Shallow Reinforcement Learning for Energy Harvesting Communications With Imperfect Channel Knowledge

Heasung Kim, *Member, IEEE*, Jungwoo Lee, *Senior Member, IEEE*, Wonjae Shin, *Senior Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—This study aims to address the power allocation problem to maximize the sum of the generalized mutual information, which refers to the achievable rate with imperfect channel state information, through a reinforcement learning (RL) approach in energy harvesting communications. In contrast to the conventional deep RL applications, which incur a large computational load on the devices due to the use of deep neural networks, we adopt shallow RL architectures involving the optimal structural properties pertaining to the optimal power allocation policy. To design the shallow architectures that can fully capture the desired power allocation policy, we derive the partial monotonicity of and bounds on the policy and value functions. These structural properties represent mathematical bases on which to construct the shallow architecture. We use a deterministic policy gradient method with monotonically shape-constrained approximators that allow us to avoid using overly complicated deep neural networks, which are not suitable for low-power devices. Through various experiments, we visualize the solutions derived from the proposed shallow architectures and demonstrate that the proposed method outperforms existing power allocation policies and exhibits a greater robustness due to optimal structural properties.**

***Index Terms*—Energy harvesting communications, reinforcement learning, power allocation, shallow neural network, generalized mutual information, rate maximization.**

Manuscript received December 31, 2020; revised May 30, 2021; accepted June 5, 2021. Date of publication June 23, 2021; date of current version October 4, 2021. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under Grant 2021-0-00467, Intelligent 6G Wireless Access System under Grant 2021-0-00106, the Basic Science Research Programs under the National Research Foundation of Korea (NRF) through the Ministry of Science and ICT under Grants 2019R1C1C1006806 and 2021R1A4A1030898, the BK21 FOUR Program of the NRF funded by the Ministry of Education under Grant NRF5199991514504, and by the U.S. National Science Foundation under Grant CCF-1908308. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. S. Ulukus. *(Corresponding author: Wonjae Shin.)*

Heasung Kim is with the Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea (e-mail: heasung1130@snu.ac.kr).

Jungwoo Lee is with Communications and Machine Learning Laboratory, Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea (e-mail: junglee@snu.ac.kr).

Wonjae Shin is with the Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, South Korea, and also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: wjshin@ajou.ac.kr).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/JSTSP.2021.3091842

I. Introduction

With the advent of the Internet of Things (IoT) and the hyper-connectivity era, the number of networks has increased significantly, thereby necessitating the development of economical communication methods for use in 5 G and more advanced networks. Since the advent of information theory, considerable progress has been made in increasing information transmission efficiency. In the context of signal analysis, coding theory has led to the development of near-optimal codes for additive white Gaussian noise (AWGN) channels to achieve close to the boundary of the *Shannon capacity* region. Moreover, to maximize the long term rate in communication systems, algorithms that optimally use the limited resources, such as bandwidth and transmission power, have been developed. In particular, the use of signal scheduling through power allocation in energy harvesting communications is emerging as a promising strategy for networks that are self-sustainable and energy-efficient. However, various realistic constraints such as those pertaining to limited energy, finite-sized battery, and time-varying channels impart randomness to the system model, hindering the development of power allocation policies for efficient communications. To overcome this complexity and randomness, policies involving learning-based algorithms have been developed. In particular, learning-theoretic approaches based on deep neural networks have demonstrated considerable potential in many optimization problems in communication fields. However, end-user devices may not have sufficient resources to support deep neural networks.

Multi-layered deep neural networks, well-known for having powerful representational capability, have been favored in various studies when information about the representational capability that function approximators should have is difficult to obtain. In contrast to these approaches, in this paper we use 2-layer shallow function approximators, which are relatively shallow in comparison to multi-layered deep neural networks that are widely used when information about the exact complexity of target functions is not available. Here, we provide proofs that are the basis for using such approximators, and exploit them with reinforcement learning techniques, which is a shallow reinforcement learning approach that we propose in this paper.

In more detail, we mathematically demonstrate the optimal structural properties of the desired policy, in contrast to the conventional ways of designing deep neural networks heuristically.

Based on the structural properties, shape-constrained function approximators are designed and trained with the deterministic policy gradient method. For training, the actor-critic architecture is built with the shape-constrained policy approximator and action-value function approximator, which is lightweight and robust. Since we use function approximators with monotonic constraints and the deterministic policy gradient method simultaneously, we call the proposed method DPGMC.

A. Related Work

Efficient power allocation policies are essential in designing energy harvesting communication systems, in which an insufficient power supply and self-sustainability are the key issues. If a transmitter has information regarding future channels or energy arrivals, then optimal policies are referred to as *offline* policies. These offline policies can only be obtained in ideal or static environments (in which the transmitter can easily forecast the channel states or energy arrivals). Also, they provide information on the maximum performance achievable in such energy harvesting communications environments. There has been relevant prior work on designing offline transmission policies to maximize throughput and minimize transmission completion time [1]. The previous works in [2], [3] considered the offline scenario along with battery constraints. Moreover, a recursive water-filling algorithm was developed in [4], and a fading channel was considered for an offline scenario. A power and rate allocation strategy for throughput maximization was proposed in the context of Gaussian relay channel in [5]. In addition, in an offline scenario, optimal power allocation strategies were investigated in a broadcast channel setting in [6].

In contrast to offline scenarios, the transmitters do not have information regarding future channels or energy arrivals in *online* scenarios. Baknina *et al.* [7] developed a closed-form expression for an online power allocation policy in a fixed channel environment with a finite-sized battery under Karush-Kuhn-Tucker conditions. Moreover, the Lyapunov optimization method was adopted to solve the throughput maximization problem in [8]. For these scenarios, energy harvesting communications systems are often modeled using a Markov decision process (MDP) [9], and iterative methods are used to overcome the randomness and lack of information. In an existing study, learning-based packet scheduling methods were proposed [10], and Q-learning, a popular method for tabular-based RL, was adopted. A value-based RL approach was used in [11] to develop an online policy. Realistic channel assumptions, including imperfect channel state feedback, were adopted in [12], and an online policy with value iteration was proposed.

However, table-based RL approaches consume considerable time to learn the target policies and require large storage space to store the state information generated in the system model. Gradient-based approaches for RL such as the policy gradient [13] and value-based approaches [14] can overcome the problem of the curse of dimensionality by using function approximators. In particular, gradient-based RL methods involving state-action-reward-state-action (SARSA) algorithms [15], [16] or policy gradient [17] have been used to manage a large number of states in communication systems. Multi-layered neural networks with gradient-based RL approaches have demonstrated excellent performance in complex tasks [14], [18] due to their representational capability. When implementing gradient-based RL, instead of storing all the system states in a tabular memory, appropriate function approximators can be designed and used; in this regard, deep neural networks are widely used due to their excellent representational capability. Moreover, these approaches have also been used in power allocation techniques for energy harvesting transmitters [19] and energy harvesting management systems [20]. The authors in [21] used the gradient-based deep Q-learning approach to solve a power allocation problem involving a modulation level adjustment.

The use of highly complex function approximators makes it challenging to interpret the obtained solutions and causes the transmitter to suffer heavy computational loads. In computational learning theory [22], the use of complex function approximators such as deep neural networks may degrade the generalization performance and lead to overfitting problems. These problems can occur not only in supervised learning but also in RL [23]. In this regard, to avoid the indiscriminate use of deep neural networks for power allocation problems, we derive the optimal structural properties of and bounds on the desired functions. In the same way as in [17], [24], which demonstrate the need for a monotonic lightweight neural network for power allocation policies with perfect channel state information, we develop interpretable shallow actor-critic architectures based on the structural properties to learn the optimal power allocation policy. Moreover, in contrast to the previous studies that involved complete channel information and one-layer neural networks [17], [24], we consider imperfect channel state information (CSI) and adopt the concept of generalized mutual information to define the reward at each time slot. Reinforcement learning agents learn the policy based on the reward, which indicates the achievable rate based on the imperfect CSI. In this setting, this paper focuses on the monotone increasing properties of the optimal power allocation policy maximizing the discounted sum of generalized mutual information with the bounds on the optimal action-value function and the optimal policy. To show the stability of the increasing function approximators for the optimal policy maximizing the discounted sum of generalized mutual information, various simulation results are provided for the random communication environments.

B. Contributions

In this study, the optimal structural properties are derived, showing how to interact between the partial monotonicity and limited bounds of the power allocation policy and value function corresponding to the actor and critic, respectively. Based on these properties, the shallow actor-critic architecture is developed, which has only sufficient representational capability to reflect the optimal structural properties for the optimal policy, as shown in Fig. 1. The shallow actor-critic architecture, which consists of shape-constrained function approximators, allows the RL agent to learn the target functions stably by preventing the framework from learning the non-optimal structural properties even under wireless communication environments with harsh

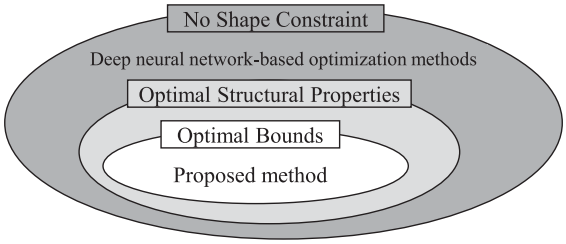


Fig. 1. Unlike the conventional optimization methods that indiscriminately use a deep neural network, the proposed method enables robust learning by exploiting optimal structural properties and bounds to construct optimized function approximators.

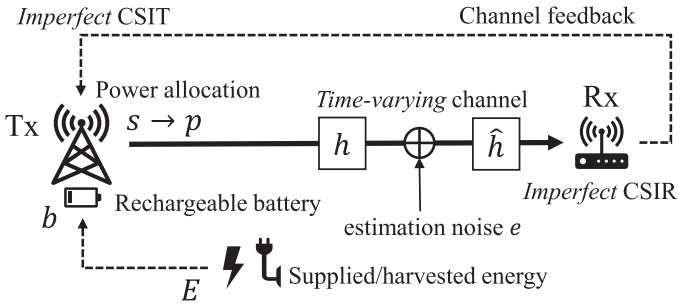


Fig. 2. An energy harvesting communications system where the transmitter has a finite-sized battery. Realistic time-varying channels and imperfect CSI are assumed. Under the conditions, our goal is to find the optimal power allocation policy π^* that optimally uses the available remaining battery every time slot.

randomness. The contributions of this work can be summarized as follows:

- To solve the practical power allocation problem of maximizing the sum of the generalized mutual information, we consider energy-constrained systems with imperfect CSI and generalized mutual information on the infinite horizon. Moreover, we specify several lemmas and theorems without any prior knowledge of the distributions of the channel gain and energy arrival, which indicates that the proposed system model can be applied to any independent and identically distributed (i.i.d.) channel models and perfect or imperfect CSI models.
- We demonstrate that the optimal power allocation policy and optimal action-value function that constitute the actor-critic architecture are limited and increasing functions of the harvested energy, imperfect channel gain, and energy reserves. Based on the mathematical proofs of the optimal structural properties, we build up shallow architectures, which involve rapid computations and impose a small computational load on the transmitter. This differs from the conventional approaches, which employ deep neural networks that can lead to the problems of overfitting and impose large loads on the transmitter without any knowledge of the structural properties of the target function.
- We compare the performance of the proposed shallow architecture and deep neural networks and initialization methods, which are widely used in the field of learning theory, and demonstrate that the proposed shallow architecture not only achieves a greater average rate but is also stable and reliable.

To the best of our knowledge, this approach represents the first technique to solve the generalized mutual information maximization problem by using shallow architectures that can provide a robust solution both theoretically and practically. Unlike naive-learning-based or iterative approaches, we provide an interpretable learning-theoretic approach based on the optimal structural properties of the target functions, which correspond to the criteria to construct the feasible function approximators.

C. Organization

The remaining paper is organized as follows. First, the generalized mutual information with imperfect CSI is introduced, and the MDP problem in the infinite horizon to maximize the generalized mutual information is formulated in Section II. Section III provides the proof for the optimal power allocation policy being an increasing function of the incoming energy, estimated channel gain, and remaining battery. Additionally, we prove that the value function, which refers to the discounted sum of the generalized mutual information, is an increasing function. The structural properties of the optimal policy and optimal value function are exploited to construct the shallow actor-critic architecture, as described in Section IV. The robustness and performance of the proposed approach are highlighted in Section V through numerical results. Concluding remarks are presented in Section VI, followed by appendices with the proofs of certain lemmas.

II. System Model and Problem Formulation

We consider an AWGN channel in which a transmitter harvests energy from external sources and transmits messages to a receiver using the harvested energy as represented in Fig. 2. The transmitter in the energy harvesting system is equipped with a finite-sized rechargeable battery with a maximum capacity of $b_{\max}$. When the transmission time is T, the transmitter consumes energy $p_i T$ at the beginning of each time slot, where p_i is the transmission power in time-slot i. The transmission power p_i can not exceed the remaining battery b_i/T at time slot i. We denote the harvested energy in time-slot i as E_i and the battery causality can then be calculated as $b_{i+1} = \min\{b_i - p_i T + E_i, b_{\max}\}$. The variables are in bounded discrete spaces $\mathcal{B}, \mathcal{E}$, and $\mathcal{P}$ as $b_i \in \mathcal{B}$, $E_i \in \mathcal{E}$, and $p_i \in \mathcal{P}(s_i)$, respectively, for all i where $\mathcal{P}(s_i) = [0, b_i] \cap \mathcal{P}$. The channel coefficient between the energy harvesting transmitter and receiver is denoted as $h_i = l_p c_i$, where $l_p = d^{-\alpha_d}$ is the path loss coefficient with a distance d and path loss exponent α_d. c_i is the channel coefficient of the distribution model, which is time-varying. Here, $H_i = |h_i|^2$ is the accurate channel gain of the transmitter in time-slot i. To consider a realistic simulation environment, it is assumed that the transmitters only have information regarding the imperfect estimated channel gain $|\hat{h}_i|^2 = \hat{H}_i$ before sending a message, and $\hat{H}_i \in \mathcal{H}$ is the estimated channel gain in time-slot i, where $\mathcal{H}$ is discrete and in bounded space. We express the state information to which the transmitter corresponds to as $s_i = (E_i, \hat{H}_i, b_i) \in \mathcal{S}$, where $\mathcal{S} = \mathcal{E} \times \mathcal{H} \times \mathcal{B}$ is the state space that is bounded and discrete. The dth element of vector s can be expressed as $s[d]$.

The *generalized mutual information* should be applied to measure the achievable rate in the imperfect CSI [25], [26]. We denote the relationship between the accurate channel state at time slot i, h_i, and estimated channel state at time slot i, $\hat{h}_i$, as follows.

$$h_i = \hat{h}_i + e \tag{1}$$

The expectation of the channel estimation error e is zero, that is, $\mathbb{E}[e] = 0$ and $\mathbb{E}[h] = \hat{h}$. When the transmitter uses the transmission power p_i in time-slot i, the immediate rate can be determined using the generalized mutual information formula as

$$I(s_i, p_i) = \log_2\left(1 + \frac{\hat{H}p}{\mathbb{E}[|e|^2]p + \sigma^2}\right) \tag{2}$$

where p is transmission power for time slot i, σ^2 is the noise density of the Gaussian channel, and $\mathbb{E}[|e|^2] = \sigma_h^2$ is variance of the channel error. If the perfect CSI is assumed, $\mathbb{E}[|e|^2] = 0$ and $\hat{H} = H$ (no error in channel estimation). In this case, (2) can be simplified as the well-known Shannon's capacity [27] which is described as

$$R(s_i, p_i) = \log_2(1 + \frac{Hp}{\sigma^2}). \tag{3}$$

For the definition of generalized mutual information [28], we do not consider the case where the variance of channel estimation error goes to infinity, and we assume that $\mathbb{E}[(|\hat{h}| - \mathbb{E}[|\hat{h}|])^2] < \infty$. In addition, it is assumed that the input has a Gaussian distribution, and a nearest neighbor decoder is used.

The key problem addressed in this study is to maximize the sum of the generalized mutual information from a long-term perspective, which is a practical form to be applied to the transmitter that operates continuously. Let us consider that the goal is to maximize the amount of information transmitted during the total time slot T_o with $\sum_{i=0}^{T_o} I(s_i, p_i)$. The transmission power in time slot i, p_i is determined by the power allocation deterministic policy π, with $\pi(s_i) = p_i$. The expected value of the achievable rate in T_o can be represented as

$$V(s) = \mathbb{E}\left[\mathbb{E}_{T_o}\left[\sum_{i=0}^{T_o} I(s_i, p_i)|s_0, \pi\right]\right]. \tag{4}$$

The objective time interval T_o can be varied and we assume that T_o follows the geometric distribution with a mean of $1/(1-\gamma)$. In this case, problem (4) can be interpreted as an optimization problem on the infinite horizon with a discount factor γ, with $(0 < \gamma < 1)$ [9, Proposition 5.3.1]. It indicates that the rate maximization problem on the finite horizon T_0 can be formulated as a discounted sum of instantaneous rate maximization problem on the infinite horizon. The objective is to maximize the rate on the infinite horizon, which is defined as

$$V(s) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i I(s_i, p_i)|s_0, \pi\right]. \tag{5}$$

V is referred to as the *value function* as it measures the value of state s. This system model for problem (5) can be modeled as an MDP, which consists of a five-element tuple $(\mathcal{S}, \mathcal{P}, p_s, r, \gamma)$ where p_s is the state transition probability function of the transmitter, and $p_s(s_{i+1}|s_i, p_i)$ is the probability of the next state s_{i+1} for a given s_i and p_i. The optimal stationary deterministic policy π^* maximizes the expected discounted reward sum (5) in the environment and can be defined as

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i I(s_i, p_i)|s_0, \pi\right] \tag{6}$$

where Π is the set of feasible policies in the MDP. By following the optimal policy π^* and allocation power, the optimal value function V^* can be obtained, which is a perfectly measure of the maximum achievable rate from state s. In this study, we primarily derive the structural properties of the optimal stationary deterministic policy π^* and the optimal value function V^* and use these values to build optimized and shallow function approximators. Using the tailored function approximators with gradient-based RL, the key goal is to obtain π^* and V^*.

III. Structural Properties of the Optimal Policy and Value Function

In this section, we demonstrate that π^* is an increasing function of E, $\hat{H}$, and b. This property is valid for the system under any i.i.d. channel distribution and energy arrival distribution. Using the monotonically increasing property, we design a function approximator optimized for π^*, as described in the previous section. To enhance the readability, we omit the time slot notation in this section. Before proving the increasing property of π^*, we introduce the key features of the generalized mutual information. Subsequently, we define the optimal value function V^* and demonstrate the increasing property of the value function. Lemmas and theorems in this section are provided to demonstrate monotonicity of the optimal policy and the optimal value function for each input element.

A. Supermodularity of the Generalized Mutual Information

In this subsection, we demonstrate that the generalized mutual information of the estimated channel gain $|\hat{H}|$ and transmission power p is supermodular.

We define the *meet* operation $\min\{x[d], y[d]\}$ and *join* operation $\max\{x[d], y[d]\}$ as $(x \wedge y)[d]$ and $(x \vee y)[d]$, respectively. The bounded discrete space $\mathcal{L} = \mathcal{H} \times \mathcal{P}$ is a lattice where $x \wedge y$, $x \vee y \in \mathcal{L}$ for every $x, y \in \mathcal{L}$. The generalized mutual information $I : \mathcal{L} \mapsto \mathcal{R}$ is supermodular if $I(x \wedge x') + I(x \vee x') \geq I(x) + I(x')$ for all $x, x' \in \mathcal{L}$. When a two-dimensional real value space domain for I is considered instaed of $\mathcal{S}$, the partial derivatives of I with respect to p can be represented as follows.

$$\frac{\partial I(\hat{H}, p)}{\partial p} = \frac{(\sigma^2\hat{H})}{\ln(2)(\sigma^2 + \mathbb{E}[|e|^2]p)(\sigma^2 + p(\hat{H} + \mathbb{E}[|e|^2]))}. \tag{7}$$

Moreover, the expression can be continuously and partially differentiated twice, as follows.

$$\frac{\partial^2 I(\hat{H}, p)}{\partial\hat{H}\partial p} = \frac{\sigma^2}{\ln 2((\hat{H} + \mathbb{E}[|e|^2])p + \sigma^2)^2}. \tag{8}$$

If (8) is non-negative, it can be considered that I is supermodular [29], in a trivial manner as the denominator and numerator in the right-hand-side term of equation (8) are positive values. This characteristic of generalized mutual information can be easily interpreted as an increasing difference, which implies that the marginal gain of information from increasing $\hat{H}$ is larger when the transmission power p is larger, as follows:

$$I(\hat{H},' p') - I(\hat{H}, p') \geq I(\hat{H},' p) - I(\hat{H}, p). \tag{9}$$

This valuable property of the generalized mutual information is exploited in the various proofs introduced in the next subsection.

B. Increasing Properties of the Value Function

The objective function (5) is the discounted reward sum on the infinite horizon. By using Bellman's equation, the maximized discounted reward sum through the optimal policy π^* can be represented in a recursive manner, as follows.

$$V^*(E, \hat{H}, b) = \max_{p \in \mathcal{P}(s)} \left\{ I(s,p) + \gamma \mathbb{E}_{\bar{E}, \bar{\hat{H}}} [V^*(\bar{E}, \bar{\hat{H}}, m(b,p,E))] \right\}. \tag{10}$$

Considering the compact domain S, which is a bounded discrete space, the upper bound of the reward function (generalized mutual information) can be identified as follows.

Corollary 1: The value function $V(s)$ is finite for all $s \in \mathcal{S}$.

Proof: The generalized mutual information has an upper bound because $\mathcal{H}$ and $\mathcal{B}$ are lattice domains with upper bounds $\max \mathcal{H}$ and $\max \mathcal{B}$, respectively. According to (2), I is an increasing function for $\hat{H}$ and p with an upper bound $\max \mathcal{B}$. The maximum value of I in $\mathcal{H} \times \mathcal{P}$ is $I_{\max} = \log_2(1 + \frac{\max \mathcal{H} \max \mathcal{B}}{\mathbb{E}[|e|^2] \max \mathcal{B} + \sigma^2})$ and

$$V(s) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i I(s_i, p_i) | s_0, \pi \right] \leq \frac{I_{\max}}{1-\gamma} < \infty. \tag{11}$$

■

An optimal stationary deterministic policy, which maximizes the long-term discounted reward sum V, exists in the system model $(\mathcal{S}, \mathcal{P}, p_s, r, \gamma)$, because the state space and action space are compact spaces and the reward function, which is I in the proposed system model, is bounded, as shown in Corollary 1 [9, Th. 6.2.10].

The value iteration algorithm can be used to obtain the optimal value function V^*, and the stationary deterministic optimal policy can be extracted from the optimal value function. The value converges to the optimal value function for a bounded discrete state and an action space with a bounded reward function. The update process of the value iteration can be represented as follows.

$$\text{For all } s \in \mathcal{S}, V_{n+1}(s) = \max_{p \in \mathcal{P}(s)} \left\{ I(s,p) + \gamma \mathbb{E}_{\bar{E}, \bar{\hat{H}}} \left[V_n(\bar{E}, \bar{\hat{H}}, m(b,p,E)) \right] \right\} \tag{12}$$

where V_n represents the value function after the nth update.

This update process requires considerable time when the state space is large because the process must be performed for all $s \in \mathcal{S}$. Consequently, we adopt a gradient-based method with the policy gradient theorem, instead of using the value iteration method. Nevertheless, we extract the valuable structural properties by using the convergence property of the value iteration update (12).

Lemma 1: The optimal value function V^* is an increasing function of the estimated channel gain $\hat{H}$ for a given E and b.

Proof: We define the optimal action p^* for s, which maximizes the nth value function, as

$$p^* = \arg\max_{p \in \mathcal{P}(s)} \left\{ I(s,p) + \gamma \mathbb{E}\left[V_n(\bar{E}, \bar{\hat{H}}, m(b,p,E)) \right] \right\}. \tag{13}$$

The same logic can be applied to state $s' = (E, \hat{H}', b)$ and we denote the optimal action p'^* for the state s' and $\hat{H}' > \hat{H}$. Let us assume that the nth value function V_n is an increasing function for $\hat{H}$, $V_n(s) \leq V_n(s')$. Then, $V_n(s')$ can be updated through (12), and

$$V_{n+1}(s') = I(s,' p'^*) + \gamma \mathbb{E}_{\bar{E}, \bar{\hat{H}}} \left[V_n(\bar{E}, \bar{\hat{H}}, m(b, p'^*, E)) \right] \tag{14}$$

$$\geq I(s,' p^*) + \gamma \mathbb{E}_{\bar{E}, \bar{\hat{H}}} \left[V_n(\bar{E}, \bar{\hat{H}}, m(b, p^*, E)) \right] \tag{15}$$

$$\geq I(s, p^*) + \gamma \mathbb{E}_{\bar{E}, \bar{\hat{H}}} \left[V_n(\bar{E}, \bar{\hat{H}}, m(b, p^*, E)) \right] \tag{16}$$

$$= V_{n+1}(s). \tag{17}$$

The inequality between (14) and (15) holds because p'^* is the optimal action for state s' although it is not necessary that p^* is the optimal action for state s'. Moreover, the inequality between (15) and (16) is satisfied because I is an increasing function of $\hat{H}$, as the partial derivative of I for $\hat{H}$ is non-negative because $\frac{\partial I(\hat{H},p)}{\partial \hat{H}} = \frac{p}{\ln 2 (p(\hat{H} + \mathbb{E}[|e|^2]) + \sigma^2)} \geq 0$. Lastly, according to the definition of the value iteration update in (12), the equality between (16) and (17) holds. When $n \to \infty$, $V_n \to V^*$. The shape of the value function can be initialized with the inequality $V_0(s') \geq V_0(s)$ and the initialization does not affect the convergence property of the value iteration. Through mathematical induction, V^* is an increasing function of $\hat{H}$ for a given E and b. ■

By using a similar approach as that to prove Lemma 1, we can demonstrate that the optimal value function has increasing properties for both E and b.

Lemma 2: The optimal value function V^* is an increasing function for the incoming energy E for a given $\hat{H}$ and b.

Lemma 3: The optimal value function V^* is an increasing function for b for a given E and $\hat{H}$.

The proofs of Lemma 2 and 3 are omitted to avoid repetition since the lemma can be proved in a similar way to the proof of 1. Combining Lemmas 1, 2, and 3, we can obtain the following theorem.

Theorem 1: Let us assume that x' dominates x if $x'[d] \geq x[d]$ for all d in $1 \leq d \leq D$. The optimal value function V^*, which is the maximum discounted sum of the generalized mutual information, is an increasing function as $V^*(s') \geq V^*(s)$ where s' dominates s.

C. Increasing Properties of the Optimal Policy

The optimal policy exhibits increasing properties in terms of its input features. In this subsection, Lemma 4, 5, and 6 are provided to show the monotone increasing property according to channel gain, remaining battery, and harvested energy, respectively. The three lemmas are based on Topkis' monotonicity theorem in [30, Th. 1] with the proof techniques from [12], which are also used in the previous studies [17], [24]. To show the monotone increasing properties of the function with arg max operator, the sufficient conditions for Topkis' monotonicity theorem are introduced and we show that the conditions are true. We first demonstrate the increasing property of the optimal policy for the estimated channel gain.

Lemma 4: The optimal policy $\pi^*(E, \hat{H}, b)$ is an increasing function for the observed channel gain $\hat{H}$ for any given E and b.

Proof: The optimal policy π^* can be represented as

$$\pi^*(s) = \min\left\{\breve{p} \in \arg\max_{p\in P(s)} \left\{I(s,p)+\gamma\mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,p,E))\right]\right\}\right\}. \quad (18)$$

In other words, if there are multiple p that can maximize the expected discounted reward sum from the given state s, it is best to choose the smallest value. The proof for the new definition of $\pi^*(s)$ can be omitted due to its clarity. To show that (18) is increasing for $\hat{H}$, Topkis' monotonicity theorem [30, Th. 1] is adopted. We define the function F, which is the term inside the arg max operation pertaining to (18) as

$$F(\hat{H}, p) = I(s,p) + \gamma\mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,p,E))\right] \quad (19)$$

where E and b are fixed. According to the monotonicity theorem [30, Th. 1], when the following two conditions are satisfied, it can be considered that (18) is an increasing function of $\hat{H}$. ■

Condition 1: $F(\hat{H}, p)$ exhibits increasing differences in $(\hat{H}, p)$. Note that an increasing difference on $(\hat{H}, p)$ indicates that the additional gain of function F when $\hat{H}$ increases is greater for a larger p.

Condition 2: The lower bound of the feasible action space of state s, $P_{lb}(s)$, and the upper bound of the feasible action space of state s, $P_{ub}(s)$, are increasing functions of s with the following constraint: $P_{lb}(s) \leq P_{ub}(s)$.

The increasing differences of F in $(\hat{H}, p)$ can be defined

$$F(\hat{H},' p') - F(\hat{H},' p) \geq F(\hat{H}, p') - F(\hat{H}, p) \quad (20)$$

$$\Leftrightarrow I(s,' p') - I(s, p') \geq I(s,' p) - I(s, p) \quad (21)$$

where $\hat{H}' > \hat{H}$, $s' = (E, \hat{H},' b)$, and $p' > p$. The equivalence between (20) and (21) indicates that (20) is true because the supermodularity and increasing difference of the generalized mutual information in (9) have been proved. The only remaining condition to prove that (18) is an increasing function of $\hat{H}$ is to prove that Condition 2 is true. The feasible action space for state s is $\mathcal{P}(s)$, and it does not depend on the observed channel gain. Consequently, the lower and upper bounds of $\mathcal{P}(s)$ and $\mathcal{P}(s')$ are identical and set as constant. This fact satisfies $P_{lb}(s) \leq P_{ub}(s)$ with equality, as well as $P_{lb}(s) \leq P_{lb}(s'), P_{ub}(s) \leq P_{ub}(s')$.

Because it is demonstrated that conditions 1 and 2 are satisfied by the supermodularity of the generalized mutual information and properties of the feasible action space, Topkis' monotonicity theorem [30, Th. 1] can be adopted as follows: To realize the proof by contradiction, we assume that the optimal power allocation policy π^* is not an increasing function of $\hat{H}$. Then, $\pi^*(s') < \pi^*(s)$ for a certain $s, s' \in \mathcal{S}$, $s' > s$. We define the optimal action for $\hat{H}'$ as p'^*, which maximizes F; then

$$0 \geq F(\hat{H},' p^*) - F(\hat{H},' p'^*) \quad (22)$$

$$\geq F(\hat{H}, p^*) - F(\hat{H}, p'^*) \quad (23)$$

$$\geq 0. \quad (24)$$

The inequality (22) holds because of the definition of the optimal actions p^* and p'^* and this aspect indicates that p'^* is the optimal action for $\hat{H}'$. The inequality between (22) and (23) is true according to Condition 1. The inequality between (23) and (24) holds according to the definitions of the optimal actions. The inequalities implies that (22), (23), and (24) must hold by equality, and thus, p'^* is an optimal action for $\hat{H}$. As we already assume that $\pi^*(s') < \pi^*(s) \Leftrightarrow p'^* < p^*$, $p^* = \arg\min F$ appears to be a contradiction because there exists another optimal action p'^* that is smaller than p^*. Therefore, the optimal policy π^* is an increasing function of $\hat{H}$ for any given E and b.

Specifically, Lemma 4 indicates that *even if the estimated channel is imperfect, for a higher estimated channel gain, more power must be used.*

In a similar approach as that used to process Lemma 4, the increasing properties of the optimal policy for the remaining battery b and harvested energy E can be demonstrated.

Lemma 5: The optimal policy $\pi^*(E, \hat{H}, b)$ is an increasing function for the remaining battery b for any given E and $\hat{H}$.

Proof: See Appendix A. ■

Lemma 6: The optimal policy $\pi^*(E, \hat{H}, b)$ is an increasing function for the harvested energy E for any given b and $\hat{H}$.

Proof of Lemma 6 is omitted to avoid repetition since the lemma can be proved in a similar way to the proof of Lemma 5. Combining Lemma 4, 5, and 6, we can obtain the following theorem.

Theorem 2: The optimal policy π^* is an increasing function as $\pi^*(s') \geq \pi^*(s)$ where s' dominates s.

Theorems 1 and 2 correspond to the structural properties of the target functions, optimal power allocation policy π^* and optimal value function V^*. As described in Sec IV-B, in contrast to the heuristic function approximator design, we build shallow function approximators which can reflect the optimal properties even before the learning begins. Note that these structural properties are derived without using any prior knowledge of the distribution of the channel gain and energy arrivals except that these aspects are i.i.d.. By exploiting Topkis Theorem [30] and the convergence property of the value iteration [9] in the same way to [17], [24], Theorems 1 and 2 reveal the structural properties of the policy and the value function. We have expanded on the

previous work by dealing with the concept of generalized mutual information.

IV. Deterministic Policy Gradient With Shallow Architecture Based on the Optimal Structural Properties

A. Actor-Critic Framework With Deterministic Policy Gradient

Before building the shallow function approximators, we describe the gradient-based RL algorithm with the actor-critic framework, which consists of the policy network and action-value function network. Unlike traditional stochastic optimization methods including value iteration and tabular-based reinforcement learning approaches, gradient-based reinforcement learning is suitable for handling a large number of states. In particular, value iteration, a well-known dynamic programming method, requires perfect knowledge of the state transition probability, and storage that can store more than the number of states. Therefore, continuous state spaces cannot be dealt with by the tabular-based method. Since we aim to deal with large discrete spaces, storing values for many states in traditional ways is inefficient and so we adopt a gradient-based learning method for the given problem with the monotone increasing properties of the optimal policy.

We adopt the Wolpertinger policy (WP) method, which is a deterministic policy gradient method for discrete spaces [31], consisting of the policy approximator π_{θ^π} and action-value function approximator Q_{θ^Q}. The two networks that compose the actor-critic framework are connected, and the weights in the networks are updated with different gradient directions. The optimal action-value function Q^*, which indicates the maximum achievable rate when the agent select action p at state s, satisfies the Bellman's optimality equation as

$$Q^*(E,\hat{H},b,p) = I(s,p) + \gamma \mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,p,E))\right]. \quad (25)$$

The goal of this model-free RL algorithm is to obtain the function approximators that have completed training as $\pi_{\theta^\pi} \approx \pi^*$ and $Q_{\theta^Q} \approx Q^*$. The actor (policy) network of the WP method maps continuous policy values from the function approximator $f_{\theta^\pi}(s)$ to the feasible discrete space as

$$\pi_{\theta^\pi}(s;\theta) = \arg\min_{p\in\mathcal{P}(s)} |p - f_{\theta^\pi}(s)|_2. \quad (26)$$

The weights in π_θ can be updated in the direction to maximize the following gradient

$$\nabla_{\theta^\pi} J \approx \mathbb{E}[\nabla_{\theta^\pi} f_{\theta^\pi}(s) \nabla_{\check{p}} Q_{\theta^Q}(s,\check{p})] \quad (27)$$

where $\check{p} = f_{\theta^\pi}(s)$. To use the fixed-Q target technique, which was proposed in [14] for a stable learning process, two function approximators must be constructed to approximate a target function, which have the same structure. We term the two approximators as the evaluation network and target network. During each learning process, the parameters of the evaluation network are updated at every step; however, the parameters of the target network are updated from the parameters of the evaluation network by considering a certain ratio τ. Let us define $\theta^{Q'}$ and $\theta^{\pi'}$ as the set of parameters in the target action-value network and target policy network, respectively. The critic network (action-value function approximator) is updated in the direction to minimize the following mean squared loss.

$$L = \mathbb{E}[(Q_{\theta^Q}(s_i,p_i) - r_i - \gamma Q_{\theta^{Q'}}(s_{i+1}, \pi_{\theta^{\pi'}}(s_{i+1})))^2]. \quad (28)$$

The gradient values for the policy and action-value function approximators are calculated using n^{batch}-sized mini-batches generated from the replay memory [14], which is a storage space for (s_i, p_i, r_i, s_{i+1}) data pairs.

B. Shallow Actor-Critic Architecture Based on the Optimal Structural Properties

This subsection describes the establishment of the actor-critic framework as a shallow architecture based on the structural properties of the optimal policy and optimal value function demonstrated in Theorems 1 and 2. Using these theorems, the size of the feasible set of parameters that constitute the function approximators that the learning agent must consider can be considerably reduced. To realize a more efficient reduction, we set upper and lower bounds for the output of each network.

Based on Theorem 1, we build the policy network θ^π as a monotonically increasing function in terms of its input features, incoming energy, estimated channel gain, and remaining battery. To effectively enforce the increasing property of the optimal policy to the function approximator, we adopt the piecewise linear calibration and interpolation method with a shape constraint [32]. This approach can provide a function approximator with various desirable structural properties by controlling a limited number of parameters. The notations used in [32, Sec. 9.3] are directly followed to describe the calibration function. The calibration function is described as $c_d(s[d];\alpha^{(d)})$, where s is the input vector, $s[d]$ is the dth element of s, and $\alpha^{(d)}$ is the set of parameters in the calibration function for the dth element. The range of $s[d]$ is divided into $C_d - 2$ parts and linearly scaled (piecewise linear approximation) to the unit range. We use a linear transformation layer with the calibrated input vector as

$$c(s;\alpha) = (c_1(E;\alpha^{(1)}),\ c_2(\hat{H};\alpha^{(2)}), c_3(b;\alpha^{(3)})) \quad (29)$$

where α is the set of all trainable parameters in the calibration layer. The calibrated state $c(s;\alpha)$ is the input of the interpolation layer which is denoted as ϕ. Specifically, we adopt the multilinear interpolation method [32, Sec. 3.1] and follow the same notations; this method involves linearly interpolating the output values according to the values of several input features with only a limited number of vertices. ϕ takes a $D = 3$-dimensional feature as an input vector, and the output is an $M = \Pi_d M_d$ dimensional vector where M_d is the number of vertices for dth feature as $\phi(s) \in [0,1]^M$. In particular, the kth component of ϕ can be expressed as follows [32, 2].

$$\phi_k(s) = \Pi_{d=1}^{D} s[d]^{v_k[d]}(1-s[d])^{1-v_k[d]} \quad (30)$$

where $v_k \in \{0,1\}^D$ is the kth vertex of the interpolation lattice. ϕ is not a trainable function approximator, and the interpolation

level can be adjusted by M-dimensional vector W^{π}, whose elements are known as lattice parameters. The lattice parameters can be trained, and the interpolation layer $l(s)$ is represented as

$$l(s) = {W^{\pi}}^T \phi(s). \tag{31}$$

By using the calibration layer, the input feature can be calibrated and the actor network (policy approximator) can be represented as follows.

$$f_{\theta^{\pi}}(s) = {W^{\pi}}^T \phi(c(s;\alpha)). \tag{32}$$

In summary, the input state $s = (E, \hat{H}, b)$ can be calibrated through the calibration layer, and the calibrated feature vector is the input vector of the interpolation layer. We exploit the structural properties demonstrated in the previous section to ensure that the calibration and interpolation layers have shape constraints. Specifically, the calibration and interpolation layers are expected to be increasing functions for their input features, and consequently, the policy approximator $f_{\theta^{\pi}}(s)$ is an increasing function for its input features due to Theorem 1 as

$${W^{\pi}}^T \phi(c(s';\alpha)) \geq {W^{\pi}}^T \phi(c(s;\alpha)) \tag{33}$$

where s' dominates s. Additionally, the transmission power is non-negative and cannot exceed the maximum battery capacity $b_{\max}$; consequently,

$$0 \leq {W^{\pi}}^T \phi(c(s;\alpha)) \leq b_{\max}. \tag{34}$$

This design scheme enables the two-layer function approximator (calibration and interpolation) [32] to possess the optimal structural properties derived from the optimal power allocation policy that is bounded and an increasing function of the corresponding input features.

In a similar manner, we can apply shape constraints to the action-value function approximator. The calibration layer $c(s, p; \alpha^Q)$ and interpolation layer ${W^Q}^T \phi$ are adopted to construct the critic network. We have

$$\begin{aligned} &c(s, p; \alpha^Q) \\ &= (c_1(E; \alpha^{(1)}),\, c_2(\hat{H}; \alpha^{(2)}),\, c_3(b; \alpha^{(3)}),\, c_4(p; \alpha^{(4)})), \end{aligned} \tag{35}$$

where α^Q is the set of all trainable parameters in the calibration layer and the $s[d]$ is divided into C_d^Q-2 parts for the action-value function. The action-value function approximator can be represented as

$$Q_{\theta^Q}(s, p) = {W^Q}^T \phi(c(s, p; \alpha^Q)). \tag{36}$$

According to Theorem 2, the optimal action-value function Q^* is a partially and monotonically increasing function for s.

Corollary 2: The optimal action-value function Q^* is a partially and monotonically increasing function for s, which indicates that $Q^*(s,' p) \geq Q^*(s, p)$ for $s' > s, s,' s \in \mathcal{S}$.

Proof: I is an increasing function for $\hat{H}$ as the derivative of I with respect to $\hat{H}$ is a non-negative value as

$$\frac{\partial I(\hat{H}, p)}{\partial \hat{H}} = \frac{p}{\ln(2)(\mathbb{E}[|e|^2]p + \hat{H}p + \sigma^2)}. \tag{37}$$

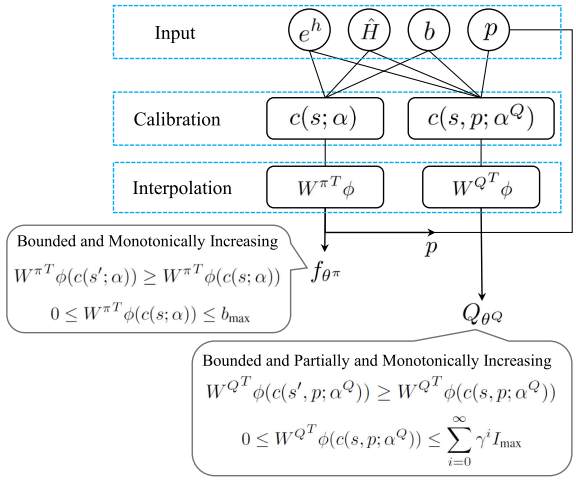


Fig. 3. Proposed actor-critic architecture. The actor (policy) network consists of the monotonic calibrators and monotonic interpolation function, and the critic network (action-value function) consists of the partially monotonic calibrators and interpolation function. This shallow structure forces the actor and critic networks to possess the optimal structural properties with only a two-layer framework.

In Theorem 1, the optimal value function V^* is an increasing function. For given p, (25) can be interpreted as a sum of the increasing functions, I and V^*. ■

Consequently, the critic network exhibits a partial monotonicity for the state s and not for the transmission power p as

$${W^Q}^T \phi(c(s,' p; \alpha^Q)) \geq {W^Q}^T \phi(c(s, p; \alpha^Q)). \tag{38}$$

Furthermore, the action-value function is bounded, with the lower and upper bounds defined as

$$0 \leq {W^Q}^T \phi(c(s, p; \alpha^Q)) \leq \sum_{i=0}^{\infty} \gamma^i I_{\max}. \tag{39}$$

The overall actor-critic architecture is illustrated in Fig. 3. Through the optimal shape constraints specified in (33), (34), (38), and (39), we can efficiently reduce the feasible space of the parameters in the function approximators.

The algorithm for the overall training phase is specified as Algorithm 1. First, the parameters in the actor and critic networks are randomly initialized. State s_i is observed, and the feasible action p_i is selected. During the training phase, p_i is generated from the current policy π with the Gaussian noise as $p_i \sim \mathcal{N}(\pi(s_i), \sigma_e^2)$ for exploration. Action p_i is applied, and the next state s_{i+1} can be obtained with a reward r_i. The data in the single loop of the training phase (s_i, p_i, r_i, s_{i+1}) are stored in the replay memory, and they can be used to approximate the gradient values in (27) and (28) through mini-batch sampling. The realization of the gradient update with the constraints specified in lines 7 and 8 in Algorithm 1 is based on the method described in [32, Sec 3.1]. The parameters constituting the networks are updated considering the ratio of τ. If no performance improvement occurs during 10 updates in Algorithm 1, the training phase is stopped early.

The calibration and the interpolation require only a small number of scalar products and summations. That calibration

Algorithm 1: Deterministic Policy Gradient With Shape Constraints.

Initialize the network weights θ^{π}, $\theta^{\pi'}$, θ^{Q} and $\theta^{Q'}$
while training phase **do**
  Observe state s_i and select p_i
  Apply action p_i and get s_{i+1}, r_i
  Store data (s_i, p_i, r_i, s_{i+1}) in replay memory
  Sample n^{batch}-sized batch from replay memory
  Update θ^{π} with the gradient ascent using (27) with shape constraints (33), (34)
  Update θ^{Q} by minimizing the loss (28) with shape constraints (38), (39)
  $\theta^{\pi'} \leftarrow (1-\tau)\theta^{\pi'} + \tau\theta^{\pi}$, $\theta^{Q'} \leftarrow (1-\tau)\theta^{Q'} + \tau\theta^{Q}$
  $s_i \leftarrow s_{i+1}$
end while

layer performs a piecewise linear transformation for each input element. Since we only have 3 input elements, a total of 3 linear transformations are performed. The interpolation function performs 3 scalar multiplications 8 times to calculate the 2^3 dimensional output. Finally, the single elementwise multiplication of the 8 dimensional vectors, W^{π} and $\phi(c(s;\alpha))$, are required for the forward propagation. Since the proposed policy approximator is based on the optimal structural properties, we can construct the approximator with only two different layers and the first calibration layer does not increase the dimensionality of the input vector. If function approximators such as deep neural networks whose representational capabilities or structural properties are difficult to analyze are applied, we cannot arbitrarily reduce the size or depth of the function approximators to reduce computation complexity.

V. Discussion

A. Simulation Environment

The transmitter is equipped with a finite-sized rechargeable battery with a capacity $b_{\max}$ and Rician fading is assumed, whose probability density function (as a function of variable x) is $x\exp(-\frac{x^2+(2K)^2}{2})I_0(2Kx)$ where $I_0(2Kx) = \sum_{k=0}^{\infty}\frac{((2Kx)^2/4)^k}{(k!)^2}$, which corresponds to the modified Bessel function with zero order. The random variables from the distribution is scaled by 0.36. Before training, the n_{replay}-sized replay memory is filled with (s_i, p_i, r_i, s_{i+1}) pairs, generated from random actions. For the gradient updates, Adam optimizer [33] is adopted. To exploit the convergence property of the value iteration algorithm, Theorems 1 and 2 are provided with a discrete state space, and they hold regardless of the size of the state space. We perform the experiment in discrete environments in which the numbers are represented with 24-bit precision, corresponding to a nearly continuous, but discrete accuracy. Since our proposed method exploits function approximators, rather than a value table, where all the values for each state are stored, it works for such discrete, but almost continuous, inputs. Note that the traditional table-based and naive RL approaches (Q-learning or value iteration) cannot be used in the case of a large state space, because a large number of states must be stored in the device memory to realize the iterative update. The various simulation parameters are listed in Table I.

TABLE I
SYSTEM MODEL AND HYPERPARAMETER SETTINGS

Description	Notation	Value
Rice shape parameter	$2K$	2.35
battery capacity	$b_{\max}$	2.0
energy harvesting probability	p^h	0.5
maximum harvested energy	$\max E$	2.0
transmission time	T	1
noise density	σ^2	1
replay memory	n_{replay}	1024
batch size	n_{batch}	128
target update ratio	τ	0.01
standard deviation of the exploration strategy	σ_e	0.1
actor learning rate	-	$1e-3$
critic learning rate	-	$1e-3$

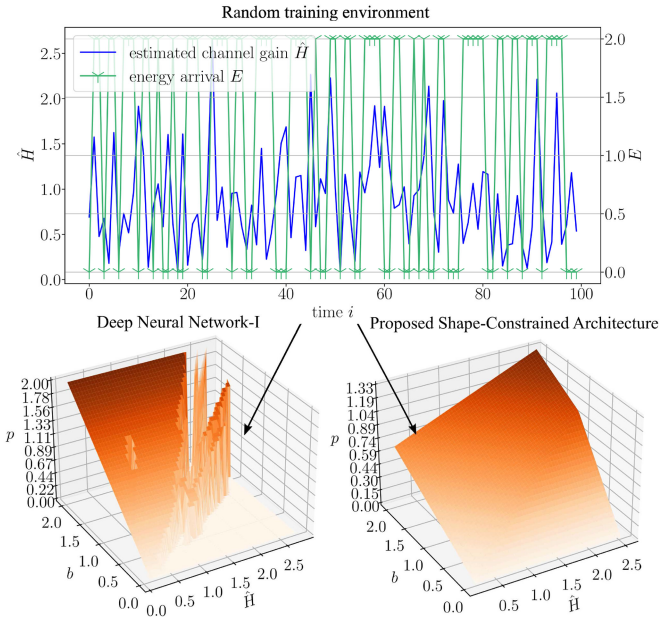


Fig. 4. Randomly generated training dataset (top) and power allocation policies after the training phase, obtained from the deep neural network (bottom left) and proposed shape-constrained network (bottom right). The uninterpretable structural properties, which an optimal policy should not have, emerge in the policy derived from the deep neural network built without deliberation.

B. Numerical Results

We name the proposed approach as the deterministic policy gradient method with the monotonic shape constraints (DPGMC). This technique is an online approach as the transmitter has no a priori information regarding the energy arrivals or channel gains. Fig. 4 illustrates the sample trajectory (training data sample) of the energy arrival and estimated channel gain. To demonstrate the overcomplexity of deep neural networks for the considered problem, deep neural networks with two hidden layers (256 and 128) were used for both the actor and critic networks, and the corresponding weights were initialized by a Gaussian distribution with zero mean and a standard deviation of

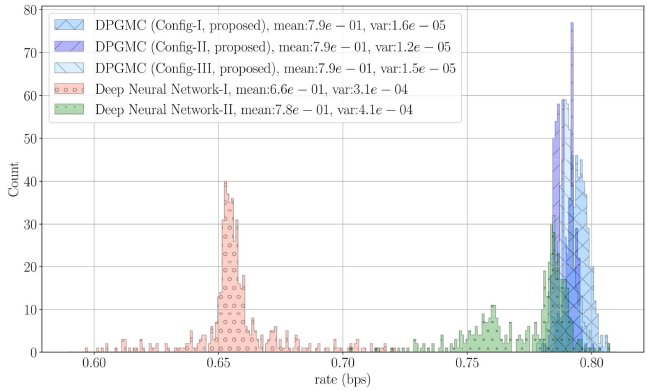


Fig. 5. Histograms of the performance of DPGMC and deep neural networks. In the harsh random environment, the randomness of the wireless communication environments and neural network initialization causes a large performance difference even if the same algorithm is used (results from deep neural networks). However, the proposed method demonstrates robustness in terms of the achievable rate with a low variance.

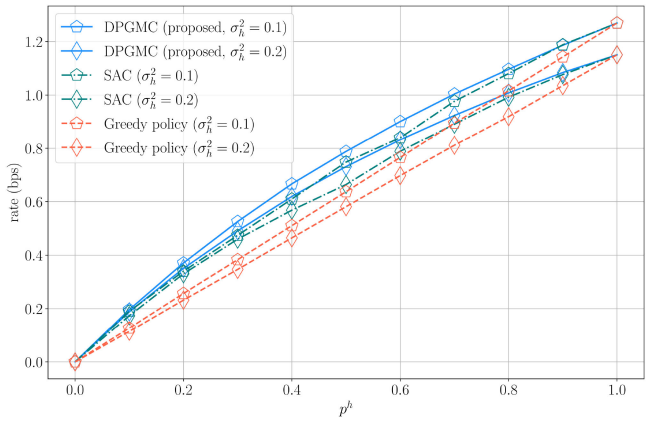


Fig. 6. Expected rate according to the harvesting probability p^h.

16. The power allocation policy obtained using this deep neural network is shown in the bottom left of Fig. 4. In the bottom right of Fig. 4, the power allocation policy after the training phase, as derived from the proposed architecture, is illustrated. In the harsh random communication environment involving a time-varying channel and energy arrival, the deep neural networks were overfitted to the partial data and thus could not represent the complete domain of the power allocation policy (bottom left). In contrast, the proposed shape-constrained network does not violate the optimal structural properties of the optimal policy. The overfitted policy from the deep neural network and the policy from the lattice network achieved values of approximately 0.68 bps and 0.79 bps, respectively. The objective of Fig. 4 is not to highlight that deep neural networks always lead to a lower performance, but to reflect that the optimal structural properties or bounds of the desired function must be accompanied at the application level. Figure 5 provides the statistical observation of the effect of the vulnerability of the overly complex neural networks to the overfitting problem in terms of the learning results. Specifically, Fig. 5 shows the histograms of the final performance (achieved rate) achieved from several actor-critic architectures. To evaluate the stability of the learning algorithm for different actor-critic architectures, we tested the same structure 500 times with random initializations. We evaluate the performance of our proposed actor-critic architecture for three different settings. The first setting is named Config-I in Fig. 5, and we set C_d to 8 for the calibration layer of the actor and $C_d^Q = 16$ for the calibration layer of the critic. In addition, we set $C_d = 8$ and $C_d^Q = 8$ for Config-II and we set $C_d = 16$ and $C_d^Q = 16$ for Config-III. Note that the greater C_d or C_d^Q are, the greater representational capability the function approximator contains. Fig. 5 shows that the performance of the implemented architectures does not change more than 0.05 bps in 500 tests, demonstrating the robustness of proposed method. Although the representational capability of the architectures is changed due to the change in the size of C_d or C_d^Q, they still contain the monotonicity constraints. Due to these constraints, the average value of each achieved rate distribution is approximately the same, even if the experiment is performed multiple times with different settings. It can also be observed that the variance of the achieved rate distribution, which is obtained from all the implemented DPGMC, is 1.6×10^{-5} or less.

Contrasting experimental results are observed when using deep artificial neural networks designed without considering system models. Deep neural network-I, which involved fully connected layers with two hidden layers, as shown in Fig. 4, was tested 500 times. The deep neural networks and weight initializer were designed without considering the system model, and these frameworks exhibited considerably different performances in each experiment (0.59 to 0.725 bps). The Glorot initialization method [34], which is a widely used initialization method in the learning theory field, was implemented for deep neural network-II. However, even when the same method was applied to the deep neural networks, and the average performance was improved, the performance corresponding to the setting ranged from 0.70 to 0.80 bps, and the variance of the achievable performance distribution compared to the previous setting was only slightly reduced. In contrast, the performance achieved using the proposed framework exhibited high mean values and low variances during the hundreds of experiments (0.78 to 0.81 bps). These experimental results are observed in all of Config-I, II, and III, which confirms the robustness of our proposed actor-critic architecture. As shown in Fig. 5, the variances of the achieved rate distributions of DPGMC are 1.6×10^{-5}, 1.2×10^{-5}, and 1.5×10^{-5}, for Config-I, II, and III, respectively, which are smaller than the variance of the achieved rate distribution from untailored deep neural networks. This finding clearly demonstrates that the function approximator design based on the structural properties considerably enhances the learning stability and ultimate average performance. Fig. 6 illustrates the comparison of the performance of the proposed approach and greedy approach, to clarify the rate increase that can be obtained through long-term scheduling. Since the experiment proceeds in a discrete environment that is nearly continuous, we adopt the Soft-Actor-Critic method (SAC)[35] for performance comparison, which is a state-of-the-art model-free reinforcement learning algorithm. Based on the gradient-based

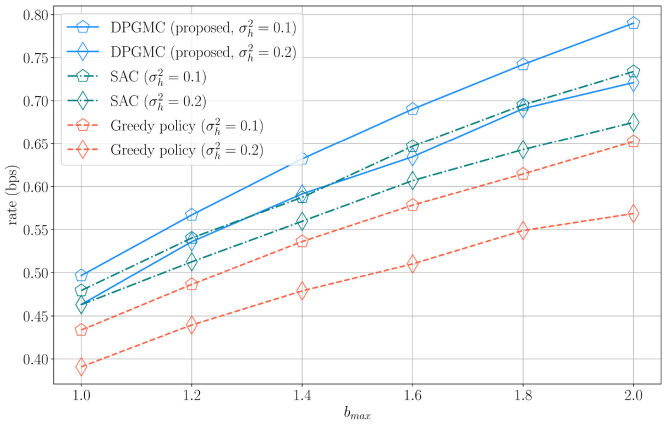


Fig. 7. Expected rate according to the battery capacity $b_{\max}$.

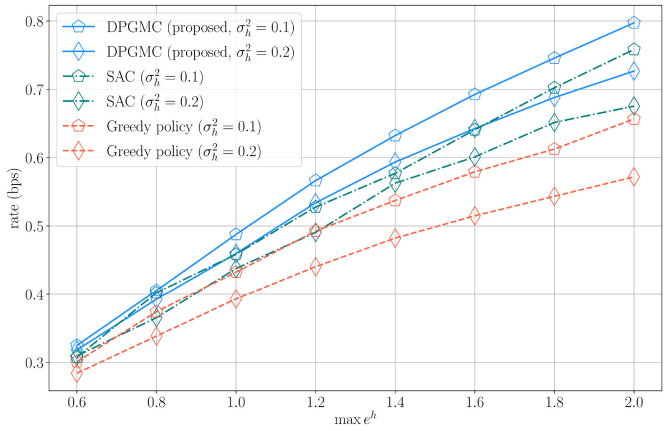


Fig. 8. Expected rate according to the maximum harvestable energy $\max E$.

reinforcement learning method and the soft MDP [35], SAC learns the stochastic policy that also maximizes the entropy of the policy with the expected sum of rewards for exploration. The implementation details are provided in Appendix C. In addition, we test a greedy policy on the environment, which consumes $p_i = b_i/T$ at time slot i to achieve the maximum rate for time slot i. We use this greedy policy as a lower bound of the performance. Higher p^h values indicate that the transmitter may have a chance to obtain energy more often. All the approaches exhibited an enhanced performance as p^h increased. Regardless of the assumed channel estimation error, the proposed approach outperformed the SAC and greedy method. The performance of SAC achieves between the performance of DPGMC and the greedy policy. When $p^h = 0$, all the algorithms operated in an environment in which no information could be transmitted, and thus, they exhibited the same performance with 0 bps. $p^h = 1.0$ refers to a situation in which the energy equal to the battery level is harvested in every time slot. In this case, because it was optimal to consume all the energy each time, all the algorithms achieved the same performance. Moreover, as shown in Fig. 7, a larger battery capacity corresponds to a larger achievable expected rate. All the tested algorithms achieved lower rates as the variance of the channel estimation error was increased. In all cases, we observe that DPGMC achieves better performance than SAC and the greedy policy. Figure 8 shows that a larger harvested energy corresponds to a greater achievable expected rate. Similar to the aspects observed in Figs. 6 and 7, the long-term scheduling approaches including DPGMC and SAC achieved a higher rate than that achieved by the greedy policy.

VI. Conclusion

In this study, the power allocation problem for rate maximization has been addressed using RL. We have proposed an online power allocation policy with a deterministic policy gradient by using function approximators. We have mainly proved that the deterministic online optimal policy, which adopts the harvested energy, estimated channel gain, and remaining battery as the input features, is a monotonically increasing function. We have also shown that the optimal action-value function is an increasing function for certain input features through the convergence property of the value iteration algorithm. By leveraging the optimized function approximators, we can avoid using overly complex neural networks, which are typically constructed in a heurisitc manner without considering the system model. In the simulations, the deep neural networks violated the structural properties that an optimal policy should possess, whereas when using the proposed method, robust learning could be realized with the optimal structural properties even in harsh random environments. The numerical results demonstrated that the proposed approach using the optimal structural properties could outperform the existing methods and improve the achievable rate in long-term operation.

It is important to note that, when implementing learning-theoretic approaches for optimization problems in communication systems, the function approximator designs must be carefully considered. Using the proposed method, the instability and problems associated with the large computation load caused by the use of deep neural networks, which have been highlighted in several existing studies, can be overcome. Future directions of this study involve analyzing performance guarantees for the tailored policy approximators. Moreover, the proposed approach can be applied to more complex and advanced communication system models including non-orthogonal multiple access-based broadcast channels that require explainable artificial intelligence technologies.

Appendix A
Proof of Lemma 5

We assume that $b' > b$ for all $b,' b \in \mathcal{B}$ and $p' > p$ for all $p,' p \in \mathcal{P}$. We define F_b as follows.

$$F_b(b,p) = I(s,p) + \gamma \mathbb{E}_{\bar{E},\bar{\bar{H}}} \left[V^*(\bar{E}, \bar{\bar{H}}, m(b,p,E)) \right]. \quad (40)$$

The controllable parameters for F_b are b and p, and the other variables are fixed to demonstrate the increasing property of b. Topkis' monotonicity theorem [30, Th. 1] is adpoted to show the increasing property of the optimal policy for b by deriving the

increasing difference of F_b, which is equivalent to

$$\gamma \mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,'p,'E))\right] - \gamma \mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,'p,E))\right] \quad (41)$$

$$\geq \gamma \mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,p,'E))\right] - \gamma \mathbb{E}_{\bar{E},\bar{\hat{H}}}\left[V^*(\bar{E},\bar{\hat{H}},m(b,p,E))\right]. \quad (42)$$

We can show that the inequality between (41) and (42) holds by following the proof in [12, Th. 2], which was used in [17], [24]; we provide the further proof following the logic in [12, Th. 2] for readability. According to the definition of m, b and p in (41) and (42) can be combined to form a single variable, and the inequality can be simplified as

$$V^*(\bar{E},\bar{\hat{H}},b'-p'+E) - V^*(\bar{E},\bar{\hat{H}},b-p'+E) \geq V^*(\bar{E},\bar{\hat{H}},b'-p+E) - V^*(\bar{E},\bar{\hat{H}},b-p+E). \quad (43)$$

Inequality (43) indicates that demonstrating the concavity of V^* for the last input variable b is sufficient to highlight the increasing difference of F_b, because $b'-p'+E > b-p'+E$, $b'-p+E > b-p+E$, and (43) is the difference in the rate change of V^* according to $b'-b$.

Lemma 7: The optimal value function V^* is a concave function of b for any given E and $\hat{H}$.

Proof: See Appendix B. ■

Next, we demonstrate that the lower and upper bounds of the action space $\mathcal{P}(s)$ are increasing functions for b. $\mathcal{P}(s)$ can be represented as $[0,b]$, and both the lower bound 0 and upper bound b can be interpreted as increasing functions for b. Subsequently, Topkis' monotonicity theorem can be adopted to demonstrate the increasing property of the optimal policy, and the remaining proof is omitted due to its similarity with the proof of Lemma 4.

Appendix B
Proof of Lemma 7

We consider mathematical induction to prove the concavity of the optimal value function V^* at b and the proof structure is similar to that of [12, Lemma. 3]. Let us assume that V_n is nondecreasing and concave for the remaining battery b. Next, we examine whether the concavity of the value function for b is preserved even after being updated through the value iteration. For b_1 and b_2 ($b_1 \neq b_2$ and $b_1, b_2 \in \mathcal{B}$), p_1 and p_2 are the optimal actions for the remaining battery levels b_1 and b_2, respectively as

$$p_1 = \underset{p\in\mathcal{P}(s_1)}{\arg\max}\left\{I(s_1,p) + \gamma \mathbb{E}\left[V_n(\bar{E},\bar{\hat{H}},m(b_1,p,E))\right]\right\}, \quad (44)$$

$$p_2 = \underset{p\in\mathcal{P}(s_2)}{\arg\max}\left\{I(s_2,p) + \gamma \mathbb{E}\left[V_n(\bar{E},\bar{\hat{H}},m(b_2,p,E))\right]\right\}. \quad (45)$$

The second derivative of function $I(s,p)$ for p is obtained in the interval $[0, b_{\max}]$ as follows.

$$\frac{\partial^2 I(\hat{H},p)}{\partial p^2} = \frac{-\hat{H}\sigma^2(2\mathbb{E}[|e|^2](\mathbb{E}[|e|^2]p+\sigma^2+\hat{H}p)+\sigma^2\hat{H})}{\ln 2((\mathbb{E}[|e|^2])p+\sigma^2)^2(\mathbb{E}[|e|^2]p+\sigma^2+\hat{H}p)^2}. \quad (46)$$

As the numerator and denominator in (46) are negative and positive values, respectively, for any $\mathbb{E}[|e|^2](>0)$ and $\sigma^2(>0)$, I is a concave function of p as

$$\lambda I(s,p_1) + (1-\lambda)I(s,p_2) \leq I(s,p_\lambda) \quad (47)$$

where $p_\lambda = \lambda p_1 + (1-\lambda)p_2$ for $0<\lambda<1$. According to the assumption and Lemma 3, the nondecreasing property of V_n is preserved through the value iteration and m is a concave function for b. Notably, $V_n(\bar{E},\bar{\hat{H}},m(b_1,p,E))$ is concave for b_1-p because V_n is a nondecreasing and concave function, and m is a concave function. By using the definition of the value iteration and the concavity of I and V_n, we can obtain

$$\lambda V_{n+1}(E,\hat{H},b_1) + (1-\lambda)V_{n+1}(E,\hat{H},b_2)$$

$$= \lambda I(s,p_1) + \lambda\gamma \mathbb{E}_{\bar{s}}\left[V_n(\bar{E},\hat{\bar{H}},m(b_1,p_1,E))\right] + (1-\lambda)I(s,p_2) \quad (48)$$

$$+ (1-\lambda)\gamma \mathbb{E}_{\bar{s}}\left[V_n(\bar{E},\hat{\bar{H}},m(b_2,p_2,E))\right] \quad (49)$$

$$\leq I(s,p_\lambda) + \gamma \mathbb{E}_{\bar{s}}\left[V_n(\bar{E},\hat{\bar{H}},m(b_\lambda,p_\lambda,E))\right] \quad (50)$$

$$\leq \max_{0\leq p\leq b_\lambda}\left\{I(s,p) + \gamma \mathbb{E}_{\bar{s}}\left[V_n(\bar{E},\hat{\bar{H}},m(b_\lambda,p_\lambda,E))\right]\right\}. \quad (51)$$

The inequality between (50) and (51) holds because the optimal action that satisfies the maximum operator in (51) ranges in $[0,b_\lambda]$ and $p_\lambda \leq b_\lambda$. The inequality between (48) and (51) indicates that the concavity of the value function is preserved during the value iteration update. The value iteration converges to the optimal point regardless of the initial function V_0. We can initialize the value function V_0 as a nondecreasing concave function for b, whereas the optimal value function $\lim_{n\to\infty} V_n = V^*$ is still a concave function for b.

Appendix C
Soft-Actor-Critic Implementation for Performance Comparison

This appendix is provided as a summary of the SAC implementation in [35]. All formulas and notation are directly adopted from [35]. The SAC algorithm aims at maximizing the maximum entropy objective function with a stochastic policy. Policy, value function, and action-value function are required to implement the algorithm. Since the algorithm deals with soft MDP [36], the objective function is defined with the expected entropy of the power allocation policy as follows.

$$J(\pi^s) = \sum_{i=0}^{T_o} \mathbb{E}_{s_t,p_t}[I(s_i,p_i) + \alpha^s \mathscr{H}(\pi^s(\cdot|s_t))] \quad (52)$$

where π^s is stochastic policy, $\mathscr{H}$ is entropy operator, and α^s is a parameter that indicates the importance of the entropy of the policy. The soft value function, V^s, is defined as follows:

$$V^s(s) = \mathbb{E}[Q^s(s,p) - \log \pi^s(p|s)]. \tag{53}$$

where $Q^s(s,p)$ is the soft action-value function. By the set of parameters, ψ, the soft value function is approximated as V^s_ψ and the estimated gradient can be calculated as follows:

$$\hat{\nabla}_\psi J_{V^s}(\psi) = \nabla_\psi V^s_\psi(s)(V^s_\psi(s) - Q_{\theta^s}(s,p) + \log \pi^s_{\phi^s}(a|s)) \tag{54}$$

where the θ^s is the set of parameters for the soft action-value function approximation and ϕ^s is the set of parameters for the soft policy $\pi^s_{\phi^s}(a|s)$. The soft action-value function also updated in the direction of minimizing mean squared error as follows:

$$J_{Q^s}(\theta^s) = \mathbb{E}[\frac{1}{2}(Q_{\theta^s}(s,p) - \hat{Q}_{\theta^s}(s,p))^2] \tag{55}$$

where $\hat{Q}_{\theta^s}(s,p)$ is a target soft action-value function which is defined as

$$\hat{Q}_{\theta^s}(s,p) = I(s,p) + \gamma \mathbb{E}_{\bar{s}}[V^s_{\bar{\psi}^s}(\bar{s})]. \tag{56}$$

The set of parameters $\bar{\psi^s}$ consists of the target soft value function and it is updated by τ^s-ratio every gradient update iteration. The stochastic policy can be updated by gradient update as

$$\begin{aligned}\hat{\nabla}_{\phi^s} J_{\pi^s}(\phi^s) &= \nabla_{\phi^s} \log \pi^s_{\phi^s}(p|s) \\ &+ (\nabla_p \log \pi^s_{\phi^s}(p|s) - \nabla_p Q_{\theta^s}(s,p)) \nabla_{\phi^s} f_{\phi^s}(\epsilon, s)\end{aligned} \tag{57}$$

where f_{ϕ^s} is a reparametrized policy. The learning rates for the soft policy, value function, and the action value function are set as 0.001, and all the function approximators are constructed as neural network with 256-256 sized hidden layers. Each hidden layer has ReLu activation functions for its output. The soft policy approximator also has 256-256 sized hidden layers with ReLu activation functions. The output of the neural network consists of two elements that include the mean of the power and the standard deviation of the power. To implement f_{ϕ^s}, we multiply a single variable that follows a Gaussian distribution, which has zero mean and unit variance, to the output representing the standard deviation of the power. The Gaussian stochastic policy is constructed based on the mean and the standard deviation values. α^s is set as 0.2 and τ^s is 0.01. The training phase of SAC is directly followed by Algorithm 1 in [35] and we halt the training phase when there is no performance improvement during 10 gradient-based updates.

REFERENCES

[1] 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.

[2] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, Mar. 2012.

[3] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4808–4818, Sep. 2012.

[4] P. He, L. Zhao, S. Zhou, and Z. Niu, "Recursive waterfilling for wireless links with energy harvesting transmitters," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1232–1241, Mar. 2013.

[5] C. Huang, R. Zhang, and S. Cui, "Throughput maximization for the gaussian relay channel with energy harvesting constraints," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1469–1479, Aug. 2013.

[6] J. Yang, O. Ozel, and S. Ulukus, "Broadcasting with an energy harvesting rechargeable transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 571–583, Feb. 2012.

[7] A. Baknina and S. Ulukus, "Optimal and near-optimal online strategies for energy harvesting broadcast channels," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3696–3708, Dec. 2016.

[8] F. Amirnavaei and M. Dong, "Online power control optimization for wireless transmission with energy harvesting and storage," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4888–4901, Jul. 2016.

[9] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.

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

[11] O. Ozel, J. Yang, and S. Ulukus, "Optimal broadcast scheduling for an energy harvesting rechargeable transmitter with a finite capacity battery," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2193–2203, Jun. 2012.

[12] S. Mao, M. H. Cheung, and V. W. Wong, "Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2862–2875, Jul. 2014.

[13] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.

[14] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," in *Proc. NIPS 13th Workshop Deep Learn.*, 2013, pp. 1–9.

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

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

[17] H. Kim *et al.*, "On the design of tailored neural networks for energy harvesting broadcast channels: A reinforcement learning approach," *IEEE Access*, vol. 8, pp. 179 678–179 691, 2020.

[18] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015. [Online]. Available: https://arxiv.org/abs/1509.02971

[19] H. Kim, H. Yang, Y. Kim, and J. Lee, "Action-bounding for reinforcement learning in energy harvesting communication systems," in *Proc. IEEE Glob. Commun. Conf.*, 2018, pp. 1–7.

[20] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (ddpg)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.

[21] M. Li, X. Zhao, H. Liang, and F. Hu, "Deep reinforcement learning optimal transmission policy for communication systems with energy harvesting and adaptive MQAM," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5782–5793, Jun. 2019.

[22] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*. Cambridge, MA, USA: MIT Press, 1994.

[23] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A study on overfitting in deep reinforcement learning," 2018. [Online]. Available: https://arxiv.org/abs/1804.06893

[24] H. Kim, T. Cho, W. Shin, J. Lee, and P. H. Vincent, "Optimized shallow neural networks for sum-rate maximization in energy harvesting downlink multiuser noma systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 982–997, Apr. 2021.

[25] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading mimo channels with channel estimation error," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[26] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: John Wiley Sons, 2012.

[28] A. Lapidoth and S. Shamai, "Fading channels: How perfect need "perfect side information" be?," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[29] D. M. Topkis, *Supermodularity and Complementarity*. Princeton, NJ, USA: Princeton Univ. Press, 1998.
[30] R. Amir, "Supermodularity and complementarity in economics: An elementary survey," *Southern Econ. J.*, vol. 71, no. 3, pp. 636–660, 2005.
[31] G. Dulac-Arnold *et al.*, "Deep reinforcement learning in large discrete action spaces," 2015. [Online]. Available: https://arxiv.org/abs/1512.07679
[32] M. Gupta *et al.*, "Monotonic calibrated interpolated look-up tables," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3790–3836, 2016.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–15.
[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[35] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
[36] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1352–1361.

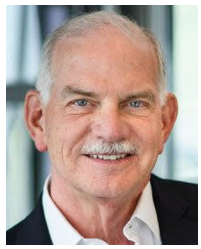
Heasung Kim (Member, IEEE) received the B.S. degree in computer and communication engineering from Korea University, Seoul, South Korea, in 2017, and the M.S. degree from the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, South Korea, in 2019.

From 2019 to 2021, he was an Engineer and a Data Scientist with the Samsung Network Business and with the Samsung Electronics Company Ltd., South Korea, where he was contributing to self-organizing networks with machine learning. He is currently a Researcher with the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea. His research interests include reinforcement learning, optimization, wireless communications, and energy harvesting systems. He was the recipient of the Best M.S. Dissertation Award from SNU, in 2019.

Jungwoo Lee (Senior Member, IEEE) received the B.S. degree in electronics engineering from Seoul National University, Seoul, South Korea, in 1988, and the M.S.E. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 1990 and 1994, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, Seoul National University. From 1994 to 1999, he was a Member of the Technical Staff working on multimedia signal processing with SRI (Sarnoff), Princeton, NJ, USA, where he was a Team Leader (PI) of an U.S. $18 million NIST ATP Program. Since 1999, he has been with the Wireless Advanced Technology Laboratory, Lucent Technologies Bell Labs, and was on WCDMA base station algorithm development, as a Team Leader. He holds 21 U.S. patents. His research interests include wireless communications, information theory, distributed storage, and machine learning. He was a TPC/OC Member of the ISITA 2005, ICC 2005, PIMRC 2008, ISIT 2009, ICC 2015, ITW 2015, and VTC 2015. He was the recipient of the Qualcomm Dr. Irwin Jacobs Award, in 2014, for his contributions in wireless communications. He was the co-recipient of the IEEE Communications Society Fred W. Ellersick Prize. From 2016 to 2017, he was the Track Chair of the IEEE ICC SPC, and in 2019, the General Chair of the JCCI. From 2008 to 2011, he was an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and from 2012 to 2016, the *Journal of Communications and Networks*. Since 2015, he has also been the Chief Editor of KICS journal and the Executive Editor of the *ICT Express* (Elsevier-KICS). Since 2017, he has been the Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS. He was the recipient of two Bell Labs technical achievement awards.

Wonjae Shin (Senior Member, IEEE) received the B.S. and M.S. degrees from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2005 and 2007, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, South Korea, in 2017. From 2007 to 2014, he was a Member of Technical Staff with the Samsung Advanced Institute of Technology and Samsung Electronics Company Ltd., South Korea, where he made contribution to next generation wireless communication networks, especially for 3GPP LTE/LTE-advanced standardizations. From 2016 to 2018, he was a Visiting Scholar and a Postdoctoral Research Fellow with Princeton University, Princeton, NJ, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea. Prior to joining Ajou University, he was a Faculty Member with Pusan National University, Busan, South Korea, from 2018 to 2021. His research interests include the design and analysis of future wireless communications, including interference-limited networks, and machine learning for wireless networks.

Dr. Shin was awarded the Fred W. Ellersick Prize and the Asia-Pacific Outstanding Young Researcher Award from the IEEE Communications Society, in 2020, the Best Ph.D. Dissertation Award from SNU, in 2017, the Gold Prize from the IEEE Student Paper Contest (Seoul Section), in 2014, and the Award of the Ministry of Science and ICT of Korea in IDIS-Electronic News ICT Paper Contest 2017. He was the co-recipient of the SAIT Patent Award, in 2010, the Samsung Journal of Innovative Technology, in 2010, the Samsung HumanTech Paper Contest, in 2010, and the Samsung CEO Award, in 2013. In 2014, he was recognized as an Exemplary Reviewer by the IEEE WIRELESS COMMUNICATIONS LETTERS, and in 2019, the IEEE TRANSACTIONS ON COMMUNICATIONS. He also was awarded several fellowships, including the Samsung Fellowship Program, in 2014, and the SNU Long Term Overseas Study Scholarship, in 2016. He is currently the Editor of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA, in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign, Urbana, IL, USA. Since 1990, he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor. From 2006 until 2016, he was the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at various other institutions, including most recently at Berkeley and Cambridge. His research interests include information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press, 2021).

Dr. Poor is a Member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences the Royal Society and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and the 2019 ASEE Benjamin Garver Lamme Award.