data buffer at the relay has to be depleted at least once in the following time intervals j, $i + 1 \leq j \leq I$, where I is the total number of time intervals. Based on these results, two novel learning approaches that consider different levels of cooperation between the transmitter and the relay are proposed. The first approach, termed independent SARSA, assumes the transmitter and the relay have no knowledge regarding the EH, data and channel fading processes associated to the other node and aim at maximizing the throughput on their own links. The second approach, termed cooperative SARSA, proposes the use of a signaling phase in which the nodes exchange their current battery and data buffer levels as well as their channel conditions. It is shown that the use of the signaling phase leads to a gain in throughput of up to 40%, compared to the no cooperation case, even if part of the time interval is dedicated to the signaling and not fully to the transmission of data. This is because by knowing the battery and data buffer levels, and channel conditions of the other nodes, the transmitter and relay can adapt their own transmission strategies in order to maximize the amount of data transmitted to the receiver. In contrast to the decode-and-forward case, when an amplify-and-forward relay is considered, the resulting throughput maximization problem is non-convex. Therefore, to find the optimal power allocation policy in this scenario, we have proposed an offline approach based on the reformulation of the objective function as the difference between two concave functions. This proposed reformulation facilitates the design of a branch-and-bound algorithm which finds the optimum solution. In the amplify-and-forward case, the communication between the transmitter and the receiver cannot be separated, but has to be considered as a single link with an effective channel that depends on the channel from the transmitter to the relay, the relay gain and the channel from the relay to the receiver. Consequently, we have proposed a centralized learning algorithm. By considering this effective channel, it is shown that the learning approach designed for the EH point-to-point case can be adjusted for this two-hop scenario by including a signaling phase in which the current values of the battery and data buffer levels, and the channel conditions are transmitted. Furthermore, through numerical simulations it is shown that the proposed centralized learning approach achieves a performance up to two times higher than the performance achieved by the reference schemes. The chapter closes with the description of how the proposed learning approaches can be extended to consider other relaying scenarios such as an EH multi-hop communication scenario with a single transmitter and a single receiver, and a EH multi-node multi-hop communication scenario with multiple transmitter-receiver pairs.

In Chapter 5, EH communications in a broadcast scenario are investigated. To this aim, a single EH transmitter which sends individual data to multiple receivers is considered and the corresponding power allocation problem for throughput maximization is formulated. It is shown that the resulting optimization problem is non-convex when an arbitrary number of receivers is considered. However, exploiting existing results from the literature, an offline approach is presented for the special case of two receivers. Furthermore, inspired by the structure of the offline optimal solution, a novel learning approach is proposed for the case when an arbitrary number of receivers are considered. The proposed learning approach, termed two-stage SARSA, separates the learning task into two sub-tasks, namely, the selection of the total power to use in each time interval and the distribution of this selected power for the transmission of the data intended for the different receivers. Through numerical simulations, it is shown that by splitting the learning task, the proposed learning approach is able to achieve a throughput up to 40% higher than for conventional learning algorithms. Moreover, by considering not only the channel coefficients, but also the battery level and the different data buffer levels, it is able to serve more receivers while achieving a higher sum throughput compared to reference schemes.

In Chapter 6, an EH multiple access scenario is investigated. In this case, the focus is on the resource allocation problem of K orthogonal resources among N EH transmitters that want to transmit data to a single receiver. The resulting throughput maximization problem is identified as a non-linear knapsack problem which is NP-hard. As a result, we have proposed an offline approach based on dynamic programming to find the offline optimum solution when perfect non-causal knowledge of the system dynamics is available. Specifically, the policy iteration algorithm is tailored to the EH multiple access scenario. Furthermore, we have proposed a novel learning approach in order to overcome the requirement of non-causal knowledge of the system dynamics and to address the combinatorial nature of the problem. In particular, the proposed learning approach, which is termed combinatorial SARSA, exploits the characteristics of the scenario to separate the original problem into $K+1$ smaller problems, thus, breaking the exponential growth of the space of possible solutions. Through numerical simulations we have shown that the proposed combinatorial SARSA achieves a performance up to 25% higher than the performance achieved by a greedy policy which allocates the available resources to the users with the best channel conditions.

7.2. Outlook

In this thesis we have focused on four different scenarios, namely, point-to-point, two-hop, broadcast and multiple access, which are the main building blocks of more complicated networks. As a result, the natural extension of this work is to consider larger networks formed by a mix of the basic scenarios addressed in this dissertation, i.e., networks consisting of multiple EH transmitters, EH relays and receivers. Furthermore, interesting research questions arise when EH harvesting is considered in conventional communication scenarios such as two-way relaying communications, multi-way communications, communication trees, multicasting scenarios and multi-hop scenarios with one or more relays. The approaches developed in this thesis lay the ground work for such extensions.

Throughout this thesis, we have assumed that only the transmitters and relays harvest energy from the environment while the receiver nodes have been assumed to be connected to a fixed power supply. However, it is an interesting research direction to consider that the receivers are also harvesting energy and are thus not always available to receive the transmitted data. When EH receivers are assumed, a careful model of the energy consumed while receiving must be developed. This model is, in general, dependent on the particular hardware that is considered. Therefore, a suitable model that describes the energy consumption at the receiver side is needed. Moreover, the design of transmission strategies should consider the EH processes of the transmitters and receivers jointly in order to find the optimal solution. As a result, special attention should be given to the signaling required between the nodes.

Another assumption made in this thesis is that the batteries are ideal. This is, no energy is lost while storing or retrieving energy from the battery. Although some works have already shed some light on the repercussion of such imperfections, they have mainly focused on offline solutions [DG12, TY12b, TY12a, BZ15]. Therefore, more work is necessary for the case of learning approaches. The main challenge in this research direction is that when learning approaches are considered, such imperfections bring additional randomness to the system to which the transmission policy should be able to adapt. This means, the learning speed of the algorithm should be sufficiently fast in order to cope with the variability in the system.

Furthermore, in this thesis, we have investigated scenarios in which the harvested energy is used solely for the transmission of data, i.e., the transmit power and the power consumed by the circuit. Considering the current trends in the wireless communication field, another possible extension of this thesis is the combination of EH with other existing technologies such as computation offloading, mobile edge computing, caching, among others. Such combination allows the exploitation of the benefits of EH in other scenarios. Nevertheless, it also brings additional constraints to the optimization problems. For example, in the context of EH computation offloading and EH mobile edge computing, the decision whether to offload a task to the server or to compute it locally, not only depends on the computational capabilities of the node, but also on the availability of energy. Furthermore, it also requires a model of the energy consumption that depends on the task at hand. Similarly, in caching applications considering EH cache servers, the energy required to store the file in memory and to send it to the interested receivers should be considered in addition to the conventional metrics, e.g., the popularity of the file.

Additionally, the offline and learning approaches developed in this thesis can be used as a baseline for throughput maximization problems in other applications. The author of this thesis has already started to work in this direction by considering learning approaches for different applications. Specifically, a learning approach for the mode selection and resource allocation problem for throughput maximization in device-to-device (D2D) systems has been proposed in [OAE+19]. Furthermore, a semi-distributed learning approach to minimize the end-to-end latency as well as to enhance the robustness against network dynamics in a self-backhauling millimeter wave scenario is investigated in [OAS+19]. Additionally, a dynamic programming approach for optimal resource allocation in multi-rate opportunistic forwarding is proposed in [HOK19].

Finally, an important aspect not covered in this thesis is the implementation of the developed algorithms in a test bed. This requires finding suitable hardware that is sufficiently flexible to enable the programming of the developed solutions. Moreover, practical aspects such as the type and architecture of the microprocessors, the availability of different types of EH modules and synchronization requirements have to be taken into account.

List of Acronyms

i.i.d.	Independent and identically distributed
ACF	Auto correlation function
AF	Amplify-and-forward
AWGN	Additive white Gaussian noise
BC	Broadcast
DC	Difference of concave functions
DF	Decode-and-forward
DP	Dynamic programming
EH	Energy harvesting
FDMA	Frequency division multiple access
FSR	Fixed sparse representation
GP	Glue pouring
IoT	Internet of Things
KKT	Karush-Kuhn-Tucker
LB	Lower bound
MAB	Multi-Armed Bandit problem
MAC	Multiple Access
MDP	Markov decision process
PL	Path loss
RBF	Radial basis functions
RL	Reinforcement learning
SARSA	State Action Reward State Action
SIC	Successive interference cancellation
SNR	Signal to noise ratio

TDMA	Time division multiple access
UB	Upper bound
WF	Water filling

List of Symbols

$(\mathbf{x})^{\mathrm{H}}$	Conjugate transpose of vector $\mathbf{x}$
$(\mathbf{x})^{\mathrm{T}}$	Transpose of vector $\mathbf{x}$
$\lvert X \rvert$	Magnitude of the complex number X
$\lvert \mathcal{X} \rvert$	Cardinality of the set $\mathcal{X}$
$\lceil x \rceil$	Rounding operation to the nearest integer greater than or equal to x
$\lfloor x \rfloor$	Rounding operation to the nearest integer less than or equal to x
$\mathbf{0}_X$	Vector of zeros with length X
$\mathbf{1}(x)$	Indicator function, it is equal to one when the event x is true and zero otherwise.
a_i	Action selected in time interval i
$\mathbf{a}_n$	Vector containing the ψ_n parameter of the autoregressive process of EH node N_n
A	Size of the action set $\mathcal{A}$
$\mathcal{A}$	Set of actions
$B_{\mathrm{max},n}$	Battery capacity of EH node N_n
$B_{n,i}$	Battery level of EH node N_n, measured at the beginning of time interval i
c_0	Speed of light
$c_{n,j}$	j^{th} parameter of the autoregressive process at EH node N_n
$\mathbf{C}_n$	Matrix containing the parameters of the autoregressive process at EH node N_n
$\mathbb{C}$	Set of complex numbers
d	Data packet size in bits
$D_{\mathrm{max},n}$	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
$e_{\mathrm{quant},u_{n,i}}$	Tolerable quantization error for parameter $u_{n,i}$
$E_{\mathrm{max},n}$	Maximum amount of energy that can be harvested by 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
$E_{n,i}^{\mathrm{Sig}}$	Energy consumed by node N_n for signaling
$E_{n,i}^{\mathrm{Tx}}$	Energy of the signal transmitted by EH node N_n in time interval i
$\mathbb{E}\left[\cdot\right]$	Expected value operator

f	Index of the feature function
f_0	Carrier frequency of the transmitted signal
$f_{D,\max}$	Maximum Doppler frequency
$\mathrm{f}_f(S_i, a_i)$	f^{th} feature function
$\mathbf{f}$	Vector containing the feature values of a given action-pair
F	Number of feature functions
$g_{n,i}$	Channel gain of the link between nodes N_n and N_{n+1} in time interval i
$\bar{g}_{n,i}$	Estimated mean value of the channel gain of the link between nodes N_n and N_{n+1} in time interval i
$\hat{g}_{n,i}$	Estimated the channel gain of the link between nodes N_n and N_{n+1} in time interval i
G	NUmber of grids
G_n	Antenna gain of node N_n
$h_{n,i}$	Complex channel coefficient of the link between N_n and N_{n+1} in time interval i
$\hat{h}_{n,i}$	Estimate of the complex channel coefficient of the link between N_n and N_{n+1} in time interval i
$\mathbf{h}_{n,i}$	Vector containing the past o channel coefficients of node N_n in time interval i
$\mathcal{H}$	Discrete set of channel gains
i	Time interval index
I	Maximum number of time intervals
$\mathbf{I}_X$	Identity matrix of size X
j	Auxiliary time interval index
J	Number of time intervals, with $J \leq I$
J_0	Zero$^{\text{th}}$ order Bessel function of the first kind
$\mathcal{J}$	Set of power sharing parameters η_n
k	Index of the resources
K	Number of available resources
$\mathbf{k}_{n,i}$	Kalman gain at node N_n in time interval i
$\mathbf{L}_X$	Lower diagonal matrix of ones with size $X \times X$
$\mathfrak{L}$	Lagrange function
m	Auxiliary index for the nodes
$M_{n,i}$	Amount of incoming data, arriving at the end of time interval i, at EH node N_n
n	Index for the nodes

N	Maximum amount of nodes in the scenario
N_n	n^{th} node
o	Order of the autoregressive process for the modeling of the channel coefficients
$O(\cdot)$	Order of the computational complexity
$p_{n,i}^{\text{Circ}}$	Power consumed by the circuit in N_n in time interval i
$p_{n,i}^{\text{Sig}}$	Üower consumed by N_n for sending the signaling in time interval i
$p_{n,i}^{\text{Rx}}$	Power of the received signal at node N_n in time interval i
$p_{n,i}^{\text{Tx}}$	Transmit power of N_n in time interval i
$P_{S_i,S_{i+1}}^{a_i}$	Transition probability from state S_i to state S_{i+1} after selecting a_i
P	Transition model
$\mathbb{P}\left[x\right]$	Probability of event x
$\mathrm{q}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\text{Tx}})$	Local action-value function at node N_n
$\hat{\mathrm{q}}_{n,i}^{\pi_n}(S_i, p_{n,i}^{\text{Tx}})$	Approximated local action-value function at node N_n
$\mathrm{Q}^{\pi}(S_i, a_i)$	Action-value function
Q^{*}	Optimal action-value function
$\hat{\mathrm{Q}}^{\pi}(S_i, a_i)$	Approximated action-value function
r	Distance between two communicating nodes N_n and N_{n+1}
r_0	Coverage radius
$R_{n,i}$	Amount of data transmitted by N_n in time interval i
R_i	Reward obtained in time interval i
$\mathbb{R}$	Set of real numbers
$\mathcal{R}$	Set of rewards
S_i	State experienced in time interval i
$\mathcal{S}$	Set of states
t	Number of tiles
T	Number of realizations
$u_{n,i}$	Variable used to represent any parameter associated with N_n in time interval i
$v_{j,k}$	k^{th} element of the vertex vector $\mathbf{v}_l$
$\mathbf{v_j}$	j^{th} vertex vector
$V_{\max,u_{n,i}}$	Maximum value parameter $u_{n,i}$ can take
$V_{\min,u_{n,i}}$	Minimum value parameter $u_{n,i}$ can take
$\mathrm{V}^{\pi}(S_i)$	State-value function
V^{*}	Optimal state-value function
$w_{n,i}$	Receiver noise plus interference at N_n in time interval i

$\mathbf{w}$	Vector of weights containing the contributions of each feature function
W	Bandwidth
$x_{n,i}$	Transmitted signal from node N_n in time interval i
$\mathbf{x}_n$	Vector containing the transmitted symbol in the autoregressive process of EH node $\mathbf{N}_n$
$\mathbf{X}_i$	Resource allocation solution in time interval i
$y_{n,i}$	Received signal at node N_n in time interval i
$z_{n,i}$	Additive white Gaussian noise at node N_n in time interval i
$Z_{u_{n,i}}$	Number of bits required for the transmission of parameter $u_{n,i}$
$\mathbb{Z}$	Set of natural numbers
α	Path loss exponent
β	Data buffer size factor
γ	Discount factor
δ	Step size
ϵ	Probability of exploring new actions in the ϵ-greedy policy
ε	Tolerance for the branch-and-bound algorithm
ζ_i	Learning rate in time interval i
η_i	Power sharing parameter in time interval i for the EH broadcast scenario
θ	Relay amplification factor
ϑ	Weights to represent any point inside a simplex as a function of the vertices
$\kappa_{n,i}$	Lagrange multiplier associated to data causality constraints of EH node N_n in time interval i
λ	Average number of packets arriving in one time interval
$\mu_{n,i}$	Lagrange multiplier associated to energy causality constraint of EH node N_n in time interval i
ν_i	Water level in time interval i
$\xi_{n,i}$	Lagrange multiplier associated to data buffer overflow constraint of EH node N_n in time interval i
π	Policy
π^*	Optimal policy
$\pi(S_i)$	Action to be selected when state S_i is encountered and the policy is deterministic
$\pi(a_i\|S_i)$	Probability of selecting action a_i when state S_i is encountered and the policy is stochastic
ρ	Average power density of the EH source

ϱ	Fraction of a time interval
σ_n^2	Noise power of node N_n
ς	Battery size factor
τ	Time interval duration
τ^{Data}	Duration of the data transmission phase
τ^{Sig}	Duration of the signaling phase
$\upsilon_{n,i}$	Lagrange multiplier associated to the transmit power of EH node N_n in time interval i
ϕ	Priority for the receiver nodes N_n in the EH broadcast scenario
φ	Artificial variable for the branch-and-bound algorithm
$\chi_{n,i,k}$	Binary variable indicating if resource k is allocated to N_n in time interval i
ψ	Parameter for the autoregressive process of EH node N_n
$\omega_{n,i}$	Lagrange multiplier associated to battery overflow constraint of EH node N_n in time interval i
Γ	Average signal to noise ratio (SNR) of a link
$\bar{\Gamma}$	Average signal to noise ratio (SNR) of all the links associated to all the receivers
$\tilde{\Gamma}$	Pseudo-SNR of a link
Δ	Prelog factor depending on the relay's transmission mode
Ω	Size of the EH panel

Bibliography

[3GP17] 3GPP, "3rd generation partnership project; technical specification group radio access network; study on new radio access technology physical layer aspects (release 14)," 3GPP TR 38.802, Tech. Rep., 2017.

[AD15] 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.

[AD16] ——, "Online power control optimization for wireless transmission with energy harvesting and storage," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4888–4901, July 2016.

[AIM10] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Comput. Networks*, vol. 54, no. 15, pp. 2787–2805, October 2010.

[AINS13] I. Ahmed, A. Ikhlef, D. W. K. Ng, and R. Schober, "Optimal resource allocation for energy harvesting two-way relay systems with channel uncertainty," in *Proc. IEEE Global Conf. Signal Inform. Process. (GlobalSIP)*, Austin, December 2013, pp. 345–348.

[ASSC02] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Comput. Networks*, vol. 38, no. 4, pp. 393–422, Dece 2002.

[ASW12] H. Al-Shatri and T. Weber, "Achieving the maximum sum rate using d.c. programming in cellular networks," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1331–1341, March 2012.

[AUBE11] M. A. Antepli, E. Uysal-Biyikoglu, and H. Erkal, "Optimal packet scheduling on an energy harvesting broadcast link," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1721–1731, September 2011.

[BBSE10] L. Buşoniu, R. Babuška, B. D. Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Linear Function Approximation.* CRC Press, 2010.

[Bel54] R. Bellman, "The theory of dynamic programming," *Bull. Amer. Math. Soc.*, vol. 60, no. 6, pp. 503–515, July 1954.

[Ber07] D. Bertsekas, *Dynamic Programming and Optimal Control Vol. 2*, 3rd ed. A, 2007, vol. 2.

[BG15] P. Blasco and D. Gündüz, "Multi-access communications with energy harvesting: A multi-armed bandit model and the optimality of the myopic policy," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 585–597, March 2015.

[BGD13] 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.

[BK13] Y. Barayan and I. Kostanic, “Performance evaluation of proportional fairness scheduling in LTE,” in *Proc. World Congr. Eng. Comput. Sci. (WCECS)*, San Francisco, Oct. 2013, pp. 1–6.

[BS02] K. M. Bretthauer and B. Shetty, “The non-linear knapsack problem - algorithms and applications,” *European J. Operational Research*, vol. 138, no. 3, pp. 459–472, 2002.

[BU16] A. Baknina and S. Ulukus, “Online scheduling for energy harvesting broadcast channels with finite battery,” in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Barcelona, July 2016, pp. 1–5.

[BZ15] A. Biason and M. Zorzi, “Energy harvesting communication system with ”soc”-dependent energy storage losses,” in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Brussels, August 2015, pp. pp. 406–410.

[CLRS09] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.

[CT06] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, 2006.

[CZ04] 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.

[Dau05] F. Daum, “Nonlinear filters: Beyond the kalman filter,” *IEEE Aerospace and Electronic Syst. Mag.*, vol. 20, no. 8, pp. 57–69, August 2005.

[DG12] B. Devillers and D. Gündüz, “A general framework for the optimization of energy harvesting communication systems with battery imperfections,” *J. Commun. Networks*, vol. 14, no. 2, pp. 130–139, July 2012.

[DLF18] M. Dong, W. Li, and Fat, “Online joint power control for two-hop wireless relay networks with energy harvesting,” *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 463–478, January 2018.

[DP10] W. Dargie and C. Poellabauer, *Fundamentals of Wireless Sensor Networks: Theory and Practice.* John Wiley and Sons, 2010.

[ECC19] ECC, “Electronic Communications Committee (ECC) within the European Conference of Postal and Telecommunications Administrations (CEPT). The european table of frequency allocations and applications in the frequency range 8.3 kHz to 3000 GHz (ECA TABLE),” March 2019.

[EOUB13] H. Erkal, F. M. Ozcelik, and E. Uysal-Biyikoglu, "Optimal offline broadcast scheduling with an energy harvesting transmitter," *EURASIP J. Wireless Commun. and Networking*, vol. 2013, no. 1, pp. 1–20, July 2013.

[FAUC16] M. Fu, A. Arafa, S. Ulukus, and W. Chen, "Delay minimal policies in energy harvesting broadcast channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[GD11] 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.

[GGV16] M. Gregori and J. Gómez-Vilardebo, "Online learning algorithms for wireless energy harvesting nodes," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[GKU16] B. Gurakan, O. Kaya, and S. Ulukus, "Energy harvesting cooperative multiple access channel with data arrivals," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[Gor01] G. J. Gordon, "Reinforcement learning with function approximation converges to a region," in *Advances Neural Inform. Process. Syst. (NIPS)*. Denver: MIT Press, 2001, pp. 1040–1046.

[GSMZ14] 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.

[GWT$^+$13] 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.

[GYGP13] 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.

[HD16] 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.

[HOK19] F. Hohmann, A. Ortiz, and A. Klein, "Optimal resource allocation policy for multi-rate opportunistic forwarding," in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC)*, Marrakech, April 2019, pp. 1–6.

[HPT00] R. Horst, P. M. Pardalos, and N. V. Thoai, *Introduction to global optimization. Nonconvex optimization and its applications.* Kluwer Academic Publishers, 2000.

[Ins17] N. Instruments, "National Instruments Specification USRP-2954," May 2017. [Online]. Available: http://www.ni.com/pdf/manuals/375725c.pdf

[Jak74] W. C. Jakes, *Microwave Mobile Communications.* Wiley-IEEE Press, 1974.

[JE15] J. Jeon and A. Ephremides, "On the stability of random multiple access with stochastic energy harvesting," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 571–584, March 2015.

[KLCL16] 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.

[KM14] M. B. Khuzani and P. Mitran, "On online energy harvesting in multiple access communication systems," *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1883–1898, January 2014.

[KPP04] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack Problems.* Springer Berlin Heidelberg, 2004.

[Küh11] 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.

[LDC16] J. Liu, H. Dai, and W. Chen, "On throughput maximization of time division multiple access with energy harvesting users," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2457–2470, April 2016.

[LG01] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels - part i: Ergodic capacity," *IEEE Trans. Wireless Commun.*, vol. 47, no. 3, pp. 1083–1102, March 2001.

[Lit94] 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.

[Lit01] ——, "Value-function reinforcement learning in Markov games," *J. Cognitive Syst. Research*, vol. 2, no. 1, pp. 55–66, October 2001.

[Liu16] Y. Liu, "Wireless information and power transfer for multirelay-assisted cooperative communication," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 784–787, April 2016.

[LLW$^+$17] X. Li, D. Li, J. Wan, A. V. Vasilakos, C.-F. Lai, and S. Wang, "A review of industrial wireless networks in the context of industry 4.0," *Wireless Networks*, vol. 23, no. 1, pp. 23–41, Jan 2017. [Online]. Available: https://doi.org/10.1007/s11276-015-1133-7

[LOAS$^+$17] X. Li, A. Ortiz, H. Al-Shatri, A. Klein, and T. Weber, "Delay-constrained data transmission for energy harvesting transmitter," in *Proc. Int. ITG Conf. Syst. Commun. and Coding (SCC)*, Hamburg, February 2017, pp. 1–6.

[LR00] 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.

[LYG09] 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.

[LZL13a] S. Luo, R. Zhang, and T. J. Lim, “Optimal save-then-transmit protocol for energy harvesting wireless transmitters,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1196–1207, March 2013.

[LZL13b] 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.

[MS05] 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.

[MSA14] 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.

[NM93] F. Neeser and J. Massey, “Proper complex random processes with applications to information theory,” *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1293–1302, July 1993.

[OAE$^+$19] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, and M. Hollick, “Cbmos: Combinatorial bandit learning for mode selection and resource allocation in d2d systems,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2225–2238, October 2019.

[OAS$^+$19] A. Ortiz, A. Asadi, G. H. A. Sim, D. Steinmetzer, A. Klein, and M. Hollick, “Scaros: A scalable and robust self-backhauling solution for highly dynamic networks,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 1–15, Dic 2019.

[OASL$^+$15] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, “Throughput maximization in two-hop energy harvesting communications,” in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Brussels, August 2015, pp. 291–295.

[OASL$^+$16a] ——, “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.

[OASL$^+$16b] ——, “Reinforcement learning for energy harvesting point-to-point communications,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[OASL$^+$17] ——, "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.

[OE12] 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.

[OE13] ——, "Throughput maximization for energy harvesting two-hop networks," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Istanbul, July 2013, pp. 1596–1600.

[OE15] ——, "Energy harvesting two-hop communication networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2658–2670, December 2015.

[OGE12] O. Orhan, D. Gündüz, and E. Erkip, "Throughput maximization for an energy harvesting communication system with processing cost," in *Proc. Inform. Theory Workshop (ITW)*, Lausanne, September 2012, pp. 84–88.

[OGE13] ——, "Delay-constrained distortion minimization for energy harvesting transmission over a fading channel," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Istanbul, July 2013, pp. 1794–1798.

[Ont13] S. Ontañón, "The combinatorial multi-armed bandit problem and its application to real-time strategy games," in *Proc. AAAI Conf. Artificial Intell. and Interactive Digital Entertainment AIIDE* , Boston, October 2013, pp. 2471–2478.

[OTY$^+$11] 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.

[OWK18] A. Ortiz, T. Weber, and A. Klein, "A two-layer reinforcement learning solution for energy harvesting data dissemination scenarios," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Apr. 2018, pp. 6648–6652.

[OWK19] ——, "Resource allocation in energy harvesting multiple access scenarios via combinatorial learning," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.

[OYU13] O. Ozel, J. Yang, and S. Ulukus, "Optimal transmission schemes for parallel and fading gaussian broadcast channels with an energy harvesting rechargeable transmitter," *Comput. Comm.*, vol. 36, no. 12, pp. 1360–1372, July 2013.

[Rip96] B. D. Ripley, *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[RN10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall, 2010.

[RWWZ09] T. Riihonen, S. Werner, R. Wichman, and E. B. Zacarias, “On the feasibility of full-duplex relaying in the presence of loop interference,” in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Perugia, June 2009, pp. 1–5.

[SB18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2018.

[Skl17] B. Sklar, *Mobile Communications Handbook.* CRC Press, 2017.

[SKN17] S. Shresthamali, M. Kondo, and H. Nakamura, “Adaptive power management in solar energy harvesting sensor node using reinforcement learning,” *ACM Trans. Embed. Comput. Syst.*, vol. 16, no. 5s, pp. 1–21, September 2017. [Online]. Available: http://doi.acm.org/10.1145/3126495

[SW15] 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.

[SZ16] F. K. Shaikh and S. Zeadallyc, “Energy harvesting in wireless sensor networks: A comprehensive review,” *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 1041–1054, March 2016.

[TV05] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[TVY13] K. Tutuncuoglu, B. Varan, and A. Yener, “Optimum transmission policies for energy harvesting two-way relay channels,” in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Budapest, June 2013, pp. 586–590.

[TY12a] K. Tutuncuoglu and A. Yener, “Energy harvesting broadcast channel with inefficient energy storage,” in *Proc. Asilomar Conf. Signals, Syst. Computers*, Pacific Grove, November 2012, pp. 53–57.

[TY12b] ——, “Optimal power policy for energy harvesting transmitters with inefficient energy storage,” in *Proc. Annual Conf. Inform. Sciences Syst. (CISS)*, Princeton, March 2012, pp. 1–6.

[TY12c] ——, “Optimum transmission policies for battery limited energy harvesting nodes,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, March 2012.

[TZW14] 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.

[UYE$^+$15] 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.

[VY13] 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.

[WAW15] Z. Wang, V. Aggarwal, and X. Wang, "Iterative dynamic water-filling for fading multiple-access channel with energy harvesting," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 382–395, March 2015.

[XHNY15] Y. Xiao, Z. Han, D. Niyato, and C. Yuen, "Bayesian reinforcement learning for energy harvesting communication systems with uncertainty," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, June 2015, pp. 5398–5403.

[XZ14] J. Xu and R. Zhang, "Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 322–332, February 2014.

[YMZM08] P. Youssef-Massaad, L. Zheng, and M. Medard, "Bursty transmission and glue pouring: on wireless channels with overhead costs," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5188–5194, December 2008.

[YOU12] J. Yang, O. Ozel, and S. Ulukus, "Broadcasting with an energy harvesting rechargeable transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 571–583, February 2012.

[YU12a] J. Yang and S. Ulukus, "Optimal packet scheduling in a multiple access channel with energy harvesting transmitters," *J. Commun. and Networks*, vol. 14, no. 2, pp. 140–150, April 2012.

[YU12b] ——, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, January 2012.

[YW15] J. Yang and J. Wu, "Online throughput maximization in an energy harvesting multiple access channel with fading," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Hong Kong, June 2015, pp. 2727–2731.

[YZGK10] 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.

[ZBM15] 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.

[ZHC$^+$15] D. Zhao, C. Huang, Y. Chen, F. Alsaadi, and S. Cui, "Resource allocation for multiple access channel with conferencing links and shared renewable energy sources," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 423–437, March 2015.

Author's Publications

[HOK19] F. Hohmann, A. Ortiz, and A. Klein, "Optimal resource allocation policy for multi-rate opportunistic forwarding," in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC)*, Marrakech, April 2019, pp. 1–6.

[LOAS$^+$17] X. Li, A. Ortiz, H. Al-Shatri, A. Klein, and T. Weber, "Delay-constrained data transmission for energy harvesting transmitter," in *Proc. Int. ITG Conf. Syst. Commun. and Coding (SCC)*, Hamburg, February 2017, pp. 1–6.

[MNK$^+$17] C. Meurisch, T. Nguyen, M. Kromm, A. Ortiz, R. Mogk, and M. Mühlhäuser, "Disvis 2.0: Decision support for rescue missions using predictive disaster simulations with human-centric models," in *Proc. Int. Conf. Computer Commun. and Networks (ICCCN)*, Calgary, July 2017, pp. 1–5.

[OAE$^+$19] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, and M. Hollick, "Cbmos: Combinatorial bandit learning for mode selection and resource allocation in d2d systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2225–2238, October 2019.

[OAS$^+$19] A. Ortiz, A. Asadi, G. H. A. Sim, D. Steinmetzer, A. Klein, and M. Hollick, "Scaros: A scalable and robust self-backhauling solution for highly dynamic networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 1–15, Dic 2019.

[OASL$^+$15] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Throughput maximization in two-hop energy harvesting communications," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Brussels, August 2015, pp. 291–295.

[OASL$^+$16a] ——, "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.

[OASL$^+$16b] ——, "Reinforcement learning for energy harvesting point-to-point communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, May 2016, pp. 1–6.

[OASL$^+$17] ——, "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.

[ODK15] A. Ortiz, H. Degenhardt, and A. Klein, "A resource requirement aware transmit strategy for non-regenerative multi-way relaying," in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC)*, New Orleans, March 2015, pp. 19–24.

[OWK18] A. Ortiz, T. Weber, and A. Klein, "A two-layer reinforcement learning solution for energy harvesting data dissemination scenarios," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Apr. 2018, pp. 6648–6652.

[OWK19] ——, "Resource allocation in energy harvesting multiple access scenarios via combinatorial learning," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.

Supervised Student Theses

Name	Title of the thesis	Thesis type	Date
Zacharias, Jithin	Reinforcement Learning in Energy Harvesting Two-receiver Broadcast Channel	Master thesis	06/2016
Geneshki, Nikola	Two-Hop Half-Duplex Energy Harvesting Throughput Enhancement based on Learning Approach	Master Thesis	02/2017
Hajrizaj, Dardan	Suboptimal Approaches for Energy Harvesting Point to Point Communications	Master Thesis	03/2017
Yilmaz, Burak Mustafa	Learning approaches for computational offloading	Master Thesis	Ongoing

Funding Acknowledgment

This work has been funded by the LOEWE Initiative (Hessen-Germany) within the priority program Networked Infrastructureless Cooperation for Emergency Response (NICER).

Curriculum vitae

Name:	Andrea Patricia Ortiz Jimenez
Geburtsdatum:	18.03.1986
Geburtsort:	Barranquilla - Kolumbien

Schulausbildung

1997-2002 Colegio Compañia de María “La Enseñanza”

Barranquilla - Kolumbien

Schulabschluß: Allgemeine Hochschulreife

Studium

2003-2008 Studium der Elektrotechnik an der Universidad del Norte, Barranquilla - Kolumbien

Studienabschluß: Bachelor of Engineering

2011-2013 Studium der Master in Information and Communication Engineering an der Technische Universität Darmstadt

Studienabschluß: Master of Science

Berufstätigkeit

2009 - 2010 RF-Ingenierurin, Applus Norcontrol. Barranquilla, Kolumbien

2010 - 2011 Spezialistin für Planung, Optimierung und RF-Design, Tigo. Barranquilla, Kolumbien

seit 2014 Wissenschaftliche Mitarbeiterin am Fachgebiet Kommunikationstechnik, Institut für Nachrichtentechnik, Technische Universität Darmstadt. Darmstadt, Deutschland.

Förderungen und Preise

2014 - 2015	Teilnahme am SciMento Mentoring-Programm für Frauen in Wissenschaft und Wirtschaft
2012 - 2013	Deutschlandstipendium, gefördert durch BASF SE
2011 - 2013	Colfuturo-Stipendium, gefördet durch Fundación para el futuro de Colombia
2008	Auszeichnung für wissenschaftliche Verdienste für Bachelorsarbeit, Universidad del Norte

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Datum und Unterschrift