Reinforcement Learning in MIMO Wireless Networks with Energy Harvesting

Hoda Ayatollahi, Cristiano Tapparello, Wendi Heinzelman
Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA
Email: {hayatoll,ctappare,wheinzel}@ece.rochester.edu

***Abstract*—Energy harvesting wireless nodes provide much longer lifetime and higher energy efficiency for wireless networks compared to battery operated systems. In this paper, we study a MIMO wireless communication link in which the nodes are equipped with energy harvesters and rechargeable batteries that are continuously charging from a renewable energy source. Since the harvested energy arrival and thus the future remaining energy of the nodes is not deterministic in practice, we propose a learning approach in order to find the most efficient transmission policy for data communication that maximizes throughput. The problem is formulated as a Markov Decision Process (MDP) with unknown transition probabilities. A Q-Learning approach is proposed to solve the MDP model and find the optimal transmission policy.**

I. INTRODUCTION

Traditional wireless networks are equipped with battery-operated nodes that have limited capacity and thus limited lifetime. To solve this issue, employing energy harvesting in these networks is essential. Wireless networking with energy harvesting is an emerging area that has received much attention during the past few years. The energy in such networks is provided using ambient sources such as the sun, wind, or vibration. Moreover, employing multi-antenna or Multiple-Input Multiple-Output (MIMO) communication in the network results in not only a longer lifetime but also a higher spectral and energy efficiency in the network.

Energy harvesting networks can be divided into two categories based on knowledge of the energy arrival process [1]. In the first category, the online energy management framework, the nodes have the knowledge about the available energy and can make an online decision about the best course of action to take for reward maximization based on the current state and prior states. This case can be modeled as a Markov Decision Process (MDP), and an optimal solution can be found through dynamic programming, or reinforcement learning methods. For example, the MDP model proposed in [2] is based on finding the optimal power allocation policy for throughput maximization. Similarly, in [3], the authors model a wireless sensor network equipped with energy harvesting as a MDP in order to find an optimal transmission policy for communication.

In the second category, the offline energy management framework, however, knowledge of the energy harvesting process is assumed to be known ahead of time. Many offline strategies are considered for throughput maximization, by adapting the transmission rate based on the energy harvesting distribution [4], the assumed battery imperfections [5], for MIMO channels [6]. Moreover, in [7] and [8], both online and offline approaches are explored. In this work, the authors formulate both the data and energy arrivals as a Markov process and solve this MDP using a proposed learning method based on Q-learning. In [7], the authors studied the maximization of total transmitted data during the transmitter activation time and they showed that as the learning time goes to infinity, the performance reaches the optimal value. In [8], an optimal power allocation policy that maximizes the throughput is the goal of the learning process. Moreover, the state space defined in [8] for the Markov process depends on the nodes' energy and thus it is a continuous state space, which is relaxed through a linear function approximation to handle the infinite number of states. In [9], the authors studied the online and offline problems with the goal of throughput maximization in a fading channel assuming that the energy arrival follows a stochastic process. In this case, they used dynamic programming to find an optimal solution.

In practical situations, however, the exact information of the harvested energy is not known. The harvested energy varies depending on different factors, such as the weather conditions. For instance, in a wireless network utilizing solar or wind power, the amount of harvested energy will be different for sunny, cloudy, or windy days. Thus, we do not have exact information about the harvested energy in reality unless the energy arrival is highly deterministic. In order to find the best transmission policy in the situation with unknown harvesting energy arrival, we propose an approach that is based on reinforcement learning [10].

To the best of our knowledge, no prior work has considered the use of learning approaches in MIMO communication in order to find the best transmission policy in terms of maximum throughput by changing the number of antennas of the nodes. In this paper, we consider a point-to-point MIMO wireless communication link in which we have two nodes equipped with energy harvesters and rechargeable batteries. The goal is to maximize the total throughput during a specific time that the system is running. We model the system using a finite MDP with unknown transmission probabilities and find an optimal transmission policy using Q-learning. We consider four transmitter-receiver antenna pairs in the MIMO system, each of which may result in different energy consumption for the nodes. Based on the energy consumption and the harvested energy arrival, we employ Q-learning to find the most energy

efficient transmission policy.

The rest of the paper is organized as follows. Section II presents the MIMO energy harvesting system model. In Section III, we derive an upper bound on the performance attainable by the system following an offline energy management formulation with complete information about the harvesting process. In Section IV, we present a theoretic model for the energy harvesting online optimization problem. In Section V, we propose a reinforcement learning approach as the solution of an MDP model in order to find the optimal transmission policy. Section VI evaluates the performance of our proposed reinforcement learning method. Finally, conclusions are drawn in Section VII.

II. System Model

We consider a wireless link between two nodes that utilize energy harvesting to recharge their batteries. The energy harvester collects the energy from an ambient source such as solar, wind, or vibration, and stores the energy in the battery. Each node is equipped with two antennas and may either use one or both of them for communication. Thus, the system may work as a Multiple-Input Multiple-Output (MIMO), Multiple-Input Single-Output (MISO), Single-Input Multiple Output (SIMO), or Single-Input Single Output (SISO) communication system. The energy consumption model for the system is adopted from the one presented in [11]. The circuit power consumptions for the transmitter (P_C^{tx}) and the receiver (P_C^{rx}) are as follows

$$\begin{aligned} P_C^{rx}(M_{rx}) =& M_{rx}(P_{\mathrm{ADC}} + P_{\mathrm{Mix}} + P_{\mathrm{Fil}}^{rx} + P_{\mathrm{Dem}} + \\ &+ P_{\mathrm{IFA}} + P_{\mathrm{LNA}}) + P_{\mathrm{Syn}}, \\ P_{\mathrm{C}}^{tx}(M_{tx}) =& M_{tx}(P_{\mathrm{DAC}} + P_{\mathrm{Mix}} + P_{\mathrm{Fil}}^{tx} + P_{\mathrm{Mod}}) + P_{\mathrm{Syn}}, \end{aligned} \tag{1}$$

where M_{tx} and M_{rx} are the number of antennas used at transmitter and the receiver, respectively, P_{ADC} represents the power consumption of the Analog-to-Digital converter (ADC), P_{Mix} is the power consumption of the mixer, P_{Fil}^{rx} is the power consumption of the receiver filter circuit, P_{Dem} is the power consumption of the demodulator, P_{IFA} is the power consumption of the Intermediate Frequency Amplifier (IFA), P_{LNA} is the power consumption of the Low Noise Amplifier (LNA) and P_{Syn} is the power consumption of the frequency synthesizer. The circuitry power consumptions are independent of the distance, bit-error-rate, and depends only on the number of antennas. Moreover, the transmit power P_{PA} can be calculated using the following formula [12]:

$$P_{\mathrm{PA}}(M_{tx}, M_{rx}) = \left(1 + \frac{\xi}{\eta}\right) \overline{E_b} R_b \left(\frac{4\pi d}{\lambda}\right)^n \frac{M_l N_f}{G_{tx} G_{rx}}, \tag{2}$$

where η is the drain efficiency of the power amplifier, while $\xi = 3\frac{W-2\sqrt{W}+1}{W-1}$ represents the Peak-to-Average Ratio (PAR) that depends on the constellation size W. For the results presented in this paper, ξ is a constant value since we only consider a BPSK modulation scheme (i.e., $K = 2$). Moreover, $\overline{E_b}$ is the energy per bit, R_b is the bitrate, d is the distance, n is the path loss exponent, λ is the carrier wavelength, M_l is the link margin, N_f is the receiver noise figure, and G_{tx} and G_{rx} are the transmitter and the receiver antenna gains.

Furthermore, the BER of the channel can be found as follows [13]:

$$p_b = \left(\frac{1}{2}(1-\zeta)\right)^L \cdot \sum_{l=0}^{L-1} \binom{L-1+l}{l} \left(\frac{1}{2}(1+\zeta)\right)^l, \tag{3}$$

where $L = M_{tx}M_{rx}$, $\zeta = \sqrt{\frac{\rho/M_{tx}}{1+\rho/M_{tx}}}$, and ρ is the average SNR.

Given the above, the energy consumption of the transmitter and the receiver for a packet with a size of m bits are:

$$E_{tx}^{M_{tx}\times M_{rx}}(t) = \frac{P_{tx}\ m}{R_b},\ E_{rx}^{M_{tx}\times M_{rx}}(t) = \frac{P_{rx}\ m}{R_b}, \tag{4}$$

where $P_{tx} = P_{\mathrm{C}}^{tx}(M_{tx}) + P_{\mathrm{PA}}(M_{tx}, M_{rx})$ and $P_{rx} = P_C^{rx}(M_{rx})$ are the total energy consumption for the transmitter and the receiver, respectively.

We consider an energy harvesting system in which the nodes know the distribution of the harvested energy. However, they don?t know the exact value of the incoming harvested energy in the future time slots. If we assume that the remaining energy of the transmitter node and the receiver node at time t are $B_{tx}(t)$ and $B_{rx}(t)$, and the amount of harvested energy for the nodes are $H_{tx}(t)$ and $H_{rx}(t)$, which follow a uniform distribution, the remaining energy at slot t is $B_X(t) = B_X(t-1) + H_X(t) - E_X^{M_{tx}\times M_{rx}}(t)$ for the transmitter ($X = tx$) and the receiver ($X = rx$). Thus, at time t we can define the number of packet that can be sent ($N_{tx}(t)$) and received ($N_{rx}(t)$) at the transmitter and the receiver, respectively, as

$$\begin{aligned} N_{tx}(t) &= \frac{B_{tx}(t)}{E_{tx}^{M_{tx}\times M_{rx}}(t) \times \frac{1}{1-p_{pkt}}} \\ N_{rx}(t) &= \frac{B_{rx}(t)}{E_{rx}^{M_{tx}\times M_{rx}}(t) \times \frac{1}{1-p_{pkt}}}, \end{aligned} \tag{5}$$

where $0 \le t \le T_s$, T_s is the total run time of the network, M_{tx} and M_{rx} are the number of antennas at the transmitter and the receiver nodes, respectively, and p_{pkt} is the packet's probability of error. Therefore, the maximum number of packets that can be successfully received by the receiver is given by $N(t) = \min\{N_{tx}(t), N_{rx}(t)\}$.

While the network is running, the nodes receive harvested energy at every time slot t. When a node runs out of energy, it stops sending/receiving packets and start recharging its battery using the harvested energy until the battery is charged enough for the communication. Assuming that at most one packet is sent at time slot t, our goal is to maximize the throughput of the communication system R, which is defined as the total number of packets that are successfully received in the network runtime T_s:

$$R = \max_{\pi} \min\left\{\frac{N_{tx}(T_s)}{T_s}, \frac{N_{rx}(T_s)}{T_s}\right\}, \tag{6}$$

where π is a particular policy that contains the sequence of (M_{tx}, M_{rx}) for all the time slots $t = 0, \ldots, T_s$.

III. Offline Optimal Policy

In the offline energy management framework, the nodes have perfect knowledge of the energy harvesting process, and hence can select the communication scheme (SISO, SIMO, MIMO or MISO) for each transmission in an optimal way. The total number of packets that can be received successfully using an offline optimal policy can be found as

$$\max \sum_{M_{tx}=1}^{M} \sum_{M_{rx}=1}^{M} \alpha_{M_{tx},M_{rx}} \tag{7}$$

s.t.

$$\sum_{M_{tx}=1}^{M} \sum_{M_{rx}=1}^{M} \alpha_{M_{tx},M_{rx}} E_{\mathrm{PKT}}^{tx} \leq B_{tx}^{0} + H_{tx}^{T_s}$$
$$\sum_{M_{tx}=1}^{M} \sum_{M_{rx}=1}^{M} \alpha_{M_{tx},M_{rx}} E_{\mathrm{PKT}}^{rx} \leq B_{rx}^{0} + H_{tx}^{T_s}$$

where T_s is the total duration that the network is running. $H_X^{T_s}$ represents the total harvested energy by the transmitter (X=tx) and the receiver (X=rx) until time T_s, and $E_{\mathrm{PKT}}^{\mathrm{X}}(M_{tx}, M_{rx})$ represents the transmitter (X=tx) and the receiver (X=rx) energy consumptions of the M_{tx}xM_{rx} MIMO scheme when the data packet is successfully transmitted. The value of $\alpha_{M_{tx},M_{rx}}$ represents the number of packets that are exchanged by the $M_{tx} \times M_{rx}$ MIMO scheme during the communication. As a result, by maximizing $\sum_{M_{tx}=1}^{M} \sum_{M_{rx}=1}^{M} \alpha_{M_{tx},M_{rx}}$, the number of received packets and thus the system throughput will be maximized.

This optimal policy works offline and requires information about the initial energy levels, the total harvested energy from the initial time until T_s, and the energy consumption for each communication scheme. Although the offline optimal policy is not reachable in practice, it provides an upper bound on the performance attainable by different communication policies. Moreover, since the number of communication schemes is small in our case, Mixed Integer Linear Programming (MILP) algorithms can efficiently solve the problem in a small amount of time.

IV. Markov Process Model for MIMO Communication with Energy Harvesting

We model the system with a finite-state continuous-time MDP, represented by the quadruplet $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}(s,a,s'), \mathcal{R}(s,a,r) \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P}(s,a,s')$ defines the probability of going from a state s to a state s' when taking action a, and $\mathcal{R}(s,a,r)$ represents the reward r of selecting action a at state s. The set of states and the reward function are defined below.

In order to find the throughput in Eq. (6), we need to know the relationship between $\frac{B_{tx}(t)}{E_{tx}(t)}$ and $\frac{B_{rx}(t)}{E_{rx}(t)}$ for all four MIMO schemes to figure out which one maximizes the throughput. Thus, the set of states $\mathcal{S}$ contains the energy intervals that the fraction $\frac{B_{rx}}{B_{tx}}$ falls into and the set of actions $\mathcal{A} = \{a_1, a_2, a_3, a_4\} = \{\text{SISO, MISO, SIMO, MIMO}\}$, which are different numbers of antenna pairs for the transmitter and the receiver.

As stated in [11], the power consumption of the receiver node depends only on the number of antennas and thus $P_{rx}^{MIMO}(t) = P_{rx}^{SIMO}(t)$ and $P_{rx}^{SISO}(t) = P_{rx}^{MISO}(t)$. Thus, we have

$$\frac{B_{rx}(t)}{E_{rx}(t)} \lessgtr \frac{B_{tx}(t)}{E_{tx}(t)} \quad \Leftrightarrow \quad \frac{B_{rx}(t)}{B_{tx}(t)} \lessgtr \frac{E_{rx}(t)}{E_{tx}(t)}$$

where

$$\frac{E_{rx}(t)}{E_{tx}(t)} = \frac{\{P_{rx}^{SISO}(t), P_{rx}^{MISO}(t), P_{rx}^{SIMO}(t), P_{rx}^{MIMO}(t)\}}{\{P_{tx}^{SISO}(t), P_{tx}^{MISO}(t), P_{tx}^{SIMO}(t), P_{tx}^{MIMO}(t)\}} \tag{8}$$

Thus, in general we have 16 cases in terms of transmitter-receiver power consumption ratios that equal $\frac{E_{rx}(t)}{E_{tx}(t)}$, which can be reduced to 8 since $P_{rx}^{SISO}(t) = P_{rx}^{MISO}(t)$ and $P_{rx}^{MIMO}(t) = P_{rx}^{SIMO}(t)$. Therefore,

$$\frac{E_{rx}(t)}{E_{tx}(t)} = \frac{E_{rx}^{a_i}(t)}{E_{tx}^{a'_j}(t)} \tag{9}$$

where $a_i, a'_j \in \mathcal{A}$ and $i = \{1, 3\}$ and $j = \{1, 2, 3, 4\}$.

As stated in [11], for a specific distance and bit-error-rate (BER), the value of $E_{rx}(t)$ and $E_{tx}(t)$ are known. Thus, $\frac{B_{rx}(t)}{B_{tx}(t)} \in \left[\frac{E_{rx}^{a_i}(t)}{E_{tx}^{a'_j}(t)}, \frac{E_{rx}^{b_i}(t)}{E_{tx}^{b'_j}(t)}\right]$ where $a_i, a'_j, b_i, b'_j \in \mathcal{A}$ and $i = 1, 3$ and $j = 1, 2, 3, 4$.

We define the state space as $\mathcal{S}$ as $\mathcal{S} = \{s_k, s_{10}\}$, where $s_k = \left[\frac{E_{rx}^{a_i}(t)}{E_{tx}^{a'_j}(t)}, \frac{E_{rx}^{b_i}(t)}{E_{tx}^{b'_j}(t)}\right]$ with $k = \{1, 2, ..., 9\}$, and $s_{10} = \{B_{rx}(t) = 0 \text{ and/or } B_{tx}(t) = 0\}$. Thus, the system state space has ten states, nine states regarding the energy fraction intervals according obtained from Eq. (8) and one state where at least one of the nodes runs out of energy.

Moreover, the reward function $\mathcal{R}$ is a function of the remaining energy, energy consumption, and the distance between the nodes:

$$\mathcal{R}(s_i, a_i) = \begin{cases} \frac{1}{E_{tx}^{a_i}(t)} + \frac{1}{E_{rx}^{a_i}(t)} & s_i \neq s_{10} \\ 0 & s_i = s_{10} \end{cases} \tag{10}$$

where R is the system throughput, and $E_{rx}^{a_i}(t)$ and $E_{tx}^{a_i}(t)$ are the receiver and the transmitter power consumptions when action $a_i \in \mathcal{A}$ is taken at state $s_i \in \mathcal{S}$.

V. Reinforcement Learning for MIMO Energy Harvesting MDP Model

Since the transition probabilities are not known in the MDP defined in the previous section, in order to find an optimal action-selection policy, we may use the Q-learning algorithm with three different action selection approaches. The first approach is a greedy approach in which the action with the maximum Q-value is considered at each state. The second approach is to employ Softmax action selection, in which an exploration versus exploitation tradeoff is explored and the action is chosen based on a probability in order to maximize the long-term reward. To take advantage of both action selection policies, we additionally consider a third adaptive approach that is a combination of the greedy and Softmax policies.

The updating rule for the Q-values for an action selection policy π is as follows:

$$Q^{\pi}(s_i, a_i) = (1-\alpha)Q^{\pi}(s_i, a_i) + \alpha[\mathcal{R}_i(s_i, a_i) + \gamma \max_{a} Q^{\pi}(s_{i+1}, a)] \quad (11)$$

where $0 \leq \alpha \leq 1$ is the learning rate, $0 \leq \gamma \leq 1$ is the discount factor, and $\mathcal{R}_i(s_i, a_i)$ is the reward at state s_i when taking the action a_i.

The optimal policy π^* can be found via Value Iteration. In particular, in every learning iteration, the Q-learning algorithm observes the current state $s_i \in \mathcal{S}$ and select the best action $a_i \in \mathcal{A}$. After applying action a_i, the algorithm observes the next state $s_{i+1} \in \mathcal{S}$ and the immediate reward value $\mathcal{R}_i(s_i, a_i)$. Finally, $Q_{\pi}(s_i, a_i)$ is updated according to Eq. (11). It can be shown that iterating this process for a sufficiently large number of learning iteration, the Q-Learning algorithm converges and returns the optimal policy π^*.

A. Greedy Action Selection Policy

Acting greedily for $Q^{\pi}(s_i, a_i)$ when the number of states is finite may result in reaching the optimal policy π^*. In a given state s_i, the greedy action selection policy selects the action that achieves the maximum $Q^{\pi}(s_i, a_i)$. In other words, it selects the a_i that leads to the highest immediate reward R_i.

The problem with the greedy policy is that it is greedy only according to the states that it explored and the energy that it is consumed, which leads to some certain remaining energy intervals of $\frac{B_{rx}(t)}{B_{tx}(t)}$. Thus, it does not have the chance to explore other remaining energy values, which leads to unexplored remaining energy intervals. One solution is to use the ϵ-greedy policy in which, with a probability $1-\epsilon$, it acts greedily and chooses the action that maximizes the Q-value. Otherwise, a random action is selected from $\mathcal{A}$ with a probability of ϵ. However, in the ϵ-greedy policy, a drawback is that it selects equally among the all actions at the time of exploration. Thus, the probability of choosing the best action is the same as the worst action.

B. Softmax Action Selection Policy

To overcome the problem in the ϵ-greedy policy, Softmax action selection can be employed to find the optimal policy $\pi^*(\tau, s, a)$ using an exploration versus exploitation tradeoff [10] with different probabilities for the action selection. In this case, the action a_i at state s_i is chosen with probability $p(s_i, a_i)$ based on the Boltzmann distribution:

$$p(s_i, a_i) = \frac{e^{\frac{Q(s_i, a_i)}{\tau}}}{\sum_{j=1}^{4} e^{\frac{Q(s_i, a_j)}{\tau}}} \quad (12)$$

where $\tau > 0$ is a parameter called temperature. With high temperatures, all actions are equiprobable while with low temperatures, the Softmax action selection policy becomes the same as the greedy policy.

Table I
SIMULATION PARAMETERS

General Parameters	
n (Path loss exponent)	2
W (Constellation size)	2
m (Packet Size)	2064 bytes
N_0	-174 dBm/Hz
$G_t G_r$	5 dBi
M_l	10dB
N_f	10 dB
η	0.35
Distance	250 m
Harvested Energy	$\mathcal{U}$[0,0.004]J
Maximum Harvested Energy(H_{Max})	0.004 J
Initial Energy of the [Transmitter, Receiver] Nodes	[3,1] J
Minimum Required Energy at the [Transmitter, Receiver] Nodes	[1,2] J

Q-Learning Parameters	
δ	0.01
Discount Factor (γ)	0.7
Learning Rate (α)	0.7
Temperature (τ)	500
Learning Iterations (k)	1440×3000
Time Frame	one day = 1440 time slots

Circuitry Power Consumption	
P_{DAC}	7 mW
P_{ADC}	7 mW
P_{Mix}	30.3 mW
P_{Syn}	50 mW
P_{Filt}^{tx}	2.5 mW
P_{Filt}^{rx}	2.5 mW
P_{LNA}	20 mW
P_{IFA}	5 mW

C. Adaptive Action Selection Policy

The third action selection policy is an adaptive action selection policy, which is given by a combination of both the Softmax and the greedy action selection policies described above. During the exploration phase, when the iteration is below a certain value in the learning process, the action a is chosen based on the Softmax policy from Eq. (12). After exploring all of the environment, when the Q-table is set, we switch to the exploitation phase in which the action is selected according to the greedy policy. Therefore, the adaptive action selection policy is as follows [14]:

$$\pi(\delta, \tau, s, a) = \begin{cases} \text{Softmax policy } \pi(\tau, s, a) \text{ according to Eq. (12)} & \Delta \leq \delta \\ \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q(s, a) & \text{Otherwise} \end{cases} \quad (13)$$

where Δ is a uniform random number drawn at each time step, and $0 \leq \delta \leq 1$.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed learning algorithm for MIMO energy harvesting systems presented in Section V, and compare it with other available methods. We assume a Rayleigh fading wireless channel with a channel data rate of $R_b = 1$ Mbps, and an average path loss that falls off with the square of distance (d^2). The simulation parameters are listed in Table I. We use the circuitry power consumption employed in [11]. We use Matlab to simulate the different options, and the results are averaged over 50 runs.

The nodes can operate in four different antenna modes: 2×2 MIMO, 2×1 MISO, 1×2 SIMO, or 1×1 SISO and they start their communication as soon as their remaining energy is greater than or equal to the minimum required energy threshold (as listed in Table I). We assume that the harvesting process follows a uniform distribution $\mathcal{U}[0, H_{Max}]$ for both nodes and their harvesting processes are either the same or

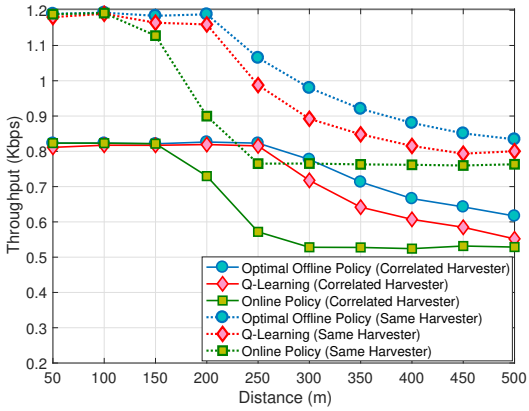

Figure 1. System throughput versus the communication distance.

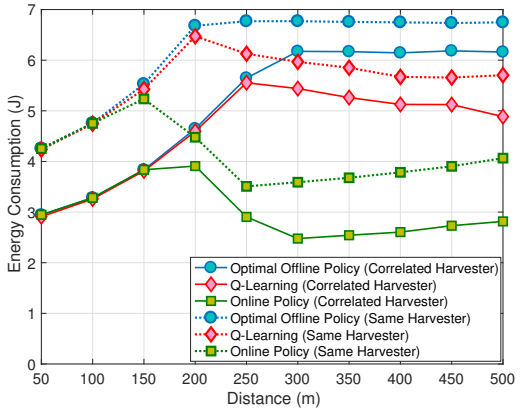

Figure 2. Total system energy consumption versus the communication distance.

correlated (harvested energy of the receiver is a multiplication of that of the transmitter). In the following figures, we consider two cases: the first one is when both nodes' harvesting processes follow a uniform distribution and have the same harvested energy at each time slot ($H_{rx} = H_{tx}$). The second one is when both nodes' harvesting processes follow a uniform distribution but the value of the harvested energy of the nodes are different but correlated (e.g., $H_{rx} = 0.8 \times H_{tx}$). In the most of the results, we employed the first type of the harvester (in which both nodes harvests the same amount of energy) unless noted otherwise.

Moreover, we compare the performance of the proposed protocol with the following policies;

- *Online Policy:* the nodes choose their number of antennas on-the-fly and based on the incoming harvested energy and their current remaining energy. At each time slot, the nodes choose the number of antennas such that the throughput is maximized by solving Eq. (6).
- *Offline Optimal Policy:* the nodes know the total harvested energy in the future and have perfect knowledge of the remaining energy and energy arrivals. Having perfect knowledge of the energy, according to Eq. (7), the nodes select the optimal number of antennas for all the slots of the entire network lifetime.

We analyze the proposed algorithm in terms of various distances, initial energies of the nodes, and network running time (i.e., time frame T_s in Eq. (6)), and energy consumption for different harvesting processes. Moreover, each time frame consists of a number of time slots in which the nodes can transmit/receive at most one packet. Each time slot equals to one minute.

In Figure 1, the throughput is measured and compared with the online policy as a function of the distance between the nodes. As distance grows, the transmit energy consumption grows, which results in having fewer packets received and thus having lower throughput. The Optimal Offline Policy provides an upper bound for the maximum throughput, but this is not necessarily achievable in practice. The throughput

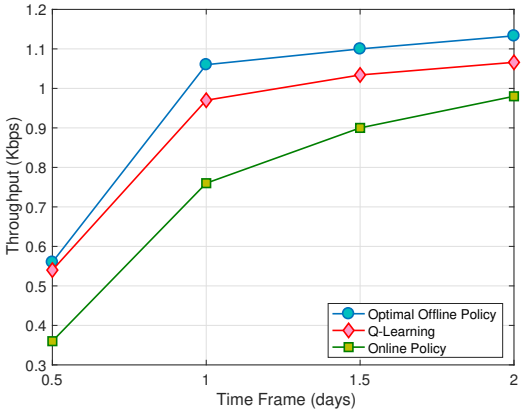

Figure 3. System throughput versus the time frame (days).

of the proposed Q-learning approach is better than that of the online policy especially for large distances due to the learning process in which the Q-values are updated based not only on the energy consumption of the different schemes, but also based on the incoming harvested energy.

In Figure 2, the total energy consumption of the transmitter and the receiver is shown as a function of the communication distance. Since the number of packets and therefore the throughput is higher for large distances for the Q-learning compared to the online policy, the total energy consumption for Q-learning is also higher. However, for small distances the throughput of the online policy is slightly higher than the Q-learning (see Figure 1) which results in having higher energy consumption for the online policy.

In Figure 3, we change the network running time T_s and measure the throughput for the proposed method. Since the maximum time frame is 2 days (2×1440 time slots), we assumed that the number of learning iterations is $2 \times 1440 \times 3000 \approx 8 \times 10^6$. As expected, as the size of the time frame gets larger, the number of received packets and thus the throughput increases as well.

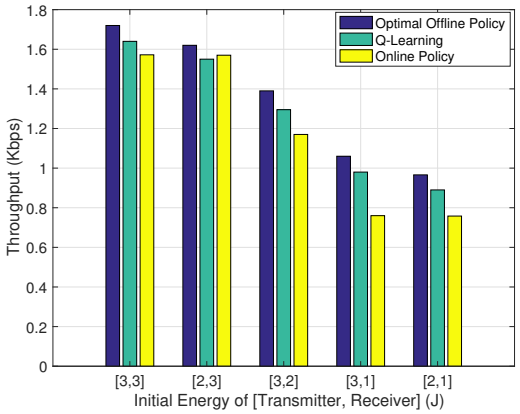

Figure 4. System throughput versus the initial energy of the nodes.

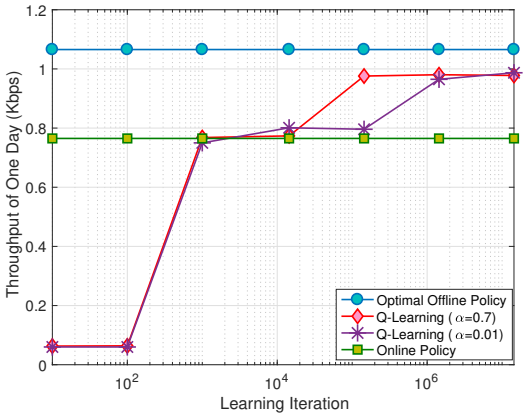

Figure 5. System throughput for one day versus the number of learning iterations.

In Figure 4, the system throughput is demonstrated versus different values of initial energy of the nodes. With higher initial energy of the nodes, the lifetime of the network and thus the throughput gets larger. Figure 5 shows a comparison of the network throughput of the Q-learning, online policy, and offline optimal policy approaches. As we increase the number of learning iterations, the Q-learning algorithm improves the Q-table by the knowledge it learned from the environment. Since we assumed that the distribution of the harvested energy is known, we learn the optimal Q-table in an offline manner for each time frame. Moreover, before the nodes start sending packets in a time frame, the Q-learning algorithm learns the optimal Q-table for the incoming time frame.

In Figure 5, it is assumed that the time frame is one day and the trend of the Q-learning is shown as the number of (offline) learning iterations grows. With a high learning rate ($\alpha = 0.7$) the Q-learning converges faster than with a low learning rate ($\alpha = 0.01$) since with higher learning rates, the algorithm gives more credit to the newly acquired rewards than the previous ones. After around 10^3 learning iterations, the Q-learning and the online policy cross each other and when the number of iterations is larger than this value, the Q-learning converges to a value near the optimal policy.

VII. Conclusions

In this paper, we introduce a new framework for throughput maximization for MIMO wireless links with energy harvesting. We model the problem as an MDP and find the solution using Q-learning. Our proposed solution achieves higher throughput for various distances, various initial energies of the nodes, different harvesting processes, and in different network operating times compared to an online policy. We also compared the proposed learning algorithm with an Optimal Offline policy, and we observe that our proposed algorithm converges to the Optimal Policy, especially for large distances, and various network time frames and the nodes' initial energy.

For future work, we intend to extend this work to a network with more than two nodes in which the problem becomes more complicated since more parameters will be added to the framework, such as the most energy efficient routing path, and the best scheduling data transmission algorithm between a node and its neighbors. We also intend to model and explore this problem in an infinite horizon in which the network may continue running indefinitely. For this case, our goal is to find the solution in terms of the optimum transmission policy and throughput maximization.

References

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

[2] A. Sinha, "Optimal power allocation for a renewable energy source," in *Proc. of IEEE NCC*, Feb. 2012, pp. 1–5.

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

[4] J. Yang and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, 2012.

[5] B. Devillers and D. Gündüz, "A general framework for the optimization of energy harvesting communication systems with battery imperfections," *Journal of Communications and Networks*, vol. 14, no. 2, pp. 130–139, 2012.

[6] M. Gregori and M. Payaró, "Optimal power allocation for a wireless multi-antenna energy harvesting node with arbitrary input distribution," in *Proc. of IEEE ICC*, Jun. 2012, pp. 5794–5798.

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

[8] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting point-to-point communications," in *Proc. of IEEE ICC*, May 2016, pp. 1–6.

[9] 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, 2011.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[11] H. Ayatollahi, C. Tapparello, and W. Heinzelman, "Transmitter-receiver energy efficiency: A trade-off in MIMO wireless sensor networks," in *Proc. of IEEE WCNC*, 2015, pp. 1476–1481.

[12] J. G. Proakis, *Digital Communications*. McGraw-Hill, 2000.

[13] T. M. Duman and A. Ghrayeb, *Coding for MIMO Communication Systems*. Wiley, 2007.

[14] M. A. Wiering, "Explorations in efficient reinforcement learning," Ph.D. dissertation, University of Amsterdam, 1999.