An Actor-Critic Reinforcement Learning Approach for Energy Harvesting Communications Systems

Ala'eddin Masadeh, Zhengdao Wang, Ahmed E. Kamal
Iowa State University (ISU), Ames, IA 50011, USA,
emails: {amasadeh,zhengdao,kamal}@iastate.edu

(Invited Paper)

***Abstract*—Energy harvesting communications systems are able to provide high quality communications services using green energy sources. This paper presents an autonomous energy harvesting communications system that is able to adapt to any environment, and optimize its behavior with experience to maximize the valuable received data. The considered system is a point-to-point energy harvesting communications system consisting of a source and a destination, and working in an unknown and uncertain environment. The source is an energy harvesting node capable of harvesting solar energy and storing it in a finite capacity battery. Energy can be harvested, stored, and used from continuous ranges of energy values. Channel gains can take any value within a continuous range. Since exact information about future channel gains and harvested energy is unavailable, an architecture based on actor-critic reinforcement learning is proposed to learn a close-to-optimal transmission power allocation policy. The actor uses a stochastic parameterized policy to select actions at states stochastically. The policy is modeled by a normal distribution with a parameterized mean and standard deviation. The actor uses policy gradient to optimize the policy's parameters. The critic uses a three layer neural network to approximate the action-value function, and to evaluate the optimized policy. Simulation results evaluate the proposed architecture for actor-critic learning, and shows its ability to improve its performance with experience.**

***Index Terms*—Energy harvesting, Markov decision process, actor-critic, reinforcement learning, neural networks.**

I. Introduction

Recently, energy harvesting (EH) has been considered as one of the promising technologies used to implement sustainable wireless communications devices. EH is defined as a technology converting ambient energy into usable electric energy [1]. Some attractive properties of EH communications are summarized as follows. The lifetimes of EH devices are determined by their hardware lifetimes due to their ability to recharge their batteries. In addition, the possibility of deploying these devices in hard-to-reach places [2].

Implementing efficient EH communications networks should consider the main challenges facing this type of networks, such as limited amount of energy that can be harvested, and changing harvestable energy and channel gain with time. To overcome such challenges, especially when the underlying EH and channel gain processes change in unknown patterns, autonomous EH devices can be deployed, which have the capability of managing the use of harvested energy efficiently. These devices are supported by algorithms enabling them to adapt with any environment, and to improve their performance with experience.

Implementing autonomous EH devices in environments with unknown EH and channel gain processes has been discussed in [3]–[9]. One of the promising approaches used in the management of harvested energy autonomously is reinforcement learning (RL). RL is known as algorithms enabling an autonomous agent to optimize its performance in unknown environments [10], [11]. RL methods can be categorized into value-based RL methods and policy gradient RL methods. Value-based RL is defined as methods learning a value function, and then, select actions according to the approximated value function. Policy gradient RL is defined as methods that learns a parameterized policy, which is able to select actions without consulting a value function. Using this type of learning, value functions may be used to learn the policy's parameters, but they are not needed for actions selection [10].

In [5], an EH point-to-point communications system is investigated. The EH and channel gain processes are Markov processes. Value-based RL is used to learn a transmission power allocation policy that maximizes the valuable received data (i.e., the data that can be utilized when it is received). The values of actions (transmission power levels) are approximated using state-action-state-action-reward (SARSA) prediction method. The approximated values of actions are used by an exploration algorithm called convergence-based algorithm to select actions at states.

In [6], an EH point-to-point communications system is discussed. The energy and data arrivals are formulated as Markov processes. Value-based RL is used to find the optimal transmission policy that maximizes the expected total transmitted data when the transmitter is active. Q-learning prediction method is used to estimate the values of actions (to transmit or to drop a packet). The ϵ-greedy exploration algorithm uses the approximated values of actions to select actions at states.

In [7]–[9], value-based RL and policy gradient are combined to implement what is called actor-critic RL. The actor uses a stochastic policy to select actions at states, and uses policy gradient to optimize the policy's parameters to maximize a performance measure. The critic approximates a value function and evaluates the policy optimized by the actor. In [7], user scheduling and resource allocation in heterogeneous networks powered by hybrid energy is studied, where part of the used

978-1-7281-1856-7/19/$31.00 ©2019 IEEE

energy is harvested renewable energy and the other part is from conventional resources. The goal is to maximize energy efficiency of the overall network. The problem is formulated as an RL problem with continuous state and action spaces. An actor-critic algorithm is proposed to obtain the optimal policy for the formulated problem. This algorithm assumes that the state distribution function under a policy is known. The actor uses policy gradient to optimize the policy, while the critic uses linear feature-based function approximation to approximate the action-value function.

In [8], the problem of energy management for EH wireless sensor nodes is considered, and formulated as an RL problem with continuous state and action spaces. An algorithm based on actor-critic learning, called RLMan, is introduced. The policy is modeled by a normal distribution with parameterized mean and fixed standard deviation to select actions stochastically at states. The actor uses policy gradient to optimize the policy's parameters. The critic uses linear function approximation to approximate the value function and evaluate the optimized policy by the actor.

In this paper, we investigate the problem of maximizing the valuable received data for a point-to-point EH communications system working in unknown and uncertain environments. The problem is formulated as an RL problem with continuous state and action spaces. An architecture based on actor-critic learning is proposed to optimize the performance of the considered system without knowing the steady state distribution under a policy. The actor optimizes a stochastic policy modeled by a normal distribution with parameterized mean and standard deviation. The critic uses a three layer neural network to approximate the action-value function and evaluate the optimized policy.

The remainder of the paper is organized as follows. Section II describes the proposed system. The problem formulation is given in Section III. Then, the proposed solution is discussed in Section IV. Numerical simulation results are presented in Section V. Finally, the paper is concluded in Section VI.

II. System Model

A point-to-point EH communication system consisting of a source (S) and a destination (D), and working in an uncertain and unknown environment is investigated. As illustrated in Fig. 1, each of S and D is capable of storing data in an infinite buffer. S is an EH node that is able to harvest solar energy and store it in a finite capacity battery. Time is slotted into equal length slots, where each slot has a duration of T_c. The maximum capacity of the used battery is $B_{\max}$ J. B_i is the battery level of S at the beginning of time slot i, where $B_i \in \mathcal{B} \triangleq [0, B_{\max}]$.

The EH and channel gain between slots are governed by Markov processes. During time slot i, S is harvesting E_i J from solar sources, where $E_i \in \mathcal{E}_n \triangleq [E_{\min}, E_{\max}]$. $f_{\mathcal{E}_n}(e'|e)$ is the transition probability density function for transiting from energy level e to energy level e'. Let H_i be the channel gain from S to D during time slot i, where $H_i \in \mathcal{H} \triangleq [H_{\min}, H_{\max}]$.

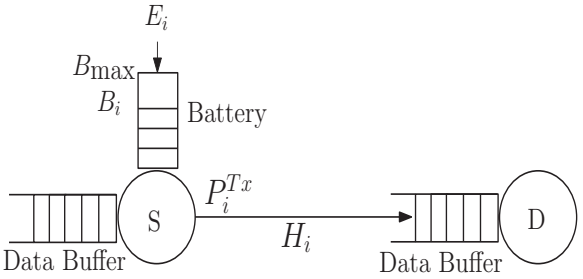

Fig. 1. Point-to-point communication system with an energy harvesting source.

$f_{\mathcal{H}}(h'|h)$ is the transition probability density function for transiting from channel gain h to channel gain h'. The channel gain at time i, H_i, is estimated using pilot signals known to both S and D. By measuring the received pilot signals, the channel gain can be estimated.

Let P_i^{Tx} be the transmission power during time slot i. $P_i^{Tx} \in \mathcal{P}^{Tx} \triangleq [0, p_{\max}^{Tx}]$. In this work, energy consumption is considered only due to data transmission, and it does not consider any other energy consumption, such as processing, circuitry, etc. P_i^{Tx} is the decision variable that will be determined in order to maximize the amount of data transmitted from S to D. This system uses harvest-store-use scheme to manage the harvested energy [12], [13].

III. Problem Formulation

Since the exact values of harvested energy levels and channel gains are unknown in the future, the decision making problem is formulated as a Markov decision process (MDP) with continuous state and action spaces. The mathematical model of an MDP with continuous state and action spaces is defined by the following principles:

- A continuous set of states $\mathcal{S}$.
- A continuous set of actions $\mathcal{A}$.
- $f(s'|s,a)$ is the transition probability density function defining the transition from state s to state s' given action a is taken at state s.
- The immediate reward, $r(s,a,s')$, is attained by taking action a at state s and then transiting to state s'.
- The discount factor γ.

The formulated MDP of our decision making problem is described as follows. Each state s is defined by the battery level b, channel gain h, and amount of harvested energy e, where $s = (b, h, e)$. The action a is the transmission power p^{Tx}. Each state s has a subset of actions $\mathcal{P}_s^{Tx}$ such that $\mathcal{P}_s^{Tx} \subseteq \mathcal{P}^{Tx}$. Battery levels evolve according to

$$b' = \min\{b + e, B_{\max}\} - T_c\, p^{Tx}. \tag{1}$$

The transition probability density function $f(s'|s, p^{Tx})$ is given by

$$f(s'|s,p^{Tx}) = \begin{cases} f_{\mathcal{E}_n}(e'|e) \cdot f_{\mathcal{H}}(h'|h), & \text{if (1) is satisfied} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where the channel gain and EH processes are independent.

The immediate reward, which is the amount of transmitted data from S to R resulting from taking action p^{Tx} at state s is given by

$$r(s, p^{Tx}) = T_c \log_2 \left(1 + \frac{p^{Tx}\,|h|^2}{\sigma_n^2}\right). \tag{3}$$

In this context, the immediate reward is a function of the current state s and the selected action p^{Tx} only, and it is independent of the next state s'.

A stochastic parameterized policy $\pi(p^{Tx}|s, \boldsymbol{\theta})$ maps states to actions stochastically, where the goal is to find the optimal policy's parameter vector that maximizes a performance measure $J(\boldsymbol{\theta})$.

IV. Proposed Solution

Due to unavailability of the transition probability density functions, and having a continuous state and action spaces, RL is used. RL Methods used for learning an optimal policy can be categorized into two classes, which are value-based RL methods, and policy gradient RL methods. Value-based RL methods learn the values of actions, and then, actions are selected according to their estimated values (i.e., policies are extracted from the estimated actions' values). Value-based methods use a sequence of policy evaluation and policy improvement cycles. Policy evaluation is used to estimate a value function for the agent's current policy, while policy improvement is used to improve the policy based on the new estimated value function [10].

Policy gradient RL methods optimize the parameters of a parameterized stochastic policy, which selects actions without consulting a value function. Value functions may be used to learn the policy's parameters, but they are not needed to select actions at states [10]. Policy gradient methods are characterized by a number advantages, which are summarized as follows. The first one is the ability to learn mixed strategies, which are balanced stochastically. The second one is their convergence properties, which are better than those of value-based methods. They are able to converge to at least a local optimal policy. The third advantage is their capability of learning in problems with continuous action spaces [14]. Actor-critic methods are learning algorithms combining value-based and policy gradient RL methods. Actor-critic algorithms mainly consist of an actor and a critic. The actor estimates a value function, while the critic optimizes the policy's parameters. Fig. 2 [15] shows the interaction between the actor and the critic in the actor-critic architecture.

In this paper, we proposed an actor-critic algorithm to maximize the valuable received data, and below, we describe our proposed implementation of the critic and the actor phases.

A. Actor

In this context, the actor uses policy gradient to optimize a parameterized stochastic policy $\pi(p^{Tx}|s, \boldsymbol{\theta})$. Policy gradient aims at maximizing the average value of the states

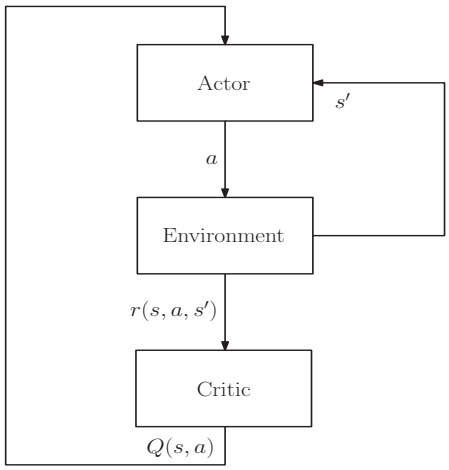

Fig. 2. Actor-critic architecture.

$$J(\boldsymbol{\theta}) = \int_{\mathcal{S}} d^{\pi}(s) \int_{\mathcal{P}^{Tx}} \pi(p^{Tx}|s, \boldsymbol{\theta})\, q_{\pi}(s, p^{Tx}) dp^{Tx}\, ds. \tag{4}$$

where transmitting data from S to D is a continuing task, $d^{\pi}(s)$ is the steady-state distribution of the underlying MDP using the policy $\pi(p^{Tx}|s, \boldsymbol{\theta})$, and $q_{\pi}(s, p^{Tx})$ is the action-value of state-action pair (s, p^{Tx}) under policy $\pi(p^{Tx}|s, \boldsymbol{\theta})$. Tasks can be classified into to episodic tasks and continuing tasks. In episodic tasks, the agent-environment interaction breaks into subsequences called episodes, such as plays in a game. On the other hand, continuing tasks refer to tasks where the agent-environment interaction does not break into episodes and continues without limit [10]. In continuing tasks, average value of the states or the average reward per time-step are used to evaluate stochastic parameterized policies when policy gradient is used [10], [16]. The policy's parameter vector $\boldsymbol{\theta}$ is updated according to

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \tag{5}$$

where $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is the gradient of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and β is the learning rate.

One of the main challenges in finding $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is to ensure improvement during changing $\boldsymbol{\theta}$, since changing $\boldsymbol{\theta}$ will change the policy and the states' distribution at the same time. The other challenge is that the effect of $\boldsymbol{\theta}$ on the states' distribution is unknown. Policy gradient theorem provides an expression for $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ that does not involve the derivative of the states' distribution with respect to $\boldsymbol{\theta}$ [10]. According to this theorem, for any differentiable policy, $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is approximated by [16]

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx E_{\pi}[\nabla_{\boldsymbol{\theta}} \ln(\pi(p^{Tx}|s, \boldsymbol{\theta}))\, Q(s, p^{Tx}, \boldsymbol{w})], \tag{6}$$

where $Q(s, p^{Tx}, \boldsymbol{w})$ is the approximated action-value function by the critic.

Due to the difficulty of finding $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, the stochastic estimate $\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}$ that approximates $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is used [10], [16], and $\boldsymbol{\theta}$ is updated according to

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}, \tag{7}$$

where

$$\widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})} = \nabla_{\boldsymbol{\theta}} \ln(\pi(p^{Tx}|s,\boldsymbol{\theta}))\, Q(s,p^{Tx},\boldsymbol{w}). \quad (8)$$

In this work, the parameterized policy $\pi(p^{Tx}|s,\boldsymbol{\theta})$ is modeled by a normal distribution with parameterized mean $\mu(s,\boldsymbol{\theta}_\mu)$ and standard deviation $\sigma(s,\boldsymbol{\theta}_\sigma)$. $\pi(p^{Tx}|s,\boldsymbol{\theta})$ is given by

$$\pi(p^{Tx}|s,\boldsymbol{\theta}) = \frac{1}{\sqrt{2*\pi*(\sigma(s,\boldsymbol{\theta}_\sigma))^2}}\, \exp\left(-\frac{(T_c\, p^{Tx}-\mu(s,\boldsymbol{\theta}_\mu))^2}{2*(\sigma(s,\boldsymbol{\theta}_\sigma))^2}\right), \quad (9)$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma]^\top$, and π is the number $\pi \approx 3.14159$.

The mean $\mu(s,\boldsymbol{\theta}_\mu)$ should be within the range of minimum and maximum values at each state. Due to this constraint, the hyperbolic tangent function, which restricts the output between -1 and 1, is used to model the parameterized mean

$$\mu(s,\boldsymbol{\theta}_\mu) = [\mu_{\max}(s) - \mu_{\min}(s)]\left(\frac{1+\tanh(\boldsymbol{\theta}_\mu^\intercal\, \boldsymbol{\phi}(s))}{2}\right) + \mu_{\min}(s), \quad (10)$$

where $\mu_{\max}(s) = b$ is the maximum value that can be assigned to the mean at state s, which is the current battery level, $\mu_{\min}(s) = 0$ is the minimum value that can be assigned to the mean at state s, and $\boldsymbol{\phi}(s)$ is a vector of features at state s. $\boldsymbol{\phi}(s)$ is a vector of two binary feature functions. The first feature is related to energy availability at state s. It is set to one if the available energy is more than zero; otherwise it is set to zero. The second feature function is related to energy overflow at state s. If the current energy level is the maximum capacity of the battery, it is set to one, otherwise, it is set to zero.

The standard deviation should be positive, so, it is modeled by an exponential with a linear exponent [10]

$$\sigma(s,\boldsymbol{\theta}_\sigma) = \exp(\boldsymbol{\theta}_\sigma^\intercal\, \boldsymbol{\phi}(s)). \quad (11)$$

$\boldsymbol{\theta}$ is updated according to

$$\boldsymbol{\theta}_\mu \leftarrow \boldsymbol{\theta}_\mu + \beta[\nabla_{\boldsymbol{\theta}_\mu} \ln(\pi(p^{Tx}|s,\boldsymbol{\theta}))\, Q(s,p^{Tx},\boldsymbol{w})], \quad (12)$$

$$\boldsymbol{\theta}_\sigma \leftarrow \boldsymbol{\theta}_\sigma + \beta[\nabla_{\boldsymbol{\theta}_\sigma} \ln(\pi(p^{Tx}|s,\boldsymbol{\theta}))\, Q(s,p^{Tx},\boldsymbol{w})], \quad (13)$$

which can be rewritten as

$$\boldsymbol{\theta}_\mu \leftarrow \boldsymbol{\theta}_\mu + \beta\left[\left(\frac{(T_c\, p^{Tx}-\mu(s,\boldsymbol{\theta}_\mu))}{\sigma(s,\boldsymbol{\theta}_\sigma)^2}\, \nabla_{\boldsymbol{\theta}_\mu}\mu(s,\boldsymbol{\theta}_\mu)\right)\cdot\, Q(s,p^{Tx},\boldsymbol{w})\right], \quad (14)$$

$$\boldsymbol{\theta}_\sigma \leftarrow \boldsymbol{\theta}_\sigma + \beta\left[\left(\left(\frac{(T_c\, p^{Tx}-\mu(s,\boldsymbol{\theta}_\mu))^2}{\sigma(s,\boldsymbol{\theta}_\sigma)^3} - \sigma(s,\boldsymbol{\theta}_\sigma)^{-1}\right)\nabla_{\boldsymbol{\theta}_\sigma}\sigma(s,\boldsymbol{\theta}_\sigma)\right)\cdot Q(s,p^{Tx},\boldsymbol{w})\right], \quad (15)$$

where

$$\nabla_{\boldsymbol{\theta}_\mu}\mu(s,\boldsymbol{\theta}_\mu) = \left(\frac{\mu_{\max}(s)-\mu_{\min}(s)}{2}\right)[1-\tanh^2(\boldsymbol{\theta}_\mu^\intercal\, \boldsymbol{\phi}(s))]\, \boldsymbol{\phi}(s), \quad (16)$$

and

$$\nabla_{\boldsymbol{\theta}_\sigma}\sigma(s,\boldsymbol{\theta}_\sigma) = \exp(\boldsymbol{\theta}_\sigma^\intercal\, \boldsymbol{\phi}(s))\, \boldsymbol{\phi}(s). \quad (17)$$

B. Critic

The critic part of RL agent is used to approximate the action-value function and evaluate the policy optimized by the actor. A neural network of three layers is used to approximate the action-value function, which is given by $Q(s,p^{Tx},\boldsymbol{w})$, where $\boldsymbol{w}$ is the weight vector used by the neural network. Backpropagation is used to minimize the squared temporal-difference (TD) error, $r(s,p^{Tx}) + \gamma\, Q(s',p^{Tx'},\boldsymbol{w}) - Q(s,p^{Tx},\boldsymbol{w})$, which is the difference between the target $r(s,p^{Tx}) + \gamma\, Q(s',p^{Tx'},\boldsymbol{w})$ and the old estimate $Q(s,p^{Tx})$.

V. Simulation Results

In this section, the proposed algorithm is evaluated. In the numerical analysis, it is assumed that each time slot is 1 second in duration. The available bandwidth BW is 1 MHz, and the noise spectral density is $N_0 = 4\times10^{-21}$ W/Hz.

It is also assumed that the S is equipped with a solar panel with an area of 25 cm^2 and 10% harvesting efficiency. An outdoor solar panel can get the benefit of 100 mW/cm^2 solar irradiance under standard conditions, and harvesting efficiency between 5% and 30%, depending on the material used in manufacturing the panel [17].

The used parameters are as follows. The discount factor γ is set to 0.9, and the learning rate β is selected to be 0.0002. The used battery has a maximum capacity of 3 J. All results are averaged over 300 runs. The starting state is selected randomly using a uniform distribution. The EH and channel gain processes are Markov processes. $E_i \in \mathcal{E}_n \triangleq [0, 0.25]$ J is a continuous random variable with a normal distribution $f_{\mathcal{E}_n}$ with standard deviation 0.1 and mean equals to E_{i-1}. $H_i \in \mathcal{H} \triangleq [0.1, 1]$ is a continuous random variable with a normal distribution $f_{\mathcal{H}}$ with standard deviation 0.1 and mean equals to H_{i-1}.

A. The cumulative discounted return for actor-critic versus hasty

In this experiment, the implemented actor-critic is compared with the hasty policy. The proposed actor-critic model uses a critic that is implemented from a 3 layer neural network. The first and second layers have 10 and 5 neurons respectively, with ReLU activation function. The third layer has one linear neuron. Using hasty policy, all available energy is allocated for data transmission each time, regardless of previous experience (i.e., using a greedy approach). The goal is to avoid energy overflow situations [3].

Fig. 3 shows the discounted return G_t (i.e., the cumulative discounted received data starting from time t). The cumulative discounted received data refers to the amount of valuable data received within a given period of time. The discounted return G_t of the hasty algorithm takes a near-constant shape all the time, since this algorithm uses one policy all the time, and the discount factor γ which restricts the discounted return to a certain value. For the actor-critic, it starts from a random policy. At the beginning, its discounted return increases significantly with experience. As time increases, the discounted return starts taking a near-constant pattern, which is due to

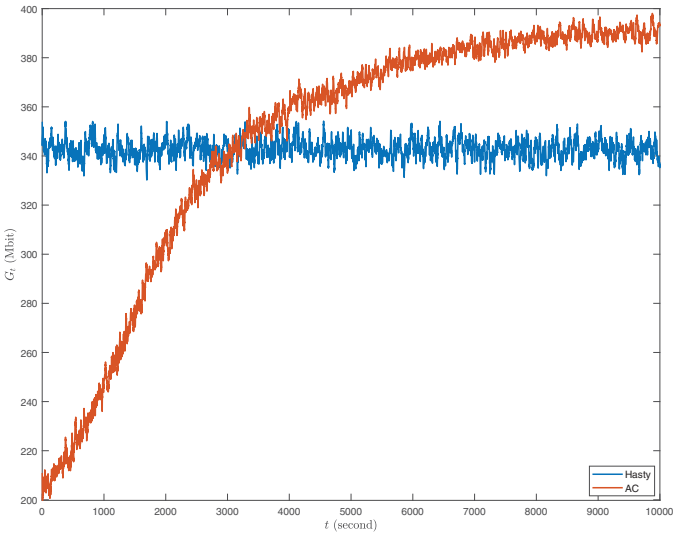

Fig. 3. The cumulative discounted return G_t versus t.

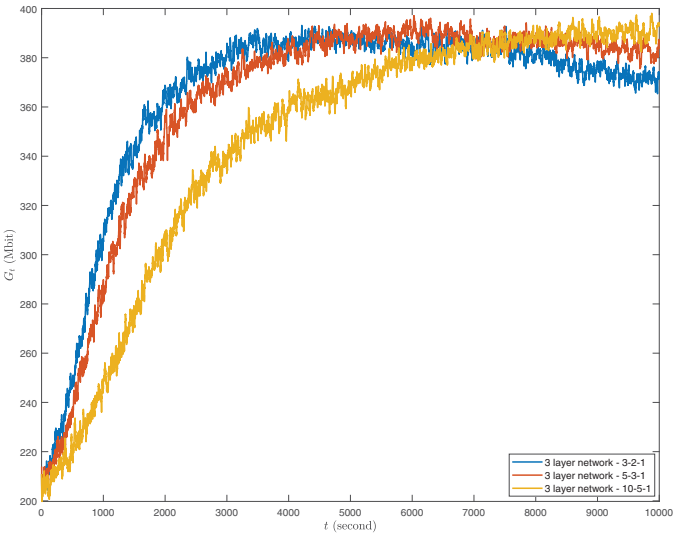

Fig. 4. The cumulative discounted return G_t versus t.

learning a suboptimal policy that can not be further improved, and the discount factor γ that restricts the discounted return to a particular value.

B. The effect of the approximated value function on the cumulative discounted return

In this part, different architectures of the neural network used by the critic are considered. The goal is to investigate the effect of the approximated action-value function on the cumulative discounted return. The considered architectures of the neural network are, three layer neural network with 3 and 2 neurons at the first and second layers, respectively, three layer network with 5 and 3 neurons at the first and second layers, respectively, and three layer network with 10 and 5 neurons at the first and second layers, respectively. The neurons in the first and second layers are neurons with ReLU activation function, while the neuron at the third layer is a neuron with a linear activation function.

Fig. 4 shows the performance of the considered architectures, which depends on their accuracy in approximating the action-value function. As the accuracy of the estimated action-value function increases, the actor will be able to optimize its policy more precisely in a direction that maximizes the cumulative discounted return. The best performance is achieved by a three layer neural network with 10 and 5 neurons at the first and second layers, respectively. This figure also shows that the performance of the other two neural networks degrades over time. As time increases, the numbers of visited states and selected actions increase, which increases the complexity of the action-value function to be approximated. Increasing number of neurons in the first and second layers handles the problem of increasing the complexity of the action-value function, and improves the accuracy of the approximated function.

C. Solar cell efficiency

In this experiment, some semiconductors used in manufacturing solar cells [18] are evaluated using our proposed algorithm, and compared with the theoretical upper bound (100% harvesting efficiency). The following results are for singlejunction cells, where the efficiencies are measured for (100 mW/cm^2) solar irradiance at 25°C. The considered materials are silicon (Si) (thin film minimodule) with $10.5 \pm 0.3\%$ efficiency, gallium arsenide (GaAs) (thin film cell) with $28.8 \pm 0.9\%$ efficiency, and cadmium telluride (CdTe) (cell) with $21.0 \pm 0.4\%$ efficiency. The critic uses a 3 layer neural network to approximate the action-value function. The numbers of neurons in the first and second layers are 10 and 5, respectively, with ReLU activation function. The third layer has one linear neuron.

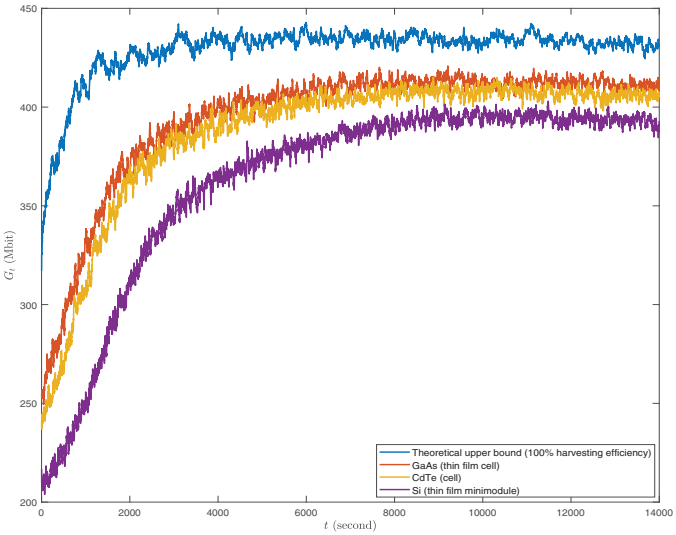

Fig. 5. The cumulative discounted return G_t versus t.

Fig. 5 shows the effect of the material used in manufacturing

solar cells. The performance patterns of all considered materials are improved with experience, and then, they become nearly constant. This is due to learning policies that cannot be further improved, and the discount factor that restricts the returns of different materials to certain values. As expected, GaAs (thin film cell) has the best performance compared with the other two materials.

As shown, increasing the harvesting efficiency increases the transmission power, which increases the discounted return. As the experience increases and the policies are optimized, the effect of increasing the harvesting efficiency on the discounted return gets smaller. At the beginning of learning, the used policies are random and the data transmission occurs randomly regardless of the channel states, which might be bad. As the experience increases, good channel states are learned. Using random policies, the data is transmitted at relatively low SNRs compared with the learned policies. The effect of increasing the transmission power at low SNRs is more significant compared with the effect of increasing the transmission power using the same values at high SNRs, which explains the reduction of the harvesting efficiency effect on the discounted return as the experience increases.

VI. Conclusions

In this work, a point-to-point EH communications system is investigated. This system consists of a source and a destination. The source is capable of harvesting solar energy and storing it in a finite capacity battery. The EH and channel gain processes are Markov processes with continuous spaces. The agent-environment interaction is modeled by an MDP with continuous state and action spaces. Actor-critic was used to optimize the performance of the considered system. The critic used a neural network of three layers to approximate the action-value function and evaluate the policy optimized by actor. The actor used a parameterized stochastic policy to map states to actions stochastically. The policy is modeled by a normal distribution with parameterized mean and standard deviation. The mean is modeled by the hyperbolic tangent function to restrict the mean by available actions at each state. The standard deviation is modeled by an exponential function with a linear exponent to guarantee positive values for the standard deviation. Policy gradient was used to optimize the policy's parameters to maximize the system throughput. The system performance was compared to hasty algorithm, where the results showed the ability of actor-critic learning to improve the performance of EH communications systems with experience, when these systems work in unknown and uncertain environments, and the state and action spaces are continuous. Then, the system performance was evaluated using different neural networks, and different materials used in manufacturing solar cells. Due to dealing with continuous state and action spaces, the probability of visiting new states and selecting new actions increases over time, which increases the complexity of the action-value function to be approximated. It was noticed that increasing the number of used neurons in the neural network handles the problem of increasing the complexity of the action-value function.

Acknowledgement

The work in the paper was supported in part by NSF Award 1827211 and NSF Award 1711922.

References

[1] M. Gregori and J. Gómez-Vilardebò, "Online learning algorithms for wireless energy harvesting nodes," in *Proc. of IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia*, May 2016, pp. 1–6.

[2] O. Ozel, K. Tutuncuoglu, S. Ulukus, and A. Yener, "Fundamental limits of energy harvesting communications," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 126–132, 2015.

[3] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting point-to-point communications," in *Proc. of the IEEE International Conference on Communications (ICC 2016), Kuala Lumpur, Malaysia*, May 2016, pp. 1–6.

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

[5] A. Masadeh, Z. Wang, and A. E. Kamal, "Reinforcement learning exploration algorithms for energy harvesting communications systems," in *Proc. of the IEEE International Conference on Communications (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

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

[7] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680–692, Jan 2018.

[8] F. A. Aoudia, M. Gautier, and O. Berder, "Rlman: an energy manager based on reinforcement learning for energy harvesting wireless sensor networks," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 2, pp. 408–417, June 2018.

[9] F. A. Aoudia, M. Gautier, and O. Berder, "Learning to survive: Achieving energy neutrality in wireless sensor networks using reinforcement learning," in *Proc. of the IEEE International Conference on Communications (ICC 2017), Paris, France*, May 2017, pp. 1–6.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT Press, 2018.

[11] C. Szepesvári, "Algorithms for reinforcement learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 4, no. 1, pp. 1–103, 2010.

[12] A. Alsharoa, H. Ghazzai, A. E. Kamal, and A. Kadri, "Optimization of a power splitting protocol for two-way multiple energy harvesting relay system," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 4, pp. 444–457, Dec. 2017.

[13] T. Li, P. Fan, Z. Chen, and K. B. Letaief, "Optimum transmission policies for energy harvesting sensor networks powered by a mobile control center," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6132–6145, Sep. 2016.

[14] D. Silver, "Reinforcement learning and simulation-based search in computer go," PhD thesis, University of Alberta, Edmonton, Alberta, 2009.

[15] X. Xu, D. Hu, and X. Lu, "Kernel-based least squares policy iteration for reinforcement learning," *IEEE Transactions on Neural Networks*, vol. 18, no. 4, pp. 973–992, July 2007.

[16] D. Silver, "Lecture 7: Policy gradient," http://www.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/pg.pdf, University College London, London, UK, 2015.

[17] R. Vullers, R. Schaijk, H. Visser, J. Penders, and C. Hoof, "Energy harvesting for autonomous wireless sensor networks," *IEEE Solid-State Circuits Magazine*, vol. 2, no. 2, pp. 29–38, Spring 2010.

[18] M. A. Green, Y. Hishikawa, E. D. Dunlop, D. H. Levi, J. Hohl-Ebinger, and A. W. Ho-Baillie, "Solar cell efficiency tables (version 52)," *Progress in Photovoltaics: Research and Applications*, vol. 26, no. 7, pp. 427–436, 2018.