A Learning Theoretic Approach to Energy Harvesting Communication System Optimization

Pol Blasco*, Deniz Gündüz† and Mischa Dohler*
* CTTC, Barcelona, Spain
Emails:{pol.blasco, mischa.dohler}@cttc.es
† Imperial College London, United Kingdom
Email: d.gunduz@imperial.ac.uk

***Abstract*—A point-to-point wireless communication system in which the transmitter is equipped with an energy harvesting device and a rechargeable battery, is studied. Both the energy and the data arrivals at the transmitter are modeled as Markov processes. Delay-limited communication is considered assuming that the underlying channel is block fading with memory, and the instantaneous channel state information is available at both the transmitter and the receiver. The expected total transmitted data during the transmitter's activation time is maximized under three different sets of assumptions regarding the information available at the transmitter about the underlying stochastic processes. A *learning theoretic approach* is introduced, which does not assume any *a priori* information on the Markov processes governing the communication system. In addition, *online* and *offline* optimization problems are studied for the same setting. Full statistical knowledge and causal information on the realizations of the underlying stochastic processes are assumed in the online optimization problem, while the offline optimization problem assumes non-causal knowledge of the realizations in advance. Comparing the optimal solutions in all three frameworks, the performance loss due to the lack of the transmitter's information regarding the behaviors of the underlying Markov processes is quantified.**

***Index Terms*—Dynamic programming, Energy harvesting, Machine learning, Markov processes, Optimal scheduling, Wireless communication**

I. Introduction

Energy harvesting (EH) has emerged as a promising technology to extend the lifetime of communication networks, such as machine-to-machine or wireless sensor networks; complementing current battery-powered transceivers by harvesting the available ambient energy (solar, vibration, thermo-gradient, etc.). As opposed to battery limited devices, an EH transmitter can theoretically operate over an unlimited time horizon; however, in practice transmitter's activation time is limited by other factors and typically the harvested energy rates are quite low. Hence, in order to optimize the communication performance, with sporadic arrival of energy in limited amounts, it is critical to optimize the transmission policy using the available information regarding the energy and data arrival processes.

There has been a growing interest in the optimization of EH communication systems. Prior research can be grouped into two, based on the information (about the energy and data arrival processes) assumed to be available at the transmitter. In the *offline optimization* framework, it is assumed that the transmitter has non-causal information on the exact data/energy arrival instants and amounts [1]–[9]. In the *online optimization* framework, the transmitter is assumed to know the statistics of the underlying EH and data arrival processes; and has causal information about their realizations [10]–[16].

Nonetheless, in many practical scenarios either the characteristics of the EH and data arrival processes change over time, or it is not possible to have reliable statistical information about these processes before deploying the transmitters. For example, in a sensor network with solar EH nodes distributed randomly over a geographical area, the characteristics of each node's harvested energy will depend on its location, and will change based on the time of the day or the season. Moreover, non-causal information about the data/energy arrival instants and amounts is too optimistic in practice, unless the underlying EH process is highly deterministic. Hence, neither online nor offline optimization frameworks will be satisfactory in most practical scenarios. To adapt the transmission scheme to the unknown EH and data arrival processes, we propose a *learning theoretic approach*.

We consider a point-to-point wireless communication system in which the transmitter is equipped with an EH device and a finite-capacity rechargeable battery. Data and energy arrive at the transmitter in packets in a time-slotted fashion. At the beginning of each time-slot (TS), a data packet arrives and it is lost if not transmitted within the following TS. This can be either due to the strict delay requirement of the underlying application, or due to the lack of a data buffer at the transmitter. Harvested energy can be stored in a finite size battery/capacitor for future use, and we consider that the transmission of data is the only source of energy consumption. We assume that the wireless channel between the transmitter and the receiver is constant for the duration of a TS but may vary from one TS to the next. We model the data and energy packet arrivals as well as the channel state as Markov processes. The activation time of an EH transmitter is not limited by the available energy; however, to be more realistic we assume that the transmitter might terminate its operation at any TS with certain probability. This can be due to physical limitations, such as blockage of its channel to the receiver, failure of its components, or because it is forced to switch to the idle mode by the network controller. The objective of the transmitter is to maximize the expected total transmitted data

This work was partially supported by the European Commission in the framework of EXALTED-258512, ACROPOLIS NoE ICT-2009.1.1 and Marie Curie IRG Fellowship with reference number 256410 (COOPMEDIA) and by the Spanish Government under SOFOCLES Grant TEC2010-21100, FPU Grant with reference AP2009-5009 and TEC2010-17816 (JUNTOS).

to the destination during its activation time under the energy availability constraint and the individual deadline constraint for each packet.

For this setting, we provide a complete analysis of the optimal system operation studying the offline, online and the learning theoretic optimization problems. The solution for the offline optimization problem constitutes an upperbound on the online optimization, and the difference between the two indicates the value of knowing the system behavior non-causally. In the learning-based optimization problem we take a more practically relevant approach, and assume that the statistical information about the underlying Markov processes is not available at the transmitter, and that, all the data and energy arrivals as well as the channel states are known only causally. Under these assumptions, we propose a machine learning algorithm for the transmitter operation, such that the transmitter learns the optimal transmission policy over time by performing actions and observing their immediate rewards. We show that the performance of the proposed learning algorithm converges to the solution of the online optimization problem as the learning time increases. The main technical contributions of the paper are summarized as follows:

- We provide, to the best of our knowledge, the first learning theoretic optimization approach to the EH communication system optimization problem under stochastic data and energy arrivals.
- For the same system model, we provide a complete analysis by finding the optimal transmission policy for both the online and offline optimization approaches in addition to the learning theoretic approach.
- For the learning theoretic problem, we propose a Q-learning algorithm and show that its performance converges to that of the optimal online transmission policy as the learning time increases.
- For the online optimization problem, we propose and analyze a transmission strategy based on the policy iteration algorithm.
- We show that the offline optimization problem can be written as a mixed integer linear program. We provide a solution to this problem through the branch-and-bound algorithm. We also propose and solve a linear program relaxation of the offline optimization problem.
- We provide a number of numerical results to corroborate our findings, and compare the performance of the learning theoretic optimization with the offline and online optimization solutions numerically.

The rest of this paper is organized as follows. Section II is dedicated to a summary of the related literature. In Section III, we present the EH communication system model. In Section IV, we study the online optimization problem and characterize the optimal transmission policy. In Section V, we propose a learning theoretic approach, and show that the transmitter is able to learn the stochastic system dynamics and converge to the optimal transmission policy. The offline optimization problem is studied in Section VI. Finally in Section VII, the three approaches are compared and contrasted in different settings through numerical analysis. Section VIII concludes the paper.

II. Related Work

There is a growing literature on the optimization of EH communication system within both online and offline optimization frameworks. Optimal offline transmission strategies have been characterized for point-to-point systems with both data and energy arrivals in [1], with battery imperfections in [2], and with processing energy cost in [3]; for various multi-user scenarios in [2], [4]–[7]; and for fading channels in [8]. Offline optimization of precoding strategies for a MIMO channel is studied in [9]. In the online framework the system is modeled as a Markov decision process (MDP) and dynamic programming (DP) [17] based solutions are provided. In [10], the authors assume that the packets arrive as a Poisson process, and each packet has an intrinsic value assigned to it, which also is a random variable. Modeling the battery state as a Markov process, the authors study the optimal transmission policy that maximizes the average value of the received packets at the destination. Under a similar Markov model, [11] studies the properties of the optimal transmission policy. In [12], the minimum transmission error problem is addressed, where the data and energy arrivals are modeled as Bernoulli and Markov processes, respectively. Ozel et al. [8] study online as well as offline optimization of a throughput maximization problem with stochastic energy arrivals and a fading channel. The causal information assumption is relaxed by modeling the system as a partially observable MDP in [13] and [14]. Assuming that the data and energy arrival rates are known at the transmitter, tools from queueing theory are used for long-term average rate optimization in [15] and [16] for point-to-point and multi-hop scenarios, respectively.

Similar to the present paper, references [18]–[21] optimize EH communication systems under mild assumptions regarding the statistical information available at the transmitter. In [18] a forecast method for a periodic EH process is considered. Reference [19] uses historical data to forecast energy arrival and solves a duty cycle optimization problem based on the expected energy arrival profile. Similarly to [19], the transmitter duty cycle is optimized in [20] and [21] by taking advantage of techniques from control theory and machine learning, respectively. However, [19]–[21] consider only balancing the harvested and consumed energy regardless of the underlying data arrival process and the cost associated to data transmission. In contrast, in our problem setting we consider the data arrival and channel state processes together with the EH process, significantly complicating the problem.

III. System Model

We consider a wireless transmitter equipped with an EH device and a rechargeable battery with limited storage capacity. The communication system operates in a time-slotted fashion over TSs of equal duration. We assume that both data and energy arrive in packets at each TS. The channel state remains constant during each TS and changes from one TS to the next. We consider strict delay constraints for the transmission of data packets; that is, each data packet needs to be transmitted within the TS following its arrival. We assume that the transmitter has a certain small probability $(1-\gamma)$ of terminating its operation at each TS, and it is interested in maximizing the expected total transmitted data during its activation time.

The sizes of the data/energy packets arriving at the beginning of each TS are modeled as correlated time processes following a first-order discrete-time Markov model. Let D_n be the size of the data packet arriving at TS n, where $D_n \in \mathcal{D} \triangleq \{d_1, \dots, d_{N_\mathcal{D}}\}$, and $N_\mathcal{D}$ is the number of elements in $\mathcal{D}$. Let $p_d(d_j, d_k)$ be the probability of the data packet size process going from state d_j to state d_k in one TS. Each energy packet is assumed to be an integer multiple of a fundamental energy unit. Let E_n^H denote the amount of energy harvested during TS n, where $E_n^H \in \mathcal{E} \triangleq \{e_1, \dots, e_{N_\mathcal{E}}\}$, and $p_e(e_j, e_k)$ is the state transition probability function. The energy harvested during TS n, E_n^H, is stored in the battery and can be used for data transmission at the beginning of TS $n+1$. The battery has a limited size of B_{max} energy units and all the energy harvested when the battery is full is lost. Let H_n be the channel state during TS n, where $H_n \in \mathcal{H} \triangleq \{h_1, \dots, h_{N_\mathcal{H}}\}$. We assume that H_n also follows a Markov model; $p_h(h_j, h_k)$ denotes its state transition probability, and the realization of H_n at each TS n is known at the receiver. Similar models have been considered for EH [12]–[14], data arrival [13], and channel state [14], [22] processes. Similar to our model, [10] also considers a strict deadline constraint and lack of data buffer at the transmitter.

For each channel state H_n and packet size D_n, the transmitter knows the amount of minimum energy E_n^T required to transmit the arriving data packet to the destination. Let $E_n^T = f_e(D_n, H_n) : \mathcal{D} \times \mathcal{H} \to \mathcal{E}_u$ where $\mathcal{E}_u$ is a discrete set of integer multiples of the fundamental energy unit. We assume that if the transmitter spends E_n^T units of energy the packet is transmitted successfully.

In each TS n, the transmitter knows the battery state B_n, the size of the arriving packet D_n, the current channel state H_n; and hence, the amount of energy E_n^T required to transmit this packet. At the beginning of each TS, the transmitter makes a binary decision: to transmit or to drop the incoming packet. This may account for the case of control or measurement packets, where the data in the packet is meaningful only if received as a whole. Additionally, the transmission rate and power are fixed at the beginning of each TS, and cannot be changed within the TS. The transmitter must guarantee that the energy spent in TS n is not greater than the energy available in the battery, B_n. Let $X_n \in \{0, 1\}$ be the indicator function of the event that the incoming packet in TS n is transmitted. Then, for $\forall n \in \mathbb{Z}$, we have

$$X_n E_n^T \leq B_n, \tag{1}$$

$$B_{n+1} = \min\{B_n - X_n E_n^T + E_n^H, B_{max}\}. \tag{2}$$

The goal is to maximize the expected total transmitted data over the activation time of the transmitter, which is given by:

$$\max_{\{X_i\}_{i=0}^{\infty}} \lim_{N\to\infty} \mathbb{E}\left[\sum_{n=0}^{N} \gamma^n X_n D_n\right], \quad \text{s.t. } (1) \text{ and } (2), \tag{3}$$

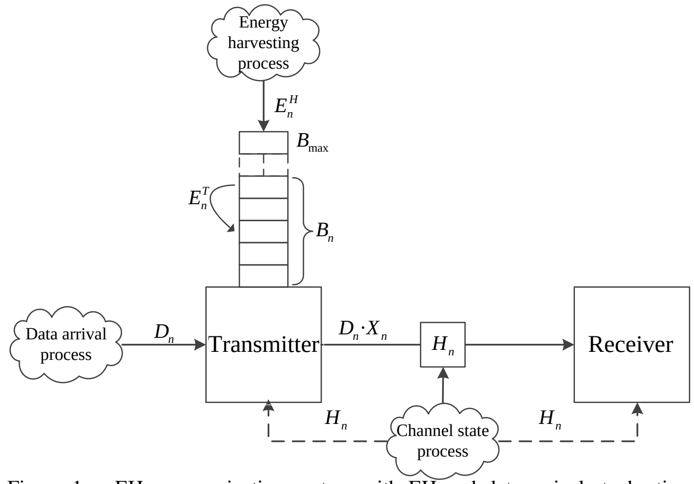


Figure 1. EH communication system with EH and data arrival stochastic processes as well as varying channel.

where $0 < 1 - \gamma \leq 1$ is the independent and identically distributed probability of the transmitter to terminate its operation in each TS. We call this problem the *expected total transmitted data maximization problem* (ETD-problem) as the transmitter aims at maximizing the total transmitted data during an unknown activation time. The EH system that is considered here is depicted in Figure 1.

We will also consider the case with $\gamma = 1$; that is, the transmitter can continue its operation as long as there is available energy. In this case, contrary to the ETD-problem, (3) is not a practical measure of performance as the transmitter operates for an infinite amount of time; and hence, most transmission policies that allow a certain non-zero probability of transmission at each TS are optimal in the expected total transmitted data criterion as they all transmit an infinite amount of data. Hence, we focus on the *expected throughput maximization problem* (TM-problem):

$$\max_{\{X_i\}_{i=0}^{\infty}} \lim_{N\to\infty} \frac{1}{N+1} \mathbb{E}\left[\sum_{n=0}^{N} X_n D_n\right], \quad \text{s.t. } (1) \text{ and } (2). \tag{4}$$

The main focus of the paper is on the ETD-problem, therefore, we assume $0 \leq \gamma < 1$ in the rest of the paper unless otherwise stated. The TM-problem will be studied numerically in Section VII.

An MDP provides a mathematical framework for modeling decision-making situations where outcomes are partly random and partly under the control of the decision maker [23]. The EH communication system, as described above, constitutes a finite-state discrete-time MDP. An MDP is defined via the quadruplet $\langle \mathcal{S}, \mathcal{A}, p_{x_i}(s_j, s_k), R_{x_i}(s_j, s_k) \rangle$, where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of actions, $p_{x_i}(s_j, s_k)$ denotes the transition probability from state s_j to state s_k when action x_i is taken, and $R_{x_i}(s_j, s_k)$ is the immediate reward yielded when in state s_j action x_i is taken and the state changes to s_k. In our model the state of the system in TS n is S_n, which is formed by four components $S_n = (E_n^H, D_n, H_n, B_n)$. Since all components of S_n are discrete there exist a finite number of possible states and the set of states is denoted by $\mathcal{S} = \{s_1, \dots, s_{N_\mathcal{S}}\}$. The set of actions is $\mathcal{A} = \{0, 1\}$, where action 0 (1) indicate that the packet is dropped (transmitted). If the immediate reward yielded by action $x_i \in \mathcal{A}$ when the state changes from S_n to S_{n+1} in TS n is $R_{x_i}(S_n, S_{n+1})$, the objective of an MDP is to find the optimal transmission policy $\pi(\cdot) : \mathcal{S} \to \mathcal{A}$ that maximizes the expected discounted sum reward (i.e., the expected total transmitted data). We restrict our attention to deterministic stationary transmission policies. In our problem, the immediate reward function is $R_{X_n}(S_n, S_{n+1}) = X_n D_n$, and the expected discounted sum reward is equivalent to (3), where γ corresponds to the discount factor, and $X_n = \pi(S_n)$ is the action taken by the transmitter when the system is in state S_n.

Given the policy π and the current state S_n, the state of the battery B_{n+1} is ubiquitously determined by (2). The other state components are randomly determined using the state transition probability functions. Since state transitions depend only on the current state and the transmitter's current action, the model under consideration fulfills the Markov property. As a consequence, we can take advantage of DP and reinforcement learning (RL) [24] tools to solve the ETD-problem.

Next, we introduce the *state-value function* and *action-value function* which will be instrumental in solving the MDP [24]. The state-value function is defined as follows:

$$V^{\pi}(s_j) \triangleq \sum_{\forall s_k \in \mathcal{S}} p_{\pi(s_j)}(s_j, s_k) \left[R_{\pi(s_j)}(s_j, s_k) + \gamma V^{\pi}(s_k)\right]. \tag{5}$$

It is, intuitively, the expected discounted sum reward of policy π when the system is in state s_j. The action-value function, defined as

$$Q^{\pi}(s_j, x_i) \triangleq \sum_{\forall s_k \in \mathcal{S}} p_{x_i}(s_j, s_k) \left[R_{x_i}(s_j, s_k) + \gamma V^{\pi}(s_k)\right], \tag{6}$$

is the expected discounted reward when the system is in state s_j, takes action $x_i \in \mathcal{A}$, and follows policy π thereafter. A policy π is said to be better than or equal to policy π', denoted by $\pi \geq \pi'$, if the expected discounted reward of π is greater than or equal to that of π' in all states, i.e., $\pi \geq \pi'$ if $V^{\pi}(s_j) \geq V^{\pi'}(s_j), \forall s_j \in \mathcal{S}$. The optimal policy π^* is the policy that is better than or equal to any other policy. Eqn. (5) indicates that the state-value function $V^{\pi}(S_n)$ can be expressed as a combination of the expected immediate reward and the state value function of the next state, $V^{\pi}(S_{n+1})$. The same happens with the action-value function. The state-value function when the transmitter follows the optimal policy is

$$V^{\pi^*}(s_j) = \max_{x_j \in \mathcal{A}} Q^{\pi^*}(s_j, x_j). \tag{7}$$

From (7) we see that the optimal policy is the *greedy policy*; that is, the policy that performs the action with the highest expected discount reward according to $Q^{\pi^*}(s_j, x_j)$. The action-value function, when the optimal policy is followed, is

$$Q^{\pi^*}(s_j, x_i) = \sum_{\forall s_k \in \mathcal{S}} p_{x_i}(s_j, s_k)\left[R_{x_i}(s_j, s_k) + \gamma \max_{x_j \in \mathcal{A}} Q^{\pi^*}(s_k, x_j)\right]. \tag{8}$$

Similarly to (5), (8) indicates that the action-value function $Q^{\pi^*}(S_n, x_i)$, when following π^*, can be expressed as a combination of the expected immediate reward and the maximum value of the action-value function of the next state.

There are three approaches to solve the ETD-problem depending on the available information at the transmitter. If the transmitter has prior information on the values of $p_{x_i}(s_j, s_k)$ and $R_{x_i}(s_j, s_k)$, the problem falls into the *online optimization* framework, and we can use DP to find the optimal transmission policy π^*. If the transmitter does not have prior information on the values of $p_{x_i}(s_j, s_k)$ or $R_{x_i}(s_j, s_k)$ we can use a *learning theoretic* approach based on RL. By performing actions and observing their rewards, RL tries to arrive at an optimal policy π^* which maximizes the expected discounted sum reward accumulated over time. Alternatively, in the *offline optimization* framework, it is assumed that all future EH states E_n^H, packet sizes D_n and channel states H_n are known non-causally over a finite horizon.

Remark 1. If the transmitter is allowed to transmit a smaller portion of each packet, using less energy than required to transmit the whole packet, one can re-define the finite action set $\mathcal{A}$. As long as the total number of actions and states remains finite, all the optimization algorithms that we propose in Sections IV and V remains to be valid. In principle, DP and RL ideas can be applied to problems with continuous state and action spaces as well; however, exact solutions are possible only in special cases. A common way of obtaining approximate solutions with continuous state and action spaces is to use function approximation techniques [24]; e.g., by discretizing the action space into a finite set of packet portions, or using fuzzy Q-learning [25].

IV. Online Optimization

We first consider the online optimization problem. We employ *policy iteration* (PI) [26], a DP algorithm, to find the optimal policy in (3). The MDP problem in (3) has finite action and state spaces as well as bounded and stationary immediate reward functions. Under these conditions PI is proven to converge to the optimal policy when $0 \leq \gamma < 1$ [26]. The key idea is to use the structure of (5), (6) and (7) to obtain the optimal policy. PI is based on two steps: 1) policy evaluation, and 2) policy improvement.

In the policy evaluation step the value of a policy π is evaluated by computing the value function $V^{\pi}(s_j)$. In principle, (5) is solvable but at the expense of laborious calculations when $\mathcal{S}$ is large. Instead, PI uses an iterative method [24]: given π, $p_{x_i}(s_j, s_k)$ and $R_{x_i}(s_j, s_k)$, the state value function $V^{\pi}(s_j)$ is estimated as

$$V_l^{\pi}(s_j) = \sum_{s_k} p_{\pi(s_j)}(s_j, s_k) \left[R_{\pi(s_j)}(s_j, s_k) + \gamma V_{l-1}^{\pi}(s_k)\right], \tag{9}$$

for all $s_j \in \mathcal{S}$, where l is the iteration number of the estimation process. It can be shown that the sequence $V_l^{\pi}(s_j)$ converges to $V^{\pi}(s_j)$ as $l \to \infty$ when $0 \leq \gamma < 1$. With policy evaluation, one evaluates how good a policy π is by computing its expected discounted reward at each state $s_j \in \mathcal{S}$.

In the policy improvement step, the PI algorithm looks for a policy π' that is better than the previously evaluated policy π. The Policy Improvement Theorem [17] states that if $Q^{\pi}(s_j, \pi'(s_j)) \geq V^{\pi}(s_j)$ for all $s_j \in \mathcal{S}$ then $\pi' \geq \pi$. Policy improvement step finds the new policy π' by applying

Algorithm 1 Policy Iteration (PI)

1. **Initialize:**
for each $s_j \in \mathcal{S}$ **do**
  initialize $V(s_j)$ and $\pi(s_j)$ arbitrarily
end for
2. **Policy evaluation:**
repeat
  $\Delta \leftarrow 0$
  for each $s_j \in \mathcal{S}$ **do**
    $v \leftarrow V(s_j)$
    $V(s_j) \leftarrow \sum_{s_k} p_{\pi(s_j)}(s_j, s_k)\left[R_{\pi(s_j)}(s_j, s_k) + \gamma V(s_k)\right]$
    $\Delta \leftarrow \max(\Delta, \|v - V(s_j)\|)$
  end for
until $\Delta < \epsilon$
3. **Policy improvement:**
policy-stable $\leftarrow$ true
for each $s_j \in \mathcal{S}$ **do**
  $b \leftarrow \pi(s_j)$
  $\pi(s_j) \leftarrow \mathrm{argmax}_{x_i \in \mathcal{A}} \sum_{s_k} p_{x_i}(s_j, s_k)\left[R_{x_i}(s_j, s_k) + \gamma V(s_k)\right]$

  if $b \neq \pi(s_j)$ **then**
    policy-stable $\leftarrow$ false
  end if
end for
4. **Check stoping criteria:**
if policy-stable **then**
  stop
else
  go to 2).
end if

the greedy policy to $Q^\pi(s_j, x_i)$ in each state. Accordingly, the new policy π' is selected as follows:

$$\pi'(s_j) = \operatorname*{argmax}_{x_i \in \mathcal{A}} Q^\pi(s_j, x_i). \tag{10}$$

PI works iteratively by first evaluating $V^\pi(s_j)$, finding a better policy π', then evaluating $V^{\pi'}(s_j)$, and finding a better policy π'', and so forth. When the same policy is found in two consecutive iterations we conclude that the algorithm has converged. The exact embodiment of the algorithm, as described in [24], is given in Algorithm 1. The worst-case complexity of PI depends on the number of states, $N_\mathcal{S}$, and actions; and in our particular model, the complexity of PI is bounded by $O\left(\frac{2^{N_\mathcal{S}}}{N_\mathcal{S}}\right)$ [27]. The performance of the proposed algorithm and the comparison with other approaches will be given in Section VII.

V. Learning Theoretic Approach

Next we assume that the transmitter has no knowledge of the transition probabilities $p_{x_i}(s_j, s_k)$ and the immediate reward function $R_{x_i}(s_j, s_k)$. We use Q-learning, a learning technique originating from RL, to find the optimal transmission policy. Q-learning relies only on the assumption that the underlying system can be modeled as an MDP, and that after taking action X_n in TS n, the transmitter observes S_{n+1}, and the instantaneous reward value $R_{X_n}(S_n, S_{n+1})$. Notice that, the transmitter does not necessarily know $R_{X_n}(S_n, S_{n+1})$ before taking action X_n, because it does not know the next state S_{n+1} in advance. In our problem, the immediate reward is the size of the transmitted packet D_n; hence, it is readily known at the transmitter.

Eqn. (6) indicates that $Q^\pi(S_n, x_i)$ of the current state-action pair can be represented in terms of the expected immediate reward of the current state-action pair and the state-value function $V^\pi(S_{n+1})$ of the next state. Note that $Q^{\pi^*}(s_j, x_i)$ contains all the long term consequences of taking action x_i in state s_j when following policy π^*. Thus, one can take the optimal actions by looking only at $Q^{\pi^*}(s_j, x_i)$ and choosing the action that will yield the highest expected reward (greedy policy). As a consequence, by only knowing $Q^{\pi^*}(s_j, x_i)$, one can derive the optimal policy π^* without knowing $p_{x_i}(s_j, s_k)$ or $R_{x_i}(s_j, s_k)$. Based on this relation, the Q-learning algorithm finds the optimal policy by estimating $Q^{\pi^*}(s_j, x_i)$ in a recursive manner. In the nth learning iteration $Q^{\pi^*}(s_j, x_i)$ is estimated by $Q_n(s_j, x_i)$, which is done by weighting the previous estimate $Q_{n-1}(s_j, x_i)$ and the estimated expected value of the best action of the next state S_{n+1}. In each TS, the algorithm

- observes the current state $S_n = s_j \in \mathcal{S}$,
- selects and performs an action $X_n = x_i \in \mathcal{A}$,
- observes the next state $S_{n+1} = s_k \in \mathcal{S}$ and the immediate reward $R_{x_i}(s_j, s_k)$,
- updates its estimate of $Q^{\pi^*}(s_j, x_i)$ using

$$\begin{aligned} Q_n(s_j, x_i) \;=\; & (1-\alpha_n) Q_{n-1}(s_j, x_i) + \alpha_n \big[R_{x_i}(s_j, s_k) \\ & + \gamma \max\nolimits_{x_j \in \mathcal{A}} Q_{n-1}(s_k, x_j)\big], \end{aligned} \tag{11}$$

where α_n is the learning rate factor in the nth learning iteration. If all actions are selected and performed with non-zero probability, $0 \leq \gamma < 1$, and the sequence α_n fulfills certain constraints[1], the sequence $Q_n(s_j, x_i)$ is proven to converge to $Q^{\pi^*}(s_j, x_i)$ with probability 1 as $n \to \infty$ [28].

With $Q_n(s_j, x_i)$ at hand the transmitter has to decide for a transmission policy to follow. We recall that, if $Q^{\pi^*}(s_j, x_i)$ is perfectly estimated by $Q_n(s_j, x_i)$, the optimal policy is the greedy policy. However, until $Q^{\pi^*}(s_j, x_i)$ is accurately estimated the greedy policy based on $Q_n(s_j, x_i)$ is not optimal. In order to estimate $Q^{\pi^*}(s_j, x_i)$ accurately, the transmitter should balance the exploration of new actions with the exploitation of known actions. In exploitation the transmitter follows the greedy policy; however, if only exploitation occurs optimal actions might remain unexplored. In exploration the transmitter takes actions randomly with the aim of discovering better policies and enhancing its estimate of $Q^{\pi^*}(s_j, x_i)$. The ϵ-greedy action selection method either takes actions randomly (explores) with probability ϵ or follows the greedy policy (exploits) with probability $1-\epsilon$ at each TS, where $0 < \epsilon < 1$.

The convergence rate of $Q_n(s_j, x_i)$ to $Q^{\pi^*}(s_j, x_i)$ depends on the learning rate α_n. The convergence rate decreases with the number of actions, states, and the discount factor γ,

[1]The constraints on the learning rate follow from well-known results in stochastic approximation theory. Denote by $\alpha_{n^k(s_j,x_i)}$ the learning rate α_n corresponding to the kth time action x_i is selected in state s_j. The constraints on α_n are $0 < \alpha_{n^k(s_j,x_i)} < 1$, $\sum_{k=0}^{\infty} \alpha_{n^k(s_j,x_i)} = \infty$, and $\sum_{k=0}^{\infty} \alpha^2_{n^k(s_j,x_i)} < \infty$, $\forall s_j \in \mathcal{S}$ and $\forall x_i \in \mathcal{A}$. The second condition is required to guarantee that the algorithm's steps are large enough to overcome any initial condition. The third condition guarantees that the steps become small enough to assure convergence. Although the use of sequences α_n that meet these conditions assures convergence in the limit, they are rarely used in practical applications.

Algorithm 2 Q-learning

1. **Initialize:**
for each $s_j \in \mathcal{S}$, $x_i \in \mathcal{A}$ **do**
  initialize $Q(s_j, x_i)$ arbitrarily
end for
set initial time index $n \leftarrow 1$
evaluate the starting state $s_j \leftarrow S_n$
2. **Learning:**
repeat
  select action X_n following the ϵ-greedy action selection method
  perform action $x_i \leftarrow X_n$
  observe the next state $s_k \leftarrow S_{n+1}$
  receive an immediate cost $R_{x_i}(s_j, s_k)$
  select the action x_j corresponding to the $\max_{x_j} Q(s_k, x_j)$
  update the $Q(s_j, x_i)$ estimate as follows:
  $Q(s_j, x_i) \leftarrow$
  $(1-\alpha_n)Q(s_j, x_i) + \alpha_n[R_{x_i}(s_j, s_k) + \gamma \max_{x_j} Q(s_k, x_j)]$
  update the current state $s_j \leftarrow s_k$
  update the time index $n \leftarrow n+1$
until check stopping criteria $n = N_L$

and increases with the number of learning iterations, N_L. See [29] for a more detailed study of the convergence rate of the Q-learning algorithm. Q-learning algorithm is given in Algorithm 2. In Section VII the performance of Q-learning in our problem setup is evaluated and compared to other approaches.

VI. Offline Optimization

In this section we consider the problem setting in Section III assuming that all the future data/energy arrivals as well as the channel variations are known non-causally at the transmitter before the transmission starts. Offline optimization is relevant in applications for which the underlying stochastic processes can be estimated accurately in advance at the transmitter. In general the solution of the corresponding offline optimization problem can be considered as an upperbound on the performance of the online and learning theoretic problems. Offline approach optimizes the transmission policy over a realization of the MDP for a finite number of TSs, whereas the learning theoretic and online optimization problems optimize the expected total transmitted data over an infinite horizon. We recall that an MDP realization is a sequence of state transitions of the data, EH and the channel state processes for a finite number of TSs. Given an MDP realization in the offline optimization approach, we optimize X_n such that the the expected total transmitted data is maximized. From (3) the offline optimization problem can be written as follows

$$\max_{\mathbf{X},\mathbf{B}} \quad \sum_{n=0}^{N} \gamma^n X_n D_n \tag{12a}$$
$$\text{s.t.} \quad X_n E_n^T \leq B_n, \tag{12b}$$
$$B_{n+1} \leq B_n - X_n E_n^T + E_n^H, \tag{12c}$$
$$0 \leq B_n \leq B_{max}, \tag{12d}$$
$$X_n \in \{0,1\}, \quad n = 0, \dots, N, \tag{12e}$$

where $\mathbf{B} = [B_0 \cdots B_N]$ and $\mathbf{X} = [X_0 \cdots X_N]$. Note that we have replaced the equality constraint in (2) with two inequality constraints, namely (12c) and (12d). Hence, the problem in (12) is a relaxed version of (3). To see that the two problems are indeed equivalent, we need to show that any solution to (12) is also a solution to (3). If the optimal solution to (12) satisfies (12c) or (12d) with equality, then it is a solution to (3) as well. Assume that $\mathbf{X}, \mathbf{B}$ is an optimal solution to (12) and that for some n, B_n fulfills both of the constraints (12c) and (12d) with strict inequality whereas the other components satisfy at least one constraint with equality. In this case, we can always find a $B_n^+ > B_n$ such that at least one of the constraints is satisfied with equality. Since $B_n^+ > B_n$, (12b) is not violated and $\mathbf{X}$ remains to be feasible, achieving the same objective value. In this case, $\mathbf{X}$ is feasible and a valid optimal solution to (3) as well, since B_n^+ satisfies (2).

The problem in (12) is a *mixed integer linear program* (MILP) problem since it has affine objective and constraint functions, while the optimization variable X_n is constrained to be binary. This problem is known to be NP-hard; however, there are algorithms combining relaxation tools with smart exhaustive search methods to reduce the solution time. Notice that, if one relaxes the binary constraint on X_n to $0 \leq X_n \leq 1$, (12) becomes a linear program (LP). This corresponds to the problem in which the transmitter does not make binary decisions, and is allowed to transmit smaller portions of the packets. We call the optimization problem in (12) the complete-problem and its relaxed version the LP-relaxation. We define $\mathcal{O} = \{0,1\}^N$ as the feasible set for $\mathbf{X}$ in the complete-problem. The optimal value of the LP-relaxation provides an upper bound on the complete-problem. On the other hand, if the value of $\mathbf{X}$ in the optimal solution of the LP-relaxation belong to $\mathcal{O}$, it is also an optimal solution to the complete-problem.

Most available MILP solvers employ an LP based *branch-and-bound* (B&B) algorithm [30]. In exhaustive search one has to evaluate the objective function for each point of the feasible set $\mathcal{O}$. The B&B algorithm discards some subsets of $\mathcal{O}$ without evaluating the objective function over these subsets. B&B works by generating disjunctions; that is, it partitions the feasible set $\mathcal{O}$ of the complete-problem into smaller subsets, $\mathcal{O}_k$, and explores or discards each subset $\mathcal{O}_k$ recursively. We denote the kth active subproblem which solve (12) with $\mathbf{X}$ constrained to the subset $\mathcal{O}_k \subseteq \mathcal{O}$ by $\mathrm{CsP}(k)$, and its associated upperbound by I_k. The optimal value of $\mathrm{CsP}(k)$ is a lowerbound on the optimal value of the complete-problem. The algorithm maintains a list $\mathcal{L}$ of active subproblems over all the active subsets $\mathcal{O}_k$ created. The feasible solution among all explored subproblems with the highest optimal value is called the *incumbent*, and its optimal value is denoted by I_{max}. At each algorithm iteration an active subproblem $\mathrm{CsP}(k)$ is chosen, deleted from $\mathcal{L}$, and its LP-relaxation is solved. Let $\hat{\mathbf{X}}^k$ be the optimal $\mathbf{X}$ value corresponding to the solution of the LP-relaxation of $\mathrm{CsP}(k)$, and I_k^{LP} be its optimal value. There are three possibilities: 1) If $\hat{\mathbf{X}}^k \in \mathcal{O}_k$, $\mathrm{CsP}(k)$ and its LP-relaxation have the same solution. We update $I_{max} = \max\{I_k^{LP}, I_{max}\}$, and all subproblems $\mathrm{CsP}(m)$ in $\mathcal{L}$ for which $I_m \leq I_{max}$ are discarded; 2) If $\hat{\mathbf{X}}^k \notin \mathcal{O}_k$ and $I_k^{LP} \leq I_{max}$, then the optimal solution of $\mathrm{CsP}(k)$ can not improve I_{max}, and the subproblem $\mathrm{CsP}(k)$ is discarded, and 3) If $\hat{\mathbf{X}}^k \notin \mathcal{O}_k$ and $I_k^{LP} > I_{max}$, then $\mathrm{CsP}(k)$ requires further exploration, which is done by branching it further, i.e., creating two new subproblems from $\mathrm{CsP}(k)$ by branching its feasible set $\mathcal{O}_k$ into two.

For the binary case that we are interested in, a branching step is as follows. Assume that for some n, the nth element of $\hat{\mathbf{X}}^k$ is not binary, then we can formulate a logical disjunction for the nth element of the optimal solution by letting $X_n = 0$, or $X_n = 1$. With this logical disjunction the algorithm creates two new subsets $\mathcal{O}_{k'} = \mathcal{O}_k \cap \{\mathbf{X} : X_n = 1\}$ and $\mathcal{O}_{k''} = \mathcal{O}_k \cap \{\mathbf{X} : X_n = 0\}$, which partition $\mathcal{O}_k$ into two mutually exclusive subsets. Note that $\mathcal{O}_{k'} \cup \mathcal{O}_{k''} = \mathcal{O}_k$. The two subproblems, $\mathrm{CsP}(k')$ and $\mathrm{CsP}(k'')$, associated with the new subsets $\mathcal{O}_{k'}$ and $\mathcal{O}_{k''}$, respectively; are added to $\mathcal{L}$. The upperbounds $I_{k'}$ and $I_{k''}$ associated to $\mathrm{CsP}(k')$ and $\mathrm{CsP}(k'')$, respectively, are set equal to I_k^{LP}.

After updating $\mathcal{L}$ and I_{max} the B&B algorithm selects another subproblem $\mathrm{CsP}(m)$ in $\mathcal{L}$ to explore. The largest upperbound associated with the active subproblems in $\mathcal{L}$ is an upperbound on the complete-problem. The B&B algorithm terminates when $\mathcal{L}$ is empty, in which case this upperbound is equal to the value of the incumbent. The B&B algorithm is given in Algorithm 3. In principle, the worst-case complexity of B&B is $O\left(2^N\right)$, same as exhaustive search; however, the average complexity of B&B is usually much lower, and is polynomial under certain conditions [31].

Remark 2. Notice that, unlike the online and learning theoretic optimization problems, the offline optimization approach is not restricted to the case where $0 \leq \gamma < 1$. Hence, both the B&B algorithm and the LP relaxation can be applied to the TM-problem in (4).

Algorithm 3 B&B

1. **Initialize:**
$I_{max} = 0$, $\mathcal{O}_0 = \mathcal{O}$, and $I_0 = \infty$
set $\mathrm{CsP}(0) \leftarrow \{$solve (12) s.t. $\mathbf{X} \in \mathcal{O}_0\}$
$\mathcal{L} \leftarrow \mathrm{CsP}(0)$
2. **Terminate:**
if $\mathcal{L} = \emptyset$ **then**
    $\hat{\mathbf{X}}_{max}$ is the optimal solution and I_{max} the optimal value
end if
3. **Select:**
choose and delete a subproblem $\mathrm{CsP}(k)$ form $\mathcal{L}$
4. **Evaluate:**
solve LP-relaxation of $\mathrm{CsP}(k)$
if LP-relaxation is infeasible **then**
    go to Step **2**
else
    let I_k^{LP} be its optimal value and $\hat{\mathbf{X}}^k$ the optimal $\mathbf{X}$ value
end if
5. **Prune:**
if $I_k^{LP} \leq I_{max}$ **then**
    go to Step **2**
else if $\hat{\mathbf{X}}^k \notin \mathcal{O}_k$ **then**
    go to Step **6**
else
    $I_{max} \leftarrow I_k^{LP}$ and $\hat{\mathbf{X}}_{max} \leftarrow \hat{\mathbf{X}}^k$
    delete all subproblems $\mathrm{CsP}(m)$ in $\mathcal{L}$ with $I_m \leq I_{max}$
end if
6. **Branch:**
choose n, such that $\hat{\mathbf{X}}_n^k$ is not binary
set $I_{k'}, I_{k''} \leftarrow I_k^{LP}$, $\mathcal{O}_{k'} \leftarrow \mathcal{O}_k \cap \{\mathbf{X} : X_n = 1\}$ and $\mathcal{O}_{k''} \leftarrow \mathcal{O}_k \cap \{\mathbf{X} : X_n = 0\}$
set $\mathrm{CsP}(k') \leftarrow \{$solve (12) s.t. $\mathbf{X} \in \mathcal{O}_{k'}\}$ and $\mathrm{CsP}(k'') \leftarrow \{$solve (12) s.t. $\mathbf{X} \in \mathcal{O}_{k''}\}$
add $\mathrm{CsP}(k')$ and $\mathrm{CsP}(k'')$ to $\mathcal{L}$
go to Step **3**

VII. Numerical Results

To compare the performance of the three approaches that we have proposed, we focus on a sample scenario of the EH communication system presented in Section III. We are interested in comparing the expected performance of the proposed solutions. For the online optimization approach it is possible to evaluate the expected performance of the optimal policy π^*, found using the DP algorithm, by solving (5), or evaluating (9) and averaging over all possible starting states $S_0 \in \mathcal{S}$. In theory, the learning theoretic approach will achieve the same performance as the online optimization approach as the learning time goes to infinity (for $0 \leq \gamma < 1$); however, in practice the transmitter can learn only for a finite number of TSs and the transmission policy it arrives at depends on the specific realization of the MDP. The offline optimization approach optimizes over a realization of the MDP. To find the expected performance of the offline optimization approach one has to average over infinite realizations of the MDP for an infinite number of TSs. We can average the performance over only a finite number of MDP realizations and finite number of TSs. Hence, we treat the performances of the proposed algorithms as a random variable, and use the sample mean to estimate their expected values. Accordingly, to provide a measure of accuracy for our estimators, we also compute the confidence intervals. The details of the confidence interval computations are relegated to the Appendix.

In our numerical analysis we use parameters based on an IEEE802.15.4e [32] communication system. We consider a TS length of $\Delta_{TS} = 10$ms, a transmission time of $\Delta_{Tx} = 5$ ms, and an available bandwidth of $W = 2$ MHz. The fundamental energy unit is 2.5 μJ which may account for a vibration or piezoelectric harvesting device [33], and we assume that the transmitter at each TS either harvests two units of energy or does not harvest any, i.e., $\mathcal{E} = \{0, 2\}$. We denote the probability of harvesting two energy units in TS n given that the same amount was harvested in TS $n-1$ by p_H, i.e., $p_H \triangleq p_e(2,2)$. We will study the effect of p_H and B_{max} on the system performance and the convergence behavior of the learning algorithm. We set $p_e(0,0)$, the probability of not harvesting any energy in TS n when no energy was harvested in TS $n-1$, to 0.9. The battery capacity B_{max} is varied from 5 to 9 energy units. The possible packet sizes are $D_n \in \mathcal{D} = \{300, 600\}$ bits with state transition probabilities $p_d(d_1, d_1) = p_d(d_2, d_2) = 0.9$. Let the channel state at TS n be $H_n \in \mathcal{H} = \{1.655 \cdot 10^{-13}, 3.311 \cdot 10^{-13}\}$ which are two realizations of the indoor channel model for urban scenarios in [34] with $d = d_{indoor} = 55$, $w = 3$, $WP_{in} = 5$, and 5 dBm standard deviation, where d is the distance in meters, w the number of walls, and WP_{in} the wall penetration losses. The state transition probability function is characterized by

$p_h(h_1, h_1) = p_h(h_2, h_2) = 0.9$.

To find the required energy to reliably transmit a data packet over the channel we consider Shannon's capacity formula for Gaussian channels. The transmitted data in TS n is

$$D_n = W\Delta_{Tx} \log_2 \left(1 + \frac{H_n P}{W N_0}\right), \quad (13)$$

where P is the transmit power and $N_0 = 10^{-20.4}$ (W/Hz) is the noise power density. In low power regime, which is of special practical interest in the case of machine-to-machine communications or wireless sensor networks with EH devices, the capacity formula can be approximated by $D_n \simeq \frac{\Delta_{Tx} H_n P}{\log(2) N_0}$, where $\Delta_{Tx} P$ is the energy expended for transmission in TS n. Then, the minimum energy required for transmitting a packet D_n is given by $E_n^T = f_e(D_n, H_n) = \frac{D_n \log(2) N_0}{H_n}$. In general, we assume that the transmit energy for each packet at each channel state is an integer multiple of the energy unit. In our special case, this condition is satisfied as we have $\mathcal{E}_u = \{1, 2, 4\}$, which correspond to transmit power values of $0.5, 1$ and 2 mW, respectively. Numerical results for the ETD-problem, in which the transmitter might terminate its operation at each TS with probability γ is given in Section VII-A whereas the TM-problem is examined in Section VII-B.

A. ETD-problem

We generate $T = 2000$ realizations of $N = 100$ random state transitions and examine the performance of the proposed algorithms for $\gamma = 0.9$. In particular, we consider the LP-relaxation of the offline optimization problem, the offline optimization problem with the B&B algorithm[2], the online optimization problem with PI, the learning theoretic approach with Q-learning[3]. We have considered a greedy algorithm which assumes a causal knowledge of B_n, D_n and H_n, and transmits a packet whenever there is enough energy in the battery ignoring the Markovity of the underlying processes.

Notice that the LP-relaxation solution is an upper bound on the performance of the offline optimization problem, which, in turn, is an upper bound on the online problem. At the same time the performance of the online optimization problem is an upper bound on the learning theoretic and the greedy approaches.

In Figure 2 we illustrate, together with the performance of the other approaches, the expected total transmitted data by the learning theoretic approach against the number of learning iterations, N_L. We can see that for $N_L > 200$ TSs the learning theoretic approach ($\epsilon = 0.07$) reaches 85% of the performance achieved by online optimization, while for $N_L > 2 \cdot 10^5$ TSs it reaches 99%. We can conclude that the learning theoretic approach is able to learn the optimal policy with increasing accuracy as N_L increases. Moreover, we have investigated the exploration/exploitation tradeoff of the learning algorithm, and we have observed that for low exploration values ($\epsilon = 0.001$) the learning rate decreases, compared to moderate exploration values ($\epsilon = 0.07$). We also observe from Figure 2 that the performance of the greedy algorithm is notably inferior compared to the other approaches.

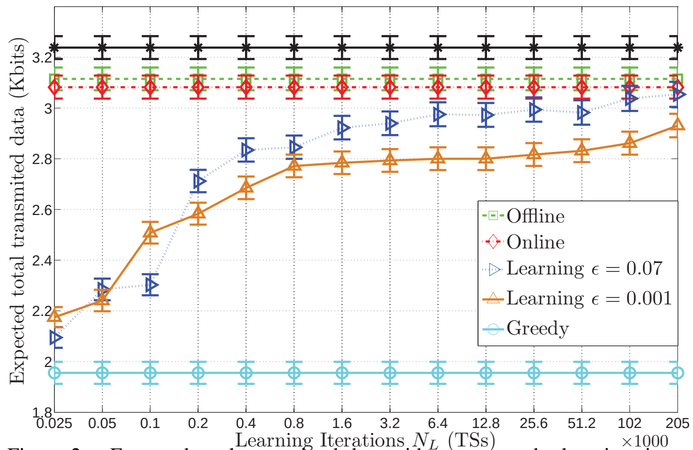


Figure 2. Expected total transmitted data with respect to the learning time N_L, with $p_H = 0.9$, and B_{max} =5.

Figure 3(a) displays the expected total transmitted data for different p_H values. We consider $N_L = 10^4$ TSs for the learning theoretic approach since short learning times are more practically relevant. As expected, performance of all the approaches increase as the average amount of harvested energy increases with p_H. The offline approach achieves, on average, 96% of the performance of the offline-LP solution. We observe that the learning theoretic approach converges to the online optimization performance with increasing p_H, namely its performance is 90% and 99% of that of the online approach for $p_H = 0.5$ and $p_H = 0.9$, respectively. It can also be seen that the online optimization achieves 97% of the performance of the offline optimization when $p_H = 0.5$, while for $p_H = 0.9$ it reaches 99%. This is due to the fact that the underlying EH process becomes less random as p_H increases; and hence, the online algorithm can better estimate its future states and adapt to it. Additionally, we observe from Figure 3(a) that the performance of the greedy approach reaches a mere 60% of the offline optimization.

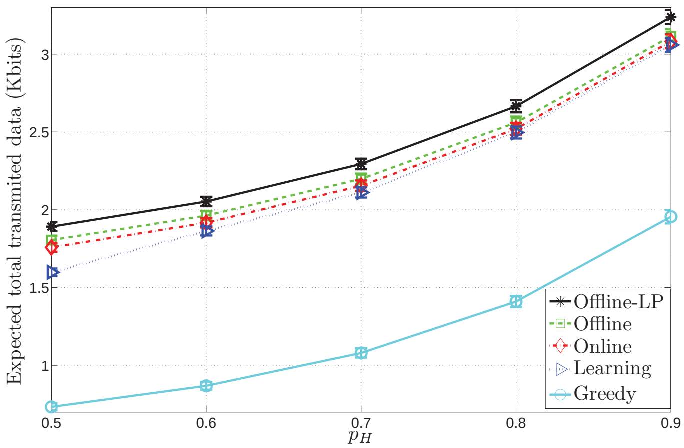
In Figure 3(b) we show the effect of the battery size, B_{max}, on the expected total transmitted data for $N_L = 10^4$ TSs. We see that the expected total transmitted data increases with B_{max} for all the proposed algorithms but the greedy approach. Overall, we observe that the performance of the online optimization is approximately 99% that of the offline optimization. Additionally, we see that the learning theoretic approach reaches at least 91% of the performance of the online optimization. Although only a small set of numerical results is presented in the paper due to space limitations, we have executed exhaustive numerical simulations with different parameter settings and observed similar results.

B. TM-problem

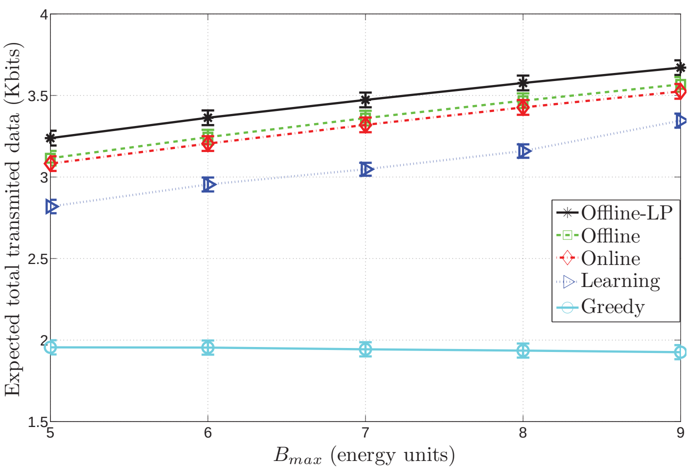
In the online and learning theoretic formulations, the TM-problem in (4) falls into the category of average reward maximization problems, which cannot be solved with Q-learning unless a finite number of TSs is specified, or the

[2]Reference [30] presents a survey on software tools for MILP problems. In this paper we use the B&B toolbox provided in [35]. In particular, B&B is set up with a 20 seconds timeout. For the particular setting of this paper, the B&B algorithm found an optimal solution, within the given timeout, 99.7% of the times.

[3]We use the ϵ-greedy action selection mechanism with $\epsilon = 0.07$, and set the learning rate to $\alpha = 0.5$.

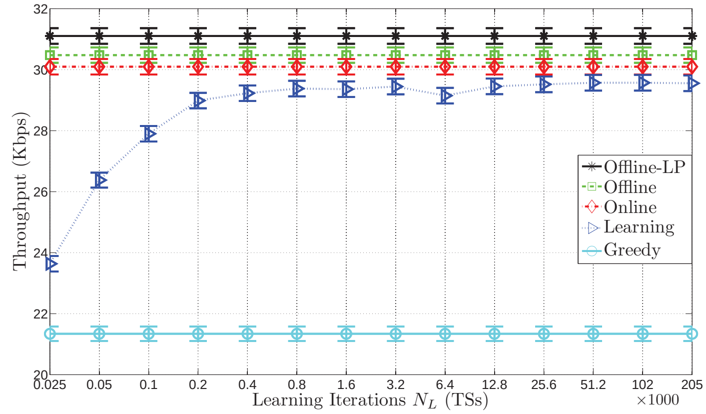


(a) $p_H = \{0.5, \ldots, 0.9\}$ and $B_{max} = 5$.

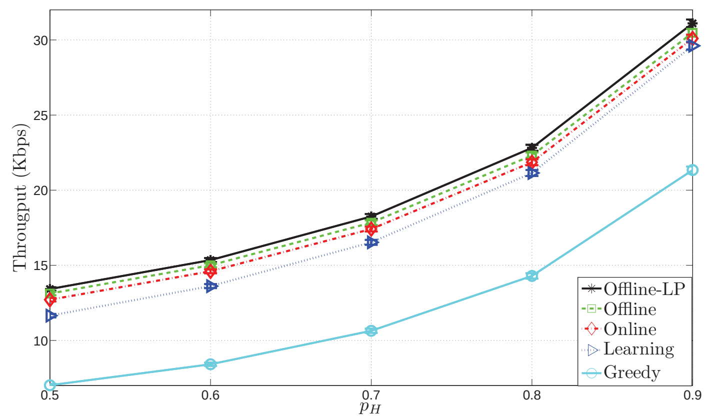


(b) $B_{max} = \{5, \ldots, 9\}$ and $p_H = 0.9$.

Figure 3. Performance comparison of proposed algorithms for the ETD-problem.



(a) Average throughput versus N_L for $p_H = 0.9$



(b) Average throughput for $p_H = \{0.5, \ldots, 0.9\}$.

Figure 4. Performance comparison of the proposed algorithms for the TM-problem for $B_{max} = 5$.

MDP presents absorbing states. Alternatively, one can take advantage of the average reward RL algorithms. Nevertheless, the convergence properties of these methods are not yet well understood. In this paper we consider R-learning[4] [36] which is similar to Q-learning, but is not proven to converge.

Similarly, for the online optimization problem, the policy evaluation step in the PI algorithm is not guaranteed to converge for $\gamma = 1$. Instead, we use relative value iteration (RVI) [26], which is a DP algorithm, to find the optimal policy in average reward MDP problems.

In our numerical analysis for the TM-problem, we consider the LP-relaxation of the offline optimization problem, the offline optimization problem with the B&B algorithm, the online optimization problem with RVI, the learning theoretic approach with R-learning[5], and finally, the greedy algorithm. For evaluation purposes we average over $T = 2000$ realizations of $N = 100$ random state transitions.

In Figure 4(a) we illustrate, together with the performance of the other approaches, the throughput achieved by the learning theoretic approach against the number of learning iterations, N_L. We observe that for $N_L > 200$ TSs the learning algorithm reaches 95% of the performance achieved by online optimization, while for $N_L > 2 \cdot 10^5$ TSs the performance is 98% of the performance of the online optimization approach. Notably the learning theoretic approach performance increases with N_L; however, in this case the performance does not converge to the performance of the online optimization approach. As before the greedy algorithm is notably inferior compared to the other approaches.

Figure 4(b) displays the throughput for different p_H values. We plot the performance of the learning theoretic approach for $N_L = 10^4$ TSs and $\epsilon = 0.07$. As expected, performance of all the approaches increase as the average amount of harvested energy increases with p_H. It can be seen that the online approach achieves, on average, 95% of the performance of the offline approach. This is in line with our finding in Figure 3(a). The throughput achieved by the learning theoretic approach achieves 91% of the online optimization throughput for $p_H = 0.5$ and 98% for $p_H = 0.9$. Similarly to Figure 3(a), the learning theoretic and the online optimization performances, compared to that of the offline optimization, increase when the underlying Markov processes are less random. Similarly to the ETD-problem, the greedy algorithm shows a performance well below the others. We observe that, although the convergence properties of the R-learning are not well understood it has a similar behavior to Q-learning, in practice.

[4]In R-learning $R_{x_i}(s_j, s_k)$ in (11) is substituted by $\hat{R}_{x_i}(s_j, s_k) = R_{x_i}(s_j, s_k) - \rho_n$, where $\rho_n = (1-\beta)\rho_{n-1} + \beta\big[R_{x_i}(s_j, s_k) + \max_{x_j \in \mathcal{A}} Q_{n-1}(s_k, x_j) - \max_{x_j \in \mathcal{A}} Q_{n-1}(s_j, x_j)\big]$, $0 \le \beta \le 1$, and ρ_n is updated in TS n only if a non-exploratory action is taken.

[5]We use the same action selection method as the Q-learning algorithm in Section VII-A.

VIII. Conclusions

We have considered a point-to-point communication system in which the transmitter has an energy harvester and a rechargeable battery with limited capacity. We have studied optimal communication schemes under strict deadline constraints. Our model includes stochastic data/energy arrivals and a time-varying channel, all modeled as Markov processes. We have studied the ETD-problem, which maximizes the expected total transmitted data during the transmitter's activation time. Considering various assumptions regarding the information available at the transmitter about the underlying stochastic processes; online, learning theoretic and offline optimization approaches have been studied. For the learning theoretic and the online optimization problems the communication system is modeled as an MDP, and the corresponding optimal transmission policies have been identified. A Q-learning algorithm has been proposed for the learning theoretic approach, and as the learning time goes to infinity its performance has been shown to reach the optimal performance of the online optimization problem, which is solved here using policy iteration algorithm. The offline optimization problem has been characterized as a mixed integer linear program problem, and its optimal solution through the branch-and-bound as well as a linear program relaxation have been presented.

Our numerical results have illustrated the relevance of the learning theoretic approach for practical scenarios. For practically relevant system parameters, it has been shown that, the learning theoretic approach reaches 90% of the performance of the online optimization after a reasonable small number of learning iterations. Accordingly, we have shown that smart and energy-aware transmission policies can raise the performance from 60% up to 90% of the performance of the offline optimization approach compared to the greedy transmission policy. We have also addressed the TM-problem and made similar observations despite the lack of theoretical convergence results.

Appendix

In the discounted sum data problem we are interested in estimating $\bar{X} = \mathbb{E}\left[\lim_{N\to\infty}\sum_{n=0}^{N}\gamma^n X_n D_n\right]$, where X_n is the action taken by the transmitter which is computed using either the offline optimization, online optimization or the learning theoretic approach, and D_n is the packet size in the nth TS. An upper bound on $\bar{X}$ can be found as

$$\bar{X} \leq \underbrace{\mathbb{E}\left[\sum_{n=0}^{N}\gamma^n X_n D_n\right]}_{\bar{X}_N} + \underbrace{D_{\max}\frac{\gamma^N}{1-\gamma}}_{\epsilon_N}, \tag{14}$$

which follows by assuming that after TS N all packets arriving at the transmitter are of size $D_{\max} \geq d_j$ for all $d_j \in \mathcal{D}$, that there is enough energy to transmit all the arriving packets, and that, $0 \leq \gamma < 1$. Notice that the error ϵ_N decreases as an exponential function of N. Then $\bar{X}$ is constrained by

$$\bar{X}_N \leq \bar{X} \leq \bar{X}_N + \epsilon_N. \tag{15}$$

Now that we have gauged the error ϵ_N due to not considering an infinite number of TSs in each MDP realization, we consider next the error due to estimating $\bar{X}_N$ over a finite number of MDP realizations. We can rewrite $\bar{X}_N$ as

$$\bar{X}_N = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\left(\sum_{n=0}^{N}\gamma^n X_n^t D_n^t\right), \tag{16}$$

where X_n^t and D_n^t correspond to the action taken and data size in the TS n of the tth MDP realization, respectively. We denote by $\hat{X}_N^T$ the sample mean estimate of $\bar{X}_N$ for T realizations as:

$$\hat{X}_N^T = \frac{1}{T}\sum_{t=0}^{T}\left(\sum_{n=0}^{N}\gamma^n X_n^t D_n^t\right). \tag{17}$$

Using the Central Limit Theorem, if T is large, we can assume that $\hat{X}_N^T$ is a random variable with normal distribution and by applying the Tchebycheff inequality [37] we can compute the confidence intervals for $\hat{X}_N^T$

$$P(\hat{X}_N^T - \epsilon_T < \bar{X}_N < \hat{X}_N^T + \epsilon_T) = \delta, \tag{18}$$

where $\epsilon_T \triangleq t_{\frac{1+\delta}{2}}(T)\frac{\hat{\sigma}}{\sqrt{T}}$, with $t_a(b)$ denoting the Student$-t$ a percentile for b samples and the variance $\hat{\sigma}$ is estimated using

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=0}^{T}\left(\sum_{n=0}^{N}X_n^t D_n^t - \hat{X}_N^T\right)^2. \tag{19}$$

Finally, the confidence interval for the estimate $\hat{X}_N^T$ of $\bar{X}$ is

$$P(\hat{X}_N^T - \epsilon_T < \bar{X} < \hat{X}_N^T + \epsilon_T + \epsilon_N) = \delta. \tag{20}$$

where ϵ_N is defined in (14). In our numerical analysis we compute the confidence intervals for $\delta = 0.9$.

Remark 3. In the throughput optimization problem we assume that, given the stationarity of the underlying Markov processes, the expected throughput achieved in a sufficiently large number of TSs is the same as the expected throughput over an infinite horizon. Thus, by setting ϵ_N to zero, the computation of the confidence intervals for the TM-problem is analogous to the ETD-problem.

References

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

[2] B. Devillers and D. Gündüz, "A general framework for the optimization of energy harvesting communication systems," *J. of Commun. and Nerworks., Special Issue on Energy Harvesting in Wireless Networks*, vol. 14, no. 2, pp. 130–139, Apr. 2012.

[3] O. Orhan, D. Gunduz, and E. Erkip, "Throughput maximization for an energy harvesting communication system with processing cost," in *IEEE Information Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012.

[4] K. Tutuncuoglu and A. Yener, "Sum-rate optimal power policies for energy harvesting transmitters in an interference channel," *J. of Commun. and Nerworks., Special Issue on Energy Harvesting in Wireless Networks*, vol. 14, no. 2, pp. 151–161, Apr. 2012.

[5] 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. 1712–1731, Sep. 2011.

[6] C. Huang, R. Zhang, and S. Cui, "Throughput maximization for the Gaussian relay channel with energy harvesting constraints," *ArXiv e-prints*, Sep. 2011.

[7] D. Gündüz and B. Devillers, "Multi-hop communication with energy harvesting," in *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, San Juan, PR, December 2011.

[8] 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, Sep. 2011.

[9] M. Gregori and M. Payaró, "Optimal power allocation for a wireless multi-antenna energy harvesting node with arbitrary input distribution," in *International Workshop on Energy Harvesting for Communication (ICC'12 WS - EHC)*, Ottawa, Canada, Jun. 2012.

[10] 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. 4, pp. 547–551, Apr. 2009.

[11] A. Sinha and P. Chaporkar, "Optimal power allocation for a renewable energy source," in *Communications (NCC), 2012 National Conference on*, Kharagpur, India, feb. 2012, pp. 1–5.

[12] Z. Wang, A. Tajer, and X. Wang, "Communication of energy harvesting tags," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1159–1166, Apr. 2012.

[13] H. Li, N. Jaggi, and B. Sikdar, "Relay scheduling for cooperative communications in sensor networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 2918–2928, Sep. 2011.

[14] C. K. Ho and R. Zang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *ArXiv e-prints*, Mar. 2011.

[15] R. Srivastava and C. E. Koksal, "Basic tradeoffs for energy management in rechargeable sensor networks," *submited to IEEE/ACM Trans. Netw.*, Jan. 2011.

[16] Z. Mao, C. E. Koksal, and N. B. Shroff, "Near optimal power and rate control of multi-hop sensor networks with energy replenishment: Basic limitations with finite energy and data storage," *IEEE Trans. Autom. Control*, vol. 57, no. 4, pp. 815–829, Apr. 2012.

[17] R. E. Bellman, *Dynamic Programming*. Princeton, N.J.: Princeton University Press, 1957.

[18] A. Kansal and M. B. Sirvastava, "An enviromental energy harvesting framework for sensor networks," in *International Symposium on Low Power Electronics and Design (ISPLED)*, Tegernsee, Germany, Aug. 2003.

[19] J. Hsu, A. Kansal, S. Zahedi, M. B. Srivastava, and V. Raghunathan, "Adaptive duty cycling for energy harvesting systems," in *International Symposium on Low Power Electronics and Design (ISPLED)*, Seoul, Korea, Oct. 2006, pp. 180–185.

[20] C. M. Vigorito, D. Ganesan, and A. G. Barto, "Adaptive control of duty cycling in energy-harvesting wireles sensor networks," in *IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, San Diego, Ca, USA, 2007, pp. 21–30.

[21] C. H. Roy, C.-T. Liu, and W.-M. Lee, "Reinforcement learning-based dynamic power management for energy harvesting wireless sensor network," in *Next-Generation Applied Intelligence*, ser. Lecture Notes in Computer Science, B.-C. Chien, T.-P. Hong, S.-M. Chen, and M. Ali, Eds. Springer Berlin / Heidelberg, 2009, vol. 5579, pp. 399–408.

[22] A. Aprem, C. R. Murthy, and N. B. Mehta, "Transmit power control with ARQ in energy harvesting sensors: A decision-theoretic apporach," in *To appear in IEEE Globecom 2012*, Anaheim, CA, USA, Dec. 2012.

[23] R. E. Bellman, "A Markovian Decision Process," *Journal of Mathematical Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learing: An Introducition*, A. B. Book, Ed. Cambridge, MA: MIT Press, 1998.

[25] P. Y. Glorennec and L. Jouffe, "Fuzzy Q-learning," in *IEEE international conference onn Fuzzy Systems*, Jul. 1997, pp. 659–662.

[26] M. L. Putterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA: Wiley-Interscience, 2005.

[27] Y. Mansour and S. Singh, "On the complexity of policy iteration," in *Proceedings of the 15th International Conference on Uncertainty in AI*, Stockholm, SE, 1999, pp. 401–408.

[28] C. J. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, Psychology Department., 1989.

[29] E. Even-Dar and Y. Mansour, "Learning rates for Q-learning," *Journal of Machine Learning Research*, vol. 5, pp. 1–25, Dec. 2003.

[30] A. Atamtürk and M. W. P. Savelsberg, "Integer-programming software systems," *Annals of Operations Research*, vol. 140, no. 1, pp. 67–124, Nov. 2005.

[31] W. Zhang, "Branch-and-bound search algorithms an theri computational complexity," USC/Information Sciences Institute, Tech. Rep., M 1996.

[32] *IEEE 802.15.4e Draft Standard: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)*, IEEE Std., March 2010.

[33] S. Chalasani and J. Conrad, "A survey of energy harvesting sources for embedded systems," in *Southeastcon, 2008. IEEE*, april 2008, pp. 442 –447.

[34] A. Galindo-Serrano, L. Giupponi, and M. Dohler, "Cognition and docition in ofdma-based femtocell networks," in *IEEE Globecomm*, Miami, Florida, USA, Dec. 2010, pp. 6–10.

[35] M. Berkelaar, K. Eikland, and P. Notebaert, "Open source (mixed-integer) linear programming system: lpsolve v. 5.0.0.0," *[Available: http://lpsolve.sourceforge.net]*, 2004.

[36] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Machine Learning, Special Issue on Reinforcement Learning*, vol. 22, pp. 159–196, 1996.

[37] A. Papoulis, *Probability, random variables, and stochastic processes.* New York: Probability, random variables, and stochastic processes, 1965.